MMEDIA 2014

The Sixth International Conferences on Advances in Multimedia

ISBN: 978-1-61208-320-9

February 23 - 27, 2014

Nice, France

MMEDIA 2014 Editors

Pascal Lorenz, University of Haute Alsace, France

MMEDIA 2014

Foreword

The Sixth International Conferences on Advances in Multimedia (MMEDIA 2014), held between February 23rd-27th, 2014 in Nice, France, was an international forum for researchers, students, and professionals where to present recent research results on advances in multimedia, and mobile and ubiquitous multimedia. MMEDIA 2014 brought together experts from both academia and industry for the exchange of ideas and discussion on future challenges in multimedia fundamentals, mobile and ubiquitous multimedia, multimedia ontology, multimedia user-centered perception, multimedia services and applications, and mobile multimedia.

The rapid growth of information on the Web, its ubiquity and pervasiveness, makes the www the biggest repository. While the volume of information may be useful, it creates new challenges for information retrieval, identification, understanding, selection, etc. Investigating new forms of platforms, tools, principles offered by Semantic Web opens another door to enable human programs, or agents, to understand what records are about, and allows integration between domain-dependent and media-dependent knowledge. Multimedia information has always been part of the Semantic Web paradigm, but it requires substantial effort to integrate both.

The new technological achievements in terms of speed and the quality expanded and created a variety of multimedia services such as voice, email, short messages, Internet access, m-commerce, mobile video conferencing, streaming video and audio.

Large and specialized databases together with these technological achievements have brought true mobile multimedia experiences to mobile customers. Multimedia implies adoption of new technologies and challenges to operators and infrastructure builders in terms of ensuring fast and reliable services for improving the quality of web information retrieval.

Huge amounts of multimedia data are increasingly available. The knowledge of spatial and/or temporal phenomena becomes critical for many applications, which requires techniques for the processing, analysis, search, mining, and management of multimedia data.

We take here the opportunity to warmly thank all the members of the MMEDIA 2014 Technical Program Committee, as well as the numerous reviewers. The creation of such a high quality conference program would not have been possible without their involvement. We also kindly thank all the authors who dedicated much of their time and efforts to contribute to MMEDIA 2014. We truly believe that, thanks to all these efforts, the final conference program consisted of top quality contributions.

Also, this event could not have been a reality without the support of many individuals, organizations, and sponsors. We are grateful to the members of the MMEDIA 2014 organizing committee for their help in handling the logistics and for their work to make this professional meeting a success.

We hope that MMEDIA 2014 was a successful international forum for the exchange of ideas and results between academia and industry and for the promotion of progress in the field of multimedia.

We are convinced that the participants found the event useful and communications very open. We also hope the attendees enjoyed the charm of Nice, France.

MMEDIA Advisory Committee:

Dumitru Dan Burdescu, University of Craiova, Romania

MMEDIA 2014

Committee

MMEDIA Advisory Committee

Dumitru Dan Burdescu, University of Craiova, Romania
Philip Davies, Bournemouth and Poole College, UK
Jean-Claude Moissinac, TELECOM ParisTech, France
David Newell, Bournemouth University, UK
Francisco J. Garcia, Agilent Technologies - Edinburgh, UK
Noël Crespi, Institut Telecom, France
Jonathan Loo, Middlesex University - Hendon, UK
Patrice Rondao Alface, Alcatel-Lucent Bell Labs - Antwerp, Belgium
Trista Chen, Fotolugu Inc, USA
Alexander C. Loui, Kodak Research Labs / Eastman Kodak Company-Rochester, USA

MMEDIA 2014 Technical Program Committee

Max Agueh, LACSC - ECE Paris, France
Hakiri Akram, Université Paul Sabatier - Toulouse, France
Musab Al-Hadrusi, Wayne State University, USA
Nancy Alonistioti, N.K. University of Athens, Greece
Giuseppe Amato ISTI-CNR, Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo" - Pisa, Italy
Maria Teresa Andrade, University of Porto / INESC Porto, Portugal
Marios C. Angelides, Brunel University - Uxbridge, UK
Stylianos Asteriadis, Centre for Research and Technology - Information Technologies Institute (CERTH-ITI), Greece
Ramazan S. Aygun, University of Alabama in Huntsville, USA
Elias Baaklini, University of Valenciennes, France
Andrew D. Bagdanov, Universita Autonoma de Barcelona, Spain
Yannick Benezeth, Université de Bourgogne - Dijon, France
Jenny Benois-Pineau, LaBRI/University of Bordeaux 1, France
Sid-Ahmed Berrani, Orange Labs - France Telecom, France
Steven Boker, University of Virginia - Charlottesville, USA
Fernando Boronat Seguí, Universidad Politecnica de Valencia, Spain
Laszlo Böszörmenyi, University Klagenfurt, Austria
Hervé Bredin, CNRS/LIMSI, France
Marius Brezovan, University of Craiova, Romania
Dumitru Burdescu, University of Craiova, Romania
Helmar Burkhart, Universität Basel, Switzerland
Nicola Capuano, University of Salerno, Italy
Eduardo Cerqueira, Federal University of Para, Brazil
Damon Chandler, Oklahoma State University, USA
Vincent Charvillat, ENSEEIHT/IRIT - Toulouse, France

Mike Matton, VRT research & innovation – Brussel, Belgium
Annett Mitschick, Technical University - Dresden, Germany
Ayman Moghnieh, Universitat Pompeu Fabra - Barcelona, Spain
Manoranjan Mohanty, National University of Singapore, Singapore
Jean-Claude Moissinac, TELECOM ParisTech, France
Mario Montagud Climent, Universidad Politecnica de Valencia, Spain
Mireia Montañola, Université catholique de Louvain, Belgium
Michele Nappi, Universita` di Salerno – Fisciano, Italy
David Newell, Bournemouth University, UK
Petros Nicopolitidis, Aristotle University of Thessaloniki, Greece
Vincent Oria, New Jersey Institute of Technology, USA
Jordi Ortiz Murillo, University of Murcia, Spain
Marco Paleari, Italian Institute of Technology / Center for Space Human Robotics - Torino, Italy
Sethuraman Panchanathan, Arizona State University, USA
Eleni Patouni, University of Athens, Greece
Tom Pfeifer, Technical University of Berlin, Germany
Salvatore F. Pileggi, University of Auckland, New Zealand
Key Pousttchi, University of Augsburg, Germany
Wei Qu, Graduate University of Chinese Academy of Sciences, China
Piotr Romaniak, AGH University of Science and Technology - Krakow, Poland
Patrice Rondao Alface, Alcatel-Lucent Bell Labs - Antwerp, Belgium
Ali Salah, Bogazici University, Turkey
Susana Sargento, University of Aveiro/Institute of Telecommunications, Portugal
Oliver Schreer, Fraunhofer Heinrich-Hertz-Institute, Germany
Christine Senac, IRIT laboratory, France
Peter L. Stanchev, Kettering University - Flint, USA
Liana Stanescu, University of Craiova, Romania
Cosmin Stoica, University of Craiova, Romania
Yu Sun, University of Central Arkansas, USA
Siyu Tang, Alcatel-Lucent Bell Labs, Belgium
Anel Tanovic, BH Telecom d.d. Sarajevo, Bosnia and Herzegovina
Tsutomu Terada, Kobe University, Japan
Georg Thallinger, Joanneum Research - Graz, Austria
Daniel Thalmann, EPFL, Switzerland
Christian Timmerer, Alpen-Adria-Universität Klagenfurt, Austria
Chien-Cheng Tseng, National Kaohsiung First University of Science and Technology, Taiwan
Kuniaki Uehara, Kobe University, Japan
Andreas Uhl, Salzburg University, Austria
Binod Vaidya, Instituto de Telecomunicações / University of Beira Interior, Portugal
Andreas Veglis, Aristotle University of Thessaloniki, Greece
Janne Vehkaperä, VTT Technical Research Centre of Finland - Oulu, Finland
Dimitrios D. Vergados, University of Piraeus, Greece
Anne Verroust-Blondet, INRIA Paris-Rocquencourt, France
Marie-Luce Viaud, French National Institute for Audiovisual (INA), France
Giuliana Vitiello, University of Salerno – Fisciano, Italy
Lei Ye, University of Wollongong, Australia
Shigang Yue, University of Lincoln, UK
Sherali Zeadally, University of Kentucky, USA

Tong Zhang, Hewlett-Packard Labs, USA
Yang Zhenyu, Florida International University, USA

Copyright Information

For your reference, this is the text governing the copyright release for material published by IARIA.

The copyright release is a transfer of publication rights, which allows IARIA and its partners to drive the dissemination of the published material. This allows IARIA to give articles increased visibility via distribution, inclusion in libraries, and arrangements for submission to indexes.

I, the undersigned, declare that the article is original, and that I represent the authors of this article in the copyright release matters. If this work has been done as work-for-hire, I have obtained all necessary clearances to execute a copyright release. I hereby irrevocably transfer exclusive copyright for this material to IARIA. I give IARIA permission or reproduce the work in any media format such as, but not limited to, print, digital, or electronic. I give IARIA permission to distribute the materials without restriction to any institutions or individuals. I give IARIA permission to submit the work for inclusion in article repositories as IARIA sees fit.

I, the undersigned, declare that to the best of my knowledge, the article is does not contain libelous or otherwise unlawful contents or invading the right of privacy or infringing on a proprietary right.

Following the copyright release, any circulated version of the article must bear the copyright notice and any header and footer information that IARIA applies to the published article.

IARIA grants royalty-free permission to the authors to disseminate the work, under the above provisions, for any academic, commercial, or industrial use. IARIA grants royalty-free permission to any individuals or institutions to make the article available electronically, online, or in print.

IARIA acknowledges that rights to any algorithm, process, procedure, apparatus, or articles of manufacture remain with the authors and their employers.

I, the undersigned, understand that IARIA will not be liable, in contract, tort (including, without limitation, negligence), pre-contract or other representations (other than fraudulent misrepresentations) or otherwise in connection with the publication of my work.

Exception to the above is made for work-for-hire performed while employed by the government. In that case, copyright to the material remains with the said government. The rightful owners (authors and government entity) grant unlimited and unrestricted permission to IARIA, IARIA's contractors, and IARIA's partners to further distribute the work.

Table of Contents

Priority-Based Routing Framework for Multimedia Delivery in Surveillance Networks

Adil A Sheikh
Science and Technology Unit
Simplicity Labs
Umm Al Qura University
Makkah, Kingdom of Saudi Arabia
aasheikh@uqu.edu.sa

Emad Felemban, Saleh Basalamah
Department of Computer Engineering
Simplicity Labs
Umm Al Qura University
Makkah, Kingdom of Saudi Arabia
{eafelemban, smbasalamah}@uqu.edu.sa

***Abstract*—Wireless sensor network consisting of nodes equipped with cameras or advanced low-cost image sensors is known as a Visual Sensor Networks (VSN). The main function of VSNs is to capture images and send them to sink nodes for processing. One of the most common applications of VSN is surveillance. Such applications require large amounts of data to be exchanged between camera nodes and sink. Image data is considerably larger than common sensor data such as temperature, humidity, pressure, etc. For data delivery in VSNs, the communication is constrained by many stringent QoS requirements like delay, jitter and data reliability. Moreover, due to the inherent constraints of wireless sensor networks such as low energy, limited CPU power and scarce memory, the architect of VSN must choose appropriate topology, image compression algorithms and communication protocols depending on his/her application. This paper focuses on one of these aspects, namely the communication protocol for VSN. In this paper, we present a new routing framework for VSN to deliver critical imagery information with system's time constraint. We have implemented our proposed framework using Contiki and simulated it on Cooja simulator to support our claim.**

Keywords-Routing Framework; VSN; Image Transmission; Priority-Based Routing; Contiki; Cooja

I. INTRODUCTION

The primary requirement of a wireless sensor network is to sense environment factors using low-power, low-cost sensors and route meaningful data to power-rich sink nodes for processing. This requirement becomes challenging in a VSN as the amount of data to be transferred is much more than a traditional wireless sensor network due to the type of data being shared. Applications of surveillance require very large amounts of data to be exchanged between camera nodes and sink. In traditional wireless sensor networks that sense light, humidity, pressure, etc. the traffic generated by a sensing node is limited to the scalar data [1]. In most cases, the memory size required to store and send is 16-bits per reading [1]. On the other hand, a VSN node, equipped with a camera generates vector data. For instance, a raw Red-Green-Blue (RGB) image of 128 x 128 pixels with 24-bits per pixel (8 bits per color) will be of 128 x 128 x 24 = 393216 bits (approximately 48 kilobytes). These are magnitudes larger than traditional sensor data.

To minimize the size of the image data, image compression techniques such as Discrete Cosine Transforms [2] or Discrete Wavelet Transforms [3][4] can be used. Although these algorithms reduce the size of an image, yet it is not comparable to traditional sensors data. Therefore, image data compression is not enough. The processing power of each node is also limited. Additionally, the topology of the network and routing protocols play a crucial role in transporting imagery information from visual sensing nodes to sink nodes. Hence, the tasks of capturing image data, compressing it and sending it to sink are some of the most challenging tasks faced by VSN architects.

As mentioned before, using image compression algorithms the size of data can be reduced to some extent. Also, a category of image compression algorithms generate multiple layers of compressed image data. The first layer contains the most prominent features of the image, for example, the edges of objects or coarse image data. The subsequent layers contain the details that when merged with the first layer, restore the original image. Some image processing algorithms consist of multiple passes requiring different levels of details of the encoded image for each pass. Using such algorithms in VSNs, system response time can be reduced. If the sink nodes receive image data required for first pass sooner than data required for subsequent passes, it can start processing the first pass and take action accordingly while data of subsequent layers arrive at the sink node. This paper helps alleviate the routing challenges of such image processing algorithms by proposing a routing framework based on four features. (1) The visual sensing nodes should be able to specify priority to outgoing packets. In this way, image data for first pass can be sent at higher priority than data for subsequent passes. (2) The intermediate or routing nodes should be aware of packet priority so that higher priority packets are forwarded before lower priority packets. (3) If packets from two nodes collide, high priority packets should be retransmitted before low priority packets. (4) Finally, in event of congestion, lower priority packets should be dropped before any high priority packet is dropped.

The next section summarizes the various communication protocols being used in VSN architectures as of today. Section III discusses typical VSN application scenario along with details of VSN components essential for delivering critical image information within system's time constraints.

 ISBN: 978-1-61208-320-9

Section IV defines our proposed priority-based routing framework. The implementation of our proposed protocol is discussed in Section V. Simulations were carried out to quantify the usefulness of the routing framework. In Section VI, simulation environment and results are discussed. Finally, the paper is concluded along in Section VII.

II. EXISTING ROUTING TECHNIQUES

The research on routing techniques for image transmission has mostly been limited to wired networks [5]-[9]. Research on QoS supported routing protocols for mobile ad-hoc networks has been summarized by Chen et al. [10] and Hanzo-II et al. [11]. Liebeherr et al. [12], Wang et al. [13], Stoica et al. [14], Younis et al. [15] and Soldatos et al. [16] discuss techniques to deliver image data on the Internet. None of these are applicable to VSNs.

Most of the work done in the field of routing techniques for VSNs has been conducted to achieve energy efficiency. The first routing protocol focused on QoS in VSNs by trying to minimize the average weighted QoS metric throughout the lifetime of the network. Sohrabi et al. [17] proposed Sequential Assignment Routing (SAR) that enforces maintenance of routing tables with status of all nodes.

RAP [18] is a priority-based routing protocol that uses velocity monotonic scheduling and geographical forwarding to achieve QoS, however, its requirement of geographical awareness can only be fulfilled by having a pre-defined network topology or additional hardware to determine geographical location.

SPEED [19], proposed by He et al., is a spatio-temporal, priority-based, QoS-aware routing protocol for sensor networks that provides soft real-time, end-to-end delay guarantees. SPEED does not provide differentiated packet prioritization. Moreover, a forwarding node can only forward the packet at a speed less than or equal to the maximum achievable speed even though the network can support it.

Real-time Power-Aware Routing (RPAR) [20] is another routing protocol that achieves application specific end-to-end delay guarantee at low power by dynamically adjusting transmission power and routing decisions based on the workload and packet deadlines. RPAR also calculates average link quality taking link variability into consideration.

Multi-path and Multi-SPEED (MMSPEED) routing protocol [21] supports probabilistic QoS guarantee by provisioning QoS in two domains, timeliness and reliability. MMSPEED adopts a differentiated priority packet delivery mechanism in which QoS differentiation in timeliness is achieved by providing multiple network-wide packet delivery speed guarantees.

III. VSN APPLICATION SCENARIO

This section explains the VSN application scenario discussed in this paper. In a typical VSN application, there are three types of nodes that make progressive image transmission possible. A brief description of each VSN node type and our network model is given below.

A. Visual Sensing Node

The visual sensing node contains the sensor that captures images. Depending on the application this sensor can be of type that captures multi-colored images, grey-scale images, thermal or infra-red images [22], etc. Nodes equipped with these sensors require more power to run additional hardware and software components such as frame-grabbers and image encoders. These nodes capture raw images, encode them and send them towards sink nodes for processing.

B. Intermediate Node

Their primary task of intermediate nodes is to send packets from camera nodes to the destination sink node. Depending on the VSN application, these nodes may also take part in sensing other scalar environmental variables such as temperature, humidity, pressure, level of certain chemicals, etc. Additionally, these nodes may also take part in encoding image data as a class of image encoding algorithms [23] offloads some processing to intermediate nodes in order to conserve power of camera nodes.

C. Sink Node

The sink nodes are responsible for processing the images captured by the camera nodes. For this purpose, sink nodes are power-rich and have high computation ability. In order to take action depending on the VSN application, these nodes may additionally contain actuators or may be connected to a fourth type of nodes called actuator nodes.

D. Network Model

The network model discussed in this paper does not restrict the number or position of any node type. One of the network topologies for a surveillance application is depicted in Fig. 1. For purposes of testing and evaluation, the network model we have used in this paper consists of one-quarter of this topology. Our visual sensing nodes are placed on the periphery of the network. The intermediate nodes are placed in the bulk of the network. In our network model, there is only one sink. It is also placed in the periphery of the network, on the opposite side of visual sensing nodes. This is depicted by the dashed-line in Fig. 1.

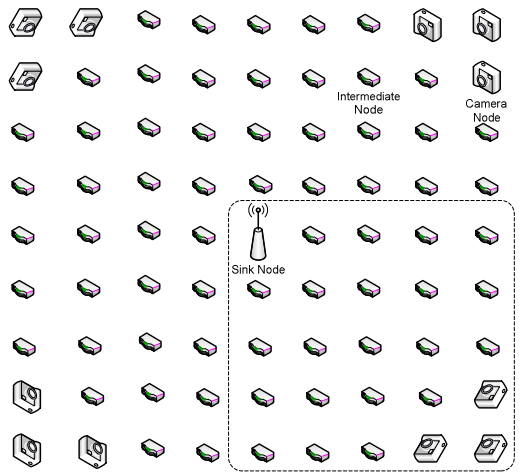


Figure 1. An Example of a Surveillance Network using VSN

In our scenario, the visual sensing nodes take images and encode them into two layers. First layer contains coarse image information collected in the first pass of image encoding and second layer contains fine image information collected in the second pass. The VSN application uses the routing framework to send this layer with high priority towards the sink. The sink uses first pass information to reconstruct the encoded image with a certain level of detail. Based on processing first pass, the sink can take action, if necessary. The VSN application uses routing framework to send second pass layer at low priority towards the sink. The sink uses the second pass layer to reconstruct a detailed image for further processing if image information from the first pass required additional image details to take action.

Most of the nodes in our network are the intermediate nodes. They are only responsible for routing packets from visual sensing nodes to sinks. They do not take part in sensing or sharing processing load of the sensing nodes or sink nodes. When the network is deployed, the intermediate nodes create routing tables that are necessary to take routing decision when packets are received. To achieve their primary task of routing image data from visual sensing nodes to sink nodes, the routing tables in intermediate nodes are updated throughout the lifetime of the network as some nodes may die due to depleted power or other environmental conditions, while other nodes may be added to the network when required. The routing framework makes sure that intermediate nodes forward high priority packets (first pass image layer) faster than low priority packets (second pass image layer). This way, the routing framework facilitates sink nodes to reconstruct first pass image much sooner than when the entire image data is received at sink. As required, the sink node can add the second pass information to the first pass to construct a more detailed image.

IV. PRIORITY-BASED ROUTING FRAMEWORK FOR VSN

This section provides detail of how the priority-based routing framework works. The framework is distributed into network layer and medium access control layer of any protocol stack. Additionally, a thin Application Interface Layer (AIL) encapsulates the details of network layer and medium access control layer. Functional details of these layers are provided in the sub-sections below.

A. Application Interface Layer

The AIL (Application Interface Layer) is the application layer component of priority-based routing framework. It is a very thin layer that provides VSN application with a set of primitives that can be used for fragmenting image data into packets, sending them, receiving them and assembling them to re-generate image data. The AIL hides the implementation details of the entire framework. The VSN application passes image data along with its priority to the routing framework using the AIL. Based on its configuration, AIL of the sending node fragments the image data into packets of size that network layer can send. AIL also inserts image number and packet fragment number into the packet. This information is used by the AIL of sink node to join the fragments to construct image data sent.

B. Network Layer

The network layer component of priority-based routing framework works in two phases explained below.

1) Network Configuration Phase

When the VSN is deployed and brought up, the VSN nodes send advertisements to their neighbors declaring identities and their number of hops from sink. These advertisements are sent periodically. Initially all nodes are configured as being infinitely away from sink node. When sink node advertises, it declare its number of hops from sink as 0. The nodes receiving this advertisement add the respective sink node to their routing tables and mark their number of hops from sink as one hop. Now when such a node sends out its own advertisement, it declares its number of hops from the sink instead of infinity. The nodes at multiple hops from sink update their routing table with sink address along with the addresses of their neighbor as next hop address from who they received the advertisement. When a node receives advertisement of a sink from more than one neighbor, it keeps only the neighbor with lesser hops to the sink in its routing table. After a number of cycles of advertising, depending on the number of VSN nodes, the network is established. Each node knows the number of hops to the sink as well as the next hop towards the sink. As the advertisements are sent out periodically, removal and addition of nodes to the network is possible dynamically. Moreover, for maintenance of routing tables, each node keeps track of live neighbors using a watchdog timer associated with each neighbor.

2) Network Operation Phase

Once the network has been established, our routing framework is ready to transport image data from camera nodes to sink nodes. When the VSN application has image data to send, it uses primitives provided by the AIL from previous section. The network layer selects the next hop towards the sink that is selected by the camera node from its routing table. If the sink address as specified by the camera node is not in the routing table, the packet is dropped. A neighbor's entry keep-alive watchdog is reset whenever a packet is received from that neighbor. If a packet is not received from a neighbor within a threshold, the neighbor's entry is deleted from the routing table. In this way, routing tables are maintained during data transmission phase.

C. Medium Access and Control Layer

At the MAC layer, the routing framework works at two levels. The first is the intra-node level where the routing framework makes sure that high priority packets are forwarded before low priority packets. The second level is the inter-node level where the routing framework makes sure that when two neighbors contest for transmission medium, the neighbor with high priority packet gets a chance to transmit its packet before the neighbor with low priority packet. The following sub-sections explain these two levels.

1) Queue Insertion

 ISBN: 978-1-61208-320-9

When a packet arrives at MAC layer for transmission, it is sent instantaneously if the MAC layer is not already receiving or sending a packet. If the MAC layer is busy, the packet is placed in a queue where it waits for its turn. Our priority-based routing framework makes use of this queue. When a packet with high priority arrives, it is placed at the head of the queue so that it is sent in the next go. If a packet of low priority arrives, it is placed at the tail of the queue. As the MAC layer always selects packets from head of the queue for transmission, it is made sure that at intra-node level a packet with higher priority is transmitted first.

2) Differentiated Back-off Window

When two nodes find the medium available and transmit at the same time, a collision occurs. In regular CSMA/CD, both nodes back off for a randomly selected time slot from a pseudo-fixed-size window. If they collide again, the window size is increased exponentially to a certain size. The priority-based routing framework maintains different windows for the different priorities. When a collision occurs, the MAC layer checks the priority of packet that collided and determines back-off times from different windows. For high priority packet, the window is smaller than for a low priority packet. This way, if the node with high priority packet gets a chance to transmit its packet within a smaller window than a node with a low priority packet. This makes sure that at the inter-node level, high priority packets transmit sooner than low priority packets.

V. PROPOSED PROTOCOL IMPLEMENTATION

To quantify the usefulness of the routing framework, a VSN application was created and simulations were run. Contiki OS [24], an open source operating system for devices such as wireless sensor network nodes, was used to implement the routing framework. Modifications were made to MAC and network layer of RIME protocol stack [25] part of Contiki OS. RIME protocol stack provides a set of basic communication primitives ranging from best-effort single-hop broadcast and best-effort single-hop unicast, to best-effort network flooding and hop-by-hop reliable multi-hop unicast. The RIME protocol stack provides multiple options for each protocol layer. The configuration of RIME used for routing framework implementation consists of hop-by-hop reliable multi-hop unicast with a user-defined network layer, CSMA/CD as MAC layer and ContikiMAC [26] as Radio Duty Cycling layer. Modifications made to each layer of RIME protocol stack of Contiki OS are explained in the sub-sections below.

A. Modifications in RIME Network Layer

The custom network layer contains a periodic timer that expires half a second. Whenever the timer expires, a node sends out an advertisement. These periodic advertisements from each node help build routing tables as explained in the previous section. A network packet in RIME protocol stack is 128 bytes long. 24 bytes of this packet are used by RIME for header and remaining 104 byes are available as payload. When used as an advertisement, the payload contains addresses of sink nodes and their corresponding hops count from the node announcing the advertisement.

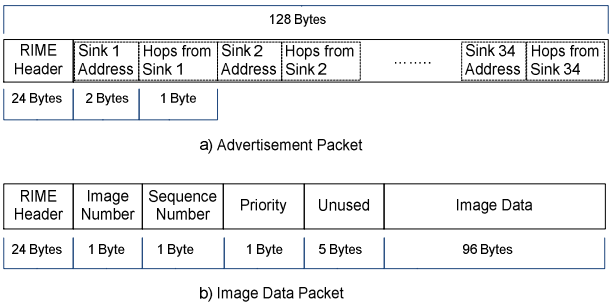


Figure 2. Types of VSN Packets

When a visual sensing node has a packet to send, it uses AIL send primitive to send it. The send primitive of AIL takes image layer, address of sink node and priority of the layer. The AIL fragments the image layer into packets. AIL also insert the image number and fragment number or packet sequence number into the data packet along with 96 bytes of image data. The image number and packet sequence number are used at the sink to reconstruct the image layer. Both advertisement and data packets are depicted in Fig. 2.

When a packet is received at the network layer of an intermediate node from a neighbor, it is checked if the packet is for the node itself or it is an image data packet that needs to be routed to some sink. In case if the packet is to be routed to the sink, the next hop is determined from the routing table that maintains next hop addresses corresponding sinks address. The neighbor is chosen as the next hop whose number of hops from sink is least. The data packet is then sent to that neighbor so that it can forward the packet to the sink or next hop towards the sink.

B. Modifications in RIME MAC Layer

The RIME MAC layer chosen for implementation of routing framework is CSMA/CD [27]. It contains a queue to store packets waiting for their turn for transmission. Modifications have been made to how a packet will be inserted into the queue. When the packet is received by MAC layer from network layer, the priority of the packet is checked. If it is a high priority packet, it is placed at the head of the queue. If it is a low priority packet, it is placed at the tail of the queue. When sending a packet, the MAC layer always picks up a packet from the head of the queue. This way if there is any high priority packet in the queue, it will be transmitted before low priority packets giving precedence to first pass image information at intra-node level.

$$E_c = \left(\frac{2^c - 1}{2} \right) \quad (1)$$

c is the number of times the packet collided

$$E_{c,p} = \left(\frac{2^{c(1+pW)} - 1}{2} \right) \quad (2)$$

$$p = \begin{cases} 0 & \text{if high priority packet collided} \\ 1 & \text{if low priority packet collided} \end{cases}$$

W is the contention window size

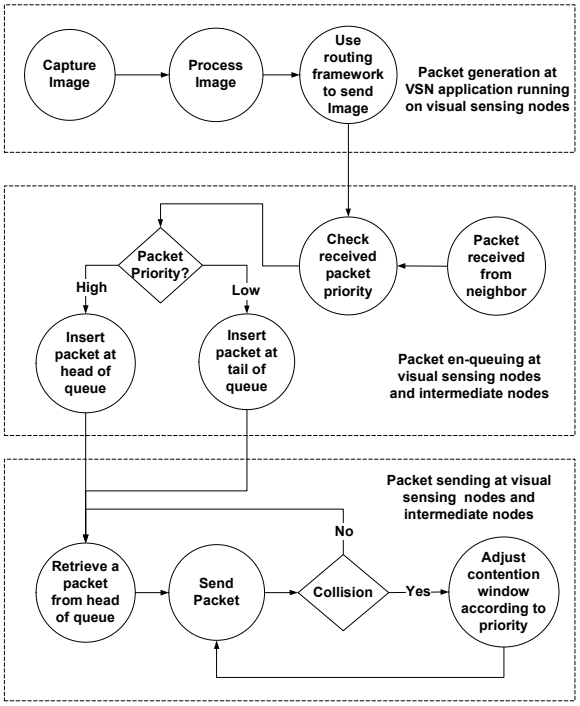


Figure 3. Routing Framework Data Flow

On sensing the medium to be free, if two nodes transmit at the same instance, a collision will occur. When this collision is detected by the MAC layer, it defers the transmission of that packet based on a random time slot out of a pseudo-fixed-sized contention window. The random time slot is selected using binary-exponential back-off algorithm. Without our modifications, the back-off algorithm maintains the same contention window for all types of packets that collide. The expected back-off time, *E(c)*, can be approximated using (1). We introduce a factor *pW* that enhances the back-off time calculation for packets of different priority levels. The factor *pW* in (2) causes contention window to shift for low priority packets, providing inter-node level precedence to high priority packets.

The flow of image data through the modified RIME stack is depicted in Fig. 3. The topmost block represents the VSN application and its usage of AIL. The middle block represents packet en-queueing into MAC layer transmission queue. The bottom block signifies the transmission of packet and calculation of contention window in case of collision.

The type of VSN applications targeted in this paper can be implemented using low-cost sensor network node such as TelosB [28]. Some motes can be equipped with CMUCam4 [29] giving them image capturing ability. The remaining TelosB nodes can be used to route image data from camera nodes to sink nodes. These applications of such VSNs can capture images and use image encoding algorithms such Discrete Cosine Transform [2] or Discrete Wavelet Transform [3][4] to encode images into different level of details for progressive image transmission. In the future, we intend to implement our proposed routing framework with real VSN application to measure its performance.

VI. SIMULATION RESULTS

For simulations, we created an application that emulates a real VSN application by generating random image layers according to user-defined configurations. The simulation configurations set to quantify the usefulness of routing framework consist of generating 90 x 90 pixels resolution image layers where each pixel is of 3 bytes, 1 byte per color. Therefore the entire image layer is 90 x 90 x 3 bytes (24 Kilobytes, approximately). One data packet can transport 96 bytes hence one image layer is transmitted in less than 256 packets. The ratio of high priority to low priority packets is kept as 50-50%. The size of MAC layer queue is set to 32 packets. The simulations consist of 25 VSN nodes arranged in a regular grid, as depicted in Fig. 4. The channel check rate is set to 64, i.e., in one second the ContikiMAC radio duty-cycling layer checks the channel 64 times to see if a neighbor is transmitting. The dotted-line represents the transmission-reception ranges. The dot-filled circles represent sink node. The empty circles represent intermediate nodes. The circles with stripes denote visual sensing nodes.

The nodes at the corner of the grid have only two neighbors in their transmission-reception range, e.g., Node-20 and Node-24 are in vicinity of Node-25. Nodes on the side have three nodes in their vicinity, e.g., Node-22, Node-18 and Node-24 are in transmission-reception range of Node-23. Finally, remaining nodes of the grid have 4 neighbors in their vicinity, e.g., Node-12, Node-8, Node-14 and Node 18 are in vicinity of Node-13.

The application can emulate different scenarios by modifying simulation configurations. The camera nodes generate packets varying from 1 to 32 packets per second.

Three network configurations, depending on the number of visual sensing nodes, have been tested with a large number of simulations for each configuration. Node-1 was selected as sink in all simulations. For each network configuration, packets were generated at rates starting from 1 packet per second to 24 packets per second. Results of each configuration are given in the below.

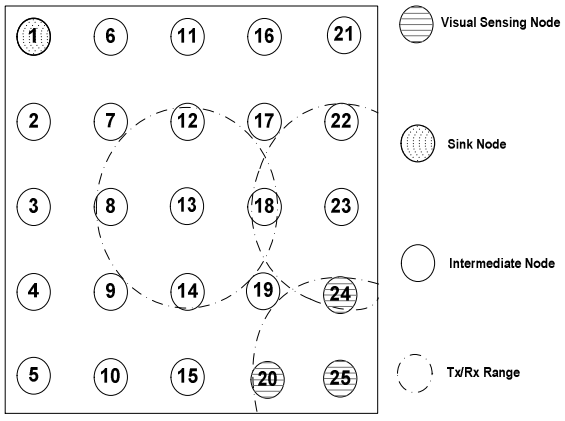


Figure 4. Grid Topology

 ISBN: 978-1-61208-320-9

The first network configuration contains one visual sensing node, Node-25, responsible for generating image layers. It is placed at 8 hops from the sink. Fig. 5 shows the average time taken by high priority packets and low priority packets to reach the sink node from visual sensing node. The lines represent average time taken with our proposed routing framework in place as compared to average time taken without our framework. The lines with circle and square symbols denote average transmission times of high priority and low priority packets, respectively, with routing framework inactive. Simulations were carried out without the routing framework in place to generate reference results. As the routing framework is not managing MAC queues and retransmission times of packets, there is no difference in routing of high and low priority packets. Both types of packets are treated the same way by the network. As a result both high and low priority packets take almost same time to reach the sink node. This is why circle symbols are not clearly visible in Fig. 5.

With the priority based routing framework actively managing MAC queues and retransmission times, high priority packets (denoted by line with triangles) take much lesser time than low priority packets (denoted by line with crosses). The legend for all figures has been kept similar to Fig. 5 for easy comparison by the reader. At lower packet generation rates the difference in average transmission times is less visible because the MAC layer queues are almost empty. Moreover, as each node has lesser packets to transmit, collisions rarely occur. As the packet generation rate is increased, the effect of routing framework becomes visible. The average transmission time for high priority packets decreases significantly as compared to the reference simulations. On the same note, average transmission times for low priority packets have increased as compared to the reference simulations.

Fig. 6 represents packet delivery ratios with and without our proposed routing framework in place at the 30 seconds deadline. Packet delivery ratio denotes the ratio of packets generated from the visual sensing nodes to packets received at the sink. At low packet generation rates, the difference in packet delivery ratios is less visible because the MAC layer queues are almost empty and as each node has lesser packets to transmit, resulting in rare cases of collisions. As the packet generation rate increases delivery ratio of high priority packets improves as compared to low priority packets. Moreover, delivery ratio of high priority packets is better than reference graphs when routing framework was inactive.

Fig. 7 represents the packets received over percentage of simulation time with and without our proposed routing framework in place. Without our framework, the number of packets received over simulation time is same for both high and low priority packets. With our framework, the number of high priority packets received is higher than number of low priority packets received. Hence, at any time in the simulation, the sink node receives more high priority packets although the packet generation rate has been kept same for both types of packets in our simulations.

To reconstruct the image at the sink node within a certain time, the image decoding algorithms running on the sink node impose deadlines for each layer. As the image encoding and decoding algorithms are not part of this paper, we have selected a deadline of 10 seconds for high priority packets corresponding to coarse image information of first pass and a deadline of 30 seconds for fine image information of second pass. In a real VSN application, these deadlines will be dependent on the image decoding algorithm. Fig. 8 represents the packet delivery ratio within these deadlines. With our proposed routing framework in place, the delivery ratio of high priority packets that reached the sink node within 10s seconds of transmission is significantly higher than without the routing framework active. With the routing framework active, the delivery ratio of low priority packets decrease as the packet generation rate increases. This decrease is due to the increase in delivery ratio of high priority packets. As the network resources remain same, the increase in packet delivery ratio of high priority packets is compensated with decrease in delivery ratio of low priority packets.

The second network configuration contains two visual sensing nodes, Node-20 and Node-24, both placed at 7 hops from the sink. Whereas the third network configuration contains three visual sensing nodes, Node-20, Node-24 and Node-25. Figs. 9 - 12 represent average transmission times, packet delivery ratios at 30 seconds simulation deadline, packets received over percentage simulation time and deadline based packet delivery ratios for two visual sensing nodes simulations, respectively. Similarly, Figs. 13 - 16 represent average transmission times, packet delivery ratios at 30 seconds simulation deadline, packets received over percentage simulation time and deadline based packet delivery ratios for three visual sensing nodes simulations, respectively. Simulations with two and three visual sensing nodes were carried out to see the effects of having more than one visual sensing node in the network.

As there is an overlapping between the paths from the visual sensing nodes to the sink node for two and three visual sensing nodes simulations, difference in average transmission times can be seen as compared to simulation results of one visual sensing node. This overlap increases the average transmission times for all packet generation rates as compared to simulations with one visual sensing node. Similarly, there is a difference in packet delivery ratios and packets received within deadlines as compared to simulation results of one visual sensing node.

The increase in average transmission times and the decrease in packet delivery ratios are because of two reasons. The first reason is that due to overlapping paths, packets collide. Collisions cause excessive retransmission. When the MAC layer's maximum retransmission threshold is achieved, the packet is discarded causing the packet delivery ratio to decrease. The packets that reach the sink take more time because of multiple retransmissions by the intermediate nodes causing the average transmission time to increase. The second reason is that as collisions increase, the lifetime of packet in the MAC layer queue also increases. This causes the queue to fill up sooner. As a result incoming packets do not find space in MAC layer queue and are dropped causing packet delivery ratio to decrease.

 ISBN: 978-1-61208-320-9

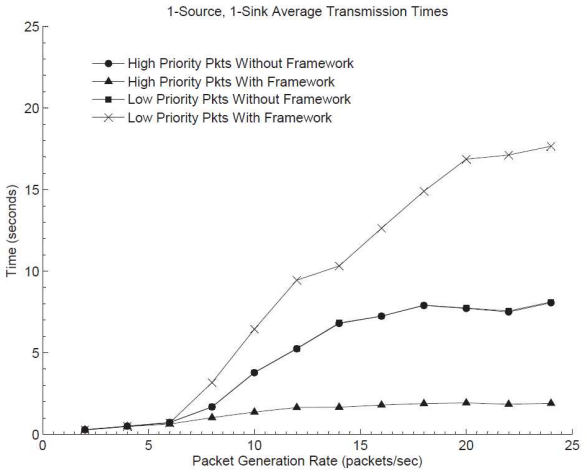


Figure 5. Average Transmission Times for 1 Source

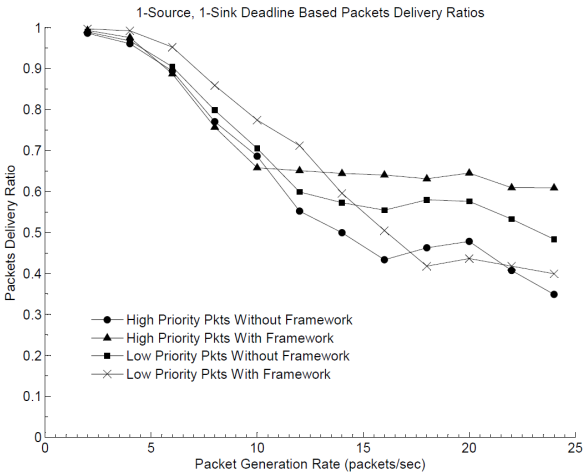


Figure 8. Delivery Ratios for 1 Source within Deadlines

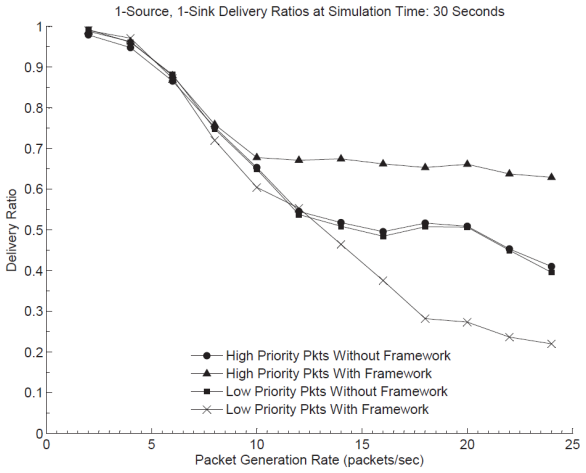


Figure 6. Delivery Ratios for 1 Source at Time: 30s

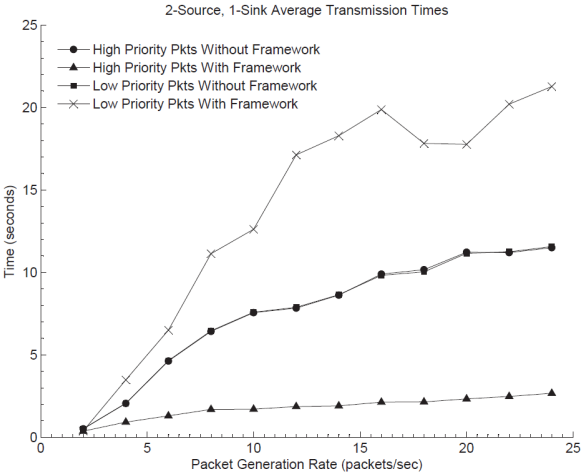


Figure 9. Average Transmission Times for 2 Sources

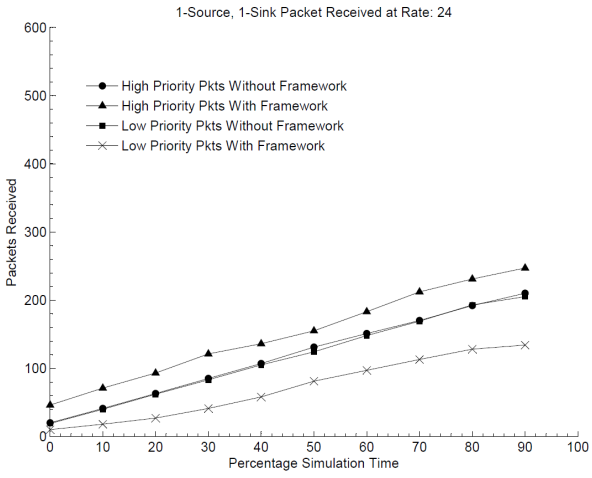


Figure 7. Packets Received for 1 Source over Simulation Time

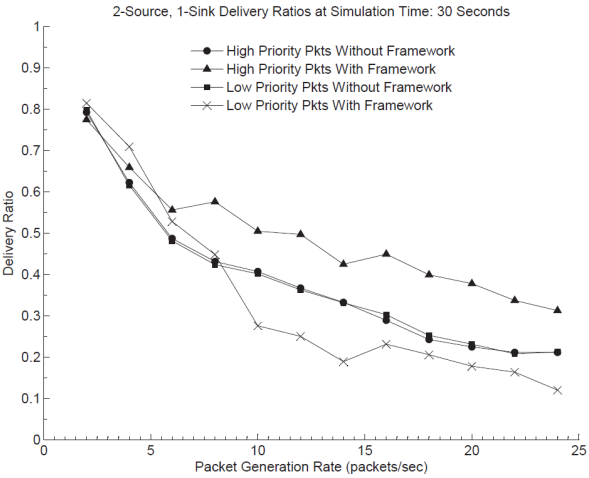


Figure 10. Delivery Ratios for 2 Sources at Time: 30s

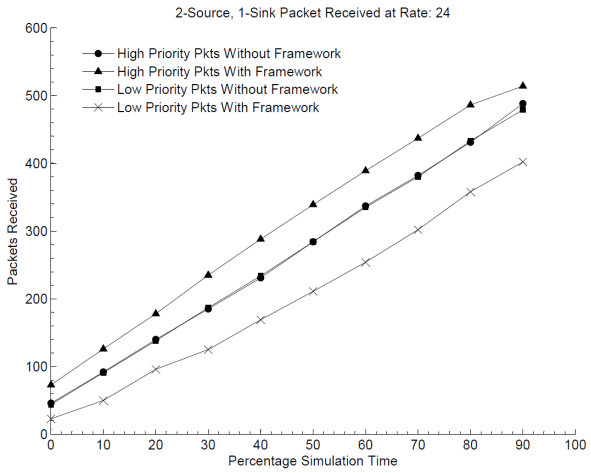


Figure 11. Packets Received for 2 Sources over Simulation Time

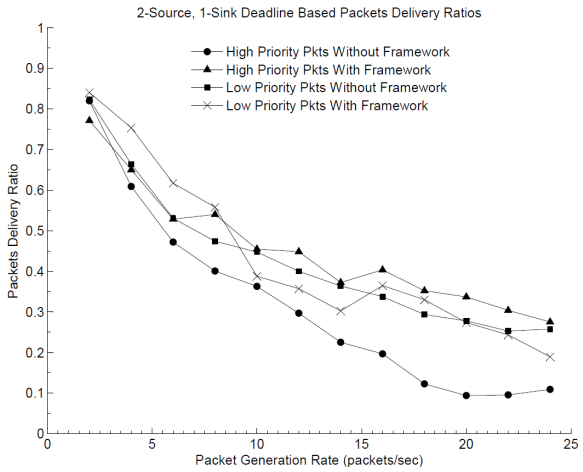


Figure 12. Delivery Ratios for 2 Sources within Deadlines

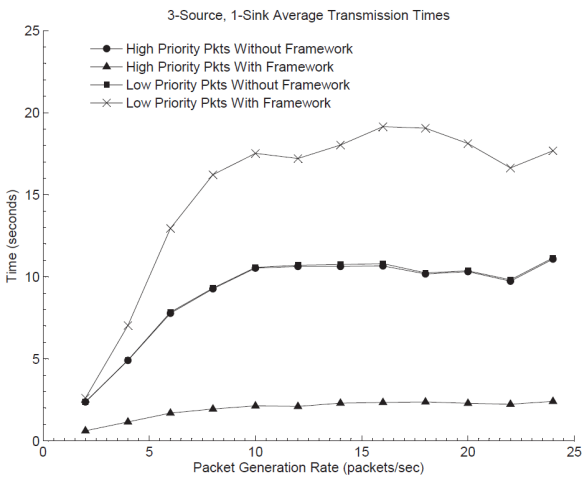


Figure 13. Average Transmission Times for 3 Sources

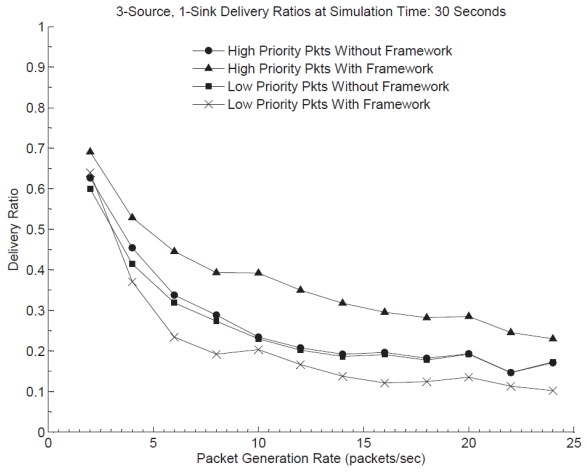


Figure 14. Delivery Ratios for 3 Sources at Time: 30s

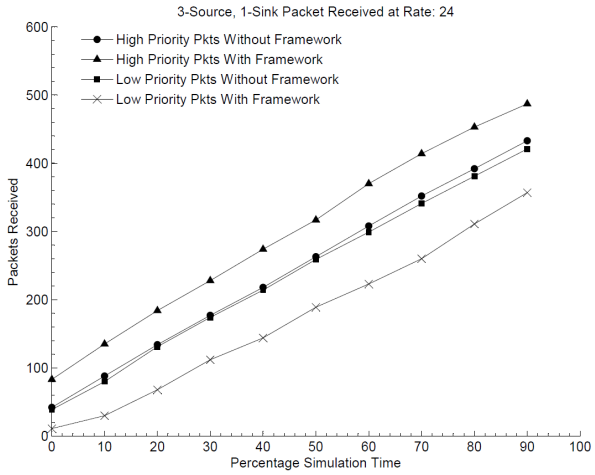


Figure 15. Packets Received for 3 Sources over Simulation Time

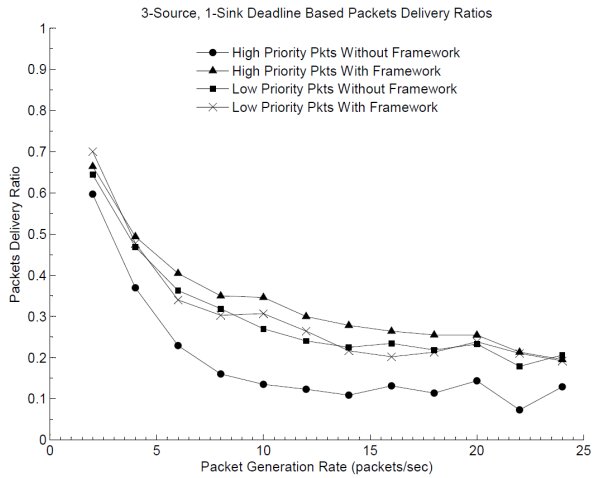


Figure 16. Delivery Ratios for 3 Sources within Deadlines

 ISBN: 978-1-61208-320-9

Hence, we prove that our framework improves system's response time in certain VSN applications.

VII. Conclusion and Future Work

Based on simulation results, we can conclude that our proposed priority-based routing framework assists progressive image transmission in VSNs. Critical imagery information from visual sensing nodes can be received at sink nodes sooner than less critical imagery information. However, there are areas of priority-based routing framework that can be improved. In the future, the authors of this paper intend to integrate this priority-based routing framework with an image encoding/decoding mechanism to measure the performance on a complete VSN platform.

Acknowledgment

This research is funded by the Institute of Consulting Research and Studies, Umm Al-Qura University, Makkah, Saudi Arabia, Grant No. S2011-9.

References

[1] I. F. Akyildiz, T. Melodia, and K. R. Chowdhury, "A survey on wireless multimedia sensor networks," Computer Networks, vol. 51, 2007, pp. 921–960

[2] Y. Huang, H. M. Dreizen, and N. P. Galatsanos, "Prioritized dct for compression and progressive transmission of images," Image Processing Transactions, vol. 1, no. 4, Oct. 1992, pp. 477–487, doi:10.1109/83.199917

[3] K. H. Talukder and K. Harada, "Enhancement of discrete wavelet transform (dwt) for image transmission over internet," in Proceedings of the 8th International Conference on Information Technology: New Generations, (ITNG '11), IEEE Computer Society, 2011, pp. 1054–1055, doi:10.1109/ITNG.2011.184

[4] F. Behnamfar, F. Alajaji, and T. Linder, "Progressive image communication over binary channels with additive bursty noise," in Proceedings of the Data Compression Conference, (DCC '02), IEEE Computer Society, 2002

[5] K. Zuberi and K. Shin, "Design and implementation of efficient message scheduling for controller area network," IEEE Transactions on Computers, vol. 49, no. 2, Feb 2000, pp. 182 –188.

[6] S.-K. Kweon and K. G. Shin, "Providing deterministic delay guarantees in atm networks," IEEE/ACM Trans. Netw., vol. 6, no. 6, pp. 838–850, Dec. 1998, doi:10.1109/90.748093

[7] C. Li, R. Bettati, and W. Zhao, "Static priority scheduling for atm networks," in Proceedings of the 18th IEEE Real-Time Systems Symposium, ser. RTSS '97. Washington, DC, USA: IEEE Computer Society, 1997

[8] W. Zhao, J. Stankovic, and K. Ramamritham, "A window protocol for transmission of time-constrained messages," IEEE Transactions on Computers, vol. 39, no. 9, 1990, pp. 1186–1203.

[9] D. Kandlur, K. Shin, and D. Ferrari, "Real-time communication in multi-hop networks," in 11th International Conference on Distributed Computing Systems, May 1991, pp. 300 –307.

[10] L. Chen and W. B. Heinzelman, "A survey of routing protocols that support qos in mobile ad hoc networks," Network Magazine of Global Internetworking, vol. 21, no. 6, Nov. 2007, pp. 30–38, doi:10.1109/MNET.2007.4395108

[11] L. Hanzo-II and R. Tafazolli, "A survey of qos routing solutions for mobile ad hoc networks," IEEE Commun. Sur. Tuts., vol. 9, no. 2, Apr. 2007, pp. 50–70, doi:10.1109/COMST.2007.382407

[12] J. Liebeherr, D. Wrege, and D. Ferrari, "Exact admission control for networks with a bounded delay service," IEEE/ACM Transactions on Networking, vol. 4, no. 6, Dec 1996, pp. 885 –901

[13] S. Wang, D. Xuan, R. Bettati, and W. Zhao, "Providing absolute differentiated services with statistical guarantees in static-priority scheduling networks," in proceedings of Seventh IEEE Real-Time Technology and Applications Symposium, 2001, pp. 127 –129.

[14] I. Stoica and H. Zhang, "Providing guaranteed services without per flow management," in proceedings of ACM Applications, technologies, architectures, and protocols for computer communication, (SIGCOMM '99), New York, NY, USA: ACM, 1999, pp. 81–94,.doi:10.1145/316188.316208

[15] O. Younis and S. Fahmy, "Constraint-based routing in the internet: Basic principles and recent research," IEEE Comm. Surveys Tuts, vol. 5, no. 1, Jul. 2003, pp. 2–13, doi:10.1109/COMST.2003.5342226

[16] J. Soldatos, E. Vayias, and G. Kormentzas, "On the building blocks of quality of service in heterogeneous ip networks," Commun. Surveys Tutorials., vol. 7, no. 1, Jan. 2005, pp. 69–88, doi:10.1109/COMST.2005.1423335

[17] K. Sohrabi, J. Gao, V. Ailawadhi, and G. Pottie, "Protocols for self-organization of a wireless sensor network," IEEE Personal Communications, vol. 7, no. 5, Oct 2000, pp. 16 –27

[18] C. Lu, B. M. Blum, T. F. Abdelzaher, J. A. StankoSvic, and T. He, "Rap: A real-time communication architecture for large-scale wireless sensor networks," Charlottesville, VA, USA, Technical Report, 2002.

[19] T. He, J. Stankovic, C. Lu, and T. Abdelzaher, "Speed: a stateless protocol for real-time communication in sensor networks," in proceedings of 23rd International Conference on Distributed Computing Systems, May 2003, pp. 46 – 55.

[20] O. Chipara, Z. He, G. Xing, Q. Chen, X. Wang, C. Lu, J. Stankovic, and T. Abdelzaher, "Real-time power-aware routing in sensor networks," in proceedings of 14th IEEE International Workshop on Quality of Service, (IWQoS 2006), June 2006, pp. 83 –92.

[21] E. Felemban, C. Lee, and E. Ekici, "MMSPEED: Multipath Multi-SPEED Protocol for QoS Gurantee of Reliability and Timeliness in Wireless Sensor Networks," IEEE transaction on Mobile Computing, vol. 5, no. 6, June 2006.

[22] T. Fang, C. Fu, B. Falkowski, and B. Wang, "Multiple dynamic range image coding for wireless sensor networks," in proceedings of IEEE International Conference on Systems, Man and Cybernetics, 2008. (SMC 2008), Oct. 2008, pp. 2944 –2949.

[23] E. Manhas, G. Brante, R. Souza, and M. Pellenz, "Energy-efficient cooperative image transmission over wireless sensor networks," in proceedings of Wireless Communications and Networking Conference (WCNC), 2012, pp. 2014 –2019.

[24] A. Dunkels, B. Gronvall, and T. Voigt, "Contiki - a lightweight and flexible operating system for tiny networked sensors," in proceedings of the 29th Annual IEEE International Conference on Local Computer Networks, (LCN '04). pp. 455–462, 2004, doi:10.1109/LCN.2004.38

[25] A. Dunkels, F. O¨ sterlind, and Z. He, "An adaptive communication architecture for wireless sensor networks," in proceedings of the 5th ACM International conference on Embedded networked sensor systems, (SenSys '07)., 2007, pp. 335–349. doi:10.1145/1322263.1322295

[26] A. Dunkels, "The contikimac radio duty cycling protocol," Swedish Institute of Computer Science, Technical Report, 2011.

[27] Farooq, M.O., Kunz, T., "Contiki-based IEEE 802.15.4 node's throughput and wireless channel utilization analysis," Wireless Days (WD), 2012 IFIP , vol., no., pp.1,3, 21-23 Nov. 2012 doi: 10.1109/WD.2012.6402828

[28] J. Polastre, R. Szewczyk, and D. Culler, "Telos: enabling ultra-low power wireless research," in proceedings of the 4th IEEE international symposium on information processing in sensor networks, (IPSN '05), Piscataway, NJ, USA, 2005

[29] L. Surhone, M. Tennoe, and S. Henssonow, "CMU-CAM", Betascript Publishing, 2010. [Online]. Available: http://books.google.com.sa/books?id=0EBfYgEACAAJ, Date Visited: 19-Dec-2013

 ISBN: 978-1-61208-320-9

Structuring Video Database using a Formal Methods Approach

Noraida Haji Ali, Fadilah Harun
School of Informatics and Applied Mathematics,
Universiti Malaysia Terengganu, Terengganu, Malaysia
e-mail: aida@umt.edu.my, elaharun@yahoo.com

***Abstract*— Formal methods provide a foundation for many of the techniques that have changed the face of software development over the last two decades. Structuralizing video streams plays an important role in the processing of video. The basic structure for video is a hierarchical structure which consists of four kinds of components, namely frame, shot, scene, and video program. Formalizing supports the reliability and accuracy of the modeling language. Various researches have been done related to formal methods, video structure, and formalizing. This paper discusses the analysis of formal methods, applications, and structuring formal methods in video structure. The output from this study is, can determine the relation in video structure using the algebraic relation.**

Keywords-formal methods; video; video structure; formalization and formal specification.

I. INTRODUCTION

A multimedia database is a structure and organizes multimedia information for content retrieval [1]. It supports the various multimedia data types like texts, images, audio and video [2]. Among all these media types, video is the most challenging one [3]. This is because video combines all other media types' information into a single data stream. This is also because the challenges faced by researchers when implementing video increase even further when one moves from images to image sequences, or video clips. Also, every single video programs have their own rules and format.

Structuring the design is important in the development phase of the system. Our research is focused on structuring video using formal speficications. Video structure used segmentation-based techniques because individual shots or scenes logically meaningful units [4]. In addition, each shot or scene consisting of a sequence of frames and each frame can be considered as the image, allowing the use of the techniques available that have been developed to model and query image data. The increase in processing power and storage capabilities of modern computers has led to the development of new multimedia applications. The structures of a videos that are commercially available are not mutually compatible and interoperable. Furthermore, there are no database support video and view schema objects. These problems led to an effort to determine the specifications of a new structure for video federation. Previous researchers have outlined the issues related to video systems and discussed the technical challenges involved in developing a general-purpose video system. There are important issues in multimedia database management, including the development of formal modeling techniques for multimedia information, especially for video and image data. This structure should be rich with the ability to capture abstractions and semantics of multimedia information. By using formal methods, we hope to improve the structure of a video and can be more efficient for their content retrieval.

Some benefits expected from this study is the improvement of the quality of the video structure. The main objective of a formal specification notation is to assist the descriptions of video in order to make sure the structures are complete, consistent and unambiguous. Therefore, we propose an algebraic relation to design the relation in video structure.

This paper is organized as follows: The following section discusses research background on previous research and problem statement. In Section III, the video structure is explained. The formalizing of video structure and video algebra relation also discussed with the given example. In Section IV, elaborate on future works in this research. Finally, our conclusions are stated in Section V.

II. RESEARCH BACKGROUND

This study presents a formal methods as a proposed method to apply in the design and structuring video system. This method requires to forces an analysis of the system requirements at an early stage.

A. Previous Research

Structuring a video used several techniques, such as temporal and spatial [5] relations to design the database. However, structuring a video has a limitation, example in size and modeling the complex object in a wide range of types for indexing, searching and organization methods [6]. Therefore, the previous research tried to solve relation issues in multimedia database using the temporal specification. The temporal relations are to determine duration relations between multimedia objects. Djeraba and Briand [7] used the power of temporal Petri net to model the temporal and interactive relations. Another research into the structured temporal composition of multimedia is based on binary operators that represent some of the previously described relations between intervals of unknown duration has been done [5]. Other research uses a novel indexing technique

 ISBN: 978-1-61208-320-9

based-on, efficient compression of the feature space for approximate similarity searching in large multimedia databases [8].

B. Problem Statement

A major problem encountered in the current database system is the lack of a natural way to define complex queries. This caused by the gap between the way users think and query language used in most systems. Multimedia data manipulation is not as easy as in a conventional database. The database structure can be represented with a clear video and can also be specified in requests for their content, but the main problem is to get the content of the video database. The difficulty arises because we must match the contents of the media data in the database with the content specified in the query. Each answer queries posted on media data, it must have an advanced technique in analyzing the contents of the data to get the different semantics associated with the media data [9]. The development of the new multimedia applications have been realized based on the improvements in processing and storage capabilities of latest computers [10]. Sometimes, available video structures are mutually incompatible and interoperability between them cannot be easily achieved. Furthermore, none of the approaches supports video and object view schemas. The issues on pertaining to video and discuss technical challenges involved in developing a general-purpose video system are specification requirements and the reference architecture. The salient issues in video system include development of formal modeling techniques for video information. These models should be high in capabilities for abstracting multimedia information and capturing semantics [6]. An important feature in accessing to databases of unknown structure is the presence of a schema repository where the database structure is explained, and a meta-model which provides a set of legal relationships and actions for entities in the database [10]. By using formal methods, we intend to improve the structure of video and it can be more efficient to retrieve their content.

III. VIDEO STRUCTURE

Video is the technology of capturing, recording, processing, storing, transmitting, and reconstructing a sequence of still images and representing scenes in motion. It helps to present the real world events to users. Video database provides random access to sequential video data. A basic video structure design it using segmentation-based techniques because individual shots or scenes logically meaningful units.

In order to help the users to retrieve relevant video materials, effectively at various semantic levels, a method defines a hierarchical structure of video material based on hierarchical data model. The structured proposes three steps of informal specification, as shown in Figure 1:

- Firstly, video is segmented into shots by shot boundary detection techniques [3].
- Secondly, key frames are selected to represent the shots.
- Finally, shots are clustered to scenes, based on the extracted shot features.

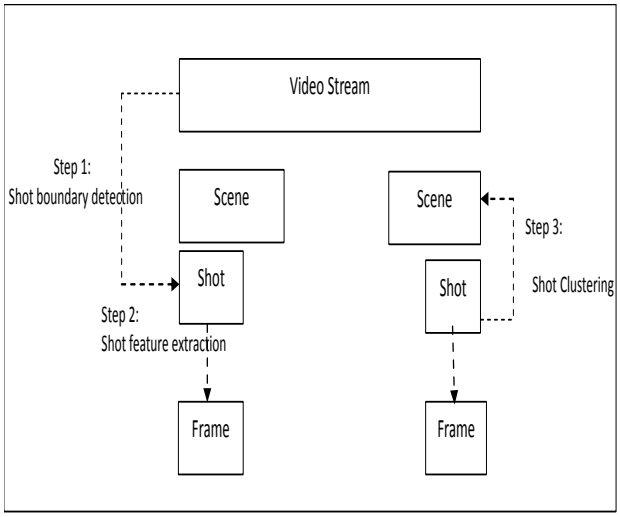


Figure 1. Video Structuring Process

A. Formalizing of Video Structure

Even though the others video do not have any specific content structure, as a scenes and videos, many videos have a fraction sequence of frames that is recorded from a single camera motion (shot) that can also express the content structure of the video [11]. A basic video structure contains scene, shot and key frame [12]. In Figure 2, a set of video structure is shown. The algebraic relationship model is illustrated through the statement mathematics [13]. A tuple is a set of relations and it structured in an easy form and this is called the schema relationships. However, this study is still in early stage. Before, we look further on video database system; we start on video basic structure.

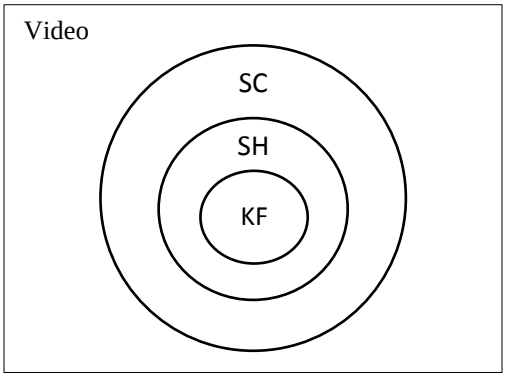


Figure 2. Set of Video Structure

A structure representation in Figure 2 can be described in the following algebraic relations.

 ISBN: 978-1-61208-320-9

V (SC, SH, KF)
SH⊆ SC
KF ⊆ SH

where:
V: video
SC: sequence of shots.
SH: unbroken sequence of frames, and
KF: selected from a shot to represent the contents of the shot.

B. Video Algebra Relation

The basic structure of the video is a hierarchical structure. It is produced by the video program, scene, shot and key frame. A video program usually consists of a few scenes, and each scene includes one or more shots [11]. A key frame is a static image and minimum logic unit of video. A shot is an uninterrupted segment of video frame sequence with static or continuous camera motion. A scene is a series of shots that are sticked together from the narrative point of view. The definition below is to explain the structure of the video.

$\overline{MM}$ (V, SC, SH, KF)

MM : Multimedia Database

Every X_i in existence contains four data, such as represented below:

$\overline{MM}$ (V_i, SC_{ij}, SH_{jk}, KF_{km}) : *i, j,k,m* is an integer number.

Assume X_i is a video structure of i that exist in multimedia database (MM). Which consists of:

$V : \{<v,i>\} \mid i \in \mathbb{N} \wedge v_i \in MM\}$
V_i : is a video for video of the i
There exists a video(V_i) , in the video structure (X_i).

$SC : \{<<sc,i> v,j>\} \mid i \in \mathbb{N} \wedge <sc, v> \in SC_{ij}\}$
SC_{ij} : j^{th} Scene for the i^{th} Video (V_i)

$SH : \{<<sh,j> sh,k>\} \mid j \in \mathbb{N} \wedge <sh, j> \in SH_{jk}\}$
SH_{jk} : k^{th} Shot for the j^{th} Scene (SC_j)

$KF :\{<<kf,k> sh,m>\} \mid k \in \mathbb{N} \wedge <kf, k> \in KF_{km}\}$
KF_{km} : m^{th} Key Frame for the k^{th} Shot (SH_k)

There exists Scene (SC_j), Shot (SC_k) and Key Frame (KF_m) in the video (V_i).

C. Example

In this section, we are given an example of notation representation based on the hierarchical structure of the video. In Figure 1, we showed that in Video (V_1) there are two scenes, two shots and two key frames, and the algebra relation are:

$V_1 \longrightarrow SC_{1j}=\{1,1..2\} \longrightarrow SH_{jk}=\{1..2,1..2\} \longrightarrow$
$KF_{km}= \{1..2,1..2\}$

If i=1, j=2, k=2 and m=2, so:

$V_1 = \{SC_{1,F}, SH_{1,1}, KF_{1,1}\}$
$= \{SC_{1,2}, SH_{2,1}, KF_{2,1}\}$

$\forall X_i \exists MM =\{V_1, SC_{1,1..2}, SH_{1..2,1..2}, KF_{1..2,1..2}\} \mid X_i \in MM$

This shows that, in Video (V_1), there are two scenes that were identified by integer value j, Scene ($SC_{1..2,1..2}$). There also had two shots that were identified by an integer value k, Shot ($SH_{1..2,1..2}$). Lastly, key frame identified by integer m, ($KF_{1..2,1..2}$).

In general:
$X : \{ X_i \mid i \in I \wedge \forall X_i \in MM \}$

where:
X : exists in multimedia database (MM)
I : integer number ($1 \leq I \geq n$)

SX_i represents the total of video in multimedia database and it may increase depending on the circumstances and situation of an existing system. Therefore:

$$SX_i = \sum_{i=1}^{n} X_i$$

IV. CONCLUSION AND FUTURE WORK

Structuring video is an important step in video data management. The basic video structure is used hierarchical structure. The structured modeling approach has evolved into segmentation-based approach. In this paper a mathematical notation used to describe the structure of the database video more clearly. It can explain the relation between the scene, shot and key frame. Mathematical notation is a combination of delegated some elements in the structure. From the representation algebraic relations, the structure of video extracted can be developed further. The video relation algebra generated will help in the process to design new structures of video database using formal methods. By using formal methods, hope can improve the structure of video and can be more efficient to retrieve their content.

Based on Figure 3 below, this paper focused on early stage in formal specification methods. We started with algebraic relation to identify the relation of set in video

structure. An algebraic relation is an algebraic structure equipped with the Boolean operations [13]. Formal specifications have two types are property-oriented and model oriented [14]. Algebraic relation is part of property oriented. Property oriented is state desired properties in a purely declarative way. Relation algebraic will be written in the form of a schematic for proving and verifying. A schema is essentially the formal specification analogous to programming language subroutines that are used to structure a system, where the schemas are used to structure a formal specification. Z [15] is physically powerful on sets and functions. Formal proving is a complete argument of mathematical representation and it is used to validate statement about system description. Formal proving can be done using theorem proving tools, i.e., Z/Eves [15].

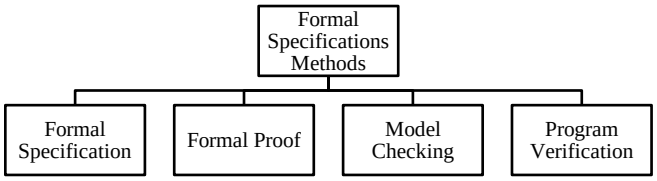


Figure 3. Formal Specification Methods [16].

ACKNOWLEDGMENT

We are grateful to Universiti Malaysia Terengganu for financial support.

REFERENCES

[1] S.K. Jalal, "Multimedia Database: Content and Structure", Workshop on Multimedia and Internet Technologies, 2001.

[2] F. Mohamed, M.N.A. Rahman, and M.L. Abdullah, "The development of temporal-based multimedia data management application using web services", in Intelligent Systems Design and Applications (ISDA), 2011 11th International Conference on. 2011.

[3] L.S. Affendey, A. Mamat, H. Ibrahim, and F. Ahmad, "Video Data Modelling To Support Hybrid Query", IJCSNS, 2007. 7(9): pp. 53.

[4] B. Günsel and A.M. Tekapl, "*Content-based video abstraction"*, In Proceedings of IEEE International Conference on Image Processing, 1998: pp. 128-131.

[5] A. Duda and C. Keramane, "Structured temporal Composition of Multimedia Data", in iw-mmdbms. 1995.

[6] A. Ghafoor, "Multimedia Support Infrastructure (MSI)", National Science Foundation 2006 [cited 2013 20 November]; Available from: http://www.cs.purdue.edu/.

[7] C. Djeraba and H. Briand, "Temporal and interactive relations in a multimedia database system", in Multimedia Applications, Services and Techniques—ECMAST'97. 1997, Springer. pp. 457-473.

[8] E. Tuncel, H. Ferhatosmanoglu, and K. Rose, "VQ-index: An index structure for similarity searching in multimedia databases", in Proceedings of the tenth ACM international conference on Multimedia. 2002: ACM.

[9] D.A. Keim and V. Lum, "Visual query specification in a multimedia database system", in Proceedings of the 3rd Conference on Visualization'92. 1992: IEEE Computer Society Press.

[10] D. Bećarević and M. Roantree, "A metadata approach to multimedia database federations", Information and Software Technology, 2004. 46(3): pp. 195-207.

[11] C.-j. Fu, G.-h. Li, and K.-x. Dai, "A framework for video structure mining", in Machine Learning and Cybernetics, 2005. Proceedings of 2005 International Conference on. 2005: IEEE.

[12] T. Catarci, M.E. Donderler, E. Saykol, O. Ulusoy, and U. Gudukbay, "BilVideo: a video database management system", MultiMedia, IEEE, 2003. 10(1): pp. 66-70.

[13] N.L. Sarda, "Algebra and query language for a historical data model", The Computer Journal, 1990. 33(1): pp. 11-18.

[14] E.M. Clarke and J.M. Wing, "Formal methods: State of the art and future directions", ACM Computing Surveys (CSUR), 1996. 28(4): pp. 626-643.

[15] J.A. Jusoh, M.Y.M. Saman, and M. Man, "Formal Validation of DNA Database Using Theorem Proving".

[16] J. M. Wing, "A Specifier's introduction To Formal Method", Defense Advanced Research Project Agency (DOD), 1990.

 ISBN: 978-1-61208-320-9

Multiexposure Image Fusion Using Homomorphic Filtering and Detail Enhancement

Hui-Chun Tsai Jin-Jang Leou Han-Hui Hsiao

Department of Computer Science and Information Engineering
National Chung Cheng University
Chiayi 621, Taiwan
{thc100m, jjleou, hhh95p}@cs.ccu.edu.tw

***Abstract*—In this study, a multiexposure image fusion approach using homomorphic filtering and detail enhancement is proposed. First, the *N* input low dynamic range (LDR) RGB color images are transformed into the HSI color space. Intensity enhancement is achieved by homomorphic filtering, gamma correction is used to compensate the nonlinear response of display devices, and "cross-image" median filtering is used to generate the reference intensity image. Guided filtering and weighted least squares (WLS) optimization are used to perform local and global detail extractions on the *N* processed LDR images, respectively. The *N* weighting maps of the *N* processed LDR images are estimated by spatial and cross-image consistencies and then refined by cross bilateral filtering. Finally, the multiresulution spline based scheme is used to perform multiexposure image fusion. Based on the experimental results obtained in this study, the performance of the proposed approach is better than those of four comparison approaches.**

Keywords-low dynamic range (LDR) image; high dynamic range (HDR) image; tone mapping; homomorphic filtering; multiexposure image fusion

I. INTRODUCTION

In the last decade, image fusion has been employed in different application areas [1-2]. Image sensors usually have a limited dynamic range and a low dynamic range (LDR) image usually contains some under-exposed or over-exposed regions. Additionally, a natural scene usually contains high dynamic range (HDR) contents. To cope with this problem, a series of LDR images with different exposures can be fused to obtain an HDR image, which will be displayed on LDR devices. There are two main types of HDR imaging, namely, typical HDR imaging and multiexposure image fusion [3].

HDR imaging consists of two main steps: HDR reconstruction and tone mapping. First, HDR reconstruction techniques [4] usually recover the camera response function (CRF) and combine the radiance maps via a weighting function from a series of LDR images. Second, tone mapping is to compress the dynamic range of HDR images in order to display on LDR devices. Existing tone mapping approaches can be classified into global and local operators [5-7].

Compared to HDR reconstruction, multiexposure image fusion usually consists of two steps: selection and blending [8]. "Selection" decides the best representative regions and exposures among all the input LDR images via assigning weights to the pixels of each LDR image. For blending, the selected regions from LDR images are fused according to their weights individually.

Multiexposure image fusion is similar to alpha blending [9]. Li, Zheng, and Rahardja [10] introduced a new quadratic optimization based image fusion approach. In [3], a mostly detailed LDR image is synthesized directly from input LDR images by solving different optimization problems. Song et al. [11] proposed a probabilistic model to preserve the calculated image luminance levels and suppress reversals in image luminance gradients.

On the other hand, in Mertens et al. [1], a weight for a pixel is determined by three quality measures: contrast, saturation, and well-exposedness. All LDR images are blended at multiple scales by using the Laplacian and Gaussian pyramidal image decompositions. Gu et al. [12] modified the gradient field iteratively with twice average filtering and nonlinearly compressing in multi-scales. Fused gradient field is derived from the structure tensor of LDR images based on multi-dimensional Riemannian geometry. Zhang and Cham [13] used the gradient information to accomplish multiexposure image composition in both static and dynamic scenes. Zhang and Cham [14] also proposed a multiexposure image fusion approach for both static and dynamic scenes using both temporal consistency and spatial consistency. Zeev et al. [15] introduced a new way to construct edge-preserving multi-scale image decompositions, based on weighted least squares (WLS) optimization.

The paper is organized as follows. The proposed multiexposure image fusion approach is described in Section 2. Experimental results are addressed in Section 3, followed by concluding remarks.

II. PROPOSED APPROACH

A. System Architecture

As shown in Fig. 1, the proposed multiexposure image fusion approach for static scenes contains six stages. First, the *N* input LDR color images are transformed from the RGB color space into the hue, saturation, and intensity (HSI) color space so that the intensity and color (hue and saturation) components can be separately processed. Intensity enhancement is achieved by homomorphic filtering in the frequency domain and gamma correction [16] is used to

 ISBN: 978-1-61208-320-9

compensate the nonlinear response of display devices. To eliminate under-exposed or over-exposed regions in LDR images, the "cross-image" median filter is used to generate the reference intensity image. The guided filter [7] and weighted least squares (WLS) optimization [15] are used to perform local and global detail extractions on the *N* processed LDR images with gamma correction, respectively. Based on the reference intensity image, spatial consistency and cross-image consistency involving five consistency measures are computed to estimate the *N* weighting maps of the *N* processed LDR images with gamma correction. The cross bilateral filter is used to refine the *N* weighting maps. Finally, the multiresulution spline based scheme [17] is used to perform multiexposure image fusion and generate the final HDR image.

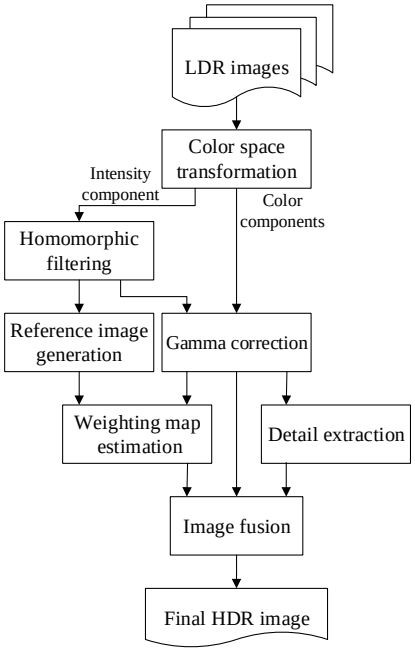


Figure 1. The framework of the proposed approach.

B. Homomorphic filtering

In this study, considering nonlinear intensity perception of the human visual system (HVS), the homomorphic filter is used to perform intensity enhancement in the frequency domain. $I_i^n(x,y)$ denotes the intensity of pixel (x,y) in the *n*-th input LDR image, $I_{i,H}^n(x,y)$ denotes the intensity of pixel (x,y) in the *n*-th homomorphic filtered image, and $H(u,v)$ is the transfer function of the homomorphic filter. The homomorphic filter processes the illumination and reflection components separately in the frequency domain (u,v) via the logarithm function. Here, $H(u,v)$ is a modified Gaussian highpass filter defined as

$$H(u,v) = (r_H - r_L) \cdot [1 - e^{-c(D^2(u,v)/D_0^2)}] + r_L, \quad (1)$$

where the constant *c* controls the sharpness of the transition slop of the filter function between r_L and r_H, D_0 is a positive constant, and $D^2(u,v)$ is the distance between a point (u,v) and the center $(W/2,H/2)$ of the frequency rectangle. Here, r_L, r_H, and *c*, are empirically set to 0, 1, and 1, respectively.

C. Gamma Correction

Gamma correction [16] is used to compensate the nonlinear response of display devices, which is defined as

$$I_C^n = (\frac{I_{C_{in}}^n}{L_{in}})^s \times L_{out}, \qquad C = R,G,B, \quad (2)$$

where *s* denotes the gamma correction coefficient in the range [0,1], $I_{C_{in}}^n$ and I_C^n are the color components of the *n*-th LDR image and the *n*-th processed LDR image with gamma correction, respectively, and L_{in} and L_{out} denote the luminances of I^n and I_H^n (the *n*-th homomorphic filtered LDR color image), respectively.

D. Reference Intensity Image Generation

In this study, the reference intensity image is generated by performing cross-image median filtering over the *N* homomorphic filtered LDR images to exclude under-exposed or over-exposed regions in the *N* input LDR images. Cross-image median filtering is performed in a pixel-by-pixel manner over the *N* homomorphic filtered LDR images to generate the "median" image of $I_H^n(x,y)$, $n = 1,2,...,N$, as the reference intensity image, i.e.,

$$I_R(x,y) = \text{median}(I_H^1(x,y), I_H^2(x,y), \ldots, I_H^N(x,y)), \quad (3)$$

where $I_R(x,y)$ denotes pixel (x,y) of the reference intensity image and *N* is the number of homomorphic filtered LDR images.

E. Detail Extraction

Edge-preserving filters, such as the bilateral filter [18], weighted least squares optimization [15], and the guided filter [7], will not blur strong edges (without ringing artifacts) in the decomposition process. Using edge-preserving filtering, detail extraction is to decompose each processed LDR image with gamma correction into two (base and detail) layers and use the detail layer to compensate image details.

1) Local detail extraction

In this study, the guided filter [7], an edge-preserving filter, is used to decompose each processed LDR image with gamma correction into a base layer and a detail layer, i.e.,

$$I_C = I_L + \hat{I}_L, \quad (4)$$

where I_L and $\hat{I}_L$ denote the base and detail layers, respectively. Here, the guided filter is applied to the three (R, G, B) color component images when edges or fine details are not discriminative in a single color component image. It is assumed that the filtering output I_L is a linear

transformation of the guidance image I_C in a local window Ω centered at pixel *k*, i.e.,

$$I_L^k = a_k^T I_C^k + b_k, \quad (5)$$

where Ω is a 3×3 sliding window, I_L^k and I_C^k denote the 3×3 pixels of I_L and I_C in window Ω, T denotes the transpose operator, and a_k and b_k are two matrices of constants in window Ω, which can be directly estimated by linear regression as

$$a_k = (\mathbf{\Sigma}_k + \varepsilon\mathbf{U})^{-1}(\frac{1}{|\Omega|}\sum_{q\in\Omega} I_C^k P^k - \mu_k \overline{P}_k \mathbf{U}), \quad (6)$$

$$b_k = \overline{P}_k \mathbf{U} - a_k^T \mu_k, \quad (7)$$

where P is the input image, P^k denotes the 3×3 pixels of P in window Ω, $\overline{P}_k$ is the mean of the 3×3 pixels of P in window Ω, μ_k is the mean of the guidance image I_C in window Ω, $|\Omega|$ is the number of pixels in window Ω, ε is a parameter empirically set to 0.16, $\mathbf{\Sigma}_k$ is the 3×3 covariance matrix of the guidance image I_C in window Ω, and $\mathbf{U}$ is the 3×3 identity matrix.

2) Global detail extraction

In this study, WLS optimization [15] is used to decompose each processed LDR image with gamma correction and extract global details. Using matrix notation, we have

$$(\mathbf{I}_G - \mathbf{I}_C)^T(\mathbf{I}_G - \mathbf{I}_C) + \lambda(\mathbf{I}_G^T D_x^T \mathbf{A}_x D_x \mathbf{I}_G + \mathbf{I}_G^T D_y^T \mathbf{A}_y D_y \mathbf{I}_G), \quad (8)$$

where $\mathbf{A}_x$ and $\mathbf{A}_y$ are two diagonal matrices containing the smoothness weights $a_x(I_C)$ and $a_y(I_C)$, respectively, and matrices D_x and D_y are two discrete differentiation operators. The vector $\mathbf{I}_G$ minimizing (8) can be uniquely determined as the solution of

$$\mathbf{I}_C = \mathbf{I}_G(\mathbf{U} + \lambda(D_x^T \mathbf{A}_x D_x + D_y^T \mathbf{A}_y D_y)), \quad (9)$$

where $\mathbf{U}$ is the identity matrix.

F. Weighting Map Estimation and Refinement

For multiexposure image fusion, weighting map estimation is used to form the desired HDR image by keeping only the "best" regions (parts) in input LDR images. Weighting maps are determined by giving weights to the pixels of all LDR images. Here, two quality measures of spatial and cross-image consistency are used to estimate the weighting map of each processed LDR image with gamma correction, which is then refined by the cross bilateral filter.

1) Weighting map estimation of spatial consistency

In this study, four image quality measures of spatial consistency, namely, contrast, saturation, well-exposedness, and saliency, are used to estimate the weighting map of each processed LDR image with gamma correction. Three quality measures of spatial consistency, namely, contrast, saturation, and well-exposedness [1], are defined as

$$W_{\text{contrast}}^n(p) = \left|L(I_C^n(p))\right|, \quad (10)$$

$$W_{\text{saturation}}^n(p) = \sqrt{\frac{\sum_{i\in\{R,G,B\}}(I_{C_i}^n(p) - avg(I_C^n(p)))^2}{3}}, \quad (11)$$

$$W_{\text{exposed}}^n(p) = \exp(-\frac{(I_C^n - 0.5)^2}{2\sigma^2}), \quad (12)$$

where p denotes a pixel in the *n*-th processed LDR image with gamma correction, $|\cdot|$ denotes the absolute value, $L(\cdot)$ is a Laplacian filter with window size 3×3, $avg(\cdot)$ denotes the average of the RGB color components, and σ is empirically set to 0.2. The fourth quality measure of spatial consistency is saliency [4]. First, Laplacian filtering is applied to each processed LDR image with gamma correction to obtain the corresponding high-pass image H^n. Then, the local average of the absolute value of H^n is used to construct the saliency map W_{saliency}^n of H^n, which is computed as

$$W_{\text{saliency}}^n(p) = \left|H^n(p)\right| * G_{\sigma_g}(p), \quad (13)$$

where $H^n(p) = I_C^n(p) * L(p)$, $G_{\sigma_g}(\cdot)$ is a Gaussian filter of size 11×11, the standard derivation σ_g is empirically set to 5, and $*$ is the convolution operator. As a summery, the weighting map of spatial consistency is determined as

$$W_{spatial}^n(p) = W_{contrast}^n(p) \times W_{saturation}^n(p) \times W_{exposed}^n(p) \times W_{saliency}^n(p). \quad (14)$$

2) Weighting map estimation of cross-image consistency

Zhang and Cham [14] found that the gradient directions of the pixels in well-exposed regions are stable in different exposures. Therefore, the weighting map of cross-image consistency can be estimated by measuring gradient direction changes between each processed LDR image with gamma correction and the reference intensity image. Here, the first derivatives of a 2-D Gaussian function in *x* and *y* directions are used to extract the gradient information as

$$\theta^n(x,y) = \arctan(\frac{I_C^n(x,y) * \frac{\partial}{\partial y} G_{\sigma_d}(x,y)}{I_C^n(x,y) * \frac{\partial}{\partial x} G_{\sigma_d}(x,y)}), \quad (15)$$

where $I_C^n(x,y)$ denotes pixel (x,y) in the *n*-th processed LDR image with gamma correction and the standard derivation σ_d is empirically set to 0.5. The weighting map of cross-image consistency is estimated as the 1-D Gaussian function of the difference between $\theta^n(x,y)$ and that $\theta^{ref}(x,y)$ of the reference intensity image, i.e.,

$$W_{temporal}^n = \exp(\frac{-(\theta^n(x,y) - \theta^{ref}(x,y))^2}{2\sigma_t^2}), \quad (16)$$

 ISBN: 978-1-61208-320-9

where σ_t controls the influence of gradient direction changes and σ_t is set adaptively to the exposure quality of the reference intensity image as

$$\sigma_t(x,y)=\begin{cases}\alpha, 1-\gamma < I_R(x,y) < \gamma,\\ \beta, \text{otherwise},\end{cases} \quad (17)$$

where the well-exposed range is $[1-\gamma,\gamma]$ with the image range normalized to $[0,1]$. Here, α controls the influence of gradient direction changes detected based on the well-exposed pixels of the reference intensity image and $\beta(\geq\alpha)$ is used to reduce the influence of the detected gradient direction changes. In this study, the three parameters α, β, and γ are empirically set to 0.02, 0.9, and 0.9, respectively.

3) Weighting map refinement

The initial weighting map directly estimated as

$$W^n_{initial}=W^n_{spatial}\times W^n_{temporal}, \quad (18)$$

may be noisy. To cope with this problem, a cross bilateral filter based refinement [19-20] is employed so that neighboring pixels having similar intensities will have similar weight values, i.e.,

$$W^n_{final}(p)=\frac{\sum_{q\in\Omega} g_{\sigma_s}(\|p-q\|)g_{\sigma_t}(|I^n_C(p)-I^n_C(q)|)W^n_{initial}(q)}{\sum_{q\in\Omega} g_{\sigma_s}(\|p-q\|)g_{\sigma_t}(|I^n_C(p)-I^n_C(q)|)}, \quad (19)$$

where p is a pixel in the n-th weighting map, Ω is a 3×3 sliding window centered at p, and q is a pixel in window Ω. Here, the standard derivations σ_s and σ_t are empirically set to 5 and 5, respectively.

G. Image Fusion

Based on the N weighting maps of the N processed LDR images with gamma correction, a composite image is generated by fusing N processed LDR images with gamma correction. Using the multiresolution spline based scheme [17] to achieve seamless image fusion, the final HDR image I_F is obtained by integrating the composite image and the extracted detail image $I_{detail}=\{\hat{I}^1_L,\ldots,\hat{I}^n_L,\hat{I}^1_G,\ldots,\hat{I}^n_G\}$ as

$$I_F=\frac{\sum_{n=1}^{N} L\{I^n_C\}G\{W^n_{final}\}}{\sum_{n=1}^{N} G\{W^n_{final}\}+\xi}\cdot\exp(\max(I_{detail})), \quad (20)$$

where $L\{\cdot\}$ and $G\{\cdot\}$ denote the Laplacian and Gaussian pyramids, respectively, and ξ is a small constant to avoid singularity.

III. Experimental Results

In this study, the proposed approach is implemented using Matlab 7.10.0 (R2010a) on Intel Core i7-2700K CPU 3.5GHz-Microsoft Windows 7 platform with 8GB main memory. To evaluate the effectiveness of the proposed approach, four comparison approaches are employed, where the source codes of Mertens et al.'s approach [1] and Shen et al.'s approach [3] are directly employed, whereas and Zhang and Cham's approach [14] and Li et al.'s approach [10] are implemented in this study. Here, nineteen LDR image sequences with different numbers of LDR images are employed.

In this study, four objective image quality measures, namely, the structural similarity (SSIM) index [21] , the saturation, the blind image quality index (BIQI) [22], and the naturalness image quality evaluator (NIQE) [23], are employed. In terms of the SSIM index, saturation index, BIQI, and NIQE, the performance comparisons between the four comparison approaches and the proposed approach for the nineteen LDR image sequences are listed in Tables I~IV, respectively. The average performances of the proposed approach are better than those of four comparison approaches. To perform subjective evaluation, subjective scores, i.e., 1 (worst) up to 10 (best), are collected from eighteen people. Here, the final images of multiexposure image fusion of each LDR image sequence are shown on an EIZO LCD color monitor (S2402W) periodically (three seconds per image) and each viewer gives his subjective scores for different final images of each LDR image sequence. The subjective performance comparisons between the four comparison approaches and the proposed approach for the nineteen LDR image sequences are shown in Table V. As two illustrated experimental results shown in Figs. 2 and 3, the overall image quality of the final images of multiexposure image fusion of the proposed approach is better than those of the four comparison approaches. In Fig. 2, more texture details of windows are persevered in the final image of the proposed approach, whereas in Fig. 3, the contrast of books in the final image of the proposed approach is better than those of the four comparison approaches.

IV. Concluding Remarks

In this study, a multiexposure image fusion approach using homomorphic filtering and detail enhancement is proposed. Based on the experimental results obtained in this study, several observations can be found. (1) Based on Tables I~IV, on the average, the objective performance measures, namely, SSIM, saturation, BIQI, and NIQE, of the proposed approach are better than those of the four comparison approaches. (2) Based on Table V, the subjective evaluation of the final HDR images of the proposed approach is better than those of the four comparison approaches. (3) Based on Figs. 2-3, the final HDR images of the proposed approach are indeed better than those of four comparison approaches.

Acknowledgment

The work was supported in part by National Science Council, Taiwan, Republic of China under Grants NSC 102-2221-E-194-028-MY2 and NSC 102-2221-E-194-041-MY3.

References

[1] T. Mertens, J. Kautz, and F. V. Reeth, "Exposure fusion: a simple and practical alternative to high dynamic range photography," Computer Graphics Forum, vol. 28, no. 1, pp. 161-171, 2009.

 ISBN: 978-1-61208-320-9

[2] S. Li,. X. Kang, and J. Hu, "Image fusion with guided filtering," IEEE Trans. on Image Processing, vol. 22, no. 7, pp. 2864-2875, July 2013.

[3] R. Shen, I. Cheng, J. Shi, and A. Basu, "Generalized random walks for fusion of multi-exposure images," IEEE Trans. on Image Processing, vol. 20, no. 12, pp. 3634-3646, Dec. 2011.

[4] P. E. Debevec and J. Malik, "Recovering high dynamic range radiance maps from photographs," in Proc. of ACM SIGGRAPH, 1997, pp. 369-378.

[5] J. Duanm, M. Bressan, and C. Dance, "Tone-mapping high dynamic range images by novel histogram adjustment," Pattern Recognition, vol. 43, no. 5, pp. 1847-1862, May 2010.

[6] T. H. Wang, C. W. Fang, M. C. Sung, and J. J. Lien, "Photography enhancement based on the fusion of tone and color mappings in adaptive local region," IEEE Trans. on Image Processing, vol. 19, no. 12, pp. 3089-3105, Dec. 2010.

[7] K. He, J. Sun, and X. Tang, "Guided image filtering," in Proc. of European Conf. on Computer Vision, 2010, pp. 1-14.

[8] T. Kartalov, Z. Ivanovski, and L. Panovski, "Full automated exposure fusion algorithm for mobile platforms," in Proc. of IEEE Int. Conf. on Image Processing, 2011, pp. 361-364.

[9] S. Raman and S. Chaudhuri, "Bilateral filter based compositing for variable exposure photography," in Proc. of Eurographics, 2009, pp. 1-4.

[10] Z. G. Li, J. H. Zheng, and S. Rahardja, "Detail-enhanced exposure fusion," IEEE Trans. on Image Processing, vol. 21, no. 7, pp. 1-6, July 2012.

[11] M. Song, D. Tao, C. Chen, J. Bu, J. Luo, and C. Zhang, "Probabilistic exposure fusion," IEEE Trans. on Image Processing, vol. 21, no. 1, pp. 341-357, Jan. 2012.

[12] B. Gu, W. Li, J. Wong, M. Zhu, and M. Wang, "Gradient field multi-exposure images fusion for high dynamic range image visualization," Journal of Visual Communication and Image Representation, vol. 23, no. 4, pp. 604-610, May 2012.

[13] W. Zhang and W. K. Cham, "Gradient-directed multiexposure composition," IEEE Trans. on Image Processing, vol. 21, no. 4, pp. 2318-2323, April 2012.

[14] W. Zhang and W. K. Cham, "Reference-guided exposure fusion in dynamic scenes," Journal of Visual Communication and Image Representation, vol. 23, no. 3, pp. 467-475, April 2012.

[15] F. Zeev, F. Raanan, L. Dani, and S. Richard, "Edge-preserving decompositions for multi-scale tone and detail manipulation," ACM Trans. on Graphics, vol. 27, no. 3, pp. 1-10, 2008.

[16] J. Tumblin, J. K. Hodgins, and B. K. Guenter, "Two methods for display of high contrast images," ACM Trans. on Graphics, vol. 18, no. 1, pp. 56-94, 1999.

[17] P. J. Burt and E. H. Adelson, "A multiresolution spline with application to image mosaics," ACM Trans. on Graphics, vol. 2, no. 2, pp. 217-236, 1983.

[18] F. Durand and J. Dorsey, "Fast bilateral filtering for the display of high-dynamic-range images," ACM Trans. on Graphics (Proc. of ACM SIGGRAPH 2002), vol. 21, no. 3, pp. 257-266, July 2002.

[19] E. Eisemann and F. Durand, "Flash photography enhancement via intrinisic relighting," ACM Trans. on Graphics, vol. 23, no. 1, pp. 673-678, 2004.

[20] G. Petschnigg, R. Szeliski, M. Agrawala, M. Cohen, H. Hoppe, and K. Toyama, "Digital photography with flash and no-flash image pairs," ACM Trans. on Graphics, vol. 23, no. 1, pp. 664-672, 2004.

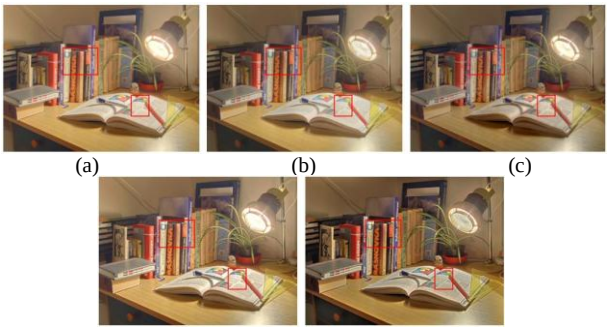
[21] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image qulaity assessment: from error visibility to structural similarity," IEEE Trans. on Image Processing, vol. 13, no. 4, pp. 600-612, April 2004.

[22] A. K. Moorthy and A. C. Bovik, "A two-step framework for constructing blind image quality indices," IEEE Signal Processing Letters, vol. 17, no. 5, pp. 513-516, May 2010.

[23] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a completely blind image quality analyzer," IEEE Signal Processing Letters, vol. 22, no. 3, pp. 209-212, March 2013.

(a) (b) (c) (d) (e)

Figure 2. The final images of multiexposure image fusion of the "Church" LDR image sequence: (a) Mertens et al. [1]; (b) Shen et al. [3]; (c) Zhang and Cham [14]; (d) Li et al. [10]; (e) proposed.



(a) (b) (c) (d) (e)

Figure 3. The final images of multiexposure image fusion of the "Desk Lamp1" LDR image sequence: (a) Mertens et al. [1]; (b) Shen et al. [3]; (c) Zhang and Cham [14]; (d) Li et al. [10]; (e) proposed.

TABLE I. IN TERMS OF SSIM, PERFORMANCE COMPARISONS BETWEEN THE FOUR COMPARISON APPROACHES AND THE PROPOSED APPROACH FOR THE NINETEEN LDR IMAGE SEQUENCES

LDR image sequences	Mertens et al. [1]	Shen et al. [3]	Zhang and Cham [14]	Li et al. [10]	Proposed
Aloe	68.2%	70.0%	71.0%	61.2%	**89.7%**
Ardeshir	76.5%	78.5%	77.6%	75.1%	**84.3%**
Belgium	40.6%	47.6%	46.6%	44.6%	**63.4%**
Bridge	80.1%	82.6%	77.9%	79.2%	**86.4%**
Church	70.3%	70.7%	**70.9%**	63.6%	67.7%
Desk Lamp1	80.8%	**81.6%**	80.5%	76.0%	79.6%
Desk Lamp2	77.6%	**79.0%**	72.5%	72.8%	75.1%
Flower8	63.5%	64.7%	63.5%	60.0%	**72.8%**
GrandCanal	62.2%	64.2%	61.5%	59.3%	**73.6%**
Hall	80.9%	**81.7%**	80.8%	79.4%	81.6%
HDRLab3	67.4%	67.2%	66.8%	66.5%	**80.4%**
House	42.1%	43.2%	41.9%	40.4%	**51.1%**
Kitchen	68.4%	71.2%	68.2%	64.5%	**76.8%**
Landscape	75.8%	76.4%	72.0%	74.0%	**85.4%**
Lighthouse	71.1%	72.8%	70.9%	68.0%	**80.5%**
Mountain	80.5%	**81.9%**	78.7%	76.1%	78.7%
Sofa	61.2%	61.4%	62.1%	57.5%	**64.1%**
Tree	70.1%	70.0%	**70.4%**	67.0%	65.0%
Wall	59.9%	61.1%	59.3%	59.4%	**67.4%**
Average	68.6%	69.8%	68.1%	65.5%	**74.9%**

TABLE II. IN TERMS OF SATURATION, PERFORMANCE COMPARISONS BETWEEN THE FOUR COMPARISON APPROACHES AND THE PROPOSED APPROACH FOR THE NINETEEN LDR IMAGE SEQUENCES

LDR image sequences	Mertens et al. [1]	Shen et al. [3]	Zhang and Cham [14]	Li et al. [10]	Proposed
Aloe	0.33	0.31	0.38	0.33	**0.45**
Ardeshir	0.38	0.32	0.40	0.38	**0.40**
Belgium	0.30	0.24	**0.33**	0.30	0.32
Bridge	0.13	0.12	0.14	0.13	**0.17**
Church	0.64	0.60	0.65	0.64	**0.69**
Desk Lamp1	0.39	0.38	0.39	0.39	**0.44**
Desk Lamp2	0.28	0.26	0.24	0.28	**0.31**
Flower8	0.28	0.24	0.27	0.28	**0.32**
GrandCanal	0.22	0.14	0.22	0.22	**0.25**
Hall	0.30	0.26	0.31	0.29	**0.33**
HDRLab3	0.27	0.24	**0.33**	0.27	0.32
House	0.33	0.25	0.35	0.33	**0.36**
Kitchen	0.46	0.39	0.47	0.46	**0.53**
Landscape	0.18	0.17	0.16	0.18	**0.22**
Lighthouse	0.39	0.34	**0.41**	0.39	0.37
Mountain	0.15	0.12	0.15	0.15	**0.18**
Sofa	0.82	0.78	**0.87**	0.81	**0.87**
Tree	0.15	0.13	0.15	0.15	**0.18**
Wall	0.20	0.11	0.19	0.20	0.17
Average	0.33	0.28	0.34	0.32	**0.36**

TABLE III. IN TERMS OF BIQI, PERFORMANCE COMPARISONS BETWEEN THE FOUR COMPARISON APPROACHES AND THE PROPOSED APPROACH FOR THE NINETEEN LDR IMAGE SEQUENCES

LDR image sequences	Mertens et al. [1]	Shen et al. [3]	Zhang and Cham [14]	Li et al. [10]	Proposed
Aloe	57.65	69.12	62.07	39.78	**32.16**
Ardeshir	24.31	29.39	**21.71**	27.41	23.47
Belgium	18.17	28.97	18.14	**12.50**	17.97
Bridge	31.33	32.64	31.53	26.94	**23.14**
Church	26.27	31.60	29.49	21.36	**19.59**
Desk Lamp1	21.97	27.22	27.23	**16.58**	17.33
Desk Lamp2	24.45	27.33	33.38	**16.41**	17.57
Flower8	29.51	33.79	29.76	25.07	**20.49**
GrandCanal	24.47	29.27	24.20	31.19	**23.99**
Hall	**18.09**	29.80	18.45	26.21	27.39
HDRLab3	29.90	30.66	31.22	**25.26**	30.84
House	**27.54**	31.66	27.98	32.78	27.80
Kitchen	26.65	31.42	27.38	22.80	**21.22**
Landscape	26.40	**25.32**	25.59	26.74	27.66
Lighthouse	**11.04**	23.64	11.36	13.57	11.98
Mountain	36.20	41.07	36.66	23.77	**17.10**
Sofa	38.55	39.79	**32.61**	41.20	39.97
Tree	26.06	27.00	27.91	14.06	**13.65**
Wall	23.68	26.71	24.30	24.49	**22.26**
Average	27.48	32.44	28.47	24.64	**22.92**

TABLE IV. IN TERMS OF NIQE, PERFORMANCE COMPARISONS BETWEEN THE FOUR COMPARISON APPROACHES AND THE PROPOSED APPROACH FOR THE NINETEEN LDR IMAGE SEQUENCES

LDR image sequences	Mertens et al. [1]	Shen et al. [3]	Zhang and Cham [14]	Li et al. [10]	Proposed
Aloe	2.99	**2.57**	2.88	2.77	2.83
Ardeshir	3.86	**3.18**	3.38	3.80	3.38
Belgium	2.38	2.45	2.24	2.19	**2.07**
Bridge	2.42	2.42	2.36	2.20	**2.13**
Church	2.09	1.90	**1.77**	1.98	1.86
Desk Lamp1	2.61	2.42	2.53	2.27	**2.25**
Desk Lamp2	2.70	2.54	2.47	2.34	**2.21**
Flower8	1.87	2.02	1.86	**1.65**	1.66
GrandCanal	2.33	2.27	2.27	2.16	**2.08**
Hall	2.49	2.39	2.41	2.39	**2.31**
HDRLab3	2.98	3.21	2.98	**2.57**	2.74
House	2.52	2.52	2.53	**2.33**	2.41
Kitchen	3.08	3.07	2.74	2.86	**2.65**
Landscape	3.06	**2.15**	3.01	2.79	2.63
Lighthouse	3.47	**2.89**	3.44	3.09	3.34
Mountain	2.07	2.14	2.17	**2.03**	2.06
Sofa	3.29	3.13	3.11	3.18	**3.05**
Tree	1.93	1.85	1.87	1.79	**1.75**
Wall	2.55	2.40	**2.15**	2.55	2.29
Average	2.67	2.50	2.54	2.47	**2.40**

TABLE V. IN TERMS OF SUBJECTIVE EVALUATION, PERFORMANCE COMPARISONS BETWEEN THE FOUR COMPARISON APPROACHES AND THE PROPOSED APPROACH FOR THE NINETEEN LDR IMAGE SEQUENCES

LDR image sequences	Mertens et al. [1]	Shen et al. [3]	Zhang and Cham [14]	Li et al. [10]	Proposed
Aloe	5.94	4.72	5.50	7.67	**8.94**
Ardeshir	**8.00**	5.67	5.56	7.33	6.78
Belgium	7.11	5.11	6.89	**7.22**	6.83
Bridge	4.94	6.28	5.78	6.56	**8.89**
Church	6.72	5.33	7.11	6.33	**8.67**
Desk Lamp1	6.44	5.67	5.61	7.11	**9.11**
Desk Lamp2	6.89	5.83	3.44	7.28	**8.22**
Flower8	6.17	5.44	6.00	7.11	**8.64**
GrandCanal	7.33	5.61	7.50	**8.33**	6.17
Hall	7.06	6.11	7.06	7.61	**8.14**
HDRLab3	6.17	5.22	7.28	6.50	**7.89**
House	8.06	5.28	7.28	**8.33**	8.06
Kitchen	7.67	5.28	6.22	7.67	**8.78**
Landscape	7.28	6.06	6.50	**7.89**	7.56
Lighthouse	7.50	5.56	7.22	**8.78**	7.00
Mountain	6.83	5.78	6.67	**8.33**	**8.33**
Sofa	7.50	5.78	6.06	7.17	**8.33**
Tree	5.56	5.61	6.22	7.11	**8.89**
Wall	6.94	6.67	6.39	7.22	**8.61**
Average	6.85	5.63	6.33	7.45	**8.10**

 ISBN: 978-1-61208-320-9

Keypoint of Interest Based on Spatio-temporal Feature Considering Mutual Dependency and Camera Motion

Takahiro Suzuki and Takeshi Ikenaga
Graduate School of Fundamental Science and Engineering
Waseda university
Tokyo, Japan
Email: takahir0@toki.waseda.jp, ikenaga@waseda.jp

***Abstract*—Recently, cloud systems start to be utilized for services to analyze user's data in the region of computer vision. In these services, keypoints are extracted from images or videos and the data is identified by machine learning with large database of cloud. Conventional keypoint extraction algorithms utilize only spatial information and many unnecessary keypoints for recognition are detected. Thus, the systems have to communicate large data and require processing time of descriptor calculations. This paper proposes a spatio-temporal keypoint extraction algorithm that detects only Keypoints of Interest (KOI) based on spatio-temporal feature considering mutual dependency and camera motion. The proposed method includes an approximated Kanade-Lucas-Tomasi (KLT) tracker to calculate the positions of keypoints and optical flow. This algorithm calculates the weight at each keypoint using two kinds of features: intensity gradient and optical flow. It reduces noise of extraction by comparing with states of surrounding keypoints. The camera motion estimation is added and it calculates camera-motion invariant optical flow. Evaluation results show that the proposed algorithm achieves 95% reduction of keypoint data and 53% reduction of computational complexity comparing a conventional keypoint extraction. KOI are extracted in the region whose motion and gradient are large.**

Keywords-Keypoint extraction; SIFT; Temporal analysis.

I. INTRODUCTION

Recently, Scale-Invariant Feature Transform (SIFT) [1] has attracted attention in computer vision because of its robustness in keypoint detection. Since SIFT can describe scale, rotation and illumination invariant features from images, matching between distinct images is executed accurately. By fully utilizing this characteristics, wide range of application is being considered. For example, it is used for object recognition [2], human or other object tracking [3], [4], recognizing panorama [5] and 3-D reconstruction [6]. In object recognition field, Bag-of-Features (BoF) was proposed by using combinations of SIFT descriptor. It generates one histogram from many keypoints which are extracted from one image. These are breakthroughes to recognize general objects. In addition, Support Vector Machine (SVM) was proposed as a machine learning algorithm. It utilizes non-linear kernel and classifies obtained keypoints with high accuracy. It needs to analyze a lot of keypoints to learn. It is also applied to recognition systems [7], [8] and shows high accuracy rate of recognition.

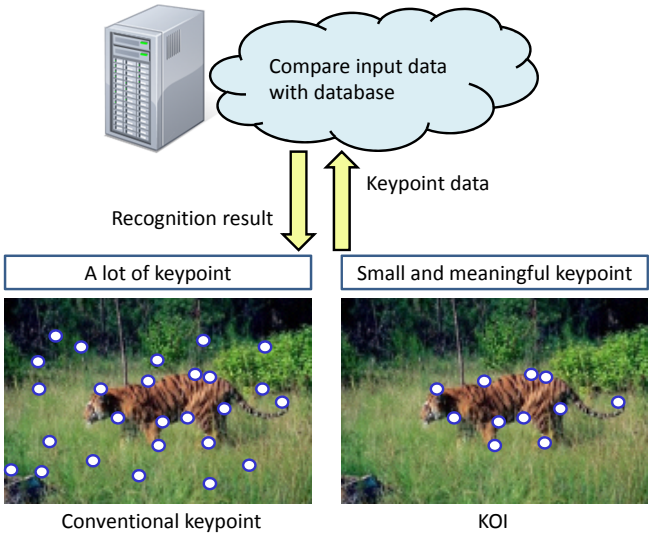


Figure 1: Conventional keypoints and KOI.

Recently, applications whose learned data is stored in cloud systems start to be released in relation to image recognition. A lot of keypoints are extracted from input images. All obtained keypoints are communicated with database. In this case, the amount of data makes it difficult to communicate data with high speed and stably. Recognition system needs only keypoints in object parts of interest. We call them Keypoints of Interest (KOI). If only KOI are extracted from input images, it generates two merits:

- Reduction of descriptor data communicated with cloud systems,
- Reduction of computational complexity of descriptor calculations.

Figure 1 shows the concept of this work. PCA-SIFT [9] which reduces the dimension of SIFT descriptor is also proposed. However, conventional keypoint extractions use only spatial data and extracts a lot of unnecessary keypoints. Other extended methods, SURF [10], GLOH [11], CSIFT [12] and ASIFT [13] , also utilize only spatial data.

This paper proposes a spatio-temporal keypoint extraction

 ISBN: 978-1-61208-320-9

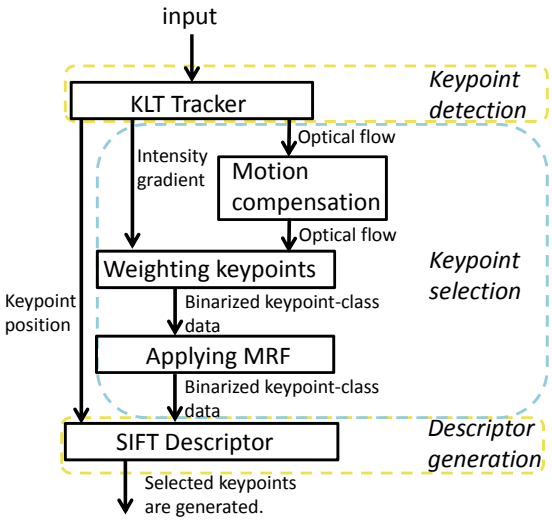


Figure 2: The flow of entire processing.

algorithm that detects only KOI based on spatio-temporal feature considering mutual dependency and camera motion. Candidate KOI are detected by Kanade-Lucas-Tomasi (KLT) tracker. KOI are selected by two kinds of features: intensity gradient and optical flow. However, important regions do not necessarily include information of large motion and gradient. Thus, we propose noise reduction by using Markov Random Field (MRF) that connects adjacent keypoints and determine the keypoint class. This algorithm extracts KOI that have many features including motion and intensity gradient. To deal with moving cameras, the compensation of motions is added. It can extract camera-motion invariant optical flow. Reduction of number of keypoints and computational complexity are evaluated in surveillance scenes.

Next section shows the proposed algorithm. Section III shows evaluation results. Finally, section IV concludes.

II. KEYPOINT EXTRACTION

Keypoint extraction is utilized for recognition and finding corresponding point between two images. The algorithm is divided into following two key parts.

- Keypoint detection
- Descriptor generation

The keypoint detection is the process which decides keypoint's position near characterized region. The SIFT descriptor generation calculates the histograms with information about neighboring region. These are common processes in all keypoint extraction methods. This paper utilizes low complexity keypoint extraction based on corner detection and plural images in database [14] as a conventional method. This algorithm also contains two key parts. It performs high-speed keypoint extraction maintaining almost same accuracy with SIFT.

III. KOI EXTRACTION BASED ON SPATIO-TEMPORAL FEATURE CONSIDERING MUTUAL DEPENDENCY AND CAMERA MOTION

In this section, we show the method that extracts KOI from an input movie. This paper proposes the following four methods.

- Calculation of optical flow by approximated KLT tracker
- Weighting keypoints by two elements: intensity gradient and optical flow
- Applying MRF to keypoint class
- Calculation of camera-motion invariant optical flow by camera motion estimation

The entire flow containing these methods is shown in Fig. 2. We choose the KLT tracker [15], [16] as a keypoint detection method because it simultaneously calculates positions of keypoints and optical flow which is utilized in keypoint selection part. This algorithm contains the keypoint selection part between the keypoint detection part and the descriptor generation part. In the keypoint selection part, first, this algorithm weights keypoint by two elements and calculates values which describe likelihood of KOI at each keypoint. Then, these values are arranged and keypoint class is determined by threshold. However, the results include a number of noise because important regions do not necessarily include motion and gradient. Thus, keypoints are connected by MRF and a graph cut algorithm is used to reduce noise from the output keypoints. In addition, to deal with moving cameras, the motion compensation is executed by camera motion estimation and camera-motion invariant optical flows are extracted. SIFT descriptor is calculated at only selected keypoints. This section shows each algorithm in more detail.

A. Calculation of optical flow by approximated KLT tracker

KLT tracker is one of the algorithms which detect keypoints and calculate optical flow. It uses filters which computes second-order difference of adjacent pixels. It needs to refer many adjacent pixels during detection from general images with noise. According to the number of referred pixels, the process time becomes very long. Thus, an integral image and box filters are utilized for speeding up.

First, keypoint detection by box filter is shown. We use Hessian matrix, $\mathbf{H}$, which is composed of second-order difference of adjacent pixels:

$$\mathbf{H} = \begin{bmatrix} L_{xx} & L_{xy} \\ L_{xy} & L_{yy} \end{bmatrix}. \quad (1)$$

In general, the elements are weighted by Gaussian function. However, it is not suitable for an integral image because weighting has to be determined at each pixel. Thus, this filter is approximated and it becomes easy to compute by integral image. An approximated filter is shown in Fig. 3. This approximation is also used by SURF. L_{xx}, L_{yy}, L_{xy} are obtained by filter process of integral image. After that, they are used to compute the function which decides corners. When the position is a corner, it satisfies the equation,

$$V = det(\mathbf{H}) - \omega tra(\mathbf{H}) > T, \quad (2)$$

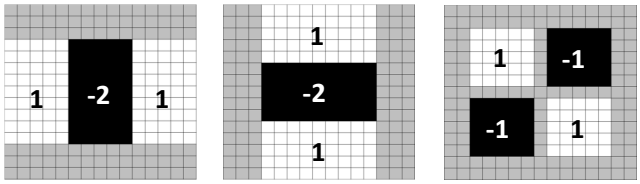


Figure 3: Approximated Filter (L_{xx}, L_{yy}, L_{xy}).

where ω is a parameter and T is a threshold. If the threshold becomes larger, corners decrease. It is adjusted to keep the number of keypoints optimal.

After the keypoint detection, The KLT tracker calculates optical flows. Optical flows are also calculated by second-order difference of adjacent pixels. Thus, the Hessian matrix is reused. The optical flow, $[u, v]$, is calculated by the equation:

$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} L_{xx} & L_{xy} \\ L_{xy} & L_{yy} \end{bmatrix}^{-1} \begin{bmatrix} L_{xt} \\ L_{yt} \end{bmatrix}. \quad (3)$$

L_{xt} and L_{yt} are calculated by the frame differences and filtering of x and y directions. In addition to gradient information, it utilizes the frame difference. This calculation also uses the integral image.

B. Weighting keypoints by two elements: intensity gradient and optical flow

In this paper, we choose two elements for weighting keypoints. The elements are intensity gradient and optical flow. With respect to intensity gradient, there is a high possibility that objects with many intensity gradients are the recognition targets. For example, book covers, posters and traffic signs are pointed out. With respect to optical flow, there is a high possibility that objects with motion are the recognition targets. For example, human, animals and vehicles are pointed out. Conventional keypoint extraction algorithms generally utilize only gradient information. Thus, it is expected to extract important keypoints including motion information if we use the temporal information. The weights of two elements are calculated at each keypoint which is obtained by the KLT tracker. The two different weights are normalized and summed up. This flow is described in Fig. 4.

Ways to obtain these values are shown next. The weight of intensity gradient is calculated by the Hessian detector [17]. The value, V, has already calculated in (1) and (2). V of (2) describes the strength of intensity gradient. It is obtained by the corner detection part of keypoint extraction. In other hand, the weight of optical flow is calculated by norm of optical flow. The value is obtained by (3). This calculation is a low complexity because the values have been already calculated. These two values are calculated at each keypoint and summed up after normalization. The weight is quantized data: $\mathbf{X} = \{x_1, x_2, \cdots, x_N\}$ where $x_i \in \{0, 1, \cdots, 255\}$. This weight data is binarized by threshold. Threshold is arranged by the number of KOI which the applications require. This process generates keypoint class $Y = \{y_1, y_2, \cdots, y_N\}$ at each

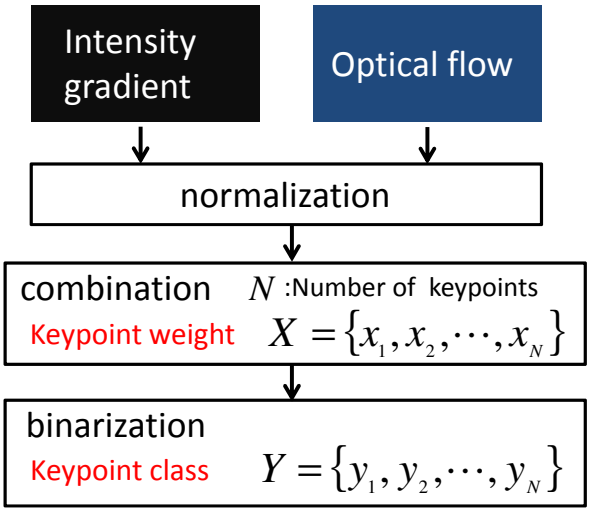


Figure 4: The flow of weight on keypoints.

keypoint where $y_i \in \{0, 1\}$. If the value of y_i is 1, the keypoint i is KOI.

However, important regions do not necessarily include motion and gradient. For example, gradient of human body is not large. It depends on their clothes. Several KOI can be extracted because the gradient of contour contains large values by proposed method in this section. Next, MRF is applied to reduce noise data and smoothing the keypoint class using the result of this section and adjacent keypoint data. The keypoint class is integrated on the each region.

C. Applying MRF to keypoint class

To solve the problem that keypoints in important region do not necessarily include large motion and gradient values, this paper applies MRF [18], [19] to keypoint class. MRF is usually used to reduce the noise of image in the region of image processing. MRF is the graph structure which represents the dependence between nodes. In this case, the nodes are keypoints and the dependency is defined in this section. Keypoints are connected by the weight of the distance from each other because the candidate keypoint whose adjacent keypoints are KOI tends to be KOI. The example of connections is shown in Fig. 5. In the circle, the keypoints are connected and they are easy to become same class. We utilize graph cut to reduce noise and determine keypoint classes. The graph cut algorithm changes keypoint class z_i and minimizes the energy equation:

$$E(Z) = \sum_i g_i(z_i) + \sum_{i,j} h_{ij}(z_j, z_i). \quad (4)$$

In this case, global solution is calculated because the keypoint class is binary. To solve this minimization problem, Min-Cut/Max-flow algorithm is used. Each function is defined as

$$g_i(z_i) = \lambda |y_i - z_i|, \quad (5)$$

 ISBN: 978-1-61208-320-9

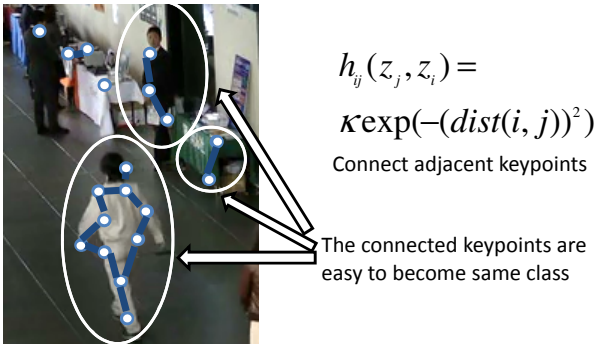


Figure 5: The connection of keypoints.

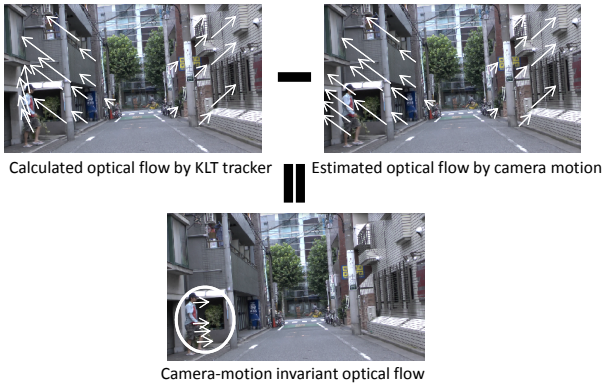


Figure 6: Calculation of camera-motion invariant optical flow.

$$h_{ij}(z_j, z_i) = \kappa \exp(-(dist(i,j))^2). \quad (6)$$

Equation (5) is data term. The outputted z_i is changed to approximate inputted y_i. Equation (6) is smoothing term. The strength of connection depends on distances between keypoints. We assume it is gaussian distribution. The nearer the keypoint distance, the stronger connection this function generates. $dist(i,j)$ represents the distance between keypoint i and keypoint j. λ and κ are parameters which are determined experimentally. They determine the strength of data term and smoothing term. If λ is larger than κ, the result approximates to the inputted data. If κ is larger than λ, the result approximates to the majority class of inputted keypoints. The calculated $Z = \{z_1, z_2, \cdots, z_N\}$ where $z_i \in \{0, 1\}$ is the output keypoint class. If the value of z_i is 1, the keypoint i is KOI. This calculation is faster than noise reduction of image which each node is a pixel because there are fewer nodes of the proposed method.

D. Calculation of camera-motion invariant optical flow by camera motion estimation

In the practical scene, cameras move like motion of pan or zoom. There are a number of scenes of zoom and pan in surveillance or in-vehicle camera. To apply this algorithm to moving cameras, this paper proposes a calculation of camera-motion invariant optical flow by camera motion estimation not to obtain large weight from the parts which do not move in fact. The overall flow is shown in Fig. 6 including zoom scenes. Optical flows are obtained by the KLT tracker at each keypoint. However, they include the influence of camera motion. For example, a number of optical flows which contain radical directions are generated like Fig. 6 from the parts which do not move in fact. Thus, we calculate the camera motion from these optical flows and the optical flows which are influenced by only camera motion is estimated. These are subtracted and the optical flow without influence of camera motion is obtained. Next, the method that estimates a camera motion is shown.

Figure 7: Test sequences.

A camera motion is estimated by all obtained optical flow by the KLT tracker. The motion vector of camera is defined as $\mathbf{T} = [t_x, t_y, t_z]^T$ where t_z represents the motion of a depth. The coordinate of the keypoint i is defined as $\mathbf{x}_i = [x_i, y_i, z_i]^T$. The optical flow, $\mathbf{v}_i = [u_i, v_i]^T$, is calculated by

$$u_i = \frac{x_i t_z}{z_i} - \frac{f t_x}{z_i},$$
$$v_i = \frac{y_i t_z}{z_i} - \frac{f t_x}{z_i}. \quad (7)$$

$\mathbf{T}$ is estimated by minimizing the function J:

$$J = \sum_{i=1}^{N} (\hat{u}_i - u_i)^T (\hat{u}_i - u_i), \quad (8)$$

where u_i is the calculated optical flow by (7) and $\hat{u}_i$ is the calculated optical flow by the KLT tracker. $\mathbf{T}$ is changed to minimize J. The result is substituted for (7) again. Estimated optical flows and obtained optical flows from inputted video are subtracted. The result is the camera-motion invariant optical flows.

 ISBN: 978-1-61208-320-9

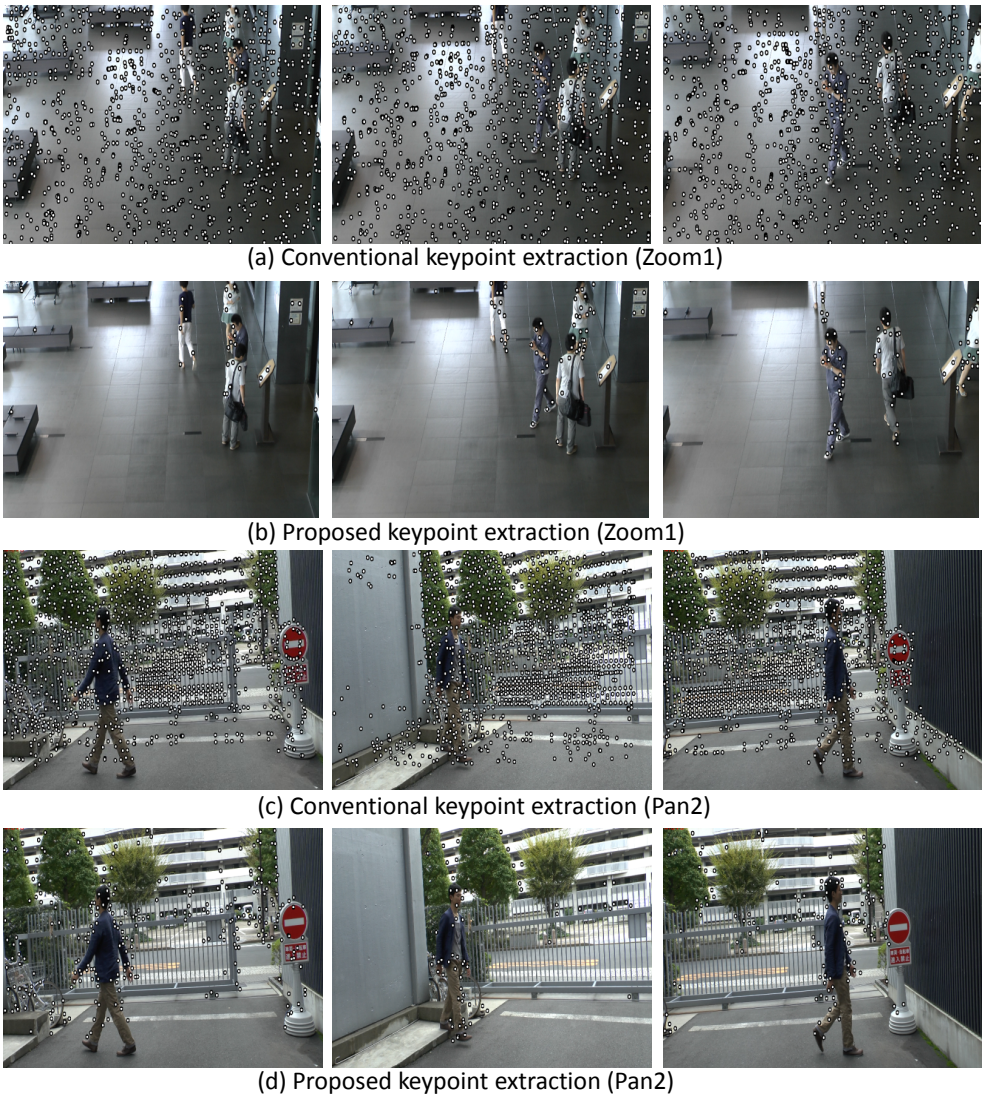


Figure 8: The comparison between (a) conventional keypoint extraction and (b) proposed algorithm in movie1, (c) conventional keypoint extraction and (d) proposed algorithm in movie2.

IV. EVALUATION RESULTS

This section shows evaluation results that compare the proposed method with the general keypoint extraction which utilizes the corner detector and SIFT descriptor introduced in section 2. The development environment on software is Visual Studio C++ 2008. CPU is Intel Core i7-2600 CPU 3.40GHz. The resolution of the video we used is Full-HD (1920×1080), 60 fps. In this paper, we evaluated six test sequences. We assume the surveillance cameras and each movie includes scenes that people walk on the path. In addition, they are taken by three camera motions including fixed cameras, zoom and pan to confirm effectiveness of camera motion estimation. The test sequences are shown in Fig. 7.

First, the number of keypoints which are detected by both methods are compared in Tab. I. It shows the average among all frames of the movie. The proposed algorithm achieves the 94%, 96%, 96%, 95%, 94% and 90% reduction of keypoints in all movies. Almost same results are obtained from Fixed1-2, Zopm1-2 and Pan1. However, Pan2 is low reduction comparing with others. It is considered that movie of Pan2 includes complex texture that has large illuminance gradients in background. In Fixed1, several keypoints in poster or other display items whose gradient is large are extracted. From these parts, many KOI are obtained in Pan2. In all video, the reduction of keypoints is confirmed. In addition, processing time is compared. The proposed algorithm reduces about 75%, 78%, 52%, 58%, 53% and 50% computational complexity than the conventional keypoint extraction in all movies. In the processing of Fixed1-2, the motion estimation parts are excepted. Thus, the complexity reduction is higher than others.

 ISBN: 978-1-61208-320-9

TABLE I: The number of keypoint and processing time between conventional method and proposed method.

		Conventional method	Proposed method
Fixed1	Number of keypoints	1192	66
	Processing time	754	190
Fixed2	Number of keypoints	1188	53
	Processing time	821	179
Zoom1	Number of keypoints	1202	47
	Processing time	759	367
Zoom2	Number of keypoints	1205	63
	Processing time	851	354
Pan1	Number of keypoints	1121	64
	Processing time	765	358
Pan2	Number of keypoints	1143	108
	Processing time	771	380

In other movies, motion estimation is calculated and almost same complexity reductions are obtained. In all video, the reduction of computational complexity is confirmed.

Figure 8 shows the video result of the conventional method and the proposed algorithm. The white circles are the keypoint obtained by each algorithm. It shows the proposal detects keypoints from only human which moves largely and outstanding texture whose gradient is large. In other video, the proposed algorithm extracts a number of keypoints from human body and the part including outstanding texture. By using only these keypoints, it is expected to analyze human or other outstanding object behaviors in surveillance and in-vehicle cameras combining motion features. In this algorithm, the parameters during weighting gradient and motion can be adjusted. Thus, if we want to obtain keypoints from only human, the parameters are adjusted to obtain the intended keypoints. The parameters can be determined by machine learning algorithm which learns correct KOI in advance. The correct KOI is determined by applications which this algorithm is applied. In this paper, application is not specified. Thus, the weight of gradient and motion is same in this evaluation.

V. CONCLUSION

Reduction of data amount of keypoints and reduction of computational complexity are required for cloud application. Conventional keypoint extractions utilize only spatial information and extract a lot of unnecessary keypoints. This paper proposes a keypoint selection algorithm from many keypoints including unnecessary ones based on spatio-temporal feature considering mutual dependency and camera motion. The proposed method includes an approximated KLT tracker to calculate the positions of keypoints and optical flow. It calculates the weight at each keypoint using two kinds of features: intensity gradient and optical flow. It reduces noise by comparing with states of surrounding keypoints. Optical flows are compensated by camera motion estimation and it calculates camera-motion invariant optical flows. Evaluation results show that the proposed algorithm achieves about 95% reduction of keypoints and 53% reduction of computational complexity. KOI are extracted in human bodies that move widely and the objects whose gradient is large. This algorithm is expected to be applied to surveillance cameras and in-vehicle cameras when they start to utilize cloud system to recognize motion of humans or other outstanding objects.

ACKNOWLEDGMENT

This work was supported by KAKENHI (23300018).

REFERENCES

[1] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” Int.Journal of Computer Vision, 60, pp. 91-110, 2004.

[2] D. G. Lowe, “Object recognition from local scale-invariant features,” In International Conference on Computer Vision, Corfu, Greece, pp. 1150-1157, 1999.

[3] Yuji Tsuzuki, Hironobu Fujiyoshi, Takeo Kanade, “Mean Shift-based Point Feature Tracking using SIFT,” Journal of Information Processing Society, Vol. 49, No. SIG 6, pp. 35-45, 2008.

[4] Huiyu Zhou , Yuan Yuan , Chunmei Shi, “Object tracking using SIFT features and mean shift,” Computer Vision and Image Understanding, v.113 n.3, pp. 345-352, 2009.

[5] Matthew Brown and David G. Lowe, “Recognising panoramas,” International Conference on Computer Vision, pp. 1218-25, 2003.

[6] S. Agarwal, N. Snavely, I. Simon, S. M. Seitz, R. Szeliski, “Building rome in a day,” In ICCV, 2009.

[7] M. Pontil and A. Verri, “Support Vector Machines for 3D Objec Recognition,” IEEE Trans. on Pattern Analysis and Machine Intelligence, vol.20, no.6, pp.637-646, 1998.

[8] G. Guo, S.Z. Li and K. Chan, “Face Recognition by Support Vector Machines,” Proceedings of the 4th IEEE Int. Conf. on Automatic Face and Gesture Recognition, pp.195-201, 2000.

[9] Y. Ke, R. Sukthankar, “A More Distinctive Representation for Local Image Descriptors,” Proceedings of Computer Vision and Pattern Recognition, pp. 506-513, 2004.

[10] H.Bay, T.Tuytelaars and L. V. Gool, “SURF: speeded up robust features,” In ECCV, pp. 404-417, 2006.

[11] Krystian Mikolajczyk, Cordelia Schmid, ”A performance evaluation of local descriptors,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 10, 27, pp. 1615–1630, 2005.

[12] A.E. Abdel-Hakim, A.A. Farag,”Csift: a sift descriptor with color invariant characteristics,” Proceedings of the Computer Vision and Pattern Recognition, pp. 1978–1983, 2006.

[13] J. M. Morel and G. Yu, “Asift: A new framework for fully affine invariant image comparison,” SIAM Journal on Imaging Sciences, 2, 2, pp. 438-469, 2009.

[14] Takahiro Suzuki, Takeshi Ikenaga, “Low Complexity Keypoint Extraction Based on SIFT Descriptor and Its Hardware Implementation for Full-HD 60 fps Video,” IEICE TRANSACTIONS on Fundamentals of Electronics, Communications and Computer Sciences, Vol. E96-A, No. 6, pp. 1376-1383, 2013.

[15] C. Tomasi and T. Kanade, “Detection and Tracking of Point Features,” Carnegie Mellon University Technical Report CMU-CS-91-132, April 1991.

[16] Bruce D. Lucas and Takeo Kanade, “An Iterative Image Registration Technique with an Application to Stereo Vision,” International Joint Conference on Artificial Intelligence, pp. 674-679, 1981.

[17] Beaudet, P. R., “Rotational invariant image operators,” In Proceedings of the 4th International Joint Conference on Pattern Recognition (ICPR), pp. 579-583, 1978.

[18] K. Tanaka, “Statistical-mechanical approach to image processing,” J. Phys. A: Mathematical and General, vol. 35, no. 37, pp.R81-R150, 2002.

[19] A. Willsky, “Multiresolution Markov models for signal and image processing,” Proceedings of the IEEE, vol. 90, no. 8, pp. 1396-1458, 2002.

 ISBN: 978-1-61208-320-9

A Learning Platform for 3D Digital Single Lens Reflex (DSLR) Camera

Seow Hui, Saw
Department of Computer Science
Faculty of Information and Communication Technology
Universiti Tunku Abdul Rahman
Kampar, Perak, Malaysia
e-mail: shsaw@utar.edu.my

Win Khai, Cheah
Department of Information Systems
Faculty of Information Communication Technology
Universiti Tunku Abdul Rahman
Kampar, Perak, Malaysia
e-mail: marcuscheah831@gmail.com

***Abstract*—The use of Digital Single Lens Reflex (DSLR) camera is becoming increasingly popular. However, the selling price for these cameras is relatively high, especially for users who are on a tight budget and just getting started with photography. Therefore, we have developed a multimedia learning platform for Digital Single Lens Reflex (DSLR) camera to provide an interactive environment between users and the camera before they make their purchasing decisions. The main contribution in our platform is the modeling of a 3D DSLR camera equipped with the necessary lens and operation buttons that allows a realistic experience for users in real-time interactions with a mouse. Furthermore, a visual simulator is built into the platform to show how different aperture and shutter settings affect the outcome of a digital photo. First, the 3D DSLR model is constructed using 3Ds MAX, a professional modeling software. It is then exported to a Virtual Reality Markup Language (VRML) file format supported by the powerful 3D interactive software called WireFusion. Finally, the entire object is published on the internet with additional learning tools and multimedia resources to enhance learning delivery. Several screenshots of the platform are also presented in this paper.**

Keywords-3D DSLR Camera; interactive visualization: multimedia application;

I. INTRODUCTION

Digital Single Lens Reflex camera (DSLR) cameras have become popular since the existence of digital photography in 1975 [12]. People like to capture and preserve their precious memories in high quality with this gigantic yet powerful device. Nevertheless, the features of Digital Single Lens Reflex (DSLR) camera are complicated to master in order to take good photos, and users, generally, lack of hands-on experience before they decide to purchase any devices. Moreover, selling prices are relatively high for the beginner who lives on a tight budget.

In response to these issues, the existing online learning platform consists of two types of contents: (1) online lessons and tutorials [2][4], (2) online lessons and tutorials, a virtual simulator that allows user to control the settings, for example, the shutter speed, in order to dictate the desire effects and outcomes of the given photos [15], and (3) online lessons and tutorials, a virtual simulator with some instant advices and tips to review the chosen settings of the camera in order to improve the quality of the capture photo [14].

Although these developed platforms are effective, a DSLR camera is displayed in a 2-dimensional picture, or, in other words, image. A plain image does not provide an interactive visualization for the user. Hence, a virtual testing on the camera is impossible before the purchasing decisions.

This paper presents a combination of multimedia, 3-dimensional graphics visualization and a web-based application in order to provide a realistic experience with easy access for the user in haptic interaction using only mouse operations [13]. Users are able to visualize the 3-dimensional DSLR camera by rotating it in different angles. The main components for the camera, such as Universal Serial Bus (USB) port can be operated and controlled by a simple click to view the descriptions (Fig. 3).

Furthermore, we have included both the tutorials and simulation tools to be in the same platform to facilitate the learning process, so users have an in-depth understanding of the respective functionalities. The users are able to operate and control the buttons of the three-dimensional model by setting different functions of the camera and capture photos to see the effects.

In addition, the platform also provides an instant feedback to the users on the improvements they would do to make the photos better. This is an effective and efficient way to learn and improve their skills rather than reading from reference books.

The 3-dimensional DSLR is first modeled with the famous professional modeling software called 3Ds MAX including the camera accessories, such as lens and a tripod. The created 3-dimensional model is then exported as Virtual Reality Markup Language (VRML) (.wrl) to be used and designed in WireFusion. WireFusion is an authoring tool for creating 3-dimensional interactive presentations that are supported in any web browser in order to provide easy access to users anywhere and anytime.

The remainder of this paper is organized as follows: Section 2 summarizes the previous platforms and implementation relevant to our approach. Section 3 explains the details and the interfaces of our platform. Users' surveys are presented in Section 4. Section 5 concludes the paper and refers to future work.

 ISBN: 978-1-61208-320-9

Figure 1. The first page of our platform [13].

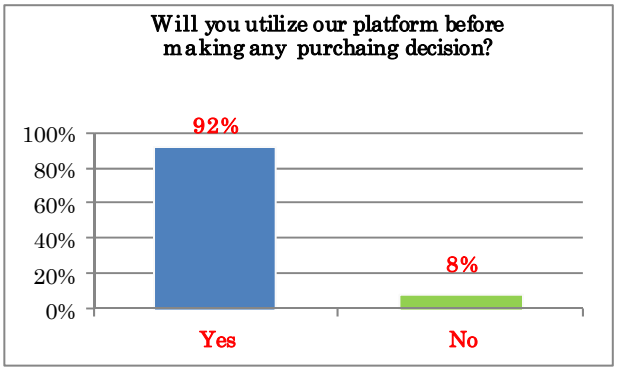


Figure 2. 92% of the respondents are interested to utilize our platform before making the purchase of a DSLR.

II. RELATED WORKS

There are several learning platforms for DSLR camera that can be found in the literature. These are the common online methods to teach beginners how to use DSLR cameras.

A survey report [1] has proposed that the best way to teach the students is through online learning. The reason is because the course materials are accessible all year long and can be shared globally across different time zones.

As described in Section I, there are several online learning platforms. McHugh [2] provides a comprehensive and all the necessary resources to guide users on the terminologies, concepts and techniques for photographing with a DSLR camera. These resources include tutorials, photography tools and fora among users to share their technical knowledge. This online platform is suitable for the beginners, while [4] is mainly for advance users since their tutorials focus on photographing skills. However, both of the online platforms lack of interaction features and simulator.

There are a few online sites combined both tutorials and an additional virtual simulator which allows user to control the basic settings and view the outcome of the photo without providing any instant feedback to compare the used settings [14][15].

Wendzich [14] offers a simple yet straightforward tutorials about the concepts and terminologies of DSLR camera. In addition, they have a real-time visual feedback so user is able to see the results of their changes without committting to the shutter. Unfortunately, user has no option to change the photo and no advices are provided in the simulator.

The proposal in [15] allows user to control only the light of the camera: shutter and aperture to view the effects of the capture photo. Although this online platform offers a very valuable teaching but an incomplete virtual simulator.

Similarly, the solution in [16] has a very simple explanation for the features of camera, very limited controls for the settings and offered only three images in their virtual simulator.

On the other hand, CameraSim [3] consists of a virtual simulator that displays the effects of the photo under various settings such as the aperture, shutter speed, International Standards Organisation (ISO) speed and etc. Nonetheless, they are lack of teaching contents, no proper advices for the settings in the virtual simulator unless the user purchases the licences, and a limited number of example shots.

Canon [17] developed an online website to describe the basic manual settings for visual effects. They provide only one image to dictate the outcome and effects of the basic settings of DSLR camera. However, they give a detailed review and feedback for the capture photo.

As summarized above, the existing online platforms implemented the necessary contents to train the beginners and give them a grasp of the photographing foundation. Nevertheless, there is no integration between the user and the DSLR camera that can enhance their tactile involvement and intutive understanding before they make a purchasing decision. Our platform [13] includes complete tutorials, the virtual simulator with extra features, and the highlight of the platform is the interactive 3D DSLR camera for the visualization assistance.

Sun et al. [5] describe a method for designing and developing a 3D virtual DSLR camera based on VRML and JavaScript languages. The authors built a 3D model by using 3Ds Max, professional modeling software, and exported it to a document supported by VRML. The purpose of this is to publish the program online for users to access anytime. Although the mentioned 3D model is available in this online learning platform, there is no visual simulator to visualize the effect of the captured photos. In addition, the platform lacked of guidance regarding the features for 3D DSLR cameras.

Recently, Moser et al. [6] presented a tangible photography education system that invites people to explore and learn about the technical settings involved in DSLR cameras. The user can physically manipulate and explore how these settings interact with one another to produce different types of photographic expressions. However, it would be more instructional for the the user to physically experience the 3D DSLR camera instead of in an online, electronic (and, thus, indirectly) context.

III. INTERFACES

A survey was conducted before the platform is established to determine the usefulness of 3D DSLR learning platform. Based on the results collected from 50 people, 92% of them will utilize the 3D DSLR learning platform before they decide to purchase a real one, as shown

 ISBN: 978-1-61208-320-9

in Fig. 2. Furthermore, a learning platform equipped with a 3D DSLR is yet a novel idea. Thus, we decided to develop this web application, as shown in Fig. 1.

A. Home

A user is able to interact with 3D DSLR camera displayed in Fig. 3 by rotating and zooming operation with the mouse. The 3D DSLR camera is first modeled by 3Ds MAX that can be downloaded for free [7] to be used for educational purposes. There are many tutorials and references which can be found in [8] and [10]. Actually the 3D DSLR camera is created based on the guidance in [9] and texture is added to the created model for a realistic effect. Whereas the bitmap image used for texturing is created with the use of Adobe Photoshop. The position for both texture image and the 3D DSLR camera must be matched exactly. Following this, a diffuse bitmap is selected in the material editor map of 3Ds MAX to diffuse the bitmap image on certain components in the 3D DSLR camera.

The second step involves, exporting the 3D content that was created into VRML97 (.wrl) file format that is supported by WireFusion 5.0. This is powerful software that is privately held by Demicron [11], a Swedish company dedicated to create an interactive 3D technology for used to visualize products and market them on the internet using Java programming language. It is a visual tool with event-driven programming. Users are able to download the manual from [11] and to learn the fundamental through a series of hands-on examples.

When users click the “Start” button shown in Fig. 3, the hotspot labels for camera components are presented as seen in Fig. 1. A brief description and a short animation are displayed when one of these hotspots is selected. For instance, the lens, mode dial, USB, lens release button and others are being labeled. Thus, users are able to gain an insight into the functionality of these components. These interactivities are completed using WireFusion 5.0.

Lastly, the end product of 3D DSLR camera is deployed as Java Applet in a Hyper Text Markup Language (HTML) file in order to publish to internet after all the necessary adjustment.

B. Tool Kits

This new tab page called “Tool Kits” is included in our learning platform that is not able to be found in other existing DSLR websites (see Fig. 5). Apart from DSLR camera, there are various accessories that can be used to improve the quality of photographs. The most common DSLR accessories including extra batteries, Ultra Violet (UV) filter, cleaning kits, lens, tripod and etc.

This page offers an interactive environment with descriptive tutorials that are useful for effectively understanding the functions of these accessories. Users are able to attach or change them by simply a mouse click. For instance, in Fig. 6, when user clicks on either macro lens or normal lens, the changing process is shown in real time. In order to see the effects of these lenses, user is required to

Figure 3. Several components of 3D DSLR camera are labeled with hotspots.

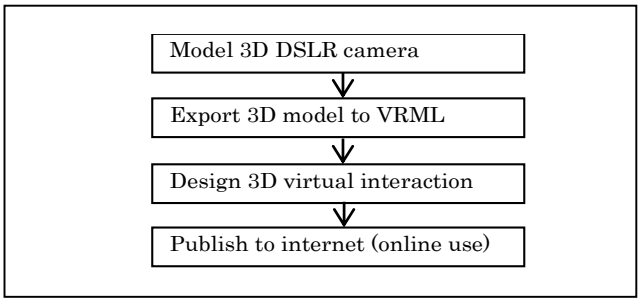


Figure 4. Several components of 3D DSLR camera are labeled with hotspots.

Figure 5. Toolkits: Lens tutorial for normal lens and macro lens.

press the button with camera icon. The background of the resulting photo is blurred and the subjects are sharped when macro lens is used as shown in Fig. 7 (left). Macro lens is suitable for portrait capturing. On the other hand, normal lens is suitable to capture a neutral photo as shown in Fig. 7 (right).

Apart from the lenses, users are able to navigate with the tripod (see Fig. 8). Fig. 9 shows the differences between photos taken with tripod and without tripod during nighttime. Thus, users are able to make a decision whether to purchase these additional accessories without actually visiting the shops.

 ISBN: 978-1-61208-320-9

Figure 6. The changing process: (Left) Macro lens is changed. (Right) Normal Lens is changed.

Figure 7. (Left) Using macro lens and the subjects are sharper than the background. (Right) Using normal lens.

Figure 8. Tool Kits: Tripod tutorial.

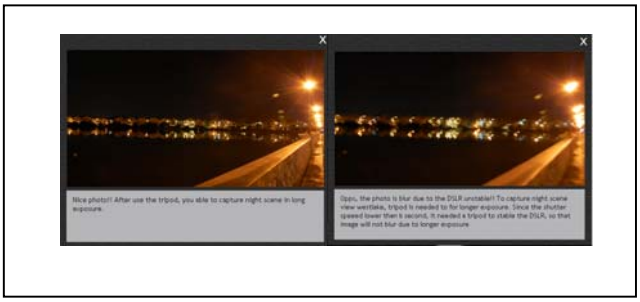


Figure 9. Picture taken at night. (Left) With tripod. (Right) Without tripod

C. Tutorials

The essential part of developing an intuition for photography is to master the use of camera exposure. A photograph's exposure determines how light or dark an image will appear when it's been captured by the camera. This is determined by just three camera settings or in other words the exposure triangle defined as follows:

- Aperture: A camera's aperture setting controls the area over which light can pass through the camera lens. It is specified in terms of f-stop value, which can at times be counterintuitive, because the area of the opening increases as the f-stop decreases. In photographer slang, when someone says they are "stopping down" or "opening up" their lens, they are referring to increasing and decreasing the f-stop value, respectively.
- Shutter Speed: A camera's shutter determines when the camera sensor will be opened or closed to incoming light from the camera lens. The shutter speed specifically refers to how long this light is permitted to enter the camera. Shutter speed and exposure time refer to the same concept, where a faster shutter speed means a shorter exposure time.
- ISO: The ISO speed determines how sensitive the camera is to incoming light. Similar to shutter speed, it also correlates 1:1 with how much the exposure increases or decreases. However, unlike aperture and shutter speed, a lower ISO speed is almost always desirable, since higher ISO speeds dramatically increase image noise. As a result, ISO speed is usually only increased from its minimum value it the desired aperture and shutter speed aren't otherwise obtainable.

The visual simulator is included in the tutorial to improve users' skills. With this simulator, users are not only limited to the facts and definitions for these settings, but they have the opportunity to visualize the effects that are caused by these exposure triangle.

As described above, aperture and shutter speed are closely related to each other. Thus, we placed their settings in the same page as shown in Fig. 10. Their settings are contradictory to get a perfect shot for the photo (see Fig. 11).

In addition, the visual simulator furnishes with an evaluator that gives useful advices and comments based on the chosen settings for further improvements. For instance, the resulting photo is dark when the user slide to F/ 22 for aperture and 1/ 1000 seconds for shutter speed using respective slider buttons (see Fig. 12). The evaluator advised user to decrease the aperture value and increase the shutter speed for a better shot, as shown in Fig. 11.

On the other hand, the resulting photo is bright, as shown in Fig. 13, when user selects F/ 5.6 for aperture and 1/ 125 seconds for shutter speed. In this situation, the evaluator suggested user to increase the shutter speed.

Technically, we applied the straightforward if-then-else logical operators for these evaluations. The following lists out all the advices and comments in the learning platform.

1) Oopps! The exposure is too long until the photo looks blank and bright. You should adjust higher aperture value and higher shutter speed value.

2) The exposure is too long due to low shutter speed value. Windmill: look dim. Background building: disappeared. You should adjust higher shutter speed value.

 ISBN: 978-1-61208-320-9

3) The exposure is too long due to low shutter speed value. Windmill: look ambiguous. Background building: disappeared. You should adjust higher shutter speed value.

4) The exposure is too long due to low shutter speed value. Windmill: look ambiguous. Background building: bright and unclear. You should adjust higher shutter speed value.

5) The exposure is too long due to low shutter speed value. Windmill: Hardly visible and ambiguous. Background building: bright and unclear. You should adjust higher shutter speed value.

6) The exposure is too long due to low shutter speed value. Windmill: Visible but out of focus. Background building: bright and unclear, yet better. You should adjust higher shutter speed value.

7) Not bad, the photo look nice! But…Windmill: look ambiguous and out of focus. Background building: clear but bright. You should adjust higher shutter speed.

8) The exposure is too long! The photo is too bright. Windmill: bright and out of focus. Background building: clear but bright. You should adjust higher shutter speed value.

9) Good! But the photo is too bright. Windmill: look bright and out of focus. Background building: clear but bright. You should adjust higher shutter speed value.

10) Great! You are achieving the correct exposure, but…Windmill: out of focus. Background building: clearest. You should adjust higher shutter speed to freeze the windmill movement.

11) Well done! The photo looks perfect. You are achieving the correct exposure and the motion of windmill is frozen. Windmill and building background: Look clearer and sharper.

12) Perfect! The photo looks perfect. You are achieving the correct exposure and the motion of windmill is frozen. Windmill and building background: look clearer and sharper.

13) Not bad! You have freeze the motion of the windmill, but the photo looks dark due to underexposed. Windmill and building background: Look clearer and sharper, the darker. You should adjust lower aperture value.

14) Great! You have freeze the motion of the windmill. Windmill and building background: Look clearer and sharper, but it looks dark. You should adjust lower aperture.

15) Great! The photo looks perfect. You are achieving the correct exposure and the motion of windmill is frozen. Windmill and building background: look clearer and sharper.

16) The photo looks dark due to the area over which light can pass through the camera lens is small, thus, underexposed. You should be decreasing the f-stop value/ adjust to lower aperture value to increase light passes through the camera lens.

17) Wow, feel like raining?! The photo looks dark due to the area over which light can pass through the camera lens is small, thus, underexposed. You should be decreasing the f-stop value/ adjust to lower aperture value to increase light passes through the camera lens.

Table I summarizes the comments that will display when the user selects the respective camera settings.

TABLE I. THE COMMENTS AND ADVICES BASED ON THE CAMERA SETTINGS.

CAMERA EXPOSURE		
Shutter speed (in seconds)	***Aperture***	***Comments/ Advices***
Less than or equal to 1/8	F/5.6	*1)*
	F/8.0	*2)*
	F/11.0	*3)*
	F/16.0	*4)*
Less than or equal to 1/30	F/5.6 F/8.0	*5)*
	F/11.0	*6)*
	F/16.0 F/22.0	*4)*
Less than or equal to 1/60	F/5.6 F/8.0	*4)*
	F/11.0 F/16.0 F/22.0	*7)*
Less than or equal to 1/125	F/5.6	*8)*
	F/8.0	*9)*
	F/11.0 F/16.0 F/22.0	*10)*
Less than or equal to 1/500	F/5.6	*11)*
	F/8.0	*12)*
	F/11.0	*14)*
	F/16.0 F/22.0	*13)*
Less than or equal to 1/1000	F/5.6	*12)*
	F/8.0	*15)*
	F/11.0	*16)*
	F/16.0 F/22.0	*17)*

Meanwhile, there is a separate visual simulator for ISO settings, as shown in Fig. 14 (left). As described above, ISO is to determine the amount of incoming light of the camera. Users are able to adjust the ISO values using the main command dial of 3D DSLR camera as displayed in Fig. 14 (Right). Furthermore, there are two types of the sample photos: indoor and outdoor. Fig. 15 (left) visualizes the effects of ISO 200 which caused darkness for this indoor photo. The evaluator suggested that user should use higher ISO values for a better picture, whereas in Fig. 15 (right) illustrated the effects of ISO 3200 that were too high for the outdoor photo. In this situation, the evaluator adviced to lower the ISO value for a balanced effects. The evaluations and comments for indoor and outdoor pictures are listed in Table II.

 ISBN: 978-1-61208-320-9

TABLE II. THE COMMENTS BASED ON THE ISO SETTINGS FOR INDOOR AND OUTDOOR PICTURE.

ISO Settings	**INDOOR PICTURE**	**OUTDOOR PICTURE**
	Comments/ Advices	***Comments/ Advices***
200	*Not Bad! But for indoor low light, you might want to adjust a higher ISO values.*	*Well done! ISO 200 is suitable for outdoor picture.*
400	*Well Done! ISO 400 is suitable for indoor picture.*	*Perfect! For outdoor with low light, ISO 400 is the most suitable.*
800	*Perfect! ISO 800 is the most suitable for indoor low light*	*Hmmm, this photo considered okay, but looked bright, you might want to lower down the ISO value*
1600	*Hmmm, this photo is considered okay, but consists of high ISO noise; you might want to adjust lower ISO values.*	*The photo is bright; you might want to lower down the ISO value.*
3200	*You should lower down the ISO values; the photo is too bright and noisy.*	*You should lower down the ISO value; the photo is too bright and noisy.*
6400	*Opps, ISO is too high! You should lower the ISO values for a better picture.*	*Opps, ISO value is too high! The photo is too bright and noisy. You should lower down the ISO values for a better picture.*

Figure 10. A visual simulator with aperture and shutter speed slider buttons.

Figure 11. A perfect shot with aperture F/2.8, shutter speed 1/1000 seconds.

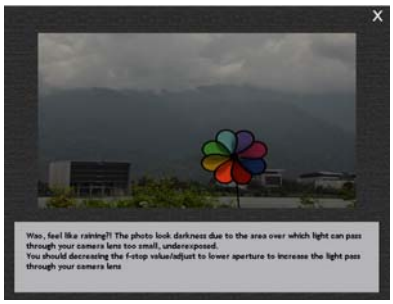


Figure 12. An example of bad settings (too dark) of aperture F/22 and shutter speed 1/1000.

Figure 13. Another example of bad settings (too bright) of aperture F/5.6 and shutter speed 1/125

Figure 14. A visual simulator with ISO settings. (Left) The starting page. (Right) Use the main command dial to adjust the ISO values.

Figure 15. (Left) An indoor photo with low ISO values. (Right) An outdoor picture with high ISO values.

The sample photos in the visual simulator are taken using the real DSLR camera. Therefore, the quality and the accuracy of the photos are preserved.

 ISBN: 978-1-61208-320-9

IV. USERS' SURVEYS – RESULTS

We have collected only 10 participants for the survey due to time limitation. Based on the results, 100% of them thought they'll have better learning about DSLR cameras through our website [13].

Moreover, we have compared our platform with several existing platform that has a visual simulator [3], [15], [16] and [17] and found out that, 89% agreed that they'll learn better if the website furnishes with a 3D model. Furthermore, the tools provided are very useful to them and, 64% enjoyed it since we have additional tools tutorials and more reality with a 3D model.

89% will visit our website before they decide to purchase a real one while 67% will revisit our website again after they bought a DSLR camera for tutorial purposes.

V. DISCUSSIONS AND FUTURE WORKS

In summary, our 3D DSLR learning platform [13] provides an interactive environment for users especially the beginners. This platform allows them to gain experience in photographing and the opportunity to test the virtual DSLR before their purchasing decision is made. There is a local company which runs a mentorship-driven startup accelerator for student entrepreneurs in Asia attracted to this benchmarking and claimed that it has a great potential as a gateway for the DSLR camera manufacturers to advertise their model in three-dimensional. They can utilize this platform to advertise their new model/s and attract their buyers.

This platform can be furthered enhanced by adding more features such as video tutorials and, higher levels of photography lessons since this platform is targeted to the beginners. Additionally, more tools can be added to the tool kits page such as flashlights, memory cards, batteries and others.

Besides the extension in commercialization, there is also a research value in this platform. We can have an investigation into modifying the software's presentation according to the user's experiences with the camera incorporating some learning algorithm. We also hope to increase the number of participants for our surveys in order to obtain a stronger justification.

ACKNOWLEDGMENT

This project implementation belongs to the Bachelor Degree student, Cheah Win Khai for his completion of the Final Year Project. This is supported by the Faculty of Information and Communication Technology in Universiti Tunku Abdul Rahman. He graduated with distinction in April 2013.

REFERENCES

[1] H. Tinti-Kane, J. Seaman, and J. Levy, "Social Media in Higher Education: The Survey," Babson Survey Research Group, Massachusetts, 2013.

[2] S. McHugh, "Cambridge in Colour: A Learning Community for Photographers," 2013. [Online]. Available: http://www.cambridgeincolour.com. [Retrieved: April, 2013].

[3] J. Arnold, "CameraSim: DSLR Photography Demystified," 2012. [Online]. Available:http://www.camerasim.com. [Retrieved: May, 2012].

[4] G. Laing, "DSLR Tips: Support this site by shopping via our partner stores," 2007. [Online]. Available: http://www.dslrtips.com/support/Support_DSLR_Tips.shtml. [Retrieved: May, 2012].

[5] Y. Sun, L. Liu, Q. Li, "Design and Development of 3D Virtual DSLR Camera Based on VRML and JavaScript," Computer Science and Education (ICCSE 2010), 5th International Conference, Aug 24– 27, 2010, pp. 1380– 1384, doi: 10.1109/ICCSE.2010.5593752.

[6] K. Moser, M. Kiechle, K. Ryokai, "Photocation: tangible learning system for DSLR photography," Proc. CHI EA '12 Extended Abstracts on Human Factors in Computing Systems, 2012, pp. 1691– 1696, doi: 10.1145/2212776.223694.

[7] Autodesk support center, "AUTODESK: Download center," [Online]. Available: http://students.autodesk.com/?nd=download_center. [Retrieved: January, 2013].

[8] Autodesk Media and Entertainment Education Video Series, "3ds Max Learning Channel," [Online]. Available: http://www.youtube.com/user/3dsMaxHowTos. [Retrieved: January, 2013].

[9] C. Hayes, "Model a Detailed, High-Poly Camera in 3ds Max," 2nd July 2009. [Online]. Available: http://cg.tutsplus.com/tutorials/autodesk-3d-studio-max/model-a-detailed-high-poly-camera-in-3ds-max/. [Retrieved: January, 2013].

[10] Autodesk Media and Entertainment Education Video Series, "Learning Center," 2011. [Online]. Available: http://www.students.autodesk.com/?nd=learning_center. [Retrieved: January, 2013].

[11] Learning Center Group, "Demicron: INTERACTIVE 3D, Learning Center," 1996-2011. [Online]. Available: http://www.demicron.com/support/learning/index.html. [Retrieved: January, 2013].

[12] D. Prakel, "The Visual Dictionary of Photography," AVA Publishing, 2010, pp. 91. [Retrieved: 24 July 2013].

[13] W.K. Cheah, "3D DSLR Learning Platform, " 2013. [Online]. Available: http:// www.webkhai.com/html_New/app/home/home.html. [Retrieved: October 2013].

[14] L. Wendzich, "Building a DSLR Simulator," 16th June 2013. [Online]. Available: http://writing.ludwignz.com/post/building-a-dslr-simulator. [Retrieved: 10 December 2013].

[15] Photonhead, "Digital Camera Tips and Reviews: The Essence of Modern Film and Digital Photography," June 2004. [Online]. Available: http://www.photonhead.com. [Retrieved: 12 October 2013].

[16] T. Strand, "Aperture, shutter and ISO value, " 2008-2013. [Online]. Available: www.kamerasimulator.se/eng/. [Retrieved: 12 October 2013].

[17] Canon Canda Inc, "Learn | Canon Explains Exposure" 2011. [Online]. Available: http://www.canonoutsideofauto.ca/learn/. [Retrieved: December 2013].

 ISBN: 978-1-61208-320-9

Extended Successive Elimination Algorithm for Fast Optimal Block Matching Motion Estimation

Changryoul Choi and Jechang Jeong
Dept. of Electronics and Communication Engineering Hanyang University
Seoul, Korea
e-mail : denebchoi@gmail.com & jjeong@hanyang.ac.kr

***Abstract*—In this paper, we propose an extended successive elimination algorithm (SEA) for fast optimal block matching motion estimation (ME). By reinterpreting the typical sum of absolute differences measure, we can obtain additional decision criteria whether to discard the impossible candidate motion vectors. Experimental results show that the proposed algorithm reduces the computational complexity up to 19.85% on average comparing with the multilevel successive elimination algorithm. The proposed algorithm can be used with other SEA to improve the ME performance.**

Keywords-motion estimation; successive elimination algorithm; block matching

I. INTRODUCTION

Motion estimation (ME) has been widely used in many video applications ranging from video compression to video segmentation, video tracking, etc [1]. The block matching algorithm (BMA) for ME is the most popular and is deployed in many video compression standards [2-3] because of its simplicity and effectiveness. In BMA, a frame is partitioned into a number of rectangular blocks and a motion vector for that block is estimated within its search range in the reference frame by finding the closest block of pixels according to a certain matching criterion, e.g. the sum of absolute differences (SAD), the sum of squared differences (SSD), etc. The full search algorithm (FSA) can give the optimal estimation of the motion in terms of minimal matching error by checking all the candidates within the search range, but the huge computational complexity of the FSA makes it inadequate for the real-time applications. Thus, many fast but optimal algorithms which provide the same accuracy as the FSA are proposed including the fast searching and the fast matching algorithms in the literature [4-9].

The fast matching technique aims at reducing the whole calculations of the matching criterion for each candidate block by comparing only a subset of the pixels in the block. In lossy fast matching, it predicts the total matching criterion based on the statistical property of it or calculates the matching criterion of the sub-sampled patterns only [4-5]. In lossless fast matching, it is based on the simple idea that since the total distortion is monotonically increasing by progressively adding the partial distortion of the pixel-by-pixel differences, a block comparison can be safely terminated if the accumulated sum of partial distortions becomes greater than the up-to-date minimum distortion [6].

The fast searching is the technique to reduce the number of searching points. In lossy fast searching, it exploits some general properties of the typical images. One of these properties is that since the real world video sequences usually vary slowly, their distribution of motion vector is center-biased. Also, many lossy fast searching algorithms assume that the matching error planes are convex to reduce the searching points [7]. In lossless fast searching, based on some mathematical inequalities, it calculates the lower bound of the matching criterion (typically the SAD) and safely skips the impossible candidate motion vectors [8-9]. In [8], the successive elimination algorithm (SEA) was proposed. The SEA estimates the SAD by calculating a block sum difference which is the lower bound of the true SAD. The estimated SAD is used as a decision criterion whether to eliminate the impossible candidate. Gao et al. extended the idea of SEA to multilevel successive elimination algorithm (MSEA) [9]. By splitting the block into small sub-blocks, closer estimations to the true SAD are given and the estimated SAD is used as decision criteria whether to discard the impossible candidate.

And there are some techniques that exploit different matching criteria instead of the classical SAD or SSD were also proposed to make the faster computation of the matching criteria using bit-wise operations [10].

In this paper, we propose an optimal fast searching algorithm which is an extension of the typical SEA. By reinterpreting the typical SAD measure, we can obtain additional decision criteria for pruning out bad motion vectors. The rest of this paper is organized as follows. Section II gives a review of the fast searching ME algorithms. Section III presents our proposed algorithm. Experimental results and analyses are provided in Section IV. Finally, Section V provides conclusions.

II. PREVIOUS ALGORITHMS

The SAD between the current block (CB) and the reference block (RB) is usually used as a matching criterion for ME and is defined as follows:

 ISBN: 978-1-61208-320-9

$$SAD(x,y)=\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}\left|CB(i,j)-RB(i+x,j+y)\right| \quad (1)$$

where $N\times N$ is the motion block size and (x,y) is the candidate motion vector within the search range.

A. Successive Elimination Algorithm

Let CB_0 and RB_0 be the sum norms of the CB and the RB which are defined as follows:

$$\begin{aligned} CB_0 &= \sum_{i=0}^{N-1}\sum_{j=0}^{N-1} CB(i,j) \\ RB_0 &= \sum_{i=0}^{N-1}\sum_{j=0}^{N-1} RB(i+x,j+y) \end{aligned} \quad (2)$$

where (x,y) is the candidate motion vector within the search range. Note that the sum norms of the RB can be calculated efficiently over the whole image [8]. From the triangle inequality, we can easily derive the following inequality:

$$\left|CB_0-RB_0\right| \le SAD \quad (3)$$

This inequality (3) shows the key idea of the SEA. If the calculated sum norm of the RB of position (x,y) and the CB does not satisfy the inequality (3) (in this case, the SAD is replaced by the up-to-date minimum SAD in the searching process), this means that the candidate motion vector of position (x,y) is not the optimal motion vector and the calculation of SAD is unnecessary and skipped [8].

B. Multilevel Successive Elimination Algorithm

The MSEA extended the idea of SEA to a multilevel case [9]. By splitting the block into small sub-blocks, closer estimations to the true SAD are given and the estimated SAD is used as decision criteria whether to discard the impossible candidate. First, the block is partitioned into four sub-blocks of size $N/2\times N/2$. Then each sub-block is partitioned into four sub-blocks of size $N/4\times N/4$. This process can be repeated until the size of the sub-blocks becomes 2×2. The maximum level of such partition is $L_{max} = \log_2 N-1$ when the motion block size is $N\times N$. Let $CB_l^{(k)}$ and $RB_l^{(k)}$ be the sum norms of the kth sub-blocks at the lth level in the CB and the RB, respectively. Based on (3), we can obtain

$$\sum_{k=0}^{N_l-1}\left|CB_l^{(k)}-RB_l^{(k)}\right| \le \sum_{k=0}^{N_{l+1}-1}\left|CB_{l+1}^{(k)}-RB_{l+1}^{(k)}\right| \le SAD \quad (4)$$

where N_l is the number of sub-blocks at the lth level. From (4), we can attain monotonically increasing SAD estimation values as the level increases. Therefore, more and more impossible candidates can be eliminated earlier as the level increases [9].

III. Proposed Algorithm

Let a_i and b_i be the sequences of length $N\times N$. As in (3), we can summarize the SEA as follows:

$$SAD=\sum|a_i-b_i| \ge \left|\sum(a_i-b_i)\right| = \left|\sum a_i-\sum b_i\right| \quad (5)$$

And we can think of the absolute operation as follows:

$$\sum|a_i-b_i| = \sum sign(a_i-b_i)\times(a_i-b_i) \quad (6)$$

In case of SEA, the equality holds only when either of the following is satisfied:

$$\begin{aligned} sign(a_i-b_i) &= 1,\ \forall i \\ sign(a_i-b_i) &= -1,\ \forall i \end{aligned} \quad (7)$$

which is rare the case. Using (5) and (6), we can think of the following inequality:

$$\begin{aligned} \sum|a_i-b_i| &= \sum sign(a_i-b_i)\times(a_i-b_i) \\ &\ge \left|\sum \alpha_i\times(a_i-b_i)\right| \end{aligned} \quad (8)$$

where α_i takes on either 1 or -1. Therefore, to estimate the SAD values more precisely, the sequence α_i must be almost the same as the signs of (a_i-b_i)s or totally inverted signs of them. Due to summations of the smaller sub-blocks and absolute operations in (4), the MSEA can be considered as the process of estimating the true SAD by forcing the signs of difference sequence into the true signs of it more and more as the level increases. In this case, the difference of the sum norms in each level plays as a one and only candidate for estimating signs of the sequences.

The basic idea of the proposed algorithm is that we can enhance the estimation of the true SAD by allowing more candidates (whose signs of the difference sequence are different) in each level. To this end, we generate more candidates which are generalized sum norms to enhance the estimation accuracy. We mean a generalized sum norm as the summation of the pixels according to the predefined basic (addition or subtraction) arithmetic.

Since the systematic and efficient calculation of the sum norms is one of the main reasons for computational complexity reduction in the SEA, the number of possible candidates for the generalized sum norms which can be calculated efficiently is limited. Fig. 1 shows the structured image arithmetic templates taking into consideration the computational complexity. We call the generalized sum

 ISBN: 978-1-61208-320-9

norms of Fig. 1 (a) as original sum norms, Fig. 1 (b) as horizontal sum norms, Fig. 1 (c) as vertical sum norms, and Fig. 1 (d) as diagonal sum norms.

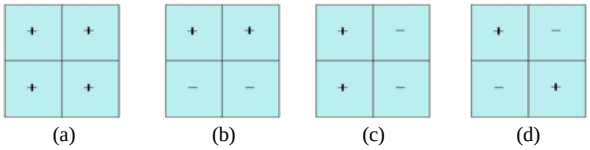


Figure 1. Structured image arithmetic templates (a) template used in SEA and MSEA (b), (c), and (d) structured image arithmetic templates horizontal, vertical and diagonal, respectively.

Due to the increased candidates (horizontal, vertical, and diagonal), the inequality in (4) is changed as:

$$\sum_{k=0}^{N_l-1} \max_{z\in\{o,h,v,d\}} \left|CB_{l,z}^{(k)} - RB_{l,z}^{(k)}\right| \le \sum_{k=0}^{N_{l+1}-1} \max_{z\in\{o,h,v,d\}} \left|CB_{l+1,z}^{(k)} - RB_{l+1,z}^{(k)}\right| \qquad (9)$$

where $CB_{l,z}^{(k)}$ and $RB_{l,z}^{(k)}$ are the generalized sum norms (o represents original sum norms, h represents horizontal sum norms, v represents vertical sum norms, and d represents diagonal sum norms) of the kth sub-blocks at the lth level in the CB and the RB, respectively. The proposed algorithm is almost the same as the MSEA except the following two differences. The first one is that the proposed algorithm gives more stop conditions than the typical MSEA (in the first level, there are 3 more stop conditions than the typical MSEA. And since these stop conditions are of the first level, if one of these stop conditions is satisfied, it reduces the total computational complexity more). The second one is that due to (9), we can attain closer SAD bound than the typical MSEA at the same level reducing the computational complexity.

Note that there is a trade-off between the increased computational complexity of calculating the increased generalized sum norms and the reduced computational complexity by pruning out the bad motion vectors in an early stage. And the increased stop conditions can be pros or cons. To estimate the actual effects of the generalized sum norms in computational complexity, we calculated the computational complexity of the proposed algorithm in terms of the searched points.

TABLE I. COMPUTATIONAL COMPLEXITY OF THE GENERALIZED SUM NORMS CALCULATIONS WHEN THE MOTION BLOCK SIZE IS 16×16 AND THE SEARCH RANGE IS ±16

	MSEA	**Proposed - All**	**Proposed - H only**	**Proposed - V only**
Overhead (points)	3.74	11.21	5.60	5.60

Table I shows the computational complexity of the generalized sum norms in terms of the SAD calculations when the motion block size is 16×16 and the search range is set to ±16. To be specific, the table shows the number of total operations divided by one SAD calculation operations, in this case we assume that the computational complexity of addition and subtraction operations and that of the absolute operations are the same.

Table II shows the average search points of the proposed algorithms when all the generalized sum norms are used (3rd column), only the horizontal sum norms are used (4th column), and only the vertical sum norms are used, respectively. Note that the original sum norms were also used in all of the proposed algorithms. The test sequences are of CIF-size and 100-frame long. The motion block size is 16×16 and the search range is set to ±16. The computational complexity of the table I is also considered.

TABLE II. AVERAGE SEARCH POINTS OF ALGORITHMS FOR CIF SEQUENCES WHEN THE MOTION BLOCK SIZE IS 16×16 (100-FRAME, SEARCH RANGE IS ±16)

	MSEA	**Proposed - All**	**Proposed - H only**	**Proposed - V only**
stefan	63.65	82.62	58.81	75.91
football	62.36	81.54	60.53	72.17
foreman	51.87	65.92	48.92	53.58
mobile	40.62	49.62	40.46	40.52
coastguard	63.25	76.65	46.83	92.75
container	42.00	53.15	37.00	47.66
flower	109.42	128.72	89.35	112.45
Avg.	**61.88**	**76.89**	**54.56**	**70.72**

From the table, we can see that using all the generalized sum norms and using only the vertical sum norms does not provide any computational reduction. Therefore, the final proposed algorithm uses the following inequality for additional stop conditions for the typical MSEA.

$$\sum_{k=0}^{N_l-1} \max_{z\in\{o,h\}} \left|CB_{l,z}^{(k)} - RB_{l,z}^{(k)}\right| \le \sum_{k=0}^{N_{l+1}-1} \max_{z\in\{o,h\}} \left|CB_{l+1,z}^{(k)} - RB_{l+1,z}^{(k)}\right| \qquad (10)$$

IV. EXPERIMENTAL RESULTS

The performance of the proposed algorithm was compared with the MSEA in terms of the SAD calculation points. The full frames of the 7 CIF (352 × 288) sequences, 4 SD sequences (704 × 576), and 4 HD sequences (1280 × 720) were used as test sequences. Motion block sizes were all 16×16 and the searching processes were in spiral order.

Tables III and IV show the average searched points of CIF and SD sequences when the search range is ±16 and ±32, respectively. The proposed algorithm outperforms the MSEA. To be specific, the performance of the proposed algorithm is better than that of the MSEA by 20.0% on average when the search range is ±32 for SD sequences. Table V shows the average searched points of HD sequences when the search range is ±16, ±32 and ±64, respectively. Note that since the ME accuracy of the

 ISBN: 978-1-61208-320-9

proposed algorithm is the same as that of the MSEA, we omit the ME accuracy in terms of the peak signal to noise ratio (PSNR) here.

TABLE III. AVERAGE SEARCH POINTS OF ALGORITHMS FOR CIF SEQUENCES WHEN THE MOTION BLOCK SIZE IS 16×16 (FULL-FRAME, S.R. = SEARCH RANGE)

	MSEA		**Proposed – H only**	
	S.R. = ±16	**S.R. = ±32**	**S.R. = ±16**	**S.R. = ±32**
Stefan	100.33	170.26	90.20	148.87
football	57.10	126.27	56.28	118.21
Foreman	57.75	97.84	57.9	93.77
Mobile	35.06	78.55	34.52	72.73
Coastguard	54.61	128.21	41.21	89.51
Container	43.25	84.04	37.94	73.68
Flower	91.46	162.33	77.57	138.36
Avg.	**62.79**	**121.07**	**56.52**	**105.02**

TABLE IV. AVERAGE SEARCH POINTS OF ALGORITHMS FOR SD SEQUENCES WHEN THE MOTION BLOCK SIZE IS 16×16 (FULL-FRAME, S.R. = SEARCH RANGE)

	MSEA		**Proposed - H only**	
	S.R. = ±16	**S.R. = ±32**	**S.R. = ±16**	**S.R. = ±32**
ICE	157.23	378.44	119.9	279.72
CITY	48.25	94.85	48.00	91.16
CREW	115.55	223.4	107.25	207.32
SOCCER	93.08	176.96	81.9	150.71
Avg.	**103.53**	**218.41**	**89.26**	**182.23**

TABLE V. AVERAGE SEARCH POINTS OF ALGORITHMS FOR HD SEQUENCES WHEN THE MOTION BLOCK SIZE IS 16×16 (FULL-FRAME, S.R. = SEARCH RANGE)

	MSEA			**Proposed - H only**		
	S.R. = ±16	**S.R. = ±32**	**S.R. = ±64**	**S.R. = ±16**	**S.R. = ±32**	**S.R. = ±64**
Big -Ships	53.57	100.53	174.95	45.26	78.89	158.67
Crew	178.57	365.27	718.07	156.08	321.34	648.33
Preak -ness	25.00	49.69	91.71	24.55	43.97	108.6
Sheriff	39.30	77.06	173.64	35.45	64.23	134.22
Avg.	**74.11**	**148.14**	**289.59**	**65.34**	**127.11**	**262.46**

As can be seen from the tables, the proposed algorithm outperforms the MSEA in terms of the computational complexity. To be specific, the proposed algorithm reduces the computation complexity by 19.85% when the search range is ±32 for SD sequences. Since the proposed algorithm can be easily plugged in other SEA based ME algorithms, additional computational reduction for this seemingly marginal gain can be expected.

V. CONCLUSIONS

By reinterpreting the typical SAD measure, we proposed an extended SEA for fast optimal block matching ME in this paper. By allowing more candidates in estimating the true SAD, we can obtain additional decision criteria whether to discard the impossible candidate motion vectors. Experimental results show that the proposed algorithm reduces the computational complexity up to 19.85% on average comparing with the typical MSEA. Since the proposed algorithm can be easily adopted in other SEA based ME algorithms, additional computational reduction can be expected. Therefore, our future research will be focused on merging the proposed algorithm with the previous SEA based ME algorithms to reduce the computational complexity a lot without degrading the ME accuracy.

REFERENCES

[1] Z. He, C. Tsui, K. Chan, and M. Liou, “Low-power VLSI design for motion estimation using adaptive pixel truncation,” IEEE Trans. Circuits Syst. Video Technol., vol. 10, no. 5, Aug. 2000, pp. 669-678.

[2] Information Technology - Coding of Audio Visual Objects - Part 2: Visual, JTC1/SC29/WG11, ISO/IEC 14496-2 (MPEG-4 Visual), 2002.

[3] Advanced Video Coding for Generic Audiovisual Services, ITU-T Recommendation H.264, May 2005.

[4] C. Cheung and L. Po, “Adjustable partial distortion search algorithm for fast block motion estimation,” IEEE Trans. Circuits Syst. Video Technol., vol. 13, Jan. 2003, pp. 100-110.

[5] Y. Wang, Y. Wang, and H. Kuroda, “A Globally Adaptive Pixel-Decimation Algorithm for Block-Motion Estimation,” IEEE Trans. Circuits Syst. Video Technol., vol. 6, Jan. 2000, pp. 1006-1011.

[6] C. Choi and J. Jeong, “New sorting-based partial distortion elimination algorithm for fast optimal motion estimation,” IEEE Trans. Consumer Electron., vol. 55, no. 4, Nov. 2009, pp. 2335-2340.

[7] X. Jing and L. Chau, “An efficient three-step search algorithm for block motion estimation,” IEEE Trans. Multimedia, vol. 6, no. 3, Jun. 2004, pp. 435–438.

[8] W. Li and E. Salari, “Successive elimination algorithm for motion estimation,” IEEE Trans. Image Process., vol. 8, Jan. 1995, pp. 105-107.

[9] X. Q. Gao, C. J. Duanmu, and C. R. Zou, “A multilevel successive elimination algorithm for block matching error estimation,” IEEE Trans. Image Process., vol. 9, Mar. 2000, pp. 501-504.

[10] C. Choi and J. Jeong, “Enhanced Two-bit Transform Based Motion Estimation via Extension of Matching Criterion,” IEEE Trans. Consumer Electron., vol. 56, no. 3, Aug. 2010, pp. 1883-1889.

 ISBN: 978-1-61208-320-9

An Efficient Event Definition Framework for Retail Sector Surveillance Systems

Fahad Anwar[1,3], Ilias Petrounias[2,3], Sandra Sampaio[4,3]
WMIC[1], Manchester Business School[2]
The University of Manchester[3]
Manchester, UK
fahad.anwar@manchester.ac.uk
ilias.petrounias@manchester.ac.uk

Vassilis Kodogiannis[4,5], Tim Morris[4,3]
School of Computer Science[4]
University of Westminister[5]
London, UK
sandra.sampaio@manchester.ac.uk, kodogiv@wmic.ac.uk,
tim.morris@manchester.ac.uk

***Abstract*— Event representation models provide a framework in which we can reason about events so as to interpret the collective behaviour of objects over time and space domains. Many are context-specific and lack flexibility when faced with unstructured video. In the past many efforts have been made to define a comprehensive event description framework (EDF), which can provide a framework to develop ontologies for semantic annotation of video events. However, it is observed that there are some areas of event modelling that were not fully explored. Hence, we extended and modified the EDF and proposed the extended version of it (EDF^E). Following are some of the major extensions we have proposed in EDF^E. 1) EDF^E extends the entity representation model of EDF by introducing three new entity classes: that of text entity, virtual entity and internal entity. II) EDF^E introduces a new set of predicates for describing more complex event scenarios and facilitating the event detection process. It also introduces granularity as a feature of temporal predicates to capture the temporal association between sub-events. III) It introduces the event evidence feature to capture the full evidence for the detected events. IV) The data structure of EDF is extended and modified to capture the properties of EDF^E and to store the results of the event detection process. V) We model complex events from real world surveillance videos using the proposed EDF^E.**

Keywords-*Multimedia event modelling; intelligent surveillance system; multimedia event annotation and data mining.*

I. INTRODUCTION

Due to the flexibility and expressive power of rich semantic models, they provide a solid ground for any semantic video event model to be used for structured and unstructured multimedia content such as surveillance videos. In [1], Gupta et. al. presented the VIMSYS model, although it mainly focuses on image contents, it provides the basic platform on which many video content management models were later based on. HMM based models presented in [2, 3, 4] have been used for event modelling. However, in this approach representation of an event is not transparent and it is difficult to generalise for new events. The work presented in [5, 6, 7, 8] mainly confronted the problem of indexing video data for efficient data extraction through user queries. In [5], the authors provide a hierarchical structure of video contents by dividing them into video objects, activity and events. The work presented in [8] takes temporal aspects of multimedia content into account in database management of video contents. Although these approaches provide conceptual understanding of how to model multimedia data for event modelling, they do not focus on modelling events in advance for event detection in a real time environment. Moreover, they do not explore the features of video event modelling which can facilitate event detection and discovery of unknown interesting events.

In [9, 10, 11] an important aspect of multimedia data management (the uncertain nature of data and its queries) was addressed. In [12, 13] mapping functions were used to define the relationships among semantic objects and explain how scene layer, object layer and concept layer can be connected by utilising the temporal aspects of multimedia data. In [14, 15] trajectory-projection information along with spatio-temporal aspects were utilised to confront the complex video data retrieval requirements. Object oriented approaches discussed above provide the basis for a video event modelling framework for unstructured multimedia contents (surveillance videos). However, these approaches mainly fall in the video indexing and retrieval category. The work presented in this paper will use an object oriented approach to propose an event modelling framework for modelling complex events and will also facilitate the event detection and event mining process on unstructured videos.

The approaches presented in [16, 17] provide an excellent hierarchical structure for unstructured video content; however, they mainly focus on video content searching and retrieval. While the approach presented in [18], identifies key semantic entities like objects and events and allows users to specify properties and relations between them, it does not make specific commitments regarding the structure of events and also does not provide mechanisms to reason with the annotations. In [19], a CASE based representation of events was extended to strengthen CASE based event representation; however, the proposed model does not focus on spatial aspects of video contents. The main drawback of the approach presented in [20], is that constructing a grammar for a relatively large domain is not feasible, especially taking into consideration the fact that for a human to have complete understanding of a specific

 ISBN: 978-1-61208-320-9

domain and anticipate all possible events is not realistic. The different approaches discussed in [21, 22, 23, 24, 25] provide a flexible hierarchical event modelling structure; however they do not fully explore the aspects of event modelling frameworks which can contain useful information to be used for optimisation of event detection and event mining processes. Moreover, they mainly deal with video contents of multimedia data. Whereas, in our research work we explore the utilisation of other multimedia data streams to model interesting events and explore their importance in event detection and the post event detection mining process.

The remainder of the paper is structured as follows: in Section 2 we will discuss the proposed multimedia event definition framework. In Section 3 we will discuss the limitation of the EDF. In Section 4 we presents Efficient Event definition framework (EDF^E) in detail. In Section 5 we will describe the data structure of EDF^E. Section 6 presents three event modelling examples using EDF^E. Lastly in Section 7 we conclude the work and discuss further research work.

II. PROPOSED MULTIMEDIA EVENT DEFINITION FRAMEWORK

The event modelling framework presented in this paper builds on the Event Description Framework (EDF) proposed in [21]. The advantage of EDF is that it provides a single template predicate for representing all events instead of defining a predicate for each event type. It also provides a set of predicates for describing spatio-temporal relationships between events and entities. Following are some of the reasons why we believe that EDF is the right candidate to provide a framework which can be extended into a promising multimedia event modelling framework for advanced surveillance systems.

- EDF is based upon the object oriented approach where a hierarchy of objects can be defined and their features can be inherited, which means the framework allows a particular event type to be defined as a subclass of another type.
- EDF provides a single template predicate for representing all events instead of defining a predicate for each event type.
- EDF allows a hierarchical decomposition of complex events into simpler events, which can be quite similar to the human approach to describing complex events in the real world.
- Another important element of the EDF framework is that it provides a set of predicates for describing spatio-temporal relationships between events and entities.

Having listed the advantages of EDF, it is also observed that there are some areas of event modelling which are not fully explored in EDF. In our proposed event modelling framework we concentrate on these areas to extend the capacity of EDF, so it can not only represent complex events in surveillance systems but can also provide valuable information for event detection and mining processes. Following are some of the major limitations of EDF which we have addressed in our proposed Efficient Event Description Framework (EDF^E).

III. LIMITATIONS TO OVERCOME

- EDF mainly focuses on the video content of a multimedia stream. However, in certain environments supporting the whole content of a multimedia stream can be very important to model interesting events. For example, in a retail store environment, utilization of ePOS data (text string) in multimedia streams can be used to deter/detect till scanning related frauds. Another example comes from Electronic Toll Collection (ETC) systems where a text stream generated by a scanner-based automatic vehicle classification system (AVC) can be used to initialize/validate the vision-based AVC.
- Although EDF provides a set of predicates for temporal relationships, it does not provide any mechanism to define different granularities of time intervals for those temporal relationships.
- In visual surveillance, there is a need to define virtual scene entities. These entities are not real observable objects but provide contextual scene information as a backdrop against which the events will take place. Two such examples are regions of interest (ROI) and tripwires. EDF does not specifically address the importance of such virtual scene entities.
- EDF provides a set of template predicates for representing video events; however it does not address the utilization of predicates to enhance the functionality of the event detection process.
- EDF does not provide a mechanism to update/store a set of entities in the system. In a surveillance system the number of entities can be large and it is generally not possible to manually store such entities. We address this problem by proposing an event mining framework which explores the relationship between entity feature-sets and associated text strings to generate appearance models of all the entities automatically (see our previous work [26]).

IV. EFFICIENT EVENT DEFINITION FRAMEWORK

In this section, a set of classes for semantic annotation of multimedia data are described along with their properties and relationships. We then present a set of predicates for describing various relationships between events and entities. While explaining each of these concepts we will also discuss how we have extended EDF to confront the limitations described above.

A. Entities

Entities are basically objects such as *car*, *person*, and *chair* observable in a particular domain. For example, while

 ISBN: 978-1-61208-320-9

describing the events occurring at a checkout area of a retail store environment, the various entities could be *till operator*, *barcode scanner*, different shopping items (*milk pack*, *sugar pack*, *butter*, *customer*, etc.) see Fig. 1. Thus for each specific domain we will have a hierarchy of entity classes, where different entity classes can be subclasses of the parent class' entity. For example '*Coke Bottle*' is sub-entity class of parent entity '*Bottle*'.

Let $\mathbf{P_E}$ be a set of parent entities $P_E = \{P_{E1}, P_{E2,}\ldots, P_{En}\}$, where each P_{Ei} can have one or more sub entities class S_E.

$$P_E=\{P_{E1}, P_{E2,}\ldots.., P_{En}\} \quad (1)$$
$$P_{En}=\{S_{E1}, S_{E2,}\ldots.., S_{En}\}$$

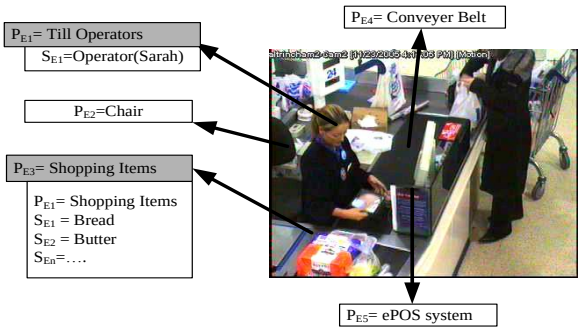


Figure 1. Parent and sub entities classes

These entity classes can have various properties or features associated with them and can also inherit the features of their parent class entity. The set of features for a specific entity class is the union of its own specific features and the features of its super class; for example if 'Coke Bottle' is a subclass of the top level entity class 'Bottle', and if shape and size are features of the super class then the sub class 'Coke Bottle' can have these features plus its own features, as defined below.

$$P_{En}=\{F_1=\text{value}, F_2=\text{value}\ldots\ldots F_n=\text{value}\} \quad (2)$$
$S_{En}=\{F_1=\text{value}, F_2=\text{value}\ldots\ldots F_n=\text{value}\}$ where F_n is feature of an entity.

1) Virtual Entities: In addition to the entities reflecting real life objects, we introduce the user defined virtual entities. These can be manually annotated using a graphical interface. These entities are not real observable objects but provide contextual scene information as a backdrop in which the events will take place. Two such examples are regions of interest (ROI) and tripwires, with their own set of features, for example the features of a tripwire are the spatial locations of start and end points.

Let $\mathbf{UD_{VE}}$ be a set of user defined virtual entity $UD_{VE}=\{UD_{VE1}, UD_{VE2}, UD_{VE3}\ldots.. UD_{VEn}\}$, each UD_{VEi} can have one or more feature values $F_{n,}$ as follows:

$$UD_{VE}=\{ UD_{VE1}, UD_{VE2}, UD_{VE3}\ldots.. UD_{VEn}\} \quad (3)$$
$$UD_{VEn} =\{F_1=\text{value}, F_2=\text{value}\ldots\ldots Fn=\text{value}\}$$

II) Text Entities: The combined knowledge derived from the combination of video and text data is more descriptive than each knowledge source considered in isolation. Based on this fact, it is our conjecture that multi-relational associations should capture more information from the combined metadata (see Fig. 2). In our event modelling framework (EDF^E) we introduce a text entity class to represent the supporting Text multimedia streams. These text entities are used not only to model interesting events but can also provide valuable information to the event mining process, where the association of text and visual entities can be explored to yield valuable information (we have discussed this in [26, 27].

Let $\mathbf{T_E}$ be a set of text entity UD_{VE} $T_E=\{T_{E1}, T_{E2}, T_{E3}\ldots..$ $T_{En}\}$, each T_{Ei} can have one or more labels (L_n) with associated values, as follows.

$$T_E=\{ T_{E1}, T_{E2}, , T_{E3}\ldots.. T_{En}\} \quad (4)$$
$$T_{En}=\{L_1=\text{val}, L_2=\text{value}, L_3=\text{value}\ldots..L_n=\text{value}\}$$
{Id=00384, item=grapes, price=1.68}

Figure 2. Text class entity

B. *Action:*

An 'Action' class refers to actions such as enter, leave, run, walk, that occur in a specific domain. Like entity classes, action classes can be organised in a hierarchy and also have features associated with them (such as speed, angle, etc). Further, they have a patient specification which describes the entity towards which the action is directed, e.g., a 'Coke Bottle' passes over the $Tripwire_1$ where $Tripwire_1$ is the patient. It is important to see that each action will have at least one object entity associated with it. For example, the 'enter' action needs two entities (object which is going to enter and the place, such as $Tripwire_1$ which that specific object is going to enter, as shown in see Fig. 3).

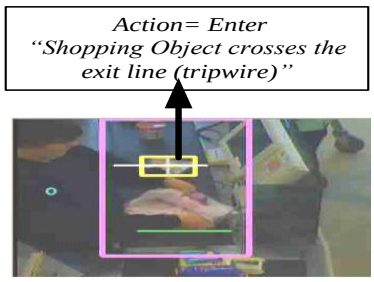


Figure 3. Action class

 ISBN: 978-1-61208-320-9

C. Events

Events are divided into two main categories, that of simple events and composite events.

I) Simple Events: These are basically (Actor, Action) tuples, where actor is a set of entities that initiate the event and action is a set of actions performed during the course of the event. For example, in the event of 'Man enters the room', man and room are entities and 'enter' denotes the action (see Fig. 4).

IV)

$$\text{Actor}=\{E_1, E_2 \ldots E_n\} \quad (5)$$
$$\text{Action}=\{A_1, A_2 \ldots A_n\}$$
$$S_{EVENT}=(\{\text{Actor}_1,\text{Actor}_2 \ldots \text{Actor}_n\}, \{\text{Action}_1, \text{Action}_2 \ldots \text{Action}_n\})$$

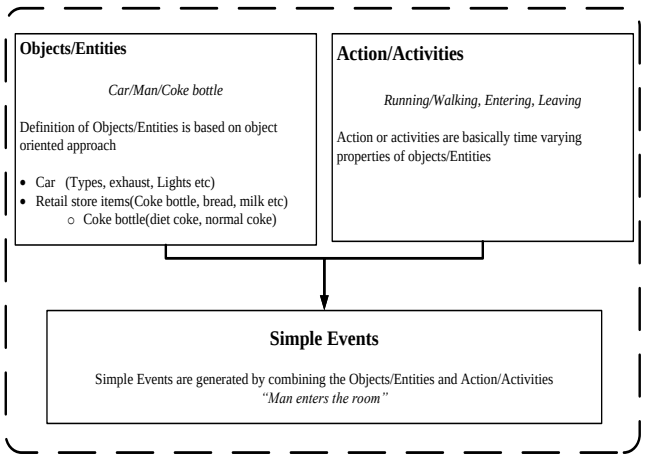


Figure 4. Simple event example

II) Composite Events: Composite events combine two or more simple events and is specified using the predicate PROCESS whose first argument is the event being defined and whose second argument is the composition of other events (see Fig. 5). The compositions of complex events are defined by using the following predicates:

$$\text{Predicate} = \{\text{SEQUENCE, AND, OR}\} \quad (6)$$

- **SEQUENCE** – represents a set of events which happen one after another in a temporal sequence.
- **AND** – represents a set of events with no particular temporal relationship between them.
- **OR** – represents a set of alternate events of which at least one should occur.

III) Additional Predicates: An extensive review of video data from a multitude of sources suggested that these three predicates cannot describe all the events that can occur in surveillance videos. Therefore, we introduce two new predicates (NOT IN, TERMINATE). These predicates not only help in modelling complex events, but they also provide valuable information to the event detection process for effective utilisation of hardware resources.

- **NOT IN:** represents a set of event/events which should not appear within a sequence of other simple events.
- **TERMINATE**: represents when the event detection should be terminated.

$$\text{Predicate} = \{\text{SEQUENCE, AND, OR, NOT IN, TERMINATE}\} \quad (7)$$

IV) Internal Entity Class: In complex events, there can be multiple entities of the same type at different times and locations. Therefore, while modelling complex events the proposed framework should be able to reference an entity which has been already defined in that event. We call such entity an internal entity (IE_{ID}) and refer to it with its ID such as IE_1, IE_4, IE_n.

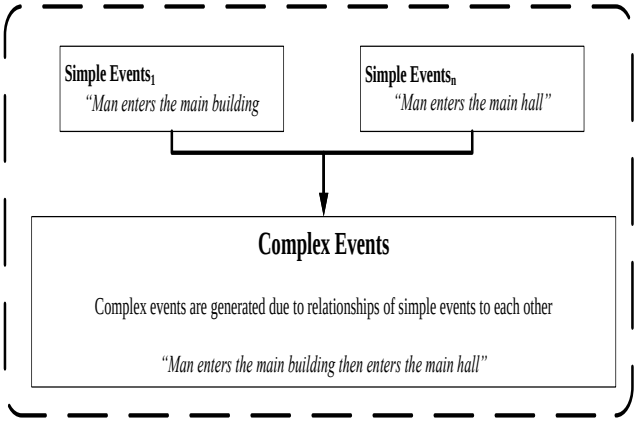


Figure 5. Composite event example

V) *Event Evidence:* Due to the volume of data in surveillance videos, it is important that only relevant information about the detected event should be stored with the constraint that full evidence of detected events is provided as backup. As there is no prior information available as to when a specific event is going to be triggered, the challenge is how to optimise the event storing process. Moreover, due to the nature of each event the requirement of evidence duration to be stored can differ as well. To overcome this challenge we introduce two new features for events: start evidence length (S_{EL}) and end evidence length (E_{EL}). These features' values determine the length of the evidence that needs to be stored for a specific event. For example if the specific event is detected with length of 1 minute and we have S_{EL}=200% and E_{EL}= 100%, then this means that 2 minutes of surveillance video will be stored prior to the detected event and 1 minute of video will be stored after the event being triggered. The main concept is to buffer a specific number of frames and only store them permanently if a

 ISBN: 978-1-61208-320-9

specific event is triggered; this provides vital evidence just prior to and after the event being triggered (see Figure 6).

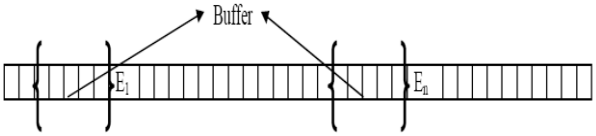


Figure 6. Event evidence buffering

D. Temporal Association Framework

Different predicates were proposed in [21] for defining temporal relationships between two events (e.g., met-by, meets, finishes, finished-by, started-by, starts, during, after, before, overlaps, overlapped-by, contains, simultaneous). Each of these predicates has two arguments and they can be either time intervals or events. These predicates are based on Interval Algebra presented by Allen in [28]; this interval temporal logic is shown in Fig. 7.

AFTER : $T_2^{start} > T_1^{end}$
MEETS : $T_1^{end} = T_2^{start}$
DURING : $(T_1^{start} < T_2^{start}) \wedge (T_1^{end} > T_2^{end})$
FINISHES : $(T_1^{end} = T_2^{end}) \wedge (T_1^{start} < T_2^{start})$
OVERLAPS : $(T_1^{start} < T_2^{start}) \wedge (T_1^{end} > T_2^{start}) \wedge (T_1^{end} < T_2^{end})$
EQUAL : $(T_1^{start} = T_2^{start}) \wedge (T_1^{end} = T_2^{end})$
STARTS : $(T_1^{start} = T_2^{start}) \wedge (T_1^{end} \neq T_2^{end})$

Figure 7. Allen's interval algebra describing temporal logic between $Time_1$ and $Time_2$ [19]

1) Temporal Predicate Granularity: Although these temporal predicates presented in [21] cover most of the temporal relationships to be found in multimedia events, they do not provide a mechanism to define different granularities of temporal associations. We therefore provide granularity (G_n) as a feature of these predicates. Moreover, we introduce two new temporal predicates that can be extensively used in events. These are: the minimum gap (MIN_GAP) and maximum gap (MAX_GAP) with given granularity and its value. These can be use to define the temporal threshold between two events before it can be considered an interesting event.

$$Temporal_{Predicates} = (Predicat_1, Predicate_2, Predicate_3 \ldots\ldots Predicate_n) \quad (8)$$
$$Predicate_n (\{S_{EVENT}/C_{EVENT}\}, \{G_n = value\})$$

E. Spatial Association Framework

We use the spatial association framework defined in EDF [21] to classify spatial relationships into three types: topological, directional and metric.

- Topological relationships are invariant under translation, rotation and scaling and basically describe the properties that characterise the relative position of objects against each other. Studies have shown that all topological relationships can be described using six basic relationships (touch, in, cover, equal, overlap and disjoint) and three operators: "b", which, when applied to an area returns the boundary, and "f" and "t" which return the end points of a line [29].
- Directional relationships are used to describe the relationship involving the relative direction between two objects, such as below, above, right, left, etc. In order to define a set of predicates for describing directional relationships, prepositions in English are used in EDF. The notion behind this is that all directional relationships can be expressed in English using prepositions. The following predicates are defined in EDF – over, upon, opposite, behind, in-front-of, left-of, right-of, above, below.
- Metric relationship involves relationship such as distance, e.g., Distance > 100 that is, applying constraints on spatial metrics. Three predicates are introduced in EDF to represent metric relationships. These are "near", "far" and "at" to specify the location of an entity or event.

V. DATA STRUCTURE

In order to manage the above mentioned ontology of our proposed event modelling framework (EDF^E), we utilise and extend the data schema presented in EDF [21]. The intention is to use relational tables for organising the elements (entities, actions, simple events, predicates and complex events etc) of EDF^E. We describe the different relational tables of the proposed data structure below:

- **viewTbl:** The purpose of this table is to store the information about different camera views available for event modelling /detection. It consists of two fields, *viewID* which is the primary key of the table (primary key is used to store the unique-ID of specific record in table) and *viewDetail* field which stores the descriptive information about the view, such as 'Entrance view', 'Checkout area view', 'Camera view covering clothing section of the shopping mall'.
- **entityTBL** This table stores the information about entities to be used in modelling different events. It consists of *entityID, entityName and parentID,* fields. The *entityID* is a primary key of the table; *entityName* stores the name of the entity such as person, car, object, coke bottle etc. The *parentID* field is used to manage the hierarchy of entities classes. Here *entityID* becomes a foreign key to the *parentID* (the example given in Table I. explains the concept of managing entity hierarchy through *entityID* and *parentID* fields).

TABLE I. EXAMPLE FOR ENTITY HIERARCHY

ID	entityName	parentID	Comments
1	Soft Drink	1	Since the *entityID* and the *parentID* are the same that means it is a top level entity in the hierarchy.
2	*Coke* Bottle	1	The *parentID* points to *entityID* of Soft Drink, that means Coke bottle is sub class of entity Soft Drink

 ISBN: 978-1-61208-320-9

- **entityProperties:** The purpose of this table is to store the different properties of entities defined in entityTbl. The table consists of seven fields: *propertyID, entityID, propertyName, propertyValue, startTime, endTime and viewID*. The *propertyID* is a primary key to the table; *entityID* refers to entityTBL and indicates the entity to which this property belongs. The *propertyName* stores the name of a specific property (for example, size, colour, shape etc). The *propertyValue* is foreign key which refers to the valueTbl table (valueTbl stores the information about the value of the property). The *startTime* and *endTime* fields contain time interval information during which this property is true, the *viewID* field refers to viewTbl and indicates for which specific camera view this property is true.
- **valueTbl:** Since properties of entities can have different types of value, for example the size property of an object can be a number of pixels, whereas the colour property can be a colour histogram stored in a file. Hence, valueTbl is used to manage different types of property values; it consists of *valueID, valueType* and *valueInfo* fields. The *valueID* is the primary key to the table and a foreign key to the entityProperties table; valueType stores the type of value (integer, string, histogram file, text file, etc) and finally *valueInfo* stores the actual value itself, it can be an integer number or file name etc.
- **actionTbl:** The action table stores information about different actions which can take place in simple and complex events. It consists of *actionID, actionName, actionDes* and *patientID* fields. The *actionID* is the primary key to the table and *actionName* specifies the action, such as enter, run, exit, move, etc. The *actionDes* stores the description of action and finally *patientID* field stores a list of *entityID's* (from entityTBL) which are patient to that specific action.
- **simpleEvent_Tbl:** This table stores the information about simple events; it consists of *simpleEventID* (primary key), *simpleEventName* which stores the name of event such as 'Man enters the room', 'object leaving the area'. The *simpleEventString* field contains the event string generated by the EDF^{E}. The *eventActor* field stores a list of *entityIDs* that take part in the event; The *eventActions* field contains a list of *actionIDs* performed during the event; whereas S_{EL} and E_{EL} fields contain the start and end length of evidence to be store for each detected instance of the event.
- **predicateTbl:** The predicateTbl stores the information about different predicates that can be used in complex events. The table predicateTbl consists of *predicateID*, the primary key of the table and used as a foreign key in complexEventTbl. The *predicateType* stores information about the type of the predicate (composite, temporal, spatial). The *predicateName* field stores the name of the predicate, whereas *predicateDes* field contains descriptive information explaining the specific predicate.
- **complexEvent_Tbl:** The complexEvent_tbl table is used to store details of complex events. It consists of *complexEventID* field which is a primary key to the table and the *complexEventName* field which stores the name of the complex event (such as 'Item scanning', 'Tail gating', etc). The *complexEventString* field holds the complex event string generated using the EDF^{E}. The *predicateID* field contains the list of predicates used in the specific complex event (this refers to predicateTbl), whereas *simpleEventsOnly* is a boolean field indicating that this complex event consists of only simple events or a combination of both simple and complex events. The *memberEventIDs* field contains the list of *simpleEventID* and/or *complexEventIDs* used in the specific complex event. Finally S_{EL} and E_{EL} fields contain the start and end length of evidence to be stored for each detected instance of the event (this overrides the S_{EL} and E_{EL} values of included simple events).
- **detectedEventsTbl:** This table stores information about each detected instance of modelled simple or complex events. The table consists of *ID* (primary key), the *eventID* refers to simple or complex event tables against which this instance is detected. The *eventClip* field stores the file name containing video evidence of the detected event, and the *startTime* and *endTime* fields contain the start and end time of each detected instance of the event.

VI. EVENT MODELLING: EXAMPLES

We now provide three examples to explain how simple and composite events can be modelled through the proposed EDF^{E}. The first two examples are of typical surveillance events in a retail store, see Fig. 8 & 9; the third example is based on a view of a secure area (see Fig. 10). In the first example (see Fig. 8), the composite event is defined using virtual entities, text entities, predicates, temporal predicate actor and action elements of EDF^{E}. After defining the main entities, the sequence of two events is defined by using the SEQUENCE predicate along with temporal predicate of MIN_GAP (to represent the minimum gap between the two events), followed by representing the Actor and Action of each events along with their feature sets (such as colour='white', size>=200).

VII. CONCLUSION AND FUTURE WORK

In this paper we presented the efficient event description framework (EDF^{E}), which builds on the event description framework (EDF) presented in [21]. After presenting a general introduction to the event modelling concept, we then discussed the reasons which make EDF the right candidate to be extended for a promising multimedia event modelling framework; this is followed by a discussion of the limitations of EDF which need to be addressed. We then

 ISBN: 978-1-61208-320-9

presented the EDF^E by describing a set of classes for semantic annotation of multimedia data along with their properties and relationships. Next, we presented a set of predicates for describing various relationships between events and entities. While explaining each of these concepts we also discussed how we have extended the framework to confront the current limitations in EDF and the different features of EDF^E which can facilitate event detection and mining aspects of surveillance systems. A modified and extended data structure was also presented to store the ontology of different events. Finally we presented examples to explain how simple and complex events can be modelled through the proposed EDF^E. In future, we will be concentrating our attention on including the event categorisation in event definition framework. That is because due to the nature of certain events in surveillance videos, the number of detected events can be very large. For such events it can be very useful to categorise them into different levels, e.g., using a traffic light system: while detecting an event of miss-scanned items on checkout area; a detected event in which the object's size is relatively large can be categorized as a "Red Event", as the miss-scanned item can be of relatively higher value. Whereas, a detected event in which the object colour is close to the skin colour can be categorized as a "Yellow Event", as this can possibly be a false detected event where a portion of the operator's hand is misclassified as object/item. The categorisation of the detected event can be based on exactness of an event matched to the modelled event as well. For example, if the event is matched with complete certainty then it is categorized as a "Red Event" and if the event is matched with partial certainty to the modelled event then it is categorized as a "Yellow Event" etc. The important question to answer here is: how to measure the exactness of an event matched that is how to know that event is matched 100%, 80% or 60%.

REFERENCES

[1] Gupta, A., T. E. Weymouth, and R. Jain. "Semantic Queries with Pictures: The VIMSYS Model. in Proceedings of the 17th International Conference on Very Large Data Bases (VLDB '91)". 1991. Morgan Kaufmann Publishers Inc.

[2] Andrew, D. W. and F. B. Aaron, "Parametric Hidden Markov Models for Gesture Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence", 1999. **21**(9): pp. 884-900.

[3] Nuria, O., R. Barbara, and P. Alex. "A Bayesian Computer Vision System for Modeling Human Interaction. in Proceedings of the First International Conference on Computer Vision Systems (ICVS '99)". 1999. London, UK: Springer Berlin / Heidelberg.

[4] Matthew, B. and K. Vera, "Discovery and Segmentation of Activities in Video. IEEE Trans. Pattern Anal. Mach. Intell.", 2000. **22**(8): pp. 844-851.

[5] Sibel, A., K. Sel, C. K. Seluck, E. Kutluhan, and V. S. Subrahmanian, "AVIS: The Advanced Video Information System, 1997, Institute for Systems Research Technical Reports.

[6] Sibel, A., K. Sel, C. Su-Shing, E. Kutluhan, and V. S. Subrahmanian, "The advanced video information system: data structures and query processing. Multimedia Syst.", 1996. **4**(4): pp. 172-186.

[7] Hacid, M. S., C. Decleir, and J. Kouloumdjian, "A Database Approach for Modeling and Querying Video Data. IEEE Trans. on Knowl. and Data Eng.", 2000. **12**(5): pp. 729-750.

[8] Carlo, C. "Modeling Temporal Aspects of Visual and Textual Objects in Multimedia Databases. in Proceedings of the Seventh International Workshop on Temporal Representation and Reasoning (TIME'00)". 2000. IEEE Computer Society.

[9] Ramazan, S., Ayg, and Y. Adnan, "Modeling and Management of Fuzzy Information in Multimedia Database Applications. Multimedia Tools Appl.", 2004. **24**(1): pp. 29-56.

[10] Duc, A. T., A. H. Kien, and V. Khanh. "VideoGraph: A Graphical Object-based Model for Representing and Querying Video Data. in ACM International Conference on Conceptual Modeling / the Entity Relationship Approach (ER2000)". 2000. Springer Berlin / Heidelberg.

[11] Duc, A. T., A. H. Kien, and V. Khanh. "Semantic reasoning based video database systems. in 11th International Conference on Databases and Expert Systems Applications". 2000. London, UK.

[12] Yong, C. and X. De. "Hierarchical semantic associative video model. in IEEE International Conference on Neural Networks and Signal Processing". 2003.

[13] Cheng, Y. and D. Xu. "Content-based semantic associative video model. in 6th International Conference on Signal Processing". 2002.

[14] Dönderler, M. E., E. Şaykol, U. Arslan, Ö. Ulusoy, and U. Güdükbay, "BilVideo: Design and Implementation of a Video Database Management System. Multimedia Tools Appl.", 2005. **27**(1): pp. 79-104.

[15] Dönderler, M. E., "Data Modeling and Querying for Video Databases (PhD Thesis), in Computer Engineering2002, Bilkent Unversity: Turkey. pp. 129.

[16] Ekin, A., "Sports video processing for description, summarization and search (PhD Thesis), in Electrical and Computer Engineering2004, The University of Rochester. pp. 166.

[17] Guler, S. and I. Pushee. "Videoviews: A Content Based Video Description Scheme and Video Database Navigator Tool. in Multimedia data mining workshop (MDM/KDD 2002)". 2002. Edmonton , Canada.

[18] Benitez, A. B., et al. "Semantics of multimedia in MPEG-7. in IEEE International Confonference on Image Processing". 2002.

[19] Hakeem, A., Y. Sheikh, and M. Shah, "CASE^E: A Hierarchical Event Representation for the Analysis of Videos, in The Nineteenth National Conference on Artificial Intelligence (AAAI)2004: San Jose, USA. pp. 263-268.

[20] Ivanov, Y. A. and A. F. Bobick, "Recognition of Visual Activities and Interactions by Stochastic Parsing. IEEE Trans. Pattern Anal. Mach. Intell.", 2000. **22**(8): pp. 852-872.

[21] Natarajan, P. and R. Nevatia, "EDF: A framework for Semantic Annotation of Video, in Proceedings of the Tenth IEEE International Conference on Computer Vision Workshops2005, IEEE Computer Society.

[22] Nevatia, R., T. Zhao, and S. Hongeng. "Hierarchical

 ISBN: 978-1-61208-320-9

Language-based Representation of Events in Video Streams. in Computer Vision and Pattern Recognition Workshop, 2003. CVPRW '03". 2003. Madison, WI.

[23] Hongeng, S. and R. Nevatia, "Multi-agent event recognition, in Proceedings Eighth IEEE International Conference on Computer Vision (ICCV 2001)2001: Vancouver, Canada. pp. 84-91.

[24] Vu, T., F. Bremond, and M. Thonnat, "Automatic Video Interpretation: A Novel Algorithm for Temporal Scenario Recognition, in The Eighteenth International Joint Conference on Artificial Intelligence2003: Acapulco, Mexico. pp. 9-15.

[25] Nevatia, R., J. Hobbs, and B. Bolles. "An Ontology for Video Event Representation. in Proceedings of the 2004 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'04)". 2004. Washington D.C, USA: IEEE Computer Society.

[26] Anwar, F., I. Petrounias, T. Morris, and V. Kodogiannis. "Entity appearance model generation for multimedia events in surveillance videos. in Intelligent Systems (IS), 2010 5th IEEE International Conference".

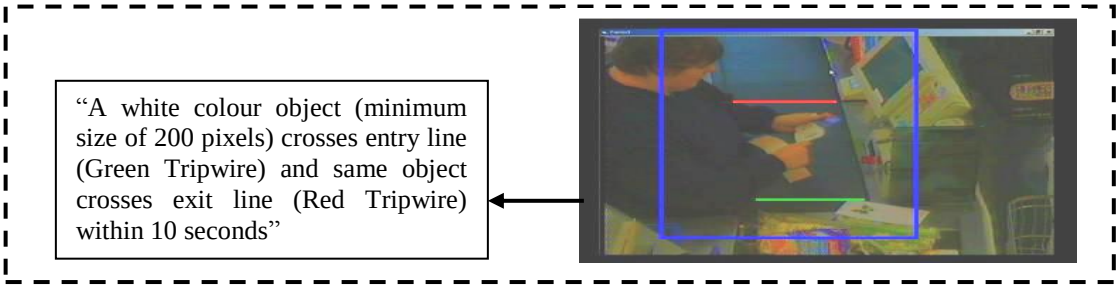
[27] Anwar, F., I. Petrounias, T. Morris, and V. Kodogiannis, "Mining anomalous events against frequent sequences in surveillance videos from commercial environments. Expert Systems with Applications". **39**(4): pp. 4511-4531.

[28] Allen, F. J. and F. George, Actions and Events in Interval Temporal Logic, in Spatial and Temporal Reasoning, O. Stock, Editor. 1997, Kluwer Acadamic Publishers: Dordrecht, Netherlands. p. 205-245.

[29] R., H. G., "An Introduction to Spatial Database Systems. The VLDB Journal", 1994. **3**(4): pp. 357-399.

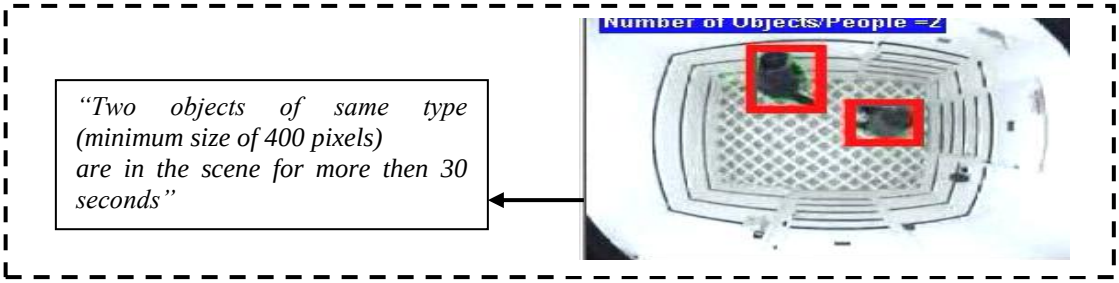
PROCESS(ObjectScanning (Item object, UDVE GreenTripwire, UD_{VE} WhiteTripwire,) SEQUENCE ($Event_1$, $Event_2$, MIN_GAP(G=Second, value<=5)), Actor($Event_1$, class = object, ID=E_1), Action($Event_1$, "Enter", class=GreenTripwire, ID=E_2), Actor($Event_2$, class=E_1, ID=E_3), Action($Event_2$, "Enter", class=WhiteTripwire, ID=E_4)

Figure. 8: Event modelling example (retail store checkout)



PROCESS($Event_{ID}$=3(Item object, UD_{VE}(GreenTripwire,RedTripwire, T_{EN} Text) SEQUENCE ((Event1, Event2), MIN_GAP(G=Second, value<=10)), Actor($Event_1$, class = object, size>=200, Colour='white' ID=E_1), Action($Event_1$, "Enter", class=GreenTripwire, ID=E_2), Actor($Event_2$, class=E_1, ID=E_3), Action($Event_2$, "Enter", class=RedTripwire, ID=E_4)

Figure. 9: Event modelling example (Secure area surveillance)



PROCESS(ObjectCounting(Item object) Event ($Event_1$, G=Second, value=>30) Actor($Event_1$,class=object, size>=400,ID=E1), Actor($Event_1$,class=E_1, ID=E_2), Action($Event_1$,"Remain", class=E_1, class=E_2, ID=E_3)

Figure 10: Event modelling example (Secure area surveillance)

A Model for Facial Activity Recognition using Metarepresentation: a Concept

Boris Knyazev and Yuri Gapanyuk
Faculty of Informatics and Control Systems
Bauman Moscow State Technical University
Moscow, Russia
emails: {bknyazev@bmstu.ru, gapyu@bmstu.ru}

***Abstract*—Recognition of the facial visual properties (physiognomy) and its static and dynamic behavioral patterns (action units) has proved to be an important part in many multimedia retrieval and analysis applications. Apart from the previous studies, where methods to extract part of the action units from an image or video have been developed, in this ongoing research project we work on a model for more accurate and detailed facial activity semantic description adaptable to new behavioral patterns and real conditions. In this paper, we address challenges of building this model and suggest its basic multilevel concept. On the low level, we propose using wavelet-based multiresolution representation of video data. On the middle level, several multiclass classifiers are being examined for the purpose of attribute learning, and a custom multiple metric is provided. On the high level, facial elements, behavioral patterns and their attributes can be connected and further extended using the ontologically-compliant architecture of this model. On the abstraction layer, all three levels of this model are seamlessly integrated via graph-based hierarchies of metavertices, metaedges and their mappings. Having this structure, the proposed model can be trained and employed to solve the problems of human behavior retrieval and human-computer multimodal interaction more efficiently. Current results, however, reveal that to be reliable, this model requires further research studies and their comprehensive experimental evaluation.**

Keywords-facial behavrior recognition; semantic annotation; video multiresolution representation; metagraph modeling

I. INTRODUCTION

Human appearance and nonverbal behavior, particularly facial visual properties (physiognomy) and its static and dynamic behavioral patterns (action units), convey a lot of overt and covert data [1, 2]. Mining of these data is inherent to tackle the problems of human intelligent monitoring and facial expression analysis more efficiently. It has also proved to be useful in psychophysiological and neurological diagnosing, e.g., autism, schizophrenia and other disorders [13], in synthesis of virtual agents [28], examining correlation between face asymmetry and brain disharmony [14] and enhancing human-computer multimodal interaction as a whole. If this mining is automatic, accurate and detailed, then its results – objective data or ground-truth – could be supplied to an expert, clinician or some logical rule-based model for the purpose of making important real-life decisions, e.g., preventing car crashes caused by drowsiness, as well as for entertainment.

The automatic behavior recognition engine could also be a core component of either general-purpose [15] or more specific [16, 17] multimedia annotators. The primary intent of this type of software is to provide means, usually via graphical user interface, to spatiotemporally bind annotations with other modalities, like audio, and with context, which may have a huge impact on interpreting behavior and making a more reliable decision [2].

This research is aimed to further develop previous studies on the automation of facial behavioral patterns recognition (e.g., [3-5]) and is inspired by the works on wavelet multiresolution decomposition [6], Gabor and Dual-Tree complex wavelets [7-9], on graph-based models [10, 11], attribute learning [12] and body segmentation [27].

The contribution of this paper is a model recognizing more action units (AUs) based on the Facial Action Coding System (FACS) [21, 23] and able to extract the new ones, including body AUs [24] and those defined by an expert. Additionally, this model is integral on the abstraction layer and expected to provide easier learning procedures, return semantic annotations improving expert-computer interaction and produce more precise results in more real conditions.

The goals of this paper are to address challenges of the development of this model, suggest its basic concept, give its basic experimental results and propose the directions of its further development.

In Section 2 of this paper, the four-level structure of the proposed model is introduced. In Section 3, we discuss the alphabet of human behavioral patterns and facial action units in particular, and suggest extracting only a specific set of attributes of these actions. Section 4 is devoted to the low level of our model, in which general properties of a metagraph, its vertices and edges are presented. In Section 5, the ways of training our model are shortly overviewed and basic evaluation results are given and discussed.

II. MODEL OVERVIEW

At this stage of research, we do not concern context-based video retrieval and real-time requirements, as well as cluttered environments and multiple persons, so that this model accepts videos (or image sequences) with only one human on a simple background as in the databases [18, 19, 20]. But, this model should be adaptable to real application

 ISBN: 978-1-61208-320-9

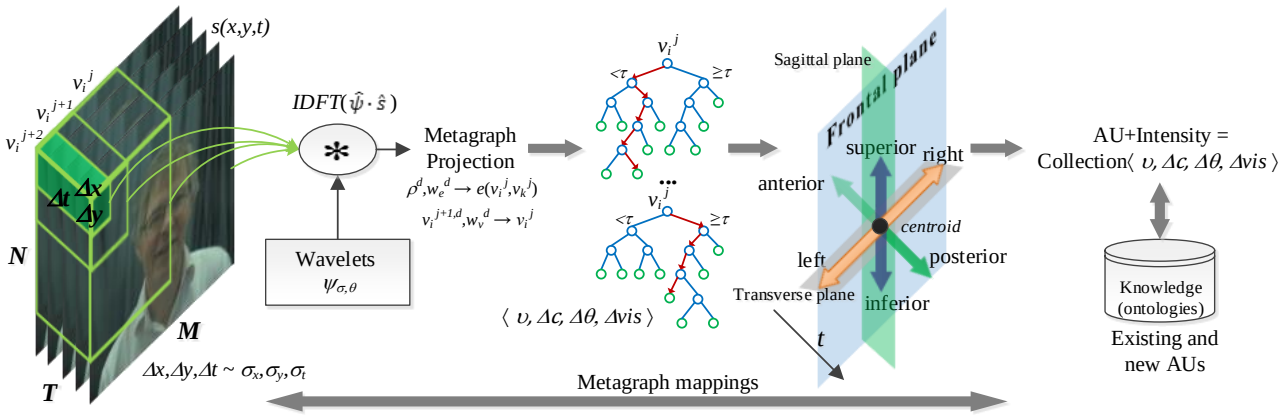


Figure 1. Model overview. Metagraph multilevel representation and relationship between human kinematics and video representation. The WT is faster calculated in the frequency domain applying the convolution theorem and the inverse Discrete Fourier Transform (IDFT); ˆ denotes the Fourier transform.

environments. Our model's structure is a three-level pipeline usually implemented in a visual recognition and understanding system plus a fourth, integration level (Fig. 1).

(1) On the low level, it maps an input video signal to a finite combination of 3D (or 2D) complex wavelets, a naturally multiresolution way to describe an N-dimensional non-stationary signal, such as a video. In addition to capturing changes in texture on different scales and orientations, a number of geometric and color properties can be computed.

(2) On the mid-level, we investigate several multiclass machine learning methods, including the unsupervised ones, such as k-means, and supervised, such as Random Forests (exemplified in Fig. 1), and suggest a multiple weighted metric to separate classes. The classes to be learned are facial primitives, attributes of the AUs as well as more abstract entities.

(3) On the high level, we propose describing facial primitives, behavioral patterns and their attributes merging existing human ontologies with our own being created on the basis of FACS, a broadly accepted facial coding scheme.

(4) To provide integrity on the abstraction layer, that is to encapsulate the three above levels making a model more flexible and scalable, metagraph-based representation of these levels is introduced.

Graphs and their extensions allow intuitive describing of hierarchical data and processes. Furthermore, graph-based models often give competitive results in vision applications, e.g., [11], [22]. One of their extensions is a metagraph, proposed in [10], which is a universal structure to describe properties (attributes), logical relations and complex mappings and, therefore, may be effectively employed to represent humans and their behaviors on the different levels.

One of the benefits of this approach is that since patterns recognition and description are two tightly related processes, detailed comprehensive description should boost the recognition of the patterns and vice versa. Even though the problem of signal reconstruction is not directly related to this study, it might be extremely important for further developments.

On the other hand, among the weaknesses of this model is its computational overload, and it is mostly the low level where optimization techniques should be applied. In spite of this cost, our experience has demonstrated that the benefits of accurate automatic description significantly overweight this negative side effect. In fact, unreliable recognition results mean double work for an expert-annotator: checking plus editing. In any case, the dependency of the accuracy of the results on the extent of detailing should be evaluated.

III. CLASSIFICATION OF BEHAVIORAL PATTERNS

Classification of human nonverbal patterns is challenging to be complete because there are many specialized versions, e.g., [32]. Nevertheless, there is a quite reliable coding system for a face, FACS, and a more recent development for a body, The Body Action and Posture Coding System (BAP) [24], which represent behavioral patterns as combinations of simple AUs and their intensities (A-E) in case of a face. These AUs can be further combined to extract more complex patterns, such as facial expressions in [5].

Training a classifier for each action unit separately is time consuming and is not straightforwardly adaptable to new action units. To remedy this, based on existing coding systems and human anatomy and kinematics, we can define all action codes using a more primitive set(s): $\langle \upsilon, \Delta c, \Delta\theta, \Delta vis \rangle$, where υ – a facial primitive, Δc – change of its centroid (translation), or other measure of spatial relocation, $\Delta\theta$ – change of its spatial orientation (rotation) or movement direction, Δvis – change of its visibility (as in AUs 43, 45, 46 and others). The latter argument must also capture appearance and disappearance of texture features like furrows, e.g., for AU 4 (Brow Lowerer) in the glabella area. To reduce biases in these changes, their values must be normalized and compared to a reference state. This state should include the current orientation and position of an upper level node (in our case, it is a face or a head), person's individual features and context.

Movement directions $\Delta\theta$ can be quantized in the human anatomic planes: left-right (intersection of the F and T planes), superior-inferior (intersection of the F and S planes)

 ISBN: 978-1-61208-320-9

and anterior-posterior (intersection of the *T* and *S* planes), where *F*, *T*, *S* – frontal (coronal), transverse (horizontal) and sagittal (medial) planes respectively (Fig. 1).

The intensity of an action unit is a measure of how distant is a certain facial primitive from a reference position, which may be weighted in Δc and how much texture is changed, which may be weighted in Δvis.

To quantize υ, facial primitives should be divided into smaller ones and include teeth, tongue and other elements involved in some facial activity [21]. To provide more flexibility, though, in addition to verbal description the facial primitives should be also defined as a collection of geometric and texture attributes.

Although, some complex action units, e.g., AU 9 – Nose Wrinkle, AU 23 – Lip Tightener, AU 28 – Lip Suck, AU 32 – Bite, AU 37 – Lip Wipe, etc., and their intensities (A-E) are laborious to be expressed in this way, this representation is more complete from a physical point of view and it is still possible, even though one of the hardest, yet tractable, challenges seems to be quantization (sampling) of the parameters υ, Δc, $\Delta\theta$ and Δvis.

IV. METAGRAPH-BASED MODEL

A. General Model

So far, metagraphs have no unified theory and in this study we adhere to the definition close to [10]:

$$MG = \langle V, E\rangle,\ v_i \in V,\ e_k \in E, \tag{1}$$

where *MG* – a metagraph, *V* – a set of vertices (metavertices), *E* – a set of edges (metaedges), v_i – a vertex of the metagraph and e_k – its edge. In contrast to simple graphs, the vertices are defined as $v_i = \langle\{v_m\},\{e_k\}\rangle$, $v_m \in V$, and can be in turn considered to be a metagraph, so these two terms are interchangeable. We should also distinguish metagraphs from other graph extensions. Compared to hypergraphs, for example, vertices of a metagraph can include both vertices and edges, forming a logical pyramid (a hierarchy). Next, metagraphs can have their own application specific properties. In our study, this pyramid is limited to a video pixel on the one side and by a spatiotemporal voxel $\langle M, N, T\rangle$ with facial activity on the other side. In other words, each metavertex v_i^j resembles an abstract basic type, instances of which include, but are not limited to, a single video pixel; group of interest pixels joined in time, space or other domains (superpixels); input video as a whole, where *j* – is a level of a metavertex v_i^j, and:

$$v_i^j = \bigcup_{m,k} v_m^{j-1}, e_k\ . \tag{2}$$

This abstraction is very convenient because one can work with metavertices in the same fashion as with base abstract types in programming frameworks, i.e., manage objects being unaware of their exact content and values of properties.

$$V_0 \underset{n\to-\infty}{\subset \ldots \subset} V_2{}^j \subset V_2{}^{j+1} = V_2{}^j \oplus O_2{}^j \underset{n\to+\infty}{\subset \ldots \subset} V_2{}^{j+n}$$

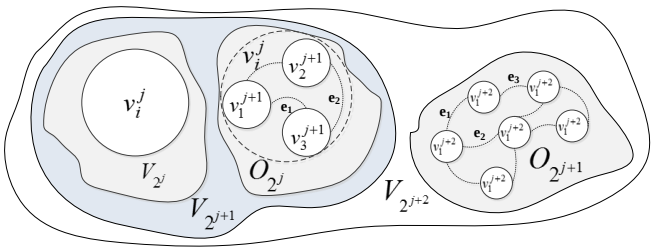


Figure 2. Connection between metagraph-based and wavelet multiresolution reprsentation. For simplicity parental relationships and all edges are not visualized.

In this research, first, a video signal *s(x,y,t)* defined on $\langle M, N, T\rangle$ is represented as a set of abstract metavertices, then mappings to several domains are iteratively constructed in a partially supervised way (see Section 5). The next subsections of this section contain details about the vertices and edges of the proposed metagraph and their properties.

B. Metavertex

Each metavertex has values in at most four domains: spatial, wavelet, color and semantic. These values are essential for further classification of υ, Δc, $\Delta\theta$ and Δvis.

1) Spatial domain (S)

From the kinematics point of view, a human, as well as a human face, can be approximated to a set of objects, the coordinates of their centers of mass and three angles in space. Instead of mass, we can only compute a geometric center (centroid). Important is that the centroid of a more complex object is a linear combination (an average) of its lower level elements.

2) Wavelet (W)

The value of a metavertex in the wavelet domain must reflect behavioral features in the spatial, temporal and frequency domains, where raw video data are difficult to be analyzed. An input signal is transformed to a higher dimensional space, in which video features are more discriminative.

There is no certain algorithm to choose a transform, also referred to as an image/video descriptor or visual words, since there are a couple tradeoffs to consider.

On the one side we want a smooth (continuous) and shift, rotation and scale invariant transform with optimal signal's time and frequency resolutions (limited due to the uncertainty principle) and perfectly reversible, on the other – we want a fast, non-redundant transform requiring less computational power. As was stated above, more accurate results are more valuable for our purpose, and, moreover, we provide optimization techniques. Thus, we suggest to be focused on the wavelet representation of a video signal for a couple of reasons. First, the wavelet transform (WT) is, generally, a type of a complete, i.e., reversible, transform [7] and does not lose information about facial behavior. Second, the wavelet theory provides methods to analyze a signal, such as a video or an image, at different scales, called

 ISBN: 978-1-61208-320-9

multiresolution analysis [6], which is a crucial ability for our metagraph-based model.

Numerous equations of 1D, 2D and 3D wavelet mother functions (wavelets), their Fourier transforms (FTs) and the wavelet criteria can be found in various forms in [6-9, 25], and are omitted here due to their lengthiness. A fast FT (FFT) and its inverse (IDFT) allow computing the WT faster applying the convolution theorem (Fig. 1). Gaussian-modulated complex exponentials, such as the Morlet and Gabor functions, are preferable to be employed as a series of wavelets $\psi_{\sigma,\theta}$, since they are continuous complex wavelets with optimal temporal and frequency resolutions [7], as well as some extensions of B-splines. We suggest binding each level j of the metagraph pyramid with a wavelet scaling (dilation) factor σ (in case of 3D it is σ_x, σ_y, σ_t) by definition included into $\psi_{\sigma,\theta}$. Together with (2) we obtain:

$$v_i^j, \psi_{\sigma,\theta} \in V_2^j, O_2^{j-1}, \quad (3)$$

where V_2^j – a space of approximate mappings, O_2^{j-1} – a complementary to V_2^{j-1} space of detailed mappings; a power of 2 means a dyadic WT [6]. Thus, the pyramidal structure of our metagraph is fixed, but the values of its vertices are assigned in respective spaces (Fig. 2).

In other words, a video volume $\langle M, N, T \rangle$ with facial activity is expressed as a sum of a "blurred" facial video plus detailed facial parts plus more detailed features of the facial parts and so forth up to single pixel values. More general approximate areas around the facial features (areas around forehead, nose, eyes, lips, etc.) and coarse movements (head shaking) can be detected and described on the high scales σ (low temporal and spatial frequencies and level j), while smaller elements (eyes, iris, mouth corners, etc.) and subtle movements (lip twitching, tics, eyes movements) can be detected and described more precisely on the lower scales σ (higher frequencies and level j). In practice, though, we do not need to build both V_2^j and O_2^j spaces on each level j, and as a result, can make scaling of the WT adaptive:

1. Apply the WT with high σ values to the whole video (or one frame) and scale down until a face low-frequency pattern is not found on the video.
2. Analogously apply the WT with lower σ values only to the facial video voxel (or image block) and scale σ down until distinct facial elements are not classified.
3. Recursively repeat step 2 with lower σ values only to specific video voxel and further scale down adaptively until all details are not extracted.

The exact σ values depend on the facial primitive and should be empirically estimated. For instance, they can be calculated on the basis of the entropy-based information gain, similar to [27]. Orientations θ of the wavelets might also vary in a similar sense.

To keep such strengths of the complex WT (CWT) as approximate shift invariance and directional selectivity, while acquiring the ability of perfect reconstruction of the real-valued discrete WT (DWT), the Dual Tree Complex Wavelet transform (DC CWT) could be employed at no extra computational cost compared to the CWT [8]. For both the CWT and DC CWT redundancy is 4:1 for 2D and 8:1 for 3D, whereas the DWT has no redundancy.

3) Color domain (V)

Data in the color domain are useful for facial segmentation. If color details are disregarded in wavelet coefficients, a separate color scheme must be kept, e.g., color histogram. Otherwise, it must be derived from the wavelet coefficients computed for each color channel independently.

4) Semantic domain (S)

The semantic structure of our model should mirror the metagraph pyramidal structure except the semantically meaningless, abstract metavertices, such as some facial regions. Semantic (verbal) terms are necessary for a more natural interaction between a clinician and lower level parts of the model. We suggest integrating existing ontologies, such as Virtual Human Ontology, Foundational Model of Anatomy Ontology and Mental Functioning Ontology to define facial AUs and more complex patterns based on a primitive set defined in Section 3.

C. Metaedge

A metaedge is an attributed multiple edge wrapping distances between two metavertices v_m^j and v_k^j of the same level j in respective domains:

$$e(v_i^j, v_\kappa^j) = \langle \rho^S, \rho^W, \rho^V, \rho^O \rangle, \quad (4)$$

where $e(v_i^j, v_k^j)$ must satisfy the three distance axioms, described in [26]. In addition, $e(v_i^j, v_k^p) = \varnothing$ for $j \neq p$, which means that a metaedge can connect only vertices within one level of hierarchy (one scale).

The distance in the spatial domain ρ^S is the Euclidean distance between the centroids of two vertices in the (x,y,t) space. The distance in the wavelet and color (visual) domains ρ^W, ρ^V can be one of the distance learning metrics or similar to ρ^S, because the values both in the wavelet and visual domains are already in the feature space. However, compared to the spatial one, in the case of nonlinear wavelets it is less trivial to compute the distance between metavertices of a lower level j (e.g., face) as a combination of the distances between the ones of a higher level j (e.g., eye, lips, nose). The distance in the semantic domain ρ^O measures the difference between the entropies of two vertices, as proposed in [26].

D. Metagraph Projection

Metagraph projection can be perceived as convolution to a metagraph of a lower dimensional space, where the values of its projected metavertices are computed separately for each of the four domains either on the basis of its children vertices or independently:

$$v_i^{j,d} = f^d(v_i^{j+1,d}), \quad (5)$$

where d – is one of the four domains S, W, V and O. The cumulative value of a metavertex is a weighted sum of the children values in the four domains:

 ISBN: 978-1-61208-320-9

$$v_i^j = \sum_d w_v^d f^d(v_k^{j+1,d}). \quad (6)$$

Similarly, the cumulative value of an edge between two metavertices is a weighted sum of the edges in the four domains:

$$e(v_m^j, v_k^j) = \sum_d w_e^d \rho^d(v_m^{j,d}, v_k^{j,d}). \quad (7)$$

The weights w_v^d and w_e^d control the impact of a value and distance (edge) in a certain domain *d* on the overall result.

E. Metagraph construction

In this work, construction of a metagraph, which represents our model, is conducted in a frame-by-frame way, however, there are no limitations to implement a voxel-by-voxel way.

First, the frame is divided into 2-4 square blocks depending on the frame size. These blocks automatically become the lowest level (highest in terms of a hierarchy) blocks. For each such block we then apply the WT adaptive algorithm (see above). After each its step the blocks are further divided into 2-4 blocks together with lowering σ. In result, we obtain the metagraph MG_1, in which some branches of its hierarchy become deeper, whereas for some of them this algorithm interrupts after two-three iterations. Simultaneously, for each metavertex independently on its level we calculate: in the spatial domain, its centroid relatively to the lower (upper in terms of a hierarchy) level metavertex and positions of the local maxima of the responses to the wavelet filters $\psi_{\sigma,\theta}$; in the wavelet domain, a distribution of the sums of these responses for different θ; in the color domain, a color distribution in the HSV color space. Currently, values in the semantic domain we keep blank ($w_v^O = 0$) and to evaluate preliminary results of our model we also assign constant values to the weights in other domains: $\{w_e^S, w_e^W, w_e^V, w_e^O\} = \{1, 1, 0.5, 0\}$. Clearly, their assignment requires more investigation, and in experimental studies various influences of each domain on the correct result depending on a metavertex and its level have been observed.

The next frame is processed analogously in order to construct the metagraph MG_2. In addition, for each its vertex we must determine whether it matches to the vertex at the same position in MG_1 or does not. In the latter case, we calculate the difference in terms of the transformations in respective value domains. Translation (Δc) is calculated by the shift of the local maxima of the responses, rotation ($\Delta\theta$) – by the difference in sums distributions, Δvis – by appearance (disappearance) of new (old) strong responses. Scale change is a change of the metavertex level in the metagraph hierarchy. However, scaling less than two times is not detected due to the dyadic WT we applied, whereas in practice, the scaling varied in a broader range (although, mostly in the range of 1.1 – 2 times), so update of our WT's power base should be considered.

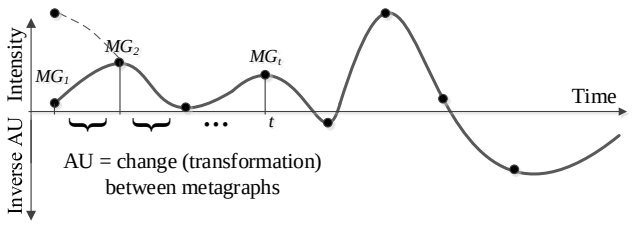


Figure 3. An action unit in terms of metagraphs. Dashed is a possible transformation to the same metagraph (MG_2).

Thus, for each frame, we construct a metagraph. All the metagraphs, excluding the first one are temporal, which means that when the difference between the first (MG_1) and the other (MG_t) metagraphs is found (see (8) below), the transformations (Δc, $\Delta\theta$ and Δvis) to get this difference are clustered to appropriate metavertices of (MG_1).

For each of the vertices of the metagraphs for the following frames we try to find the transformations from either the previous frame only, or from some combination of the previous frames.

In this sense, an action unit is a transformation of a certain metavertex corresponding to a facial element (Fig. 3). Formally, having only one frame in the middle of the frame sequence we cannot certainly infer the current AU, previous states must be known to avoid ambiguities.

Facial activity at some timestamp t is then a set of action units' intensities (a set of transformations) at this timestamp.

V. Training and Evaluation of the Model

Training of our model implies solving the following optimization problem:

$$\arg \min_{w_v^d, w_e^d, T_i} e(MG_1, MG_t). \quad (8)$$

where MG_1 – a metagraph of the first frame, MG_t – a metagraph of the frame at the timestamp t, T_i – a transformation (mapping). Thus, we need to find such mappings and weights, which can transform MG_1 to some MG^*_1, so that $e(MG^*_1, MG_t)$ would be a minimum. In general, this is a complex graph matching problem. In this study, to solve it we assume small changes between frames and as a result, metavertices at the same positions do not differ too much and can be matched more confidently.

Our solution of (8) is composed of two trainings: training the model to recognize attributes and to recognize AUs by these attributes.

A. Learning the attributes

The first training is inspired by the works on attribute learning, such as [12], as one of the ways to deal with the zero-shot learning problem emerged in this research. In our case, it means that no training data for new action units is available. To classify a newly defined AU, the model has to be able to multiclassify its attributes: v, Δc, $\Delta\theta$ and Δvis (see Section 3) during Δt, given metavertices and metaedges associated with a video voxel $\langle \Delta x, \Delta y, \Delta t \rangle$ (or, at least, two

frames) (Fig. 1). Consequently, we need to implement either a supervised or unsupervised multiclass learning method; afterwards, feed the model with training attributive data. In both approaches, weights must be first assigned manually or randomly before a training procedure for the cumulative value (6) and/or for the edge measure (7).

1) Supervised mode

The advantage of Random Forests (exemplified in Fig. 1) among supervised multi classifiers is that their trees are hierarchical by definition and might be associated with the hierarchy of our model further integrating it. Additionally to assigning the weights in (6), thresholds τ for every node of each tree must be also assigned, e.g., as in [27], and then respectively compared with projected values. Training assumes iterative changing of these weights and thresholds for every node of each tree until accuracy is increasing, the trees are not too deep and the gain in information is not sufficient.

2) Unsupervised mode

Among unsupervised methods, k-means, self-organizing maps and other classifiers can be trained. In any case, training assumes an iterative grouping of metavertices with smaller distances (7) between each other closer until no improvement (in the sense of some error function) can be reached. The advantage of these methods is a less tedious training process, since no labeling is required; and facial segmentation can be more objective if a metric is properly chosen, because a human expert is less involved.

B. Learning the AUs

There are two ways to solve the second task, i.e., to recognize AUs by their attributes. First, after semantic values for all facial elements are assigned in a supervised way, each AU can be defined as a set of rules in an xml/owl file. These rules must be written in close accordance with the FACS manual. Another prospective way of learning AUs is to recognize the same attributes from videos of the MMI dataset [33], in which a lot of AUs are labeled, and to infer these rules automatically.

At this stage, all attributes and action patterns (APs) were just clustered in an unsupervised way and we can apply either of the methods in the next works. Note, that feature extraction using the family of unsupervised methods can be tuned to be reliable, but, as it will be shown below, unsupervised classification of AUs themselves is not as reliable, because it is difficult to relate output clusters (APs) with the required classes (AUs).

C. Dataset

Our model can be trained and tested using a labeled dataset with real video [18, 19], images sequences [20] or a synthetic one with inherently labeled action units, for instance generated by the means of [29-31]. In [27], real and synthetic datasets complemented each other, which led to high recognition scores of body pose recognition, and therefore, this approach should be adopted in this study in the future studies.

TABLE I. RESULTS OF THIS WORK FOR THE DATASET [18]

Session Id	Subject Id	No. of AUs (TP)	No. of APs (TP)	No. of false APs (FP)	Overall No. of positives
21	2 (Operator)	6	21	15	36
	3 (User)	9	24	12	36
29	3 (Operator)	9	23	12	35
	16 (User)	7	17	19	36
64	7 (Operator)	11	16	8	24
	11 (User)	9	15	13	28

D. Evaluation

This work is in progress and to determine and, perhaps, correct the further direction of its development we collected qualitative results of a demo version of the model presented above for several subjects from the Semaine Database [18], which seemed to be closer to real environment compared to other datasets.

A simple .NET Framework (ver. 4.5) application integrated with the MATLAB API (ver. 2012a) was developed. The first part was used for object oriented metagraph implementation and abstract manipulation, and the second part was used for wavelet decomposition and calculations of transformations and was compiled for .NET using NE Builder.

The number of facial action patterns (No. of APs, Table 1) that our model clustered turned out to be far more than the number of facial AUs from FACS (No. of AUs), even though they overlapped partially, e.g., 6 from 21.

The coincidences mostly occurred when a particular expression and a respective AU was very intensive, whereas during substantial periods of time expressions were unclear, but it does not mean there was no AU. Another set of ambiguities were observed when a person was talking, which resulted in a lot of APs, which we could not always correspond to one of the AUs. Since the database that we used is labeled only using feeltrace annotations, it was difficult to check our results correctly, therefore we measured them categorically. To calculate the categorical error rate we counted the overall number of action unit types present in a video by watching it, and compared it to the number of output clusters returned by our model, which we could attribute to some AU with high confidence, even though the exact AU was unclear.

TABLE II. RESULTS OF SOME PREVIOUS WORKS

AUs	Method	Dataset	CR/F_1/PR, %
15+	Multi-state geometric face model [3]	CK	82-96.7/-/-
27	Free-form Deformations (FFD) + GentleBoost + Hidden Markov Model (HMM) [4]	MMI	94.3/65.1/59.7
18		CK	89.78/72.14/70.25
15	Viola-Jones + ASM + Gabor filters + GentleAdaboost [5]	private	95.9 (agreement rate)/-/-
9+	This work	Semaine	-/-/58.3

 ISBN: 978-1-61208-320-9

We also counted the number of false positive APs (No. of false APs) which included mismatched metavertices or invalid transformations. Having no results about negatives, we were only able to calculate the precision rate (PR), Altogether, the challenges described above led to an indecent number of false positive errors and PR compared to some previous works (Table 2, in which CR is the classification rate, AUs – the number of analyzed AUs), even though we did not measure them frame by frame.

VI. CONCLUSION AND FUTURE WORK

In this paper, a concept of the model for multilevel facial activity recognition and description is presented, and emerged technical and scientific challenges are discussed. Even though the model is not formalized and not explained in detail here, it promises to fulfill the requirements of this research: (1) recognize more FACS-based action units more accurately and in more real conditions compared to the previous studies; (2) be able to recognize the new ones, including body AUs and those defined by an expert using primitive attributes from the ontological network; (3) be integral and scalable for multiple persons. Among the hardest challenges are reliable quantization of classes and attributes and the complexity of classifying posterior-anterior movements, since they are not intrinsic to a video.

The suggested four level model should not be perceived as an overcomplete application of heavy methods. On the contrary, the metagraph model allows representing low-, mid- and high-level methods using mappings of metavertices making the model more homogeneous and flat. Indeed, this is an important theoretical implication, that many models can be represented using metagraphs and their mappings, even though the mappings are not always trivial.

The preliminary evaluation results demonstrated that our model needs further development, tuning and more comprehensive evaluation to solve the problems of human behavior retrieval and human-computer multimodal interaction more efficiently, which must be the focus of further research studies.

ACKNOWLEDGMENT

This research is partially supported by Bauman Moscow State Technical University. Portions of the research in this paper use the Semaine Database collected for the Semaine project (www.semaine-db.eu) [18].

REFERENCES

[1] T. Kanade, “Visual Processing and Understanding of Human Faces and Bodies”, 9th International Conference (ICVS 2013), Jul. 2013, Keynote Talk, URL: http://workshops.acin.tuwien.ac.at/ICVS/downloads/Kanade_ICVS2013.pdf (December 16, 2013)

[2] P. Ekman, “What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System”, New York: Oxford University Press, 2005.

[3] Y. L. Tian , T. Kanade, and J. F. Cohn, "Recognizing action units for facial expression analysis", IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 23, no. 2, Feb. 2001, pp. 97-115.

[4] S. Koelstra, M. Pantic, and I. Patras, “A Dynamic Texture Based Approach to Recognition of Facial Actions and Their Temporal Models”, IEEE Trans. on Pattern Analysis and Machine Intelligence, Nov. 2010, pp. 1940-1954.

[5] J. Hamm, C. G. Kohler, R. C. Gur, and R. Verma, "Automated facial action coding system for dynamic analysis of facial expressions in neuropsychiatric disorders", J. Neurosci. Methods, vol. 200, no. 2, Sep. 2011, pp. 237-256.

[6] S. Mallat, "A theory for multiresolution signal decomposition: The wavelet representation", IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 11, Jul. 1989, pp. 674 -693.

[7] T. S. Lee "Image representation using 2D Gabor wavelets", IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 18, no. 10, Oct. 1996, pp. 959 -971.

[8] I. W. Selesnick, R. G. Baraniuk, and N. C. Kingsbury, "The dual-tree complex wavelet transform", IEEE Signal Process Mag., vol. 22, no. 6, Nov. 2005, pp. 123-151.

[9] B. Solmaz, S. M. Assari, and M. Shah, “Classifying Web Videos using a Global Video Descriptor”, Machine Vision and Applications (MVA), vol. 24, iss. 7, Oct. 2013, pp. 1473-1485.

[10] A. Basu and R. W. Blanning, “Metagraphs and Their Applications”, Integrated Series in Information Systems, Vol. 15, 2007, 172 p.

[11] L. Wiskott, J.-M. Fellous, N. Kruger, and C. von der Malsburg, “Face Recognition by Elastic Bunch Graph Matching”, IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 17, no. 7, Jul. 1997, pp. 775-779.

[12] V. Ferrari and A. Zisserman, “Learning visual attributes”, Advances in Neural Information Processing Systems, Dec. 2007, pp. 433-440.

[13] M. S. Bartlett and J. Whitehill, “Automated facial expression measurement: Recent applications to basic research in human behavior, learning, and education”, In Oxford Handbook of Face Perception, Oxford University Press, 2011, pp. 489-514.

[14] A. N. Anuashvili, “Fundamentals of Objective Psychology”, Moscow-Warsaw, 2005. (in Russian)

[15] A. Heloir, M. Neff, and M. Kipp, “Exploiting Motion Capture for Virtual Human Animation: Data Collection and Annotation Visualization”, Proc. Workshop on "Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality", 2010, URL: http://embots.dfki.de/doc/Heloiretal10.pdf (December 16, 2013)

[16] C. Delgado, R. Garcia, J. I. Navarro, and E. Hinojo, “Functional analysis of challenging behaviours in people with severe intellectual disabilities using The Observer xT 10.0 software”, Proc. Measuring Behavior, Aug. 2012, pp. 365-367.

[17] B. Knyazev, “Human nonverbal behavior multi-sourced ontological annotation”, Proc. International Workshop on Video and Image Ground Truth in Computer Vision Applications (VIGTA '13), Jul. 2013, Article 2 , 8 p.

[18] G. McKeown, M. Valstar, R. Cowie, and M. Pantic, “The Semaine Corpus of Emotionally Coloured Character Interactions,” Proc. IEEE Int'l Conf. Multimedia and Expo (ICME), Jul. 2010, pp. 1079-1084.

[19] X. Zhang et al., “A High-Resolution Spontaneous 3D Dynamic Facial Expression Database”, Proc. 10th IEEE Int'l Conference and Workshops on Automatic Face and Gesture Recognition (FG'13), Apr. 2013, pp. 1-6.

[20] P. Lucey, J. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression," Proc. 2010 IEEE Computer Society Conference on CVPR Workshops, Jun. 2010, pp. 94-101.

 ISBN: 978-1-61208-320-9

[21] P. Ekman and W. Friesen, “Facial Action Coding System: A Technique for the Measurements of Facial Movements”, Consulting Psychologists Press, 1978.

[22] C. Y. Chen and K. Grauman, “Efficient activity detection with max-subgraph search”, Proc. CVPR, Jun. 2012, pp. 1274-1281.

[23] C. H. Hjortsjö, “Man's face and mimic language”, Studentlitteratur, 1969.

[24] N. Dael, M. Mortillaro, and K. R. Scherer, “The Body Action and Posture Coding System (BAP): Development and Reliability”. Journal of Nonverbal Behavior, vol. 36, iss. 2, Jun. 2012, pp. 97-121.

[25] R. J. E. Merry and M. Steinbuch, "Wavelet theory and applications", literature study, Eindhoven University of Technology, 2005.

[26] J. Calmet and A. Daemi, “From Entropy to Ontology”, Proc. 4th Int'l Symp. “From Agent Theory to Agent Implementation”, Apr. 2004, URL: http://www.iks.kit.edu/fileadmin/User/calmet/papers/AT2AI4.pdf (December 16, 2013)

[27] J. Shotton et al., “Real-time human pose recognition in parts from single depth images”, Proc. CVPR 2011, Jun. 2011, pp. 1297-1304.

[28] I. A. Essa, “Analysis, Interpretation and Synthesis of Facial Expression”, PhD thesis, MIT, Media Lab, 1995.

[29] Autodesk MotionBuilder, URL: http://www.autodesk.com/products/motionbuilder/overview (December 16, 2013)

[30] Di3D Inc., URL: http://www.di3d.com (December 16, 2013)

[31] E. Roesch, L. Tamarit, L. Reveret, D. Grandjean, D. Sander, and K. Scherer, "FACSGen: A Tool to Synthesize Emotional Facial Expressions Through Systematic Manipulation of Facial Action Units," Journal of Nonverbal Behavior, vol. 35, 2011, pp. 1-16.

[32] The Nonverbal Dictionary - Center for Nonverbal Studies, URL: http://center-for-nonverbal-studies.org/6101.html (December 16, 2013)

[33] M. F. Valstar and M. Pantic, “Induced Disgust, Happiness and Surprise: an Addition to the MMI Facial Expression Database”, Proc. International Language Resources and Evaluation Conference, Malta, May 2010, pp. 65-70.

 ISBN: 978-1-61208-320-9

Face Recognition Using Histogram-based Features in Spatial and Frequency Domains

Qiu Chen[1], Koji Kotani[2], Feifei Lee[3], and Tadahiro Ohmi[3]

[1] Department of information and Communication Engineering, Faculty of Engineering, Kogakuin University
[2] Department of Electronics, Graduate School of Engineering, Tohoku University
[3] New Industry Creation Hatchery Center, Tohoku University
e-mail: chen@cc.kogakuin.ac.jp

Abstract—**Previously, we proposed an efficient algorithm using vector quantization (VQ) histogram for facial image recognition in low-frequency DCT domains. In this paper, we newly utilize Local Binary Pattern (LBP) histogram in spatial domain. These two histograms, which contain both spatial and frequency domain information of a facial image, are utilized as a very effective personal feature. Publicly available AT&T database is used for the evaluation of our proposed algorithm, which is consisted of 40 subjects with 10 images per subject containing variations in lighting, posing, and expressions. It is demonstrated that face recognition using combined histogram-based features can achieve much higher recognition rate.**

Keywords-Face recognition; Vector quantizaiton (VQ); Local Binary Patterns (LBP); DCT coefficients.

I. INTRODUCTION

In recent years, face recognition has been hot research topic due to its potential applications in many fields, such as law enforcement applications, security applications and video indexing, etc. Many algorithms have been proposed for solving face recognition problem [1]-[11].

These algorithms can be roughly divided into two categories, namely, statistics-based and structure-based approaches. Statistics-based approaches [5], [6], [7] attempt to capture and define the face as a whole. The face is treated as a two dimensional pattern of intensity variation. Under this approach, the face is matched through finding its underlying statistical regularities. Based on the use of the Karhunen-Loeve transform, PCA [5] is used to represent a face in terms of an optimal coordinate system which contains the most significant eigenfaces and the mean square error is minimal. However, it is highly complicated and computational-power hungry, making it difficult to implement them into real-time face recognition applications. Structure-based approach [3], [4] uses the relationship between facial features, such as the locations of eye, mouth and nose. It can implement very fast, but recognition rate usually depends on the location accuracy of facial features, so it cannot give a satisfied recognition result.

There are many other algorithms have been used for face recognition, such as Local Feature Analysis (LFA) [11], neural network [1], local autocorrelations and multi-scale integration technique [2], and other techniques have been proposed.

Discrete Cosine Transform (DCT) is not only widely used in many image and video compression standards [12], but also for pattern recognition as a means of feature extraction [13]-[21]. The main merit of the DCT is its relationship to the KLT [18]. It has been demonstrated that DCT best approach KLT [23], but DCT can be computationally more efficient than the KLT depending on the size of the KLT basis set.

In our previous work [27], we present a simple, yet highly reliable face recognition algorithm using vector quantization (VQ) method for facial image recognition in compressed DCT domain. Feature vectors of facial image are firstly generated by using DCT coefficients in low frequency domains. Then, the codevector referred count histogram, which is utilized as a very effective facial feature value, is obtained by VQ processing.

This algorithm can be considered utilizing the phase information of DCT coefficients by applying binary quantization on the DCT coefficient blocks. If we could combine spatial information of the facial image, the composite features of face are expected to be more robust and effective. In this paper, we utilize Local Binary Patterns (LBP) to represent facial features in spatial domain. These two histograms, which contain spatial and frequency domain information of a facial image, are utilized as a very effective personal value. Recognition results with different type of histogram features are first obtained separately and then combined by weighted averaging.

This paper is organized as follows. A brief introduction to DCT as well as LBP histogram is given in Section II. Our proposed face recognition method will be described in detail in Section III. Experimental results will be discussed in Section IV. Finally, we make a conclusion in Section V.

II. RELATED WORKS

A. Discrete Cosine Transform (DCT)

Discrete Cosine Transform (DCT) is used in JPEG compression standard. The DCT transforms spatial information to decoupled frequency information in the form of DCT coefficients.

2D DCT with block size of N × N is defined as follows:

 ISBN: 978-1-61208-320-9

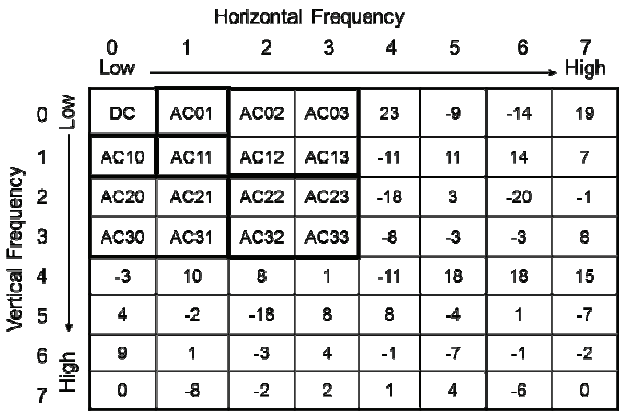


Figure 1. Generation of Low-frequency DCT coefficients (used as phase information)

$$C(u,v)=\alpha(u)\alpha(v)\sum_{x=0}^{N-1}\sum_{y=0}^{N-1}f(x,y)\cdot\cos\left(\frac{(2x+1)u\pi}{2N}\right)\cos\left(\frac{(2y+1)v\pi}{2N}\right) \quad (1)$$

$$f(x,y)=\sum_{u=0}^{N-1}\sum_{v=0}^{N-1}\alpha(u)\alpha(v)C(u,v)\cdot\cos\left(\frac{(2x+1)u\pi}{2N}\right)\cos\left(\frac{(2y+1)v\pi}{2N}\right) \quad (2)$$

$$\text{where, } \alpha(\omega)=\begin{cases}\frac{1}{\sqrt{N}} & : \quad for \quad \omega=0\\ \frac{2}{\sqrt{N}} & : \quad for \quad \omega=1,2,...,N-1\end{cases} \quad (3)$$

B. Face recognition using Binary vector quantization in low-frequency DCT domains

In our previous work [27], we proposed a feature extraction algorithm for face recognition using binary vector quantization (VQ) to generate feature vectors of facial image from DCT (Discrete Cosine transform) coefficients in low frequency domains.

First, low-pass filtering is carried out using 2-D moving filter. Block segmentation step, in which facial image is divided into small image blocks with an overlap, namely, by sliding dividing-partition one pixel by one pixel, is the following. Then the pixels in the image blocks (typical size is 8x8) are transformed using DCT according to the equation (1).

A typical sample of transformed block is shown in Figure 1. The DCT coefficients of the image block are then used to form a feature vector. From left to right and top to bottom, the frequency of coefficients changes from low to high as shown in Figure 1. Because low frequency component is more effective for recognition, we only use the coefficients on the left and above to extract features. The equation for calculation is shown below.

$$\begin{aligned} a[0] &= AC01;\\ a[1] &= AC11;\\ a[2] &= AC10;\\ a[3] &= (AC02 + AC03 + AC12 + AC13) / 4;\\ a[4] &= (AC22 + AC23 + AC32 + AC33) / 4;\\ a[5] &= (AC20 + AC21 + AC30 + AC31) / 4 \end{aligned} \quad (4)$$

where a[i] is the element of extracted feature vector, and d[i][j] is the coefficient value at point (i, j), respectively.

After that, quantization of the feature vectors is implemented. There are only 2 types of value for each a[i], so the number of combination of 6-dimensional vector is 64, which is very easy and fast to be determined. The number of vectors with same index number is counted and feature vector histogram is easily generated, and it is used as histogram feature of the facial image. In the registration procedure, this histogram is saved in a database as personal identification information. In the recognition procedure, the histogram made from an input facial image is compared with registered individual histograms and the best match is output as recognition result.

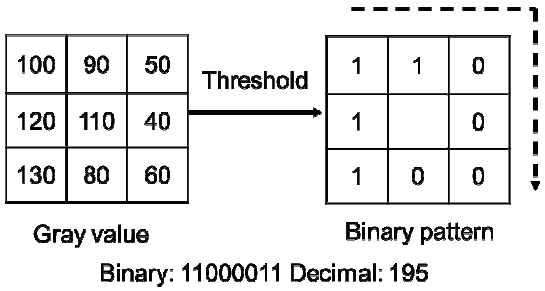


Figure 2. Fundamental LBP operator.

C. Local Binary Patterns (LBP) histogram

The original LBP operator proposed by Ojala et al. [28], is used for robust texture description. The operator labels the pixels of an image by thresholding the 3x3-neighbourhood of each pixel with the center value and considering the result as a binary number. Then the histogram of the labels can be used as a texture descriptor. Figure 2 shows an illustration of the basic LBP operator.

The limitation of the fundamental LBP operator is its small 3x3 neighborhood which can not capture dominant features with large scale structures. Hence, the operator later is extended to use neighborhood of different sizes. As shown in Figure 3, LBP(P, R) means P sampling points on a circle of radius of R to get LBP features. For instance, LBP(8, 2) means comparing a neighborhood of 8 on the circle of radius of 2 to get LBP features.

After labeling an image with the LBP operator, the histogram of the labeled image $p(x,y)$ can be defined as

$$H_u=\sum_{x,y}U(p(x,y)=u), u=0,1,...,n-1 \quad (5)$$

 ISBN: 978-1-61208-320-9

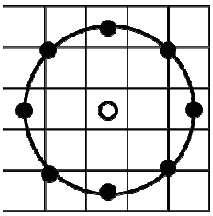


Figure 3. The circular (8,2) neighborhood.

Where n is the number of different labels produced by the LBP operator and

$$U(A) = \begin{cases} 1, A = ture \\ 0, A = false \end{cases} \quad (6)$$

A LBP histogram can effectively describe the distribution of the local micro-patterns over a whole face image without any indication about their locations. For efficient face representation, one should also retain spatial information. Thus, a face image can be equally divided into small regions. And then, the LBP features extracted from each sub-region are concatenated into a single histogram as

$$H_{u,v} = \sum_{x,y} U(p(x,y) = u) U\{(x,y) \in R_v\} \quad (7)$$

where $u = 0,1,..., n-1$ *and* $v = 0,1,..., m-1$.

III. PROPOSED METHOD

As described in Section II (B), we have proposed a face recognition algorithm by applying binary quantization on the low-frequency DCT coefficient blocks, which was demonstrated to be effective for face recognition by experimental results. Actually, it can be thought that phase information of low-frequency DCT coefficients is extracted by this algorithm. If we could combine spatial information of the facial image, the composite features of face are expected to be more robust and effective.

We utilize LBP to represent facial features in spatial domain. In this paper, we propose an improved face recognition algorithm using combined histogram-based features. Figure 4 shows proposed face recognition process steps. First, low-pass filtering is carried out using 2-D moving filter. This low-pass filtering is essential for reducing high-frequency noise and extracting most effective low frequency component for recognition.

Block segmentation step, in which facial image is divided into small image blocks with an overlap, namely, by sliding dividing-partition one pixel by one pixel, is the following. Then the pixels in the image blocks (typical size is 8x8) are transformed using DCT according to the equation (1). After generations of low-frequency DCT coefficients, binary quantization of the feature vectors is implemented as

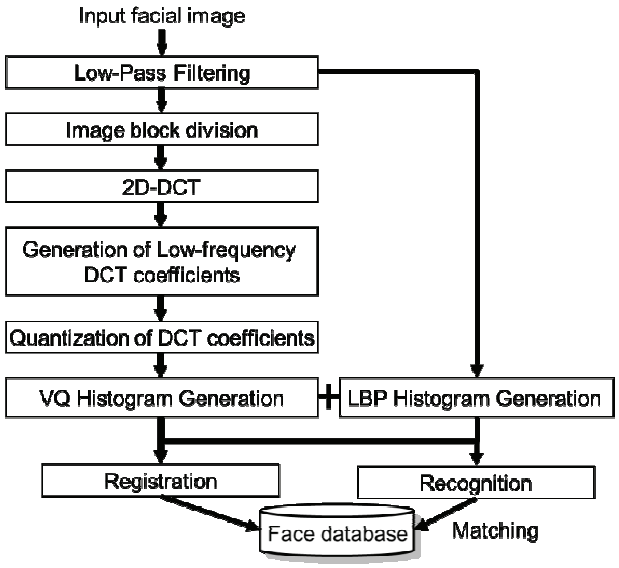


Figure 4. Face recognition process using combined histogram-based features.

described in Section II (B), and then VQ histogram of low-frequency DCT coefficients is created.

On the other hand, LBP histogram of facial image in spatial domain is generated after filtering processing. Once the features have been selected, LBP histogram is created by using formula (7) as described in Section II(C).

These two histograms, which contain both spatial and frequency domain information of a facial image, are utilized as a very effective personal feature. Recognition results with different type of histogram features are first obtained separately and then combined by weighted averaging.

IV. EXPERIMENTAL RESULTS AND DISCUSSIONS

A. ORL database

Face database of AT&T Laboratories Cambridge [25], [26] is used for recognition experiments. In the database, 10 facial images for each of 40 persons (totally 400 images) with variations in face angles, face sizes, facial expressions, and lighting conditions are included. Each image has a resolution of 92x112. Five images were selected from each person's 10 images as probe images and remaining five images are registered as album images. Recognition experiment is carried out for 252 ($_{10}C_5$) probe-album combinations by rotation method. The algorithm is programmed by ANSI C and run on PC (Pentium(R)D processor 840 3.2GHz).

B. Results and discussions

Figure 5 shows the comparison of the recognition results with different features. The average recognition rates obtained by each case with block size of 8x8 are shown here.

 ISBN: 978-1-61208-320-9

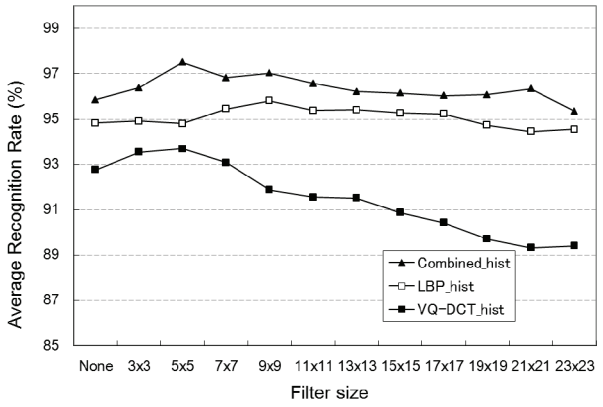


Figure 5. Comparison of recognition results

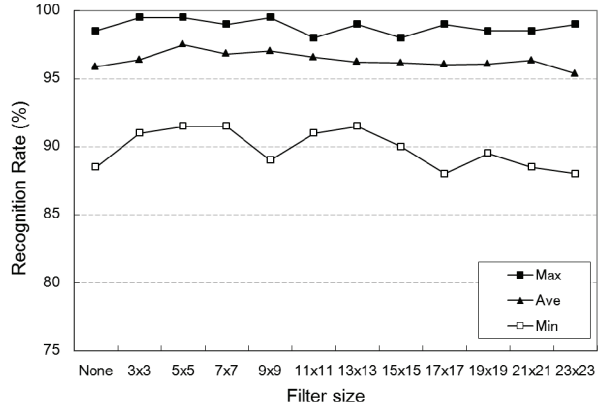


Figure 6. Recognition rate as a function of filter size (image block size is 8x8 for DCT coefficients here)

Recognition success rates are shown as a function of filter size. Recognition results only using LBP histogram ("LBP_hist") achieved 95.8% at the filter size of 9x9, average recognition rate increases combined with VQ histogram of low-frequency DCT coefficients ("Combined_hist"). The maximum of the average rate 97.5% is achieved, which is 3.8% higher than that only using VQ histogram in our previous work ("N8_VQ_hist", the maximum of the average rate is 93.7%) [27].

Figure 6 shows recognition results using combined features with the same weighting coefficient of two histogram features. Recognition success rates are shown as a function of filter size. "Max," "Min" and "Ave" stand for the best case, worst case, and average results in 252 ($_{10}C_5$) probe-album combinations, respectively. The highest average recognition rate of 97.5% is obtained at the filter size of 5x5. Low pass filter is effective for eliminating noise component and extracting important frequency component for recognition.

By combining these two different features, namely spatial and frequency domain information of a facial image, the most important information for face recognition can effectively be extracted.

V. CONCLUSIONS AND FUTURE WORK

We have developed a very simple yet highly reliable face recognition method using features extracted from low-frequency DCT domain and spatial domain of a facial image, which is combined with VQ histogram and LBP histogram. Excellent face recognition performance has been verified by using publicly available ORL database. The effect of the image block size will be discussed in our future work, as well as the performance evaluation of the face recognition using larger face database.

REFERENCES

[1] R. Chellappa, C. L. Wilson, and S. Sirohey, "Human and machine recognition of faces: a survey," Proc. of IEEE, vol. 83, no. 5, 1995, pp.705-740.

[2] S. Z. Li and A. K. Jain, "Handbook of face recognition," Springer, New York, 2005.

[3] R. Brunelli and T. Poggio, "Face recognition: features versus templates," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 15, no. 10, 1993, pp. 1042-1052.

[4] L. Wiskott, J. M. Fellous, N. Kruger, and C. Malsburg, "Face recognition by elastic bunch graph matching," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 19, no. 10, 1997, pp.775-780.

[5] M. Turk and A. Pentland, "Eigenfaces for recognition," Journal of Cognitive Neuroscience, vol. 3, no. 1, 1991, pp. 71-86.

[6] W. Zhao, "Discriminant component analysis for face recognition," Proc. ICPR'00, Track 2, 2000, pp. 822-825.

[7] K.M. Lam, H. Yan, "An analytic-to-holistic approach for face recognition based on a single frontal view," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 20, no. 7, 1998, pp. 673-686.

[8] M. S. Bartlett, J. R. Movellan, and T. J. Sejnowski, "Face recognition by independent component analysis," IEEE Trans. on Neural Networks, vol. 13, no. 6, 2002, pp. 1450-1464.

[9] B. Moghaddam and A. Pentland, "Probabilistic visual learning for object representation," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 19, no. 7, 1997, pp. 696-710.

[10] S. G. Karungaru, M. Fukumi, and N. Akamatsu, "Face recognition in colour images using neural networks and genetic algorithms," Int'l Journal of Computational Intelligence and Applications, vol. 5, no. 1, 2005, pp. 55-67.

[11] P. S. Penev and J. J. Atick, "Local feature analysis: a general statistical theory for object representation," Network: Computation in Neural Systems, vol. 7, no. 3, 1996, pp. 477-500.

[12] W. B. Pennebaker and J. L. Mitchell, "JPEG still image data compression standard," Van Nostrand Reinhold, New York, 1993.

[13] H. B. Kekre, T. K. Sarode, P. J. Natu, and S. J. Natu, "Transform based face recognition with partial and full feature vector using DCT and Walsh transform," Proc. of the Int'l Conf. & Workshop on Emerging Trends in Technology, 2011, pp. 1295-1300.

[14] Z. Liu and C. Liu, "Fusion of color, local spatial and global frequency information for face recognition," Pattern Recognition, vol. 43, Issue 8, Aug. 2010, pp. 2882-2890.

 ISBN: 978-1-61208-320-9

[15] H. F. Liau, K. P. Seng, L. M. Ang, and S. W. Chin, "New parallel models for face recognition," Recent Advances in Face Recognition, Edited by K. Delac etc., InTech, 2008, pp. 15-26.

[16] R. Tjahyadi, W. Liu, S. An and S. Venkatesh, "Face recognition via the overlapping energy histogram," Int'l Joint Conf. on Artificial Intelligence, 2007, pp. 2891-2896.

[17] D. Zhong and I. Defee, "Pattern recognition in compressed DCT domain," Proc. of Int'l Conf. on Image Processing, vol. 3, 2004, pp. 2031 - 2034.

[18] Z. M. Hafed and M. D. Levine, "Face recognition using the Discrete Cosine Transform," Int'l Journal of Computer Vision, vol. 43, no. 3, 2001, pp. 167-188.

[19] S. Eickeler, S. Müller and G. Rigoll, "Recognition of JPEG compressed face images based on statistical methods," Image and Vision Computing Journal, Special Issue on Facial Image Analysis, vol. 18, no. 4, Mar. 2000, pp. 279-287.

[20] S. Eickeler, S. Müller, and G. Rigoll, "High quality face recognition in JPEG compressed images," Proceeding of Int'l Conf. on Image Processing, vol. 1, Oct. 1999, pp. 672-676.

[21] V. Nefian and M. H. Hayes, "Hidden Markov models for face recognition," Int'l Conf. on Acoustics, Speech, and Signal Processing, May 1998, pp. 2721-2724.

[22] M. Shneier and M Abdel-Mottaleb, "Exploiting the JPEG compression scheme for image retrieval," IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 18, no. 8, Aug. 1996, pp. 849-853.

[23] A. Jain, Fundamentals of Digital Image Processing, Prentice: Englewood Cliffs, NJ, 1989.

[24] A. Gersho and R. M. Gray, "Vector quantization and signal compression," Kluwer Academic, 1992.

[25] AT&T Laboratories Cambridge, "The database of faces," at http://www.cl.cam.ac.uk/research/dtg/attarchive/ facedatabase. html [retrieved: Dec. 2013].

[26] F. Samaria and A. Harter, "Parameterisation of a stochastic model for human face identification," 2nd IEEE Workshop on Applications of Computer Vision, 1994, pp. 138-142.

[27] Q. Chen, K. Kotani, F. F. Lee, and T. Ohmi, "Face recognition using VQ Histogram in compressed DCT domain," Journal of Convergence Information Technology, vol. 7, no. 1, 2012, pp. 395-404.

[28] T. Ojala, M. Pietikainen, T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 7, 2002, pp. 971-987.

[29] K. Kotani, Q. Chen, F. F. Lee, and T. Ohmi "Region-division VQ histogram method for human face recognition," Intelligent Automation and Soft Computing, vol. 12, no. 3, 2006, pp. 257-268.

[30] P. J. Phillips, H. Wechsler, J. Huang, and P. Rauss, "The FERET database and evaluation procedure for face recognition algorithms," Image and Vision Computing J, vol. 16, no. 5, 1998, pp. 295-306.

 ISBN: 978-1-61208-320-9

Understanding Users' Continued Use of Online Games: An Application of UTAUT2 in Social Network Games

Xiaoyu Xu
Turku School of Economics
University of Turku
Turku, Finland
e-mail: xiaoxu@utu.fi

***Abstract*—Online gaming market is thriving but accompanied with fierce competitions. Players' continued use of online games is critical for the success of online game providers. This study applies UTAUT2 as the research framework to identify the key determinants of social network game (SNG) players' continued use intention, and to explore the moderating effects of individual characters (such as age, gender, and user experience) on the proposed hypotheses as well. The research model is examined by analyzing 3919 valid empirical data collected among SNG players in China. The results suggest that social influence is the most important determinant of continuance intention to use SNG, followed by habit, fantasy, enjoyment, achievement and price value. User experience and age are not moderators, whereas gender exerts moderating effects on the paths from social influence, perceived enjoyment and price value to continuance intention.**

Keywords- IS Continuance, UTAUT2, Social Network Games, Online Games

I. INTRODUCTION

Online gaming is emerging as one of the fastest growing online entertainment industries with a continued increase in the number of participants [24]. Online gaming has become powerhouses of electronic-commerce and largely changed how the Internet users spend their leisure time [5]. However, the growing popularity and soaring revenue are accompanied with fierce competitions in online gaming industry. The features of Internet make it easy for online gaming players to access to and switch to alternative online games. Hence, how to retain the existing players and to prolong their playing duration in online games have attracted the attention of both practitioners and researchers [41].

Prior research on online games explored players' usage behavior (e.g., adoption, continued usage, and stickiness) in online games was mainly based on the dominant information systems (IS) theories, such as theory of reasoned action (TRA)[14], technology acceptance model (TAM) [11], and theory of planned behavior [2]. However, these theories were mainly developed in the work related settings to study employees' IT usage for utilitarian purposes. They might fall short in explaining individual usage of hedonic IS in home settings [42].

UTAUT2 was recently developed based on UTAUT which has been developed to explain users' technology adoption behavior in organizational context [37]. UTAUT2 was selected since it can provide several advantages in the current research context. Venkatesh et al. [38] indicated that the objective of developing UTAUT2 was to focus on individual consumers' use context. Thus, comparing to theories build in the organizational setting for studying utilitarian oriented IS usage; UTAUT2 may provide more insights to investigate online gaming player's behavior in home settings. Further, UTAUT2 was developed based on a rigorous theoretical model UTAUT which has superior performances comparing to other eight IS models in explaining individual IS usage. Venkatesh et al. [38] argued that compared to UTAUT, UTAUT2 showed significant improvements in explaining the variance of consumers' technology use intention. Moreover, several constructs, such as hedonic motivation, price value, and user habit were added into UTAUT2. These constructs were repeatedly examined in prior studies as the important determinants of individual IS continuance usage in home settings, and have not been theoretically incorporated and examined in UTAUT.

In the work of Venkatesh et al., the importance to extend or adapt UTAUT2 to different research contexts is highlighted. Venkatesh et al. argued that "compare to general theories, theories that focus on a specific context are considered to be vital in providing a rich understanding of a focal phenomenon and to meaningfully extend theories" (pp. 158). Therefore, it is critical to examine how UTAUT2 can be generalized to different research contexts. In prior literatures, little research has attempted to apply UTAUT2 in the research context of online gaming, e.g., social network games (SNG) defined as "a type of browser game distributed through social networks fitting to multiplayer and asynchronous game playing "[27]. Thus, our theoretical choice of examining the extension of UTAUT2 in online gaming is further justified.

In addition, it is indicated that when applying UTAUT2 to different research contexts, modification or extension of UTAUT2 might be needed in order to understand a focal phenomenon better. Venkatesh et al. advocated the examination of other key constructs that were salient to different research contexts when applying UTAUT2 to build the models, since new constructs can result in important changes in theories in different context. Online gaming is different from mobile Internet technology investigated in

 ISBN: 978-1-61208-320-9

UTAUT2. Prior researchers have suggested that the explanatory power of a particular model or theory would depend on the characteristic of the technology [21]. Thus, in the current study, UTAUT2 is selected as the research framework, and some modification is done in order to understand the phenomenon of continuous play of SNG.

The rest of the paper is arranged as follows. Research background and research model are discussed in the next section, followed by the presentation of the research method in Section III. Subsequently, research results are illustrated in Section IV. Then, the paper goes on with discussions towards the research findings in Section V. Finally, we present the conclusion of this study, and discussion of limitations in Section VI.

II. RESEARCH BACKGROUND AND RESEARCH MODEL

A. Social Network Games

Nowadays, social networks services (SNS) (e.g., Facebook, MySpace) have become popular among the Internet users. People are using SNS for different purposes, such as for entertainment and communication. For example, Facebook, the most popular SNS, until June in 2013, it has 1.5 billion monthly active users with an increase of 21 per cent compared to last year [13]. Meanwhile, there are millions of apps run on SNS. And among these apps, SNGs have made great success on SNS by attracting an increasing number of players all over the world, and "have spawned a whole new subculture" [6]. In spite of the huge popularity and rapid growth of SNG, research on SNG is still in an infant stage [38].

SNGs usually have some features in common. SNG players mainly play social games with people in their existing social networks, such as friends, family, and co-workers instead of virtual players meet through the game [29]. Most SNGs are designed to be easy for players to play [27], and SNG players can interact with others without the constraints of time as SNGs are asynchronous [28]. SNGs combine multiple elements from both SNS and online gaming.

B. Research Model

In UTAUT2, seven constructs are identified as the main determinants of continuous intention, namely performance expectancy, effort expectancy, social influence, facilitating conditions, hedonic motivation, price value, and habit. This study aims at investigating individual player's continued intention to use SNG. Thus, some modifications have been made in order to make the model fit better to explore the research context of SNG as discussed in Section I.

Venkatesh et al. [38] have suggested that hedonic motivation is one of the key factors determining IS users' behavioral intention in non-organizational contexts. In this study, perceived enjoyment and fantasy are employed as the two factors reflecting hedonic motivations in the SNG context. Li et al. [22] found that the hedonic gratification, such as perceived enjoyment and fantasy, determined individuals' continued intention to use SNG. Emotional response (such as enjoyment) and imaginary response (such as fantasy) have also been suggested to be important motivations for individual to conduct hedonic consumption [20].

Performance expectancy represents the utilitarian value of IS usage and emphasizes the benefits provided to consumers by using the technology [38]. The utilitarian benefit players expect to gain is the sense of achievement by engaging in kinds of activities in SNGs, such as gaining power or accumulating in-game symbols of wealth, competing with other players in the SNG, and achieving higher game levels [22][43]. Therefore, in the current study, achievement is used to reflect the utilitarian value driving individuals' continuance intention to use the SNG.

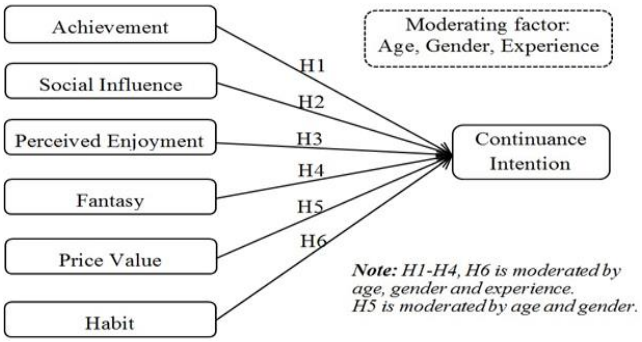
Effort expectancy is similar to perceived ease of use and means the degree of ease associated with consumers' use of technology. However, this construct has been argued to lose its influence on continuance intention when users accumulate experience during their continued use stage [18]. Moreover, a SNG is usually designed for players to obtain the game rules and skills easily. Thus, effort expectancy is not included in this study to explore continued use of SNG.

Facilitating conditions refer to consumers' perceptions of resources and support available to perform behavior. The purpose of the current study is to examine the players who have accumulated experience in SNG use. Venkatesh et al. [38] pointed out that the users with more experience depend less on external support. Furthermore, SNGs are featured as easy-learning curve, free-to-play pattern via no matter PC or mobile devices, and requiring less continuous time and effort [27],[28]. These features enable the SNG players to require little additional support for learning, device, location and time to continue playing a SNG. Hence, we assume the influence of facilitating condition can be marginal in the current research context and it is not included in our research model.

In this research, we also explore the moderating roles of individual characters (age, gender and user experience) on the relationships from independent variables to dependent variable as proposed in UTAUT2. The moderating effects exerted by age, gender and user experience have attracted attention in online gaming studies. Prior online gaming research claims different results. For example, Lin et al. [31] reported the moderating effect of gender on perceptions of online game loyalty, whereas, Ha et al. [17] claimed age was a more significant moderator on perceptions of online gaming loyalty, gender only exerted marginal moderating effect. Hence, the moderating effects of these individual characters should be examined in the current study.

Based on the above ground, six constructs are proposed to predict continuance intention, including achievement, social influence, perceived enjoyment, fantasy, price value, and habit. Since behavioral intention has been examined to be the dominant determinant of IS actual use in the prior IS research, in this study, we focus on exploring the determinants of continuance intention. The research model is presented in Figure1.

 ISBN: 978-1-61208-320-9



Figue 1. Research Model

C. Research Hypotheses

The achievement components refer to playing SNG to gain power, to progress rapidly, to accumulate in-game symbols of wealth or status, and to compete with others [7]. Suznjevic and Matijasevic [35] found that achievement was the most important motive for players to play the MMORPG. Prior research results towards online games have supported this argument and indicated that achievement positively predicts continuance intention to play online games [7], [22]. Thus, it is reasonable to expect that achievement will influence SNG continuance intention, and the following hypothesis is proposed:

H1. Achievement is associated with SNG continuance intention positively.

Social influence refers to the extent to which players perceive that important others believe they should continue playing a particular SNG [38]. Social influence is included as a major predictor of behavioral intention in UTAUT2. Lee [29] claimed that many players decided to play online games was just because their friends recommended them to do so. Similarly, Hsu and Lu [23] conducted an empirical study with 233 responses in the context of online games and supported the argument that social influence had significant impact on intention to play online games. Thus, it is reasonable to argue that SNG players are more likely to continue playing SNG if their friends encourage them to continue playing it. Hence, the following hypothesis is suggested:

H2. Social influence is associated with SNG continuance intention positively.

Perceived enjoyment in the current study refers to the extent to which the activity of playing the SNG is perceived to be enjoyable in its own right, apart from any performance consequences that may be anticipated [12]. Perceived enjoyment is theorized to predict behavioral intention directly [19]. In online games settings, Ha et al. [17] indicated that "games must, of course, provide players with enjoyment, as part of their basic nature". Players are more willing to persist in playing online games in the future when their behavior is prompted by intrinsic motivation, such as perceived enjoyment [40]. Thus, it seems reasonable to argue that SNG players are more likely to continue playing SNG if they perceive there perceive more enjoyment during their game playing process, and the following hypothesis is proposed:

H3. Perceived enjoyment is associated with SNG continuance intention positively.

Hedonic consumption studies in marketing discipline suggest that seeking product-related fantasy and imagine is an important determinant for pleasure-oriented consumption behavior [20]. Prior research in the online gaming context also identified fantasy as a major motive for individuals to play online games [26], [32]. In the SNG settings, players can construct and realize their fantasy by trying different identities, fancy avatars and conducting activities, which are not possible for them to do in the real life. It seems that SNG players are more likely to continue playing the SNG if they perceive the SNG is with fantasy. Thus, it is proposed that:

H4. Fantasy is associated with SNG continuance intention positively.

In UTAUT2, price value is proposed as a direct key determinant of behavioral intention and is conceptualized as consumers' cognitive trade-off between the perceived benefits of the applications and the monetary cost for using them [38]. SNGs are usually provided for free to register and basic play. However, players have to pay for fancy decorations and powerful equipment, or if they want to achieve higher game levels quickly. Hence, SNG players are also consumers and will be affected by price value. Therefore, we follow the trend of UTAUT2, and assume that:

H5. Price value is associated with SNG continuance intention positively.

Ajzen and Fisherbein [1] claimed that habit was a driver of continuance intention and explained the effect of habit on behavioral intention from the instant activation perspective (IAP). IAP suggests that the relationship from habit to behavioral intention is equivalent to and is an expedited form of conscious processing theory. The formed habit triggered by the attitude objects or environmental cues can activate the behavioral intention which is well-established and restored. Venkatesh et al. [38] supported this argument and verified the significant impact of users' habit on behavioral intention. Hence, it is postulated that:

H6. Habit is associated with SNG continuance intention positively.

In the research model, the postulations of moderating effects in UTAUT2 are followed and examined. The paths from social influence perceived enjoyment, fantasy, achievement and habit to behavioral intention are hypothesized to be moderated by age, gender and experience. The path from price value to continuance intention is postulated to be moderated by age and gender.

III. RESEARCH METHOD

A. Instrument Development

The study employed survey as the research method for gathering empirical data. Each construct in the research model was measured with multiple items adapted from extant literatures to improve the content validity [35]. Items were slightly modified according to the research context. Each item was measured with a five-point Likert scale, ranging from disagree (1) to agree (5).

 ISBN: 978-1-61208-320-9

The research model includes seven constructs. Items measuring social influence, price value and habit are adopted from Venkatesh et al. [38]. Continuance intention is measured by two items adopted from Lee [29]. The 4 items developed by Wu et al. [41] were employed to measure achievement. Fantasy (FA) was measured by three items adapted from the work of Sherry and Lucas [34]. Three items developed by Ghani and Deshpande [16] were used to measure perceived enjoyment.

The questionnaire was developed in English, and then translated to Chinese by one of the researchers in the research project, who is a native Chinese speaker. Then, the questionnaire was sent to 7 participants for pilot study. The participants were consisted of 3 IS researchers, a manager of the SNG provider and 3 current players of the SNG. Some phrases and words were revised according to the feedback from the respondents. The clarity and the overall quality of the questionnaire were improved.

B. Data Collection

Data was collected via a web-based survey from the current SNG players of one popular SNG in China. The SNG is offered by one of the biggest Chinese social network service providers which have multi-million users. Before the data collection, the SNG has been running for 6 months and is distributed via the social network sites.

With the help of the company, the questionnaire was distributed to registered players of the SNG. The survey aimed at studying individual SNG players' continued usage and switching behavior among SNGs respectively. This study attempted to explore the continued usage behavior among SNG players. This study identified the most potential continuous players by asking them whether they have been playing the SNG in the recent one month.

220,000 invitations for answering the questionnaire were sent out to a random sample from registered players of the SNG from Nov. 23rd to 27th, 2012. No rewards were offered to the respondents for answering the questionnaires. All respondents provided their respondents voluntarily. As a result, 7769 respondents were collected including continuous players, switching players and discontinuous players. 3919 valid responses were from continuous players. In the survey on Chinese online game players conducted by iResearch (2012), 67.8 per cent of online game players in China are male and 32.2 per cent are female. 37 per cent of players are below 18 years old, 63per cent of players are above 18 years old [25]. From the demographic information of the respondents presented in Table 1, it can be seen that the sample largely fits to the online game users in China.

TABLE 1. DEMOGRAPHIC INFORMATION OF RESPONDENTS

Measure	Items	Frequency	%
Gender	Male	2357	60.1
	Female	1562	39.9
Age	Adolescence	1083	27.6
	Adult (over 18 years old)	2836	72.4
Experience of playing the SNG	Less than 1 month	1860	47.5
	1-3months	1177	30.0
	3-6months	881	22.5

C. Data Analysis

A two-step approach suggested by Anderson and Gerbing [3] was adopted to analyze the empirical data. This study first analyzed the measurement model to examine the reliability and validity of the instruments, and then tested the structural model to investigate the research hypotheses.

Amos 20. was employed to conduct confirmatory factor analysis (CFA) to examine the measurement model including convergent validity and discriminant validity. Several common used model-fit indices were adopted to estimate the measurement model. All indices exceed the acceptance level (>0.9): GFI=0.956, AGFI=0.939, IFI= 0.981, NFI=0.990, CFI=0.981, TLI=0.977 and RMSEA=0.050 [9]. $\chi 2/df$ is not considered, because the value is very sensitive to sample size, and current study has a very large sample size.

Convergent validity and discriminant validity is presented in Table 2 and Table 3.

TABLE 2. RELIABITLITY AND CONVERGENT VALIDITY STATISTICS

Construct (no. of items)	α	Composite reliability	Minim. factor loading	AVE
SI(3)	0.97	0.97	0.93	0.91
AC(4)	0.94	0.94	0.81	0.81
PE (3)	0.93	0.93	0.88	0.82
FA(3)	0.85	0.86	0.73	0.67
PV(3)	0.93	0.93	0.86	0.82
HA(3)	0.94	0.94	0.88	0.84
CI (2)	0.95	0.95	0.95	0.91

TABLE 3. DISCRIMINANT VALIDITY

Construct	SI	AC	PE	FA	PV	HA	CI
SI	**0.95**						
AC	0.66	**0.90**					
PE	0.21	0.18	**0.91**				
FA	0.71	0.69	0.26	**0.82**			
PV	0.61	0.51	0.12	0.60	**0.91**		
HA	0.64	0.62	0.23	0.66	0.55	**0.92**	
CI	0.70	0.62	0.35	0.69	0.53	0.66	**0.95**

Convergent validity evaluates whether a particular item is developed to measure the construct which is supposed to be measured. Factor loading, average variance extracted (AVE) [9]; composite reliability (CR) and Cronbach's alpha values are usually used to examine convergent validity [9]. The values of the indices in our model are presented in Table 2. All of the values exceed the acceptance level: factor loadings are all over 0.7, composite reliability are over 0.7, AVE are over 0.5, and Cronbach's alpha are over 0.7. Discriminant validity reflects whether two constructs are statistically distinguished from each other. The results in Table 3 show that discriminant validity is achieved, since the square roots of AVE on the diagonal are higher than the correlations between constructs [9].

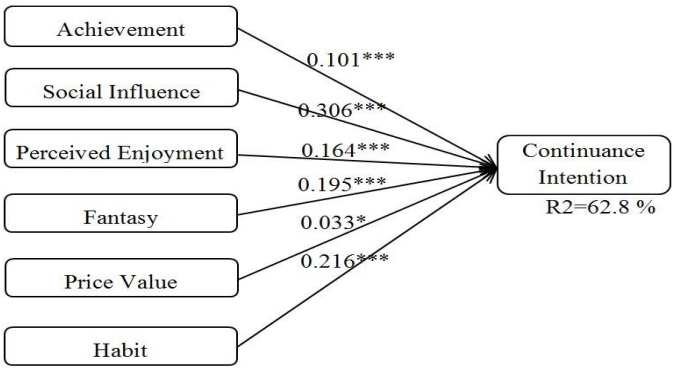
Two tests are conducted to examine common method bias. Harmon's one-factor test is performed to test common method bias. No factor is found to account for the majority of the covariance in the variables [33]. In addition, single factor model test is performed by modeling all items as

 ISBN: 978-1-61208-320-9

indicators of one factor representing common method bias impact. The single-factor model showed a poor fit (GFI = 0.476; AGFI = 0.371; NFI = 0.546; IFI = 0.547; TLI = 0.502; CFI = 0.547; RMSEA = 0.222). With results from two tests, common method bias is not likely to be a significant problem in this study.

IV. RESEARCH RESULTS

The analysis results on the structural model of multi-group model were presented in Figure 2. The model fit indices suggest a good model fit between the data and the research model in current study. The values of indices are presented as following: GFI=0.956, AGFI=0.939, IFI= 0.981, NFI=0.980, CFI=0.981, TLI=0.977 and RMSEA=0.050. Achievement ($\beta = 0.101$, $p < 0.001$), social influence ($\beta = 0.306$, $p < 0.001$), enjoyment ($\beta = 0.164$, $p < 0.001$), fantasy ($\beta = 0.195$, $p < 0.001$), price value ($\beta = 0.033$, $p < 0.050$), habit ($\beta = 0.217$, $p < 0.001$) are positively associated with continuance intention significantly. 62.8 per cent of variance of continuance intention is explained by the research model, which indicates a good explanatory power of the research model [9]. According to the analysis results presented in Table 4, age and user experience are not significant moderators. Gender exerts significant moderating effect on the paths from social influence, perceived enjoyment and price value to continuance intention.



*Note. * p<0.05, **p<0.005, ***p<0.001, n.s: not significant.*

Figue 2. Structure Model Results of Multi-group Model

TABLE 4. STRUCTURE MODEL RESULTS OF MODERATORS

Hypothesis	**Age** (Basic model, $\chi2$= 2105,122, df =336)		**Gender** (Basic model, $\chi2$=2087.445, df =336)		**Experience** (Basic model, $\chi2$=2318.762 , df = 504)		
	Adolescence	***Adult***	***Male***	***Female***	***Less than1 month***	***2-3months***	***3-6months***
AC→CI	β=0.128***	β =0.085***	β =0.085***	β = 0.114***	β =0.087***	β =0.099**	β =0.136***
	$\chi2$= 2106.013, $\Delta\chi2$=0.891, n.s.		$\chi2$=2088.028, $\Delta\chi2$=0.583, n.s.		$\chi2$=2320.316, $\Delta\chi2$=1.554, n.s.		
SI→CI	β =0.347***	β =0.296***	β =0.363***	β =0.230***	β =0.318***	β =0.304***	β =0.288***
	$\chi2$= 2106.957, $\Delta\chi2$=1.835, n.s.		$\chi2$=2100.328, $\Delta\chi2$=12,883, **p<0. 01**		$\chi2$=2319.729, $\Delta\chi2$=0.967, n.s.		
PE→CI	β =0.138***	β =0.178***	β =0.124***	β =0.232***	β =0.163***	β =0.174***	β =0.160***
	$\chi2$= 2106.628, $\Delta\chi2$=1.506, n.s.		$\chi2$=2108,563, $\Delta\chi2$=21.118,**p<0.01**		$\chi2$=2319.134, $\Delta\chi2$=0.372, n.s.		
FA→CI	β =0.213***	β =0.210***	β =0.210***	β =0.171***	β =0.217***	β =0.193***	β =0.167***
	$\chi2$= 2106.289, $\Delta\chi2$=1.167, n.s.		$\chi2$=2088.051, $\Delta\chi2$=0.606, n.s		$\chi2$=2320.432, $\Delta\chi2$=1.67, n.s.		
PV→CI	β =0.029, n.s.	β =0.039*	β =0.10,n.s.	β =0.069**			
	$\chi2$= 2105.161, $\Delta\chi2$=0.039, n.s.		$\chi2$=2091.225, $\Delta\chi2$=3.78, **p<0.05**				
HA→CI	β =0.159***	β =0.230***	β =0.204***	β =0.221***	β =0.187***	β =0.203***	β =0.229***
	$\chi2$=2107.336, $\Delta\chi2$=2.214,n.s.		$\chi2$=2087.490, $\Delta\chi2$=0.045, n.s.		$\chi2$=2320.165, $\Delta\chi2$=1.403, n.s.		
R2(CI)	65.2	61.4	65.6	59.4	64.6	59.1	60.3

V. DISCUSSION

This study aims to test the explanatory power of a research model extended from UTAUT2 in predicting SNG player's continuance intention. As presented in Figure 2, all hypotheses in multi-group model are supported. Factors including achievement, social influence, perceived enjoyment, fantasy, price value, and habit all have significant and direct influences on continuance intention to play the SNG.

In this study, the effect of social influence is stronger than other factors in predicting continuance intention to play the SNG. Findings of prior studies have indicated that user's intention was significantly affected by other important referees' opinions when they made decisions of IS usage [4], [36], [37], [38]. In this study, players mainly play the SNG with real friends/families in their existing social networks. SNG players usually connect with these friends/families both in real life and the virtual SNG world. Hence, the recommendations from important others exert a strong influence on player's continuance intention to play the SNG.

Fantasy (β=0.195) and perceived enjoyment (β=0.164) are found to exert strong influence on continuance intention in this study. The present studies indicate that fantasy affects the intention to continue playing SNGs, because players would like to try out new identities and to be absorbed in the virtual fantasy world [28]. In the current study, players can play the SNG to reflect their own imaginations when they manage and decorate the virtual spaces, avatars and various activities in the SNG. The study sheds light on the importance of fantasy in predicting continuance intention. It implies that players would like to engage in the SNG, if they can continually construct and realize their fantasies which cannot be performed in real life.

The finding of perceived enjoyment in this study concurs with the arguments that perceived enjoyment is an important determinant of behavioral intention in the context of hedonic settings [19], [37]. Since the players mainly play the SNG to

 ISBN: 978-1-61208-320-9

obtain the entertainment, this result suggests that the players are not likely to continue playing the SNG if they do not enjoy it [29].

One interesting result emerging from the findings is that habit is the second important driver of continuance intention. Prior studies reported that "habitual previous preferences to use a specific IS directly and strongly increase user intentions to continue using the same IS again" [15], [38]. The result implies that player's decisions on whether they should continue playing the SNG is based on both their perceptions of the desirable outcomes of playing the SNG and their habit of playing the SNG. Players are more intended to play the SNG when playing the SNG becomes habitual to them.

The result towards the influence of achievement is in accord with the findings from prior studies. Present studies report that achievement exerts significant influence on continuance intention in the context of online gaming [40], [42]. It can be inferred that players would like to continue playing the SNG if they can obtain the sense of achievement by participating various kinds of activities in the SNG, such as acquiring superior power, and defeating other players. In this study, achievement exerts a relatively less effect on continuance intention to play the SNG. It implies that getting the sense of achievement might not be the premier target for players to play the SNG. Instead, players might engage in the SNG for other reasons, such as for realizing the fantasy, and experiencing the enjoyment during the process.

The study provides interesting findings on how individual characteristics (e.g., age, gender and experience) moderate the effects of achievement, social influence, perceived enjoyment, fantasy, price value, and habit on continuance intention. User experience is not a significant moderator according to the test results presented in Table 4. Among the groups with different use experience, there is no significant difference on the effect of their perceptions (including achievement, social influence, perceived enjoyment, fantasy and habit) on continuance intention. In other words, the players who started playing the SNG earlier and the players who started playing the SNG later do not have different perceptions on continuance intention to play the SNG. Since a player with longer playing history does not necessarily suggest that the player is a heavy user who plays the SNG frequently.

In addition, no statistically significant differences between different age groups are found according to analysis results shown in Table 4. The result is consisted with the work of Lee [29] which reports that no paths are significantly moderated by age in online gaming. It seems that in the current study players with different ages, no matter they are adolescences or adults, do not have different perceptions towards the effects exerted by factors (including achievement, social influence, perceived enjoyment, fantasy, price value and habit) on continuance intention to play the SNG in the post-adoption stage.

Gender plays moderating role on the paths from social influence, perceived enjoyment and price value to continuance intention according to the analysis result presented in Table 4. The finding indicates that the path coefficient from social influence to the continuance intention for males was significantly larger than that for females. This finding suggested that the effect of social influence on the intention to play online games is stronger for males than females. It can be inferred that male players are more likely to be influenced by the most important people around them when they make decisions on continuing or not continuing using the SNG. It is probably because males are usually more interested in playing online games than females do. Hence, male players concern the information towards online games more than females do, including recommendations from important others.

Moreover, the effect of enjoyment of playing SNG on continuance intention is stronger for females than males. It can be inferred that female players concern more about the entertainment obtained in playing SNG compared to male players. It might be due to the design of less violent and less competitive features of the SNG. These features enable females to have pleasant experience from the process of playing. Finally, as we expected, female players concerns more about the price value compared to male players when they make decision on continuing playing SNG. The result is consistent with prior findings which suggest that women are likely to pay more attention to the prices of services and will be more cost conscious than man [38].

VI. IMPLICATIONS FOR THEORIES AND PRACTICES

The study provides some implications for both theories and practitioners.

From a theoretical perspective, in the prior literature, little research has explored continuance intention in the online game settings based on UTAUT2, especially SNG [6]. This study filled the gap by developing the research model based on UTAUT2 and identified the key determinants of continuance intention in SNG. Furthermore, this study contributes to a theoretical understanding of the explanatory power of the extended model based on UTAUT2. By explaining a relatively high proportion of variance in the continuance intention, this study suggests that the tailored UTAUT2 is suitable for investigating continuance intention in SNG.

Through the examination of the research model, this study highlights the important factors in influencing continuance intention to play the SNG, namely social influence, habit and hedonic motivations (e.g., fantasy and perceived enjoyment), followed by achievement and price value. The importance of fantasy and achievement in predicting SNG continuance intention offers new insights into explaining the utilitarian and hedonic motivations in hedonic IS research, especially in SNG games. The analysis results on the moderators of age, gender, and use experience reveals that individual characteristics of online game players can still be the moderators, and its moderating effect, such as age and user experience are diminishing in the online gaming context.

From a practical perspective, this study emphasizes the strong impact of social influence to continuance intention. Thus, SNG providers should try to use the networks of SNG

 ISBN: 978-1-61208-320-9

players to facilitate players' continuance behavior via different online communication channels, such as the popular social network sites, Renren, Sina Microblog, and QQ.

The importance of habit in predicting SNG continuance intention suggests that SNG providers should raise some strategies to help the development of SNG players' habit, such as offering players rewards for repeated and prolonged usage to foster the habit.

The finding on fantasy in predicting continuance intention suggests that SNG providers should offer more fancy themes, diverse imaginary identities and activities, and novel virtual worlds in their SNG design in order to retain their SNG players. Meanwhile, the SNG providers can strengthen player's sense of achievement by providing more opportunities for players to gain more in-game wealth, compete with other players, and achieve higher game levels.

Finally, the moderator test findings in age (a moderator) and gender (not a moderator) suggest that the SNG providers should try to balance the preferences of both male and female in their SNG design, but not the user age yet. The finding that user experience is not a moderator offers the SNG providers suggestion that they should focus more on the heavy SNG players who play the SNG quite often, but not those with long time use experience.

VII. CONCLUSION AND LIMITATION

The main purpose of this study was to investigate the determinants of a player's continuance intention in online gaming. By applying and tailoring UTAUT2 to study continuance intention in online gaming, we found UTAUT2 to be a useful theoretical model in our context. Thus, the explanatory power of UTAUT2 is expanded in the new research context, since all the constructs in the research model are statistically significant. Further, comparing to age and experience, gender exerts more significant moderating effects. These unexpected results also contribute to a better understanding of a player's continuance intention, and provide practical suggestions to online game service providers.

This study is subject to some limitations. Firstly, we conducted the research in China which has different culture from other countries. The examination of the results in other countries may provide richer insight in understanding continued use of SNG. Secondly, the research setting in current study is SNG as one form of online games. This study need to be replicated in other types of online games. Finally, we only examined the moderating role of age, gender and experience. Studies investigating other moderators (e.g., education level, income level, social status) may provide more understandings on continued usage in online games

REFERENCES

[1] I. Ajzen and M. Fishbein, "Attitudes and the Attitude-Behavior Relation: Reasoned and Automatic Processes," European Review of Social Psychology, vol. 11, Jan. 2000, pp. 1–33, doi: abs/10.1080/14792779943000116.

[2] I. Ajzen, "From Intentions to Actions: A Theory of Planned Behavior, Action-Control: From Cognition to Behavior," J. Kuhl and J. Beckmann, Eds. Berlin: Springer-Verlag, 1985, pp. 11-39, doi: 10.1007/978-3-642-69746-32.

[3] J. C. Anderson and D. W. Gerbing, "Structural Equation Modeling in Practice: A Review and Recommended Two-step Approach," Psychological Bulletin, vol. 103, May. 1988, pp. 411–423, doi: 10.1037/0033-2909.103.3.411.

[4] R. Bagozzi and K. Lee, "Multiple Routes for Social Influence: The Role of Compliance, Internalization, and Social Identity," Social Psychology Quarterly, vol. 65, Sep. 2002, pp. 226–247, doi: dx.doi.org/10.2307/3090121.

[5] E. Boyle , T. M. Connolly, T. Hainey, and J.M. Boyle, "Engagement in Digital Entertainment Games: A Systematic Review," Computers in Human Behavior, vol. 28, May. 2012, pp. 771–780, doi: 1016/j.chb.2011.11.020.

[6] C. C. Chang, "Examining Users′ Intention to Continue Using Social Network Games: A Flow Experience Perspective," Telematics and Informatics, vol. 30, Nov. 2013, pp. 311 - 321. doi: 10.1016/j.tele.2012.10.006.

[7] J. H. Chang and H. Zhang, "Analyzing Online Game Players: From Materialism and Motivation to Attitude," Cyberpsychology & Behavior, vol. 11, Dec. 2008, pp. 711–714. doi: 10.1089/cpb.2007.0147.

[8] C. Chou and M. J. Tsai, "Gender Differences in Taiwan High School Students' Computer Game Playing," Computers in Human Behavior, vol. 23, Jan. 2007, pp. 812–824, doi: 10.1016/j.chb.2004.11.011.

[9] C. Fornell and D. F. Larcke, "Evaluating Structural Equation Models with Unobservable Variables and Measurement Error," Journal of Marketing Research, vol. 18, Feb. 1981, pp. 39–50, dio: dx.doi.org/10.2307/3151312.

[10] J. Colwell, "Needs Met Through Computer Game Play Among Adolescents," Personality and Individual Differences, vol. 43, Dec. 2007, pp. 2072–2082, doi: 10.1016/j.paid.2007.06.021.

[11] F. Davis, "Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology". MIS Quarterly, vol. 13, Sep. 1989, pp. 319–340, retreived from: http://www.jstor.org/stable/10.2307/249008.

[12] F. D. Davis, R. P. Bagozzi, and P. R. Warshaw, "Extrinsic and Intrinsic Motivation to Use Computers in the Workplace," Journal of Applied Social Psychology, vol. 22, Jul. 1992, pp. 1111–1132, doi: 10.1111/j.1559-1816.1992.tb00945.x.

[13] Facebook Reports Second Quarter, "Facebook Reports Second Quarter 2013 Results". 2013. [retrieved: Sep. 2013] http://investor.fb.com/releasedetail.cfm?ReleaseID=780093.

[14] M. Fishbein, I. Ajzen, "Belief, Attitude, Intention, and Behaviour: An Introduction to Theory and Research Reading," Mass Don Mills, Ontario: Addison-Wesley Pub. Co, 1975.

[15] D. Gefen, "TAM or Just Plain Habit," Advanced Topics in End User Computing, vol. 3, Jul. 2004, pp. 1–13, doi: 10.4018/978-1-59140-257-2.ch001.

[16] J. Ghani and S. Deshpande, "Task Characteristics and the Experience of Optimal Flow in Human—Computer Interaction," The Journal of Psychology, Vol. 12, July. 1994, pp. 1143–1168, doi: abs/10.1080/00223980.1994.9712742.

[17] I. Ha, Y. Yoon, and M. Choi, "Determinants of Adoption of Mobile Games under Mobile Broadband Wireless Access Environment," Information & Management, vol. 44, Apr. 2007, pp. 276–286, doi: 10.1016/j.im.2007.01.001.

[18] G. Hackbarth, V. Grover, and M. Y. Yi, "Computer Playfulness and Anxiety: Positive and Negative Mediators of the System Experience Efect on Perceived Ease of Use," Information & Management, Vol. 40, Jan. 2003, pp. 221–232, doi: 10.1016/S0378-7206(02)00006-X.

[19] H. Heijden Van der, "User Acceptance of Hedonic Information Systems," MIS Quarterly, vol. 28, Dec. 2004, pp. 695–704.

 ISBN: 978-1-61208-320-9

[20] M. Holbrook and E. Hirschman, "The Experiential Apects of Consumption: Consumer Fantasies, Feelings, and Fun," Journal of Consumer Research, vol. 9, Sep. 1982, pp. 132–140.

[21] J. H. Hu, Y. K. Chau, O. Sheng, and K.Y. Tam, "Examining the Technology Acceptance Model Using Physician Acceptance of Telemedicine Technology," Journal of Management Information Systems, vol. 16, Fall, 1999, pp. 91-112.

[22] H. Li, Y. Liu, X. Xu, and J. Heikkilä, "Please Stay with Me! An Empirical Investigation of Hedonic IS Continuance Model for Social Network Games," International Confernce on Information Systems (ICIS) 2013, in Press.

[23] C. L. Hsu and H. P. Lu, "Why Do People Play Online Games? An Extended TAM with Social Influences and Flow Experience," Information & Management, vol. 41, Sep. 2004, pp. 853–868, doi: 10.1016/j.im.2003.08.014.

[24] C. L. Hsu and H. P. Lu, "Consumer Behavior in Online Game Communities: A Motivational Factor Perspective," Computers in Human Behavior, vol. 23, May. 2007, pp. 1642–1659, doi: 10.1016/j.chb.2005.09.001.

[25] iResearch, "China Online Game Players' Behaviour Research, 2011-2012," 2012.[retrieved: May, 2013]. http://wenku.baidu.com/view/372853200722192e4536f642.html.

[26] J. Jansz, C. Avis, and M. Vosmeer, "Playing The Sims2: An Exploration of Gender Differences in Players' Motivations and Patterns of Play," New Media & Society, vol, 12, Jan. 2010, pp. 235–251, doi: 10.1177/1461444809342267.

[27] A. Järvinen, "Game Design for Social Networks," Proceedings of the 2009 ACM SIGGRAPH Symposium on Video Games - Sandbox '09, ACM Press, 2009, pp95-102, doi: 10.1145/1581073.1581088.

[28] J. Lee, M. Lee, and I. H. Choi, "Social Network Games Uncovered: Motivations and Their Attitudinal and Behavioral Outcomes," Cyberpsychology, Behavior and Social Networking, vol. 15, Dec. 2012, pp. 643–648, doi: 10.1089/cyber.2012.0093.

[29] M. C. Lee, "Understanding the Behavioural Intention to Play Online Games: An Extension of the Theory of Planned Behaviour," Online Information Review, vol. 33, Jul. 2009, pp. 849–872, doi: 10.1108/14684520911001873.

[30] Y. H. Lee and D. Y. Wohn, "Are There Cultural Differences in How We Play? Examining Cultural Effects on Playing Social Network Games," Computers in Human Behavior, vol. 28, Jul. 2012, pp. 1307–1314, doi: 10.1016/j.chb.2012.02.014.

[31] W. K. Lin, C. K. Chiu, and Y. H. Tsai, "Modeling Relationship Quality and Consumer Loyalty in Virtual Communities," Cyberpsychology & Behavior : the Impact of the Internet, Multimedia and Virtual Reality on Behavior and Society, vol. 11, Oct. 2008, pp. 561–564, doi: 10.1089/cpb.2007.0173.

[32] K. Lucas, J. L. Sherry, "Sex Differences in Video Game Play: A Communication-Based Explanation," Communication Research, vol. 31, Oct. 2004, pp. 499–523, doi: 10.1177/0093650204267930.

[33] P. M. Podsakoff, S. B. MacKenzie, J. Y. Lee, and N. P. Podsakoff, "Common Method Biases in Behavioral Research: a Critical Review of the Literature and Recommended Remedies," The Journal of Applied Psychology, vol. 88, Oct. 2003, pp. 879–903, doi: 10.1037/0021-9010.88.5.879.

[34] J. Sherry and K. Lucas, "Video Game Uses and Gratifications as Predictors of Use and Game Preference," in Playing Video Games: Motives, Responses, and Consequences, P. Vorderer, and J. Bryant, Eds. 2006, pp.213–224.

[35] D. W. Straub and M. C. Gefen, " Validation guidelines for IS positivist research," Communications of the Association of Information Systems, vol.13, 2004, pp.380-427.

[36] M. Suznjevic and M. Matijasevic, "Why MMORPG Players Do What They Do: Relating Motivations to Action Categories, " International Journal of Advanced Media and Communication, Nov. 2010, pp.1–20, doi: 10.1504/IJAMC.2010.036838.

[37] V. Venkatesh, M. Morris, G. Davis, and F. Davis, "User Acceptance of Information Technology: Toward a Unified View," MIS Quarterly, vol. 27, Sep. 2003, pp.425–478.

[38] V. Venkatesh, J. Thong, and X. Xu, "Consumer Acceptance and Use of Information Technology: Extending the Unified Theory of Acceptance and Use of Technology," MIS Quarterly, vol. 36, Mar. 2012, pp.157–178.

[39] D. Y. Wohn and Y. H. Lee, "Players of Facebook Games and How They Play," Entertainment Computing, vol. 4, Aug. 2013, pp.171–178, doi: 10.1016/j.entcom.2013.05.002.

[40] J. Wu and D. Liu, "The Effects of Trust and Enjoyment on Intention to Play Online Games, " Journal of Electronic Commerce Research, vol. 8, 2007, pp.128–140.

[41] J. Wu, S. Wang, and H. Tsai, "Falling in Love with Online Games: The Uses and Gratifications Perspective," Computers in Human Behavior, vol. 26, Nov. 2010, pp. 1862–1871, doi: 10.1016/j.chb.2010.07.033.

[42] C. Xu, S. Ryan, V. Prybutok, and C. Wen, "It Is Not for Fun: An Examination of Social Network Site Usage," Information & Management, vol. 49, Jul. 2012, pp. 210–217, doi: 10.1016/j.im.2012.05.001.

[43] N. Yee, "Motivations For Play in Online Games," Cyberpsychology & Behavior : the Impact of the Internet, Multimedia and Virtual Reality on Behavior and Society, vol. 9, Dec. 2006, pp. 772–775, doi: 10.1089/cpb.2006.9.772.

 ISBN: 978-1-61208-320-9

Kinect Skeleton Coordinate Calibration for Remote Physical Training

Tao Wei, Yuansong Qiao, Brian Lee
Software Research Institute, Athlone Institute of Technology
Athlone, Ireland
{twei, ysqiao}@research.ait.ie, blee@ait.ie

Abstract—**With the advent of the Microsoft Kinect sensor, skeleton coordinate systems have become an active part of interactive multimedia applications. The skeleton coordinate data captured by the Kinect sensor can be used to compare the similarity of remote users' motions in remote training systems. However, this approach is limited in that the remote users' initial positions have to be at the same position and face the sensor with the same angle. This paper proposes a Kinect Skeleton Coordinate Calibration (KSCC) algorithm to calibrate the remote user's arbitrary initial positions, thereby removing the above limitations on the initial positions and angles. It collects the remote user's initial position data, and calculates the initial centre coordinate of the user and initial angle between user and Kinect sensor. After the collection and calculation, all skeleton coordinates are transformed to a universal coordinate system according to the initial centre coordinate and rotated by a quaternion rotation. An evaluation test has been performed to assess the accuracy and limiting fields of the system. The results show that our approach is able to calibrate the Kinect skeleton coordinates with a high accuracy, with the requirements that the user's initial positions only need to be in the detection zone of the Kinect sensor.**

Keywords - Kinect skeleton coordinate; calibration; remote physical training; interactive multimedia

I. INTRODUCTION

With advances in human body recognition technologies, some 3D sensors have been able to capture and analyse the human body without marker-based systems, which require the users to wear obtrusive devices. One such device is the Microsoft Kinect [1], a marker-less motion capturing sensor, which can track a user skeleton and capture data at a rate of 30 frames per second using the Microsoft Kinect for Windows SDK [1]. Each such tracked skeleton contains twenty joints' 3D coordinates [2].

Our goal is to compare the similarity of remote users' motions. Various approaches [3][4][5] have been proposed to use the skeleton coordinates captured by Kinect sensor to do body motion comparison. In these approaches, they compare two skeletons with recorded data. These recorded data have several limitations, such as the initial positions and angles of the users are the same; the length and the content of the recorded data are constant. It is easy to control their experiments using the recorded data. However, for a practical remote body motion comparison system, the users may not act exactly according to the recorded data which are used in experimental environment. The users may stand at different positions relative to Kinect sensor, i.e., different angles facing the Kinect and/or different distances from the Kinect. Users may thus get very different coordinate data for the skeleton position, even if they do the same motion. It is difficult to compare two user-motions using these data.

This paper proposes a KSCC algorithm. The main features of this approach are to "pull" the user's skeleton to the centre of the Kinect sensor and then rotate it to face the Kinect sensor. KSCC treats this position as the initial position in a universal coordinate system. In this way all users' initial positions are normalized in the universal coordinate system, irrespective of their original standing position and angle. After reconstruction in the universal coordinate system, all movements of the skeleton are referenced in the universal coordinate system. In this approach, the user needs to stand motionless for about four seconds to enable the KSCC system to collect the data of the user's initial position. This data are then used to calculate the initial angle between user and Kinect sensor (let us call it 'initial angle') and the initial centre coordinate of the user's skeleton. Then KSCC transforms the twenty joints' coordinates of the user's skeleton according to the initial centre coordinate, and rotates them about the initial centre point's y-axis by the initial angle to a universal coordinate system. Consequently, all the skeletons are transformed to the same location with the same angle relative to the sensor. As a result, when different users do the same motions, they can get the same coordinate data for their skeletons, even if their initial positions are different.

The following experiment is designed to test the KSCC algorithm. As it is difficult for two people to do exactly the same motion at the same time, we use one person as experimenter. We setup two Kinect sensors to capture the user's skeleton separately, and then run two calibration systems to detect the person at the same time. The two skeleton images captured by the two calibration systems are drawn in the same canvas. It is intuitive to compare the calibrated results. Next, the coordinate data of left and right shoulders are recorded to find out the two calibration systems' differences of value. The experimental results show that the skeleton coordinate data are accurately calibrated by the KSCC algorithm and the differences of the value are less than 4cm, given that both initial angles of the skeleton are all less than 20°.

The rest of the paper is organized as following. In Section 2, the related work is presented and compared. The Kinect system environment and details of the KSCC algorithm are described in Section 3. Section 4 presents the experiments that have been made to evaluate the KSCC algorithm. Section 5 is dedicated to the conclusion and the future work.

 ISBN: 978-1-61208-320-9

II. RELATED WORK

Multimedia devices have been widely applied in remote physical training. Huang et al. [6] present a Multiple-video-based E-learning Platform for Physical Education. Their system records videos and voices in three different angles synchronously. Then, the user reviews the records to teach or learn the sports actions. Li et al. [7] propose a tennis e-learning system. In that system, a Nintendo Wii Remote is used as the input device to capture the motion of a tennis swing. The system is aimed at differentiating different types of swings. Muller et al. [8] propose a pre-processing method substantially accelerating the cost-intensive classical dynamic time warping techniques for the time alignment of logically similar motion data streams.

With the advent of the Microsoft Kinect sensor, a lot of attention has been focused on skeleton coordinate system. Tamura et al. [9] propose a three-dimensional motion capture and feedback system for flying disc throwing action learners. Their system captures learners' body movement, checks their skeleton positions in pre-motion/motion/post-motion in several ways, and displays feedback messages to refine their actions. However, they set presupposed motion in the system. It only supports throwing motion comparison. Essid et al. [3][4] propose a virtual dance performance evaluator based on 17 skeletal joints positions. It can be used to evaluate a student's performance and provide him/her with meaningful feedback to aid improvement. In their system, Kinect sensors are used to acquire a "choreography" dance rating. Three choreography scores are calculated by considering the modulus of the Quaternion Correlation Coefficient for each pair of joint position signals. The 3D coordinates of each joint are used to be input data directly. The angular skeleton representation of Raptis et al. [5] is a good method to remove dependence on Kinect position. They treat the torso as a vertically elongated rigid body. Their approach is to fit the full torso with a single frame of reference, and to use this frame to parameterize the orientation estimates of both the first-degree and second-degree limb joints. However, these approaches compare experimenter's skeleton coordinates with recorded data, not with the real people.

Due to different users' skeleton coordinates belonging to different Kinect skeleton coordinate systems, finding the relationship between the two coordinate systems is useful. A closed-form solution [10][11] is to calibrate a number of points' coordinates in two different Cartesian coordinate systems. Their approach transforms the points from one coordinate system to another using a 4 × 4 transformation matrix. However, their solution is used to solve local multiple cameras fusion problems, i.e., the multiple cameras must shoot the same object, and the system needs to know the points' coordinates data from all cameras system before calibration.

In this paper, a novel calibration algorithm KSCC that calibrates all remote users' skeleton coordinates into a universal coordinate system is presented. Unlike [3][4][5][9] the KSCC algorithm is used for a practical remote body motion comparison system rather than recorded data or presupposed motions. It reduces the limitation of the users' initial positions and angles which are constant in recorded data. Moreover, unlike the closed-form solution [10][11], the KSCC algorithm does not transform the skeleton coordinate from one skeleton to another. The remote calibration systems of both users do not need to know each other's skeleton coordinates data before calibration. As a result, the remote calibrated skeletons can be compared in a universal coordinate system for the practical remote body motion comparison system.

III. SKELETON CALIBRATION

A. Kinect Skeleton Coordinate System

As shown in Fig. 1, the Kinect sensor has a practical ranging limit of 82cm~400cm [2]. The Kinect sensor also can maintain tracking through an extended range of approximately 70cm~600cm in the context of ignoring some accuracy. The field of view of the sensor is pyramid shaped. It has an angular field of view of 57° horizontally and 43° vertically, while the motorized pivot is capable of tilting the sensor up to 27° either up or down [12].

Skeleton data contain 3D position data for human skeletons. Each joint position in the skeleton space is represented as (x, y, z). The skeleton space coordinates are expressed in meters. As illustrated in Fig. 2, it is a right-hand coordinate system that places a Kinect sensor at the origin. More specifically, the positive x-axis extends to the left of the Kinect, and the positive y-axis extends upward. The positive z-axis is extending in the direction in which the Kinect is looking at.

It assumes that the distance between the Kinect sensor and floor is 100cm. And the surface on which the Kinect sensor is placed parallels the floor. Also, the Kinect sensor is not tilted, i.e., the z-axis of the skeleton coordinate system also parallels the floor.

In the context of detecting the whole body of the user, the available active area is an isosceles trapezoid area. The minimum distance D_{min} between Kinect sensor and the user can be calculated by

$$D_{min} = h / \tan (\theta_v / 2). \quad (1)$$

where h is the distance between Kinect sensor and floor. h equals 100cm. θ_v is the vertical angular field of view. θ_v equals 43°.

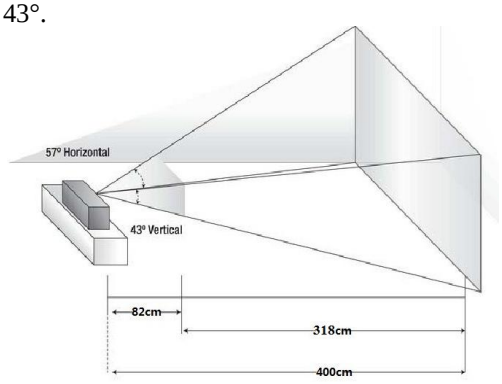


Figure 1. Detecting range of Kinect [2]

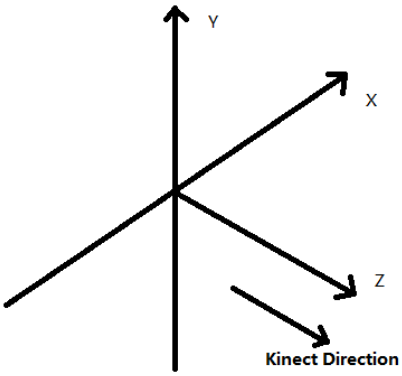


Figure 2. Kinect skeleton space

The minimum distance D_{min} equals 254cm, and the maximum D_{max} equals 400cm which is limited by the Kinect sensor.

The two bases of the isosceles trapezoid area can be calculated by

$$B = D \times \tan (\theta_h / 2). \quad (2)$$

where θ_h is the horizontal angular field of view, θ_h equals 57°. The short base is calculated when D equals D_{min}; the long base is calculated when D equals D_{max}.

Finally, the available active area is a isosceles trapezoid whose height is 146cm ($D_{max} - D_{min}$), and two bases are 138cm and 217cm respectively.

B. Initial Data Collection and Calculation

In order to calibrate the initial position of the user, the system collects the first 120 frames as initial data. These initial data are used to calculate the initial angle and the initial centre coordinate of the user. Since the Kinect products about 30 frames data per second [2], it suggests that the user remains standing still around four seconds.

1) Initial Angle between User and Kinect Sensor

KSCC assumes that all joints of the user are in the same plane when the user stands straight. Also, the line between left shoulder and right shoulder is treated as the horizontal line in the user's body plane. According to the initial data, it obtains the average coordinate values of left and right shoulders as:

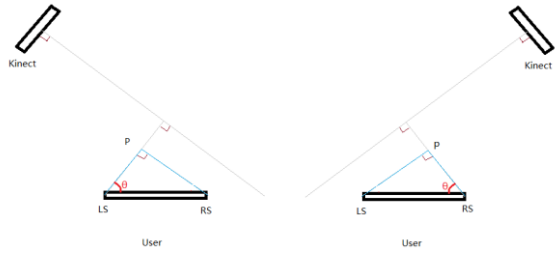
$$LS = (X_l, Y_l, Z_l) \quad (3)$$
$$RS = (X_r, Y_r, Z_r) \quad (4)$$

The initial angle has three situations:

1. The user's body plane is perpendicular to the z-axis direction of the skeleton coordinate system.

2. The user's body plane faces right direction (Fig. 3-(a)).

3. The user's body plane faces left direction (Fig. 3-(b)).



(a) (b)

Figure 3. (a) The left shoulder is closer to the Kinect; (b) the right shoulder is closer to the Kinect.

In the first situation, as the initial angle θ is 0, i.e., the user parallels the Kinect, it is not necessary to consider the angle problem. The other two situations are shown in Fig. 3. First, when the body plane faces right, the z-axis value of left shoulder (LS) is less than the right shoulder (RS). Second, when the body plane faces left direction, the z-axis value of LS is larger than the RS.

The lengths of (LS, P) and (RS, P) in the right triangles can be calculated by

$$D = Z_r - Z_l \quad (5)$$

where D is (RS, P) in Fig. 3-(a); D is (LS, P) in Fig. 3-(b).

$$W = X_r - X_l \quad (6)$$

where W is (LS, P) in Fig. 3-(a); W is (RS, P) in Fig. 3-(b).

Then, the initial angle θ is:

$$\theta = \text{Atan} (D / W) \quad (7)$$

where θ is positive in the situation 2, is negative in the situation 3, and equals 0 in the situation 1.

2) Initial Centre of User's Skeleton

In order to get the initial centre coordinate of user's skeleton, the sum of all joints coordinates in one frame is calculated by

$$\vec{S}(X, Y, Z) = \sum_j \vec{J}(X, Y, Z), \quad j = 0,\ldots, 19 \quad (8)$$

where, j is the index of joints in a skeleton.

Then, the average of the twenty joints' coordinates is treated as the centre of the skeleton in one frame:

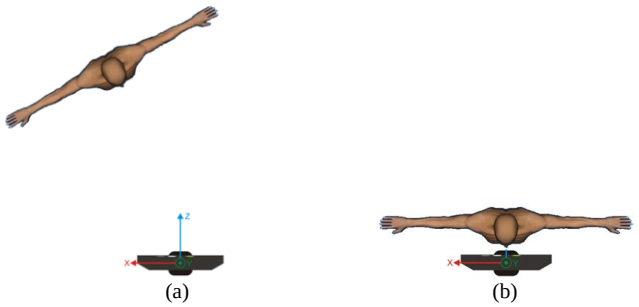
$$\vec{A}(X, Y, Z) = \vec{S}(X, Y, Z) / 20 \quad (9)$$

Finally, the initial centre coordinate of the user's skeleton in the period time can be calculated by the average of the 120 centre coordinates:

$$\vec{C}(X_C, Y_C, Z_C) = \{\sum_t \vec{A}(X, Y, Z)\} / T, \quad t = 1,\ldots,T \quad (10)$$

where T is the total frames in the period, here T equals 120.

 ISBN: 978-1-61208-320-9



(a) (b)

Figure 4. (a) User's skeleton position before calibration; (b) user's skeleton position after calibration.

C. Transform and Rotation

After the collection and calculation of the initial position data, the initial angle θ and the initial centre coordinate vector $C(X_C, Y_C, Z_C)$ are obtained. These two results are the foundation of the calibration, and will be utilized all the time, unless the Kinect sensor is moved or the system is restarted. Then, any joint coordinates can be transformed and rotated to a universal coordinate system that places the initial centre at the origin.

As illustrated in Fig. 4, the calibration process can be regarded as pulling the original skeleton to the centre of the Kinect sensor, and rotating it to face the Kinect sensor. This is the initial position in a universal coordinate system. Thus, all users' initial positions are the same in the universal coordinate system, wherever they stand at. After reconstruction in the universal coordinate system, all movements of the skeleton are in the universal coordinate system.

Firstly, all joints are transformed to the origin of the universal coordinate system according to the initial centre coordinate:

$$\vec{P}_j(X_P, Y_P, Z_P) = (X_j - X_C, Y_j - Y_C, Z_j - Z_C,), j = 0,\ldots,19 \quad (11)$$

where X_j, Y_j, Z_j are coordinates of joint j; X_C, Y_C, Z_C are coordinates of the initial centre.

Secondly, a quaternion rotation [13] is used to rotate the coordinate vector $P_j(X_P, Y_P, Z_P)$ about the y-axis of the initial centre by the initial angle θ. The quaternion rotation is a right handed rotation. The thumb points the direction of unit rotation axis vector R which is the y-axis of the initial centre.

$$\vec{R} = (X_R, Y_R, Z_R) \quad (12)$$

where $\| \vec{R} \| = 1$. Thus, $X_R = 0$, $Y_R = 1$ and $Z_R = 0$.

The rotation quaternion [13] is defined to be:

$$Q = \cos(\frac{\theta}{2}) + X_R \sin(\frac{\theta}{2})i + Y_R \sin(\frac{\theta}{2})j + Z_R \sin(\frac{\theta}{2})k \quad (13)$$

The point $\vec{P}_j(X_P, Y_P, Z_P)$ is viewed as a quaternion without scalar part [14]:

$$Q_P = 0 + X_P i + Y_P j + Z_P k \quad (14)$$

Then, to rotate Q_P about the axis $\vec{R}$ by the angle θ, the quaternion rotation function [13] is defined to be:

$$Q_{PR} = Q \times Q_P \times Q^{-1} \quad (15)$$

$$Q^{-1} = \frac{Q^*}{Q \cdot Q} \quad (16)$$

where Q^{-1} is the reciprocal [14] of the rotation quaternion Q, Q* is the conjugation [14] of the rotation quaternion Q.

The result of Q_{PR} is a quaternion without scalar part:

$$Q_{PR} = 0 + (X_P \cos\theta + Z_P \sin\theta) i + Y_P j + (Z_P \cos\theta - X_P \sin\theta) k \quad (17)$$

Finally, the vector part of the quaternion Q_{PR} is the coordinate of the new point:

$$N_{PR} = (X_P \cos\theta + Z_P \sin\theta, Y_P, Z_P \cos\theta - X_P \sin\theta) \quad (18)$$

The system can utilize the calculated initial angle and initial centre coordinate as long as the Kinect remains in the same position.

IV. Experimental Result

In order to evaluate the accuracy of this system, a skeleton calibration experimental system (Fig. 5) is used to show and compare calibrated results from two Kinect calibration systems. The reason for using two Kinect to test one person is that it is difficult for two people to do the completely same motion at the same time. In our experimental system, for the two Kinect calibration systems, the person does the same motion all the time. Consequently, comparing the calibrated results of both systems is an effective method to test the KSCC algorithm.

Firstly, when the two systems finish the initial data collection, one system starts to send the skeleton calibrated results to another system via TCP. According to the calibrated skeleton coordinates, two skeletons which come from two systems respectively are displayed in one canvas.

Shown in Fig. 6 is an example of the image results after calibration. The left skeleton (Fig. 6-(a)) captured by the left Kinect is facing right. While the middle skeleton (Fig. 6-(b)) captured by the right Kinect is facing left. Fig. 6-(c) shows the result after calibrating the two skeletons. Both skeletons are calibrated to face the same direction, and the coordinates of corresponding joints are very close to each other. The result indicates that the KSCC algorithm is able to calibrate the Kinect skeleton coordinate system.

 ISBN: 978-1-61208-320-9

TABLE I. AVERAGE DIFFERENCES OF TWO KINECT SKELETON CALIBRATION SYSTEMS

Test Index	Left Angle	Right Angle	Angle Gap	Average Difference before Movement			Average Difference after Movement		
				X	Y	Z	X	Y	Z
Test1	-3.4648°	0.7889°	4.2537°	0.00cm	1.00cm	2.00cm	2.00cm	1.28cm	3.78cm
Test2	16.3570°	0.5013°	15.7557°	0.50cm	0.00cm	3.00cm	1.71cm	1.57cm	2.92cm
Test3	21.2544°	0.1923°	21.0621°	0.50cm	1.00cm	3.00cm	1.21cm	1.50cm	2.14cm
Test4	26.4897°	0.9105°	25.5792°	0.50cm	1.50cm	1.00cm	1.85cm	1.92cm	3.87cm
Test5	14.9705°	-12.1837°	27.1542°	0.50cm	1.00cm	1.00cm	2.14cm	1.50cm	1.64cm
Test6	30.6423°	2.0577°	28.5846°	0.50cm	0.50cm	2.50cm	4.50cm	1.50cm	5.57cm
Test7	21.0891°	-17.7214°	38.8105°	0.00cm	0.50cm	2.50cm	2.00cm	1.78cm	3.21cm
Test8	24.5130°	-19.8589°	44.3719°	1.00cm	1.00cm	1.50cm	4.28cm	1.64cm	5.28cm
Test9	24.5566°	-25.8503°	50.4069°	1.50cm	0.50cm	2.50cm	6.35cm	2.21cm	6.00cm
Test10	29.3320°	-26.0919°	55.4239°	0.00cm	1.00cm	2.00cm	4.50cm	3.21cm	5.71cm

Note: 1° = 1 degree of arc, 1cm = 1 centimetre.

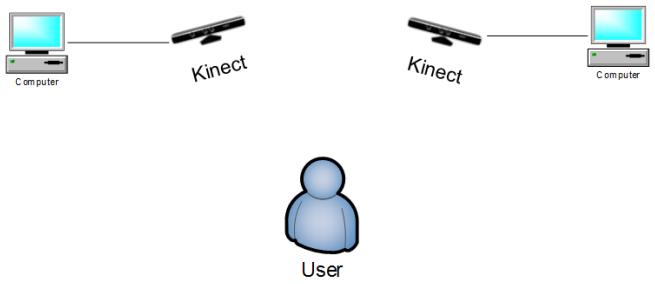
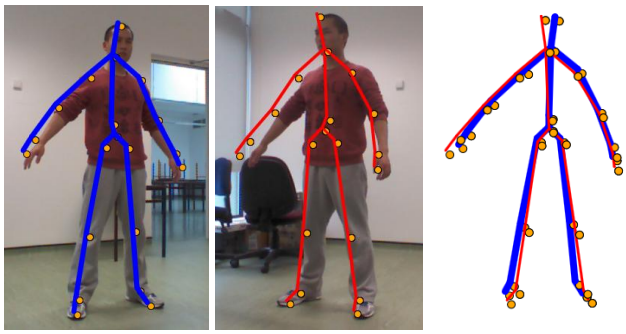


Figure 5. Skeleton calibration testing system



(a) (b) (c)

Figure 6. Original skeletons and calibrated skeletons:
(a) skeleton captured by left Kinect; (b) skeleton captured by right Kinect;
(c) calibrated skeletons result

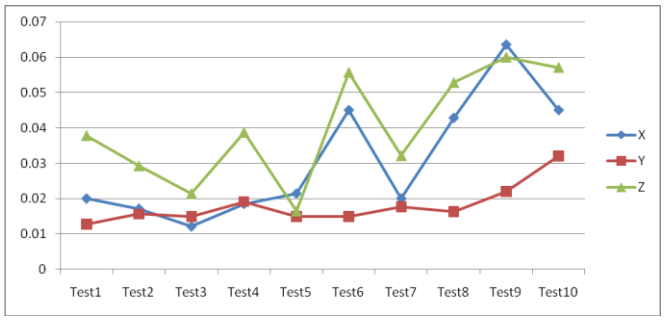


Figure 7. Average differences of 10 Tests

In order to evaluate out the accuracy of the KSCC algorithm, the relative differences for the two calibrated skeletons are also evaluated by using the coordinate recorded for the left and right shoulders as estimation of the measurement. Ten different Kinect angle classes with a total of 420 pairs of recording are evaluated. Each Kinect angle class represents different pairs of initial angles. In each test, seven different locations of experimenter (initial location, move a step forward, move a step to the left, move a step to the right, move a step backward, move a step backward and a step to the left, and move a step backward and a step to the right. All movements start from the initial location) are used to record the coordinates.

Table I shows ten tests' average differences of two Kinect skeleton calibration systems. The second column (Left Angle) is the initial angle between experimenter and the left Kinect; the third column (Right Angle) is the initial angle between experimenter and the right Kinect; the fourth column (Angle Gap) is the gap of the Left Angle and Right Angle; the following three columns are the average differences of coordinates (X, Y, Z) before movement, i.e., after calibration the experimenter still stand at the initial location; the last three columns are the average differences of coordinates (X, Y, Z) after movement, each value is calculated by fourteen pairs of recording (two shoulders and seven locations). We test the initial angle in a range from 0.1923° to 30.6423°. Since the experimenter is human, it is difficult to find absolute 0 degree. We also test the maximum workable initial angle, when the Kinect can detect all twenty joints of the experimenter. The maximum angle is up to 50°. The variation of the maximum angle depends on different standing location of the experimenter. Standing at the middle of the Kinect detection zone has smaller maximum angle than standing at the edges. And the accuracy of the Kinect will be decreased when the angle is increased. Moreover, it will increase the probability of self-occluded other body parts [15][17].

The initial angle gap is a range from 4.2537° to 55.4239°. The average differences in Table 1 show that the differences are very small (1cm~3cm). After movement, the difference increases with the increasing of the initial angle gap. The

 ISBN: 978-1-61208-320-9

average differences of Y are much smaller than X and Z, and more stable. The maximum average difference is 6.35cm (Test9-X).

Test1 is intended to get the least differences, but due to the two Kinect sensors being very close in Test1, the interference with each other will result in a larger error [11]. The average differences of Test6 become large abruptly. It illustrates that the average differences not only depends on the initial angle gap, but also depends on the value of the Left Angle or Right Angle.

A random error of depth measurement increases with increasing distance to a Kinect sensor, and ranges from a few millimetres up to about 4cm at the maximum range (82cm~400cm) of the sensor [15][16]. If setting 4cm as boundary in Fig. 7, there are two kinds of solutions. Solution 1: the initial angle of one Kinect sensor is set to be around 0 degree and the initial angle of another Kinect sensor must be less than around 26.4897° (Test4). Solution 2: the two initial angles (Left Angle and Right Angle) are all less than approximately 21.2544° (Test3). Consequently, in order to control the average differences to less than 4cm, the solution of satisfying the two situations is that the two initial angles must be all less than approximately 20°.

V. CONCLUSION AND FUTURE WORK

In this paper, we proposed a KSCC algorithm to calibrate the Kinect skeleton coordinate for remote physical training applications. The calibration approach is based on the user's initial position data detected by the Kinect sensor. The user's initial centre coordinate and initial angle can be calculated by the initial position data. Twenty joints' coordinates of the user are transformed according to this initial centre coordinate, and rotated by quaternion rotation to a universal coordinate system. An experiment is designed to assess the accuracy and limitation of the system. The experiment results show that the proposed method removes the common constraint in traditional motion comparison systems that the users' initial positions have to be at the same position and facing the sensor with the same angle. Instead, users are allowed to stand at any position in the detection zone of the Kinect sensor, and the initial angle is extended to 20° for a high accuracy result. This improves the consumer experience and gives users more freedom.

As a foundation work for body motion comparison for remote physical training, the KSCC algorithm still needs improvement to reduce the constraints of user's position and motion. As mentioned previously, the accuracy of the Kinect will be decreased when the initial angle is increased. And some body parts often self-occluded due to the limitation of single Kinect sensor. In future work, multiple sensors can be utilized in a 360° environment to reduce the limitation of the single Kinect sensor and extend the available active area of the users.

The Kinect based body motion comparison for remote physical training is not discussed in this paper. In the future, it will be shown that this can be done simply and easily with the KSCC algorithm. Finally, the system will be tested with real physical training motions under remote environment.

REFERENCES

[1] Microsoft Kinect, [Online] Available: http://www.microsoft.com/en-us/kinectforwindows/, December 2013.

[2] J. Webb and J. Ashley, "Beginning Kinect Programming with the Microsoft Kinect SDK," , 2012, pp. 52, pp. 67-100.

[3] S. Essid, et al., "An advanced virtual dance performance evaluator," IEEE Conference on Acoustics, Speech and Signal Processing (ICASSP), 2012, pp. 2269 - 2272.

[4] D. Alexiadis, P. Daras, P. Kelly, N. E. O'Connor, T. Boubekeur and M. B. Moussa, "Evaluating a dancer's performance using Kinect-based skeleton tracking," in ACM Multimedia, 2011, pp. 659-662.

[5] M. Raptis, D. Kirovski and H. Hoppe, "Real-Time clissification of dance gesture from skeleton animation," Eurographics ACM SIGGRAPH Symposium on Computer Animation, 2011, pp. 147-156.

[6] C. H. Huang, T. L. Won, C. Y. Liu, Y. D. Chen and Y. H. Chen, "Multiple-video-based E-learning platform for physical education," Pervasive Computing (JCPC), 3-5 Dec. 2009, pp. 21-26.

[7] K. F. Li, T. Kosuke and G. J. Mark, "Motion tracking and processing for multimedia sport E-Learning," International Conference on Broadband and Wireless Computing, Communication and Applications, 2011, pp. 75-82.

[8] M. Muller, T. Roder, and M. Clausen, "Efficient content-based retrieval of motion capture data," in ACM Siggraph, 2005, pp. 677-685.

[9] Y. Tamura, K. Yamaoka, M. Uehara, and T. Shima, "Capture and feedback in flying disc throw with use of Kinect," Word Academy of Science, engineering and Technology, Feb. 2013, pp. 313-317.

[10] B. K. P. Horn, "Closed-form solution of absolute orientation using unit quaternions," Journal of the Optical Society of America, April 1987, Vol. 4, Issue 4, pp. 629-642.

[11] M. Caon, Y. Yue, J. Tscherrig, E. Mugellini, and O. A. Khaled, "Context-Aware 3D gesture interaction based on multiple Kinects," The first International conference on Ambient Computing, Applications, Services and Technologies, October 2011, pp. 7-12.

[12] Kinect Wikipedia, [Online] Available: http://en.wikipedia.org/wiki/Kinect, December 2013

[13] C. H. John, K. F. George and H. K. Louis, "Visualizing quaternion rotation", ACM Transactions on Graphics, July 1994, Vol. 13, No. 3.

[14] Quaternion Wikipedia, [Online] Available: http://en.wikipedia.org/wiki/Quaternion, December 2013

[15] J. Shotton, et al, "Real-time human pose recognition in parts from single depth images," in CVPR, 2011, pp. 1297-1304.

[16] K. Khoshelham and S. O. Elberink, "Accuracy and resolution of Kinect depth data for indoor mapping applications," Sensors (Basel) 2012, 12(2): 1437-1454.

[17] S. Obdrzalek, et al, "Accuracy and robustness of Kinect pose estimation in the context of coaching of elderly population," Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE, Aug. 28, 2012-Sept. 1, 2012, pp. 1188 - 1193.

 ISBN: 978-1-61208-320-9

Directional Variances Based Demosaicing Method

Joohyeok Kim
Electronics and Computer Engineering
Hanyang University
Republic of Korea
kjh76363@gmail.com

Gwanggil Jeon
Embedded Systems Engineering
Incheon National University
Republic of Korea
gjeon@incheon.ac.kr

Jechang Jeong*
Electronics and Computer Engineering
Hanyang University
Republic of Korea
jjeong@hanyang.ac.kr

***Abstract*—In this paper, we propose a new demosaicing method, which has the improved edge detection method and the refinement scheme. The proposed method finds the interpolation direction based on the directional variances, and then interpolates the missing green components. The missing red and blue components are populated with the use of the fully interpolated green components and color differences. According to the edge direction, two or six neighboring pixels are used to interpolate the red and blue channels. A full colored image, after that, is refined by using median filter with 5×5 cross-shaped kernel. The experimental results show that the proposed algorithm provides a better demosaiced image with relatively low computational complexity.**

Keywords- demosaicing; color filter array; adaptive color plane interpolation

I. INTRODUCTION

A pixel of a full color image is composed of three colors; hence, three separate spectrally selective sensors are required to capture a particular color channel. However, the sensor is one of the most expensive components of a camera system; specifically, it takes about 10-25% of the total cost [1]. For this reason, most cameras use a single sensor covered with a color filter array (CFA) in order to reduce the cost. Fig. 1 shows a popular CFA pattern, known as the Bayer CFA, which is composed of red (*R*), green (*G*), and blue (*B*) filter elements. As one can observe from this CFA, each pixel has only one color component and accordingly the two missing components at each pixel must be estimated. Such an estimation process is called as CFA interpolation or demosaicing.

The simplest method for demosaicing is to use conventional interpolation methods such as bilinear or cubic interpolation [2]-[4]. However, such methods produce some color artifacts because each color channel is independently interpolated without the use of the inter-channel correlation. One solution to consider the inter-channel correlation is to use the color difference rule, which is based on the assumption that the color differences such as *G-R* and *G-B* are quite constant over small regions [2],[5].

Adaptive color plane interpolation (ACPI) proposed in [2] has provided a framework of demosaicing. In order to interpolate a missing green component, ACPI uses the mean term of two neighboring green components and the second order directional Laplacian term of red or blue components two pixels apart on the same column or same row. Effective color interpolation (ECI) calculates the estimates of color differences and utilizes them to interpolate missing components by averaging [6]. Enhanced ECI (EECI) is another method that utilizes color differences for interpolation. In EECI, weight factors on neighbor color differences are utilized for interpolation [7]. EECI shows better results and its complexity is comparable to that of ECI. Recently, voting-based directional interpolation (VDI) is proposed, which adds the voting strategy to determine interpolation direction [8]. A missing pixel is interpolated by using the gradient weights.

In this paper, we propose a new demosaicing method based on ACPI. The proposed method uses variance of directional neighboring pixels to determine interpolation direction of a missing component. Interpolation is performed along the determined direction using the same predictors as those of ACPI.

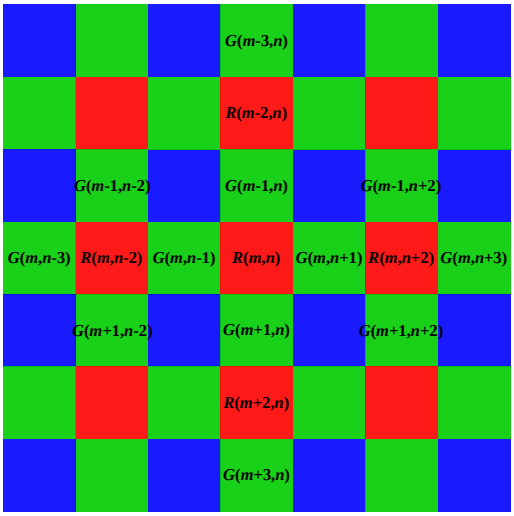


Figure 1. Bayer color filter array.

The remainder of the paper is organized as follows. In Section II, we explain the details of the proposed algorithm, including each color component interpolation and the refinement. Some simulation results are presented and analyzed for comparison in Section III. Finally, we conclude the paper in Section IV.

* Corresponding author

 ISBN: 978-1-61208-320-9

II. PROPOSED METHOD

In the proposed method, we first interpolate the missing green components because the green channel contains important spatial information. After green channel interpolation, red and blue channels are interpolated with the use of the populated green channel. Finally, a fully populated image is refined by median filter.

A. Green Component Interpolation

In order to interpolate the missing green components at the red sampling positions in Fig. 1, three predictors are used as follows:

$$\begin{aligned}\hat{G}_H(m,n) &= \frac{G(m,n-1)+G(m,n+1)}{2} \\ &\quad + \frac{2R(m,n)-R(m,n-2)-R(m,n+2)}{4}, \\ \hat{G}_V(m,n) &= \frac{G(m-1,n)+G(m+1,n)}{2} \\ &\quad + \frac{2R(m,n)-R(m-2,n)-R(m+2,n)}{4}, \\ G_A(m,n) &= \frac{\hat{G}_H+\hat{G}_V}{2}.\end{aligned} \tag{1}$$

These predictors are same as in ACPI. One of them is chosen by edge detection, and used as the estimate of the missing green component at (*m*,*n*). For edge detection, we use the directional variance of neighboring pixels. When calculating the horizontal variance, we exploit green components on the upper and lower lines, the *m*-1th and *m*+1th row, and green and red components on the line that the target pixel is belonging, the *m*th row. Let sets of the positions $\Omega_{H,R0}$, $\Omega_{H,G0}$, $\Omega_{H,G+}$, and $\Omega_{H,G-}$ be defined as

$$\begin{cases}\Omega_{H,G-} = \{(-1,-2),(-1,0),(-1,2)\}, \\ \Omega_{H,G0} = \{(0,-3),(0,-1),(0,1),(0,3)\}, \\ \Omega_{H,G+} = \{(1,-2),(1,0),(1,2)\}, \\ \Omega_{H,R0} = \{(0,-2),(0,0),(0,2)\},\end{cases} \tag{2}$$

and then the variances of pixels on the sets are calculated as follows:

$$\begin{aligned}\sigma^2_{H,R0} &= \frac{1}{3}\sum_{(i,j)\in\Omega_{H,R0}} \left[\mu_{H,R0} - R(m+i,n+j)\right]^2, \\ \sigma^2_{H,G0} &= \frac{1}{4}\sum_{(i,j)\in\Omega_{H,G0}} \left[\mu_{H,G0} - R(m+i,n+j)\right]^2, \\ \sigma^2_{H,G+} &= \frac{1}{3}\sum_{(i,j)\in\Omega_{H,G+}} \left[\mu_{H,G+} - R(m+i,n+j)\right]^2, \\ \sigma^2_{H,G-} &= \frac{1}{3}\sum_{(i,j)\in\Omega_{H,G-}} \left[\mu_{H,G-} - R(m+i,n+j)\right]^2\end{aligned} \tag{3}$$

where $\mu_{H,R0}$, $\mu_{H,G0}$, $\mu_{H,G+}$, and $\mu_{H,G-}$ are the mean values of those pixels, and they are calculated as

$$\begin{aligned}\mu_{H,R0} &= \frac{1}{3}\sum_{(i,j)\in\Omega_{H,R0}} R(m+i,n+j), \\ \mu_{H,G0} &= \frac{1}{4}\sum_{(i,j)\in\Omega_{H,G0}} R(m+i,n+j), \\ \mu_{H,G+} &= \frac{1}{3}\sum_{(i,j)\in\Omega_{H,G+}} R(m+i,n+j), \\ \mu_{H,G-} &= \frac{1}{3}\sum_{(i,j)\in\Omega_{H,G-}} R(m+i,n+j).\end{aligned} \tag{4}$$

The cost for the horizontal direction is defined as sum of the variances as

$$C_H = \sigma^2_{H,R0} + \sigma^2_{H,G0} + \sigma^2_{H,G+} + \sigma^2_{H,G-}. \tag{5}$$

The cost for vertical direction is obtained analogously by using (3)-(5) with the sets defined by

$$\begin{cases}\Omega_{H,G-} = \{(-2,-1),(0,-1),(2,-1)\}, \\ \Omega_{H,G0} = \{(-3,0),(-1,0),(1,0),(3,0)\}, \\ \Omega_{H,G+} = \{(-2,1),(0,1),(2,1)\}, \\ \Omega_{H,R0} = \{(-2,0),(0,0),(2,0)\}.\end{cases} \tag{6}$$

After calculating variances for the horizontal and vertical directions, the missing green component is estimated by

$$\hat{G}(m,n) = \begin{cases}\hat{G}_H(m,n), & \text{if } C_H + \delta < C_V \\ \hat{G}_V(m,n), & \text{else if } C_V + \delta < C_H \\ \hat{G}_A(m,n), & \text{else.}\end{cases} \tag{7}$$

where δ is an offset. With this offset, we can distinguish clear edge from flat or omni-directional edge region. The missing green components on the blue sampling positions are interpolated with the same process except for swapping *R* for *B*.

B. Red/blue Component Interpolation

There are two configurations for interpolating red/blue components at green components as shown in Fig. 2. Because the interpolation process is same for each configuration, we explain only the case of Fig. 2 (a).

The directional variances are also utilized in red/blue component interpolation. That is, we first determine the edge direction at *G*(*m*,*n*) using variances described in the previous subsection. If the direction is horizontal, we interpolate the missing red component at (*m*,*n*) by horizontal average of two color differences at red sampling positions. Because there is no blue component on the horizontal line, we average color differences at six closest blue sampling positions. If the direction is determined as vertical one, we interpolate the

 ISBN: 978-1-61208-320-9

missing red component with six color differences, and the missing blue component with two color differences. When the pixel at (*m*,*n*) is not in edge region, two color differences are used as in ACPI. Let *KR* be the color difference between green and red, and *KB* be the color difference between green and blue. Then, this process can be represented by

$$\text{If } C_H + \delta < C_V$$

$$\hat{R}(m,n) = G(m,n) + \frac{KR(m,n-1) + KR(m,n+1)}{2}$$

$$\hat{B}(m,n) = G(m,n) + \frac{1}{6}\sum_{j=-2,0,2} KB(m-1,n+j) + KB(m+1,n+j)$$

$$\text{elseif } C_V + \delta < C_H$$

$$\hat{R}(m,n) = G(m,n) + \frac{1}{6}\sum_{i=-2,0,2} KR(m+i,n-1) + KR(m+i,n+1)$$

$$\hat{B}(m,n) = G(m,n) + \frac{KB(m-1,n) + KB(m+1,n)}{2}$$

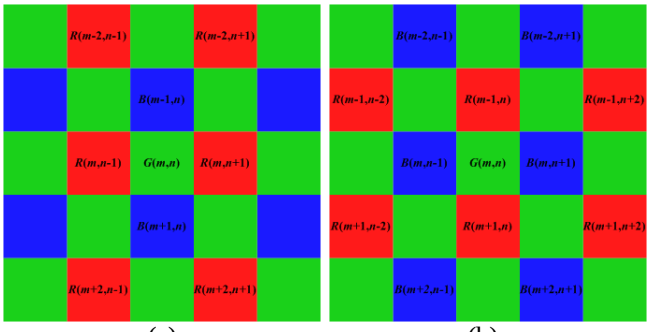
$$\text{else}$$

$$\hat{R}(m,n) = G(m,n) + \frac{KR(m,n-1) + KR(m,n+1)}{2}$$

$$\hat{B}(m,n) = G(m,n) + \frac{KB(m-1,n) + KB(m+1,n)}{2}$$

For interpolating the missing red components at blue sampling positions and the missing blue components at red sampling positions, the average on color differences of four diagonal neighbors is used as the estimates as:

$$\begin{aligned} R(m,n) &= G(m,n) + \big[KR(m-1,n-1) + KR(m-1,n+1) \\ &\quad + KR(m+1,n-1) + KR(m+1,n+1)\big]/4 \\ B(m,n) &= G(m,n) + \big[KB(m-1,n-1) + KB(m-1,n+1) \\ &\quad + KB(m+1,n-1) + KB(m+1,n+1)\big]/4. \end{aligned} \tag{8}$$



(a) (b)

Figure 2. Two configurations for red/blue component interpolation: (a) horizontal GR line, (b) horizontal GB line.

C. *Refinement*

The fully populated image is refined so as to improve image quality. In the proposed method, we use a simple refinement scheme instead of iterative schemes proposed in [9]-[11]. The following median filter is applied to the estimated green component at the red sampling position as shown in Fig. 1 to suppress color artifacts:

$$\tilde{G}(m,n) = \text{median}\begin{bmatrix} KR(m-2,n), KR(m-1,n), KR(m,n), \\ KR(m+1,n), KR(m+2,n), KR(m,n-2), \\ KR(m,n-1), KR(m,n+1), KR(m,n+2) \end{bmatrix}. \tag{9}$$

We choose not a 3×3 window, but a 5×5 cross-shaped window because the 3×3 window includes the *KR* values generated by the estimated green and red components. The green components at the blue sampling positions are similarly obtained by exchanging *KR* for *KB*. After refining all the estimated green components, we refine the estimated red and blue components using the line average of the color differences proposed in ACPI.

III. Experimental Results

To evaluate the performance of the proposed demosaicing method, we simulated it and the four existing methods: ACPI [2], ECI [6], EECI [7], and VDI [8]. The 24 digital color images shown in Fig. 3 were utilized as a set of testing images, each having 768×512 pixels. As a measure, the color peak signal-to-noise ratio (CPSNR) were used as defined as

$$\text{CPSNR} = 10\log_{10}\left(\frac{255^2}{\text{CMSE}}\right) \tag{10}$$

where

$$\text{CMSE} = \frac{1}{3HW}\sum_{k=r,g,b}\sum_{m=1}^{H}\sum_{n=1}^{W}\big(I_o(m,n,k) - I_d(m,n,k)\big)^2. \tag{11}$$

In this equation, I_o and I_d represent the original and the demosaiced images of size $H \times W$. All the testing images are sampled according to the Bayer CFA pattern, and then interpolated. To measure the reconstructed image quality, the interpolated images are compared to the original images.

Table I tabulates the CPSNR results of different methods. It is shown that the proposed algorithm achieved higher PSNR measures than the other methods. The proposed method achieved the best CPSNR results among the compared methods while ACPI, which is the base of the proposed method, showed the worst CPSNR scores. The difference was of 2.95 dB on average and the proposed method showed higher PSNR scores for all the testing images. This signifies that the edge detection scheme of the proposed method outperforms the original one.

 ISBN: 978-1-61208-320-9

Figure 3. Test images (referred to as image 1 to image 24, enumerated from left-to-right and top-to-bottom.)

TABLE 1. Comparison of CPSNR Results

Image	ACPI	ECI	EECI	VDI	Prop.
1	33.48	33.43	37.81	34.13	**38.35**
2	38.50	36.32	**40.36**	39.30	40.26
3	40.44	38.35	**42.74**	40.92	42.39
4	38.56	38.41	**40.53**	38.88	40.15
5	34.59	34.60	**38.09**	35.23	37.96
6	34.78	34.19	38.11	35.99	**38.97**
7	40.55	38.70	**42.60**	41.28	41.99
8	31.85	30.53	35.22	33.03	**36.22**
9	40.12	38.77	42.76	41.00	**42.92**
10	39.47	39.07	**42.43**	40.60	42.05
11	36.06	35.95	39.54	36.91	**39.61**
12	40.45	38.85	**42.67**	41.30	42.25
13	29.89	31.30	34.27	30.09	34.81
14	35.51	34.59	**37.60**	36.11	36.68
15	37.36	35.94	39.15	37.44	**39.27**
16	38.29	36.35	41.28	39.88	**41.62**
17	38.60	39.05	**41.79**	38.87	41.75
18	33.68	35.18	36.82	33.61	**37.02**
19	36.87	35.45	40.12	37.82	**40.31**
20	36.83	36.02	**40.53**	38.44	40.22
21	34.89	35.47	38.90	35.84	**39.19**
22	35.92	36.28	**38.40**	36.69	38.19
23	38.84	38.78	**41.16**	40.77	41.12
24	32.03	33.56	34.58	31.93	**34.64**
average	36.56	36.05	39.48	37.34	**39.51**

Fig. 4 shows the demosaiced images of image 19. Fig. 4 (a) is the cropped version of the original image, and Fig. 4 (b)-(f) are images reconstructed by different methods. The demosaiced image from ECI suffered from color artifacts, and that from EECI had the relatively reduced artifacts but they are still serious. VDI showed the much reduced results, but it cannot avoid zig-zag artifacts. The proposed method reduced those artifacts: color artifacts and zig-zag artifacts; hence, the demosaiced image was smoother than that of VDI. Compared to ACPI, the proposed method showed much better result although both used the same predictors. This demonstrates again that the edge detection and the refinement scheme of the proposed method are excellent.

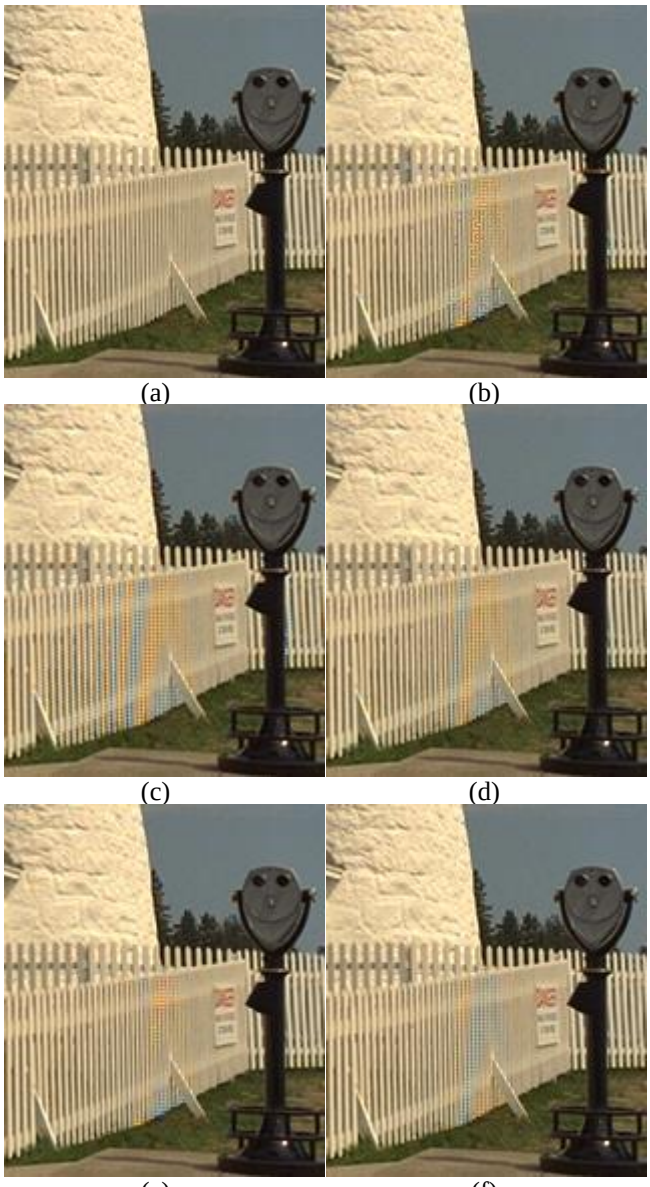


Figure 4. Comparison of Reconstructed Images: (a) original, (b) ACPI, (c) ECI, (d) EECI, (e) VDI, (f) the proposed method.

 ISBN: 978-1-61208-320-9

In order to evaluate the computational complexity, we compared the number of arithmetic operations required for generating a fully populated image from a CFA image. To reduce the complexity, the variance calculation was performed as

$$\begin{aligned}\sigma^2 &= \frac{1}{N}\sum_{i=1}^{N}\left[x_i - \bar{x}\right]^2 \\ &= \frac{1}{N}\sum_{i=1}^{N} x_i^{\,2} - \bar{x}^2,\end{aligned} \tag{12}$$

and the other computations were also optimized. As one can observe from Table II, the proposed method required 6*HW* multiplication operations, which was much less compared to those of EECI and VDI, each requiring 28*HW* and 13*HW*. In the simulations using MATLAB, the proposed method consumed 0.671s while EECI and VDI took 0.925s and 0.702s, respectively, to populate an image. The proposed method required 6.295*HW* comparison operations. In our analysis, we found that it was from the refinement step. In a simulation without the refinement, the number of comparison operations required was 2*HW*, CPU time was reduced to 0.191s, and the average CPSNR result was 37.62dB.

TABLE II. Comparison of Computational Complexity

Image	ACPI	ECI	EECI	VDI	Prop.
ADD	6.75 *HW*	10 *HW*	58 *HW*	42 *HW*	24.5 *HW*
SHT	2.5 *HW*	4 *HW*	2 *HW*	4.5 *HW*	7 *HW*
CMP	0.5 *HW*	0 *HW*	0 *HW*	1.5 *HW*	6.295 *HW*
MUL	0 *HW*	0 *HW*	28 *HW*	13 *HW*	6 *HW*

IV. CONCLUSION

In this paper, we presented a new demosaicing method. In order to determine the interpolation direction, the proposed method used the horizontal and vertical directional variances. Because the predictors proposed in ACPI is fast and has a good performance, we interpolated a missing pixel using the predictors of ACPI along the direction determined by our proposed edge detection method. For interpolating the red and blue channels, we used again the directional variances. Based on the direction determined by the directional variances, two or six neighboring pixels were used for interpolation. A fully interpolated image, then, went through the refinement step. In the simulation for comparison with other conventional methods, it was demonstrated that the proposed algorithm has low computational complexity, while showing better quality than the tested conventional methods.

ACKNOWLEDGMENT

This work was supported by the National Research Foundation of Korea(NRF) Grant funded by the Korean Government(MOE) (NRF-2011-0011312).

REFERENCES

[1] J. Adams, K. Parulski, and K. Spaulding, “Color processing in digital cameras,” IEEE Micro, vol. 18, no. 6, Nov.-Dec. 1998, pp. 20–30.

[2] J. H. Hamilton and J. E. Adams, “Adaptive Color Plane Interpolation in Single Sensor Color Electronic Camera,” U.S. Patent 5 629 734, 1997.

[3] T. Sakamoto, C. Nakanishi, and T. Hase, “Software pixel interpolation for digital still camera suitable for a 32-bit MCU,” IEEE Trans. Consum. Electron., vol. 44, no. 4, Nov. 1998, pp. 1342–1352.

[4] J. Adams, “Interactions between colorplane interpolation and other image processing functions in electronic photography,” Proc. SPIE, vol. 2416, Mar. 1995, pp. 144–151.

[5] E. Chang, S. Cheung, and D. Y. Pan, “Color filter array recovery using a threshold-based variable number of gradients,” Proc. SPIE, vol. 3650, Mar. 1999, pp. 36–43.

[6] S. C. Pei and I. K. Tam, “Effective color interpolation in CCD color filter arrays using signal correlation,” IEEE Trans. Circuits Syst. Video Technol., vol. 13, no. 6, June 2003, pp. 503–513.

[7] L. Chang and Y. P. Tam, “Effective use of spatial and spectral correlations for color filter array demosaicing,” IEEE Trans. Consum. Electron., vol. 50, no. 1, Feb. 2004, pp. 355–365.

[8] X. Chen, G. Jeon, and J. Jeong, "Voting-based directional interpolation method and its application to still color image demosaicking," IEEE Trans. Circuits Syst. Video Technol., accepted, 2013.

[9] R. Kimmel, “Demosaicing: Image reconstruction from color CCD samples,” IEEE Trans. Image Process., vol. 8, Sep. 1999, pp. 1221–1228.

[10] B. K. Gunturk, Y. Altunbasak, and R. Mersereau, “Color plane interpolation using alternating projections,” IEEE Trans. on Image Process., vol. 11, no. 9, Sep. 2002, 997–1013.

[11] K. Hirakawa and T. W. Parks, “Adaptive homogeneity-directed demosaicking algorithm,” IEEE Trans. Image Process., vol. 14, no. 3, Mar. 2005, pp. 360–369.

 ISBN: 978-1-61208-320-9

Video Watermarking Based on Interactive Detection of Feature Regions

Asma Kerbiche, Saoussen Ben Jabra, Ezzeddine Zagrouba
Lab. RIADI - Team of Research SIIVA
Higher Institute of Computer Science
University Tunis El Manar
Ariana, Tunisia
asma.kerbiche@gmail.com saoussen.bj@laposte.net
ezzeddine.zagrouba@fsm.rnu.tn

Vincent Charvillat
Lab. IRIT - Team of Research VORTEX
ENSEEIHT - INP TOULOUSE
University of Toulouse
Toulouse, France
vincent.charvillat@enseeiht.fr

***Abstract*—Video watermarking is very important in many areas of activity and especially in multimedia applications. Therefore, security of video stream has recently become a major concern and has attracted more and more attention in both the research and industrial domains. In this perspective, several video watermarking approaches are proposed but, based on our knowledge, there is no method which verified the compromise between invisibility and robustness against all usual attacks. In our previous work, we proposed a new video watermarking approach based on feature region generated from mosaic frame and multi-frequential embedding. This approach allowed obtaining a good invisibility and robustness against the maximum of usual attacks. In our future work, we propose to optimize the choice of the region of interest by using crowdsourcing technique. This last one is an emerging field of knowledge management that involves analyzing the behavior of users when they view a video to automatically deduct the regions of interest.**

Keywords-watermarking; crowdsourcing; robustness; invisibility.

I. INTRODUCTION

The fast development of the Internet in recent years has eased the process of coping, transmitting, and distributing digital data such as image, 3D meshes and video. In fact, the recent decade has seen the emergence of video-based application technologies, such as wireless video, video conferencing, and videophones. As with other types of multimedia data, it was necessary to find new techniques for visualization, compression, indexing, and also the security of this type of media to enable the protection of rights, authentication and data integrity during transfer on a given communication channel. Therefore, the need for protection from piracy and illegal use rises more and more. The best technique to protect digital video from this manipulation is watermarking. It consists of embedding a signature into data and to try to detect it after any manipulation done on marked data. Usually, signature must be imperceptible and should resist to malicious attacks which try to destroy or to remove it. Several studies have been made to develop robust and invisible video watermarking methods [18]. These methods differ mainly in the insertion area [27] (spatial or frequency) and the type of the treated stream [26] (compressed or uncompressed). To improve the invisibility and the robustness of the watermark, some methods are based on the choice of region of interest where the embedding will be done. This choice can be based on many tests applied on different feature regions [17] or using a specific technique [13]. In our work, we propose to use crowdsourcing technique that involves analyzing the behavior of users when they view a video to automatically deduct the regions of interest. The remainder of this paper is organized as follows: in Section 2, video watermarking and crowdsourcing technique are presented while Section 3 describes the proposed watermarking method based on interactive detection of regions of interest using crowdsourcing technique. Finally, conclusion is drawn in Section 4.

II. RELATED WORK

A. Watermarking

Several techniques of watermarking video have been proposed in the literature. We chose to classify them according to two criteria. The first one is the original video format which can be compressed or uncompressed. In fact, watermarking can be applied to compressed stream where the insertion is done during the compression process or after compression [1],[2],[3]. For the uncompressed video, four classes of embedding methods can be applied (some of them can also be used on compressed video). The first class is derived from frame by frame watermarking that consider video as a succession of images and consists of applying still images watermarking algorithms [4],[5]. The second one is the spatio-temporal schemes where video is defined as a 3D signal considering the temporal dimension in video sequences. These schemes decompose the video by performing spatial 2D transform on individual frames followed by 1D transform in the temporal domain [6],[7],[8]. The third class is the temporal schemes that insert the signature in the temporal domain by modifying only the low spatial frequencies [9],[10],[11]. Finally, the last class is based on mosaic frame generated from the original video [12],[13]. This last one selects an interesting area where the mark should be embedded. In fact, mosaïcing allows the insertion of the same mark into the same pixels which represent the same physical point.

 ISBN: 978-1-61208-320-9

TABLE I. ADVANTAGES AND INCONVENIENCES OF EACH CLASS ACCORDING TO VIDEO FORMAT

Methods	Advantages	Inconveniences	Invisibility
Frame by frame watermarking	Minimum insertion time and Avoids compression and decompression steps which can degrade the images of the sequence.	May be ineffective against a MPEG-4 compression that will destroy the inserted signature. Possibility of artifacts	+
Spatio-temporal schemes	Robustness against temporal changes	Unblinded detection and low insertion capacity	++
Temporal Schemes	The mark value is constant for a given image, but vary from one image to another.	Fragile watermark against temporal changes.	+
Compressed Video	If the encoding is complex detection is not.	Fragile watermark against motion estimation and may disappear at re-coding.	++
Mosaic Schemes	Robust against MPEG-4 compression, collusion attacks and deleting images	Not robust against MPEG-2 compression.	++

The second criterion is embedding domain where signature can be inserted directly on video by modifying its pixels [14] or by modifying some video transformations like DCT (Discrete Cosine Transform**)**, DWT (Discrete Wavelet Transform) and SVD transform (Singular Value Decomposition) [15],[16]. The scheme disperses the watermark in the spatial domain of the video frame, hence making it very difficult to remove the embedded watermark.

TABLE II. COMPARISON OF EMBEDDING DOMAIN CLASSES

	Methods	Advantages	Invisibility
Frequency Domain	DCT	The most robust against compression	++
	DWT	The most robust against compression	++
	SVD	Robust against compression, rotation, noise and deleting frames	++
Spatial Domain	LSB	Not robuste against collusion, compression, noise...	+
	Correlation-Based techniques		

In our previous work, after a comparative study of these classes [17], we proposed a new watermarking schema based on mosaic and multi-frequentiel embedding algorithm using wavelet, DCT and SVD transforms. In fact, it consists to generate mosaic from original video. Then, the region where the objects move is selected to be marked. Finally, signature is inserted using wavelet, DCT and SVD transforms. This choice allows obtaining robustness against various types of attacks, such as geometric transformations (rotation, zooming), cropping vertices, MPEG4 part X (h.264 Advanced Video Coding (AVC)) compression, noise, frame suppression and collusion.

B. Crowdsourcing

With the sweeping progress of Web 2.0 technologies and capabilities, many socio-technical systems have attracted attention from both practitioners and scholars. Crowdsourcing is a new emerging Web 2.0 based phenomenon and becomes a recognized sourcing mechanism for problem solving in organizations and societies by outsourcing problems to an undefined entity or the 'crowd'. For that, crowdsourcing research has become a dynamic and vibrant research area, and has been steadily growing over the years.

The term crowdsourcing was first coined by Howe, in a Wired Magazine article in June 2006 [19]: "Simply defined, crowdsourcing represents the act of a company or institution taking a function once performed by employees and outsourcing it to an undefined (and generally large) network of people in the form of an open call. This can take the form of peer-production (when the job is performed collaboratively), but is also often undertaken by sole individuals. The crucial prerequisite is the use of the open call format and the large network of potential laborers".

In essence, crowdsourcing is based on a simple, but powerful, concept: virtually everyone has a potential to plug in valuable information [20]. It seeks to mobilize competence and expertise, which are distributed among the crowd and has different forms [21].

Crowdsourcing is not exclusive for business purposes. In fact, many non-profit organizations have adopted it as an effective model for problem-solving [22],[23]. In addition to having gained great attention and interest from the industry, crowdsourcing has also gained attention from the academic community.

Indeed, several recent studies are based on crowdsourcing technique. Xie et al. [24] propose new method to detect user interest maps and extract user attention objects from the image browsing log using crowdsourcing where ten subjects were selected to take part in this study, and the criterion for their selection was that they should be very familiar with the use of computers and cell phones before the study. A smart image viewer was then developed based on user interest analysis and a second experiment was carried out to study how users behave with such a viewer. This approach is more efficient than image-analysis based methods and can better represent users' actual interest. Based on the fact that the viewing experience on the mobile devices can be improved by determining important and interesting

 ISBN: 978-1-61208-320-9

regions within the video (regions of interest, or ROIs) and displaying only the ROIs to the viewer, Carlier et al [25] propose an alternative paradigm to infer ROIs from a video by crowdsourcing from a large number of users through their implicit viewing behavior using a zoom and pan interface, and infer the ROIs from their collective wisdom. A retargeted video, consisting of relevant shots determined from historical users' behavior, can be automatically generated and replayed to subsequent users who would prefer a less interactive viewing experience. A user study with 48 participants shows that this automatically retargeted video is of comparable quality to one handcrafted by an expert user.

III. PROPOSED METHOD

The study of watermarking and crowdsourcing recent works shows that these two areas can be combined to propose a new watermarking approach which presents a high level of robustness against the most important attacks. In fact, our work aims to develop new robust approaches to introduce signatures in videos. Our idea consists in a first step to understand the visual content of the original video and then to select feature regions to embed signature. To achieve our goal, crowdsourcing technique will be used. Although, the concept of crowdsourcing is based on sharing media to the public, this can cause confidentiality problems and can damage watermarking process. To avoid this problem, we thought to generate a video summary and to share it to a fixed number of selected users.

The proposed approach is decomposed to three main stages: video summarization, interactive detection of feature regions using crowdsourcing technique and signature insertion. General architecture of the proposed approach is presented by Figure 2.

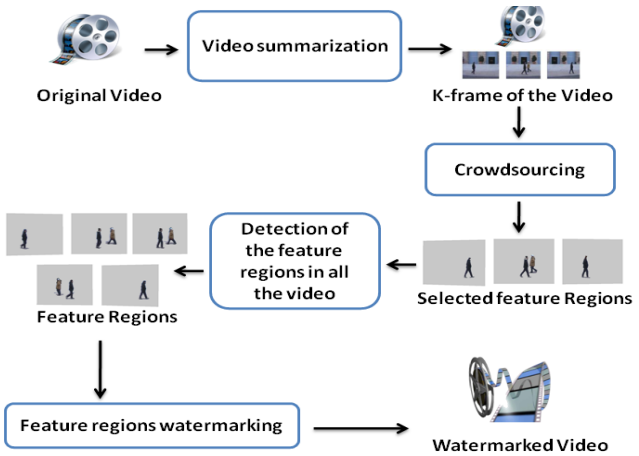


Figure 1. General architecture our proposed method.

A. 1st Step: Video Summarization

To avoid confidentiality problems that can be caused by sharing the original video, its summary will be generated using the approach proposed in [29]. This last one is an object-based technique which allows extracting a relatively small number of still key-frames in order to summarize the salient visual content of a video. This method is based on spatial segmentation of each frame in order to detect the important events. Indeed, the extraction of key-frames will be facing a much more semantic criterion so that each extracted key-frame present an important event such as the appearance and/or the disappearance of significant objects.

B. 2nd Step: Feature Regions Detection

After video summarization, feature regions must be selected to embed signature. To detect these regions, crowdsourcing technique is chosen. In fact, this technique is an emerging field of knowledge management which allows analyzing the behavior of users when they watch a video to automatically deduct the regions of interest. Indeed, the combination of the analysis of visual content and interactive use of the identified data can improve the detection of interest visual objects in a video. The method proposed in [26] and explained in the previous section, has been used for this step. To ensure the confidentiality of our video and to avoid video hacking, we choose to interact with an X number of selected users in our laboratory.

C. 3rd Step: Watermarking

After extracting user attention objects, the signature will be inserted in the N feature regions selected from second step. To embed signature, our multi-frequential watermarking scheme [17] based on DWT, DCT and SVD transforms will be used. In fact, the regions selected by users are the most important regions in the video and their destruction will cause the destruction of the whole video. Thereby, the insertion in these selected regions will ensure a greatest robustness of our watermark. General architecture of the embedding step is presented by Figure 2.

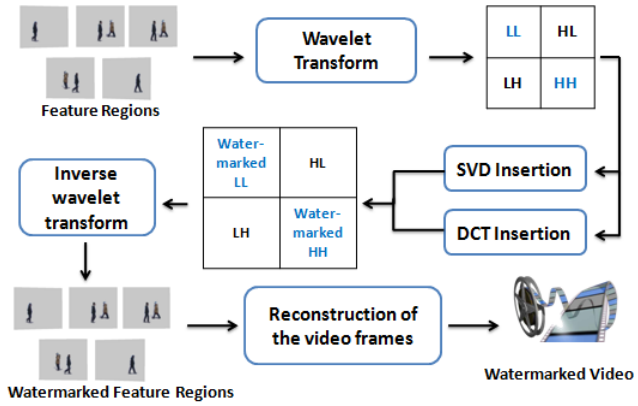


Figure 2. General architecture of the embedding step.

D. Test

The reliability of the developed method will be tested by evaluating the two main criterions: robustness and invisibility. In fact, for the robustness we will test the resistance of the watermark against compression phases, collusion, Cropping and several main types of video attacks that could destroy the watermark. For invisibility criterion

 ISBN: 978-1-61208-320-9

and in order to enhance our tests, crowdsourcing process will be used. A user study will be carried.

IV. CONCLUSION AND FUTUR WORK

Crowdsourcing has become a dynamic and vibrant research area, and has been steadily growing over the years. In fact, it can be applied to a wide variety of problems and ever more classes of applications. For this reason, we proposed in this paper a new video watermarking based on crowdsourcing technique to select the most interesting feature regions. These last ones will be used to insert signature. In fact, interactive detection of feature regions using Crowdsourcing technique will guaranty a high level of robustness and invisibility of marked video. The robustness of the proposed method will be verified after application of various types of attacks such as geometric transformations (rotation, zooming), cropping vertices, MPEG4 compression, noise, frame suppression and collusion. For invisibility, we will use three metrics (peak signal to noise ratio (PSNR), Haussdorff distances and correlation) and a user study will be done to measure marked video quality.

This method can be extended to 3D video and can be used by all watermarking applications such as video indexing.

REFERENCES

[1] N. Mohaghegh and O. Fatemi, "H.264 copyright protection with motion vector watermarking," in Proc. Int'l Conf. on Audio, Language and Image Processing, July 2008, pp. 1384-1389.

[2] P. Bas, J. M. Chassery, and B. Macq, "Image watermarking; An evolution to content based approaches," Pattern Recognition, Special Issue on Image/Video Communication, vol 35, march 2002, pp 545-561.

[3] H. Joumaa and F. Davoine, "Performance of an application video watermarking scheme using informed techniques," in IEEE International Conference on Image Processing, vol 1, Sept. 2005, pp. 261-365.

[4] J. J. Chae and B. S. Manjunath, "Data hiding in video," in 6th IEEE International Conference on Image Processing (ICIP'99), Kobe, Japan, vol 1, Oct. 1999, pp. 311-315.

[5] S. N. Merchant, A. Harchandani, S. Dua, H. Donde, and I. Sunesara, "Watermarking of video data using integer-to-integer discrete wavelet transform," in Proc. IEEE TENCON Conference on Convergent Technologies for the Asia-Pacific Region, vol 3, Oct 2003, pp. 939-943.

[6] F. Deguillaume, G. Csurka, J. J. O'Ruanaidh, and T. Pun, "Robust 3D DFT video watermarking," in Proc. Security and Watermarking of Multimedia Contents, SPIE, vol. 3657, no. 1, 1999, pp. 113-124.

[7] J. H. Lim, D. J. Kim, H. T. Kim, and C. S. Won, "Digital video watermarking using 3D-DCT and intracubic correlation," in Proc. SPIE Security and Water- marking of Multimedia Contents III, vol. 4314, no. 1, 2001, pp. 64-72.

[8] S. J. Kim et al., "A new digital video watermarking using the dual watermark images and 3D DWT," in Proc. IEEE Region 10 TENCON, vol. 1, 2004, pp. 291-294.

[9] P. Vinod and P. K. Bora, "Motion-compensated inter-frame collusion attack on video watermarking and a countermeasure," in IEEE Proceedings on Information Security, vol. 153, no. 2, June 2006, pp. 61-73.

[10] P. Vinod, G. Doerr, and P. K. Bora, "Assessing motion-coherency in video watermarking," in Proc. ACM Multimedia and Security, 2006, pp. 114-119.

[11] K. Su, D. Kundur, and D. Hatzinakos, "Statistical invisibility for collusion resistant digital video watermarking," in IEEE Trans. Multimedia, vol. 7, no. 1, Feb 2005, pp. 43-51.

[12] G. Doerr and J. Dugelay, "Secure background watermarking based on video mosaicing," in Proc. SPIE 5306 Electronic Imaging, 2004, pp. 304-314.

[13] M. Koubaa, M. Elarbi, C. Ben Amar, and H. Nicolas, "Collusion, MPEG4 compression and frame dropping resistant video watermarking," in International Journal of multimedia tools and applications MTAF, Springer Netherlands, Vol. 56, 2012, pp. 281-301.

[14] H. Seddik, M. Sayadi, F. Fnaiech, and M. cheriet, "A New Spatial Watermarking Method, based on a Logarithmic transformation of An Encrypted embedded Mark", in 17th IMACS World Congress on Scientific Computation, Applied Mathematics and Simulation, Paris, France, July 2005.

[15] L. S. Liu, R. H. Li, and Q. Gao, "A robust video watermarking scheme based on DCT," in Machine Learning and Cybernetics, Proceedings of 2005 International Conference, Vol. 8, Aug. 2005, pp. 5176-5180.

[16] P. P. W. Chan, and M. R. Lyu. "A DWT-based Digital Video Watermarking Scheme with Error Correcting Code," in ICICS, Vol. 2836 Springer, 2003, pp. 202-213.

[17] A. Kerbiche, S. B. Jabra, and E. Zagrouba, "A robust video watermarking based on image mosacing and multi-frequential embedding," in IEEE International Conference on Intelligent Computer Communication and Processing ICCP, Aug. 2012, pp 159-166.

[18] R. T. Paul, "Review of Robust Video Watermarking Techniques," in IJCA Special Issue on Computational Science New Dimensions and Perspectives, Vol. 3, 2011, pp. 90-95.

[19] J. Howe, "The rise of crowdsourcing," in Wired Magazine 14(6), June 2006, pp. 1-4.

[20] S. Greengard, "Following the Crowd," in Communications of the ACM 54(2), Feb. 2011, pp 20–22.

[21] Y. Zhao and Q. Zhu, "Evaluation on crowdsourcing research: Current status and future direction." in Information Systems Frontiers, 2012, pp 1-18.

[22] D. C. Brabham, "Crowdsourcing as a model for problem solving: an introduction and cases," in The International Journal of Research into New Media Technologies 14(1), 2008, pp 75-90.

[23] D. C. Brabham, "Moving the crowd at threadless: motivations for participation in a crowdsourcing application," in Information, Communication and Society 13(8), 2010, pp 1122–1145.

[24] X. Xie , H. Liu , S. Goumaz , and W. Y. Ma, "Learning User Interest for Image Browsing on Small-formfactor Devices" in SIGCHI Conference on Human Factors in Computing Systems, 2005, pp. 671-680.

[25] A. Carlier, V. Charvillat, W. T. Ooi, R. Grigoras, and G. Morin, "Crowd-sourced Automatic Zoom and Scroll for Video Retargeting," in ACM Multimedia, 2010, pp. 201-210.

[26] P. Campisi and A. Neri, "Video watermarking in the 3D-DWT domain using perceptual masking," in Proc. IEEE ICIP, vol. 1, 2005, pp. 997-1000.

[27] L. Rajab, T. Al-Khatib, and A. Al-Haj, "Video Watermarking Algorithms Using the SVD Transform," in European Journal of Scientific Research, 2009, pp. 389-401.

[28] N. Mlik, W. Barhoumi, and E. Zagrouba, "Object-based event detection for the extraction of video key-frames," in International Conference on Multimedia Computing and Systems, Tangier, Morocco, May 2012.

 ISBN: 978-1-61208-320-9

Tone Reproduction based on Singular Value Decomposition for High Dynamic Range Imaging

Changwoo Ha, Wonkyun Kim, Cheonghee Kang, and Jechang Jeong
Department of Electronics and Computer Engineering
Hanyang University
17 Haengdang-dong, Seongdong-gu, Seoul, Korea
hahanara@hanyang.ac.kr, wonkyun.kim@gmail.com, cheonghee.kang@gmail.com, and jjeong@hanyang.ac.kr

***Abstract*— This paper presents a tone reproduction method with the appropriate tone and details of high dynamic range images on the conventional low dynamic range display devices. The proposed algorithm mainly consists of image decomposition using singular value decomposition, a singular value based luminance adjustment, and an image composition process. In the final tone reproduction process, the proposed algorithm combines color and luminance components in order to preserve the color consistency. The experimental results show that the proposed method achieves good subjective quality while enhancing the contrast of the image details.**

Keywords-high dynamic range imaging; tone reproduction; tone mapping; singular value decomposition

I. INTRODUCTION

The area of high dynamic range (HDR) imaging is an attractive method with the improved image capturing hardware, and it is an image processing method that produces a large dynamic range. There are many practical applications including scientific, medical visualization [1], satellite imagery [2], and digital camera [3]. In particular, the ability to capture the real world luminance in a scene has become a necessary function of digital camera. However, compared with the 10^8:1 range of real world luminance from bright sunlight to starlight, the HDR image display devices, such as CRT, LCD, and LED have a low dynamic range (LDR) of 10^2:1 cd/m^2 [4].

Although HDR monitors have been existed and researched, they are not only rare and costly, but also difficult to calibrate [5]. Therefore, the tone reproduction (or tone mapping) method is necessary to display on the conventional display devices, for which the luminance range should be transformed to a displayable range that below two orders of magnitude. In tone reproduction, the HDR image contains details in the extremely dim and extremely bright regions; therefore it is difficult to preserve the details while compressing the HDR into the LDR. The discrepancy between the wide range of luminance captured by HDR techniques and the small range of luminance reproduced within displayable ranges from the HDR causes an inaccurate display of the images. Due to this discrepancy, the proper tone reproduction techniques are required. These techniques not only transform HDR images into the displayable range, but also preserve the details, considering the characteristics of the original HDR images.

In the previous works, most tone reproduction techniques have focused on compression and quality evaluation for visualizing HDR images. The tone reproduction can be classified as a global operator [6]-[9] which is the same reproduction function is applied in all regions, or as a local operator [10]-[12] which is the different tone reproduction functions are applied throughout the modeling of spatial adaptations.

This paper is organized as follows. Section II provides a basic concept that is used for the proposed algorithm. In Section III, the proposed method is described. Experimental results prove the performance of the proposed method in Section IV. The paper is concluded with an overall discussion in Section V.

II. PRELIMINARY KNOWLEDGE

The singular value decomposition (SVD) of a rectangular matrix A has many important properties and useful applications [13]-[16]. Without loss of generality, for every $m\times n$ ($m\geq n$) matrix A, the SVD can be written as

$$A = USV^T = \sum_{k=1}^{n} u_k s_k v_k^T \qquad (1)$$

where $U=[u_1, u_2, ..., u_m] \in R^{m\times m}$ and $V=[v_1, v_2, ..., v_n] \in R^{n\times n}$ are two-column orthogonal matrices, and V^T denotes the transpose of V. S is a diagonal matrix with elements s_i, i=1,2,…,n. The diagonal elements can be sorted in a decreasing order, and these are the singular values of matrix B:

$$S = \begin{pmatrix} s_1 & & \\ & \ddots & \\ & & s_n \end{pmatrix} = diag\{s_1, s_2, \cdots, s_n\}. \qquad (2)$$

The matrix S represents the luminance information of the given image. Note that increasing the magnitudes of the singular values of matrix S will increase the image luminance range, whereas lowering the magnitudes will decrease the image luminance range. Therefore, we can

 ISBN: 978-1-61208-320-9

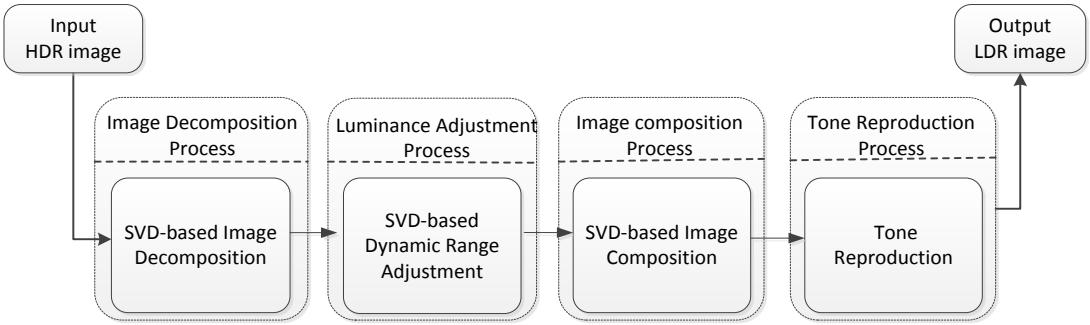


Figure 1. Overall structure of the proposed algorithm.

use a singular value instead of luminance range. For example, the SVD-based contrast enhancement technique is used while preserving the original image characteristic [13, 14]. In other words, SVD-based method will control the dynamic range maintaining the characteristic of the original image.

III. PROPOSED METHOD

The basic principle of the proposed algorithm is to perform a luminance adjustment using SVD. The objective of the proposed SVD-based method is to maintain the important features in the transformation from HDR to LDR. The proposed luminance adjustment method works on the luminance component of the image only which is described in the logarithmic space. From the HDR image, the run-length encoded input values are converted to floating point numbers with linear *RGB* values. The *RGB* values are then converted to a CIE *XYZ* color space [17]. The standard matrix in logarithmic space is obtained by

$$L_i = \log(0.2126 \times R + 0.7152 \times G + 0.0722 \times B) \quad (3)$$

where L_i is the input luminance value and *R*, *G*, and *B* are the red, green, and blue components.

Firstly, we calculate an anchor image, which offers the dynamic range of displayable devices. The process of displaying an HDR image is composed of quantization and mapping as shown in Fig. 1. Since there are too many discrete values in the high dynamic scene, the tone mapping methods must reduce the number of candidate values by quantization. Quantization is a well-known clustering method [18]. Let $x(k)$ with $k = 1, 2, \ldots, N$ be the luminance elements of the HDR image. A quantizer is represented by an encoder Q, which maps $x(k)$ to an index $n \in N$. One of a small collection of mapping values (candidates) $C = \{c_n ; n \in N\}$ is used for mapping, where N is set to 256, which is the number of displayable levels in the LDR image. The quantizer is described as

$$Q(x(k)) = n, \quad if \left\| x(k) - c_n \right\| \le \left\| x(k) - c_i \right\| \ \forall i. \quad (4)$$

A pixel in the HDR image is assigned to the available candidate close to the original value. A set of all pixels assigned to the same candidates is defined as a cluster of pixels. In other words, all HDR pixels belong to the same cluster are displayed at the same LDR level. Pixels in the cluster of a larger candidate value are expected be brighter than those of a smaller candidate value. Note that the propose method only works in logarithmic space, treats each pixel individually, and uses the scalar quantization.

The HDR log image, L_i, is decomposed through the use of (1). Then, the proposed method controls the luminance adjustment by using (2). The adjustable singular value can be seen as a solution of a bi-criteria optimization problem. The object is to find the adjustable singular value $\tilde{s}$ that is close to the singular value of the quantized image s_q, while also reducing the size of the s_i. This adjustable singular value can then be used to obtain the composition using (5) below. This adjustable singular value can be formulated as a weighted sum of the two objectives as

$$\tilde{s} = \arg\min_s \left| s - s_i \right| + \lambda \left| s - s_q \right| \quad (5)$$

where s, s_i, and s_q are the maximum singular value, the maximum singular value of the original HDR image, and the maximum singular value of the quantized image, respectively. The parameter λ varies over $[0, \infty]$, and the solution of (5) gives the optimal trade-off curve between the two objectives. When the squared sum of the Euclidean norm is used, the following analytical solution of (5) can be obtained:

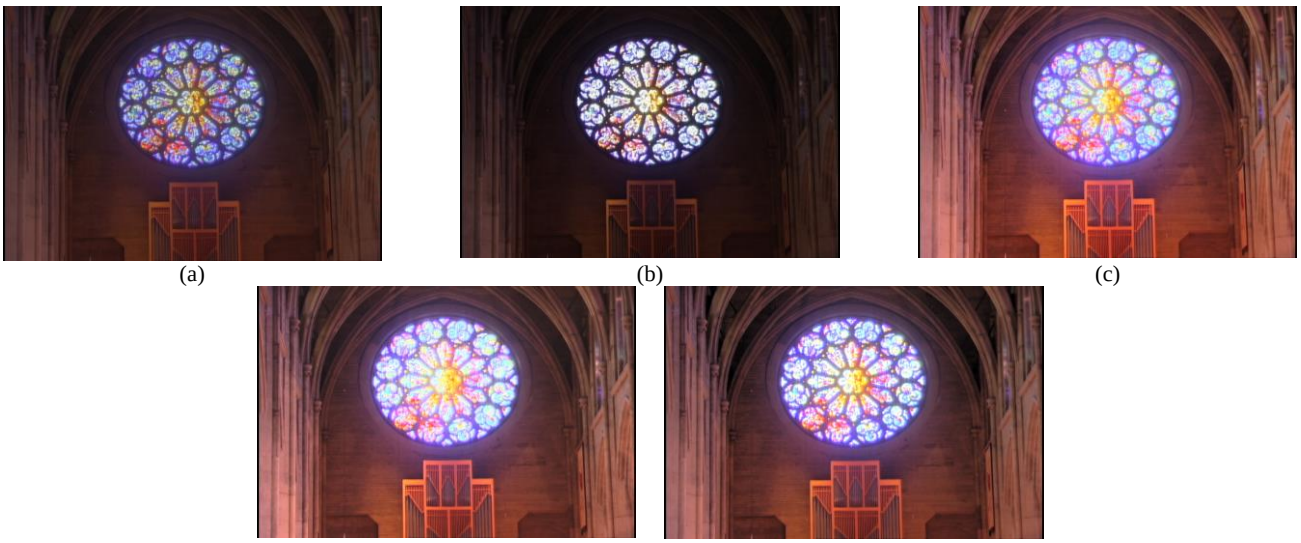
$$\tilde{s} = \arg\min_s (s - s_i)^2 + \lambda (s - s_q)^2. \quad (6)$$

The solution to this minimization problem is obtained by

$$\tilde{s} = (1 + \lambda)^{-1} (s_i + \lambda s_q). \quad (7)$$

The adjustable singular value $\tilde{s}$ is a weighted average of s_i and s_q. If λ is zero, $\tilde{s}$ is the singular value of the input image. As λ approaches infinity, it approaches the singular value of the quantized image. Thus, various levels of luminance can be adjusted by simply changing the parameter λ. Therefore, the displayable luminance adjustment L_d is composed through the use of the optimal singular value ratio to obtain the following:

 ISBN: 978-1-61208-320-9



(a) (b) (c) (d) (e)

Figure 2. The quality comparison with Rosette (720×480): (a) Photoreceptor method, (b) Segmentation-based method, (c) Logarithmic method, (d) Photographic method, and (e) Proposed method. Image courtesy of Paul Debevec.

(a) (b) (c) (d) (e)

Figure 3. The quality comparison with Memorial (512×768): (a) Photoreceptor method, (b) Segmentation-based method, (c) Logarithmic method, (d) Photographic method, and (e) Proposed method. Image courtesy of Paul Debevec.

$$L_d = U\tilde{S}V^T, \quad where\ \tilde{S} = \frac{\max(\tilde{s})}{\max(s_i)} S. \tag{8}$$

Conventional methods reproduce images by using the luminance information only. However, the reproduced luminance needs to be combined with the color information to produce good quality tone reproduced images. In the proposed algorithm, the modified color and luminance components are combined to obtain the final tone reproduced values in (9), because the color shifts will be minimized if the ratio among the color channels before and after the compression is maintained. The scale of the color components is determined by L_d, and the ratio among the color channels (C = R, G, B) is preserved by the fractions C_i/L_i as in

$$C_d = \left(\frac{C_i}{L_i}\right)^{\gamma} \times L_d \tag{9}$$

where C_d represents the final tone reproduced R, G, and B values for the display, and C_i represents the input HDR data with L_i. To control the amount of saturation in an image, the fraction is fitted with an exponent, γ, resulting in a per-channel gamma correction, which is given as a user parameter in the range of 0 to 1. Experimentally, γ was set as 0.45.

 ISBN: 978-1-61208-320-9

IV. EXPERIMENTAL RESULTS

The experimental results are comparable in quality and performance to the manually reproduced images. In this simulation, the parameter λ was set to 99. In Figs. 2 and 3, the proposed algorithm is compared to the conventional tone reproduction operators from a subjective quality point of view. With default parameter settings for each technique, the proposed algorithm is compared using four global operators: the photoreceptor method of Reinhard *et al* [7], the segmentation-based technique of Yee *et al* [8], the logarithmic method of Drago *et al* [9], and the photographic method of Reinhard *et al* [10], are used. Brief explanations of the methods used to compare against the performance of the proposed algorithm are described as follows. In the global operators, the photoreceptor method first mimics the photoreceptor physiology with some user parameters, but also applies a simple local operator in the spatial domain. Next, the segmentation-based method computes the local adaptation luminance for a global operator that makes use of the image segmentation. The logarithmic method is a simple global tone reproduction algorithm that compresses the luminance range according to the base of the logarithm chosen for each pixel. The photographic method has the global and local operators. In this experiment, the photographic method uses a global operator without a Gaussian filter with a parameter estimation technique.

Figs. 2 and 3 show global tone operators that produce good subjective results, but cannot successfully preserve both tone and local details in an image, particularly in the dim and bright regions such as the cropped regions. In Figs. 2(a) and 3(a), the details are enhanced but the tone is lost to low brightness. In Figs. 2(b) and 3(b), the tone and details are washed out in the whole image. The tone is not natural because the details are over enhanced and the tone is missing, as shown in Figs. 2(c)-(d) and 3(c)-(d). In Figs. 2(e) and 3(e), the proposed technique achieves the natural tone and preserves the details simultaneously.

V. CONCLUSION

The most important aspect of reproducing an HDR image as an LDR image is to preserve the tone and details of the HDR image without causing undesirable artifacts, such as halo effects. The proposed algorithm adjusts the luminance values of an image in order to preserve the tone and details based on the SVD-based luminance adjustment. By combining the modified color components with the displayable scale, the color consistency is preserved. The experimental results show that the proposed algorithm achieves good subjective results, especially the balance of the details and subtle tones in both dark and bright regions.

ACKNOWLEDGMENT

"This research was supported by the MKE(The Ministry of Knowledge Economy), Korea, under the ITRC(Information Technology Research Center) support program supervised by the NIPA(National IT Industry Promotion Agency)" (NIPA-2013-H0301-13-1011).

REFERENCES

[1] A. A. Bell, J. Brauers, J. N. Kaftan, D. Meyer-Ebrecht, A. Bocking, and T. Aach, "High dynamic range microscopy for cytopathological cancer diagnosis," IEEE J. Sel. Topics Signal Process., vol. 3, no. 1, Feb. 2009, pp. 170-184.

[2] A. R. Varkonyi-Koczy and A. Rovid, "High-dynamic-range image reproduction methods," IEEE Trans. Instrum. Meas., vol. 56, no. 4, Aug. 2007, pp. 1465-1472.

[3] S. Battiato, A. Castorina, and M. Mancuso, "High dynamic range imaging for digital still camera: an overview," J. of Elect. Imaging, vol. 12, no. 3, Jul. 2003, pp. 459-469.

[4] E. Reinhard, G. Ward, S. Pattanaik, P. Debevec, W. Heidrich, and K. Myszkowski, High dynamic range imaging: acquisition, display, and image-based lighting, 2nd ed., Morgan Kaufmann: San Francisco, 2010.

[5] Y.-K. Cheng and H.-P. D. Shieh, "Colorimetric characterization of high dynamic range liquid crystal displays and its application," J. Display Technol., vol. 5, no. 1, Jan. 2009, pp. 40-45.

[6] J. Lee, G. Jeon, and J. Jeong, "Piecewise tone reproduction for high dynamic range imaging," IEEE Trans. Consumer Electron., vol. 55, no. 2, pp. 911-18, May 2009.

[7] E. Reinhard and K. Devlin, "Dynamic range reduction inspired by photoreceptor physiology," IEEE Trans. Vis. Comput. Graphics, vol. 11, no. 1, Jan. 2005, pp. 13-24.

[8] Y. H. Yee and S. Pattanaik, "Segmentation and adaptive assimilation for detail-preserving display of high-dynamic range images," Visual Comput., vol. 19, no. 7-8, Dec. 2003, pp. 457-466.

[9] F. Drago, K. Myszkowski, T. Annen, and N. Chiba, "Adaptive logarithmic mapping for displaying high contrast scenes," Comput. Graph. Forum, vol. 22, no. 3, Sep. 2003, pp. 419-426.

[10] E. Reinhard, M. Stark, P. Shirley, and J. Ferwerda, "Photographic tone reproduction for digital images," ACM Trans. Graph., vol. 21, no. 3, Jul. 2002, pp. 267-276.

[11] J. Lee, G. Jeon, and J. Jeong, "Chromatic adaptation-based tone reproduction for high-dynamic-range imaging," Optical Eng., vol. 48, no. 10, Oct. 2009, pp. 107002.

[12] G. Guarnieri, S. Marsi, and G., Rampon "High dynamic range image display with halo and clipping prevention," IEEE Trans. Image Process., vol. 20, no. 5, May 2011, pp. 1351-1362.

[13] C. Ha, W. Kim, and J. Jeong, "Remote sensing image enhancement based on singular value decomposition," Optical Eng., vol. 52, no. 08, Aug. 2013, pp. 083101.

[14] C. Ha, G, Jeon, and J. Jeong, "Contrast enhancement and noise elimination using singular value decomposition for stereo imaging," Optical Eng., vol. 51, no. 09, Sep. 2012, pp. 090504.

[15] H. Demirel, C. Ozcinar, and G. Anbarjafari, "Satellite image contrast enhancement using discrete wavelet transform and singular value decomposition," IEEE Geosci. Remote Sens. Lett., vol. 7, no. 2, Apr. 2010, pp. 333-337.

[16] D. Kalman, "A Singularly valuable decomposition: The SVD of a matrix," Coll. Math. J., vol. 27, no. 1, 1996, pp. 2-23.

[17] ITU, *Parameter values for the HDTV standards for production and international program exchange*, ITU-R recommendation BT.709, 1998.

[18] G. Qiu, J. Duan, and G. D. Finlayson, "Learning to display high dynamic range images," Pattern Recognition, vol. 40, 2007, pp. 2641–2655.

 ISBN: 978-1-61208-320-9

Making a Travel Diary from GPS Traces Using an Area-Based Reverse Geocoder

Kiichi Hikawa
Department of Computer Science and Engineering
Graduate School of Engineering
Nagoya Institute of Technology
Gokiso–cho, Syowa–ku,Nagoya–shi, Aichi, 466–8555 Japan
Email: hikawa@moss.elcom.nitech.ac.jp

Daisuke Yamamoto
Department of Computer Science and Engineering
Graduate School of Engineering
Nagoya Institute of Technology
Email: daisuke@nitech.ac.jp

Naohisa Takahashi
Department of Computer Science and Engineering
Graduate School of Engineering
Nagoya Institute of Technology
Email: naohisa@nitech.ac.jp

***Abstract*—This paper presents a system for creating a travel diary that shows a list of the locations visited on a trip with durations as well as a travel trace on a map. It also describes the implementation methods for the system. Proposed system determines the visited locations and infers the names of the sites and managed places with boundaries such as fences and roads, including the locations, from a GPS trace using an area-based reverse geocoder (ARG) newly developed for the system. Because it infers the names of sites using a site-boundary database, the ARG is more precise than conventional reverse geocoders, particularly when multiple sites are close to a portion of the travel trace. Proposed system introduces a sequence matching method that simultaneously matches consecutive GPS points with a boundary provided by the ARG. Because it uses site boundaries to find a sequence of noisy GPS points across the boundaries, the method effectively reduces the impact caused by GPS noise.**

***Keywords*—Web map, Reverse Geocoder, GPS, Travel Diary , Travel Trace, Area Database**

I. INTRODUCTION

Using embedded GPS (Global Positioning System) receivers, many mobile devices make it possible for users to record their locations and make travel tracing easy. A user can use GPS data with many GIS (Geographic Information System) applications, and share travel traces with other users. By storing GPS traces and sharing them with others, users can recall past travels, understand the travels of other people, and make new travel plans.

Several GIS applications draw a GPS trace as a line that follows the GPS points over a map view and display it on a screen. We have developed systems displaying travel traces [1]–[3]. These help users to easily see an overview of a trip. However, it can be difficult to grasp detailed information such as the name of a place visited (‘ stayed at ’) from the GPS trace on a map alone. In several studies, determining significant places from GPS traces has been addressed [4]–[6].

A GPS trace with a list of locations and visit (‘ stayed ’) durations for a trip will help users better understand their travel. The names of the sites, including the locations, are essential. Although conventional reverse geocoders provide the identification of a location, they provide only an address corresponding to the location and a list of names of sites close to that location. In other words, they cannot provide the name of a site including the specific location precisely.

Many WEB map services such as Google Map [7] provide the functions of a geocoder and reverse geocoder [8]. They can connect locations to sites, areas managed by someone and surrounded by boundaries like fences and roads, such as precincts of shrines, farms, and parks, which travelers may visit. They have pairs that include a site name and a representative point in their database. Given the name of a site, a geocoder will return the coordinates of the site-representative point, a latitude and longitude pair. On the other hand, given the coordinates of a point, a latitude and longitude pair, a reverse geocoder will return a list of candidate site names with the sites sorted by the distances between the point and the representative point of the sites. These functions are useful to find a site on a map. They have a problem, however, finding the name of a site precisely. They connect the name of a site to the coordinates of the site-representative point. Given the coordinates of a point inside a site, they may return the name of a different site from the site including the point, one that is close to the point.These problems occur when the point is inside a large site and close to its boundary when another small site is close to the point.

To resolve this problem, we have developed a site-boundary database and an area-based reverse geocoder, called ARG that uses this database. The site-boundary database connects a site to its name and boundary. Given the coordinates of a point, the ARG returns the name and boundary of the site that includes this point. In general, a significant effort is required to create a site-boundary database because it is more difficult to capture boundary data than representative point data for sites. To compensate for the lack of boundary data, we have developed a method for approximating this automatically as follows. The method calculates a loop road [9]–[11], a path surrounding the representative point of the site in the road network, and uses it as an approximate boundary of the site. If other site-boundaries overlap the loop road, it removes the overlapped areas from the loop road to improve the approximation.

The objective of our research is to precisely determine the sites visited on a trip with durations in each site using a GPS trace, and generate a travel diary list including a travel trace

 ISBN: 978-1-61208-320-9

on a map. The travel diary enables users to understand the trip intuitively and precisely.

The remainder of this paper is organized as follows.First, we describe the systems used to make travel diaries from GPS points in Section Ⅱ.We propose a system for making a precise travel diary using ARG in Section Ⅲ.In Sections Ⅳ, we describe the implementation methods for two key aspects of the proposed system, an area-based reverse geocoder and a sequence matching method, which matches a sequence of GPS points with the boundary of a site.We describe an experimental prototype of the proposed system in Section V and evaluate the proposed system by comparing the ARG and a conventional reverse geocoder through experiments determining locations and durations from GPS traces in Section VI.We conclude our presentation in Section Ⅶ.

II. Making a travel diary from GPS traces

A. Overview

We propose a system for generating a travel diary that analyzes GPS data that a traveler records during a trip, and shows the locations visited. It precisely determines the sites visited on the trip and durations spent in the sites from a GPS trace, and shows a travel diary listing with these sites as well as a travel trace on a map, as shown in Fig.1.

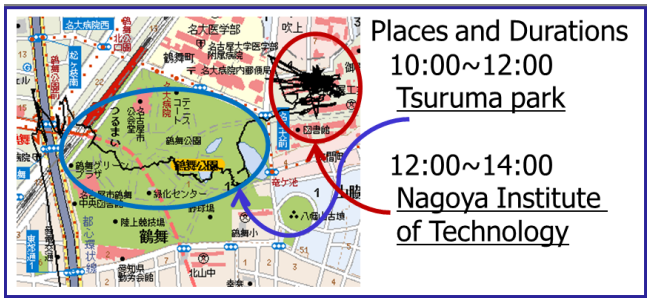


Fig. 1. Travel diary

B. Implementation Issues

In making a travel diary, extracting the portion of the GPS trace, that was recorded during the stay in the site and detecting the visited sites (‘ stayed ’) is important. Conventionally, reverse geocoders can be used for transforming a point in a GPS trace into the name of the location corresponding to the point.The duration that the traveler visited (‘ stayed at ’) the site is calculated from the sequence of consecutive GPS points corresponding to the site.

A GPS trace is a sequence of time-stamped points, for example, $P_i(i = 1, 2, ..., n)$, a 4-tuple of latitude, longitude, altitude, and time, that represents the geographic location, latitude, longitude, altitude, and the time the point was logged. Using a reverse geocoder, we can select the site shown by the reverse geocoder nearest to the GPS point of the representative point. This means that we can define a Voronoi diagram using the center coordinates (site representative points) of all the sites. We can then divide the area including the sites shown by the reverse geocoder into sub-areas determined according to the Voronoi diagram, and set boundaries by considering each sub-area as a site. Point matching classifies to which sub-area each point in the GPS trace belongs.

The above method may be unable to estimate the locations and durations of visits precisely for the following reasons. Consider the example of the GPS data recorded during the movement in the neighborhoods of sites A, B, and C as shown in Fig. 2. In this figure, X marks show the site-representative point and the dotted lines show the Voronoi diagram defined by the site-representative point. In this case, the sequence of the GPS points passes inside sites A and C. Although the sequence does not pass inside site B, point matching with a conventional reverse geocoder will record site B in a travel diary because it determines that the three GPS points in the middle of the sequence are inside the area of site B according to the Voronoi diagram as shown in a Fig. 2.

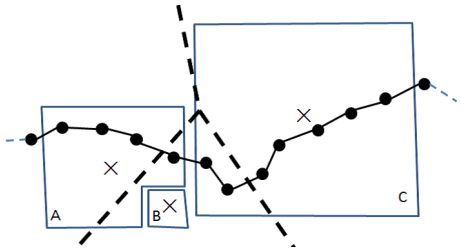


Fig. 2. Representative point of sites and Voronoi diagram

Figure 3 shows an example of a GPS trace that may include noise caused by large buildings or by narrow roads. It shows the movement between sites B and C. When moving frequently near the boundaries of the sites, as shown in this figure, point-matching will provide the location of a visit list showing that the traveler goes back and forth between sites for short intervals, such as site C→site A→site C→site A→site C. Erroneous position coordinates may be included in the GPS trace data because of noise. When a sequence fluctuates between sites, we can say that the sequence includes noise and the result of the point matching has a high possibility of being incorrect. This is because in many cases there are physical boundaries such as fences and roads between two adjacent sites and the traveler cannot cross these boundaries so frequently.

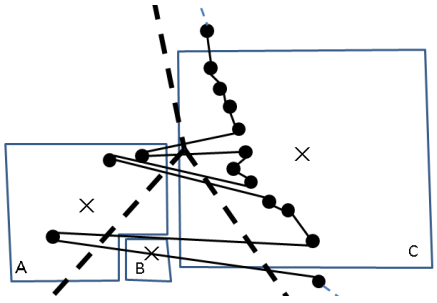


Fig. 3. GPS points which go to and return near a boundary

The requirements to implement a travel diary system are summarized as follows.

Requirement 1: It is required to solve the above-mentioned problem and to be able to estimate visited locations with high accuracy, even when two or more sites border.

 ISBN: 978-1-61208-320-9

Requirement 2: Because the results of matching a noisy sequence with site boundaries can be erroneous, it is necessary to implement a matching method that can reduce the impacts caused by GPS noise.

III. PROPOSED SYSTEM

In order to satisfy the above requirements, we propose a travel diary system using GPS traces with the following features.

Feature 1 To satisfy Requirement 1, the system should estimate the location of visits and determine the name of the sites, managed places with boundaries such as fences and roads, including the locations from a GPS trace, using the area-based reverse geocoder (ARG) newly developed for this system. ARG is more effective than conventional reverse geocoders, particularly when multiple sites are close to some portion of the travel trace, because it determines the name of the sites more precisely using a site-boundary database.

Feature 2 To satisfy Requirement 2, the system should introduce a sequence matching method that simultaneously matches consecutive GPS points with a boundary provided by the ARG. The method must be able to reduce the impact caused by GPS noise, by using site boundaries to find a sequence of noisy GPS points across the boundaries.

The proposed system consists of a site-boundary database and the functions shown in Fig.4.

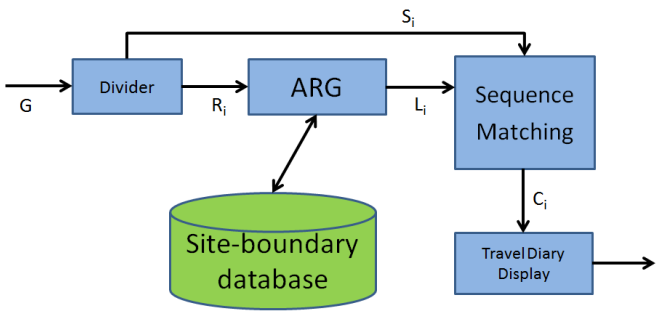


Fig. 4. Structure of the Proposed System

Divider: This function transforms a GPS log file, G, into a GPS trace, a sequence of points with five attributes, latitude, longitude, time, address, and name, where the first three attributes are set by copying the data of the corresponding point in the GPS log file and the last two attributes are empty. It also identifies an address for each point of the GPS trace using the ARG, and attaches it to that point as its attribute. Consecutive points having the same address attribute are gathered and formed into a sequence, S_i, where i=1,2,...,n and n is the number of sequences. The Divider calculates the minimal rectangle, R_i, which includes all the points of S_i for each i=1,2,...,n.

ARG: By searching a site-boundary database, this function obtains a list set of L_i, area data, where each entry is a pair of site name and boundary, for each set of all the sites in R_i for i=1,2,..,n.

Sequence Matching: This function matches the points in S_i with the boundaries stored in L_i and creates a chunk C_i of consecutive points with the three attributes, name, address, and duration. The name and the address attributes show the name and the address of the site that the traveler visited (' stayed at'). The duration attribute shows a time that the traveler stayed within the site.

Travel Diary Display: Using the chunks from the sequence matching, this function creates and displays a list of records, each a pair of the name of the site where the user visited and the time spent at the site.

Using these functions, the proposed system creates a travel diary from a GPS log file, G, by the following steps:

STEP1: The Divider function divides G into multiple sequences, $\{S_i\}$, and calculates a minimal rectangle R_i for each S_i in $\{S_i\}$.

STEP2: A list of the area data for all of the sites in R_i, L_i, is generated using ARG.

STEP3: Using S_i and L_i, the Sequence Matching builds a sequence of chunks, C_i.

STEP4: The Travel Diary Display function displays the attributes, a triplet of duration, site name, and address for each chunk in C_i.

IV. IMPLEMENTATION OF THE PROPOSED SYSTEM

A. Implementation of Area-Based Reverse Geocoder

In order to create ARG, it is necessary to gather significantly more site area boundary data than is found in a typical site database. In a site database, many site area records include only a pair of name and representative point. Only a small number of site area records have boundary data as well. For the site records without boundary data, we build site boundaries automatically by calculating a loop road [9]–[11], a path surrounding the representative point of the site in the road network, and consider the boundary of the site area to be the loop road. If an area is surrounded by a loop road and used for multiple purposes, such as a school and park, we divide it into smaller areas using land usage data from a GIS database. These methods can generate boundary data for any site data consisting of a site name and its representative point.The result is that we can create a site-boundary database, a set of site data consisting of a site name and its boundary. We implement an area-based reverse geocoder (ARG) using this database and use it to generate our travel diary.

The site-boundary database has site data consisting of site name, boundary, representative point, and purpose.A boundary is shown by a polygon and the boundary data is a sequence of position coordinates of a polygonal vertex.

A1-A16 of Table I are classified according to the existence of each component of the site data. If data exists, it is indicated with the mark, ○, otherwise, it is marked, ×. Even when there is no data, it can be created easily from other data. For example, a representative point can be created from boundary data. This is expressed with the mark, △ .

In order for ARG to output a site name, a representative point, boundary, and purpose, such as sites A1-A4 of TableI, is required. Data such as A1 and A2 can be created by combining A5 and A6 that does not have boundary data, although it has a site name, with A9-A12 that has boundary data, but does not have a site name. That is, A5+A12 → A1 and A6+A12 → A2. In the case of overlapping boundaries where the use differs, a new boundary is generated by removing the overlapping portion.

 ISBN: 978-1-61208-320-9

TABLE I. SITE DATA

	Site name	Boundary	Representative Point	Purpose	Example of Data
A1	○	○	○	○	Tourist attractions (867 places in Japan)
A2	○	○	○	×	Address
A3	○	○	△	○	Named land usage data
A4	○	○	△	×	
A5	○	×	○	○	Public facility (9082 places in Aichi prefecture). Park (3938 places in Aichi prefecture)
A6	○	×	○	×	
A7	○	×	×	○	Shopping street, Busy street. Business district
A8	○	×	×	×	
A9	×	○	○	○	
A10	×	○	○	×	
A11	×	○	△	○	Unnamed land usage data
A12	×	○	△	×	Loop road
A13	×	×	○	○	
A14	×	×	○	×	
A15	×	×	×	○	
A16	×	×	×	×	

TABLE II. DEFINITION OF THE AREA DATA TABLE

column	type	description
id	integer	ID of the area data
name	text	site name of area data
lat	real	latitude of representative point
long	real	longitude of representative point
north	real	northernmost latitude of area data
south	real	southernmost latitude of area data
west	real	westernmost longitude of area data
east	real	easternmost longitude of area data
coordList	text	list of coordinates

The site-boundary database is indexed by the columns north, south, west, and east, in TableII. Given the coordinates of a point *p*, a pair of latitude *p*.lat and longitude *p*.long, the ARG returns an area $\bar{a}$ that includes this point from set A of all areas in the site-boundary database using the following STEPs.

STEP1: S1 = { $a \in A \mid a.\text{west} < p.\text{long} \wedge a.\text{east} > p.\text{long} \wedge a.\text{north} < p.\text{lat} \wedge a.\text{south} > p.\text{lat}$ }

STEP2: Obtain the area data *a* ,which includes the point *p* , from the all elements in S1

B. Implementation of Sequence Matching

Point matching methods match each point of the GPS trace with a site boundary one at a time, and determine to which area they belong. Using ARG, methods simultaneously match a sequence of consecutive GPS points with a boundary provided by ARG and determine in which site each point of the sequence belongs. This is based on the hypothesis that many sites have boundaries that separate the inside and outside clearly, such as a fence and a road. When the sequence of consecutive GPS points cross the boundary provided by the ARG frequently, we consider that the crossing of the boundary is caused by the influence of GPS noise. When a sequence of GPS points crosses a road frequently as shown in Fig.5, the sequence matching method judges that it is an influence of noise. It

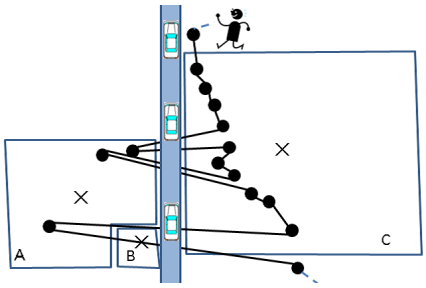


Fig. 5. GPS points that go and return near a boundary

determines that the minority points, points in site A in this figure, belong to the same site as the majority points, the points in site C in the figure, if the number of minority points is significantly smaller than the majority points.

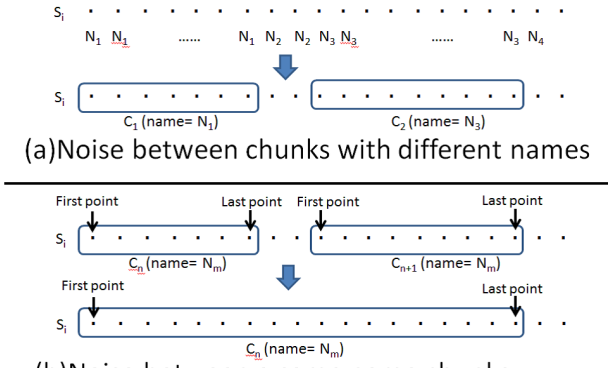


Fig. 6. Noise removed

In Fig.6 (a), given a sequence of points, S_i, the sequence matching function identifies a site name for each point of S_i, and attaches it to that point as the name attribute, N . It organizes the points of S_i by their name attribute. Consecutive points with the same attribute are gathered and formed into a chunk, C_i. The same name and address attributes as those of the gathered points is attached. When a chunk has only a few points, it is regarded as noise and discarded.In Fig.6 (b), when the name attribute of a chunk is the same as that of an adjacent chunk discarded in the previous step, these chunks are gathered and formed into a single chunk. The sequence matching calculates the duration stayed at the chunk for each chunk of C_i. Assume the time of the first point of C_i is t1, and the time of the last point of C_i is t2, the duration *d* is calculated by *d* = t2 - t1 . We introduce *d0* as the minimum duration of a stay in a chunk because we regard a short length chunk as noise. If $d < d0$ then C_i is considered as noise. If $d > d0$, we set the value of the duration attribute of C_i to *d*.

V. PROTOTYPE SYSTEM

We implemented a prototype system of the proposed method using JAVA. The database used for the prototype system was MySQL. The following data was used as site data: address data（A2）and land usage data (A3, A11) from AlpsMAP [12], public facility data (A5) from the Ministry of Land, Infrastructure and Transport [13], and loop road (A12). The map picture used is from AlpsMAP.

The prototype system loaded the GPS trace data shown in Fig.7. It drew a GPS trace (①) as a line that followed the

 ISBN: 978-1-61208-320-9

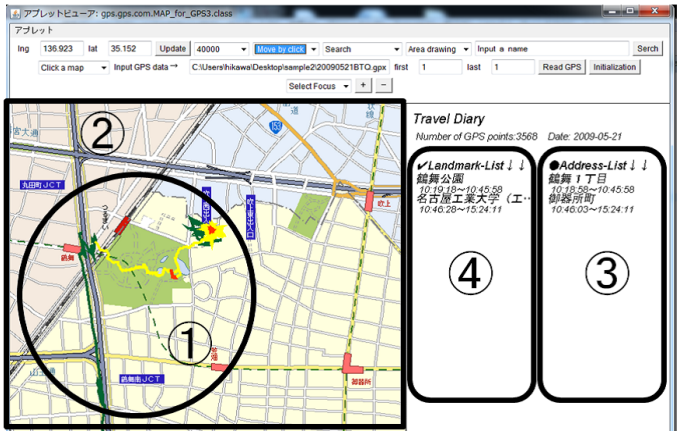


Fig. 7. Prototype System

GPS points over a map (②) on the left-hand side of the applet screen.
The list of locations and durations of each visit is shown on the right-hand side of the applet screen. First, the list of addresses is displayed on the extreme right-hand side (③). Then, the names of the sites are displayed, one by one, on the left-hand side (④) of the address name of the corresponding address. Nothing is displayed on the left-hand side of the address when a site is not found.

VI. EVALUATION

We evaluated the accuracy of the locations and durations of each visit that the proposed system determined to generate the travel diary. We checked the list of the locations and durations of each visit that the prototype system created, and verified whether the site at which the traveler stayed was detected. This evaluated the usefulness of our ARG.

A. The Method of an Experiment

We incorporated the following two methods into the prototype system for evaluation purposes.

ARG-Method（Proposed method）: Using our ARG, it determines the location and duration of the visit.

RG-Method（Baseline）: Using a conventional reverse geocoder, it determines the location and duration of the visit. RG-Method simply selects the site shown by the reverse geocoder nearest to the GPS point of the representative point.

In both method, we set d0 in the same value, which introduced as the minimum duration of a stay. We used the thirty GPS traces that were recorded by moving inside Aichi.

First, we drew each GPS trace on a map, and showed this to the traveler. Next, in the GPS trace, the traveler reported the name of the site visited, and identified point P1 showing the entrance and point P2 showing the exit. We defined the reported site as the Correct Answer Site, that was actually visited. We defined the difference of the time between P2 and P1 as the duration Correct Answer Duration. When a traveler visited (‘ stayed at ’) in the same site more than once, we appended a numerical value showing the number of times stayed at the site name, and regarded it as a different site.

Next, using the ARG-Method (proposed method) and the RG-Method, we detected visits or stay sites, which we call Detect Sites. We determined the time stayed at these sites, which we call Detect Duration. When a traveler stayed in the same site more than once, we appended a numerical value to the site name, as we did with the Correct Answer Site and regarded it as a different site. Moreover, similar to the Correct Answer Duration, in the GPS trace, we determined P3, which was the first point and P4, which was the last point, in the Detect Site. We defined the difference of the time between P4 and P3 as the duration Detect Duration, that is, the time that the traveler stayed in the Detect Site.

Experiment 1: We obtained the following sets with each method from the GPS data.

V_1 : Set of Detect Sites

V_2 : Set of Correct Answer Sites

$V_3 = V_1 \cap V_2$

We obtained the number n_1,n_2,n_3 of elements of the sets V_1,V_2,V_3 and calculated P and R using the following formulas:

Precision (p_v) $= n_3/n_1$

Recall (r_v) $= n_3/n_2$

We calculated F_v using the following formula:

$F_v = 2 / (1/r_v + 1/p_v)$

Experiment 2: In the GPS trace, we evaluated the validity of the Detect Duration detected by the two methods by comparing the GPS points in Correct Answer Duration with the GPS points in Detect Duration as follows. We obtained the following sets that consisted of the chunks showing the duration for each site from each site contained in the set V_3 of Experiment 1 as follows:

T_1 : Set of GPS points in Detect Duration

T_2 : Set of GPS points in Correct Answer Duration

$T_3 = T_1 \cap T_2$

We obtained the number m_1,m_2,m_3 of elements of the sets T_1,T_2,T_3 and calculated P and R using the following formulas:

Precision (p_t) $= m_3/m_1$

Recall (r_t) $= m_3/m_2$

We calculated F_t using the following formula:

$F_t = 2 / (1/r_t + 1/p_t)$

B. Results

The results of Experiment 1 for the thirty GPS traces are shown in FIg.8, where the horizontal axises are GPS trace numbers and the vertical axises are the Precisions, Recalls and F values. It shows that the proposed ARG-Method detected the stay sites correctly with Precision=1 and Recall=1 for the GPS traces numbers 1, 2, 7, 8, 19, 24, 27, and 28. However, the Precision and Recall are low for the GPS trace numbers 4, 5, 9, 11, 13, 14, 15, 20, 22, 25 and 26. On the other hand, although the RG-Method detected all the stay sites (Recall=1) for the GPS traces numbers 2, 7, 9, 11, 12, and 19, the Precision is very low. We can say that the RG-Method provides low precision and the RG-Method detected a site that was not actually visited for many GPS traces. Comparing the F_v

 ISBN: 978-1-61208-320-9

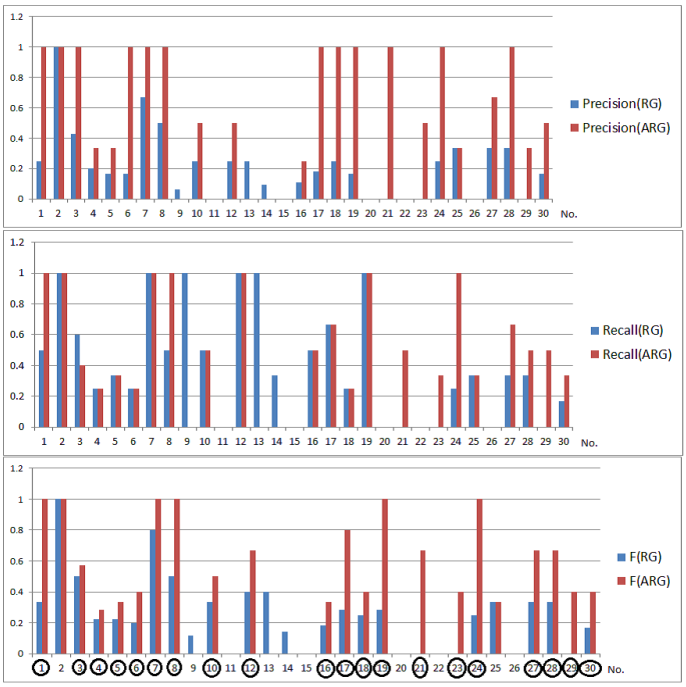


Fig. 8. Results of Experiment 1

values, the ARG-Method is better than the RG-Method for many GPS traces. Based on this, in this experiment, the results indicate that the stay site accuracy determined by the ARG-Method is higher than that of the RG-Method. However, the Precision, Recall, and F value were also low for the ARG-Method in some cases. This was caused by the following: (1) Because there was only a small amount of data in the database and there was no area data for a stay site, the system could not detect the site. (2) When two or more sites exist in one area, the system could not detect a stay site correctly. In other words, to improve the effectiveness of the site detection of the proposed method, it is necessary to expand the site data database. We present the results of Experiment 2 in Table III. As shown in Table III, the proposed ARG-Method recorded

TABLE III. RESULTS OF EXPERIMENT 2

No	p_v		r_v		F_v	
	RG-Method	ARG-Method	RG	ARG	RG	ARG
Average	0.883	0.951	0.779	0.857	0.828	0.902

a higher value for Precision, Recall and F value, compared to the RG-Method. This shows that when the ARG-Method identifies a site correctly, it can calculate the duration very accurately. The value of Recall for the RG-Method is low, and even when a stay site is detected correctly, there is a tendency to calculate a shorter duration than actual. By comparing the F_t values in Table III, we determined that the accuracy of the duration calculated by the ARG-Method is higher than that of the RG-Method. The above experiments with thirty GPS data showed that the proposed method is effective. However we must evaluate the proposed method with more GPS data in order to clarify its effectiveness.

VII. CONCLUSIONS

In this paper, we presented an area-based reverse geocoder (ARG) and described an implementation method. Moreover, we presented a system using ARG to create a travel diary and described the implementation method. Furthermore, we compared and evaluated the proposed system and a system using a conventional reverse geocoder, using locations and visit durations of a system-generated travel diary. As a result, we were able to derive the sites and durations at which the user stayed from the GPS trace with high precision using the ARG-Method, better than when the conventional reverse geocoder was used. There were. however, cases where a site was undetectable. To resolve this, we must expand the site data, which includes the area data. The subject of future research follows. When the representative point of two or more sites exist in the boundary of an area, it is difficult for ARG to show the exact position of the site. When detecting the site as ‘ stayed ’ , a device using a combination of ARG and reverse geocoding by distance of points, is necessary.

ACKNOWLEDGMENTS

We would like to thank Yahoo! Japan Corporation for supporting us in the development of the prototype system. This work was also supported by JSPS KAKENHI 23500084 and 25700009.

REFERENCES

[1] Pablo Martinez Lerin, Daisuke Yamamoto, and Naohisa Takahashi “Making a Pictorial and Verbal Travel Trace from a GPS Trace”, Proceedings of the 11th International Symposium on Web and Wireless Geographical Information Systems (W2GIS 2012), LNCS 7236, Springer, pp. 98-115 , 2012.

[2] Pablo Martinez Lerin, Daisuke Yamamoto, and Naohisa Takahashi, “Inferring and Focusing Areas of Interest from GPS traces”, Proceedings of the 10th International Symposium on Web and Wireless Geographical Information Systems (W2GIS 2011), LNCS 6574, Springer, pp. 176-187 , 2011.

[3] Pablo Martinez Lerin, Daisuke Yamamoto, and Naohisa Takahashi, “ Pace-Based Clustering of GPS Data for Inferring Visit Locations and Durations on a Trip,” IEICE Transactions on Information and Systems, 2014 to be published.

[4] YU ZHENG and XING XIE , “Learning Travel Recommendations from User-Generated GPS Traces”, ACM Transaction on Intelligent Systems and Technology, Vol. 2, No. 1, Article 2, 2011.

[5] Changqing Zhou, Dan Frankowski, Pamela Ludford, Shashi Shekhar, and Loren Terveen, “Discovering personally meaningful places: An interactive clustering approach”, ACM Transaction on Intelligent Systems and Technology, Vol. 25, No. 3, Article 12, 2007.

[6] Nobuharu Kami, Nobuyuki Enomoto, Teruyuki Baba, and Takashi Yoshikawa, “Algorithm for Detecting Significant Locations from Raw GPS Data”, Discovery Science Lecture Notes in Computer Science Vol. 6332, pp 221-235, 2010.

[7] GoogleMaps, January 2014. http://maps.google.co.jp/

[8] Google Maps API Web Services, January 2014. https://developers.google.com/maps/documentation/geocoding/

[9] Hiroki Ito,Daisuke Yamamoto, and Naohisa Takahashi, “Creating a database of roads around the object and method of fast derivation of the roads around the object by using database”, DEIM Forum 2011, B4-3, 2011.

[10] Hiroki Ito, Daisuke Yamamoto, and Naohisa Takahashi, “A Loop-Road Database Building Method Supporting Multiple Meshes” ,The 74th National Convention of IPSJ ,1P-6, 2012.

[11] Daisuke Yamamoto, Hiroki Itoh, and Naohisa Takahashi, “One Click Focusing: An SQL-based Fast Loop Road Extraction Method for Mobile Map Service”, Proceedings of the 4th International Conference on Advanced Geographic Information Systems, Applications, and Services (GEOProcessing 2012), pp. 7-16, Valencia, Spain, 2012.

[12] AlpsMAP, November 2013. http://alpsmap.yahoo.co.jp/

[13] National Land Numerical Information Download Service, November 2013. http://nlftp.mlit.go.jp/ksj/

 ISBN: 978-1-61208-320-9

Secure and Anonymous Multimedia Content Distribution in Peer-to-Peer Networks

Amna Qureshi, Helena Rifà-Pous, David Megías
Department of IT, Multimedia and Telecommunications, Internet Interdisciplinary Institute,
Universitat Oberta de Catalunya,
Barcelona, Spain
{aqureshi, hrfia, dmegias}@uoc.edu

Abstract—**In recent years, the distribution of large scale multimedia contents has become easier and more efficient than ever before. The unauthorized distribution of copyright-protected content has emerged as a major concern. A buyer of a multimedia content does not want to reveal his/her identity whereas the seller or owner of the content does not want the buyer to further distribute the content illegally. Accordingly, this paper presents a secure and privacy-aware multimedia content distribution mechanism based on Tardos Codes for Peer-to-Peer networks. Our proposed system offers copyright protection by means of a traitor-tracing capability, collusion-resistance, content owner and buyer's security with mutual anonymity during secure data exchange.**

Keywords-privacy; security; anonymity; trust; collusion-resistant fingerprinting;Peer-to-Peer networks

I. INTRODUCTION

Peer-to-Peer (P2P) is often described as a type of decentralized computing system in which nodes, referred to as peers, use the Internet to communicate with each other directly. All the peers in this interconnected network provide resources to other peers, including bandwidth, storage space, and computing power. It greatly helps to establish multimedia communication networks and enables efficient distribution of digital content at a large scale. However, in the absence of security mechanism, P2P networks lack the ability to maintain content protection and access control towards copyright material, hence leading to piracy commitment and copyright infringement. The enforcement of copyright protection mechanisms in P2P content distribution systems, however, poses serious privacy threats to end users, i.e., monitoring of their activities by keeping record of downloaded files, IP address through which files are being downloaded, history of files shared or downloaded, or a list of the peers with whom a user has interacted in the past. Thus, there is an inherent conflict of interest between copyright protection supporters and privacy advocates. P2P developers need to balance security and privacy needs when designing content distribution systems.

However, security and privacy in P2P content distribution systems are achieved by various cryptographic mechanisms, (Homormorphic encryption, multi-party computation, etc) and digital watermarking/fingerprinting technology. Most of the research work related to copyright protection and privacy in content distribution system incurs high computational and communicational burden at the buyer's and/or at the multimedia owner's/merchant's end. Literature review shows that a few P2P content distribution systems [1-2] exist that provides both secure and anonymous content distribution with reduced computational burden at merchant's end. The review of some recent P2P content distribution systems satisfying either security or privacy properties can be found in [3]. Hence, it is of utmost importance to build an integrated platform which can enable content owners' to distribute their large-sized digital contents without a fear of copyright violation at reduced delivery costs and simultaneously enable end users to receive legal content without fear of privacy breach.

In this paper, we propose a secure and anonymous content distribution system for P2P networks that provides both copyright protection and privacy. Also, the proposed system reduces the computational cost of a merchant.

A. Contributions

A newly proposed privacy-aware content distribution scheme for P2P networks which would benefit multimedia owners (merchants) to share their large-sized copyrighted files and end users/peers (buyers) to download their content legally. In the proposed system, the original multimedia file is partitioned by the merchant into a base and supplementary file. The base file is much smaller than the original file and contains the most important information. Without this information, the file reconstructed by the end user would be unusable. The base file is dispensed by the merchant on payment from the user and a supplementary file is distributed to the P2P network. Thus, it reduces the burden of the merchants by only sending the small-sized base file and making use of the P2P network infrastructure to support most of the file transfer process. In this paper, we adopt a collusion-resistant Tardos fingerprint code [4] to confirm the distribution scheme's security against collusion attacks. A Trusted Third Party (TTP) Monitor is used, which is assumed to be an honest entity and is trusted by both the merchant and the buyer. Since Tardos codes used by a merchant for the generation of collusion-resistant fingerprint depend on a secret vector which cannot be revealed to the buyer, this secret vector is committed to a Monitor. Moreover, to provide security to a buyer from a malicious merchant who might frame an honest user for illegal-redistribution, this secret vector is committed to a Monitor along with the pseudo identities of a merchant and a buyer. The proposed scheme can protect the buyer's privacy until the illegal re-distribution is proven. The illegal re-distributor is identified by the merchant using Tardos Code's piracy tracing algorithm. Also, the scheme provides resistance to man-in-middle

 ISBN: 978-1-61208-320-9

attacks by employing mutual anonymity and key agreement and pseudonym-based trust management.

B. Outline of the paper

In this paper, we focus on a new solution to provide security, anonymity and privacy to both merchant and end-users of P2P content distribution system. In Section II, we highlight the design requirements for our system. The building blocks of our system are introduced in Section III. In Section IV, we describe our proposed system and, in Section V, we draw a conclusion and discuss future work.

II. DESIGN REQUIREMENTS AND ASSUMPTIONS

This section defines the system requirements and certain assumptions made during the design process of the system.

A. Design Requirements

For secure and privacy-aware P2P content distribution scheme, we have the following requirements depending on security, privacy, anonymity, trust, robustness and imperceptibility constraints.

- The identity of a buyer should remain anonymous during transactions until he/she is proven to be guilty of copyright violation.
- The merchant should be able to trace and identify the illegal re-distributor in case of finding a pirated copy.
- The scheme should be collusion resistant against a specific number of colluders' c as specified by Tardos codes.
- The buyer accused of re-distributing an unauthorized copy should not be able to claim that the copy was created by the merchant.
- The judge with the help of a Monitor should be able to resolve the disputes without the buyer revealing his/her identity.
- The fingerprinting process should be blind and the inserted fingerprint should be imperceptible and robust against common signal processing attacks.
- None of the other peers in the system should know about the source provider peer and requesting peer's identity or an item being exchanged.
- The real identity of a user should be protected during authentication process thus, enabling each user to verify the authenticity of each other anonymously.

B. Assumptions

The assumptions for the proposed system are as follows:

- There are four major players involved: merchant, buyer, Monitor and Judge.
- The merchant and the buyer do not trust each other but they both trust the Monitor.
- Each entity is supposed to have a public key K_p and a private key K_s.
- *SP* is selected on the basis of its reputation and resources and is thus assumed to be honest.
- Each user can have only one pseudonym against its secret information.
- The base file transfer between the merchant and the end user is secure and efficient by using the Anonymous two-party Authenticated Key Exchange (AKE) protocol.
- A P2P system maintains end-to-end data integrity. The reconstruction of the original file from the base and supplementary files is performed at the user end without assistance of a user. The supplementary file is only shared within the P2P network whereas the base file is not shareable.

III. BUILDING BLOCKS

Digital Signature Algorithm (DSA), Symmetric key cryptography, Zero-Knowledge Proof (ZKP) of identity, hash function, fingerprint collusion-resistance algorithm and Quantization Indexed Modulation (QIM) scheme are useful technologies for the proposed P2P content distribution system. In our proposed system, DSA and ZKP of identity are used by the users to authenticate each other's identity and generate a one-time session key for secure data exchange. An unforgeable and verifiable pseudonym for each entity of the system is generated by using standard hash function. Symmetric key encryption is used to encrypt and decrypt the data exchanged anonymously between the two parties. Tardos Codes are used in our system in order to provide collusion-resistant against colluders and these fingerprints are embedded into the content using a QIM based watermarking scheme.

A. Digital Signature Standard (DSS)

The DSS [5] is the Digital Signature Algorithm (DSA) which is used to generate digital signature for the authentication of an identity of the message sender and data integrity. The algorithm works in conjunction with the Secure Hash Algorithm (SHA). DSA uses the following three parameters which are publicly known:

- P: a large prime number (at least 1024 bits)
- Q: a sufficiently large prime number (at least 160 bits) that is also a divisor of (p-1)
- g: a generator for the multiplicative subgroup of order Q of integers modulo P

A DSA private key is an integer x taken modulo Q and the public key is the integer $y = g^x \bmod P$.

B. Symmetric key cryptography

To prevent illegal copying of digital content, the content is generally encrypted using a symmetric key algorithm. Symmetric key encryption is a cryptography technique that uses a shared secret key to encrypt and decrypt data. Symmetric encryption algorithms (AES, DES, etc) are very efficient at processing large amounts of information and computationally less intensive than asymmetric encryption algorithms (RSA, El Gamal, etc).

C. Zero-knowledge proof of identity

Zero-knowledge proof of identity [6] system is a cryptographic protocol between two parties whereby, the first party wants to prove his/her identity to the second party, without revealing anything about his/her identity to the

 ISBN: 978-1-61208-320-9

second party. Following are the three main properties of zero knowledge proof of identity:

1) Completeness: The honest prover convinces the honest verifier that the secret statement is true.

2) Soundness: Cheating prover can't convince the honest verifier that a statement is true (if the statement is really false).

3) *Zero-knowledge*: Cheating verifier can not get anything other than prover's public data sent from the honest prover.

In our proposed scheme, we employ zero-knowledge proof of identity scheme of [7] in Section IV. In this algorithm, the central authority being used in basic ZKP of identity protocol [6] has been eliminated to make it adaptable to P2P networks.

D. SHA-1

SHA-1 is a secure hashing algorithm which is used to output a 160-bit message digest of any input file less than 2^{64} bits. The SHA-1 hash is called secure because it is computationally infeasible to find a message which corresponds to a given message digest, or to find two different messages which produce the same message digest

We employ SHA-1 to use it with the DSA as specified in the DSS. The initiator and receiver of a message compute and verify a digital signature using SHA-1. In the spirit of Pseudo-Trust [7], we use a one-way function to generate pseudonyms together with a proof.

E. Tardos Codes

A variation of Tardos codes, i.e., Nuida's codes [8] are used in the proposed system for fingerprint generation. These codes are based on the Marking assumption, i.e., set of colluders *c* can only alter those bits of the codeword that differ between colluders. The fingerprinting code *F* and a secret vector *s* are outputs of this algorithm. The details of Nuida's codes construction and traitor tracing algorithm can be found in [8].

F. Quantization Indexed Modulation (QIM)

QIM [9] is a relatively recent watermark embedding technique. It has become popular because of the high watermarking capacity and the ease of implementation. The basic QIM scheme embeds a fingerprint bit *f* by quantizing a Discrete Wavelets Transform (DWT) coefficient *W* by choosing between a quantizer with even or odd values, depending on the binary value of *f*. It is important to consider an optimal selection of the embedding quantizer step size Δ and scaling factor, so that the best tradeoff between robustness and minimum quality degradation can automatically be achieved.

IV. PROPOSED SYSTEM

The proposed privacy-aware content distribution scheme shown in Fig. 1 involves five entities:

- A merchant *M* is an entity that distributes the copyrighted content to end users (peers) in the P2P system. It is involved in fingerprint generation and embedding, base and supplementary file distribution, traitor tracing and dispute resolution.
- A peer is an entity that can either play a role of a content requester or provider. Some of the peers with additional facilities act as super-peers. A peer is involved in acquisition of a base file from the merchant, distribution of a supplementary file in the system and a dispute resolution, in case he/she is found guilty of copyright violation.
- A super peer *SP* (a.k.a index server) is a peer with additional facilities which is assigned a role of content server for a small portion of the network. Each peer on registration with P2P system may upload their files to *SP*. *SP* maintains a list of the peers connected to the network and act as central servers. However, instead of peers' address, their pseudonyms are stored. The peers send their queries to the *SP* for downloading their files of interest. The *SP* receives a supplementary file from *M*. On receiving the file, it divides the content into multiple fragments and transmits these fragments to users. Initially, *SP* is booted with a supplementary file by *M*.
- A Monitor *MO* functions as a trusted party to keep the encrypted hash of the secret vector used in the fingerprint generation with the pseudonyms of the merchant and the peer. It also keeps the record of transactions between *SP* and the peer. In case of traitor tracing, the *MO* reveals the encrypted hash of the secret vector to *M*.
- A judge *J* is assumed to be a trusted party which resolves the disputes between *M* and a peer with the cooperation of *MO*.

Setup: The setup of the system involves two things, (1) A Hybrid structured P2P design is opted which provides efficient data search and consists of multiple coordinators called super-peers. At the system startup, the bootstrapping can be done via a well-known booting node. (2) To protect real identities of entities in the system, each entity is required to generate a pseudo identity *P* (pseudonym) and a pseudo identity certificate (*PC*) based on [7] before joining the system. A peer with *PC* can verify what it claims to be to the other party. Each peer uses its unique secret as an input for a pseudo identity generation. The complete description of the generation of *P* and *PC* can be found in [7]. On registration, each peer transfers its *P*, *PC* and public key to the connected *SP*. *P* and *PC* are used by peers for an anonymous communication in the system.

 ISBN: 978-1-61208-320-9

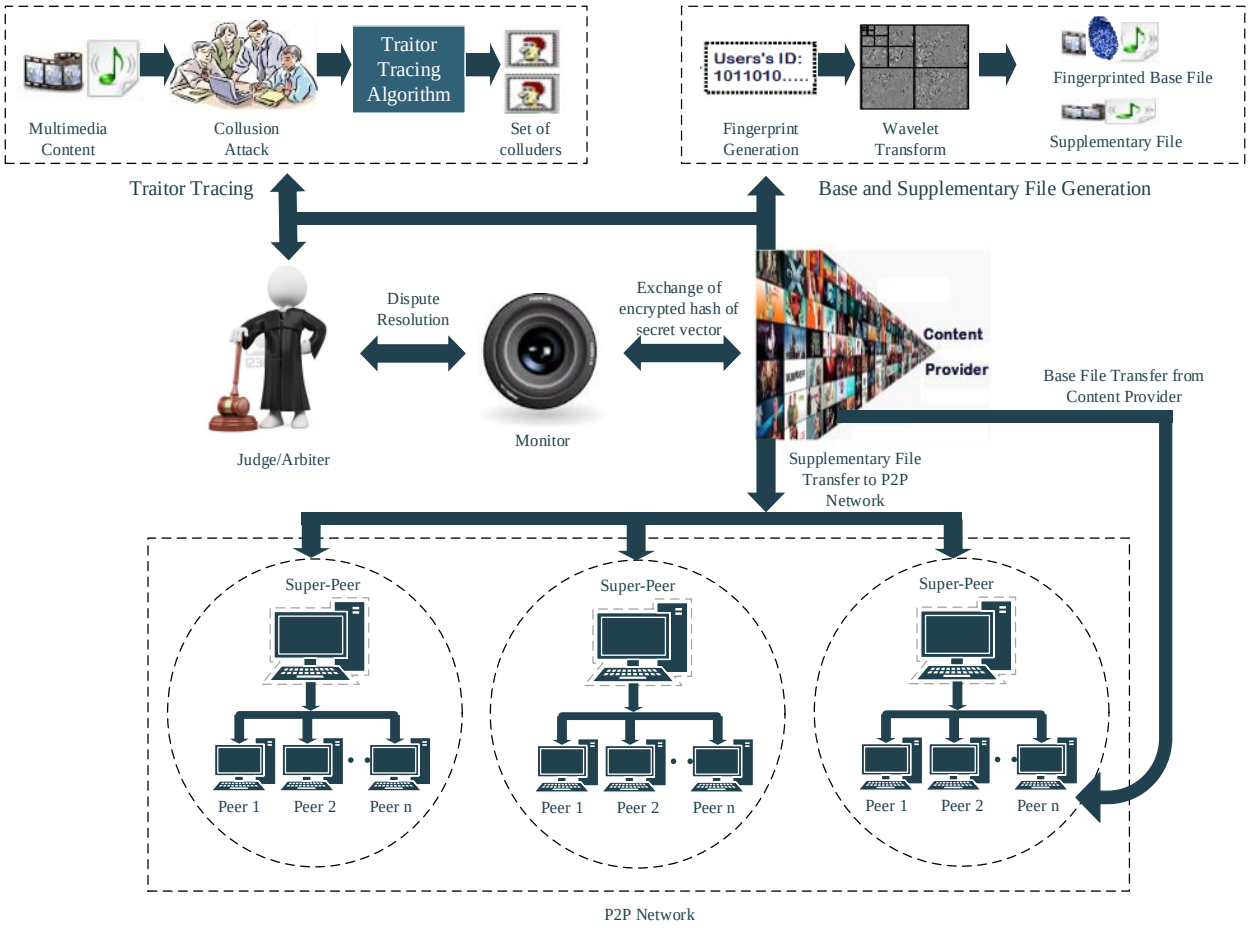


Figure 1. Secure and Privacy-aware P2P Content Distribution System

The proposed scheme as shown in the Fig. 1 consists of the following sub-protocols.

A. *Fingerprint generation process*

The algorithm for fingerprint generation takes a parameter ε for error probability and the total number N of users as an input, and outputs a collection $F = (f_1, .., f_N)$ of binary codewords f_i of length m and a secret vector s. The codeword F_i is sent to the user i, while the encrypted vector s and a hash of s is sent to a *MO* to be used later in traitor tracing.

B. *Base file and supplementary file generation*

The base file (*BF*) is designed to have a small size and is distributed from *M* to all the peers on receiving a payment for the requested file. The proposed method employs DWT to split the content into low-frequency (approximate coefficients) and high-frequency (detail coefficients) components. In case of audio DWT decomposition, we get approximate and detail (C_A, C_D) coefficients whereas in case of an image file obtained from a frame of a video, we get one approximate component (*LL*) and three detail (horizontal (*HL*), vertical (*LH*) and diagonal (*HH*)) components. An approximate coefficient is then itself split into a second-level approximation and detail coefficients, and the process is repeated. For example, for four-level DWT decomposition of an image of size 1024x768, the coefficient set is, $W=[W_{A4},W_{H4},W_{D4},W_{V4},...,W_{A1},W_{H1},W_{D1},W_{V1}]$. At the fourth level, the size of the approximate coefficient W_{A4} is significantly reduced to 64x48. DWT level-4 decomposition has been chosen for our design considering the trade-offs between robustness, capacity and transparency properties of watermarking. W_{A4} provides more robustness and reduced file size with a lesser amount of effect on quality. The fingerprint *fi* of length *m* is then imperceptibly embedded into this W_{A4} using a blind, robust and secure QIM-based watermarking technique [8]. This constitutes a *BF* which is sent to the end user. The remaining detail coefficients constitute a supplementary file (*SF*), which is sent to the *SP* for distribution in P2P system. Eventually, the system at the user end can apply the inverse DWT to get a fingerprinted copy.

C. *Distribution of the base file*

On receiving a file request from a peer P_i (P_i is the pseudo identity of a peer *i)*, *SP* directs P_i's request to the *M (pseudo identity of a merchant). M* first initiates the authentication procedure to verify that the peer claiming to

be the holder of P_i is not lying. An anonymous two-party AKE protocol based on PseudoTrust [7] is established between *M* and P_i. *M* sends an authentication request to P_i. P_i replies to *M* with its PC_i (pseudo identity certificate of peer P_i). Having the PC_i of P_i, *M* initiates the authentication process. The authentication process includes two main phases: (1) P_i acts as a prover to prove its validation to *M* and (2) *M* proves its authenticity to P_i.

For generation of a session key for secure *BF* exchange between P_i and *M*, Diffie-Hellman Key Exchange protocol is used in the authentication process. A DSS is adopted to simplify the AKE protocol.

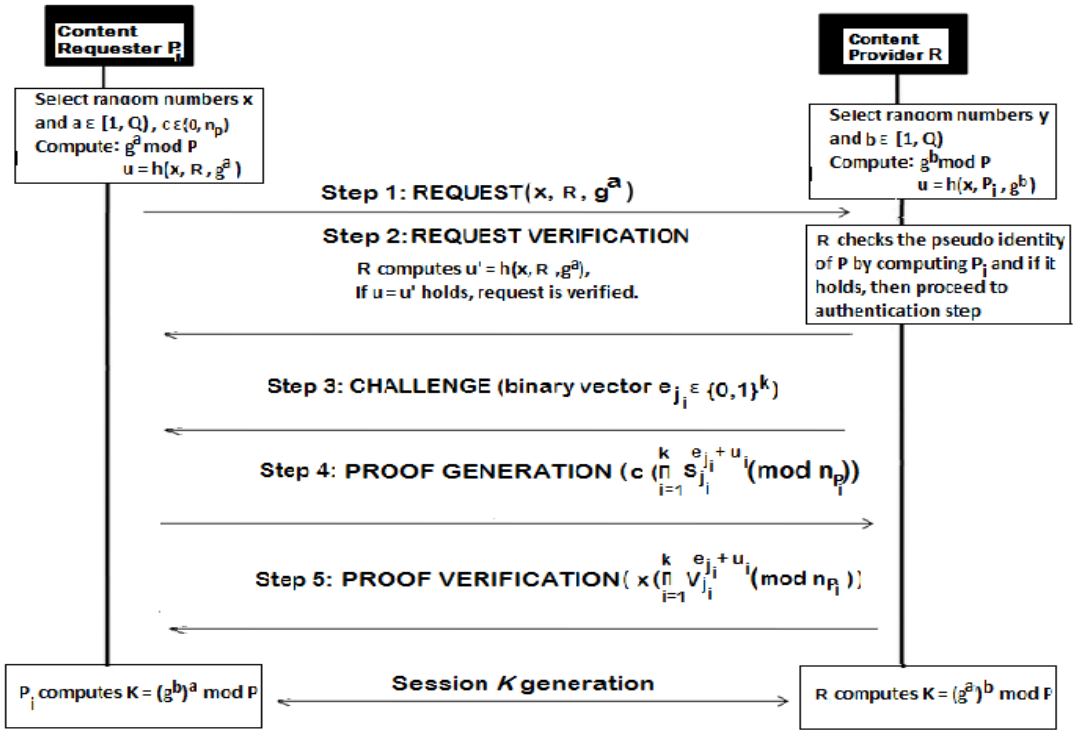


Figure 2. Two-party anonymous AKE protocol based on PseudoTrust [7]

Fig.2 illustrates the authentication process between *M* and P_i. Similarly, for secure exchange of *h(p)*, *M* and *MO* undergo a two-party AKE to exchange one time secret session key K_2 as illustrated in the Fig.2. M encrypts *s* with its public key K_{PM} and further encrypts (E_{KPM} *[s], h(s) M, P_i)* with *MO*'s public key (K_{Pm}) and obtains E_1= E_{KPm} (E_{KPM} *[s], h(s), M, P_i)*. Then it encrypts it again with the session-key K_2 to get $E_{K2}(E_1)$. *MO*, upon receiving the encrypted content, decrypts it with the session key K_2 and further decrypts it with its private key K_{Sm}. *MO* stores E_{KMP}*[s]* and *h(s)* along with the pseudonyms of *M* and P_i.

After mutual authentication between P_i and *M*, the session key K_1 is generated to encrypt *BF*: E_{BF} =E_{K1} (*BF)* and send the encrypted content to the requesting P_i. *M* generates *h(BF)* and saves *h(BF)*, P_i against a transaction ID. P_i decrypts it with K_1 and gets a decrypted *BF*.

D. Distribution of the supplementary file

SP searches for a particular file requested by a peer within its group. If found, it displays the list of peers having that particular file and also displays randomly selected peer nodes to act as middle nodes between content providing peer and the requesting peer. On joining the system, peers construct anonymous paths with existing peers and find their tail nodes based on the APFS protocol [10]. These middle nodes form a covert path for the content-providing peer to send the content to the requesting peer. The content-providing peer pre-constructs an onion-path which points to it and add this path to the *SP*. By doing so, a requesting peer can utilize the onion-path to contact the content-providing peer while knowing nothing about provider's identity. If *SP* is unable to find the file within its group, it sends a request for the file to other connected *SP*s. The other *SP* on finding the particular content-provider sends the response to the requesting *SP*. The *SP* then establishes a path between the requesting peer and that content-providing peer.

Let's assume that *a* is a requesting peer (*P_a*), *b* is the providing peer (*P_b*) and *T_p* acts as a tail node to relay a message for the requesting peer *P_a*. When a peer *P_b* receives the file request and it holds the requested file and decided to be the file provider, it replies to the query through its tail-node *T_b*. The requesting peer *P_a* initiates the authentication process to verify the identity of *P_b*. *P_a* sends an authentication request to *P_b* through the anonymous path, *P_a* → *T_a*→ *T_b*→ *P_b*. Using the two-party anonymous AKE as illustrated in Fig.2, both parties authenticate each other identities and generates a

one-time session key *K_ab* to encrypt the content of the file. *P_b* sends the encrypted file *F* and hash of the file ($E_{K\_ab}(F,h(F))$ to *P_a* through *T_b* and *T_a*.

In order to transfer *P_a*, *P_b*, *h* (*F)* along with the ID of *SP* securely to *MO*, a session key is generated using two-party AKE as described in Fig. 2. At the completion of *SP* transfer to *P_a*, *MO* concatenates all the *h(F)*s and stores this concatenated hash and *P_a* against a transaction ID.

E. Traitor-Tracing

Once a pirate copy *Y* of content *X* is found, *M* extracts the fingerprint by decomposing the pirated content *Y* with the same wavelet basis used in the fingerprint insertion step. This gives the approximate coefficient matrix in which pirated code $pc \in \{0, 1\}^*$ is embedded. After extraction of a pirated code from *Y*, the tracing algorithm of Nuida's codes is employed to identify colluder(s).

In the tracing algorithm, a colluded fingerprint *pc* and secret vector *s* are given as an input. The encrypted secret vector *s* is obtained from *MO* in order to generate the fingerprint matrix for identification of the colluder(s). *M* further decrypts *s* using its private key K_{SM}. The score of users are calculated as per algorithm described in [8] and the output of this tracing algorithm is the user with the highest score. *M* cannot accuse any user of producing a forged copy due to commitment of secret vector *s* to *MO*, which is being trusted by both merchant and buyer.

F. Arbitration

If the user has been identified as a colluder through detection algorithm and he/she denies that an unauthorized copy has been originated from his/her copy and want to prove his/her innocence, a judge *J* can be requested to solve this conflict. *J* is assumed to be a trusted party, which is able to resolve conflicts without the user revealing his/her identity. Moreover, *J* does not require the whole fingerprinted content. On receiving request from a user to deny accusation of piracy, *J* gets a pirated copy's hash from traitor tracing step. *J* requests *MO* to provide the hash of the file registered against the user's pseudonym. If both the hashes have high correlation, it means end user is guilty else he/she is innocent. If found guilty, the real identity of the user can be traced with the help of *MO*. Since *MO* has kept record of each transaction between a peer and *SP*, it can identify the *SP* with whom the accused peer was connected to and thus through its public key, its real identity can be revealed to *J*.

V. CONCLUSIONS

In this paper, we have proposed a privacy-aware content distribution mechanism which provides security and anonymity to both the merchant and buyer. The newly proposed scheme is specific for P2P networks and can benefit multimedia owners to share their big files without fear of copyright violation, such as video files, utilizing the convenience of P2P networks. This scheme reduces the burden of the media owner's server by only sending a small-sized base file and making use of the P2P network to support the majority of the file transfer process. The wavelet technique makes the base file into necessary information for the customer and fingerprint generated using Tardos codes, offers collusion-resistance against a chosen number of colluders and illegal-redistributors' identification whilst preserving the privacy of honest users.

Future Work: Preliminary results of our proposed system show that the base file of both the audio and video files is considerably small in size as compared to the original size. In our forthcoming paper, we will discuss the security and performance analysis of the system. Also, we will compare its performance with similar P2P content distribution systems in terms of security and privacy properties and computational cost.

ACKNOWLEDGMENT

This work was partly funded by the Spanish Government through projects TIN2011-27076-C03-02 ''CO-PRIVACY'' and CONSOLIDER INGENIO 2010 CSD2007-0004 ''ARES''.

REFERENCES

[1] D. Megías and J. Domingo-Ferrer, "Privacy-aware peer to peer content distribution using automatically recombined fingerprints", Multimedia Systems, (In press)

[2] J. Domingo-Ferrer and D. Megías, "Distributed multicast of fingerprinted cntent based on a rational peer-to-peer community", Computer. Communication, vol. 36, no. 5, Elsevier, 2013, pp. 542-550.

[3] Amna Qureshi, Helena Rifà Pous and David Megías, "Security, Privacy and Anonymity in Legal Distribution of Copyrighted Multimedia Content over Peer to Peer Networks A Brief Overview", Proc. Fifth International Conference on Multimedia Information Networking and Security (MINES), IEEE, 2013.

[4] G. Tardos, "Optimal probabilistic _fingerprint codes," Proc. Symposium on Theory of Computing (STOC), ACM, 2003, pp. 116-125.

[5] Digital Signature Standard, FIPs PUB 186-4, [Online]. Available: http://csrc.nist.gov/publications/fips/fips186-4/fips186-4 change1.pdf, accessed 14.12.2013.

[6] U. Fiege, A. Fiat and A. Shamir, "Zero Knowledge Proofs of Identity," Journal of Cryptology, vol. 1, no. 2, Springer, 1988, pp. 77-94.

[7] L. Lu *et al.* , "Pseudo Trust: Zero-knowledge authentication in anonymous P2Ps," IEEE Trans. Parallel and Distrib. Syst., vol. 19, no. 10, IEEE, 2007, pp. 1325- 1337.

[8] K. Nuida, "Short collusion-secure fingerprint codes against three pirates", International Journal of Information Security, vol. 11, no. 2, Springer, 2012, pp. 85-102.

[9] B. Chen and G.W.Wornell, "Quantization index modulation: A class of provably good methods for digital watermarking and information embedding," IEEE Transactions on Information Theory, vol. 47, no. 4, IEEE, 2001, pp. 1423–1443.

[10] V. Scarlata, B. N. Levine and C. Shields, "Responder Anonymity and Anonymous Peer-to-Peer File Sharing", Proc. Ninth International Conference on Network Protocols (ICNP), IEEE, 2001, pp. 272-280.

 ISBN: 978-1-61208-320-9

Adaptive Search Range Determination for Fast Motion Estimation

Wonjin Lee
Department of Electronics and Computer Engineering
Hanyang University
17 Haengdang-dong, Seongdong-gu, Seoul, Korea
veronica0083@gmail.com

Jechang Jeong
Department of Electronics and Computer Engineering
Hanyang University
17 Haengdang-dong, Seongdong-gu, Seoul, Korea
jjeong@hanyang.ac.kr

***Abstract*—The motion estimation (ME) in video codec is extremely important, having acritical effect on encoding time and video quality. Although the full search algorithm is the most fundamental ME method which shows the best video quality, it has high computation complexity. To alleviate this issue, many literatures have been proposed to improve the computational speed and maintain the video quality. In this paper, we propose the new method which determines the search range by using the sum of absolute difference between macroblocks of current frame and reference frame. Experimental results show that proposed method can achieve nearly 209 times computation reduction and can maintain its mean square error performance very close to full search method.**

Keywords- Block matching, partial distortion search, video coding, motion estimation, search range determination.

I. INTRODUCTION

Motion estimation (ME) in many video standards uses the block matching algorithm (BMA). It efficiently removes the temporal redundancy between frames [1-3]. BMA divides a frame into macroblocks and searches the most similar prediction block with the macroblock of current frame.

Instead of sending all information of current macroblock, BMA only encodes difference between current macroblock and prediction block and motion vector that indicates the relative location of prediction block. The full search (FS) algorithm has high computation complexity because of comparing the sum of absolute difference (SAD) in all locations within search range to find the most similar block with current macroblock. Therefore, FS occupies the most of whole encoding time, while it gives the smallest distortion.

In order to reduce the computation complexity of FS, many literatures have been proposed.These algorithms can be classified into fast searching approaches and fast matching approaches.The fast searching approaches can achieve the high speed-up by reducing the search point in search window.

The representative methods are diamond search (DS) [4], four-step search (4SS) [5], new three-step search (N3SS) [6].The representative methods to reduce the encoding time using sub samples are partial distortion search (PDS), normalized partial distortion search (NPDS) [7], adjustable partial distortion search (APDS) [8], and efficient two step edge based partial distortion search for fast block motion estimation (TS-EPDS) [9]. Previous algorithms tried to improve the search speed while maintaining the peak signal-to-noise ratio (PSNR) performance.

In this paper, we present the method that determine the search range by using SAD between macro block of current frame and prediction block of reference frame to reduce the computation complexity and maintain the video quality. The rest of this paper is organized as follows. Section 2 introduces the previous idea and preliminaries. Section 3 explains the proposed method and experimental results are exhibited in Section 4. Finally, concluding remarks are given in Section 5.

II. PREVIOUS ALGORITHMS

NPDS separates 16x16 macro block into 4x4 blocks that do not overlap and obtain the partial distortion between current macro block and prediction block. Partial distortion is SAD for one group which consists of a total of 16 pixels extracted to each 4x4 block, as shown in Figure 1.

The SAD for one group is defined as

$$d(p)(k,l;u,v) = \sum_{i=0}^{3}\sum_{j=0}^{3}\left|\begin{matrix} I_n\left(k+4i+s(p), l+4j+t(p)\right) \\ -I_{n-1}\left(k+4i+s(p)+u, l+4j+t(p)+v\right)\end{matrix}\right|, \quad (1)$$

where pixel position of p is given in TABLE 1, I is a frame, k and l are the positions of current macro block, u and v are position of candidate blocks in search range, n and n-1 are the current frame and the previous frame, respectively. The $t(p)$ and $s(p)$ are offset of the location for the p^{th} partial distortion. $d(p)$ is partial distortion that is the SAD for one group and it is accumulated $D(p)$ and $D(p)$ is compared with full SAD of starting search point.

$D(p)$ is the accumulated d(p), which is given by

$$D(p) = \sum_{i=0}^{p} d(i) \quad . \quad (2)$$

 ISBN: 978-1-61208-320-9

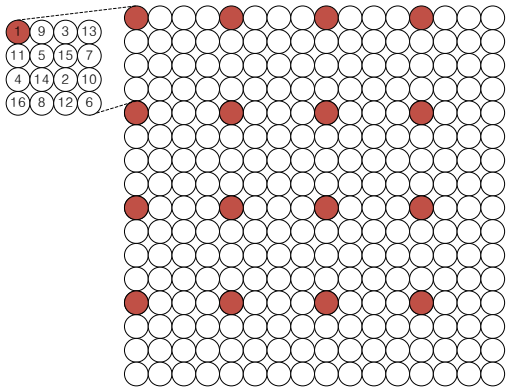


Figure 1. Pixel group for partial distortion.

TABLE I. PIXEL POSITION OF P

p	(s(p), t(p))	p	(s(p), t(p))
1	(0,0)	9	(1,0)
2	(2,2)	10	(3,2)
3	(2,0)	11	(0,1)
4	(0,2)	12	(2,3)
5	(1,1)	13	(3,0)
6	(3,3)	14	(1,2)
7	(3,1)	15	(2,1)
8	(1,3)	16	(0,3)

NPDS is compared with initial SAD that is calculated from all the pixels for 16x16 block, while accumulating the partial SAD for each 16 group. It can improve the match speed by comparing initial SAD with $D(p)$*16/p. The $D(p)$ can estimate initial SAD by multiplying 16/p. For instance, if SAD of each 16 group is similar, we can estimate the full SAD with SAD of one group. However, if SAD of each 16 group is not similar, NPDS can find the motion vector incorrectly. The APDS that is proposed to solve this demerit separates the first group of 16 groups of NPDS into 4 groups to increase search speed. Instead of increasing the search speed by dropping the PSNR, APDS presents the quality factor to increase search speed and maintain the PSNR of FS.

Equation of the quality factor is given by

$$f(n,k) = (1-k)n + kN^2, \qquad (3)$$

where N is the normalization factor and n is the number of accumulated set of pixels. The n is from 1 to 16. $k$has the value from 0 to 1. If k=1, APDS has the same performance as PDS, if k=0, APDS has the same performance as NPDS.

TABLE II. PROBABILITY OF EACH SEARCH RANGE

SAD / SR	128	256	512	1000	1500
1	100	97	94	93	92
2	100	98	95	94	93
3	100	98	96	95	95
4	100	99	97	96	96
5	100	99	97	97	97
6	100	99	97	97	97
7	100	99	98	97	97
8	100	99	98	98	98
9	100	99	98	98	98
10	100	99	99	98	98
11	100	100	99	99	99
12	100	100	99	99	99
13	100	100	99	99	99
14	100	100	99	99	99
15	100	100	100	99	99
16	100	100	100	100	100

III. PROPOSED ALGORITHM

APDS algorithm calculates the SAD for the macroblock at the same position in reference frame with the macroblock in current frame. Then, in the next candidate blocks, APDS predicts the SAD of all the pixels by using the small samples and performs while macroblock moves whole search range. If the SAD of the first candidate is small enough, the true motion vector may be near the current position because difference between macroblocks of current and reference is small.

TABLE 2 presents the probability of that there is a true motion vector at each search range (SR) according to the first calculated SAD. we can find that probability of true motion vector is high in small search range when the first calculated SAD is small enough.

From this motivation, we propose a new method that searches the partial range, it does not perform whole search range. However, this method may find inaccurate motion vector by falling into local minimum point.

To solve this demerit, after we calculate the SAD of median of three neighboring motion vectors and the SAD of origin, we set the position of SAD with a smaller value to the starting point (SP). Then we adjust the size of search range according to the initial SAD. The equation of determination of SR according to the initial SAD is defined as

$$SR = ceil\left(\frac{SAD_{initial}}{TH}\right), \qquad (4)$$

where we experimentally defined TH by 256 to maintain the approximately 95% in the TABLE 4.

 ISBN: 978-1-61208-320-9

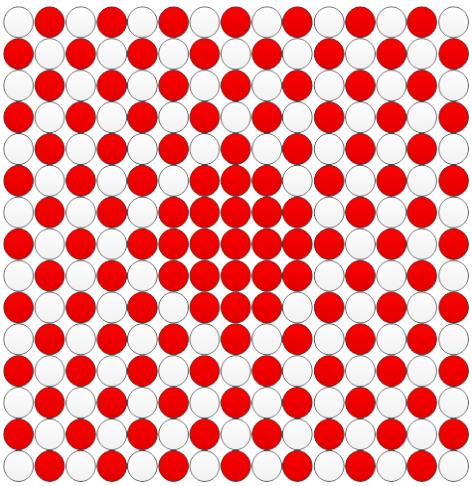


Figure 2. Search point of the proposed method.

We set the search point to increase the search speed as shown in Figure 2. For instance, we find the motion vector roughly and search the neighboring skipped point. Flowchart of proposed method is given by Figure 3.

The Proposed method is performed as follows :

Step 1)
Find the minimum SAD between the SAD of median of three neighboring motion vectors and the SAD of origin.

Step 2)
Determine the search range with the use of SR equation and starting point that is position of minimum SAD decided in Step 1.

Step 3)
Perform APDS through the use of proposed search pattern in Figure 2.

Step 4)
If there are neighboring skipped points, perform APDS in skipped points of neighboring 8 points, otherwise, go to Step 5.

Step 5)
Get the motion vector with the minimum SAD of current block.

IV. EXPERIMENTAL RESULTS

We conducted experiments through the use of 10 sequences (Akiyo, Bridge_close, Children, Hall, Mother, News, Silent, Singer, Stefan, and Paris). The proposed method and BMAs are performed to only motion estimation, not codec such as MPEG-2 and H.264/AVC. Each sequence has 300 frames and the sequence format is CIF(352x288). All implemented BMAs are programmed by visual C++. The proposed method is compared with four conventional methods: FS, PDS, NPDS, APDS, and TS-EPDS. The size of macroblock used in motion estimation is 16x16. We set the default search range by ± 16.

The PSNR, speed-up are used to evaluate the objective performances. The speed-up is computed as operations of

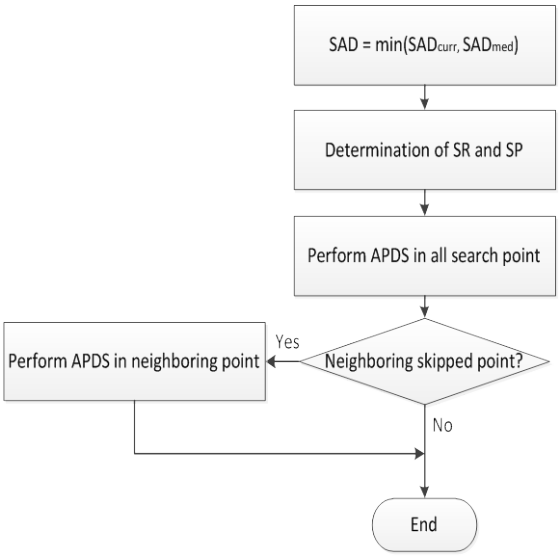


Figure 3. Flowchart of the proposed method.

FS divided by operations of BMA. The total number of operations is the sum of addition, comparison, absolution, and multiplication.

As shown in TABLE 3, the proposed method resulted in the average of 0.02 dB PSNR degradations compared to that of FS. The PSNR performance of APDS and proposed method show almost same with FS, but TS-EPDS has significant PSNR drop on News, Singer, and Stefan sequences.

As tabulated in TABLE 4, theaverage of speed up of the proposed method was 209 times faster than FS, 15 times faster than NPDS and 8 times faster than APDS, and 2 times faster than TS-EPDS.

The proposed algorithm maintained the similar level of PSNR with previous fast search algorithm and could check that computational complexity is considerably less than the previous algorithms. Furthermore, if the background of video sequence is stopped, we could confirm that complexity is reduced exceptionally.

V. CONCLUSION

In this paper, we proposed the algorithm changing the search range for fast search. The proposed method can adjust search range considering that if SAD between current and candidate macroblocks is small enough, the true motion vector may be near the current position. As the experimental results suggest, the proposed method reduced the encoding time by maintaining the similar level of PSNR compared to previous fast search methods. If starting point is determined more accurately, our proposed method can be further improved.

 ISBN: 978-1-61208-320-9

TABLE III. COMPARISON OF THE AVERAGE PSNR WITH THE CONVENTIONAL METHODS

	PSNR					
	FS	**PDS**	**NPDS**	**APDS**	**TS-EPDS**	**Proposed**
Akiyo	42.94	42.94	42.85	42.94	42.92	42.94
Bridge_close	35.23	35.23	35.23	35.23	35.21	35.23
Children	29.79	29.79	29.58	29.77	29.77	29.76
Hall	34.83	34.83	34.70	34.80	34.77	34.79
Mother	40.44	40.44	40.35	40.43	40.41	40.41
News	36.90	36.90	36.68	36.88	36.82	36.86
Silent	35.85	35.85	35.69	35.83	35.83	35.84
Singer	36.87	36.87	36.62	36.82	36.76	36.81
Stefan	24.59	24.59	24.46	24.59	24.40	24.58
Paris	31.90	31.90	31.75	31.89	31.88	31.88
Avg.	34.93	34.93	34.79	34.92	34.88	34.91
Diff.	-	0.00	0.14	0.02	0.06	0.02

TABLE IV. COMPARISON OF THE AVERAGE SPEED WITH THE CONVENTIONAL METHODS

	Speed Up					
	FS	**PDS**	**NPDS**	**APDS**	**TS-EPDS**	**Proposed**
Akiyo	1	10.78	14.57	34.95	183.14	400.98
Bridge_close	1	3.52	6.25	7.03	68.67	188.50
Children	1	8.24	14.38	29.45	90.64	152.66
Hall	1	3.54	14.14	16.09	41.59	158.89
Mother	1	3.65	13.66	14.83	66.02	187.80
News	1	7.95	14.45	28.86	120.92	232.93
Silent	1	6.59	14.37	27.36	84.06	176.96
Singer	1	10.58	14.55	33.88	170.21	337.43
Stefan	1	2.92	13.29	12.74	25.88	54.69
Paris	1	8.09	14.50	31.44	124.23	208.82
Avg.	1	6.58	13.42	23.66	97.54	209.97

ACKNOWLEDGMENT

"This work was supported by the National Research Foundation of Korea (NRF) Grant funded by the Korean Government (MOE) (NRF-2011-0011312).

REFERENCES

[1] Dufaux. F. and Moscheni. F.,“Motion estimation techniques for digital TV:a review and a new contribution,” Proc. IEEE, vol. 83,June 1995, pp. 858-876.

[2] JTC1/SC29/WG11, ISO/IEC 14496-2: Information Technology -coding of audio visual objects - Part 2: visual, (MPEG-4 Visual), 2000.

[3] JVT G050r1: Draft ITU-T recommendation and final draft internationalstandard of joint video specification (ITU-T Rec. h.264/ISO/IEC 14496-10 AVC), May 2003.

[4] J. Y. Tham, S. Ranganath, M. Ranganath, and A. A. Kassim, “A novelunrestricted center-biased diamond search algorithm for block motionestimation,” IEEE Trans. Circuits Syst. Video Technol., vol. 8, Aug. 1998, pp.369-377.

[5] L. M. Po and W. C. Ma, “A novel four-step search algorithm for fast block motion estimation,” IEEE Trans. Circuits Syst. Video Technol.,vol. 6, June 1996, pp. 313-317.

[6] R. Li, B. Zeng, and M. L. Liou, “A new three-step search algorithm forblock motion estimation,” IEEE Trans. Circuits Syst. Video Technol.,vol. 4, Aug. 1994, pp. 438-443.

[7] C. H. Cheung and L. M. Po, “Normalized partial distortion search algorithm for block motion estimation,” IEEE Trans. Circuits Syst. Video Technol., vol. 10, no. 3, Apr. 2000, pp. 417-422.

 ISBN: 978-1-61208-320-9

[8] C. H. Cheung and L. M. Po, “Adjustable partial distortion search algorithm for fast block motion estimation," IEEE Trans. Circuits Syst. Video Technol., vol. 13, no. 1, Jan. 2003, pp. 100-110.

[9] M. G. Sarwer and Q. M. Jonathan Wu, “Efficient Two Step Edge based Partial Distortion Search for Fast Block Motion Estimation,” IEEE Trans. Consumer Electron., vol. 55, no. 4, Nov. 2009, pp. 2154-2162.

 ISBN: 978-1-61208-320-9

Hair Segmentation for Color Estimation in Surveillance Systems

Aleš Křupka, Jiří Přinosil, Kamil Říha, Jiří Minář
Department of Telecommunications
Brno University of Technology
Brno, Czech Republic
e-mail: {akrupka,jiri.minar}@phd.feec.vutbr.cz
{prinosil,rihak}@feec.vutbr.cz

Malay Kishore Dutta
Amity School of Engineering and Technology
Amity University
Uttar Pradesh, India
e-mail: mkdutta@amity.edu

***Abstract* — The paper proposes a novel method for hair segmentation, that can be used in real-time video surveillance systems or multimedia services. The method utilizes an approach of video subtraction to obtain a person silhouette. Subsequently, the head part can be identified on the silhouette by exploiting information about face position. From the head, the skin is separated using floodfilling procedure and the hair area is determined as the difference between the head and the skin. The precision of the method is evaluated using manually extracted hair masks. The purpose of the segmentation method is to specify a hair area which can be then used for a hair color estimation. Therefore, the usability of the hair segmentation procedure is tested by a proposed scheme of hair color estimation.**

Keywords-segmentation, hair, color, face detection, video subtraction, floodfilling.

I. INTRODUCTION

In recent years, computer vision has become an inherent part of modern surveillance systems. Algorithms for pedestrian detection, people tracking or identity verification are common parts of these systems. This paper deals with analysis of hair in video sequences. Hair is classified to the category of soft biometric traits. This means that it cannot be used for a person identification by itself, but it can help the identification together with other soft biometric traits. Further, it can be used for division people into sub-groups, for example by assigning a hair color index to a person in a video-sequence. Then, such indices can be used as an additional filtering clue when searching in multimedia databases.

Firstly, we mention relevant works in this topic. In [1], hair area is detected based on sliding window which evaluates the hair of color. In [2], color and frequency information is used for creating seeds. Hair is then extracted using matting process using the seeds. In [3], hair is segmented using Graph-Cut and Loopy Belief Propagation. In [4], hair seeds are detected and a growing of hair region is applied based on color and texture features. In [5], the approach of seed identification and consecutive propagation is used. This procedure is done in two stages where the second stage uses a specific hair model based on the first stage results. In [6], the hair seed patches are obtained via active shape and active contours. These areas are then used to train a model of hair color and texture. According to this model, the final hair area is determined. In [7], selected hair and background seed regions are used for online support vector machines (SVM) model training. This model is then used to differentiate between other hair/background pixels. In [8], the coarse hair probability map is estimated and this map is consequently refined using Isomorphic Manifold Inference Method to get optimal hair region. In [9], part-based model is proposed together with a way of modeling relations between the parts of head and hair which helps to a better hair identification.

The previous works are designated to work with static images and therefore the need to distinguish between a head and a background exists. The motivation of this work is to be able to estimate a hair color of people in a video-sequence, so this soft biometric trait can be extracted in real-time. The fact of using video sequence significantly simplifies the solution of head/background separation and therefore the hair segmentation procedure can be simplified in advance of shortening the processing time.

The paper is organized as follows: Section II describes the proposed method for hair region selection. Section III presents the experimental setup and the results of hair region selection and its usability for hair color estimation. Section IV then concludes the results and according to them proposes a direction of the future work.

II. HAIR SEGMENTATION METHOD

The method presumes the usage of video-sequences. The scheme of the method can be seen in Fig. 1. A hair is determined as a difference between head and skin area of a head. Every frame of the video-sequence is examined by a face detector for a face occurrence and it is also supplied to a background subtractor. If a face in the frame is detected, the position of the face is used for a segmentation of head. A silhouette of the person is obtained from the background subtractor and the head is given as the part of the silhouette. This part is specified by the position returned by the face detector. As the head mask is given, the skin area in the head is needed to be found. This is performed utilizing information about eyes position and nose position. This information is obtained during the face detection stage and it is used for a selection of proper points, which are used as seeds for a flooding procedure. Using the flooding procedure, the skin area is defined. Finally, the hair mask is given as the difference between the head segment and the skin segment.

 ISBN: 978-1-61208-320-9

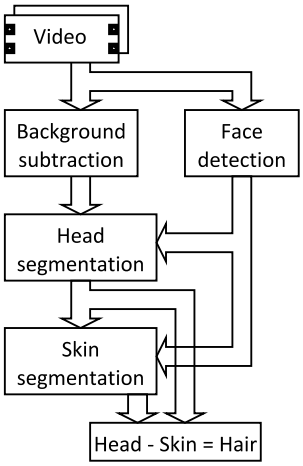


Figure 1. Scheme of the proposed method

A. *Face detection*

For the face detection, the widely used cascade Viola-Jones detector [10], implemented in OpenCV library, is used. Using the detector, a face occurrence in a frame can be located and a sub-window containing face is specified. Then, in the selected sub-window, eyes and nose are also located by the Viola-Jones detector, as it can be seen in Fig. 2(b). The position of eyes and nose is then used for further processing. In this stage, a successful extraction of face, eyes and nose is required for the further processing. Thus, if no face is detected in a frame or if eyes and nose positions are not obtained successfully, the processing procedure of the current frame is cancelled and the current frame is only supplied for the background subtractor. The processing procedure is then started with the following frame in the video-sequence.

B. *Background modeling*

Background modeling is a big advantage when using a video-sequence for the hair segmentation. The frames of the sequence allow to model a background scene and thus a silhouette of a moving human subject can be obtained. For this purpose, the method using Gaussian mixtures for background modeling [11] implemented in OpenCV library is used. This implementation also addresses shadow detection, thus shadows appearances can be eliminated during silhouette extraction. However, because the segmentation of a moving object in a frame using this method is still not perfect, the morphological opening is applied to remove small spurious segments in the frame. The moving object is then separated from the scene, as shown in Fig. 2(c).

C. *Head segmentation*

A head in the frame is obtained using the silhouette from the background subtractor and the face position from the face detector. The face position is presented by a rectangle with face. This rectangle is enlarged in order to cover the whole head area, as illustrated in Fig. 2(b). The size of the rectangle is enlarged by factor 1.5 which was empirically selected as optimal value. This rectangle thus contains a part of the silhouette corresponding to the head. Usually, the silhouette obtained from the background subtractor is not ideal. Concretely, it can consist of unconnected regions and thus the part corresponding to the head cannot be used as a head mask directly, as can be seen Fig. 2(c). Thus, a convex hull is constructed from the point set which is given as the union of regions in the head area. This convex hull then represents the head mask. The convex hull is constructed only from the pixels of the upper rectangle part to avoid including pixels of arms, the resulting head mask is shown in Fig. 3(a). Although the head mask does not cover all the head area, for the purpose of color estimation it is sufficient to have hair from the top of the head.

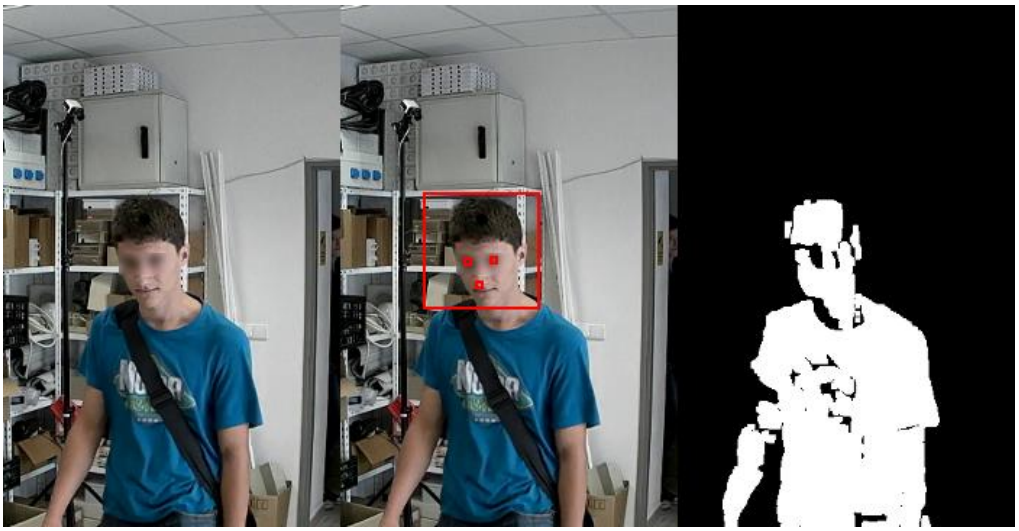


Figure 2. Illustration of initial video frame processing: (a) original frame, (b) face, eyes and nose detection, (c) silhouette extraction

 ISBN: 978-1-61208-320-9

D. Skin segmentation

The skin segmentation is the most crucial part of the processing. When a skin of a head is segmented correctly, then a difference between the head mask and the skin mask defines a hair mask. As mentioned in the beginning, the floodfilling approach is used for the skin segmentation. The flooding starts repeatedly from the different seed-points. For the floodfilling, an optimal RGB range was empirically selected. A pixel belongs to this range if the distance between its value and the value of the current seed-point is not greater than 30 (for the corresponding 8-bit color channels). The seed-points can be seen in Fig. 3(b), and they are given as points above and under eyes. Their positions are calculated utilizing the positions of the eyes and the nose. The reason for the usage of multiple seed-points is that the skin color varies in the different places of a face and thus the floodfilling would not work satisfactorily. An example of the color variation can be seen in Fig. 3 where one half of the face is darker due to shadows. When using multipleseed-points, the area with similar color around a particular seed-point is flooded. Then, the skin area is obtained as an union of the areas flooded from the different seed-points.

Similar as in the case of head segmentation, such the union of flooded areas does not give an ideal skin mask, as can be seen in Fig. 3(c). For example, the eyes' area is not flooded because the pixels' values are too different from seed point values. Thus, the same approach as during head segmentation, i.e., a convex hull of an union of flooded areas, is constructed. The final hair area is then given as the difference between the head and the skin area, as in Fig. 3(d).

Sometimes, the upper seed-points can fall into the hair area. In this case, such the seed-point cannot be used to initiate the floodfilling procedure. For a selection of proper seed-points, the color modeling method, described in [12], is used. The pixels of the line going through the lower seed-points are considered to have a color of skin, because the pixels are in the area around the nose. Thus, the color model of the skin is created using these pixels. If a color of a particular upper seed-point fits into this model, than the seed-point can be used to initiate the floodfilling procedure.

Finally, one more fact needs to be regarded. When a color of hair is similar to a skin color, the flood can also propagate into an area of hair, as can be seen in Fig. 4. Therefore, to obtain the correct skin segment, only a part of flooded area around the current seed-point is selected. The pixels around the seed-point are considered to be a skin as long as the shape of flood shrinks when going away from the seed point. This way, only a relatively compact part of the flooded area is selected, the selection is illustrated as blue areas in Fig. 4. This is based on the assumption that a texture of skin is more homogeneous than a texture of hair, so theflooding procedure forms more compact shapes on skin parts. Here follows the selection principle of a compact part: The flooded area is represented by the binary shape as shown in Fig. 5(a). Further, the distance transform of morphological type [13] is applied on this shape to get distance image illustrated in Fig. 5(b). In this image, the appropriate regional maximum is found according to the position of the current seed-point (black pixel in Fig. 5(b)). From this regional maximum, the process of descending to lower levels of distance image is performed in individual steps. At the beginning, the appropriate regional maximum is marked as positive. The other regional maxima are marked as negative. Then, the descending starts. In every step, pixels of a current level are marked as positive or negative. The pixels neighboring to negative pixels are marked as negative. The other pixels of the current level connected with positive pixels are marked as positive. These steps are repeated until the level of one is reached. The compact part is then composed from positive pixels as can be seen in Fig. 5(c).

III. EXPERIMENTS

The performance of the method was tested using 28 video sequences of average length of 4 seconds and 25 frames per second. Each video-sequence contains one person passing through a room, as can be seen in Fig. 2. On average, 37 hair masks are segmented from each video-sequence (a mask is obtained, if a face together with positions of eyes and nose is detected in a frame). These masks represent a hair of a person passing through the room. Totally, 1035 hair masks (extracted from all 28 video-sequences) were evaluated in this experiment.

The stages of face detection and background subtraction were performed on frames of size 854x480 pixels. To provide good stability of the background subtractor, several seconds of a video-sequence capturing the typical changes in the scene are supplied to the background subtractor to better model the background scene. When executing the part of skin segmentation, a square with detected face is normalized to size of 350x350 pixels. The part of selecting compact flooded parts around seed-points (Fig. 4) is due to computational reasons performed on down-sampled squares of 130x130 pixels. This step of down-sampling shortens the processing time.

A. Hair segmentation

As a reference, the area of hair was manually labeled.

Figure 3. (a) head segmentation, (b) seed-points, (c) union of flooded areas, (d) hair and skin segment

Figure 4. Selection of compact subpart of flood for different seed-points

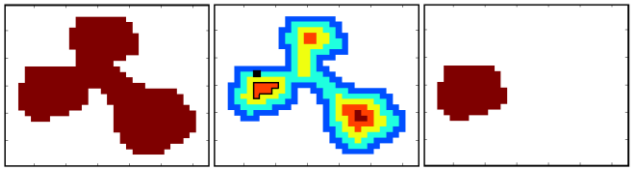


Figure 5. (a) binary shape, (b) distance image, (c) resulting selection

A mask extracted by using the described segmentation procedure is then compared with label mask and pixel counts such as true positives (*TP*), false positives (*FP*), false negatives (*FN*) and true negatives (*TN*) are given. Using these counts, the measures

$$precision = \frac{TP}{TP + FP} \quad (1)$$

and

$$recall = \frac{TP}{TP + FN} \quad (2)$$

can be computed.

Resulting hair masks often include on their border parts also non hair pixels. To alleviate this, the morphological erosion is applied on the resulting mask. Although this operation reduces recall measure, it increases precision measure. In the context of the utilization of segmented hair area for hair color estimation, precision can be considered the important measure, because a big number of false positive pixels (low precision) can more influence the final color estimation. On the other hand, the situation is not critical, when a subpart of hair is not included into the final color estimation (low recall), if we assume, that the hair area does not have more parts of different colors. The results of the experiments can be seen in Table. I.

TABLE I. SEGMENTATION RESULTS

Measure	***Average***	***Median***
Precision	0.84%	0.92%
Recall	0.49%	0.48%

The obtained results can be considered as good. Despite recall being quite low, it can be stated, that still approximately one half of hair area is marked by the segmentation procedure. On the other hand, the precision value is very good and it is comparable with the state of the art techniques referenced in the introduction. The main contribution is that the method was intended to be applicable in real-time systems and it is able to work sufficiently fast, see processing times in Table II.

TABLE II. PROCESSING TIMES

Phase	***Average time* [ms]**
Face detection	121
Background subtraction	55
Head and skin segmentation	38

The test was run on machine with Intel Core i5 M560 processor. Although the face detection stage is the most time consuming, the need of face position can be satisfied by frame down-sampling or face detection with rate lower than actual video frame rate. A face position in frames, where the detection is not done, are obtained using face tracking via optical flow as proposed in [14]. If we have a position of face and a silhouette by hand, the hair segmentation procedure is very fast. Therefore, the method can be implemented as an additional processing in various surveillance systems, where the face detection and background subtraction is also a part of processing and therefore, the hair segmentation method will not present a big additional computational burden.

B. Color estimation

In this stage, the suitability of the hair segmentation procedure for color estimation was tested. A hair color of 28 people in available video-sequences was estimated. Firstly, the color of hair was estimated from the area which is determined by hair segmentation procedure. Secondly, the color of hair was estimated from the area which is given by manually extracted hair label. To get a reference, a hair color of each person was subjectively classified. The colors, which are estimated from the pixels of extracted hair masks and hair labels, are then compared with this reference. The color can be classified into 5 different categories (black, brown, blond, red, gray/white).

For the automatic color classification, the following scheme is used. The RGB space can be considered as a cube. When it is equally spaced to 10 ranges for every dimension, 1000 sub-cubes are obtained. The color from the central RGB triplet of each sub-cube was subjectively evaluated using the previously mentioned color classes (plus a class, when a color is not a color of hair). Then, according to this evaluation, every sub-cube presents a range of RGB values corresponding to a given color - this approach assumes, that all possible colors in one sub-cube are similar.

A color of pixel, which is determined by the segmentation procedure as a hair pixel in a particular frame, is estimated using the scheme described above. When examining all hair pixels in the frame, a hair color histogram, showing numbers of pixels of defined hair colors, is obtained. For one person, every frame of the video-sequence is examined this way. The numbers of corresponding colors are then summed for all examined frames in the video-sequence. The biggest number determines the hair color of the person captured in the video-sequence. This procedure was conducted for all 28 videos. The estimation results, which are based on the hair masks extracted by the proposed hair segmentation method, are shown in Table III. For a comparison, Table IV provides results, when the color estimation is based on the pixels determined by manually extracted hair labels. It can be seen that the results are very similar, they differ only in two cases. Thus, we can state, that the hair segmentation method can be considered as good for purposes of hair color estimation, because the low value of recall only slightly influences the results of the color estimation procedure.

 ISBN: 978-1-61208-320-9

TABLE. III COLOR ESTIMATION RESULTS USING EXTRACTED HAIR MASKS

		Estimation					
		Blk.	Brw.	Bld.	Red	Wht.	Perc.
Subj. Evaluation	Blk.	**13**	1	0	0	0	92.9%
	Brw.	2	**5**	0	0	1	62.5%
	Bld.	0	0	**0**	0	3	0%
	Red	1	0	0	**0**	0	0%
	Wht.	0	0	0	0	**2**	100%

TABLE. IV COLOR ESTIMATION RESULTS USING MANUALLY EXTRACTED HAIR LABELS

		Estimation					
		Blk.	Brw.	Bld.	Red	Wht.	Perc.
Subj. evaluation	Blk.	**13**	1	0	0	0	92.9%
	Brw.	1	**6**	0	0	1	75.0%
	Bld.	0	1	**0**	0	2	0%
	Red	1	0	0	**0**	0	0%
	Wht.	0	0	0	0	**2**	100%

For the discussion about usability of our hair color estimation model, we use Table IV. When we aim at discriminability of individual colors using our proposed model, we can see that the black and white/gray color estimation is the most successful (the people subjectively evaluated as having black or white/gray hair were correctly estimated in 92.9% or 100%, respectively). On the other hand, the subjectively perceived blond color was not correctly estimated in any case. According to the results, the blond color is estimated as white or brown. The reason is that subjectively evaluated blond hair can be in fact blend of more colors, the most common are white and brown. The brown color is correctly estimated in 75.0%. The red color is not correctly estimated in any case, however, only one person with subjectively evaluated red hair was present in the database, so the result of the red color cannot be considered as representative.

Although the color estimation scheme is rather simple, it was intended mainly for the evalution of usability of the developed hair segmentation method. However, as can be seen from the results, hair color estimation itself is a very challenging topic and therefore it will need further effort to develop sufficiently robust method, but this development is out of the scope of this paper. Generally, retrieval of color from image is not a trivial task. RGB values in image can be for example influenced by various sources of illuminations. Thus, an illuminant estimation and its inclusion into the color estimation procedure is needed as mentioned in survey [15]. For the future work, more sophisticated schemes for hair color estimation need to be developed such as the scheme presented in [16], where the color values are classified into predefined categories specified by google search engine.

IV. CONLUSION AND FUTURE WORK

The main objective of this paper was the proposal of the hair segmentation method which will be suitable for the following task of hair color estimation. This method analyses a hair of human from the frontal view and it is applicable on video-sequences. The main contribution of this proposal is its applicability in real-time systems because of its low computational requirements.

The method was tested on 28 labeled video-sequences and the results showed that precision, which is important measure in context of color estimation, has very good values. The relatively low recall of the proposed segmentation method has no big importance in the context of hair color estimation. This was supported by the experiment of the estimation of hair color. The resulting hair color was based on pixels which were determined by automatically extracted hair masks in the first case and by manually extracted hair labels in the second case. For both cases, the color estimation was nearly the same. For the estimation, the hair color model using equidistantly divided RGB space was applied. However, the hair segmentation method can be further developed especially to be able to recognize bald people. Currently, due to the segmentation errors, it is still not easy to distinguish bald people form the people with big forehead and thin layer of hair (bald people were not contained in video-sequences for evaluation).

From the results, some suggestions for the future work can be made. The color estimation scheme was rather simple and it was used to evaluate usability of the proposed hair segmentation procedure for the purpose of hair color analysis. As mentioned earlier, it is not a trivial task to automatically determine a color of hair as it is commonly done by humans. Therefore, a more sophisticated color estimation scheme should be proposed. This scheme will take into account a typical color composition of different hair color and also consider the influence of illumination. For example, some kind of supervised machine learning techniques combined with known methods of illuminant estimation could be utilized for these purposes. These techniques could take into account percentage composition of colors for a particular hair color class. Further, the reference hair colors were obtained by a subjective classification. Thus, because of the subjective classification, these values should be acquired from more subjects in standardized conditions to compare opinions of hair color from different evaluators. Eventually, the bigger experimental database should be acquired. This database should be balanced, when considering different hair colors, to better evaluate the obtained results.

ACKNOWLEDGMENT

This research was supported by part of the project reg. no CZ.1.07/2.3.00/20.0094 which is co-financed by the European Social Fund and the state budget of the Czech Republic and by the project VG20102014033 financed by the Ministry of the interior of the Czech Republic and by the project SIX CZ.1.05/2.1.00/03.0072.

 ISBN: 978-1-61208-320-9

REFERENCES

[1] Y. Yacoob and L. S. Davis, "Detection and Analysis of Hair," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 28, no. 7, July 2006, pp. 1164-1169.

[2] C. Rousset, "Frequential and color analysis for hair mask segmentation," 15th IEEE International Conference on Image Processing, Oct. 2008, pp. 2276 – 2279.

[3] K-C. Lee, D. Anguelov, B. Sumengen, and S. B. Gokturk, "Markov random field models for hair and face segmentation," 8th IEEE International Conference on Automatic Face & Gesture Recognition, Sept. 2008, pp. 1-6.

[4] U. Lipowezky, O. Mamo, and A. Cohen, "Using integrated color and texture features for automatic hair detection," IEEE 25th Convention of Electrical and Electronics Engineers in Israel, Dec. 2008, pp. 51-55.

[5] D. Wang, S. Shan, W. Zeng, H. Zhang, and X. Chen, "A novel two-tier Bayesian based method for hair segmentation," 16th IEEE International Conference on Image Processing, Nov. 2009, pp. 2401-2404.

[6] P. Julian, C. Dehais, F. Lauze, V. Charvillat, A. Bartoli, and A. Choukroun, "Automatic Hair Detection in the Wild," D. Wang, S. Shan, W. Zeng, H. Zhang, and X. Chen, "A novel two-tier Bayesian based method for hair segmentation," 20th International Conference on Pattern Recognition, Aug. 2010, pp. 4617-4620.

[7] D. Wang, X. Chai, H. Zhang, H. Chang, W. Zeng, and S. Shan, "A novel coarse-to-fine hair segmentation method," IEEE International Conference on Automatic Face & Gesture Recognition and Workshops, March 2011, pp. 233-238.

[8] D. Wang, S. Shan, H. Zhang, W. Zeng, and X. Chen, "Isomorphic Manifold Inference for hair segmentation," 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, April 2013, pp. 1-6.

[9] N. Wang, H. Ai, and F. Tang, "What are good parts for hair shape modeling?," IEEE Conference on Computer Vision and Pattern Recognition, June 2012, pp. 662 - 669.

[10] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Dec. 2001, pp. 511-518.

[11] Z. Zivkovic, "Improved adaptive Gausian mixture model for background subtraction," Proceedings of the 17th International Conference on Pattern Recognition, Aug. 2004, pp. 28-31.

[12] T. Horprasert, D. Harwood, and L. S. Davis, "A Statistical Approach for Real-time Robust Background Subtraction and Shadow Detection," International Conference on Computer Vision FRAME-RATE Workshop, 1999, pp. 1- 19.

[13] L.Vincent, "Granulometries, Segmentation, and Morpho-logical Algorithms," Lecture Notes for Morphological Image and Signal Processing Workshop, September 1995, pp. 37-41.

[14] K. Říha, J. Přinosil, D. Fu, M. Zukal, and J. Karásek, "Method for Real-Time Face Tracking in a Video Sequence", Advances in Sensors, Signals, Visualisation, Imaging and Simulation. Recent Advances in Electrical Engineering Series, 2012, pp. 182-187.

[15] A. Gijsenij, T. Gevers, and J. van de Weijer, "Computational Color Constancy: Survey and Experiments," IEEE Transactions in Image Processing, vol. 20, no. 9, Sept. 2011, pp. 2475 - 2489.

[16] J. van de Weijer, C. Schmid, J. Verbeek, and D. Larlus, "Learning Color Names for Real-World Applications," IEEE Transactions in Image Processing, vol. 18, no. 7, July 2009, pp. 1512 - 1523.

 ISBN: 978-1-61208-320-9

Creative Applications of Microvideos

Angus Forbes and Javier Villegas
School of Information: Science, Technology, and Arts
University of Arizona
angus.forbes@sista.arizona.edu, javier.villegasp@gmail.com

***Abstract*—This paper introduces ongoing work on video granular synthesis. The strategies traditionally used in granular synthesis in order to granulate audio signals are extended to streams of video data. We present initial techniques that are made possible through transforming a video signal into a large array of microvideos, or video grains. These involve the dynamic resynthesizing of the video based on spatial and temporal elements of the video grains. Initial explorations in the creative manipulation of videos using these methods are described. We show that video granular synthesis strategies facilitate novel video processing techniques that could lead to new creative effects.**

Keywords - Video granulation; Video processing; Granular synthesis

I. Introduction to Video Granular Synthesis

Granular synthesis is a common method for creating new sonic textures [1]. For instance, it is one of the preferred strategies to manipulate the duration of existing sounds without changing their pitch, or changing the pitch without affecting their length [2]. The fundamental elements of a granular synthesizer are small acoustic objects, sounds of short duration that can barely be perceived as individual sonic events. By manipulating the position in time of the grain, the overlapping factor of adjacent grains, or individual characteristics of the grain (e.g., frequency), a composer can create different sonic atmospheres. Interesting transformations can also be obtained with granular techniques if the sound grains are captured from real-world signals, rather than computationally generated. Grains extracted from the source signal can be re-arranged, eliminated, repeated, or otherwise manipulated in order to create compelling effects. This approach is known as micromontage or *granulation* [1], [2]. Granular synthesis, based on granulation, and applied to real-world signals, is an extremely successful technique; many of the most popular audio manipulation software suites now include tools for granular synthesis and transformation (see [3] for a extensive list of software tools).

Creative approaches using granular synthesis strategies are not as common in the video domain. However, some previous works do present spatial-temporal manipulations of video signals which are related to the ones we present here. For instance, one popular technique known as *slit-scan* is also based on a transmutation of the time and space axes. A repository of artworks based on the slit-scan technique has been put together by Golan Levin [4]. Our video granulator software, described below, can be used to produce visual outputs similar to the ones obtained with an slit-scan (see

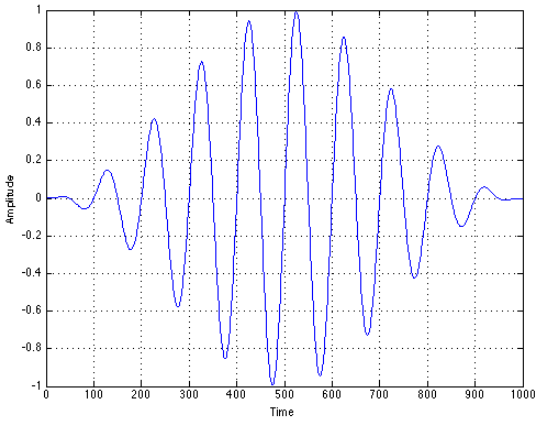


Figure 1: An audio grain

Fig. 7), but it can also be generalized to include geometric manipulations, for example, to allow the signal to be perceived from unusual, i.e., not front-facing, points of view. In many image-based Non-Photorealistic (NPR) effects, sets of pixels are grouped and replaced by synthesis elements that can, for instance, simulate brushstrokes or provide other kinds of creative manipulations. But with most NPR techniques the internal pixel information is usually "smoothed over" and does not remain part of the output [5]. Fels et al. [6] showed a re-interpretation of the space-time cube of a video signal and presented different alternatives to shuffle these two domains. Our technique differs in that it manipulates a larger perceptual entity, the video grain, rather than individual pixels. Alvaro Cassinelli also created a pixel-based interactive piece that allows navigation in time and space using a tangible surface [7], and his examples and experiments with moving objects are relevant to our investigation. A combination of NPR synthesis with space-time analysis is presented by Klein et al. [8]. It uses a set of different "rendering solids" to recreate a NPR version of the input. In some sense, their rendering solids are similar to the time-varying envelope that we use to create spatial grains, but their method is not intended as an extension of granulation techniques. The work described in this paper also relates to previous work by the authors on video processing and analysis, including [9] and [10].

Human perception of audio and video streams has strong

 ISBN: 978-1-61208-320-9

Figure 2: A video grain, the windowing function is applied on the spatial and temporal dimensions.

differences. Extending the concept of granulation to video domains demands a new exploration of the creative possibilities of such techniques. Below we present our ongoing work on the exploration of such alternatives. We present our initial implementation of a *video granulator* and show how basic audio techniques like cloning, skipping, or grain-shuffling can also be creatively applied to video signals. Finally, we provide examples how how the spatial-temporal organization of grains can be manipulated and demonstrate content-dependent manipulations on video signals.

II. System Overview

Similar to an *audio grain*, a *video grain* is a portion of an input video signal windowed by an envelope. In audio granular synthesis, different envelopes are used to overlap regions of the input signal, and can be chosen by a composer for particular effects [1]. In adapting granular synthesis to the video domain, we applied a Hann window envelope to video signals since they create grains with a uniform overlap characteristic [11]. A audio grain is depicted in Fig. 1, and Fig. 2 shows the spatial-temporal windowing of a video grain.

Fig. 3 provides an overview diagram of our video granulator system, transmuting an input video signal into a creatively manipulated output video signal. An input video is interpreted as a *video cube* of three dimensional data (step 1), made up of video frames – the x and y coordinates – extended through time – the z coordinate. According to parameters that define the input window size and various factors that define the amount of overlap, the starting position of each grain in each of the three dimension is calculated and stored in a *grain map* (step 2). This grain map contains information about how the position of each grain created from the input video cube is mapped to an output signal. It is in the construction of the grain map where different manipulations such as cloning, skipping, shuffling grains, or changing overlapping factors can be generated. A scheduler component looks at the output of the grain map to determine what portion of each grain should be used to build the current output frame (step 3).

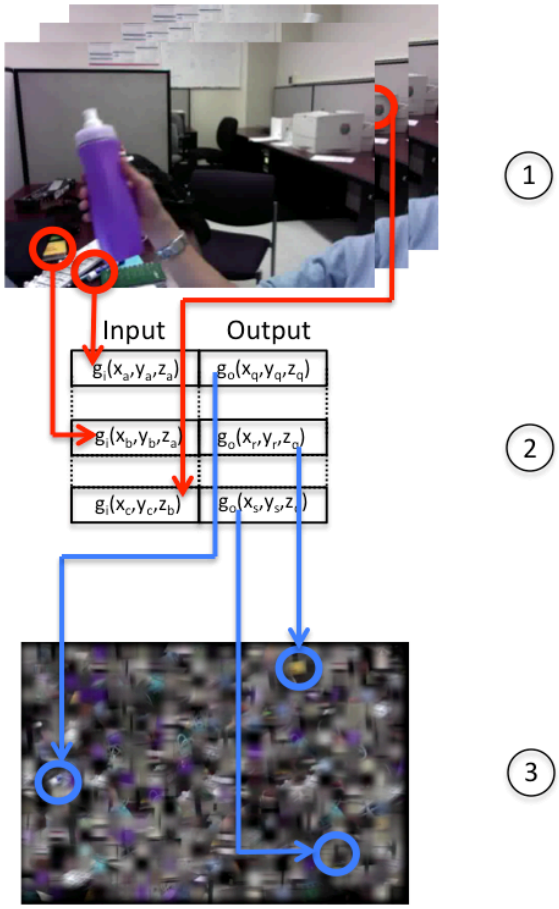


Figure 3: A overview of our *video granulator* system. Step 1 shows the input video cube; step 2 shows the grain map; and finally step 3 shows one frame of the final manipulated video output. The red arrows show example grains from the video cube being added to the grain map, and the blue arrows show how these grains are repositioned temporally and spatially into an output frame.

III. Creative Manipulations

In this section we will present some of the video manipulations that can be obtained with our granular approach.

A. Grain-Based Manipulations

Specific grains can be *cloned* (added multiple times to the grain map) or *skipped* (not included at all in the grain map). We explored examples where we created video grains from the input video cube using a 50% overlapping factor. On the creation of the output, we either chose some grains to be skipped or cloned, effectively changing the size of the image while preserving the spatial-temporal frequencies. Fig. 4 shows a frame of the video granulator after a cloning operation was performed. The image has a curious resemblance to the "op art" artwork created by Julio Le Parc [12].

We also explored randomizing the position of the grains. That is, we altered the spatial and temporal aspects of the grains in different ways. In Fig. 5, we show an arbitrary permutation along all axes. Fig. 6 shows an example frame from an output video which used a grain map that shuffled gains only along the temporal axis. We plan to explore more controlled manipulations of the position of the grains that should lead to interesting visual effects.

B. Spatio-Temporal Reinterpretation

By considering the input video as a cubic array of grains, new interpretations of the data can be created simply by relocating the *point of view* of the array. That is, we can imagine the video being played from a different direction. Fig. 7 shows a

 ISBN: 978-1-61208-320-9

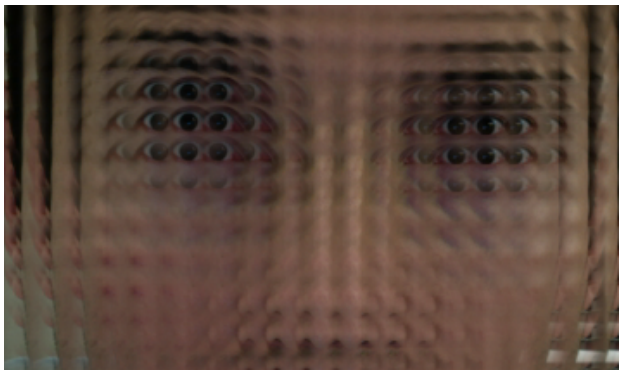


Figure 4: A frame showing the cloning of video grains.

Figure 5: A frame showing arbitrary permutation of grains along the spatial and temporal axes.

Figure 6: Shuffling the time position of some of the grains.

frame obtained while looking at the video array from one side (where a space axis and the time axis, the x and z axes, are interchanged). Fig. 8 shows a view from one of the corners.

C. Image Dependent Manipulations

The position of grains can also be modified by other, non-procedural strategies. For example, higher level information from the video stream can be used to determine the behavior of the grains. In one of our explorations we changed the spatial

Figure 7: The video cube of grains viewed from the side.

Figure 8: The video cube of grains viewed from one corner.

position of the grains according to its temporal variance. The squared root of the temporal variance for a grain spatially placed at x, y is:

$$\bar{\sigma_G}\left(x_0, y_0\right) = \frac{1}{G_s^2} \sum_{x=x_0}^{x_0+G_s} \sum_{y=y_0}^{y_0+G_s} \sqrt{\text{VAR}\left[p_{x,y}\left(t\right)\right]} \qquad (1)$$

Where G_s is the grain size (assumed to be equal in all dimensions), $p_{x,y}(t)$ is the gray-scaled pixel value at position x, y and time frame t, and the variance is calculated over time. The position of the grains that have a variance greater than a predefined threshold is altered in a random (but pre-calculated) direction by an amount proportional to the square of the averaged temporal variance of the grain. Fig. 9 shows a frame of a video of person's hand waving back and forth, illustrating this dynamic manipulation.

IV. Conclusion and Future Work

This initial work shows that creative manipulations with a granular approach are also possible with video signals. Although there are strong differences in the way in which human perception works in the two domains, some of the strategies can be extended in a very straightforward way. This is the case when we clone or skip grains. In the audio domain, this strategy is commonly used to modify the length of an audio signal without changing its pitch. But this trade-off is not as meaningful when processing visual information. While, for instance, relative small changes on the reproduction rate of a voice signal can immediately sound artificial, we are used to seeing faces at different scales. However, although this pitch-preserving time-modification is not as important

 ISBN: 978-1-61208-320-9

Figure 9: The position of grains in the regions with greater temporal variance is modified.

in images (an exception might involve periodic textures), even a straightforward application incorporating cloning and/or skipping grains produces visually interesting results.

Randomizing grain placements is another strategy that is also used in audio. Our initial experiments lead us to believe that applying this strategy to video signals has significant narrative potential. Different video atmospheres can be created by controlling the amount of randomization and the dimensions (spatial or temporal) to which it is applied. Through presenting the video cube of grains from different directions, diverse interpretations of the same block of visual information can be generated. The spatial and temporal dimension can interchange roles reveal interesting differences about the perception of time-evolving time versus space-changing data.

Our last example illustrates the versatility of our proposal. The dynamics of the grains can be controlled with high level features that introduce many possibilities for *interactive* systems. In this example we used the temporal variance of the grains, but in future explorations we will use more sophisticated measurements, such as optical flow, to control video manipulation parameters. Other image characteristics that can be used to condition the grain behavior include luminance, chrominance, frequency content, spatial position, or even features that measure the size of a face or how close it is to the camera. It is our intention to continue exploring other creative possibilities, such as, for instance, using a variable grain size, filtering the grains in various ways, utilizing a swarming or flocking algorithm on the grains, among others. We also believe that these techniques would enhance some of our existing research into creative video processing techniques and applications [13]–[15].

We have discussed the differences and similarities of granulation in the audio and video domains, but an interesting field for future exploration is the interaction *between* both streams. Some of the effects explained on this document, including cloning and randomization, can be applied to audio and video signals. Highly coupled audio-visual pieces can be generated if both streams undergo similar operations simultaneously. Moreover, an audio granular synthesizer can be used to generate the soundtrack of a granular video, and audio events can be triggered by the video grains. Characteristics such as the density or frequency content of the audio grains can be mapped to other features in the video grains. Finally, spatialized sound and 3D audio compositions that place audio grains at different positions within a virtual space could be complemented by video grains moving with similar dynamics throughout an immersive environment.

One outcome we are currently pursuing is the creation of a more versatile graphical tool similar to the ones that exist for audio granulators and granular synthesizers [3]. A tool for the real-time granular manipulation of video streams would be useful for promoting novel techniques, and would enable their use in a variety of creative situations, including the creation of musical videos, or in live performances by VJs and other visual experimenters.

REFERENCES

[1] C. Roads, Microsound. The MIT Press, Sep. 2004.

[2] U. Zölzer, et al, DAFX: Digital audio effects. Wiley, May 2002.

[3] T. Opie, "Granular synthesis: A granular synthesis resource website," Available: http://granularsynthesis.com/software.php, Retrieved: December, 2013.

[4] G. Levin, "An informal catalogue of slit-scan video artworks and research," Available: http://www.flong.com/texts/lists/slit_scan/, Retrieved: December, 2013.

[5] T. Strothotte and S. Schlechtweg, Non-photorealistic computer graphics: Modeling, rendering and animation. Morgan Kaufmann, 2002.

[6] S. Fels, E. Lee, and K. Mase, "Techniques for interactive video cubism (poster session)," in Proceedings of the eighth ACM international conference on Multimedia (MM). ACM, 2000, pp. 368–370.

[7] A. Cassinelli. and M. Ishikawa, "Khronos projector," in Emerging Technologies SIGGRAPH 2005. Los Angeles, CA: ACM, 2005.

[8] A. W. Klein, P.-P. J. Sloan, A. Finkelstein, and M. F. Cohen, "Stylized video cubes," in Proceedings of the 2002 ACM SIGGRAPH/Eurographics Symposium on Computer Animation (SCA). New York, NY, USA: ACM, 2002, pp. 15–22.

[9] J. Villegas and A. G. Forbes, "Interactive non-photorealistic video synthesis for artistic user experience on mobile devices," in Proceedings of the International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM), Scottsdale, Arizona, January 2014.

[10] A. G. Forbes, C. Jette, and A. Predoehl, "Analyzing intrinsic motion textures created from naturalistic video captures," in Proceedings of the International Conference on Information Visualization Theory and Applications (IVAPP), Lisbon, Portugal, January 2014.

[11] G. Heinzel, A. Rudiger, and R. Schilling, "Spectrum and spectral density estimation by the discrete fourier transform (dft), including a comprehensive list of window functions and some new at-top windows," Max-Planck-Institut fur Gravitationsphysik, Tech. Rep. 395068.0, February 2002.

[12] E. Guigon and A. Pierre, L'oeil moteur: Art optique et cinétique, 1950-1976. Strasbourg, Germany: Muses de Strasbourg, 2005.

[13] C. Roberts, A. G. Forbes, and T. Höllerer, "Enabling multimodal mobile interfaces for musical performance," in Proceedings of the International Conference on New Interfaces for Musical Expression (NIME), Daejeon, Korea, May 2013.

[14] A. G. Forbes, T. Höllerer, and G. Legrady, "Generative fluid profiles for interactive media arts projects," in Proceedings of the International Symposium on Computational Aesthetics in Graphics, Visualization, and Imaging (CAe), Anaheim, California, July 2013, pp. 123–129.

[15] J. Villegas and A. G. Forbes, "Double-meaning: Interactive animations with simultaneous global and local narrative," in Proceedings of the Renew Digital Arts Festival, Copenhagen, Denmark, October-November 2013, pp. 300–304.

 ISBN: 978-1-61208-320-9

Using Grammar Induction to Discover the Structure of Recurrent TV Programs

Bingqing Qu
University of Rennes 1
French National Audiovisual Institute
bqu@ina.fr

Félicien Vallet, Jean Carrive
French National Audiovisual Institute
fvallet@ina.fr, jcarrive@ina.fr

Guillaume Gravier
CNRS
IRISA & INRIA Rennes
guillaume.gravier@irisa.fr

***Abstract*—Video structuring, in particular applied to TV programs which have strong editing structures, mostly relies on supervised approaches either to retrieve a known structure for which a model has been obtained or to detect key elements from which a known structure is inferred. In this paper, we propose an unsupervised approach to recurrent TV program structuring, exploiting the repetitiveness of key structural elements across episodes of the same show. We cast the problem of structure discovery as a grammatical inference problem and show that a suited symbolic representation can be obtained by filtering generic events based on their reoccurring property. The method follows three steps: *i*) generic event detection, *ii*) selection of events relevant to the structure and *iii*) grammatical inference from a symbolic representation. Experimental evaluation is performed on three types of shows, viz., game shows, news and magazines, demonstrating that grammatical inference can be used to discover the structure of recurrent programs with very limited supervision.**

Keywords*—TV program structuring; symbolic representation; structural grammar induction; unsupervised approach.

I. INTRODUCTION

Large scale audiovisual archives offer an extremely abundant digital TV program library for users and content providers. For instance, the French National Audiovisual Institute [1], a repository of French radio and television audiovisual archives, has more than five million hours of radio and television programs stored. However, in order to be useful for later usage such as Internet diffusion, browsing and sharing, such large-scale archives need to be appropriately indexed. In particular, structuring programs, i.e., obtaining a temporal segmentation of programs into their basic constituents, is a crucial step for high-quality indexing, enabling better description as well as direct access to meaningful excerpts.

TV program structuring consists in automatically recovering the structure of a program from the video material, where structure refers to the way in which a program is organized by editors. In the video, the underlying program structure is often reflected in editing rules. Also, the structure of a program is consistent across the different episodes. For example, as described in [2], TV news programs usually start with a brief outline of the reports announced by the anchorperson. Headlines are followed by an alternation of anchorperson's announcement of the upcoming topic and news reports. Most news programs end with interview segments, weather forecasts or program trailers. Program structuring aims at detecting the existence and the temporal boundaries, i.e., the start and end frames, of such constituting elements designated as the structural elements, or events, of the program. In the framework of recurrent programs, i.e., of programs for which several episodes are available, a structural element refers to a video segment with a particular syntactic meaning and which can be regularly found in different episodes.

In this paper, we report ongoing work investigating grammatical inference to discover the basic structural elements as well as their temporal ordering, i.e., the temporal structure, by analyzing a collection of episodes from the program with minimal prior knowledge about the program genre and about the type of structural elements which might be present. In particular, we make very limited assumptions on what structural elements should be, as opposed to supervised approaches which seek to retrieve structural elements previously deemed as relevant for a type of program. To skirt the issue of not knowing which structural elements to look for, we exploit the repetitive nature of recurrent TV programs. A recurrent TV program [3] is a program with multiple episodes which are periodically broadcasted, e.g., daily, weekly. News, entertainments, games and magazines are typical recurrent programs. Most episodes, if not all, follow the same editorial structure: structural elements appear in almost the same order with very similar durations, separated by sequences which repeat across episodes at more or less the same time instants. This last property is successfully used in [3] to detect separators. We adopt a similar idea, further exploiting separators to yield a symbolic representation of episodes suited for grammatical inference. By searching for recurrent elements throughout the episodes and selecting the ones which are relevant to the structure, one can infer the grammar of the show, i.e., the time ordered sequence of structural elements that each episode follows. Assuming such a grammar can be found, a model of the structure of the show can then be established to process additional episodes.

As a proof of concept, we focus here on grammar inference, implementing a three steps approach based on the sole visual modality. Firstly, a batch of broad scope event detection tools are used to find various types of events in all episodes. Secondly, events detected are analyzed across episodes to select the ones relevant to the structure of the program. Finally, a symbolic representation is derived from the segmentation given by structural element s and grammatical inference is applied to yield a grammar of the program by analyzing the symbolic time-ordered representation of each episode.

The rest of the paper is organized as follows. Section II

 ISBN: 978-1-61208-320-9

reviews the existing techniques for TV program structuring. Section III explains the overall method and details each step towards grammar induction for unsupervised TV program structuring. Experimental evaluations are reported in Section IV, followed by conclusions and perspectives in Section V.

II. Related work

TV program structuring has been extensively studied, almost exclusively relying on supervised approaches, which can be classified in two categories, depending on whether a particular program is targeted or not.

Previous work on TV program structuring mainly focused on the case where information on the structure is available as prior knowledge, thus enabling the use of supervised classification techniques. This is typically the case when targeting news or sports, two domains which have received tremendous attention (see, e.g., [2], [4], [5], [6]). Assuming the entire structure of the program is relevant, models of the structure can be learned from annotated data. Hidden Markov models, in multiple variants, have been widely used to this end [7], [8]. Event detection has also been used as an alternative to structure modeling. In this case, models are designed for the events of interest, e.g., goal or penalty in soccer, violent scenes in movies, and parameters are estimated on annotated training data.

Resorting to prior knowledge on the structure offers relatively accurate structuring algorithms but is limited by nature to specific types of programs, requiring training data for each new type. As an alternative, research has targeted the detection of typical structural elements, i.e., of events related to the structure independently of the different program types. Such elements are very diverse, such as anchorperson [9], [10], typical scenes or separators [3], [11] as defined by their repetitiveness. Separators, which are short sequences separating the different parts of a program, are of particular interest to unsupervised structuring. However, program-independent structural elements are insufficient to yield a complete program structure.

Work reported in this paper attempts to make the best of both worlds, i.e., being independent of the program type and obtaining a complete model of the structure.

III. Structure discovery with grammatical inference

The global objective we are targeting consists in inducing a structural grammar for a recurrent TV program, taking advantage of the existence of a collection of episodes of the same program. To avoid any confusion, the term *program* refers to the name of the show, thus being disconnected from video material, while *episode* refers to an exemplar of the program. A structural grammar is comprised of the set of structural elements, representing the different elements which compose each episode, and their temporal order. For example, report and anchorperson's announcement are structural elements for a news program, on top of which a temporal model with occurrence probabilities can be built to form a structural grammar.

A natural idea to discover the temporal model is to use grammatical inference on a symbolic representation of the video material representing each episode. The main difficulty

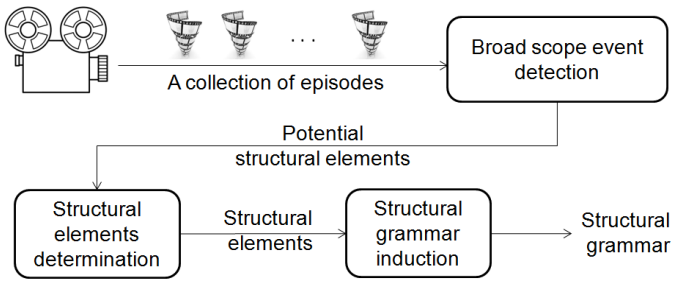


Fig. 1: General architecture of the three steps approach for the grammatical inference of a program structure.

therefore lies in obtaining a symbolic representation of the episodes in the absence of prior knowledge on what structural elements are to be considered. To solve this problem, we take advantage of the redundancy of information across all episodes to obtain symbols in an almost unsupervised manner. Redundancy appears in the structural elements, as well as in the so-called separators which recurrently occur between structural elements. The key idea of our method is therefore to analyze recurrent separators to in turn find recurrent elements which are deemed structural elements and which can be used for grammatical inference.

Targeting this general idea, we propose a three steps approach as illustrated in Figure 1. In the first step, a set of broad scope event detectors are used to find events within each episode, which might be of interest as a potential structural element or as a potential separator. In the second step, we assess the set of events detected along two lines. Density estimation is used to assess repeatability, i.e., to find events which recurrently occur at about the same instant in each episode. Role recognition is further used to assign properties to structural elements so as to help in deriving a symbolic representation. Finally, we induce the structural grammar of the program by leveraging multiple sequence alignment techniques.

One of the benefits of this approach is the very limited supervision that is required, thus making it possible to virtually apply the method on any collection quite straightforwardly. In particular, no data annotation is required at any step of the process. Apart from the selection of episodes, minimal prior knowledge on the program genre is required to select relevant event detectors in the first step and, in the second step, to deduct a meaningful symbolic representation from the set of structural elements found.

We detail in turn each of the three steps.

A. Detection of broad scope events

Ideally, a large number of event detectors should be used, which are generic enough to apply to a large number of shows. This is however not very practical because of implementation and run time issues. A number of key event detectors must therefore be selected based on a trade-off between the type of programs to process, the complexity at run time and, to a lesser extent, the implementation complexity. In this work, five event detectors were used as described below. Note that only minimal knowledge on the type of programs to process was considered, resulting in general purpose detectors, which were

 ISBN: 978-1-61208-320-9

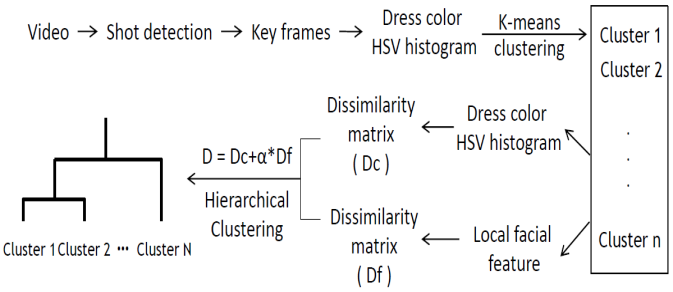


Fig. 2: Flowchart for person clustering

applied to all types of programs in evaluation.

Shot detector. Shots are basic units for program structure which are required for structuring purposes. Shot boundaries were detected using the implementation of J. Mathe et al. [12] which seeks for hard cuts.

Dissolve detector. Gradual transitions are also common in TV programs, which usually signal the start and end of a scene. We focused on dissolve transitions, the most common gradual transition. Dissolve were detected using an extension of [13] in which the dissolve feature description was improved using double chromatic difference.

Monochrome image detector. Monochrome images—mostly white or black—are usually added to the TV streams at edition time to separate the different parts of a program. Detecting monochrome images is therefore an obvious choice for structure discovery in TV programs, in particular due to the simplicity of implementation. In this work, monochrome images are detected by simply verifying the histogram variance of the images.

Person clustering. Persons are essential features for almost any type of TV program. Especially, in many programs, a few number of key persons appear and are strong structure cues, such as the anchorperson in news shows or the host in game shows. TV show conductors usually appear as the most dominant face in a program, i.e., the one which appears most. Dominant person is usually implemented using person clustering based on faces and clothing. Taking news as an example, the anchorpersons clothes are usually carefully chosen so as to be easily distinguishable from guests (and one from another in case of multiple anchors) and obviously do not change within an episode. Person clustering was implemented using Viola and Jones face detection [14] and dress bounding box determination [15]. A key frame of each shot with a face is firstly otained. For each key frame, face features [16] and clothing histograms are then used in a two-step clustering algorithm, shown in Figure 2, based on K-means and hierarchical clustering to obtain person clusters.

Shot reverse shot detector. Shot reverse shot is a classical video technique depicting shots alternating between two characters facing one another, usually engaged in some face to face interaction. In the case of TV programs, we assume that a segment of shot reverse shot represents a dialog and that such interaction between two characters are relevant to the structure. Based on the results of person clustering, we detect shot reverse shot segments by a straightforward analysis of the cluster interlacing.

B. Determination of structural elements

Events detected in the first step are obviously not all relevant to the structure of a program and cannot be used as is to obtain a symbolic representation of episodes. For instance, monochrome images appear between two parts of a game show as a separator, i.e., a structural element, but can also be found at other places such as in a night scene. The second step therefore consists in selecting valid structural elements from which the symbolic representation is obtained. Selection of structural events implements two complementary strategies. On the one hand, role recognition is used to further characterize the outcome of person clustering and identify important persons. On the other hand, the temporal distribution of the events across episodes is analyzed to find elements which repeat with relative temporal stability.

Role recognition. Role recognition enables determining the role of each person cluster resulting from face and clothing clustering. We mostly focus here on the conductor, or anchor, role which is clearly a strong cue with respect to structure. The first characteristic of the anchor is that he/she appears frequently, at more or less regular intervals and at key places, e.g., start and end of the each episode.

We used similar features as those defined in [17] to characterize a cluster, viz.: total duration of appearance (TFL); total number of distinct appearances, i.e., number of non consecutive clusters (TFT); duration of the longest segment in which the person appears (LFL); time range between the first and last occurrence (FR); duration in which the speaker is engaged in a dialog (DPT). Given such features, a dominant person is assumed to ideally have the following properties: he/she is the one that appears most (highest TFL); he/she is filmed the most frequently (highest TFT); he/she participates the most frequently in dialogs (highest DPT); his/her range of appearance (FR) and longest time of appearance (LFL) should not be the lowest. To account for varying episodes and programs lengths, all five features are scaled in $[0, 1]$. Decision on the dominant person is taken based on the sum of the five normalized features, the cluster for which the sum is maximal being identified as the dominant person.

Density filtering. In addition to role recognition, we exploit the property of repeatability across episodes to filter out events which occasionally appear in some episodes and which are therefore not relevant for the global structure of the program. Since different episodes of a recurrent program share a common temporal structure and have almost identical structural elements, the latter should appear in the vast majority of episodes and at about the same time.

Identifying elements which appear frequently at roughly the same time in all episodes is performed using temporal density analysis with a kernel function. For each type of event, we project onto the temporal axis the occurrences across all episodes of the event considered, counting the number of occurrences. To smooth out limited time variations, a kernel density estimation is used based on the following function

$$f(x;h) = \frac{1}{nh^d} \sum_{i=1}^{n} K(\frac{x - x_i}{h}) \ , \qquad (1)$$

where $f(x;h)$ is the estimated density function, h the bandwidth and $K()$ is the kernel. A Gaussian kernel was used with

 ISBN: 978-1-61208-320-9

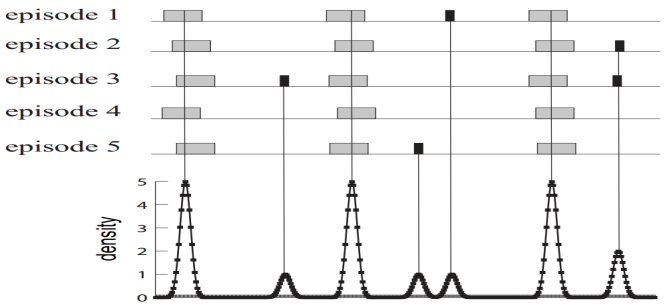


Fig. 3: Example of temporal density filtering to select structural events corresponding to separators (from [3])

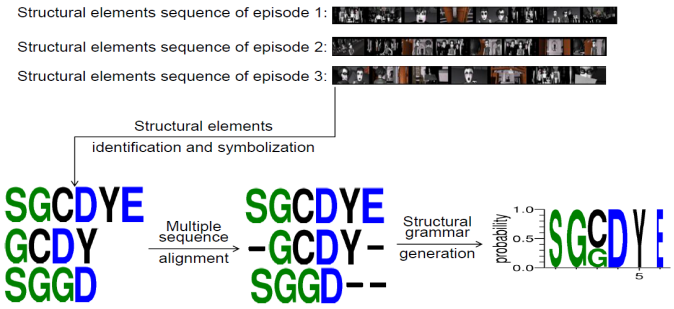


Fig. 4: Illustration of structural grammar induction from three episodes with symbols SGCDYE.

optimal bandwidth automatically chosen [18]. Peaks in the density estimation identify segments with high structural power due to their repetitiveness across episodes. Structural elements are selected based on thresholding of the density function to retain only relevant separators. This process is illustrated in Figure 3, where events in black are removed while the ones in gray are retained because of their high temporal concentration

C. Structural Grammar Induction

As a result of structural element determination, each episode can be represented as a time-ordered sequence of symbols with one symbol per structural element. Selecting and identifying valid structural elements for a program requires semantic interpretation of the structural elements detected via role recognition and temporal density filtering. This identification in turn requires minimal prior knowledge. For instance, a structural element corresponding to a sequence of white frames is a separator, while a long duration shot containing the dominant person at the beginning of the program is the conductor's opening.

Based on simple rules to identify valid structural elements, each episode is represented as a symbolic sequence depicting the succession of valid structural elements. The symbolic sequences corresponding to the different episodes of a program are usually similar due to the temporal stability of recurrent TV program structure. However, slight differences still exist between different episodes. Inferring a grammar from such sequences can be done by discovering the common patterns across symbolic sequences, a problem which is straightforwardly handled via grammatical inference. We adopt multiple sequence alignment techniques which can align the symbolic sequences in the way that alphabet symbols, i.e., valid structural elements, in a given position are homologous, superposable or play a common functional role. Many multiple sequence alignment tools have been developed, e.g., T-Coffee, MAFFT and ClustalW. We adopt in this work the ClustalW algorithm [19], which is one of the most widely used sequential tools for multiple sequence alignment due to its high accuracy, effectiveness and free availability [20]. While more complex grammatical inference techniques exist, based on regular expressions or context free grammars [21], we limited ourselves to multiple sequence alignment to study the meaningfulness of symbolic representations derived in an unsupervised manner.

The process of grammar induction from a symbolic representation is illustrated in Figure 4 with three episodes. Symbols, i.e., SGCDYE, represent the valid structural elements

TABLE I: Description of the data used for evaluation

Dataset	Date	Episodes	Type	Average duration
GAME	Sep. - Dec. 1991	12	Game	37.4 m
NEWS	Jan. - Dec. 2007	12	TV news	36.9 m
MAG	Jan. - Jul. 1997	12	Magazine	56.2 m

that were identified. A graphical representation of the resulting grammar was obtained using WebLogo [22]. A stack of symbols is used to illustrate each position in the grammar: The height of objects within the stack indicates the relative frequency of each symbol while the stack width is proportional to the fraction of valid symbols in that position.

IV. Experimental results

Experiments are conducted on three recurrent programs from different types, viz., game, news and magazine. Evaluation considers in turn the three steps of our method for structural grammar inference. We first measure performance of the event detectors considered. Second, we evaluate structural element determination. Finally, we discuss the grammar inferred for each of the three programs to assess their relevance.

A. Data set description

Three different programs, with 12 episodes each, are used for inference and evaluation, as given in Table I. Two programs, *Que le meilleur gagne* (GAME) and *20h News* (NEWS), were taken with episodes selected over a large time period spanning 1991 to 2007. *Que le meilleur gagne* is a game show with four parts divided by separators. The program, leaded by a conductor, mainly contains interview scenes and questions/answers scenes with full text segments. The daily news show follows a standard pattern for such shows. The third program *Thalassa* (MAG) is a magazine about sea stories, where episodes were taken over a single year (1997). A conductor leads the show which is composed of reports and discussions. While the same conductor appears in all the episodes of GAME, two distinct conductors can be found both for NEWS and MAG.

B. Performance of broad scope detectors

Initial general purpose detectors are a key to subsequent steps and must therefore exhibit an acceptable level of accuracy. We report here evaluation of dissolve transition detection and person clustering. Shot detection and monochrome image detection are based on standard techniques which yield very

 ISBN: 978-1-61208-320-9

TABLE II: Performance of dissolve transition and person clustering

Dataset	Dissolve			Person
	Prec.	Rec.	F	Purity
GAME	0.62	0.79	0.70	72.7%
NEWS	0.64	0.81	0.71	49.5%
MAGS	0.52	0.95	0.67	69.6%

high accuracy. Performance for shot reverse shot detection is directly linked to performance of person clustering.

Results are presented in Table II. Dissolve detection is evaluated in terms of recall and precision. Detectors perform similarly for the three programs with correct recall and precision rates. Person clustering is evaluated by means of cluster purity. While correct purity values are obtained for GAME and MAGS, purity is rather low on NEWS. This is likely due to the fact that a fairly large number of persons appear in news programs and that scenes of reports in news are very variable with non discriminative clothing. Nevertheless, clustering results were found reliable enough for structure inference.

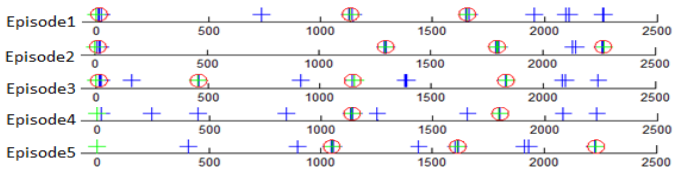
C. Structural elements detection

Selection of structural elements from the output of event detectors is evaluated both qualitatively and quantitatively.

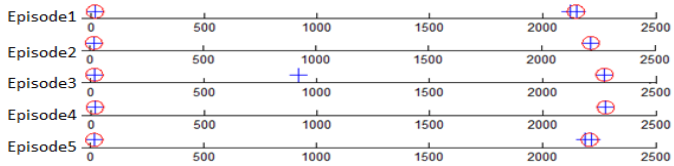
Regarding temporal density filtering to select recurrent events, we observed for GAME that peaks in the density coincide for monochrome frames and for dissolve transitions, with a relatively consistent number of such regions, i.e., 3 or 4 per episode. For NEWS, monochrome images were found to be often around the same temporal positions. Additionally, two short sequences of monochrome images are found in each episode, resp. at the beginning and end of the episode. Positions of events and separators are illustrated in Figure 5. Based on their characteristics, such elements are considered as separators on both programs. For evaluation purposes, a separator is considered as correctly detected when overlapping with a reference separator as annotated in the data. Precision and recall in GAME are resp. 94 % and 67 %. In NEWS, where we have a stronger structure, a recall of 96 % is achieved with similar precision as in GAME.

Dominant person detection succeeded in 83 % of the episodes for GAME, 92 % for NEWS and 75 % for MAG. NEWS are obviously easier with clear features identifying the anchorperson as the dominant speaker: fewer gestures, large area showing clothing, neutral and still facial expression, etc. In spite of relatively low cluster purity, dominant person detection is accurate. At the other end of the spectrum, MAG is a difficult content with long interviews for which occurrence time of the interviewee is almost as long as that of the anchor. Results in grammatical inference below indicate that limited accuracy in MAG did not hurt structure inference, even in spite of the limited number of episodes considered.

Moreover, identifying monologues of the dominant person turned out to be fairly easy based on the duration of the shots containing the dominant person. Dialog segments were also accurately determined combining person clustering and shot reverse shot detection, with performance directly proportional to the purity of the person clusters.



(a) GAME



(b) NEWS

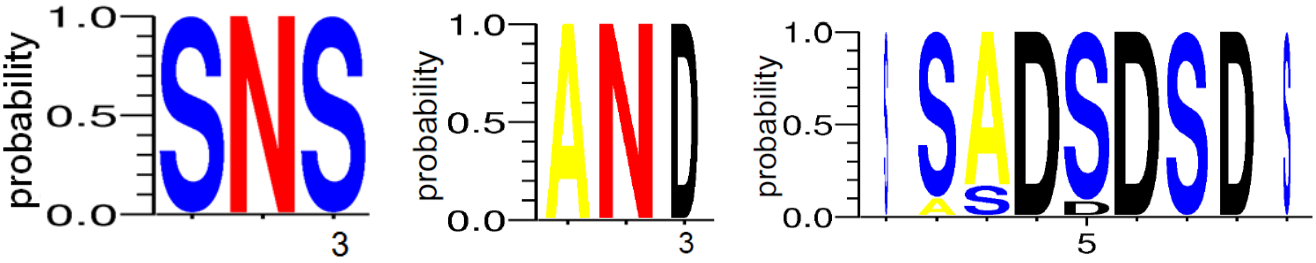
Fig. 5: Examples of separators for five different episodes of GAME and NEWS, where "+" in green (resp. blue) represents monochrome images (resp. dissolve) and "o" represents separators

D. Structural grammar inference

Results for grammar induction are illustrated in Figure 6 for the three programs. For NEWS, separators, denoted as S, are first determined. The segment between two separators, accounting for most time of the program, is considered as news content, denoted as N. The grammar of NEWS is therefore that of an introduction, followed by news reports, followed by a conclusion, a grammar shared by all shows. For MAG, selected symbols are based on dominant person's monologue and dialogs segments, yielding a simple deterministic grammar. A continuous segment with long duration, denoted N, is considered as a report while A and D represent anchorperson's monologue and dialogs respectively. For GAME, separators, dominant person (i.e., monologues of the conductor) and dialogs are the valid structural elements, resp. denoted S, D and A. The grammar inferred is more complex than for NEWS and MAG, reflecting the greater variability across shows. The main syntax is as follows: The game starts with an introduction (separator) followed by a dialog (between the anchor and the participants). We then have an alternating pattern of anchor (dominant person) and game phases (appearing as separator).

With very limited supervision, i.e., basic prior knowledge about TV program structure, possible structure is yielded for each program. All the identified structural elements in this ongoing work are the most common ones, so little bias caused by prior knowledge influences the final structural grammars. Obviously, all three grammars are concise because of the simple symbolic representation that was adopted. Yet, each one represents the structure of the corresponding program, thus demonstrating that grammar inference can efficiently handle structure inference in videos.Results on the game program interestingly prove that non deterministic video structures can be discovered. This last result opens the door to inferring structures at a finer grain. Considering symbolic description at a finer granularity will increase grammar diversity and complexity, which we believe can be handled via probabilistic grammar inference. Taking NEWS as an example, the news content, coarsely denoted N, can be divided into a repeating alternating an anchor's introduction and a report. This can be reflected in a grammar, either using multiple sequence align-

 ISBN: 978-1-61208-320-9



(a) Grammar of NEWS (b) Grammar of MAG (c) Grammar of GAME

Fig. 6: Grammars induced for resp. NEWS, MAG and GAME. See text for details.

ment as considered here (assuming the number of reports is comparable across news episodes) or relying on more complex grammatical inference techniques with better factoring and generalization capabilities.

V. Conclusion

Preliminary work described in this paper shows how a symbolic representation suited for grammatical inference can be obtained from a collection of episodes of the same program, with almost no supervision and no specific training data. Starting from a set of general purpose event detectors, structural elements are derived with minimal prior knowledge by exploiting role recognition and recurrence across episodes. Grammatical inference finally brings a final layer of abstraction by evidentiating the overall structure of the program from the joint analysis of multiple episodes. Experimental evaluation on three types of programs shows that coarse yet relevant structures can be discovered from examples, even for non deterministic programs structures.

Results reported here mostly hint that unsupervised video structuring in recurrent collections using grammatical inference is viable and deserves further attention. The framework proposed in this paper as a proof of concept remains general and can be extended in a number of directions. Obviously, obtaining a symbolic description at a finer grain with limited supervision has yet to be achieved. Increasing the number of general purpose detectors and targeting multiple modalities seems like the most natural path to follow. But adding detectors will challenge the determination of structural elements and the grammar induction step, requiring more elaborate grammars to be considered.

Acknowledgment

The authors wish to acknowledge the help of François Coste and Vincent Claveau, IRISA, for their advise. They in particular have been very useful in guiding the choices made regarding grammar induction.

References

[1] *http://www.institut-national-audiovisuel.fr/en/home*.

[2] H. Zhang, Y. Gong, S. W. Smoliar, and S. Y. Tan, "Automatic parsing of news video," in *Multimedia Computing and Systems, 1994., Proceedings of the International Conference on*. IEEE, 1994.

[3] A. E. Abduraman, S.-A. Berrani, and B. Merialdo, "An unsupervised approach for recurrent tv program structuring," in *of the European Interactive TV Conference*, 2011.

[4] M. Bertini, A. Del Bimbo, and P. Pala, "Content-based indexing and retrieval of tv news," *Pattern Recognition Letters*, 2001.

[5] H. Li, J. Tang, S. Wu, Y. Zhang, and S. Lin, "Automatic detection and analysis of player action in moving background sports video sequences," *Circuits and Systems for Video Technology, IEEE Transactions on*, 2010.

[6] L.-Y. Duan, M. Xu, Q. Tian, C.-S. Xu, and J. S. Jin, "A unified framework for semantic shot classification in sports video," *Multimedia, IEEE Transactions on*, 2005.

[7] L. Xie, P. Xu, S.-F. Chang, A. Divakaran, and H. Sun, "Structure analysis of soccer video with domain knowledge and hidden markov models," *Pattern Recognition Letters*, 2004.

[8] E. Kijak, G. Gravier, L. Oisel, and P. Gros, "Audiovisual integration for tennis broadcast structuring," *Multimedia Tools and Applications*, 2006.

[9] A. Hanjalic, R. Lagensijk, and J. Biemond, "Template-based detection of anchorperson shots in news programs," in *Image Processing, 1998. ICIP 98. Proceedings. 1998 International Conference on*. IEEE, 1998.

[10] X. Gao and X. Tang, "Unsupervised video-shot segmentation and model-free anchorperson detection for news video story parsing," *Circuits and Systems for Video Technology, IEEE Transactions on*, 2002.

[11] A. E. Abduraman, S.-A. Berrani, and B. Merialdo, "Audio/visual recurrences and decision trees for unsupervised tv program structuring," in *Proceedings of the 8th International Conference on Computer Vision Theory and Applications (VISAPP)*, 2013.

[12] *http://johmathe.name/shotdetect.html*.

[13] A. Mittal, L.-F. Cheong, and L. T. Sing, "Robust identification of gradual shot-transition types," in *Image Processing. 2002. Proceedings. 2002 International Conference on*. IEEE, 2002.

[14] P. Viola and M. Jones, "Robust real-time object detection," *International Journal of Computer Vision*, vol. 4, 2001.

[15] G. Jaffré, P. Joly *et al.*, "Costume: A new feature for automatic video content indexing," in *Proceedings of RIAO*. Citeseer, 2004.

[16] J. Sivic, M. Everingham, and A. Zisserman, "who are you?-learning person specific classifiers from video," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009.

[17] D. B. Jayagopi, S. Ba, J.-M. Odobez, and D. Gatica-Perez, "Predicting two facets of social verticality in meetings from five-minute time slices and nonverbal cues," in *Proceedings of the 10th international conference on Multimodal interfaces*. ACM, 2008.

[18] Z. Botev, J. Grotowski, and D. Kroese, "Kernel density estimation via diffusion," *The Annals of Statistics*, 2010.

[19] J. D. Thompson, D. G. Higgins, and T. J. Gibson, "Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice," *Nucleic acids research*, 1994.

[20] M. Larkin, G. Blackshields, N. Brown, R. Chenna, P. A. McGettigan, H. McWilliam, F. Valentin, I. M. Wallace, A. Wilm, R. Lopez *et al.*, "Clustal w and clustal x version 2.0," *Bioinformatics*, 2007.

[21] Y. Sakakibara, "Efficient learning of context-free grammars from positive structural examples," *Information and Computation*, 1992.

[22] G. E. Crooks, G. Hon, J.-M. Chandonia, and S. E. Brenner, "Weblogo: a sequence logo generator," *Genome research*, 2004.

 ISBN: 978-1-61208-320-9

Person Tagging in Still Images by Fusing Face and Full-body Detections

Vlastislav Dohnal and Alexander Matecny
Faculty of Informatics
Masaryk University
Brno, Czech Republic
Email: dohnal@fi.muni.cz / shanio@mail.muni.cz

***Abstract*—We address the problem of organizing personal photo albums by assigning tags/names to people present in photographs. Our proposed framework improves similar systems such as Google+ Photos (Picasa) or Apple's iPhoto by incorporating not only a face detector, but also a full-body detector. Both these modalities are combined together to provide the user with tags of people whose face has not been detected or is not even present in the photograph. An implementation of the proposed framework is evaluated on a sample of real life photographs. This paper is a "work in progress" contribution to the conference.**

***Keywords*—photo tagging; face detection; full-body detection; feature extraction; multi-modal search**

I. Introduction

The development in portable devices, which are nowadays equipped with a digital camera led to a need for users to organize their photographs in an effective and efficient manner. There are many public web photo galleries that offer management of photo collections, where some of them provide users with advanced functionality such as automatic people tagging. Examples are Google+ Photos (Picasa) and Apple's iPhoto with the iCloud service where face recognition is used.

In this paper, we propose a framework that goes further and exploits not only face recognition, but also figure (full-body) recognition for automatic management of tags of people. The motivation for such a system is to improve tagging of people who were captured in a posterior position (looking away from camera) so their face cannot be obtained. Assuming the fact that people present at an event usually do not change their clothing during that event, we can take detections of people in the same clothing as the presence of the same person and associate them with a tag saying his or her name, so easing the process of tagging people in photo collections.

The remaining part of this paper is organized as follows. We discuss related work in the next section. In Section III, we describe our proposal and its substantial parts in detail. Evaluation on real-life datasets is given in Section IV. The paper concludes with future directions drawn in Section V.

II. Related Work

The MediAssist system proposed in [1] exploits the idea of using body clothing to improve person identification too. However, they do not detect and recognize human figures in photos but rather extract body patch from the location of person's head. Clothing in the body patch region is used to improve quality of face recognition. By analogy, the authors in [2] define a body region based on the person's face position in an image. An RGB histogram is then obtained from the clothing in body region. Finally, an extrapolation technique to obtain upper-body bounding box is given in [3].

Since the figure detection in still photos is much more challenging than detection in richer sources, e.g., thermal images [4] or video streams [5], we focus on figure detectors in more detail. Many figure or pedestrian detectors exploit Histogram of Oriented Gradients (HOG) features [6]. Follow-up papers improve it by combining HOG with other features, e.g., color channels and histograms of flow [7]. An optimization of HOG to multiscale gradient histograms is introduced in [8], where the scale-space image pyramid is approximated to increase the detection speed. Figure detection based on independent body-part detectors is proposed in [9]. They define a deformable model of parts to create a detector not only for figures but in general for various kinds of objects. This principle is used to segment people in 3D movies [10]. Another approach [11] is based on local binary patterns and its compressed variants. The authors show that this technique outperforms HOG. Figure detection reliability is greatly improved by applying tracking to solve people occlusions effectively [5]. A recent survey [12] of figure detectors gives the reader a complete insight into this problem.

III. Framework for Person Recognition

We propose a generic framework for fusing face and figure feature modalities to significantly improve effectiveness of automatic tagging persons in still images organized in personal photo collections. Figure 1 depicts the proposed framework. The users shown in the figure communicates with the framework by making several requests.

First, the photo-collection-upload request and its processing represent the core of framework. It issues an automatic process of recognition and eventual person tagging. It consists of *detection phase* where faces and figures are localized in the photos, *visual feature extraction phase* where specific descriptors capturing visual appearance in detected regions are obtained, and *clustering phase* where such descriptors are compared by a similarity function to create clusters of the same person. These phases are implemented in independent modules emphasized in blue in the figure.

Second, the result of automatic clustering of faces and figures of the same person may not be perfect, so is not even in Google+ Photos, so requests to manage the tags can be made. It includes naming not-yet-known people, removing false positives in clusters (pictures of different people) and merging separate clusters of the same person.

Third, since the process of person tagging is inherently based on comparing visual features in detected regions, i.e., on similarity, the last request a user can make is a similarity search (image content-based retrieval).

In the following, we focus on the automatic cluster creation and its phases, which form the core of the whole proposed framework.

 ISBN: 978-1-61208-320-9

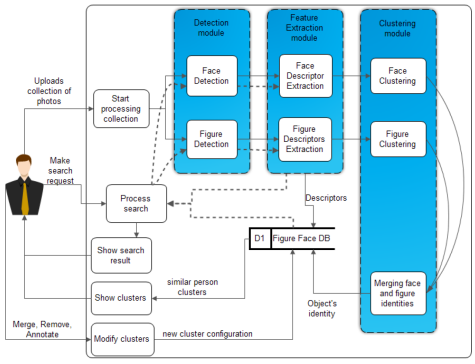


Fig. 1. Schema of framework for people tagging in general photo collections.

A. Detection Module

We use two independent detectors to localize person faces and their figures, which is convenient twofold. It allows running detections in parallel and their implementations can be any of the state-of-the-art methods. The output of the detectors is assumed to be a list of image regions that contain a face or a figure. To detect person faces, we use the Luxand SDK [13] that provides good detection quality, but any other face detector can be used.

For the task of figure detection, we decided to use a method based on Edgelets [14]. This method was designed to work in crowded scenes and an individual human is modeled as an assembly of natural body parts, for each a particular detector is trained by adaptive boosting. This method offers good performance when person is occluded and it is more tolerant to pose and viewpoint change. After training several strong classifiers for detecting individual figures, each image is scanned by a window in scale-space and the classifiers are evaluated. Due to this windowing approach, more detections of the same body can appear, so we join such detections if their overlap exceeds 72%, as defined in (1).

$$\frac{area(R_1 \cap R_2)}{min(area(R_1), area(R_2))} \geq 0.72 \qquad (1)$$

An example of detections and the final merged region is given in Figure 2(b) and Figure 2(c). This merging procedure has a negative effect when large occlusions of bodies appear without all faces being detected, see Figure 2(b). But, this is referred to as a hard problem [7]. To decrease the number of false detections, we filter out all detected regions that cannot be merged with another region. On the other hand, if there is a face detected that coincide with a region of figure detection, it is not filtered out. Both these cases are depicted in Figure 2(e) where both the detections are candidates for filtering out by the first rule. But, the left region is not discarded thanks to its intersection with a face detection, so we take it as reliable enough.

B. Feature Extraction Module

Regions containing persons' faces or their figures detected in uploaded images are passed to the extraction module where descriptors covering visual characteristics are obtained. We use the Luxand SDK again since it offers a high quality face descriptor and a similarity function that perform person identification effectively. This is also confirmed by a comparative study [15] where Luxand SDK in ver. 2.0 exhibited very good values of false rejection and false acceptance rates in a person identification task. Its competitor VeriLook, ver. 4.0 by Neurotechnology, commercial software, offers the same performance but we did not have it licensed. The publicly-available software OpenCV exhibited worse false acceptance rate.

For figure extraction we used the clothing patch covering the central part of body torso. In [14], the torso is defined as the middle part of the whole detected body constrained from top and bottom. To capture the person's clothing as precisely as possible, we have modified this constraint to 0.32–0.58 of the full-body region height and 0.30–0.70 of its width. This was experimentally verified that it maximizes the area of clothing patch while minimizes the influence of background. An example is given in Figure 2(c). A more sophisticated technique to extract clothing based on segmentation can be used [10], [16]. Having obtained a region with clothing patch, we extract one visual descriptor capturing the colors and edges in the clothing patch. In particular, we use a combination of descriptors from the MPEG-7 standard [17]. An experimental evaluation on selecting the best combination is given in Section IV.

Finally, the extraction module produces descriptors consisting of the position and extent of the detected region and the visual descriptor itself for each of the detected faces and human figures. The position and extent are important not only for the clustering module, but also for displaying detections to the user.

C. Clustering Module

This module is responsible for fusing individual detections and their feature descriptors to form groups of images capturing the same person, so a final tag (e.g., person's name) can be assigned to it.

First, all detected faces are separated into clusters by evaluating the Luxand's distance function and the face descriptors whose pair-wise distance is less than 0.14 form a cluster, i.e., describe the same person face. This constant has been experimentally set. In case a different face descriptor or a similarity function is used, this constant must be updated appropriately. Next, the database of known person faces can be searched to identify them and assign their names directly. Currently, we have not implemented such identification yet.

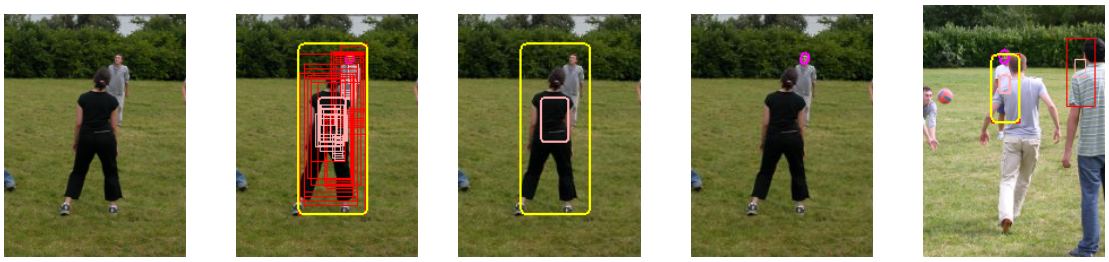
Second, the module proceeds to cluster all detected figures by analogy. For the specific setting of clustering threshold constant on distance and the distance function, please refer to experiments in the next section. Next, the figure clusters are identified by finding correspondences between figure regions and face regions. In particular, we test each figure region in a cluster whether a face region in the same image can be associated with it or not by applying the formula in (2).

$$\frac{area(R_{face} \cap R_{top\_body})}{min(area(R_{face}), area(R_{top\_body}))} \geq 0.10 \qquad (2)$$

It takes the top third of the figure region containing head and shoulders (R_{top_body}) and the face region (R_{face}) and tests their overlap to be at least 10%.

In both the clustering phases, the original image ID from which the detections come, is respected. It obviously assumes that the same person cannot reappear within the same photo, so no two figure nor face detections within the same image can emerge in the same cluster.

 ISBN: 978-1-61208-320-9



(a) (b) (c) (d) (e)

Fig. 2. Figure and face detection result on an image from INRIA person dataset [6]. All detections returned by the Edgelets-based detector are emphasized in red boxes, while the final figure detections after merging are in yellow boxes. The clothing patches (pink boxes) are defined as a region 0.32–0.58 and 0.30–0.70 of height and width of the detection box, respectively. In (a), the original image. In (b), all figure detections with the final detection after merging in yellow; the occluded people are incorrectly merged here. In (c), resulting figure detection with the clothing patch region situated in the middle. In (d), all detected faces. In (e), filtering out figure detections that cannot be merged except detections intersecting a face detection.

IV. Experimental Results

In this section, we give details about training the figure detector and clothing patch extraction since these parts we had to research in order to implement the whole framework proposed in this paper. We also include experience from the clustering people for the task of assigning tags.

A. Figure Detection

To train the figure detector, we used the MIT pedestrian dataset [18] consisting of 914 person images as the positive examples and 1,886 images from the INRIA person dataset [6] as the negative examples.

We tested the quality of trained detector on the ETH person dataset [5], which also includes ground-truth files containing annotations of full-body regions. There are 1,201 person figures in 196 images and our detector correctly identified 817 person figures (68%) and had 29.6% precision (there were other 1,943 detections not containing person figure). We attribute the high number of false positive detections to using a small training set during training detector classifiers. Deeper analysis of this is our future research direction.

B. Clothing Patch Feature Extraction

The task of identifying the person based on their clothing was next challenge. We decided to use a set of global visual descriptors defined in MPEG-7 standard [17]. First, we defined the clothing patch region (see Figure 2(c)) based on our experience with the detector. Second, we picked color structure, scalable color, color layout and edge histogram, since they work on small images and capture not only color and also other visual features. We tested them on the ETH person dataset. From various trials ranging from individual descriptors to their weighted combinations and following the paper [19], we concluded with the combination of scalable color, color structure and edge histogram normalized and weighted in the ratio 5:2.5:1, respectively. This combination reached 0.797 value of Mean Average Precision (MAP), see Figure 3. For space constraints, we do not include other results. The distance constant we used to cluster figures clothed similarly, was set to 1.28.

C. People Tagging

We tested the quality of tagging people on a subset of ETH dataset. We selected 477 images taken from the BAHNHOF sequence. The values of distance used to cluster face and

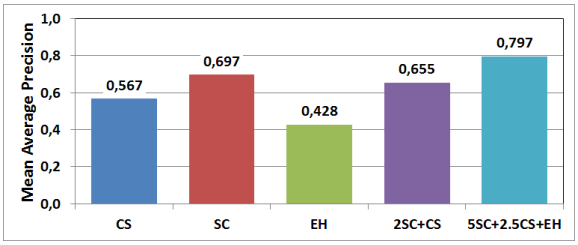


Fig. 3. Comparison in MAP of different global descriptors and their combinations. The last column depicts variant used by ourselves.

figures were set quite strictly, so the automatic process created 290 clusters of faces and figures. There were on average 7.8 clusters per person, a value which was obtained by manual checking the resulting clustering. On the other hand, eight clusters (out of 290) contained different persons, so the clustering failed here. It was caused mainly by people in black coats and their occlusions with tree trunks. The implemented system then allows the user to manage the clusters manually, i.e., to merge clusters of the same person, to name the person, etc.

We have also tested the prototype implementation on a small personal holiday collection that contained 24 photos of people indulging winter activities. Original photographs were of 18 megapixels but we had had to down-sample them to around 400 by 300 pixels since our figure detector was trained on small-resolution images, as have been mentioned above. This toy dataset contains 76 figures, each at least 220 pixels tall (in original resolution), and 54 faces. Many faces are covered with skiing goggles, which makes it very challenging for common face detectors. In these photographs, 26 distinct people were shot, 15 of which only once.

Our face detector revealed 7 faces only, where all were true positives. The figure detector correctly bounded 54 figures out of 78 detections. They were grouped in 5 correct clusters, thus they contained photos of the same person. Three clusters contained different people, but they were all wearing very similar clothes – red, black or red/black jackets. One cluster consists of a face-figure pair of one person. Next, seven clusters can be considered as mixtures of false-positive and true-positive detections, where the clothing patches are alike. Finally, the other detections were not grouped at all, so they resulted in "one-detection" clusters. An example of an automatically created cluster is given in Figure 4. The complete results are

 ISBN: 978-1-61208-320-9

Fig. 4. A cluster of grouped images of the same person. The person's tag has been generated from database cluster ID.

available at http://disa.fi.muni.cz/mmedia2014/.

We tested Google+ Photo by creating an album out of the original high-resolution photos. Google's software detected 16 faces and clustered them into 12 clusters. By manual verification, these detections correspond to 7 people. We attribute these results to the Google's policy to provide its users with high-precision face detections. Surprisingly, no faces were detected in the down-sampled photos.

V. Conclusions

We proposed a framework that combines a face and figure (full-body) detector to recognize people with the aim of providing people tagging in user photo collections. The contribution of this paper is in testing various descriptors for comparing clothing patches to recognize similar clothing, which in other words, leads to identification of the same person in different photographs. This, of course, requires the assumption that people do not change their clothes within a short period of time in which a social event takes place. The other and main contribution is in implementing a prototype using state-of-the-art techniques for face and figure detections and fusing these two modalities into one system. The prototype is available at http://disa.fi.muni.cz/mmedia2014/.

This preliminary prototype can be improved fourfold. First, the face detection module can be changed to use a multi-view face detector [20], which was successfully used in a recent paper on finding actors in movies and assigning their name and actions from movie scripts [3]. Second, preparing an Edgelets detector for not only full-body detections, but also an upper-body detector on a bigger training data set is our next goal. Third, the personal holiday photos do not contain overcrowded scenes very often, so a better alternative would be to apply a detector based on histogram-of-gradient features and latent support vector machines [8]-[9]. Fourth, person occlusion (see Figure 2(b)) can be partly eliminated by training a separate head detector to avoid merging two figure detection if they contain two different heads.

Finally, the proposed system can be used to assign person tags very easily having a better figure detector with low rate of false positives.

Acknowledgment

Vlastislav Dohnal was supported by the Czech Science Foundation, project no. GBP103/12/G084. Alexander Matecny was supported by the Ministry of the Interior, Czech Republic, project no. VG20122015073.

References

[1] N. O'Hare and A. Smeaton, "Context-aware person identification in personal photo collections," Multimedia, IEEE Transactions on, vol. 11, no. 2, 2009, pp. 220–228.

[2] L. L. Presti, M. Morana, and M. L. Cascia, "A data association algorithm for people re-identification in photo sequences," Multimedia, International Symposium on, 2010, pp. 318–323.

[3] P. Bojanovski, F. Bach, I. Laptev, J. Ponce, C. Schmid, and J. Sivic, "Finding actors and actions in movies," in IEEE International Conference on Computer Vision, ICCV 2013, Sydney, Australia, December 3-6, 2013. IEEE, 2013, pp. 1–8.

[4] M. Correa, G. Hermosilla, R. Verschae, and J. Ruiz-del Solar, "Human detection and identification by robots using thermal and visual information in domestic environments," Journal of Intelligent & Robotic Systems, vol. 66, no. 1-2, 2012, pp. 223–243.

[5] A. Ess, B. Leibe, K. Schindler, , and L. van Gool, "A mobile vision system for robust multi-person tracking," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE Press, June 2008, pp. 1–8. [Online]. Available: http://www.vision.ee.ethz.ch/~aess/dataset/[retrieved:Dec.,2013]

[6] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in International Conference on Computer Vision & Pattern Recognition, vol. 2, June 2005, pp. 886–893. [Online]. Available: http://pascal.inrialpes.fr/data/human/[retrieved:Dec.,2013]

[7] S. Walk, N. Majer, K. Schindler, and B. Schiele, "New features and insights for pedestrian detection," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2010, pp. 1030–1037.

[8] P. Dollar, S. Belongie, and P. Perona, "The fastest pedestrian detector in the west," in Proceedings of the British Machine Vision Conference (BMVC). BMVA Press, 2010, pp. 68.1–68.11, doi:10.5244/C.24.68.

[9] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 32, no. 9, 2010, pp. 1627–1645.

[10] K. Alahari, G. Seguin, J. Sivic, and I. Laptev, "Pose estimation and segmentation of people in 3d movies," in Proc. IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1–8.

[11] Y. Mu, S. Yan, Y. Liu, T. Huang, and B. Zhou, "Discriminative local binary patterns for human detection in personal album," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2008, pp. 1–8.

[12] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI), vol. 34, no. 4, 2012, pp. 743–761.

[13] Luxand, Inc., "Luxand face SDK 4.0," 2005-2013. [Online]. Available: http://www.luxand.com/facesdk/[retrieved:Dec.,2013]

[14] B. Wu and R. Nevatia, "Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors," in IEEE International Conference on Computer Vision, ICCV 2005, vol. 1, 2005, pp. 90–97.

[15] N. Degtyarev and O. Seredin, "Comparative testing of face detection algorithms," in Image and Signal Processing, ser. LNCS. Springer Berlin Heidelberg, 2010, vol. 6134, pp. 200–209.

[16] A. Gallagher and T. Chen, "Clothing cosegmentation for recognizing people," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2008, pp. 1–8.

[17] B. Manjunath, P. Salembier, and T. Sikora, Eds., Introduction to MPEG-7: Multimedia Content Description Interface. New York, NY, USA: John Wiley & Sons, Inc., 2002.

[18] C. Papageorgiou, T. Evgeniou, and T. Poggio, "A trainable pedestrian detection system," in Proceeding of Intelligent Vehicles, October 1998, pp. 241–246. [Online]. Available: http://cbcl.mit.edu/software-datasets/PedestrianData.html[retrieved:Dec.,2013]

[19] M. Batko, P. Budkov, and D. Novk, "Cophir image collection under the microscope," in Proceedings of the 2009 Second International Workshop on Similarity Search and Applications. Washington, DC, USA: IEEE Computer Society, 2009, pp. 47–54.

[20] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012, pp. 2879–2886.

 ISBN: 978-1-61208-320-9

A Semi-Automatic Multimodal Annotation Environment for Robot Sensor Data

Konstantinos Tsiakas
Department of CS & Engineering
University of Texas at Arlington
Arlington, TX, USA
Email: konstantinos.tsiakas@mavs.uta.edu

Theodoros Giannakopoulos
Computational Intelligence Lab
Institute of Informatics & Telecomm.,
NCSR 'Demokritos', Athens, Greece
Email: tyiannak@gmail.com

Stasinos Konstantopoulos
Software and Knowledge Engineering Lab
Institute of Informatics &
Telecomm., NCSR 'Demokritos'
Email: konstant@iit.demokritos.gr

***Abstract*—In this paper, we present RoboMAE, a multi-modal sensor data annotation environment that allows humans to concentrate on high-level decisions producing full frame-by-frame annotations. Multi-modal annotation tools focus on interpreting a scene by annotating data on separate modalities. In this work, we focus on the cross-linking of the same object's recognition across the different modalities. Our approach is based on exploiting spatio-temporal co-occurrence to link the different projections of the same object in the various supported modalities and on automatically interpolating annotations between explicitly annotated frames. The backend automations interact with the visual environment in real time, providing annotators with immediate feedback for their actions. Our approach is demonstrated and evaluated on a dataset collected for the recognition and localization of conversing humans, an important task in human-robot interaction applications. Both the annotation environment and the conversation dataset are made publicly available.**

Keywords-multimodal annotation; robotic sensors; human computer interaction

I. INTRODUCTION

Manual annotation is already a laborious, but essential, task in the development of any multimedia analysis system that attempts to assign human-interpretable labels to data; treating *multimodality* makes the annotation task harder, as the alignment of the projections of the same object in the different modalities needs to also be marked. In other words, fully annotating multimodal data requires more effort than the sum of the effort needed for the individual modalities since it is also necessary, for example, to link the speaker recognized in the audio modality with a human figures present in the visual channel.

Several general-purpose multi-modal annotation tools have been designed in the past. For example, ANVIL [1] is one of the most widely used and advanced free video annotation tools, mostly used in the area of multimodal communication research and usually focusing on the modality of speech. In [2] ANVIL has been used for creating a multimodal corpus of particular human actions. Lately [3], ANVIL has been extended by Kinect-based motion analysis procedures. In addition, VisSTA (Visualization for Situated Temporal Analysis, [4]) also focuses on natural multi-modal language annotation.

In this work, we pursued an approach towards a semi-automatic annotation tool for robot sensor data that turns the tables and makes an *advantage* out of the need to simultaneously annotate multiple modalities. We emphasize the need to both *internally represent* and *graphically visualize* the data in a manner that stresses the space and time each individual object and event occupies. In this representation, we exploit spatio-temporal coincidence in order to automatically infer initial annotations and cross-modality object correspondences. Human annotators confirm or correct the automatic annotations in any of the visualized modalities they find more convenient and the cross-modality correspondences carry these over to all modalities. To give a concrete example: if in a scene it is easier to tell who is speaking given his/her voice, then the annotator should only annotate the audio modality and let that carry over to that person's appearance in the other modalities; if in another, more noisy scene it is easier to tell who is speaking from lip movement, then the annotator should only annotate the image modality and let that carry over to that person's appearance in the other modalities.

This achieves a more judicious allocation of annotation effort allowing human annotators to concentrate on high-level decisions regarding the interpretation of a scene, while at the same time producing full frame-by-frame annotations with the same object's appearances across the different modalities cross-linked. In order to make the above more concrete, let us consider the task of scene interpretation for a robot featuring a fairly common sensor inventory: (a) *camera* for obtaining RGB images (b) a passive *stereoscopic camera* or an active *structured light sensor* for obtaining depth images (c) a *microphone* and (d) a *laser range finder* for obtaining planar range scans. Creating a unified perception from these modalities presents us with both an opportunity and a challenge: the opportunity to exploit straightforward, unambiguous recognitions in one modality in order to annotate another and the challenge of how to best represent annotations across modalities and the link between the appearances of the same real-world object in the different modalities.

There will be different levels of natural overlap that can be exploited in order to align modalities into this unified perception. Our particular mixture of modalities exemplifies all of full, partial, and no overlap. More specifically, *full overlap*, as in aligning RGB with depth data, is straight-forward since both modalities are typically recorded from sensors on the same device and are analyzed into objects that almost fully overlap in their shared frame of spatial reference. Compare, for

 ISBN: 978-1-61208-320-9

example, the RGB and depth image in the center of Figure 3. *Partial overlap* occurs in aligning the above with data from the laser range finder: range data is the planar contours of objects at a low height from the ground, typically used for obstacle avoidance. Mapping range data to the RGB-D frame (or vice versa) and looking for overlapping objects is not straightforward and often these contours are outside the field of vision of the RGB-D sensors. Compare, for example, the RGB-D images in the center with the range data in the bottom left of Figure 3: the three pairs of curves in the latter are the contours of the legs of the three people seen in the former, but at a height below what is visible in the RGB-D images. Finally, aligning data from a different space altogether, such as the *audio signal* that only has a temporal dimension and cannot be positioned at all in space. Even using microphone arrays to localize sound would only give us a rough angular position, which cannot be used to geometrically calculate spatial overlap between the sound source and the objects in the RGB-D images or range data.

In the remainder of this paper, we first present the use case and the data collection procedure (Section II) and the RoboMAE multi-modal sensor data annotation environment we have developed (Section III). We then proceed to evaluate our environment (Section IV) and conclude with discussion and future research directions (Section V).

II. Use Case and Data Collection

Our use case is the interpretation by the robot of a *human conversation* scene, an important task in any human-robot interaction application. In order to support the development and evaluation of the relevant sensor data analysis components, we envisaged a graphical tool that facilitates the following cycle:

- the different modalities are visualized simultaneously and in synchronization, including initial automatically derived annotations also presented visually
- the human annotator edits annotations in any individual modality as well as the linking across modalities
- manual edits are used to improve the automatically derived annotations

The cycle repeats until the annotator is satisfied with the quality of the annotations, so that they can be exported for training and testing the robot's recognition components.

The data has been recorded using Sek (Figure 1), a custom-made robot at NCSR 'Demokritos' that has all four sensing modalities mentioned above.RGB and depth from an Xbox 360 Kinect, audio from an Andrea microphone, laser range data from a Hokuyo 30LX laser range finder. The laser scanner is placed almost 10cm above the ground, while the height of the Kinect sensor's position is around 80cm. (For more details please see http://roboskel.iit.demokritos.gr/personnel/sek) We have made nine different recordings, with a total run-time of almost 25 min, where ten volunteers were asked to play out different conversation scenarios of varying difficulty for automatic recognition.

The recorded modalities are synchronized by global timestamps and formatted as follows: audio is 1 sec-long WAV files,

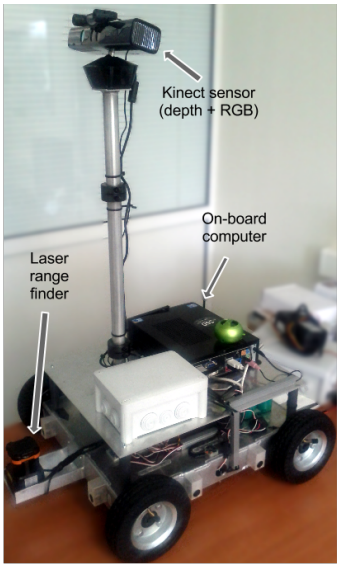


Figure 1: SEK. The robot platform used to record the data.

RBG frames are JPEG files, kinect depth frames are raw binary files, and laser scan data is in a single text file.

III. Semi-Automatic Multimodal Annotation

A. Graphical user interface and manual functionalities

The annotation tool focuses on providing a user-friendly interface for multi-modal annotation of audio, visual and laser data and a set of semi-automatic methods that will utilize the annotation process. It visualizes the different modality data recorded from the robot's different sensors. The user is able to see the data frame from each sensor at any time, as a synchronization procedure of different modality data is embedded.

In Figure 3, we present a screenshot of the implemented annotation tool. The slider control at the bottom of the GUI is used to select the time frame. The upper left display presents a 2-second window of audio, that can be played back. The annotator can zoom in and out of the display to change the size of audio window size. The upper right display is the visual modality while the bottom right display is the depth modality, visualized as gray-scale video.

Finally, the range data display on the bottom left visualizes a planar laser scan. This display can be toggled between two alternative visualizations, showing either the raw polar coordinates or their Cartesian transform.

In case the annotator performs a fully manual annotation task, the annotation can be divided in three main tasks: visual, audio and laser track–depth image mapping. The user has to annotate all frames by using the respective controls. Regarding the labeling of the annotated humans, either the default names (Speaker1, Speaker2, etc.) or any other name can be used. There are two ways to complete a face annotation task. Either by drawing bounding boxes on each frame or by using an interpolation procedure as an assisting tool. For simple cases where the positions of the face bounding boxes do not dramatically change for a particular time period, the annotator

 ISBN: 978-1-61208-320-9

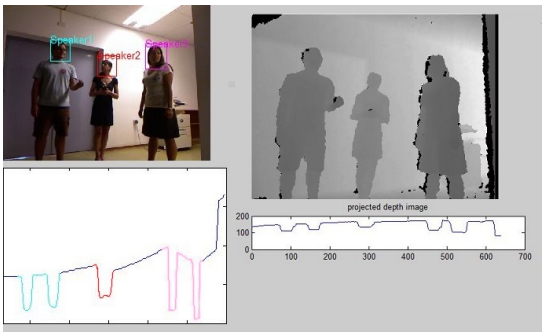


Figure 2: Depth image projection and laser scan mapping.

can use an interpolation procedure applied on each face's position over time. Specifically, the user can annotate some certain frames and use cubic interpolation to automatically annotate the intermediate frames. In this case, the annotator should check the 'Interpolation' choice in the Face Annotation panel for 'Faces', annotate some certain frames and select them as 'Interpolation Points'. After executing this method, the intermediate frames will be annotated. The accuracy of this method depends on the frames' selection.

As far as the sound annotation sub-process is concerned, the user can interactively choose and play a specific audio segment and finally match it to any of the annotated faces. In this way, audio annotation is linked to the RGB annotation described above. The annotator is able to see if a segment is annotated by checking the corresponding text box next to the audio segment figure. Moreover, the user is given the choice of a speaker color-coded view of audio segments.

Regarding the depth image information, each depth image value depicts the distance from the sensor to the specific point. Our purpose in the context of a unified multimedia annotation tool is to associate this depth image information with the laser scan output. This is achieved through a projection calculation of the bottom third of the depth image to the horizontal axis. In the sequel, taking advantage of the similarity between the laser scanner and the depth sensor projection, we define a mapping function that assigns each point of the projected depth image to the laser scan curve. The user can click either on the projection, the depth image or the laser scan curve and select an area of interest with equivalent meaning (Fig.2).

B. Automatic annotation

The safest way (in terms of annotation accuracy) to complete an annotation task is to follow a fully-manual annotation procedure. That means, for example, that the user needs to draw bounding boxes on each frame, annotate each audio segment and each laser scan plot. As this is a tedious process, apart from the manual annotation functionalities in the GUI, RoboMAE integrates recognition techniques, such as face detection, face tracking, speaker diarization, image projection.

1) Visual: Instead of defining face bounding boxes for every single frame (or simply use the interpolation procedure described before), users can employ a face detection approach based on the Viola-Jones algorithm [5]. This can be used as an initial estimate of the face positions in each frame. Apart from the automatic face detection approach, we have also used the Mean Shift algorithm for automatic *face tracking* [6]. The annotator must choose 'Tracking' in the Face Tracking Panel, choose 'Faces' and annotate speakers in a particular frame, either by drawing bounding boxes or by using the Face Detector. The user can then complete the face annotation by choosing 'Tracking frames' to point to the last frame to track and executing the tracking method. Naturally, accuracy depends on the accuracy of the individual manually annotated faces.

2) Audio: Speaker diarization partitions an audio stream into segments denoting speaker identity. In other words, a speaker diarization algorithm answers the question 'Who speaks when?' [7], [8], [9]. Most of the proposed methods on speaker diarization are only based on audio information, however there are also multimodal approaches [10], [11]. Here, we employ semi-supervised learning in order to cluster the audio segments into speakers [8]. The idea is to have the user annotate speaker identity in a small part of the audio signal and then use this information to 'guide' the semi-supervised speaker diarization algorithm. In other words, the user annotates a small number of speech segments and the semi-supervised algorithm returns a fully-annotated stream.

3) Laser track and depth image mapping: Laser scan data is currently annotated fully manually or by interpolation. The user can choose 'Laser' in the 'Face Tracking' panel and—choosing certain annotated frames—to execute the cubic interpolation method described previously.

IV. Usability and Annotation Performance

We have evaluated the *usability* of the implemented tool in terms of the time needed for identically annotating the same data using either the fully manual or semi-automatic approaches. The average annotation time was reduced by 60%, dropping from $562\,\mathrm{min}$ for the fully manual annotation to $219\,\mathrm{min}$ for the semi-automatic annotation.

In order to measure how close the initial automatic annotations were to the fully manual ones, we measured the performance of the *face tracking* and *speaker diarization* modules assuming the fully manual annotations as ground truth. In this experiment, *face tracking* achieves an $\mathrm{F}_{\beta=1}$ measure of 68% and *speaker diarization* a *cluster accuracy rate* of 74%.

V. Conclusion

We have presented RoboMAE, a visual playback and annotation editing environment for multi-modal sensor data. The major innovation in our tool is that it exploits sparse manual annotations in order to *interpolate* a complete frame-by-frame annotation and to *transfer* object recognitions across modalities. By interacting with the visual environment in real time, the backend facilitates starting out with a sparser and effortless annotation that only delves into details where necessary in order to converge to a satisfactory result. Our contribution comprises the complete MATLAB code for RoboMAE and the annotated dataset used in the experiments described here, both publicly available at http://roboskel.iit.demokritos.gr.

We are currently integrating more advanced pattern recognition methods over the laser range data [12], in order to

 ISBN: 978-1-61208-320-9

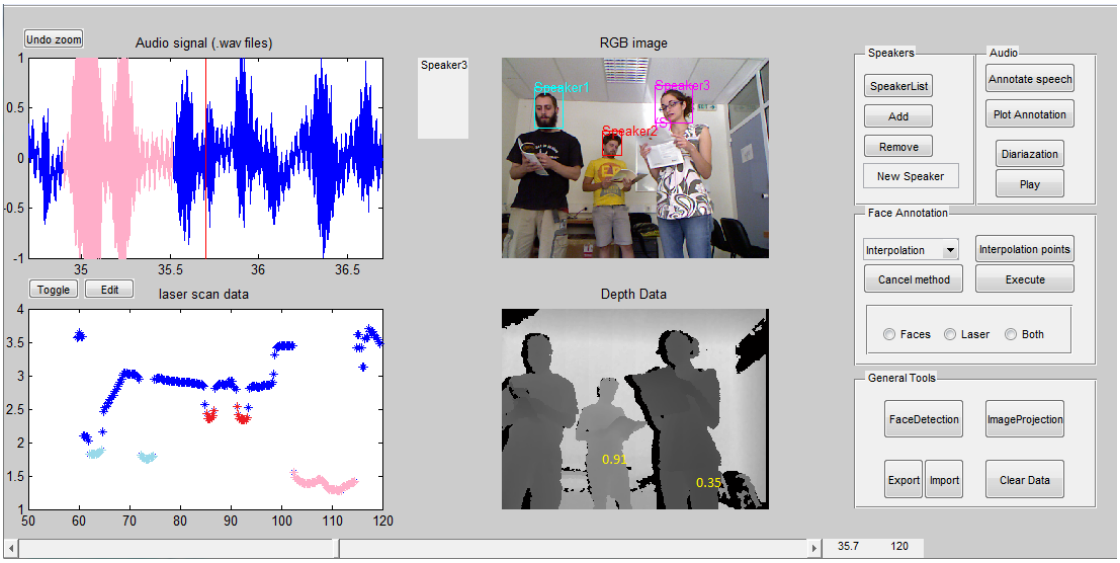


Figure 3: RoboMAE supports manual and semi-automatic techniques to help the user complete multimodal annotation tasks accurately and efficiently.

enhance the automatic annotations that currently only rely on interpolation (cf. Section III-B3). Furthermore, we are extending the heuristics used to transfer annotations across modalities, e.g. by experimenting with skeleton models extracted from depth and used to guide face tracking.

Longer-term plans include taking advantage of the experience gained by developing this first prototype to re-design the architecture of RoboMAE. The further aim is that RoboMAE is not tied to any particular sensor type and automatic recognition method, but to define generic interfaces for the recognition tools used in the back-end.

We will also develop annotation quality metrics that will assist the users decide whether the current annotations are 'good enough' for their purposes or further refinement is needed. One idea is to support a cycle where a small, 'ground truth' portion of the material is annotated in detail and checked thoroughly. As annotation over the rest of the material progresses, this is used to re-train recognition tools and test them over the ground truth, providing an indication of the quality of the current annotations in the larger portion of the data.

ACKNOWLEDGEMENTS

The work described here was partially carried out at the 2013 International Research-Centred Summer School (http://irss.iit.demokritos.gr). and in the context of *Roboskel*, the robotics activity of the Institute of Informatics and Telecommunications, NCSR 'Demokritos' (For more details please see http://roboskel.iit.demokritos.gr).

We would also like to gratefully acknowledge the participation of colleagues and IRSS students in the data collection.

REFERENCES

[1] M. Kipp, "ANVIL: The video annotation research tool," in *The Oxford Handbook of Corpus Phonology*. Oxford University Press, 2014, to appear, pre-print at http://www.anvil-software.org.

[2] M. Swift, G. Ferguson, L. Galescu, Y. Chu, C. Harman, H. Jung, I. Perera, and et al., "A multimodal corpus for integrated language and action," in *Proc. Workshop on MultiModal Corpora for Machine Learning*, 2012, held at LREC 2012, Istanbul, Turkey, 22 May 2012.

[3] M. Kipp, "Annotation facilities for the reliable analysis of human motion," in *Proc. 8th Intl Conf. on Language Resources and Evaluation (LREC), Istanbul, Turkey*, 2012, pp. 4103–4107.

[4] Y. Shi, T. Rose, and F. Quek, "A system for situated temporal analysis of multimodal communication," in *Proc. Workshop on Multimodal Corpora*, 2004, held at LREC-08, Lisbon, Portugal, 25 May 2004.

[5] P. Viola and M. J. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.

[6] D. Comaniciu, V. Ramesh, and P. Meer, "Real-time tracking of non-rigid objects using mean shift," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, vol. 2. IEEE, 2000, pp. 142–149.

[7] S. E. Tranter and D. A. Reynolds, "An overview of automatic speaker diarization systems," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 5, pp. 1557–1565, 2006.

[8] T. Giannakopoulos and S. Petridis, "Fisher linear semi-discriminant analysis for speaker diarization," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 7, pp. 1913–1922, 2012.

[9] X. Zhu, C. Barras, S. Meignier, and J.-L. Gauvain, "Combining speaker identification and BIC for speaker diarization," in *INTERSPEECH*, vol. 5, 2005, pp. 2441–2444.

[10] G. Friedland, H. Hung, and C. Yeo, "Multi-modal speaker diarization of real-world meetings using compressed-domain video features," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*. IEEE, 2009, pp. 4069–4072.

[11] K. Otsuka, S. Araki, K. Ishizuka, M. Fujimoto, M. Heinrich, and J. Yamato, "A realtime multimodal system for analyzing group meetings by combining face pose tracking and speaker diarization," in *Proc. of the 10th Intl Conf. on Multimodal Interfaces*. ACM, 2008, pp. 257–264.

[12] T. Varvadoukas, I. Giotis, and S. Konstantopoulos, "Detecting human patterns in laser range data," in *Proc. 20th European Conference on Artificial Intelligence (ECAI 2012)*, 2012.

 ISBN: 978-1-61208-320-9

Classification of Human Skin Color and its Application to Face Recognition

Marwa Jmal, Wided Souidene Mseddi, Rabah Attia
Electronic Systems and Communication Network Dept,
TUNISIA Polytechnic School
marwa.jmal@telnet-consulting.com,
wided.souidene@ept.rnu.tn, rabah.attia@enit.rnu.tn

Anis Youssef
Head of Innovation Activities
TELNET Group, Tunisia
Technological Park El Ghazala,Tunisia
anis.youssef@groupe-telnet.net

***Abstract*—This paper presents a novel human skin color classification into skin color tones: White and Black. This is performed by developing a skin color classifier based on pixel-based classification using RGB model. Our proposed method is classified under the category of an explicitly defined skin region model. The skin classifier divides our database formed by some images from the FERET set of faces into two sub-databases according to the skin color. The skin color classification method is then applied on a face recognition technique by reducing the number of trained images in the matching process. The performance of the proposed human skin color classifier is evaluated perceptually. Experimental results showed that our proposed skin color classifier is able to classify a face into its possible skin color tone and reaches 87% as hit rate.**

Keywords-Skin color classification; Face recognition; Pixel-based classification; RGB model.

I. INTRODUCTION

Face recognition is a biometric technique, which aims to identify a person from a digital image by comparing its extracted facial features with the ones of images in database. This field has presented for the past decades, the center of extensive research and it was mainly used for security and access control. However, human face image is vulnerable to a lot of variations caused by aging, illumination changes, facial expressions and low resolution, which make it harder for face recognition techniques to acquire interesting discrimination results.

Several approaches of face recognition were developed as a solution to this problem. In this paper, we employ Lowe's Scale Invariant Feature Transform (SIFT) descriptors [6] [13] to detect facial features.

SIFT descriptors are known as the most local invariant feature descriptors. First, they were developed for object recognition systems and have become, recently, the core of many algorithms in computer vision applications. This method transforms an image into local feature vectors, which are invariant to image translation, scaling and rotation, and partially invariant to illumination changes and 3D projective transform. However, SIFT descriptors were designed only for gray images [9]. Thus, the color component in an image grants weighty information for object classification (animals, flowers, faces, etc.). For some sort of applications, such as face recognition, color may be an important distinction tool for discrimination and it has been proven that it is very salutary and robust for applications applied on faces (detection, tracking and recognition).

Human skin color classification finds out to which color tone the skin belongs. The simplest and most employed technique for skin modeling is to explicitly define skin region [12]. The advantage of this method is the simplicity of detection rules which leads to building a very fast classifier. Other skin modeling techniques employing statistical based approaches are involved such as neural networks [7], k-means clustering [3] and Bayesian networks [8].

Unlike it seems to be, skin modeling is complex and quite challenging. In fact, skin color in an image depends mostly on illumination conditions which affect the distribution model of the skin color. Other problems facing skin color classification are shade and shadow occlusions, resolution as well as skin tone variation between races.

The purpose of this study is to develop a skin color classifier into skin color tones in order to improve the face recognition results based on SIFT descriptors.

The rest of the paper is organized as follows: Section 2 is dedicated to the human skin color classification in which our proposed method is detailed. The application of our classifier to face recognition using SIFT descriptors is presented in Section 3. Experimental results are covered in Section 4. In Section 5, the conclusion is drawn.

II. HUMAN SKIN COLOR CLASSIFICATION

In general, the purpose of this study is to enhance a human skin classifier that is able to effectively classify skin color tones. To reach this objective, this section will present the pursued methodology which is divided into two steps: skin segmentation and skin color modeling.

A. Skin segmentation

Precise skin segmentation aims to remove all "non-skin" pixels in order to acquire good results in skin classification. Each image in the database was segmented according to the method proposed in [1]. In this article, a skin segmentation scheme based on RGB (Red, Green, Blue) pixels' color is developed. The model presented in [1] is divided into two rules. A pixel is considered as skin if:

$$0.0 \leq \frac{R-G}{R+G} \leq 0.5 \qquad (1)$$

 ISBN: 978-1-61208-320-9

and

$$\frac{B}{R+G} \leq 0.5 \quad (2)$$

The RGB values of skin pixels were preserved as they are, while the RGB values of non-skin pixels were mapped to [0 0 0]. Fig. 1 exemplifies the skin segmentation process to preserve only skin pixels.

(a) (b)

Figure 1. Example of skin segmentation: (a) Original image, (b) Mask image.

This process excludes different parts like eyes, hair and accessories. Thus, it is not totally effective since it confuses, for example, skin with bright hair such as in Fig. 1 where some parts of the hair were not removed.

This method outputted a mask image (Fig. 1 (b)) and stored it for further applications.

B. Skin color modeling

The fundamental goal of skin color modeling is to create a decision rule that will distinguish between skin color tones. To achieve this objective, this section will describe the methodology which is divided into three main steps as follows:

- Selection of the color space,
- Choice of skin color tones,
- Creation of the decision rule.

1) Selection of the color space: The most important step is to select the color space in order to acquire more accurate classification results. Many color spaces have been involved in the problem of skin color representation and recognition such as RGB, YCrCb (Luminance, Chroma: Red, Chroma: Blue) and HSV (Hue, Saturation, Value). Shin et al. [4] suggest a comparison of the performance of eight color spaces for skin detection. As a conclusion, RGB and YCrCb were the best classified color spaces when dealing with separability. Based on this result, the RGB color space will be used for skin color mapping.

2) Choice of skin color tones: In general, skin color tones are white, yellow, brown and black. In [10], the skin tone set was classified as white, brown and black since the yellow had its tone too close to the white one. However, experiment results had shown that the highest incidence of error was found in the brown skin color set: almost half of brown colored skin was misclassified. From this result, in the current paper, we present a new method for skin color classification into white and black tones based on the RGB model. Brown tone is classified under the black set.

3) Creation of the decision rule: The human skin color classification technique is derived from histograms. In fact, histogram-based segmentation approach is an efficient method for image segmentation given its rapidity in training. Moreover, Vezhnevets, Sazona and Andreeva [2] revealed that this approach is independent from the shape of skin distribution. A new developed RGB ratio histogram was plotted to elicit new threshold for skin color tones. This ratio was formed by mixing RGB values in order to define new colors. It is defined as follows:

$$\text{Ratio:} \quad \frac{B-G}{R+B+G} \quad (3)$$

The skin tone classification will be based on computing the distance between two histograms, histogram of a reference skin tone and histogram of a query skin tone, and comparing it to some thresholds.

In literature, two methodologies exist in histogram distance measure: probalistic and vector. Probalistic based approach measures the distance between probability density functions. Examples of distances used in this approach are the Bhattacharyya distance or B-distance [14]. However, vector based measures between fixed histograms are more used in image indexing and retrieval [15] [11] such as city block, euclidean, correlation or intersection. In this paper, correlation has been used as a distance measure between two histograms h_1 and $h2$. The range of values of this measure is always included between 0 and 1. The closer the distance to 1, more similarity is detected. This measure is given by:

$$d(h_1,h_2) = \frac{\sum_{i=1}^{N} \overline{h_1}(i)\overline{h_2}(i)}{\sqrt{\sum_{i=1}^{N} \overline{h_1}^2(i) \sum_{i=1}^{N} \overline{h_2}^2(i)}} \quad (4)$$

Where

$$\bar{h}(i) = h(i) - \frac{1}{N}\sum_{i=1}^{N} h(i) \quad (5)$$

In order to improve the classification results, we have added a second rule, which is based on the interpretation of the color distribution of an image as a probability distribution. In fact, the color distribution can be

characterized by three moments on each channel: mean, variance and skewness. Skewness measures the asymmetry of the probability distribution. If the value of the c color channel at the (x, y) image pixel is $f_c(x,y)$ and the number of pixels in the image is M × N, then skewness is given by:

$$(skewness)_c = (\frac{1}{M \times N}\sum_{x=1}^{M}\sum_{y=1}^{N}(f_c(x,y) - m_c)^3)^{\frac{1}{3}} \quad (6)$$

Where

$$m_c = \frac{1}{M \times N}\sum_{x=1}^{M}\sum_{y=1}^{N} f_c(x,y) \quad (7)$$

We computed the mean and skewness of red component values from the RGB color space of each image and compare them to some boundary. This choice was based on the fact that a majority of skin colors cluster in red channel.

C. *Human skin color classification process*

Skin color classification is composed from a pre-processing step and classification step. The entire process is introduced in Fig. 2.

The pre-processing consists first in computing ratio given in equation (3) of the reference image, which represents black or white skin color tone and it is segmented so that the only parts left are skin regions. Then, histogram $hist_{ref}$ is plotted.

The classification phase starts by first processing the query image: face detection, using the Viola-Jones face detector [17], and segmentation using the method described in paragraph II.A. After that, ratio (3), mean and skewness of red component values are computed.

Then, histogram $hist_{query}$ is plotted and distance between histogram of query image and the one of reference image is computed. Let d_{hist} the distance between $hist_{ref}$ of reference image and $hist_{query}$ of query image.

Finally, obtained values are compared to some defined thresholds in order to classify the skin color into its corresponding tone where at least two rules should be satisfied.

Experiments of this proposed method have led to a new rule. Skin color tone is classified as black if at least two out of the following three conditions are satisfied:

$$d_{hist} > T_1 \quad (8)$$

$$Mean < T_2 \quad (9)$$

$$Skewness > 0 \quad (10)$$

Where T_1 and T_2 are thresholds experimentally determined.

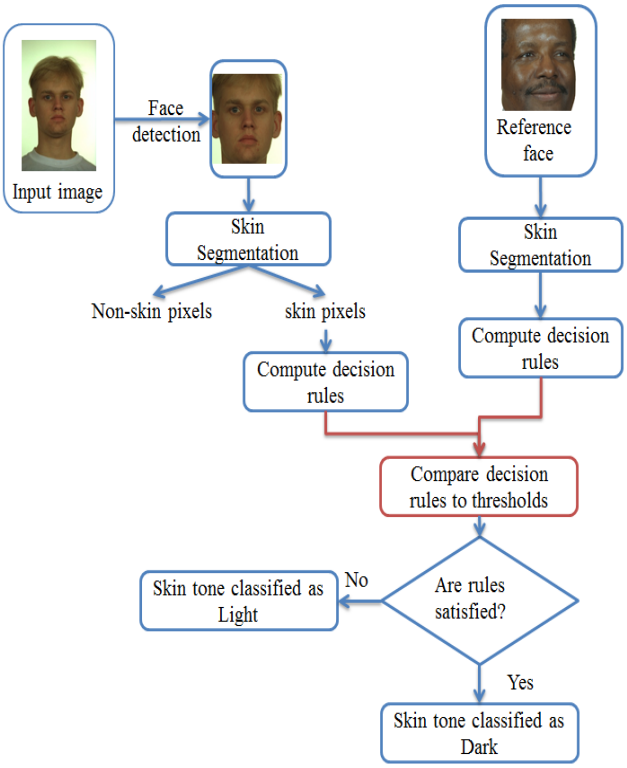


Figure 2. Human skin color classification process

III. APPLICATION OF THE HUMAN SKIN CLASSIFIER TO FACE RECOGNITION

Human skin color classification may be considered as a preprocessing operation to many applications. In our proper application, we applied it on face recognition using SIFT descriptors.

A. *SIFT descriptors definition*

SIFT is an algorithm developed by David Lowe [6]. It aims to detect and identify similarities between extracted features of different digital images. SIFT features extraction consists mainly in four steps:

- First step is to detect points of interest which correspond to the extreme points in an image. Those points are calculated from plane subsets of Difference of Gaussian (DoG) filters applied to the image at different scales.
- Then, points of interest with low extreme of DoG are discarded.
- After that, one or many orientations are given to the relevant points of interest.
- Finally, digital descriptors derived from these orientations are modeled with a set of 128-length feature vectors.

SIFT descriptor outputs large number of features with different scales and locations that cover the whole image.

 ISBN: 978-1-61208-320-9

Once extracted, feature vectors of an image may be compared to the query ones to find the most relevant. This is called matching process.

B. Matching features

Feature matching has presented a major concern in computer vision and pattern recognition for several decades. For image matching and recognition, extracted features from the input image are compared with ones extracted from training images in database in order to identify the most similar image to the input one. This process is based on an image similarity measure between two images. Many factors can affect the performance of the matching such as the matching measure criterion and the type of used features. In this paper, we employ the fast approximate near neighbors measure [5].

In [5], Muja and Lowe compare many algorithms for Fast Approximate Nearest Neighbor (FANN) search. As a result, two algorithms showed the best performance. This algorithm used either the hierarchical k-means tree or multiple randomized kd-trees.

C. Face recognition

The human skin color classification process is first applied on our database in order to divide it into two sub-databases and to compute and save SIFT descriptors for each image. In the recognition phase, the skin color tone of the query image is determined, SIFT descriptor is computed and finally, the feature matching distance between SIFT descriptor of query image and ones of each of trained images in the appropriate database is evaluated. The resulting distance matches the query image to the nearest ones in database.

IV. RESULTS AND DISCUSSION

Our experiments were carried out with a set of FERET images [16]. This database contains 1564 sets of images for a total of 14,126 images that includes 1199 individuals and 365 duplicate sets of images.

In this paper, we selected a set of 200 (near) frontal FERET faces with different skin color tones: 100 samples of faces classified as black and 100 samples of faces classified as white. We first present the evaluation results of our skin color classifier followed by the face recognition results before and after classification.

1) Human skin color classifier: The obtained results of our classifier applied on each skin tone are shown in table I.

TABLE I. HIT AND ERROR RATES

Skin color	white	black	Total
Hit percentage (%)	90	84	87
Error percentage (%)	10	16	13

For the 100 white tone sample images, 10 were classified incorrectly, while for the 100 black tone sample images, 16 were classified incorrectly. In fact, misclassification is caused mainly by bad illumination conditions. Fig. 3 shows some of misclassified skin color tones. The two faces belong to white tone but due to dark illumination they were classified as black tone.

Moreover, when considering the miscalssified black skin tones, we note that most of them belong to brown skin tones.

Figure 3. Examples of incorrectly classified black skin color tone

Figure 4. Examples of incorrectly classified skin color tones

Our developed classifier is able to classify 83% of faces successfully. This result is higher than the hit rate reached in [10] where only 70% of skin color tones were well classified. In that study, faces were classified into black, brown and white skin tones using also a pixel-based classification based on the RGB model.

The time elapsed for the classification of our database is around 100sec (0.5sec for each image). This elapsed time includes: reading image from database, face detection, face segmentation, average of red component computing, histograms computing and comparison to thresholds.

2) Face recognition using SIFT descriptors: An example of the performance of the employed face recognition technique before and after skin classification is presented. Fig. 5 presents the query image. In our database, five images belong to this person.

Figure 5. Query image

 ISBN: 978-1-61208-320-9

The face recognition results before and after classification are illustrated in Fig. 6 and Fig. 7 where the first most similar images to the query one are displayed.

Figure 6. The first similar images to the query before skin classification

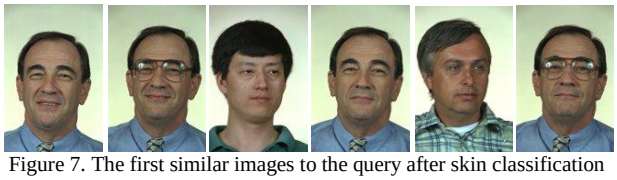


Figure 7. The first similar images to the query after skin classification

According to the qualitative evaluation of the retrieved images, retrieval results are ameliorated and become more accurate. In fact, in the presented example, 90% of faces belonging to the same person appeared in the first ten retrieved images.

V. CONCLUSION AND FUTURE WORK

Human skin color classification into skin tones is a hard operation since skin color may be easily affected by environmental effects especially illumination (light, shade, etc.). Moreover, it is considered as a delicate operation since it is employed as a preprocessing step in many systems such as face recognition. As a consequence, those systems' performance is highly related to the results obtained in the classification step. In spite of those facts, our proposed human skin color classifier based on RGB model succeeded to reach a hit rate of 87%. Also, its high speed and accuracy makes it appropriate for real time applications. Hence, classification reduces the processing time, but can degrade the recognition performance.

In the future, this work should focus on overcoming the effect of illumination in skin color classification. In fact the luminance histogram skewness is correlated with surface brightness. When the image of a surface has positively skewed statistics, it tends to appear darker than a similar surface with lower skewness. Thereby, image illumination can be enhanced basing on the skewness value.

ACKNOWLEDGMENT

This research and innovation work is carried out within a MOBIDOC thesis funded by the EU under the PASRI project.

REFERENCES

[1] G. Osman, M. S. Hitam and M. N. Ismail, "Enhanced skin colour classifier using RGB Ratio model," arXiv preprint arXiv:1212.2692, 2012.

[2] V. Vezhnevets, V. Sazonov and A. Andreeva, "A survey on pixel-based skin color detection techniques," Proc. Graphicon, 2003, pp. 85-92.

[3] K. Ravichandran and B. Ananthi, "Color skin segmentation using k-means cluster," International Journal of Computational and Applied Mathematics, vol. 4, no. 2, 2009, pp. 153-157.

[4] M. C. Shin, K. I. Chang and L. V. Tsap, "Does colorspace transformation make any difference on skin detection?," in Applications of Computer Vision, 2002. (WACV 2002). Proceedings. Sixth IEEE Workshop on, IEEE, 2002, pp. 275-279.

[5] M. Muja and D. G. Lowe, "Fast approximate nearest neighbors with automatic algorithm configuration," In VISAPP(1), 2009, pp. 331-340.

[6] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," International journal of computer vision, vol. 60, no. 2, 2004, pp. 91-110.

[7] N. Bourbakis, P. Kakumanu, S. Makrogiannis, R. Bryll and S. Panchanathan, "Neural network approach for image chromatic adaptation for skin color detection," International journal of neural systems, vol. 17, no. 01, 2007, pp. 1-12.

[8] D. Chai, S. L. Phung and A. Bouzerdoum, "A Bayesian skin/non-skin color classifier using non-parametric density estimation," in Circuits and Systems, 2003. ISCAS'03. Proceedings of the 2003 International Symposium on, pp. II-464, 2003.

[9] A. E. Abdel-Hakim and A. A. Farag, "Csift: A sift descriptor with color invariant characteristics.," In Computer Vision and Pattern Recognition,2006 IEEE Computer Society Conference on, vol. 2, 2006, pp. 1978-1983.

[10] I. Boaventura, V. Volpe, I. da Silva and . A. Gonzaga, "Fuzzy Classification of Human Skin Color in Color Images," IEEE Int. Conf. on Systems, Man and Cybernetics, vol. 6, 2007, pp. 5071-5075.

[11] M. J. Swain and D. H. Ballard, "Color indexing," International journal of computer vision, vol. 7, no. 1, 1991, pp. 11-32.

[12] J. Kovac, P. Peer and F. Solina, "Human skin color clustering for face detection," vol. 2, 2003.

[13] D. G. Lowe, "Object recognition from local scale-invariant features," Computer vision, 1999. The proceedings of the seventh IEEE international conference on, vol. 2, 1999, pp. 1150-1157.

[14] T. Kailath, "The divergence and Bhattacharyya distance measures in signal selection," Communication Technology, IEEE Transactions on, vol. 15, no. 1, 1967, pp. 52-60.

[15] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic and others, "Query by image and video content: The QBIC system," Computer, vol. 28, no. 9, 1995, pp. 23-32.

 ISBN: 978-1-61208-320-9

[16] P. J. Phillips, H. Wechsler, J. Huang and P. J. Rauss, "The FERET database and evaluation procedure for face-recognition algorithms," Image and vision computing, vol. 16, no. 5, 1998, pp. 295-306.

[17] P. Viola et M. Jones , "Rapid object detection using a boosted cascade of simple features," In Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on, IEEE, 2001, pp. I-511.

 ISBN: 978-1-61208-320-9