MMEDIA 2015

The Seventh International Conferences on Advances in Multimedia

ISBN: 978-1-61208-400-8

April 19 - 24, 2015

Barcelona, Spain

MMEDIA 2015 Editors

Philip Davies, Bournemouth and Poole College, UK

MMEDIA 2015

Foreword

The Seventh International Conferences on Advances in Multimedia (MMEDIA 2015), held between 19th-24th, 2015 in Barcelona, Spain,, 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 2015 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 2015 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 2015. 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 2015 organizing committee for their help in handling the logistics and for their work to make this professional meeting a success.

We hope that MMEDIA 2015 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 also hope Barcelona provided a pleasant environment during the conference and everyone saved some time for exploring this beautiful city.

MMEDIA 2015 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 2015

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 2015 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
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
Bruno Checcucci, Perugia University, Italy

Hongyu Li, Tongji University - Shanghai, China
Anthony Y. H. Liao, Asia University, Taiwan
Pascal Lorenz, University of Haute Alsace, France
Alexander Loui, Kodak Alaris Inc. - Rochester, USA
Massudi Mahmuddin, Universiti Utara Malaysia, Malaysia
Erik Mannens, Ghent University, Belgium
Sathiamoorthy Manoharan, University of Auckland, New Zealand
Xiaoyang Mao, University of Yamanashi, Japan
Michael Massoth, University of Applied Sciences - Darmstadt, Germany
Mike Matton, VRT research & innovation – Brussel, Belgium
Annett Mitschick, Technical University - Dresden, Germany
Ayman Moghnieh, Universitat Pompeu Fabra - Barcelona, Spain
Manoranjan Mohanty, Swedish Institute of Computer Science (SICS) Lund, Sweden
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
Vincent Oria, New Jersey Institute of Technology, USA
Jordi Ortiz Murillo, University of Murcia, Spain
Marco Paleari, Fondazione Istituto Italiano di Tecnologia | Center for Space Human Robotics, Italy
Sethuraman Panchanathan, Arizona State University, USA
Neungsoo Park, Konkuk University, South Korea
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
Amna Qureshi, Univeristat Oberta de Catalunya, Spain
Piotr Romaniak, AGH University of Science and Technology - Krakow, Poland
Patrice Rondao Alface, Alcatel-Lucent Bell Labs - Antwerp, Belgium
Angel D. Sappa, Computer Vision Center, Spain
Susana Sargento, University of Aveiro/Institute of Telecommunications, Portugal
Oliver Schreer, Fraunhofer Heinrich-Hertz-Institute, Germany
Christine Senac, IRIT laboratory, France
Kimiaki Shirahama, University of Siegen, Germany
Xubo Song, Oregon Health & Science University, USA
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
Tamas Sziranyi, MTA SZTAKI (Institute for Computer Science and Control) | Budapest University of Technology and Economics, Hungary
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

Table of Contents

Blind Tamper Detection to Copy Move Image Forgery using SURF and MSER

Kelsey Ramirez-Gutierrez, Mariko Nakano-Miyatake, Gabriel Sanchez-Perez, Hector Perez-Meana
Instituto Politecnico Nacional
Mechanical and Electrical Engineering School
Mexico, City, Mexico
hmperezm@ipn.mx

Abstract— **The sharing of digital images has become a common practice in our daily life, with the risk that these images can be accessed and easily modified by malicious people with the intention of causing moral or economic damage; or even to incriminate innocent people in legal issues. This paper proposes an algorithm to authenticate digital images by means of blind tampering detection against one of the principal manipulations that an image is put through, i.e. the *Copy-Move* which intends to erase or replicate a part of the image. The development and evaluation results of this proposal are presented in this paper.**

Keywords-Tamper detection; SURF; MSER; copy-move

I. INTRODUCTION

Nowadays, a huge amount of digital images, with or without commercial value, are easily shared among the general public via Internet or stored using any of the several available digital formats. Such images, which include private pictures or confidential images, have in general high quality and can be easily manipulated using computational tools such as: Photoshop®, Corel Paint Shop®, etc. Such kind of malicious attacks can be divided in copy-move and cut-and paste attacks. The copy-move is one of the most studied forgery techniques which consist in copying a portion of an arbitrary size and shape of a given image and pasting it in another location of the same image. Clearly, this technique is useful when the forger wants either to hide or duplicate something that is already present in the original image [1][2]. On the other hand, in the cut-and-paste attack or splicing, the attacker firstly chooses a region of a given image and pastes it into a second one, usually to alter its content and meaning. Splicing is probably more common than the copy-move attack, because it is far more flexible and allows the creation of images with a very different content with respect to the original image [2].

The image authentication has been a topic of active research during the last several years, because the tampered images may cause moral or economic damages to the persons related to the maliciously modified images, giving as a result the publication of several image authentication techniques, which can be broadly classified into two types: active and passive image authentication methods. The main difference among them is that in the active methods some useful information is extracted from the image to be authenticated and embedded in it or stored separately. This information is then used during the authentication process. On the other hand, in the passive methods, also called forensic methods, the authentication must be carried out without previous information about the processing that the image to be authenticated had passed through [1][2].

The active methods can be classified into two categories: the watermarking-based and the image hashing-based schemes, both of them with advantages as well as some drawbacks. The watermarking-based techniques embed an imperceptible signal into the image to be authenticated to create a watermarked image. The embedded signal can be a random signal or a signal derived from the image to be authenticated. During the authentication process, the watermark is extracted from the watermarked image to be used for authentication purpose [3] or even to restore the tampered image. Several high performance methods for embedding information into digital images have appeared in the literature [3][4][5]. These methods perform fairly well and in several cases have the ability to restore the tampered regions [3]. However, if the parameters are not properly chosen some distortion may be introduced in the image to be protected [3]. On the other hand, the image hashing-based techniques, or multimedia fingerprinting, take out a set of robust features from the image to be authenticated to create a compact code, called perceptual hashing code, which is stored or transmitted separately. During the authentication process, employing the same method, an authentication code is extracted from the suspicious image, which is then compared with the stored code and if the difference between both codes is smaller than a given threshold the suspicious image is considered as authentic; otherwise it is determined as a tampered image. It is necessary to point out that the perceptual hashing technique is different from the cryptographic hashing since in the last one, any change in the image to be authenticated, even if it is perceptually similar to the original one, produces a quite different hash value [6]; while the perceptual hashing technique has the capacity of discriminating between malicious attacks and distortions resulting for standard image processing tools. Because these methods have proved to be very efficient, several image hashing algorithms [6][7][8] have been suggested. These methods do not distort the image, although the authentication code must be stored or transmitted separately.

ISBN: 978-1-61208-400-8

In many practical situations the investigators have only the image under analysis, such that passive image authentication schemes are required, which carry out the authentication process analyzing the processing artifacts to infer the potential alterations introduced during the image generation process [1]. This paper analyzes image authentication schemes to deal with images tampered using the copy-move scheme. This tampering method creates a forged image by copying a certain portion of an image and moving it to another part of the same image. [1]. The main characteristic of this kind of tampered images is that, because the duplicated region is picked from the image itself, the noise components, texture and color patterns are compatible with the rest of the image. This fact makes it not easy to detect this kind of forgeries.

The authentication of this kind of tampered images has many important practical applications giving as a result the proposal of several authentication algorithms during the last several years. Among them, there is the feature matched technique proposed by Pan and Lyu [10], which employs local statistics features together with a verification step which tries to find duplicated regions using normalized correlations maps and thresholding. The main weakness of this method is its lack of accuracy [10]. Jaberi et al. [11] proposed a copy-move detection method in which firstly the Scale Invariant Feature Transform (SIFT) [12][13] is used to extract the key points, then the affine transform of a region around each key point is estimated and finally, to reduce the false detection, the Dense Mirror Invariant Feature Transform (MIFT) is estimated [11]. This scheme performs fairly well although it does not work well if the duplicated region corresponds to a flat surface where not key points are located. Kumar et al. [14] propose an image authentication scheme in which the image under analysis is divided in overlapping sub blocks which are then transformed to the frequency domain using the Discrete Cosine Transform (DCT), keeping only the lowest frequency components. These components are then ordered in a lexicographic way to carry out the evaluation of each sub block. This scheme performs well when the duplicated regions do not presents scaling or rotational distortions. A similar approach was also proposed by Fridrich [15] which presents the same advantages and disadvantages. Popescu [16] proposes to replace the DCT by a Principal Component Analysis (PCA) to reduce the dimensionality of features vector. However this method lacks of robustness against even small rotations of the copy moved regions. Several other methods have been proposed, some of them based on intensity method, which assume that the image may be under any JPEG, rotation or scaling operations [17]. To solve some of the problems still present in the image authentication algorithms describe above, this paper proposes an algorithm that allows the authentication of digital images that have gone under copy-move tampering attacks. Evaluation results show that the proposed scheme performs fairly well when it is required to authenticate tampered images, even when the duplicated region has been rotated and scaled.

The rest of the paper is organized as follows. Section II provides a detailed description of proposed algorithm. In Section III the evaluation results using the CASIA database [21] are given. Finally Section IV provides the conclusions of the paper.

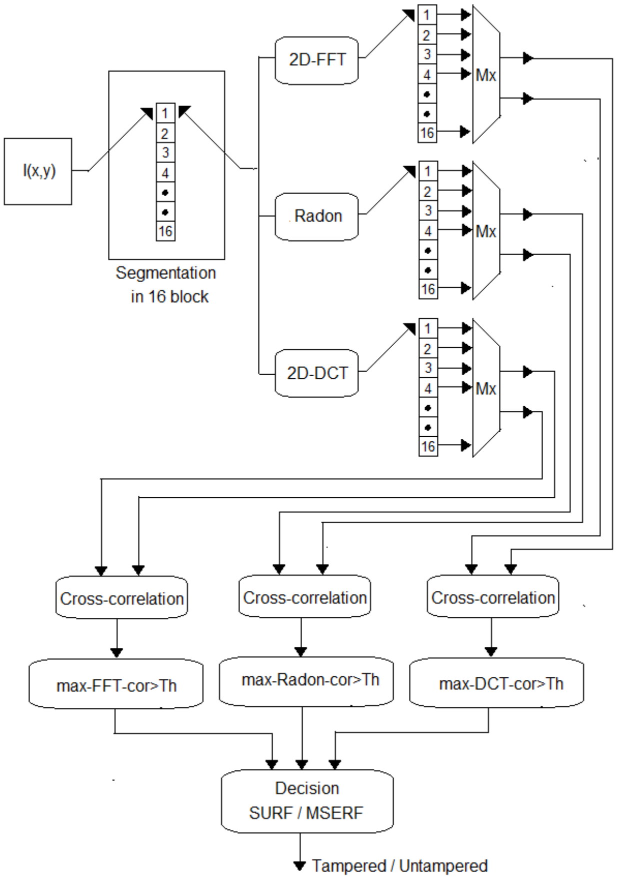
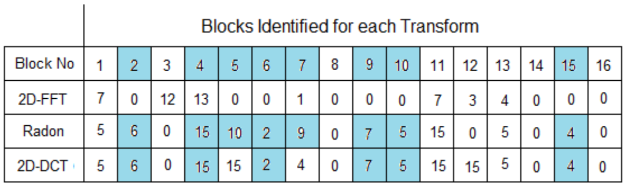


Figure 1 Proposed image authentication system

II. PROPOSED IMAGE AUTHENTICATION SYSTEM

The proposed image authentication system is shown in Fig. 1. Here the image under analysis is converted to a gray scale image and divided in 16 blocks. Next, the magnitude of the bi-dimensional Fast Fourier Transform (2D-FFT) [18], the Discrete Radon Transform (DRT) [7][9] and 2D-DCT [18] of each block are estimated. The main idea behind the proposed schema is to take advantage of the translation invariance property of the 2D-FFT, the rotation and scaling properties of the DRT and the compression capability of the DCT. Next, the cross correlation between the 16 blocks in each domain is calculated and the block index with the higher correlation values greater than a given threshold, are kept to form a matrix of 3×16 elements. Here, the threshold is given by the highest correlation value between the 16 blocks. At the end of this process a matrix of 3×16 elements is obtained containing the possibly tampered blocks. The second part of the authentication process can be more easily explained with the example shown in Fig. 2. Here, we look for a block which can be found as tampered by at least two of the three frequency transformations applied and that also correspond to the block that is being compared to.



	Blocks Identified for each Transform															
Block No	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
2D-FFT	7	0	12	13	0	0	1	0	0	0	7	3	4	0	0	0
Radon	5	6	0	15	10	2	9	0	7	5	15	0	5	0	4	0
2D-DCT	5	6	0	15	15	2	4	0	7	5	15	15	5	0	4	0

Figure 2. Block identified as tampered by each transform.

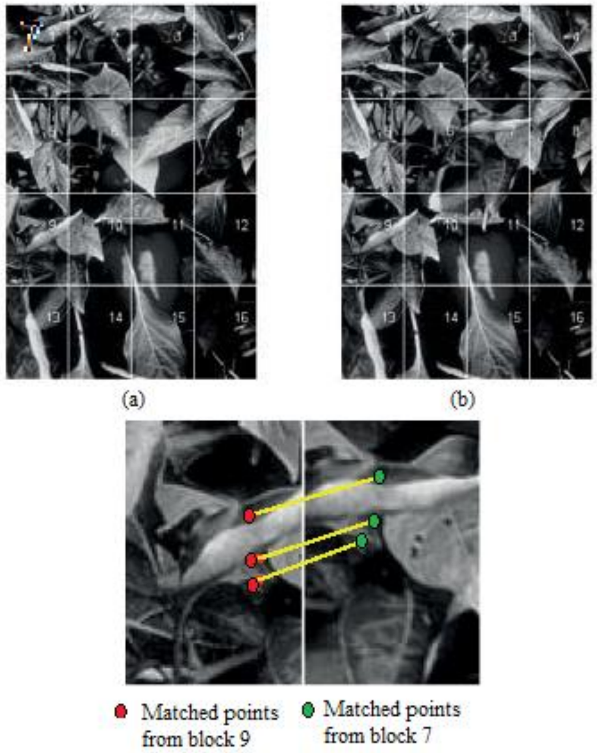


Figure 3. (a) Original image, (b) tampered image, (c) matched points in blocks 7 and 9.

For example, block 2 is highly related with block 6, according to Radon and 2D-DCT transforms. In a similar form block 6 appears to be related with block 2. Thus, blocks 2 and 6 are the first blocks to be compare using a SURF detector [19]. After this evaluation, if at least one matched point is found the block is considered as tampered; otherwise the block is labeled as untampered. The system continues analyzing the next block that is found to be related with other block according to at least two transformations. For example, block 4 is related with block 15. Again, blocks 4 and 15 are compared among them using the SURF (Speeded Up Robust features) [19] and Maximally Stable Extremal Regions, (MSER) [20]; and labeled as tamper or untampered depending if there are matched points or not inside them. This process continues until all blocks of the image related among them according to at least two transforms are labeled as tampered or untampered. Next, if after all blocks are analyzed the decision is that all of them are untampered, the blocks related with other block according to only one transform are analyzed. For example block 4 is related with the block 13 according with the 2D-FFT and block 9 and with block 7. After applying the SURF [19] and MSER [20], it was found that in block 9 and block 7 at least one matched point is found, as shown on Fig. 3 and then the system decides that the image was tampered. This process can be repeated in each one of the 16 regions for a more accurate evaluation to detect region duplications inside each sub-block. This fact reduces the computational complexity avoiding the use of overlapping blocks. Next subsections describe each stage of proposed system.

A. 2D- Discrete Fourier Transform

The bi-dimensional Discrete Fourier Transform (2D-DFT) has found a large amount of applications in several fields, because it and its inverse allow analyzing the frequency spectral characteristics of images [18]. The 2D-DFT pair is given by

$$F(u,v)=\frac{1}{NM}\sum_{x=0}^{M-1}\sum_{y=0}^{N-1}f(x,y)e^{-j2\pi\left(ux/M+vy/N\right)} \qquad (1)$$

$$f(x,y)=\frac{1}{NM}\sum_{x=0}^{M-1}\sum_{y=0}^{N-1}F(u,u)e^{j2\pi\left(ux/M+vy/N\right)} \qquad (2)$$

Some general statements can be made about the relationship between the frequency components of the Fourier transform and spatial features of an image. For instance, because the frequency is directly related to the spatial changes rate, it is not difficult intuitively to associate frequencies in the 2D-DFT with intensity variations patterns in an image, because the low frequencies correspond to the slowly varying intensity components of an image and the higher frequency components correspond to the faster intensity changes in the image [18]. Other important feature is the fact that the magnitude of the 2D-DFT is translation invariant, i.e.

$$\begin{aligned}&IF \quad f(t)\leftrightarrow F(\omega)\\&and\ \ f_1(t)=f(t-T)\leftrightarrow F_1(\omega)=F(\omega)e^{-j\omega T}\\&Then\ |F(\omega)|=|F_1(\omega)|\end{aligned} \qquad (3)$$

B. Radon Transform

The Radon Transform [7] is used in this proposal because it is robust against rotation, scaling and translation. The Radon transform for a set of parameters (ρ, θ) is the line integral through the image $f(x,y)$, where the line corresponding to the value of (ρ,θ) is given by (4)

$$g(\rho,\theta) = \sum_{m-s_{\max}}^{s_{\max}} \sum_{n=-s_{\max}}^{s_{\max}} f(m,n)\delta(\rho - m\cos\theta - n\sin\theta) \qquad (4)$$

where $\delta(\eta)$ is the Dirac delta function which is equal to one when η=0 and zero for all other arguments [7][9]. The definition of the Radon transform forces the summation of $f(m,n)$ along the line ρ=$m cos\theta$+$n sin\theta$ and consequently the value $g(\rho,\theta)$ for any (ρ,θ) is the sum of the value of $f(m,n)$ along this line [7]. The Radon transform has the following useful properties for the affine transformations of an image [7][9].

1. The translation of an image by (xo,yo) causes the Radon transform to be translated in the direction of s

$$f(m-m_0, n-n_0) \leftrightarrow g(s - m_0\cos\theta - n_0\sin\theta, \theta) \qquad (5)$$

2. Scaling (retaining aspect ratio) of an image by a factor ρ (ρ>0) causes the Radon transform to be scaled through the same factor

$$f(\rho m, \rho n) \leftrightarrow \frac{1}{|\rho|} g(\rho s, \theta) \qquad (6)$$

3. Rotation of an image by an angle θ causes the Radon transform to be shifted by the same amount

$$\begin{aligned} &f(m\cos\theta_r - n\sin\theta_r, m\sin\theta_r + n\cos\theta_r) \\ &\qquad \leftrightarrow g(s, \theta - \theta_r) \end{aligned} \qquad (7)$$

C. 2D-Discrete Cosine Transform

The 2D Discrete Cosine Transform is widely used in image compression applications, because it is able to represent a given image with a reduced number of coefficients, besides that the DCT of a real a valued signal is also real valued. The general equation of a 2D-DCTof an image of N×M pixels, $f(m,n)$, is given by [18]

$$\begin{aligned} F(u,v) = (4/MN)^{1/2} \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} A(i)A(j)f(m,n) \times \\ \cos(\pi u(2i+1)/2N)\cos(\pi v(2j+1)/2M) \end{aligned} \qquad (8)$$

where

$$A(\xi) = \begin{cases} \dfrac{1}{\sqrt{2}}, & for\,\xi = 0 \\ \xi = 0 & otherwise \end{cases} \qquad (9)$$

D. SURF Detector

The *SURF* [19] employs integral images and efficient scale space construction to generate key points and descriptors very efficiently. SURF uses two stages namely key point detection and key point description. The detector is based on the Hessian matrix with the Laplacian-based detector. It relies on integral images to reduce the computation time and therefore call it the “Fast-Hessian” detector. The descriptor, on the other hand, describes a distribution of Haar-wavelet responses within the interest point neighborhood. In the first stage, integral images allow the fast computation of approximate Laplacian of Gaussian images using a box filter. The computational cost of applying the box filter is independent of the size of the filter because of the integral image representation. The determinants of the Hessian matrix are then used to detect the key points, because of its good performance in computation time and accuracy. It relies on the determinant of the Hessian for both location and the scale. Given a point p=(x,y) in an image $I(x,y)$, the Hessian matrix section $H(x,\sigma)$ in p at scale σ is defined as follows

$$H(p,\sigma) = \begin{bmatrix} L_{xx}(p,\sigma) & L_{xy}(p,\sigma) \\ L_{yx}(p,\sigma) & L_{yy}(p,\sigma) \end{bmatrix}, \qquad (10)$$

where $L_{xx}(p,\sigma)$, $L_{xy}(p,\sigma)$, $L_{yx}(p,\sigma)$, $L_{yy}(p,\sigma)$ are the convolution of the Gaussian second order derivative with respect to x and y, respectively, with the image $I(x,y)$ in the point p [19]. The SURF builds its scale space by keeping the same image size while varying only the filter size. In the final stage, to each detected key point is firstly assigned a reproducible orientation. For orientation, the Haar wavelet responses in x and y directions are calculated for a set of pixels within a radius of 6σ where σ refers to the detected key point scale. The SURF descriptor is then computed by constructing a square window centered on the key point and oriented along the orientation obtained before. This window is divided into 4 x 4 regular sub-regions and Haar wavelets of size 2σ are calculated within each sub-region. Each sub-region contributes 4 values thus resulting in 64D descriptor vectors which are then normalized to unit length. The resulting SURF descriptor is invariant to rotation, scale and contrast; besides that it is also partially invariant to some other transformations [19].

E. MSER Detector

The Maximally Stable Extremal Regions (MSER), proposed by Matas et al. [20], estimates a set of distinguished regions that are detected in a gray scale image and defined by an extremal property of the intensity function in the region and on its outer boundary. The MSER has properties that allow it to achieve a superior performance as stable local detector compared with other local point detectors. Two of the main properties of the set of MSER are that it is closed under continuous geometric transformations and invariant to affine intensity changes. Furthermore the MSER regions are detected at different scales. Some other important properties of the *MSER* detector are:

a) Invariance to affine transformation of image intensities.
b) Covariance to adjacency preserving (continuous) transformation *T*: *D* on the image matched point domain.
c) Stability of the detected regions which means that only the regions whose support is nearly the same over a range of thresholds is selected.
d) Multi-scale detection without any smoothing involved, thus both fine and large structure is detected.

The set of all extremal regions that can be enumerated in worst-case is of $O(n)$, where n is the number of pixels in the image.

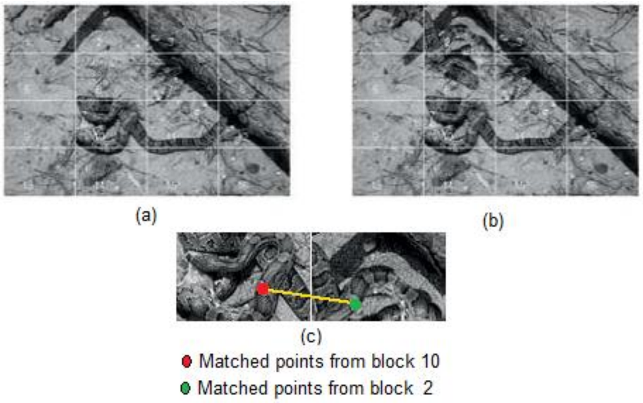


Figure 4. Performance of proposed scheme (a) Original image, (b) tampered image and c) matched points.

III. Evaluation Results

To evaluate the proposed images authentication system, the CASIA Image Tampering Detection Evaluation Database [21] was used, which consists of 102 Images. To obtain a correct operation, the parameters of proposed system are set as follows: The threshold value used to determine if two blocks in each transformation domain are similar using the cross correlation among them is the mean value of the highest correlation among the 16 blocks. For the Radon Transform the projection of the image intensity along the 180 angles equally spaced in the interval $0 \leq \theta < \pi$ are analyzed. For the SURF detector the number of octaves is set equal to 3, which will give a filter size of 27x27. Finally for the MSER detector the step size between intensity threshold levels is set equal to 0.8.

Table I shows the detection performance of proposed system when it is required to evaluate both tampered and original images, where a false positive is an error in which the test result indicates that an image is tampered when it is an original one, while a false negative is an error produced when the test result indicates that the image is original, although it is in fact tampered.

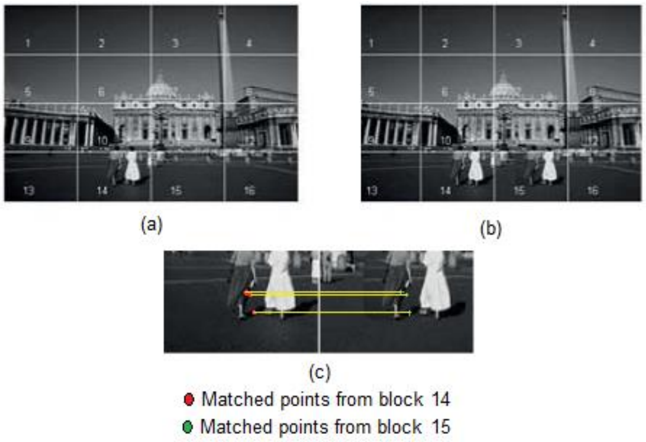

1
2
3
4
5
6
10
13
14
15
16
(a)
(b)
(c)
● Matched points from block 14
● Matched points from block 15

Figure 5. Performance of proposed scheme (a) Original image, (b) tampered image and c) matched points.

Figures 4 and 5 show the evaluation results in which the proposed scheme correctly detects a tampered image. In both cases the original and tampered images are shown in (a) and (d), while in (c) the matched points obtained by the SURF and MSER are shown which confirm that the image under analysis was tampered is shown in (c). In some cases, depending on the alteration introduced on the original image, for example if the pasted object suffer some affine transformation, the SURF features are not robust enough to detect these changes, so in this case to use the MSER features may allow to detect that the image under analysis was tampered, as shown in Figs. 6 and 7. Finally Table II shows the evaluation results obtained using the Mean Opinion Scoring (MOS) criterion in which several images were presented to 100 peoples who were asked to determine if the images were tampered or untampered.

The total time of calculation evaluating the three transformation techniques and the SURF and MSER detector is of 3.53 minutes.

TABLE I. TAMPER DETECTION PERFORMANCE OF PROPOSED ALGORITHM

Success rate	False positive	False negative
75%	10%	15%

TABLE II. TAMPER DETECTION PERFORMANCE OF PROPOSED ALGORITHM USING THE MOS CRITERION

Success rate	False positive	False negative
53%	20%	27%

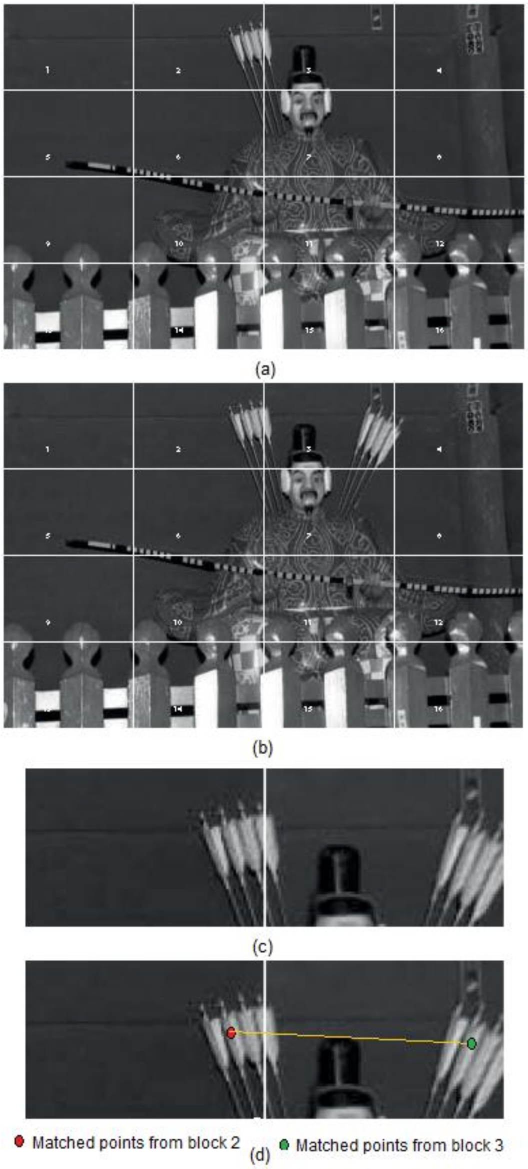


Figure 6. Performance of proposed scheme (a) Original image, (b) tampered image and c) matched points detected using SURF, (d) matched points detected using MSER.

IV. Conclusions

This paper proposes a copy-move tamper detection algorithm in which firstly the image under analysis is divided in 16 blocks and the 2D-DCT, 2D-FFT and DRT of each block is estimated. Then, the similitude between such blocks, in each domain, is estimated using the maximum of the cross correlation value together with the SURF detector and MSER features to determine if the image was tampered. From the evaluation results presented in this paper we can observe that the proposed scheme is able to identify the copy-move regions of the image under analysis. We must add that this method is not trying to identify any particular type of copy-move forgery mechanism, like rotation, or scaling, or JPEG compression. Instead it is intended to be a more general method able to operate in almost any situation and that, combined with other methods can lead to an accurate detection of a specific type forgery attack.

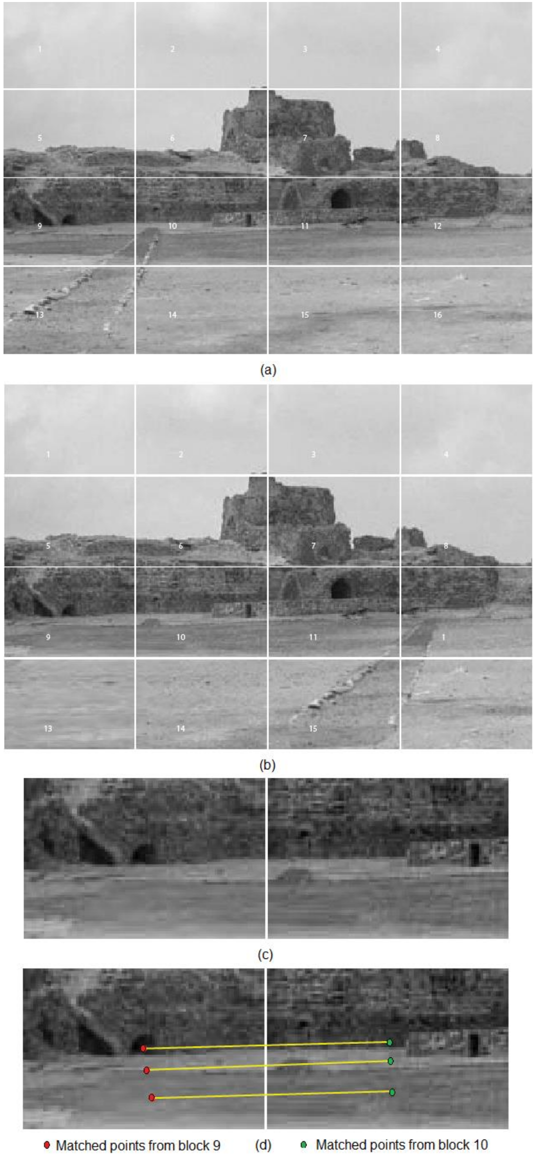


Figure 7. Performance of proposed scheme (a) Original image, (b) tampered image and c) matched points detected using SURF, (d) matched points detected using MSER.

Acknowledgment

We thank the National Science and Technology Council (CONACYT) and to the National Polytechnic Institute of Mexico (IPN) for the financial support provided during the realization of this research.

References

[1] M. Kirchner, “Notes on digital image forensics and counter-forensics,”http://dud.inf.tu-dresden.de/~kirchner/Documents /image_forensics_and_counter_forensics.pdf, Oct. 2011.

[2] A. Piva, *An Overview on Image Forensics*, ASRN Signal Processing, Hindawi, http://dx.doi.org/10.1155/2013/496701, 2012.

[3] L. Rosales-Roldan, M. Cedillo-Hernandez, M. Nakano-Miyatake, H. Perez-Meana, B. Kurkoski, "Watermarking-basewd image authentication with recovery capability using halftoning technique", Signal Processing: Image Communications, vol. 28, pp. 69-83, Jan 2013.

[4] M. Cedillo-Hernández · F. García-Ugalde · M. Nakano-Miyatake, H. Manuel Perez-Meana, "Robust hybrid color image watermarking method based on DFT domain and 2D histogram modification", Journal of Signal Image and Video Processing, April, 2013, doi 10.1007/s11760-013-0459-9.

[5] I. Ismali, S. El-Zoghdy, and A. Abdo, "A novel techinque for data hiding," International Journal of Computers and Applications, vol. 32, pp. 119–124, Jan. 2010.

[6] A. Swaminathan, Y. Mao, and M. Wu, "Robust and secure image hashing," *IEEE Transactions on Information Forensics and Security*,, vol. 1, pp. 215–230, June 2006.

[7] J. S. Seo, J. Haitsmab, T. Kalkerb, and Y. C. D., "A robust image fingerprinting system using the radon transform," *Signal Processing: Image Communication*, vol. 19, pp. 325–339, April, 2004.

[8] Y. Li, Z. Lu, C. Zhu, and X. Niu, "Robust image hashing based on random gabor filtering and dithered lattice vector quantization," IEEE Transactions On Image Processing, vol. 99, pp. 1–14, Jan. 2011.

[9] D. Q. Nguyen, L. Weng, and B. Preneel, "Radon transform-based secure image hashing," International Conference on Communications and Multimedia Security, Springer-Verlag, pp. 186–193, Oct. 2011.

[10] X. Pan and S. Lyu, "Region duplication detection using image feature matching", IEEE Trans. On Forensic Security, pp. 857-867, Dec. 2010.

[11] M. Jaberi, G. Bebis, M. Hussain, G. Muhammad, "Improving the detection and localization of duplicated regions in copy-move image forgery", Proc. Int. Conf. on Digital Signal Processing, pp. 1-6, 2013.

[12] H. Bay, T. Tuytelaars, , and L. V. Gool, "Surf: Speeded up robust features," ETH Zurich and Katholieke Universiteit Leuven, Tech. Rep., 2006.

[13] N. Y. Khan, B. McCane, and G. Wyvill, "SIFT and SURF performance evaluation against various image deformations on benchmark dataset," International Conference on Digital Image Computing: Techniques and Applications, pp. 501–506, Dec. 2011.

[14] S. Kumar, J. Desai, S. Mukherjee, "A fast DCT based method for copy move forgery detection", Proc. Int. Conf. on Image Information Processing, (CIIP'13), pp. 1-6, 2013

[15] J. Fridrich, D. Soukalm, J. Luka, "Detection of copy move forgery in digital images," Proc. Digital Forensic Research Workshop, pp. 19-29, 2003.

[16] A. Popescu, H. Farid, "Exposing Digital Forgeries by detecting duplicated regions," Tech. Report TR2004-515, Dartmouth Collage, Hanover, 2004.

[17] M. A. Qureshi, M. Deriche, "A review on copy move image forgery detection techniques," Proc Multi-Conference Signal, Systems and Devices, pp. 1-5, 2014.

[18] R. Gonzalez and R. Woods, *Digital Image Processing*, Prentice Hall, Englewood Cliffs, NJ, 2008.

[19] H. Bay, T. Tuytelaars, L. Van Gool, "SURF: Seed Up Robust Features", Proc. Int. Conf. on Computer Vision (ECCV´06), May 2006

[20] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide baseline stereo from maximally stable extremal regions," Image Vision and Computing, vol. 22, pp. 761-767, 2004.

[21] CASIA Image Tampering Detection, Evaluation Database http://forensics.idealtest.org:8080/index_v1.htm. June 2014.

Electronic Payment and Encouraged Cooperation in a Secure and Privacy-Preserving P2P Content Distribution System

Amna Qureshi, Helena Rifà-Pous and David Megías
Estudis d'Informàtica, Multimèdia i Telecomunicació, Internet Interdisciplinary Institute (IN3)
Universitat Oberta de Catalunya (UOC), Barcelona, Spain
E-mail:{aqureshi,hrifa,dmegias}@uoc.edu

***Abstract*—In this paper, we propose a peer-to-peer (P2P) content distribution system that allows the efficient distribution of large-sized multimedia contents while preserving the security and privacy of content providers (merchants) and end users (buyers), respectively. However, the privacy of a buyer can be revoked as soon as he/she is found guilty of copyright violation. A payment protocol is also proposed that provides a secure payment mechanism, where personal information and order information cannot be exposed to an unauthorized third party. In addition, a reputation-based system is introduced for the selection of the proxy peers required for secure delivery of the fingerprinted content from the merchant to the buyer. The paper presents a thorough security analysis of the system against several security compromising attacks.**

***Keywords*—privacy; security; collusion-resistant fingerprinting; permutation; peer-to-peer; e-payment; reputation.**

1 Introduction

The low-cost, scalability and ease of content dissemination provide a lucrative opportunity for content providers to generate revenues through P2P systems. However, the content providers have been reluctant in adopting P2P systems as a distribution vehicle to monetize digital content, since these systems are plagued with piracy. The ability to make perfect copies and the ease with which these copies can then be distributed has given rise to significant problems regarding the misuse, illegal copying and redistribution. The content providers apparently fear losing control of content ownership in the sense that they are no longer in control of the content distribution and worry about the promotion of illegal activity. Also, tracing a copyright violator in a P2P system with millions of connected users is an immense task. Therefore, ensuring the appropriate use of copyrighted multimedia content in P2P systems has become increasingly critical. This copyright infringement problem motivates the development of content protection techniques. Among various content protection techniques, digital fingerprinting addresses the problems of copyright protection and traitor tracing.

Digital fingerprinting gives merchants more options to control the distribution of their content. Fingerprinting techniques involve the generation of a fingerprint (a buyer-specific identification mark), the embedding operation and the realization of traceability from re-distributed copies. In traditional fingerprinting schemes, it is assumed that the merchants are trustworthy and always perform embedding honestly [1]. Thus, a dishonest merchant could frame an innocent buyer, while a cheating buyer would be able to deny his/her responsibility for a copyright violation act. Asymmetric fingerprinting schemes [2] were introduced to overcome this problem. In these schemes, only the buyer obtains the exact fingerprinted content, and hence the buyer cannot claim that a pirated copy was originated from the merchant. However, most of the asymmetric fingerprinting schemes in the literature incur high computational and communicational burdens at the merchant's and/or at the buyer's end, due to the use of cryptographic protocols such as homomorphic encryption or committed oblivious transfer.

Though the content protection techniques enable the merchants to enforce copyrights in the content, these techniques are often criticized for breaking buyers' privacy by collecting information about the buyers, such as the transaction history or the purchasing behavior. A priori, copyright protection places the buyer into an adversarial relation with the merchant. Hence, the incorporation of a content protection mechanism in a P2P system can have serious effects on the privacy interests of the buyers. Recent years have drawn increasing attention from the research community towards the preservation of the merchants' ownership property and buyers' privacy in P2P content distribution systems. To date, very few P2P distribution systems have been proposed that provide both copyright protection and privacy preservation.

Megías and Domingo-Ferrer [3] introduced a novel concept of a recombination fingerprinting mechanism for P2P content distribution. The proposed scheme provides copyright protection, collusion resistance and traitor tracing. However, this system is implemented with a two-layer anti-collusion code (segment level and fingerprint level), that results in a longer codeword. Furthermore, honest and committed proxies are required for the generation of valid fingerprints at the buyer's end. Megías [4] proposed an improved version of [3], in which a four-party anonymous communication protocol is proposed to prevent malicious proxies to access clear-text fingerprinted contents. However, the system still requires a two-layer anti-collusion code. Domingo-Ferrer and Megías [5] proposed a P2P protocol for distributed multicast of fingerprinted content in which each receiver obtains a different fingerprinted copy of the content, which allows the provider to trace re-distributors without affecting the privacy of honest buyers. However, an implementation of a secure multi-party protocol results in increased computational and communication costs at the buyer end. Qureshi, Megías and Rifà-Pous [6] proposed a P2P content distribution framework for preserving privacy and security of the user and the merchant based on homomorphic encryption. In the framework, some discrete wavelet transform (DWT) low-frequency (approximation) coefficients are selected according to a secret key for embedding an encrypted fingerprint to prevent data expansion due to homomorphic encryption. Although the selective public-key encryption of the multimedia content results in lesser data expansion, it imposes computational burden on a merchant and an increased complexity in file reconstruction at the buyer's end.

In this paper, we present a P2P content distribution system that provides copyright protection and conditional privacy to the merchant and the buyer, respectively. In the proposed system, the original multimedia file is partitioned by the merchant into a small-sized base file and a large-sized supplementary file. This enables to reduce the communication bandwidth and the computation power required by the merchant in delivering the large-sized multimedia file. The base file contains the most important information and is transmitted in a semi-centralized way. The supplementary file is unusable without the base file and is distributed through a P2P network. A merchant forms a base file by using a pre-computation-based secure embedding mechanism in which the DWT approximation coefficients are embedded in parallel with all 1s and all 0s bit streams. An asymmetric fingerprinting protocol based on collusion-resistant codes and a robust embedding scheme is performed between a merchant, a buyer and a set of proxies in the presence of a third party (monitor), in such a way that the merchant does not know the fingerprint or the fingerprinted content, and the proxies are unable to frame honest buyers by combining their assigned permuted fingerprint bits. A reward and punishment mechanism is also proposed to ensure that each proxy peer's best strategy is to loyally follow the prescribed fingerprinting protocol. The system also enables buyers to purchase digital contents anonymously by using dynamic pseudonyms based on a one-way hash function instead of their real IDs.

The paper is organized as follows. In Section 2, the building blocks of the system are introduced. In Section 3, the proposed P2P content distribution system is described in detail. In Section 4, we discuss the security analysis of the system's protocols through a number of attack scenarios. Section 5 presents the comparative analysis of the proposed system with related P2P content distribution systems. Finally, Section 6 summarizes the conclusions.

2 Building Blocks

In this section, a brief overview of the building blocks (embedding domain and algorithm, collusion-resistant fingerprinting codes, PseudoTrust model and permutation) of the system is presented.

A. *Embedding domain*

In the signal processing research area, the wavelet transform has gained widespread acceptance in recent years. The DWT is used in the system to embed the collusion resistant fingerprint into a multimedia content. The DWT of a signal results into approximation and detail coefficients. Since the low frequency coefficients can effectively resist various signal processing attacks, the fingerprint bits are typically embedded into the approximation coefficients of the signal after the DWT. Moreover, the original signal can be reconstructed from the approximation and detail coefficients through the inverse discrete wavelet transform (IDWT).

B. *Embedding algorithm*

An embedding algorithm is used to embed a fingerprint into different copies of the same content. Quantization index modulation (QIM) [7] is a relatively recent embedding technique that 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 DWT coefficient W by choosing between a quantizer with even or odd values, depending on the binary value of f. The proposed system employs a QIM-based watermarking technique to embed the collusion-resistant fingerprint into the content.

C. *Collusion-resistant fingerprinting codes*

Nuida *et al.*'s c_0-secure codes [8] are used in the system for the generation of the collusion-resistant code. Nuida *et al.* proposed a discrete distribution of state-of-the-art collusion-resistant Tardos codes with a δ-marking assumption (the number of undetectable bits that are either erased or flipped is bounded by δ-fraction of the total code length m) to reduce the length of the codewords and the required memory amount without degrading the traceability. The tracing algorithm of Nuida *et al.* outputs one user with the highest accusation score. The details of Nuida *et al.*'s fingerprint generation and traitor-tracing algorithms can be found in [8].

D. *PseudoTrust model*

The PseudoTrust model proposed by Lu *et al.* [9], based on a zero-knowledge proof-of-identity, is used in the system to provide revocable anonymity and unlinkability properties. The PseudoTrust model enables pseudonym-based trust management such that the real identities of the peers are protected during the authentication. In addition, the communication between two peers is anonymized using onion routing within the system. In the PseudoTrust model, the pseudo-identities are generated by the peers without any trusted third party, which leads to an accountability problem in the system. Thus, to add accountability to our system, an internal certificate authority (CA_R) is incorporated in the PseudoTrust model. Each peer is authenticated by CA_R before he/she joins the network. Hence, each peer has a private key, a public key and a public-key certificate signed by CA_R. The details of generation of pseudo-identities and anonymous authentication process are provided in [6].

E. Permutation

In the proposed system, the buyer's security and non-repudiation (merchant's security) are provided by using the concept of permutation. The permuted fingerprint generated by the monitor is permuted using different permutation keys and is then assigned to a set of proxy peers Pr_j in such a way that the merchant cannot predict about the fingerprint and the fingerprinted content, and Pr_j are unable to frame honest buyers by combining their information bits.

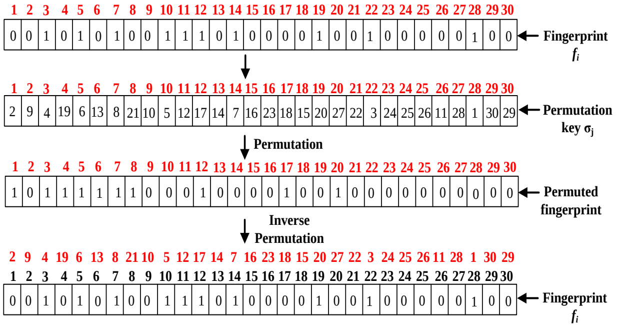


Figure. 1: Permutation of a fingerprint

Figure. 1 illustrates the permutation concept of a fingerprint in the system. Figure. 1 shows a fingerprint f_i of 30 bits, and a random permutation key σ_j of 30 elements. σ_j is applied to f_i such that the bit position 1 of the fingerprint corresponds to the bit position 2 of a permuted fingerprint ($1 \rightarrow 2$), the second bit position corresponds to the bit position 9 of the permuted fingerprint ($2 \rightarrow 9$), and so on. On applying the inverse permutation key σ_j^{-1} to a permuted fingerprint, the original fingerprint f_i is obtained.

3 Proposed system

This section describes the design and functionality of the system. In Section 3-A, we define the role of each entity. Section 3-B defines the functionality requirements and the security assumptions.

A. System entities

The system involves seven entities and the function of each entity is defined as follows:

- A merchant M is an entity that distributes the copyrighted content to the buyers in the P2P system. It is involved in the fingerprint generation, the file partitioning, the distribution of base and supplementary files, the traitor tracing and the dispute resolution protocols.
- A buyer B_i is an entity that can either play the role of data requester or provider. B_i is involved in the registration protocol, acquisition of a base file (*BF*) from the merchant, the distribution of a supplementary file (*SF*) through the system, the file reconstruction protocol and a dispute resolution, in case he/she is found guilty of copyright violation.
- A super peer *SP* acts as a coordinator for a small portion of the group of peers (buyers). However, instead of peers' addresses, their pseudonyms are stored. *SP* facilitates B_i's acquisition of *BF* from M, and *SF* from the buyers present in the system.
- A Certification Authority CA_R is a trusted party that is responsible of issuing certificates to the buyer for the acquisition of *BF* from M, and *SF* from other buyers.
- A monitor *MO* functions as a trusted party, which is responsible for the registration of buyers and merchants, the generation of collusion-resistant fingerprint codes, the distribution of *BF*, the file reconstruction, the traitor tracing and the dispute resolution protocol. *MO* also acts as a bank that assists B_i to download *BF* from M after making a payment. In addition, *MO* manages the rewards and punishments mechanism in the system.
- A proxy peer Pr is responsible for querying content of *BF* available at M's end with the pre-assigned bits of a fingerprint codeword and transferring the retrieved content to B_i.
- A judge J is assumed to be a trusted party, which resolves the disputes between M and B_i with the cooperation of *MO* and CA_R.

B. Design requirements and assumptions

In this section, the design requirements and security assumptions of the system are described.

- **Design Requirements:**

- M should be able to trace and identify an illegal re-distributor in case of finding a pirated copy with the help of *MO*, J and CA_R.
- The scheme should be collusion-resistant against a given number of colluders c_0 as specified by Nuida *et al.* codes [8].
- The possible collusion of Pr_j should be unable to frame an honest B_i. Also M should not be able to frame an honest B_i of illegal re-distribution.
- A B_i accused of re-distributing an unauthorized copy should not be able to claim that the copy was created by M or a collusion of the proxies Pr_j.
- The real identity of a buyer should remain anonymous during transactions unless he/she is proven guilty of copyright violation.
- J, with the help of *MO*, should be able to resolve the disputes without involving B_i in the process.

- The reconstruction of the original file from *BF* and *SF* should be performed at the buyer's end. *BF* cannot be shared within the buyers of the system.
- The buyers should register to *MO* with a subscription fee at a system start-up.
- The coin generated by *MO* should be revocable, thus enabling *MO* to refund the money to B_i in case of incomplete *BF* delivery to B_i.

- **Security Assumptions:**

- M and B_i do not trust each other but they both trust *MO*.
- In order to deliver *BF* from M to B_i, *MO* selects a fixed number (n) of proxy peers. These proxy peers follow each other in a sequential manner to transfer *BF* to B_i from M.
- The permutation keys σ_j (for $j = 1, \ldots, n$) are generated by B_i to perform permutation of a fingerprint codeword to be assigned to the proxy peers (Pr_j).
- Pr_j are not trusted and the content transferred through them is encrypted.
- Each entity (M, *MO*, Pr_j, B_i, CA_R, J) is supposed to have a public key K_p, a private key K_s. Public-key cryptography is restricted to the encryption of small-length binary strings, such as symmetric session and permutation keys.
- Before joining the system, B_i is authenticated by CA_R of the system. Once authenticated, B_i obtains a private key and a public key certified by CA_R. CA_R generates a random number r and shares it with an authenticated B_i for the generation of a pseudo-identity. Each buyer can have multiple pseudo-identities.
- M is assumed to be registered with *MO* at a system start-up.

4 Model

In this section, we detail the system designing and how to motivate the proxy peers in the base file distribution protocol to rationally play their corresponding roles.

A. Registration

Before joining the system, each buyer is assumed to be authenticated by CA_R and also the pseudo-identity of each buyer is assumed to be generated (Section 3-B). On joining the system, B_i sends a registration request to *MO* with his/her pseudo-identity. On receiving the request, *MO* verifies the pseudo-identity of B_i from CA_R. On verification, *MO* opens up an account of B_i and sends him/her the details of the subscription fee payment. B_i deposits the subscription fee and sends the signed payment receipt to *MO*. *MO* acknowledges the payment, creates a transaction identity TID in his/her database and generates a digital coin C_{B_i}. Then, *MO* signs C_{B_i} and sends it to B_i. M is also assumed to be registered with *MO*. Once registered with *MO*, the buyers connect with *SP* to obtain the multimedia content. In case the same buyer B_i joins the system with another pseudo-identity, he/she must send the old pseudo-identity to *MO* along with the new pseudo-identity in the registration request. Figure. 2 illustrates the registration protocol between MO and B_i.

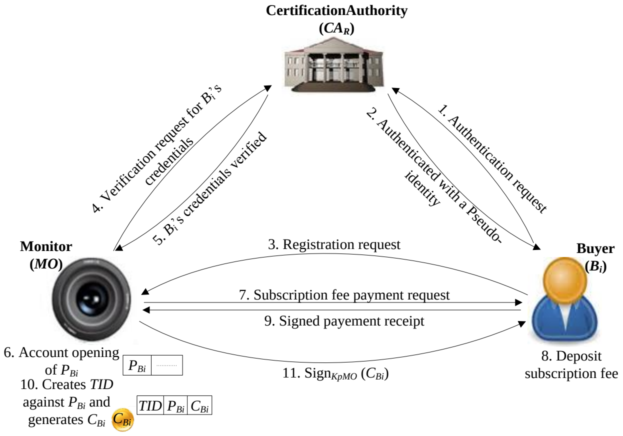

CertificationAuthority
(CA_R)
1. Authentication request
2. Authenticated with a Pseudo-identity
4. Verification request for B_i's credentials
5. B_i's credentials verified
Monitor
(*MO*)
Buyer
(B_i)
3. Registration request
7. Subscription fee payment request
9. Signed payement receipt
11. Sign$_{KpMO}$ (C_{Bi})
6. Account opening of P_{Bi}
P_{Bi}
10. Creates *TID* against P_{Bi} and generates C_{Bi}
C_{Bi}
TID
P_{Bi}
C_{Bi}
8. Deposit subscription fee

Figure. 2: Registration protocol

B. Fingerprint generation

The algorithm for fingerprint generation takes a parameter ε for error probability, the total number N of users and c_0 colluders as inputs, and outputs a collection $F = (f_1, \ldots, f_N)$ of binary codewords (f_i) of size m and a secret bias vector p. The details of the fingerprint generation algorithm can be found in [8].

C. File partitioning

The DWT decomposition on a file results in approximation (a) and detail (d) coefficients. The 3-level approximation coefficients (a_3) are used to imperceptibly embed f_i using a blind, robust and secure QIM-based watermarking scheme. M uses a_3 twice to create *BF* in such a way that it employs an embedding algorithm to insert a codeword of all ones into a_3 and simultaneously using the same embedding scheme embeds a codeword of all zeros into a_3. The two variants of a_3' form *BF* in a binary form. The detail coefficients d are used to form *SF*. Figure. 3 shows the partitioning of a multimedia file into *BF* and *SF*.

D. Base file distribution

When B_i requests *SP* for a particular content, *SP* provides B_i all the details of M having a requested content. Before the transaction, B_i generates a one-time anonymous key pair ($K^*_{pB_i}, K^*_{sB_i}$) and sends an anonymous certificate request to CA_R. On receiving an anonymous certificate $\mathrm{Cert}_{CA_R}(K^*_{pB_i}, P_{B_i})$ from CA_R, B_i negotiates with M to set-up an agreement (*AGR*) that explicitly states the rights and obligations of both parties and specifies the price and the multimedia content (X). During *AGR* set-up, B_i uses his/her pseudonym P_{B_i} and $\mathrm{Cert}_{CA_R}(K^*_{pB_i}, P_{B_i})$. M verifies the received certificate from CA_R and, on verification, generates a transaction ID (*TID*) for keeping a record of

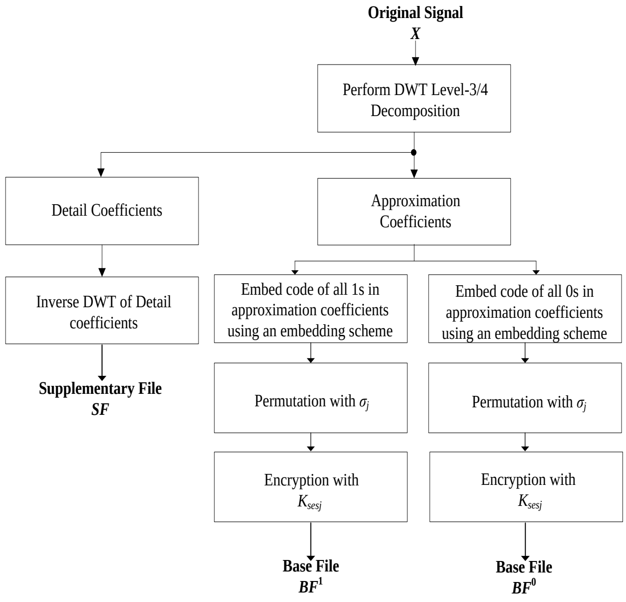

Original Signal
X
Perform DWT Level-3/4 Decomposition
Detail Coefficients
Approximation Coefficients
Inverse DWT of Detail coefficients
Embed code of all 1s in approximation coefficients using an embedding scheme
Embed code of all 0s in approximation coefficients using an embedding scheme
Supplementary File
SF
Permutation with σj
Permutation with σj
Encryption with Ksesj
Encryption with Ksesj
Base File
BF1
Base File
BF0

Figure. 3: File partitioning

the transaction between him/her and B_i. Then, M sends a request for f_i to *MO* by sending $\mathrm{Cert}_{CA_R}(K^*_{pB_i}, P_{B_i})$, $\mathrm{Cert}_{CA_R}(M)$, *AGR*, P_{B_i} and $\mathrm{Sign}_{K^*_{pB_i}}(AGR)$. *MO* validates the certificates and signatures of M and B_i from CA_R. After verification, *MO* generates a Nuida's c_0-secure codeword f_i of length m and randomly selects n proxy peers Pr_j for the delivery of a fingerprinted *BF* from M to B_i. *MO* then sends a request of permutation keys σ_j to B_i. B_i then generates n random σ_j of length $l = \lfloor m/n \rfloor$. B_i sends $E_{K_{p_{MO}}}(\sigma_j)$ to *MO*. *MO* decrypts $E_{K_{p_{MO}}}(\sigma_j)$ with $K_{s_{MO}}$ and obtains σ_j. *MO* generates n session keys K_{ses_j} and divides f_i into n segments (s_j) of length l and permutes s_j using σ_j in the same order as received by B_i. *MO* then sends $E_{K_{p_M}}(\sigma_j)|E_{K_{p_M}}(K_{ses_j})$ to M. M performs permutation on both pre-computed variants of *BF* with σ_j. It then encrypts the permuted variants of *BF* with K_{ses_j}. *MO* assigns contiguous permuted fingerprint segments to Pr_j, who then contact M in a sequential manner to obtain the fragments of the encrypted and permuted approximation coefficients fa_j. M sends a set of encrypted and permuted fragments of pre-computed coefficients to Pr_j. Pr_j selects the correct pre-computed approximation coefficients from the received coefficients using the assigned permuted fingerprint segments.

E. Supplementary file distribution

Initially, *SP* is fed with *SF* by M. On joining the system, a buyer constructs an onion path with existing peers, which points to it and adds this path to *SP* of its group. By doing so, a content requesting peer R can use this onion path to contact the content-providing peer P while knowing nothing about the provider's identity. The peer requests for a particular file to *SP* of its group. If found, it displays the list of the peers having that particular file; else it sends a request for the file to other connected *SP*s. The other *SP*s, on finding the particular content provider, send the response to the requesting *SP*. *SP* then establishes a path between R and P. After receiving a positive reply from P, R initiates a two-party authenticated key exchange (AKE) protocol to authenticate each other identities and exchange the content of *SF* anonymously. The details of *SF* distribution can be found in [6].

F. File reconstruction protocol

On delivering fa_j to B_i, Pr_j generates a one-time hash of fa_j, encrypts it with the public key of *MO* $(h(fa_j))$ and sends $E_{K_{p_{MO}}}(h(fa_j))$ to *MO*. When B_i receives fa_j from Pr_j, he/she also generates a one-time hash of fa_j, encrypts it with the public key of *MO* $(h(fa_j))$ and sends $E_{K_{p_{MO}}}(h(fa_j))$ to *MO*. *MO* stores $h(fa_j)$ in his/her database against TID that includes date, time, *AGR* and pseudo-identities of B_i and M. On receiving all the fragments of the *BF* from Pr_j, B_i sends a request for the session keys from *MO* by sending him/her a signed digital coin $\mathrm{Sign}_{K_{p_{MO}}}(C_{B_i})$. *MO* charges B_i for *BF* and sends the signed receipt and encrypted session keys $E_{K*_{p_{B_i}}}(K_{ses_j})$ to B_i. *MO* puts C_{B_i} in spent-transaction database, credits M's account and sends the payment confirmation to M. B_i decrypts $E_{K*_{p_{B_i}}}(K_{ses_j})$ with his/her $K^*_{s_{B_i}}$, then decrypts the received fragments of *BF* with K_{ses_j}, and finally applies the inverse σ_j^{-1} on the decrypted fragments of *BF*. B_i recombines all the un-permuted and decrypted fragments to form a single *BF*. B_i receives *SF* in parallel to *BF* through P2P network. Once both files are available at B_i's end, an inverse L-level DWT is performed on the approximation (embedded *BF*) and detail (*SF*) coefficients to form a fingerprinted multimedia content X'.

G. Traitor tracing

Once a pirate copy Y of content X is found, M extracts the pirated codeword pc by decomposing Y with the same wavelet basis used in the fingerprint embedding protocol. This gives the approximation coefficient matrix in which pc is embedded. The watermark detection technique is applied on the approximation coefficient matrix to extract pc. Then M sends pc to *MO*, which performs the tracing algorithm of Nuida's *et al.* codes to identify the colluder(s). The output of this tracing algorithm is the buyer with the highest score. The details of the tracing algorithm can be found in [8].

H. Dispute resolution

The goal of the dispute resolution protocol, performed between M, *MO*, CA_R and J, is to reveal the real identity of the traitor or reject the claims made by M. In order to reveal the real identity of the traitor, *MO* sends $(Y,\ pc,\ K_{p_{MO}}(f_i))$ and M sends $\mathrm{Cert}_{CA_R}(K^*_{pB_i}, P_{B_i})$, $\mathrm{Cert}_{K_{p_{B_i}}}(K^*_{p_{B_i}})$, AGR, $K^*_{P_{B_i}}$ and $\mathrm{Sign}_{K^*_{pB_i}}(AGR)$ to J. J verifies the validity of all the certificates and the signatures. If valid, it asks *MO* to decrypt $E_{K_{p_{MO}}}(f_i)$. If pc and f_i match with a high correlation, it requests CA_R to provide the real identity of the buyer. Otherwise, the buyer is proved innocent.

I. Rewards and punishments

In an attempt to induce Pr_j to correctly follow the *BF* distribution protocol, a reputation-based mechanism is introduced for a proxy peer who delivers the content correctly and honestly to the buyer or behaves maliciously and deviates from his/her course of the *BF* distribution protocol. *MO* is responsible for awarding or punishing a proxy peer. The reputation of a proxy peer is calculated using the following data: the collection of feedback about Pr_j from each buyer after reconstruction of his/her multimedia file, the collection of feedback about Pr_j from the merchant after completion of the *BF* distribution protocol, the collection of feedback about Pr_j from other peers selected by *MO* for an anonymous *BF* delivery to a buyer and the evaluation of the transaction history of each proxy peer maintained at *MO*'s end.

Based on above parameters, *MO* calculates a score of each proxy peer over a period of a time, e.g., one month, in terms of positive and negative values. A proxy peer with a positive score is rewarded with a discount coupon for his/her future content purchases, whereas, a proxy peer with a negative score is punished by *MO* in terms of money deduction from his/her account and other penalties (e.g., black listing of proxy peer's pseudo-identity). Thus, in terms of game theory, the dominant strategy solution for each proxy peer is to honestly and correctly follow the *BF* distribution protocol.

5 Security Analysis

In this section, possible security and privacy attacks on the protocols are discussed.

- **Buyer's security:** The possible collusion of Pr_j cannot frame an honest B_i and held him/her responsible for illegal re-distribution due to the fact that Pr_j would need to compute $l!$ combinations each on the colluded fingerprint. Thus, with more m-bits in f_i, Pr_j would need to carry out an increased number of permutations, which would be computationally infeasible. Also, if all Pr_j combine their fa_j, they cannot decrypt these fragments since the fragments can only be decrypted with K_{ses_j}, which are known only to M and *MO* and finally to B_i after making the payment.

  In another scenario, if B_i is unable to obtain all the fragments from M through Pr_j, he/she can request *MO* for digital coin's revocability. Since *MO* keeps the details of all the signed fragments sent by B_i, he/she can accept or deny the request of B_i.

- **Merchant's security:** From the perspective of M, the system is secure because B_i has no idea about the original digital content and the embedded f_i in the purchased copy. Also, B_i cannot claim that Y is created by M since f_i is generated by *MO*, which is trusted by both B_i and M. Also a possible B_i and Pr_j collusion is prevented by assigning the task of selecting Pr_j to *MO* using a reputation-based mechanism. Moreover, a claim made by B_i about receiving invalid fragments from M is repudiated by *MO*. *MO* could deny this claim since he/she stores the hashes of fa_j sent by Pr_j and B_i in the file reconstruction protocol. Thus, in case of a piracy claim made by B_i, *MO* could compare the hashes received from Pr_j with the hashes received from B_i. If the hashes are not equal, *MO* can investigate to determine the cheating party (either Pr_j or B_i).
- **Unlinkability:** Despite the fact that anonymous certificates provide anonymity to B_i, the transactions carried out by the same pseudo ID can be linked to one another. The solution to this problem is to allow a buyer to apply for multiple pseudonyms and anonymous certificates.
- **Coin integrity:** The integrity of C_{B_i} is guaranteed due to the signature of *MO* that generated that coin. Such a signature cannot be computed by anybody else, as the private key of *MO* is never disclosed.
- ***BF* security:** In case a malicious buyer E steals *BF* from another buyer's machine and requests his/her *SP* for *SF* only, this security attack is withstood by our system. After Pr_j deliver fa_j to B_i, both B_i and Pr_j generate a one-time hash of fa_j, encrypt it with $K_{p_{MO}}$ and send $E_{K_{p_{MO}}}(h(fa_j))$ to *MO*. *MO* saves the received $E_{K_{p_{MO}}}(h(fa_j))$ in his/her database along with other transaction details. When E sends a request to *SP* for *SF* only, then *SP* asks E to send the chain of the encrypted hashes of the fragments of the *BF* that he/she had sent to *MO*. In this scenario, E has the *BF* but he/she does not have the chain of the encrypted hashes of the *BF* fragments. In case E generates fake hashes and sends it to *SP*, the *SF* request from E would be denied due to verification of the hashes stored in *MO*'s database.
- **Buyer's privacy:** The attempt of de-anonymization attack by E is withstood by the collusion resistance of the hash function that is used for generation of a pseudo-identity of a buyer. Moreover, E cannot use the pseudo-identity of another buyer because he/she does not know the secret number r shared by the buyer with CA_R. Also, in the *BF* distribution protocol, an attempt by M to find an identity of the buyer by relating proxies to each buyer is withstood by considering a fixed number n of Pr_j for *BF* delivery. Moreover, to ensure anonymous *BF* delivery, *MO* selects random peers and creates an anonymous path in such a way that Pr_j are unable to predict that the next peer in the path is the buyer or some other peer.

6 Comparative Analysis

This section presents a comparative analysis of the proposed system with [3]–[6] in terms of security, privacy and performance. Table I presents the functionality comparison among our proposed system and related P2P content distribution systems.

TABLE I: Comparison of the proposed system with related P2P content distribution systems

Properties	[3]	[4]	[5]	[6]	Our Scheme
Buyer's security	Yes	Yes	Yes	Yes	Yes
Merchant's security	Yes	Yes	Yes	Yes	Yes
Buyer's privacy	Yes	Yes	Yes	Yes	Yes
Traceability	Yes	Yes	Yes	Yes	Yes
Unlinkability	Yes	Yes	Yes	Yes	Yes
Payment mechanism	Yes	Yes	No	No	Yes
Length of anti-collusion codeword	Large	Large	N/A	Small	Small
Computational complexity	Low	Low	High	High	Low

From Table I, it can be seen that the proposed system and the systems in [3]–[6] provide security against customer's rights problem (buyer's security), non-repudiation (merchant's security), piracy tracing, unlinkability and anonymity to a buyer. Our system and the systems in [3], [4] provide an electronic payment protocol between a buyer, a trusted monitor and a merchant in a centralized manner. The systems in [5], [6] do not explicitly consider payment by the buyers to the merchant. While the fingerprinting protocol in our proposed system and the system in [6] are based on Nuida's *et al.* [8] collusion-resistant fingerprinting codewords that result in small length fingerprint codewords, the systems in [3], [4] are implemented with a two-layer anti-collusion code, which results in a longer codeword. Authors in [5] have not considered the collusion resistance of the scheme against collusion attacks. The lower computational complexity of our system and systems in [3], [4] is due to the fact that these systems do not require highly demanding technology (public-key encryption of the content and secure multi-party protocols, among others) unlike the systems in [5], [6]. The proposed system utilizes the idea of permutation and file partitioning to avoid an increased computational costs at the merchant's end, whereas the systems proposed by [3], [4] provide recombined automatic fingerprints, which are generated as contents are downloaded by the buyers from other peers of the system.

7 Conclusions

In this paper, we have proposed a P2P content distribution system, which provides security and privacy to the merchant and the buyer, respectively. The newly proposed scheme can benefit merchants to distribute their contents such as video files, without fear of copyright violation, using the convenience of P2P networks. This scheme reduces the burden of the merchant by only sending a small-sized base file and making use of the P2P network to support the majority of the file transfer process. For distribution of a base file, an asymmetric fingerprinting protocol is performed between the merchant, the proxy peers and the buyer in the presence of a trusted monitor. The buyer's privacy is preserved until he/she is found guilty of illegal re-distribution. The buyer can access the received base file for file reconstruction once he/she makes a payment of the requested content to the monitor. The reputation-based mechanism enables the monitor to select the reputed proxy peers for secure delivery of the fingerprinted content from the merchant to the buyer.

Acknowledgment

This work was partly funded by the Spanish Government through a project TIN2011-27076-C03-02 "CO-PRIVACY".

References

[1] I. J. Cox, J. Kilian, T. Leighton, and T. Shamoon, "Secure Spread Spectrum Watermarking for Multimedia," IEEE Transactions on Image Processing, vol. 6, no. 12, 1997, pp. 1673-1687.

[2] B. Pfitzmann and M. Schunter, "Asymmetric Fingerprinting," Proc. 15th Annual International Conference on Theory and Application of Cryptographic Techniques, EUROCRYPT96, Springer, 1996, pp. 84-95.

[3] D. Megías and J. Domingo-Ferrer, "Privacy-aware Peer-to-Peer Content Distribution using Automatically Recombined Fingerprints," Multimedia Systems, vol. 20, no. 2, 2013, pp. 105-125.

[4] D. Megías, "Improved Privacy-Preserving P2P Multimedia Distribution based on Recombined Fingerprints," IEEE Transactions on Dependable and Secure Computing, vol. PP, no. 99, 2014, pp. 1.

[5] J. Domingo-Ferrer and D. Megías, "Distributed Multicast of Fingerprinted Content Based on a Rational Peer-to-Peer Community," Computer Communications, vol. 36, no. 5, 2013, pp. 542-550.

[6] A. Qureshi, D. Megías, and H. Rifà-Pous, "Framework for Preserving Security and Privacy in Peer-to-Peer Content Distribution Systems," Expert Systems with Applications, vol. 42, 2015, pp. 1391-1408.

[7] 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, pp. 1423-1443, 2001.

[8] K. Nuida, "Short Collusion-secure Fingerprint Codes against Three Pirates," International Journal of Information Security, vol. 11, 2012, pp. 85-102.

[9] L. Lu *et al.*, "Pseudo Trust: Zero-knowledge Authentication in Anonymous P2Ps," IEEE Transactions on Parallel and Distributed Systems, vol. 19, no. 10, 2007, pp. 1-10.

3D Virtual Image Composition System based on Depth Map

Hye-Mi Lee
Dept. of Computer Science
Sunchon National University,
Sunchon, the Republic of Korea
e-mail:lhrooh@sunchon.ac.kr

Nam-Hoon Ryu
WithusBiz Co., Ltd.
Gwangyang, the Republic of Korea
e-mail: nhryu@sunchon.ac.kr

Eung-Kon Kim
Dept. of Computer Engineering
Sunchon National University,
Sunchon, the Republic of Korea
Corresponding Author
e-mail: kek@sunchon.ac.kr

***Abstract*— To complete a film, it needs to go through the process to capture the actual actor's motion and compose it with virtual environment. Due to the excessive cost for production or lack of post-processing technology to make the film, however, it is mostly conducted by manual labor. The actor plays his role depending on his own imagination at the virtual chromakey studio, and at that time, he has to move and consider the possible collision with or reaction to an object that does not really exist. And in the process of composition applying Computer Graphics(CG), when the actor's motion does not go with the virtual environment, the original image may have to be discarded and it is necessary to remake the film. This paper presents and realizes depth-based real-time 3D virtual image composition system to reduce the ratio of remaking the film, shorten the production time, and lower the production cost. As it is possible to figure out the mutual collision or reaction by composing the virtual background, 3D model, and the actual actor in real time at the site of filming, the actor's wrong position or action can be corrected right there instantly.**

Keywords-3D Image; Composition; Chromakey; Depthmap

I. Introduction

It is common that the digital technology is used to produce the video in the broadcasting media including movie. The computer graphic technology, which can realize the creative idea, removed the limitations in the production space and allowed the production of diverse contents by composing the actual image with the virtual environment. In case of the video that the live action is hard, the background image is composed with the actual actor using the computer graphic technology, for which composing the virtual background image or objects with the actual actor should be made seamlessly and the sense of difference in the images should be minimized. Since the realistic and high quality video invested with high cost raised the demand of the audiences for the quality, it is now that the development of low cost and high efficient video production technology is needed to meet such demand of the audiences.

In this article, the system, which can monitor the screen at the same time as filming is made by composing the virtual environment applied with 2-D background and 3-D virtual model with the actor, is intended to be designed and realized. Since the conflicts and the reaction between the actor and 3D virtual objects can be identified during the filming, the filming can be completed correcting the wrong action and location of the actor. In Section II, the depth information-based real-time virtual image composite system will be explained. And in Section III, the results of its realization will be examined, in Section Ⅳ, the conclusion and the expected effects will be presented.

II. Real-Time Virtual Video Compositing System

In this section, the video composing system, which replaces the 2D background by converging the depth information and the actual object and the 3D object can respond each other in real-time, is designed excluding the fragmentary composing by chromakey. Since the purpose of this system is to produce the video according to the story or scenario, all the works like the animation of the virtual environment and 3D object should be performed in advance. This system is divided into the 2-layer composing module, 3D virtual space generation module and the 2D-3D virtual video composing engine.

In case of the foreground and background separation and the background superimposing by the existing chromakeying, the depth buffer is not generated. Since to link with the 3D model, 3D space data should be generated, the foreground and background are separated through the depth-keying method.

As the hardware to obtain precise depth data is expensive equipment, to obtain the depth image and color image at the low cost simultaneously, the RGB image and the depth stream are entered using Kinect. Kinect is the device that generates the 3D space data and provides diverse user experiences by measuring the depth of the video entered through the infrared pattern recognition [1][2]. Figure 1 shows the pipeline of the 2-layer composing module.

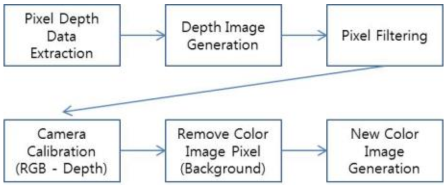

Pixel Depth Data Extraction
Depth Image Generation
Pixel Filtering
Camera Calibration (RGB - Depth)
Remove Color Image Pixel (Background)
New Color Image Generation

Figure 1. Pipeline of the 2-Layer Composition Module

To separate the foreground from the depth image generated, the filtering should be performed. Kinect returns the depth value from 0 to 4095 and each value is responded to the length unit in mm. In 2-layer composing module, the threshold value for the foreground position is set to between 800 ~ 4000 mm, and the pixels in the region, whose depth threshold exceeds 4000, are processed to background through filtering. Figure 2(a) shows the color video entered through RGB camera, and Figure 2(b) shows the depth image, which visualizes the depth of space.

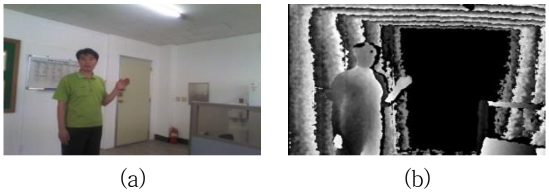

(a) (b)

Figure 2. The process of Separating the Foreground from the Depth Image

The pixel of the depth image that entered through the depth camera has a problem that it is not converted 1 vs 1 with the color stream entered through the RGB camera. The reason for this is because the positions of two cameras are different and therefore, they convert the position of the depth image pixel position into the pixel position of the color image through the camera calibration [3][4]. And then our system removes the pixels, which are not the foreground, generates the new color image and draws it to the area corresponded to the background. Figure 3(a) shows the RGB image of the foreground after filtering the background and Figure 3(b) shows the image that the background of actual filming image entered through the device was converted and composed with the foreground.

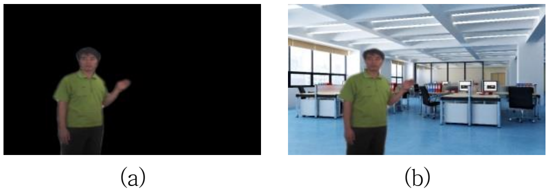

(a) (b)

Figure 3. Synthesis of the Foreground and Background Image based on the Depth Information

At this moment, since the noise is generated around the boundary of the foreground, unless the smoothing process is not performed, the pixels around the foreground remove the noise by compensating the edge.

III. RESULTS OF SYSTEM REALIZATION

Since this system is the technology to reduce the refilming rate when filming the virtual video, the purpose of this system is to prevent the mismatching between the position and the eye line of the actual actor according to the virtual objects that appear in video. Figure 4 shows that the virtual background, actual actor, 3D virtual model are continuously moving. Figure 4(a) shows the results realized as if the 3D virtual model is located closer to the camera than the actor and Figure 4(b) shows the results realized as if the actor is located far away from the camera than 3D virtual model. Each object is moving continuously and is rendering on the screen in real-time.

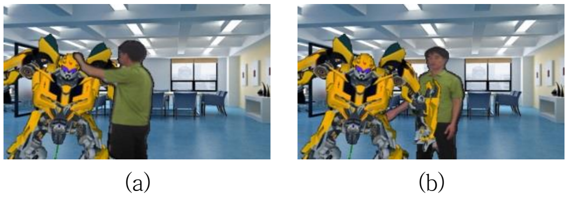

(a) (b)

Figure 4. Screen of the Final Implementation Results

The actor acts in the virtual studio where the 3D objects to be arranged in final outcome do not exist, but in the video, the virtual environment and 3D objects appear together. Therefore, the movement and the flow of the eye line of the actual actor matched with the virtual environment can be observed.

IV. CONCLUSIONS

In the dangerous scene or in the production of video using the computer graphic technology for the SF movie where the live action is difficult, there are many problems such as technical problem, work time, and cost. To complete the high quality video, the work should be done manually but if the action of the actor is not delicately matched with the 3D virtual environment, the entire work done by the actor might have to be redone. About 10% of footage would need to be re-shot, which results in additional time and cost.

Since the depth-based real-time 3D virtual image composition system can verify the position where the virtual environment and the 3D model are inserted in the video in advance and monitor the outcome that the action of the actor and the virtual environment are composited through the real-time screen in the filming site, it can reduce the error that might occur in the post production process. If the number of additional takes can be reduced through this system, the production period of high quality video can be reduced and significant amount of production cost can be saved. Therefore the vitalization of diverse and experimental video production will be expected.

ACKNOWLEDGMENT

This work was supported by the Human Resource Training Program for Regional Innovation through the Ministry of Education and National Research Foundation of Korea (NRF-2013H1B8A2032217).

REFERENCES

[1] C.-K. Lee and G. Park, “A Study on Comparison of background chroma studio for Virtual Studio,” Journal of KOSST. vol. 7, no. 2, pp. 36–41, 2012.

[2] J. Park and J. Park, "Upper Body Exercise Game using a Depth Camera," Journal of Korean Society For Computer Game, vol. 25, no. 1, pp. 61–66, 2012.

[3] J. Webb and J. Ashley, Beginning Kinect Programing with the Microsoft Kinect SDK. California: O'Reilly, 2012.

[4] K.-I. Kim, S.-H. Han, and J.-S. Park, "A Distortion Correction Method of Wide-Angle Camera Images through the Estimation and Validation of a Camera Model," Journal of the Korea institude of Electronic Communication Sciences, vol. 8, no. 12 , pp. 1923–1932, 2013.

[5] S.-C. Kwon, W.-Y. Kang, and Y.-H. Jeong, "Stereoscopic Video Composition with a DSLR and Depth Information by Kinect," Journal of. Korea Inst. of Commun. Inform Sci. (KICS), vol. 38C, no. 10, pp. 920–927, 2013.

Runner’s Jukebox: A Music Player for Running Using Pace Recognition and Music Mixing

Sanghoon Jun, Jehyeok Rew, Eenjun Hwang
Department of Electrical & Computer Engineering
Korea University
145 Anam-Ro, Seongbuk-Gu, Seoul 136713, Korea
e-mail: ysbhjun@korea.ac.kr, rjh1026@korea.ac.kr, ehwang04@korea.ac.kr

***Abstract*—In this paper, we propose a smartphone music player called Runner’s Jukebox (RJ) that is especially effective for walking and running activity. RJ consists of user pace speed recognizer, music archive with feature data, and dual music player. To show its effectiveness, we assume two scenarios in this paper: (i) users are required to catch up with music tempos of predefined playlist. (ii) Songs are played dynamically or changed to match with the speed of user pace. We present novel pace recognition algorithm that enables RJ to recognize user pace in any orientations and even in the pocket. Various methods for playing and selecting songs based on user pace speed are presented. In addition, in order for seamless playback of songs, two mixing schemes are presented for music player. To evaluate the performance of our proposed scheme, we carried out several experiments on an Android platform. We report some of the results.**

Keywords- Interactive music; Mobile application interface; Pace tracking; Context-aware computing; Wearable computing.

I. Introduction

Music can be used as a good stimulus for increasing effects of physical exercise [1]-[3]. It helps exercisers to have positive mood while workout and hence leads to better performance. In other words, exerciser can use music as a means to improve their exercise workout.

Explosively growing popularity of smartphone and wearable devices has brought entirely new interfaces and applications to human beings. Many studies have been done recently based on these state-of-the-art technologies. For instance, in the traditional music streaming services, users play songs by just clicking or selecting the desired ones. Hence, they should decide the playback list beforehand or continuously, which makes them easily fed up with the tasks or unsatisfied with the outcome.

In this paper, we present a smartphone music player called Runner’s Jukebox (RJ), which takes different approaches to playing music compared to the traditional one. By combining the sensor data analysis of smartphone (or wearable devices) and music feature extraction, the music player can play songs to be matched with user pace and adjust the playback speed dynamically to catch up with the pace afterwards. The idea of music player tracking user pace is not novel in itself. A few studies have been done for tracking and synchronizing a song to user’s pace [4-8]. The contribution of this paper focuses on the practical implementation of music player integrating aforementioned technologies on Android smartphone and wearable devices and some evaluations of actual user behaviors. The rest of this paper is organized as follows: In Section 2, we present a brief overview of the related works. Section 3 presents the overall system architecture, music play modes and song mixing method. Section 4 describes the experiments that we performed and some of the results, and the last section concludes the paper.

II. Related works

Wijnalda et al. [6] proposed the personalized music system called IM4Sport. It helps select music that suits a training program, changes playback to reflect or guide current sport performance, and collects data for adapting training programs and music selections. They proposed the prototype system that consists of personal computer, a portable music player, a heart sensor strap and a pedometer. Moens et al. [7] presented D-Jogger, a music interface that makes use of body movement to dynamically select music and adapt its tempo to the user’s pace. While implementing the interface, they focused on the entrainment which is the synchronization of music and walking. Oliver et al. [8] presented MPTrain, a mobile phone based system that helps users to achieve their exercise goals. Using extra physiological sensors of mobile hardware, the proposed system allows the user to select desired exercise pattern such as heart-rate goal. While several studies are based on their own devices and systems, we focus on the development of algorithms on popular and general mobile platforms such as android platform without any additional devices

A few applications are already available for this purpose. Recently, applications such as CruiseControl, TempoRun and TrailMix are released to play and approximately match a song to user pace. In the case of RockMyRun, mixed tracks with certain tempo are provided to the user for better efficiency of running or walking. However, it is designed to provide pre-mixed track without any customization or dynamic pace matching. In this paper, we focus on more advanced features such as mixing songs continuously and adaptively based on accurate user pace recognition to

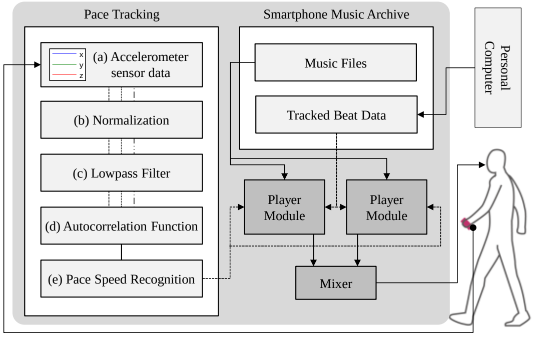


Figure 1. Overall system architecture

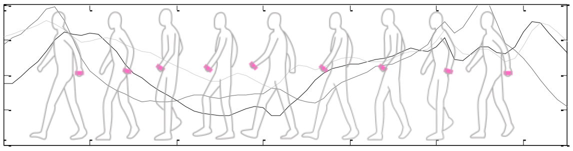


Figure 2. 3-axis acceleration changes of human pace

provide better exercise efficiency and convenience for joggers.

III. RUNNER'S JUKEBOX

A. Implementation

The RJ application consists of three main components: (1) Pace recognizer based on device sensor data. (2) Music archive that contains music features such as beat and tempo. (3) Two music players and one mixer that support diverse playing modes and music mixing. Its overall architecture is shown in Fig. 1.

B. Pace Recognizer

In RJ, we suppose that the user holds the smartphone on his hand or armband while the user walks or runs. Since the smartphone gets acceleration while swinging his/her arm, the pace can be detected by analyzing acceleration sensor data of smartphone. Fig. 2 shows an example of acceleration changes of human pace. Acceleration data may change unpredictably while he/she walks or runs.

The Android platform provides hardware-based sensor data for monitoring the motion of a device. For detecting the pace, the accelerometer sensors in x, y and z axis are used. The platform can monitor sensor data with short delays (0ms to 60ms).

In this paper, we propose an algorithm for recognizing the pace. Fig. 3 shows the detailed steps for pace recognition. First, 3-axis acceleration sensor data in sample window is acquired from device. In order to remove gravity or unnecessary acceleration, three dimensional data are normalized to zero-center, respectively. Similarly to the method presented in [7], low-pass filter based on Butterworth IIR filter is applied to remove noises. In order to observe the periodicity of user pace, an auto-correlation function is computed from each axis sensor data. Three auto-correlation functions (ACF) are fused together using the following equation to calculate their ACFF:

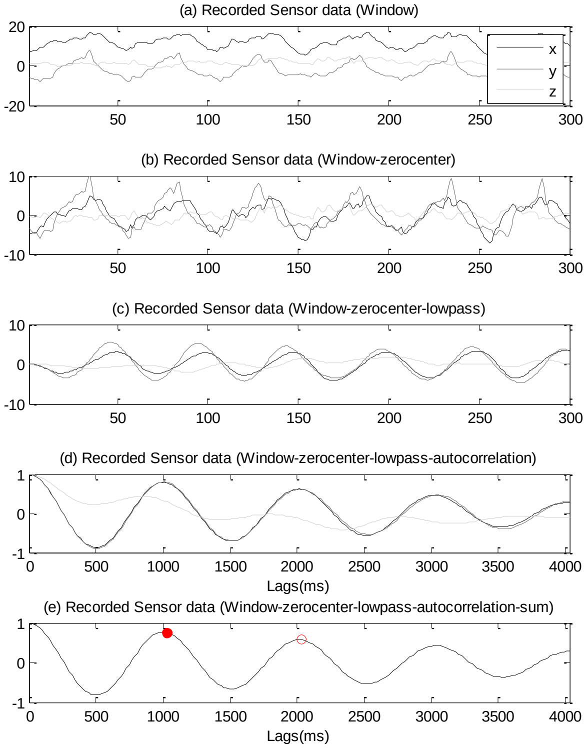


Figure 3. Recognizing pace from acceleration sensor data

$$ACFF(\tau) = \sum_{d \in \{x,y,z\}} \mathrm{var}(d) \times ACF_d(\tau) \tag{1}$$

where var(d) is the variance of d-axis sensor data within a window. Among the three axis sensor data, we consider the axis with the largest variance has more primary information for recognizing periodicity. This approach makes the pace recognition independent of any orientation of the sensors. This is important since user's gripping style and orientations of wearable device can be different.

In order to represent the speed of user pace quantitatively, we define SWing Per Minute (SWPM).The period of swings can be observed by finding the lag having the first local maximum value of auto-correlation function. The following equation shows how to calculate the SWPM from the period.

$$SWPM = \frac{60}{FirstPeakLag(ACFF)} \tag{2}$$

The value of SWPM indicates how fast the user walks or runs. For instance, we empirically found that the SWPM of typical walking adults is between 10 and 40. The detailed experiment for SWPM values will be described in later section.

Typical time taken for a human to swing his arm is less than 3 seconds. Therefore, we calculated the SWPM using 3 seconds of window in every 20 milliseconds. User can speed up or down instantly for synchronizing his/her pace to the playing song. To make the SWPM closer to user speed, we used the moving average window method.

C. Music Archive Construction

In order to synchronize song to user pace or play music continuously, beats within a song should be tracked. In this paper, we employed BeatRoot, the java based beat tracking system that proposed in [9]. Beats are tracked and their timestamps are stored as a text file in a personal computer as shown in Fig. 1. Subsequently, tempo in terms of beats per minutes (BPM) is calculated by the average intervals of beats.

D. Music Player

While songs are played in RJ, their tempo needs to be adjusted according to the user's SWPM. Since the SWPM is correlated to an arm swing of two footsteps, the proper tempo of song is calculated by the following equation

$$BPM = 2 \times SWPM \qquad (3)$$

SoundTouch, which is a sound processing library implemented on C++ platform, is employed for adjusting playback speed without pitch variation [10]. Java native interface is used to enable the library on java platform. Using the library, songs can be played with its playback speed adjusted dynamically according to user SWPM.

The RJ player provides two modes for music selection and two options in each mode for adapting its playback to the user pace.

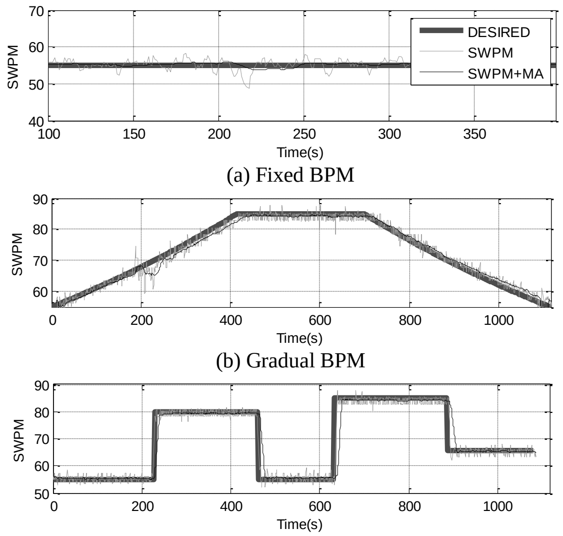
1) User Mode

In this mode, the player plays songs in the playlist which was made by the user beforehand. This mode provides two particular options: fixed and dynamic pace. With the fixed pace option, tempo of all songs in the playlist is changed to some fixed BPM. On the other hand, with the dynamic pace option, the songs in the playlist are played sequentially and its playback speed is changed dynamically depending on user SWPM.

2) System Mode

In this mode, the playlist is made automatically by the system and two options are supported: Beat-aware and program. With the beat-aware option, the system monitors user SWPM continuously. When any change is detected in the user SWPM, the player finds songs having the most similar tempo. In other words, if the user changes his/her speed up or down beyond the adjustable tempo range of current music, the player stops current music and plays another song with corresponding tempo. Another option is program where songs are played based on the predefined programs. For instance, if the user defines a program that consists of 5 minute run and 5 minute walk with 5 repeats, the system automatically prepares a playlist of songs appropriate for the program.

3) Song Mixing



(a) Fixed BPM

(b) Gradual BPM

(c) Stepwise BPM

Figure 4. Examples of pace matching to music tempo

RJ has two player modules for music mixing to support seamless play. In other words, they play songs seamlessly in such a way that songs are sounded like one. The two player modules synchronize their music play based on the beat feature in the music archive. There are two options for song mixing: crossfading and cutting. In the crossfading mix, two songs are mixed gradually by volume transition over time. This method can be used when the SWPM is steady. In the cutting mix, a new song starts as soon as current song ends. This method is used for mixing music when the SWPM changes suddenly.

IV. EVALUATION

To evaluate the performance of our scheme, we performed several experiments for 10 females and 10 males. 15 of them are 20s and the others are 30s. Also, 5 of them are daily exerciser and 12 of them exercise regularly. The others exercise rarely. We collected 100 songs (MP3 files) of various genres from a song streaming service, which include electronic, hip-hop, pop and rock. Since our evaluation is related to exercise, we collected the user's custom playlists that were programmed for their exercise or workout.

A. SWPM Ranges

Based on the proposed pace metric (SWPM), we collected the actual walking/running records from 20 subjects. The subjects were asked to walk or run about 2km with their own pace. Average SWPMs of walking and running are 52.6 and 82.9, respectively. We implemented a prototype Runner's jukebox and tested it on two android smartphones: LG Optimus G (Snapdragon S4 Pro 1.5GHz Quad-core processor and Android OS v4.1) and Samsung Google Nexus S (Cortex-A8 1GHz processor Android OS v2.3).

B. User Pace Matching to Music Tempo

In the first experiment, we evaluated the performance of the pace recognizer. In the experiment, users are required to catch up with music tempos of predefined playlist. Based on the average SWPM, we generated music clips having dynamic BPM ranging from 110 to 170 and played to them. We defined three different types of BPM modes: fixed, gradual and stepwise. We generate music clips according to the BPM modes and then asked subjects to walk or run to the music clips. Fig. 4 (a) shows the result of pace recognition under fixed BPM. Fig. 4 (b) shows the result of user pace recognition under gradual BPM. Finally, Fig. 4 (c) shows the result of user pace recognition under stepwise BPM. To evaluate their accuracy, we calculate their average difference using the following equation.

$$AvgDifference = \frac{\sum_{i \in SWPM\ samples} \left| SWPM_i - (bpm/2) \right|}{\#of\ SWPM\ samples} \tag{4}$$

In addition, we measured their standard devation and minimum -maximum differences. The results are shown in Table I. SWPM+MA is the result of moving average (MA) window after SWPM measurement.

TABLE I. SWPM DIFFERENCES MEASURED

	Average Difference	Standard Deviation	Min ~ Max Difference
SWPM	2.05%	1.654	-49.5 ~ 15.0
SWPM+MA	1.88%	0.905	-20.7 ~ 12.1

C. Pace Recognizing Conditions

Since our proposed scheme aims to assist exercise or workout, its robustness to the external conditions is very important. To this end, we evaluated the accuracy of pace recognition under three different conditions: smartphone is held in exerciser’s hand, armband and pocket. Subjects are asked to keep pace 85 SWPM while jogging. Fig. 5 shows the SWPM tracking results of three different conditions. Table 2 shows their comparison in terms of average SWPM difference.

TABLE II. COMPARISON OF RECOGNITION CONDITIONS

	Average SWPM difference (85 SWPM)
Hand	1.74%
Armband	1.35%
Pocket	4.51%

As shown in Fig. 5 and Table II, the proposed scheme can recognize user pace quietly correctly in all conditions. Even with the device in the pocket, our scheme can recognize user pace quite accurately. This is interesting because this is not related to user’s swinging action. According to the experiment result, we got the best pace recognition accuracy using the armband.

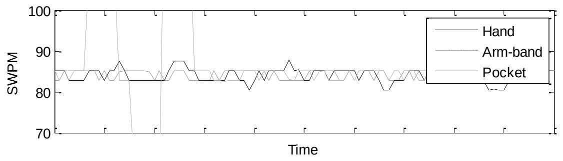


Figure 5. SWPM tracking result in different recognition conditions

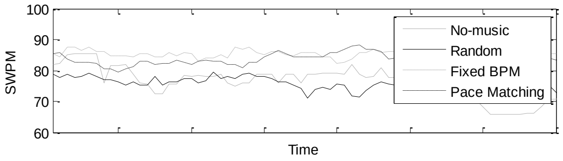


Figure 6. SWPM tracking result of a subject

D. Song Playing Method

In this experiment, we measured the effect of music playing on the exercise or workout. Subjects are asked to keep their own pace as much as possible while jogging 2km course. For the comparison, we considered 4 different jogging conditions: no music, randomly selected music, fixed bpm music and user pace matching music. The measurement is carried out every other day for each subject. Fig. 6 shows SWPM tracking result of a subject for 4 different conditions. Table 3 shows their average SWPMs and the standard deviation (STD).

TABLE III. AVERAGES AND AVERAGE STANDARD DEVIATIONS OF SWPM

	SWPM	SWPM STD
No music	78.8	4.33
Randomly selected music	75.8	3.81
Fixed bpm(170) music	85.4	1.17
Pace matching music	85.1	1.91

From the result, we can observe that controlled music plays such as fixed bpm and pace matching give better exercise effect in term of SWPM. Also, they show less variation than randomly selected music or no music. This means that music can be used to maintain or keep up the effectiveness of exercise or workout.

V. CONCLUSION

In this paper, we proposed Runner’s Jukebox, a smartphone music player to improve the user’s experience and performance of typical exercises such as walking and running. Based on user pace detection and music feature analysis, our proposed music player retrieves songs and adjusts their playback speed according to the detected user pace. We presented a practical implementation on an Android platform. In addition, for seamless music playback, we proposed music mixing methods. To evaluate the effectiveness of our system, we performed several experiments and showed that our system can work on Android platform with accurate recognition and positive effect on user workout.

ACKNOWLEDGMENT

This work was supported by Basic Science Research Program through the Ministry of Education of the Republic of Korea and the National Research Foundation (NRF) of Korea (NRF-2013R1A1A2012627)

REFERENCES

[1] G. Tenenbaum, R. Lidor, N. Lavyan, K. Morrow, S. Tonnel, A. Gershgoren, J. Meis, and M. Johnson, “The effect of music type on running perseverance and coping with effort sensations,” Psychology of Sport and Exercise, vol. 5, no. 2, pp. 89–109, Apr. 2004.

[2] C. I. Karageorghis and P. C. Terry, “The psychophysical effects of music in sport and exercise: a review,” Journal of Sport Behavior, vol. 20, no. 1, pp. 54–68, 1997.

[3] F. Styns, L. van Noorden, D. Moelants, and M. Leman, “Walking on music,” Human Movement Science, vol. 26, no. 5, pp. 769–785, Oct. 2007.

[4] G. T. Elliott and B. Tomlinson, “PersonalSoundtrack: context-aware playlists that adapt to user pace,” CHI ’06 Extended Abstracts on Human Factors in Computing, pp.736–741, Apr. 2006.

[5] J. A. Hockman, M. M. Wanderley, and I. Fujinaga, “Real-Time Phase Vocoder Manipulation by Runner’s Pace,” Proceedings of the International Conference on New Interfaces for Musical Expression, pp. 90–93, June 2009.

[6] G. Wijnalda, S. Pauws, F. Vignoli, and H. Stuckenschmidt, “A Personalized Music System for Motivation in Sport Performance,” IEEE Pervasive Computing, pp. 26–32, 2005.

[7] B. Moens, L. van Noorden, and M. Leman, “D-Jogger: Syncing Music with Walking,” Proceedings of SMC Conference 2010, Barcelona, pp. 451–456, 2010.

[8] N. Oliver and F. Flores-Mangas, “MPTrain: A Mobile, Music and Physiology-based Personal Trainer,” Proceedings of the 8th Conference on Human-computer Interaction with Mobile Devices and Services, pp. 21–28, 2006.

[9] S. Dixon, “Evaluation of the Audio Beat Tracking System BeatRoot,” Journal of New Music Research. vol. 36, no. 1, pp. 39–50, 2007.

[10] SoundTouch Sound Processing Library. Available from: http://www.surina.net/soundtouch/. [retrieved: Feb. 2015].

The DCP Bay: Toward an Art-House Content Delivery Network for Digital Cinema

Nicolas Bertrand, Jean Denis Durou and Vincent Charvillat

IRIT, UMR CNRS 5505
Université de Toulouse, France
email: nicolas.bertrand@isf.cc, durou@irit.fr, vincent.charvillat@enseeiht.fr

***Abstract*—Cinema theaters have arrived in the digital era. The Digital Cinema Initiatives has chosen Digital Cinema Package (DCP) as format for the distribution of feature films. No suitable economical nor technological model is proposed for DCP content delivery to art-house theaters. The existing solutions are too expensive or not adapted. Therefore, we conduct this research activity in cooperation with Utopia cinemas, a group of art-house French cinemas. Utopia's main requirement (besides functional ones) is to provide free and open source software for DCP distribution. In this paper, we present a Content Delivery Network for DCP adapted to art-house. This network is operative since mid 2014 and based on torrent peer-to-peer technology inside a multi-point VPN.**

Keywords–Digital cinema; Content Delivery Network; Peer-to-peer; VPN.

I. INTRODUCTION

Cinema theaters switched from 35 mm prints to digital era. The Digital Cinema System Specification (DCSS) [1], provided by Digital Cinema Initiatives (DCI), is now a worldwide standard. The specification describes how to create, distribute and project a Digital Cinema Package (DCP).

Our research is conducted in collaboration with Utopia cinemas, a network of five independent theaters in France. More independent theaters support this activity through the Indépendants, Solidaires et Fédérés (ISF) association. Those theaters initiated this research because they need to understand the implications of changing to digital. Primarily, they are concerned about becoming dependent on a single company's technology for presentation, and want us to provide Free and Open Source Software (FOSS) for Digital Cinema (DC). As exhibitors, they are mainly concerned in two parts: a projection system and a system for DCP reception in theaters in a dematerialized way.

The first topic is addressed in [2]. It is a full software projection system running under VLC, with JPEG2000 decoding optimization for DC (the VLC-DCP part in Figure 1). This paper is on the second part: the DCP transmission system. We designed a Content Delivery Network (CDN), peer-to-peer based, for DCP delivery to exhibitors. Our goal is to deliver on time a DCP to be used in the projection system (not to stream the content). A major requirement is also to design a CDN which fits independent distributors and theaters ecosystem: deliver many contents for many theaters in a cost effective way.

In Section II, we present the environmental reasons to propose our CDN. In Section III the CDN architecture is presented. In Section IV the CDN is evaluated, and in Section V we analyse the evaluation results.

II. WHY AN ALTERNATIVE CDN FOR ART-HOUSE THEATERS?

Once a movie is realized by producers, the movie shall be distributed. We study the case of distribution to theaters (not the video distribution as DVD or VOD). The distributors acquire the film rights per country. They negotiate per theater when and how long the movie will be screened (cf. Figure 1, blue arrows for right owners negotiation, red for DCP data flow). Our work is based on film distribution in France. Film industry in France is ruled by the Centre National du Cinéma et de l'image animée (CNC) who delivers permissions to distribute a film. CNC distinguishes three classes of theaters (big, medium and small) and sets labels to films (art-house, research, etc.). We focus on distribution to medium and small theaters, with CNC art-house labels. Big (multiplex) theaters already have distribution solutions (mainly via satellites) adapted to their business (for instance distribution of more than 900 copies to exhibitors). At the opposite 60 % of the movies in 2013 have been screened in less than 100 theaters (CNC source) and the great majority of these films were art-house ones.

Technically, distributors don't create (except for the very small ones) DCPs themselves. DCP mastering is done by cinema industry laboratories. Basically DCP mastering consists in creating a DCP according to DCSS standard (3 big steps: compress video, create container and encrypt DCPs). The DCP can then be delivered to exhibitors. In France, 2012 is the year which represents the switch from 35 mm prints to digital copies. At this time DCPs were transported like 35 mm prints: via mail transporters. The switch to dematerialized transfer is currently in progress. The dematerialized transfer was initiated by big Network Service Providers (NSP): in France, Orange via Globecast and TDF via Smartjog.

Globecast, for some theaters (big or representative ones), proposes free equipment (rent of 1, 2 or 4 ADSL network links and reception box), the DCP transfers are cost free also. The distributor is in charge to pay the transfers to theaters. The transfer cost is expensive for medium and small distributors, so a few switched to dematerialized. Taking en example, Utopia Bordeaux theater is on the top 5 of French cinemas playing art-house movies. The cinema was free of charge equipped with Globecast system: after 3 years of usage, they can receive less than 10 % of the screened film via this system. The remaining 90 % are received by hard disk via traditional mail delivery.

ISBN: 978-1-61208-400-8

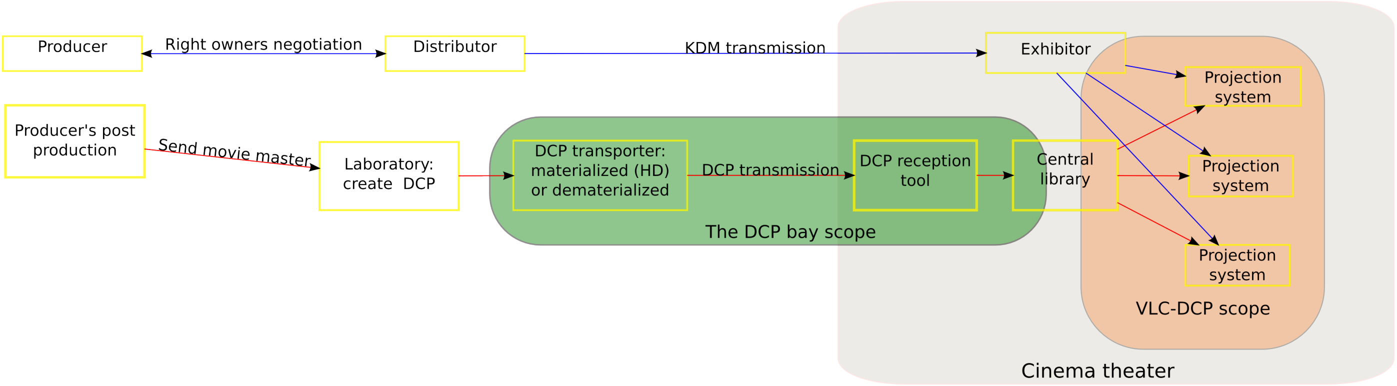


Figure 1. Overview of DC system. The figure represents the DCP distribution from distributors to exhibitors.

Furthermore, depending on the theater, Globecast offer is free or not (installation cost and DCP transfer cost).

Other issue: transporters try to deal with distributors for exclusivity on delivery. A theater shall then be equipped with several reception mechanisms to receive all DCPs. So, we have here an inter-operability problem. Inter-operability is one of the keywords of DCSS: creation of DCPs, projection system, security are largely described in the document. The transportation chapter is only one page long in a 160 pages document.

Affordable costs for art-house cinema and inter-operability are two main reasons to starting this project. In next section we will present our CDN: The DCP Bay.

III. The DCP Bay

The fundamentals are to propose an open network design for DCP delivery to solve inter-operability issues and build it in a cost effective way for distributors and exhibitors. By inter-operability, we mean a standard public protocol for DCP transfer based on FOSS and independent (from NSP, from distributors and from theaters projection system).

A DCP can grow up to 400 GB (3 hours of a movie with a compression ratio of 250 Mb/s). The observed DCP mean size is around 200 GB. So, DCP distribution is challenging due to the large size of the content and on time delivery for first screening. The distributors can make available the DCPs to distribute from one month to 24 hours before the projection.

We excluded solutions based on JPEG2000 scalability [3] and multiple-description-based distributed system [4] because the majority of DCPs to distribute are encrypted: it is not possible to access JPEG2000 data.

Our transfer system is peer-to-peer (torrent) based. Usage of torrent files is common and implementation is open. Torrent allows transfer of large files and ensures also the integrity of the system. For security and maintenance simplification all the peers are in a multipoint VPN called tinc. Tinc is a well known open source VPN simple to set up and with numerous functionalities [5].

A distributor delivers the DCPs to our platform (by secure FTP or by torrent or by traditional mail delivery). When a DCP is received in one of our servers, the integrity is verified (the information for hash verification is present in DCP meta-data) and the torrent file is created. The DCP is replicated in other servers to create seeders. Then, the DCP can be delivered to theaters, the exhibitor selects the torrent to download in our torrent tracker (also inside the VPN), and starts the transfer via his torrent client interface. Once the DCP received, the exhibitor can transfer it to the DC library or DC projector server (usually via FTP). For running The DCP Bay (TDB) in a cinema, the exhibitor needs a dedicated machine. We do not impose to the theater a specific hardware, neither a specific NSP, but we ask to use a dedicated machine to not be dependent on any DC equipment provider.

Thanks to our VPN and peer-to-peer architecture, we can propose aggregation of network links to increase theater bandwidth. The aggregation solution is presented in Figure 2. For each network link, we have a modem. In the TDB theater machine at VPN level, two connections are made to two TDB servers, and we create separate routes via each modem. So, we receive data from seeder 1 via modem A, and data from seeder 2 via modem B. At torrent client level both seeders are viewed, via the VPN tunnel, and download rate for each seeder is limited by each network link. This mechanism can be extended to multiple seeders by balancing the connections between modems. This aggregation solution is easy to set up in the theater and does not need any specific connection from an NSP. We can even add network robustness by selecting separates NSP for each network link.

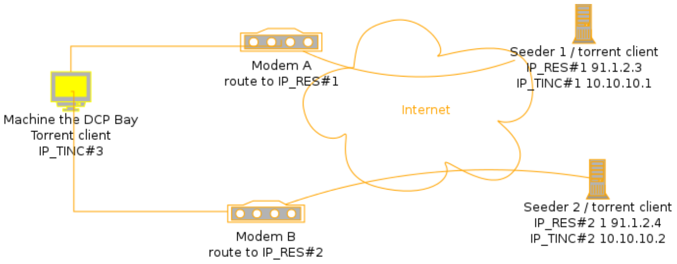


Figure 2. TDB network link aggregation. The TDB machine can receive torrent data from the 2 network links by specifying routes at VPN level. For the torrent client all the machines are in the same VPN network (10.10.10.XXX).

TDB have server machines in data-centers. We work with non profit NSPs (tetaneutral, aquilenet, LDN). In tetaneutral one machine, with 9 To of disk storage, is used for DCP re-

TABLE I. PROGRESSION OF TRANSFERRED DCPs

Month	Number of transfers	Transferred data (in GB)
Jun 2014	1	153.04
Jul 2014	24	3450.19
Aug 2014	12	1656.76
Sep 2014	33	4000.11
Oct 2014	76	10290.37
Nov 2014	91	10355.58
Dec 2014	66	9046.71

ception via FTP, integrity verification, torrent creation, torrent tracker hosting and seeding. The second machine replicates the DCP of the first one and acts as a seeder also. The third one is a virtual machine for archiving 'old' DCPs, and seeds only the archived DCP (16 To of disk storage). The machines in aquilenet and LDN act as cache machines. We send a DCP to these machines when we know that it will be transferred to many theaters at the same time.

IV. THE DCP BAY EVALUATION

TDB is operative since June 2014. Table I presents the number of DCPs transferred by month to the theaters. All the connected theaters (from one to five screens) are art-house ones but some of them have also mainstream movies. The theaters are connected with different connection types (optic fiber, VDSL, ADSL) or usage (shared, dedicated or aggregated lines). They can be equipped only by TDB, but also by others DCP transporters (Globecast, Smartjog, Cinego). This represents heterogeneous bandwidth and according to theater screening it implies several distribution scenarios.

A. Tinc VPN evaluation

All the DCPs transit (except the FTP transfers) via VPN and tunnels. All the traffic is made on the same tunnel. Data transit inside tunnels does not have the same performance as direct network transfer. We connected two machines directly via an Ethernet wire and both with gigabit connections and running under the same Linux kernel. We performed transfers of 1 GB files via wget commands. With direct connections we measured rates of 960 MB/s. Via the tunnel we measured rates of 220-250 MB/s. We deactivated the ciphering to not reduce VPN performances and only evaluate the tunnel performances.

At a more macroscopic level, we measured also non constant rates between distant machines via tinc connections. This problem is probably due to NAT firewall rules on network link modems as mentioned in [6].

B. Torrent client evaluation

The selected torrent client is transmission, because it is a FOSS torrent application. And the software provides also a web interface to manage the downloads. All the theaters have their own torrent clients accessible via a web interface (example http://32.cinema.tdcpb.org for Utopia Tournefeuille). Compared to other FOSS torrent clients transmission is slower (in download time) than rtorrent in high speed networks. We measured the download bandwidth inside the tetaneutral local network. We downloaded a DCP from a server to a virtual machine. With transmission the maximum measured bandwidth was 96 MB/s. With the same DCP we measured with rtorrent download rates of 230 MB/s. After code reading in transmission client there are some timer loops which explain the slower bandwidth.

C. Global TDB evaluation

We use a torrent tracker, named XBTT, to log all the transfers. We can have the start and end times for the transfer of each torrent. A sample of DCP transfer is presented in Table II, which corresponds to the time window of Mr Turner transfer (2 days, 8:45:45). All the DCP transfers presented in this table have been ingested in a machine (utopian7) in the laboratory network and then transmitted to server tdcpb31 and to theaters (cineXX). In that case the DCPs are available to download for all the theaters. Since a DCP is available to download in utopian7, this machine can be considered as a primary seeder. The machine tdcpb31 is where all the DCPs are finally stored and located at tetaneutral. In Table II, DCPs are transferred to five theaters (noted cine35, cine69, etc.). The theaters have different download bandwidths, the cine36 is connected to 100 Mbps optical fiber. For cine36 in that case the maximum reached rate is 21.30 Mbps, so 5 times less than the theoretical bit-rate of 100 Mbps. This is partly explained by the optic fiber NSP (we did not reach more than 80 Mbps in direct download tests), due also to several DCPs downloaded at the same time for instance (Timbuktu), and also due to VPN tunnel limitations. The machine tdcpb31 is connected to a high speed network (with a 1Gbps Ethernet card). It is downloading (and partially uploading to other theaters) DCP from utopian7 (MrTurner, Timbuktu, Quand vient la nuit) and at the same time seeding to theaters (Men women, White God). The download rates of tdcpb31 are slow (11.68 Mbps) regarding the expected 1Gbps connections. The VPN tunnel does not support traffic rates greater than 10/15 Mbps when traffic is in both directions (upload and download). In a simpler case: no upload traffic in tdcpb31, and utopian7 is only uploading to tdcpb31. We measured download rate of 30 Mbps (not represented in Table II) and if we add a virtual machine in tetaneutral network, which also downloads the same DCP as tdcpb31 from utopian7 we reached rate of 58 Mbps for tdcpb31. The limitation is in the tunnel, not in the connection between utopian7 and tetaneutral.

V. TDB EVALUATION ANALYSIS

We designed a CDN for DCP delivery. Functionally, the system is fully operative, and in “beta-production” since 6 months. We found one experimental solution for DCP delivery. Now the problem is to find an optimized solution for an increase of payload (more theaters connected and more DCPs to transfer). Putting aside the problem of tunnel transfer limitation, which shall be addressed at a more “network” level, we need to evaluate if the system can distribute all the DCPs to all the theaters at an expected date (the first screening day).

Figure 3 represents 8 months of film screening at cinema Utopia Bordeaux. Each Wednesday new films are projected. The number (and the size) of projected movies change each week. The blue bars represents in GB the amount of new films each week. The calculated size is the real one, sum of DCP sizes. The size of a DCP varies according to film duration and compression ratio. The red line represents in GB the total amount of data that can be transferred on a network link per week. From the logged transfers we extract a typical download payload of 8 Mbps. During the 8 months represented

TABLE II. DCP TRANSFER SAMPLE.

Machine	Start	End	Duration	Movie	Size (GB)	Rate (Mbps)
cine35	2014-12-04 12:49:43	2014-12-06 16:19:13	2 days,3:29:30	Men women	202.34	8.73
cine69	2014-12-04 21:45:31	2014-12-07 03:06:57	2 days, 5:21:26	White God	203.16	8.46
tdcpb31	2014-12-05 11:07:25	2014-12-06 11:26:57	1 day, 0:19:32	Timbuktu	127.80	11.68
tdcpb31	2014-12-05 12:56:53	2014-12-07 10:32:29	1 day, 21:35:36	Quand vient la nuit	186.73	9.10
cine36	2014-12-05 13:46:24	2014-12-06 10:24:20	20:37:56	Quand vient la nuit	186.73	20.11
tdcpb31	2014-12-05 16:01:45	2014-12-08 00:47:30	2 days, 8:45:45	Mr Turner	135.50	5.30
cine36	2014-12-05 19:48:15	2014-12-06 14:49:54	19:01:39	Timbuktu	127.80	14.93
cine36	2014-12-06 10:07:05	2014-12-07 00:15:25	14:08:20	Mr Turner	135.50	21.30
cine56	2014-12-06 23:23:40	2014-12-07 21:59:46	22:36:06	Mr Turner	135.50	13.32
cine34	2014-12-07 00:07:30	2014-12-08 09:42:40	1 day, 9:35:10	Mr Turner	135.50	8.97
cine36	2014-12-07 10:36:22	2014-12-09 23:00:05	2 days, 12:23:43	White God	203.16	7.48

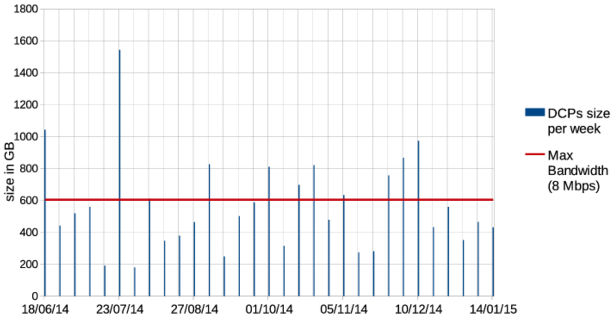


Figure 3. Movies distribution at theater Utopia Bordeaux. 8 months of DCP screening are represented (real situation). Each week new films are presented. We represent in the blue bars the amount in GB of DCP to receive.

in Figure 3 in 10 cases we cannot send all the DCPs in a week. A solution is some cases, can be to use the free payload of the previous weeks to start sending DCPs. This solution is limited in time by the availability of the DCPs from the distributor.

The availability of the DCPs can be short (from two weeks to less than one week) for movies in first screening week. Another solution is to increase the transfer payload, which can be achieved by adding a new network link and using the TDB aggregation method. Adding a new link is not always possible, it depends on theater localisation and infrastructure.

For TDB, our goal is to distribute all the movies to a theater. If we cannot achieve all the transfers in dematerialized, we will send the remaining DCPs by physical transporter, which will increase our costs. A simple solution to minimize the costs is to send by transporter the largest DCPs, to free payload on the network link. For each theater, we have a distribution to achieve in time and the same movie can be distributed to multiple theaters at the same time. To achieve this goal we can adapt the mathematical model from [7] to our CDN model. The aim of the proposed model is to minimize the network operational cost and respect bandwidth constraints and download time. We will add to this model the DCP availability from distributor as input parameter.

VI. Conclusion

In this paper, we have presented our operative CDN for DCP distribution. The CDN will continue growing by connecting more theaters. We have demonstrated that we can create a FOSS CDN, which respects all the constraints of DC content delivery. Our system does not depend on any NSP provider and can be deployed in any network. And we also propose a network link aggregation for theaters without usage of dedicated NSP service.

We have raised some networking issues: the bandwidth limitation in the VPN tunnel. We will continue to investigate this issue to deeply understand tunnel limitations and increase our CDN global bandwidth. The proposed solution shall not (or slightly) increase the network operational cost and respect the FOSS requirement.

Future work will be to adapt the proposed model by [7] to peer-to-peer distribution and increase the storage capacity in data-centers. We are configuring a server with storage capacity based on ceph [8] distributed file system and erasure coding algorithms.

References

[1] DCI, Digital Cinema System Specification, 1st ed., Digital Cinema Initiatives, Mar. 2012.

[2] N. Bertrand, J.-D. Durou, and V. Charvillat, “Lecture de DCP pour le cinéma numérique avec le lecteur multimédia VLC et libav/ffmpeg,” in Actes de la conférence CORESA’13, vol. 1, Le Creusot, France, May 2013, pp. 185–190, in french.

[3] H. Sparenberg, T. Joormann, C. Feldheim, and S. Foessel, “Adaptive RAID: Introduction of optimized storage techniques for scalable media,” in Proceedings of the 20th IEEE International Conference on Image Processing, Sep. 2013, pp. 1826–1830.

[4] N. Blefari-Melazzi, D. Di Sorte, M. Femminella, L. Piacentini, and G. Reali, “Performance evaluation of a multicast-based solution for wireless resources discovery,” in Proceedings of the IEEE International Conference on Communications, vol. 5, May 2005, pp. 3254–3260.

[5] S. Khanvilkar and A. Khokhar, “Experimental evaluations of open-source Linux-based VPN solutions,” in Proceedings of the 13th International Conference on Computer, Communications and Networks, Oct. 2004, pp. 181–186.

[6] G. Sliepen, “The difficulties of a peer-to-peer VPN on the hostile Internet,” in Proceedings of the Free and Open Source Software Developers’ European Meeting - FOSDEM, 2010.

[7] D. Di Sorte, M. Femminella, G. Reali, and L. Rosati, “Definition and performance evaluation of a request routing algorithm to distribute digital cinema contents,” in Proccedings of the 4th International Telecommunication Networking Workshop on QoS in Multiservice IP Networks, Feb. 2008, pp. 27–32.

[8] Ceph. [Online]. Available: http://ceph.com

Development of an Adaptive Learning System

Philip Davies
Software Systems Research Group
Bournemouth University
Bournemouth, UK
daviesp@bournemouth.ac.uk

David Newell
Software Systems Research Group
Bournemouth University
Bournemouth, UK
dnewell@bournemouth.ac.uk

***Abstract* - We investigate the requirements for an adaptive learning system. A conceptual model is explored which links together a student model, a tutor model and a knowledge model. We further consider the use of an adaptive engine which allows the system to respond to the needs of individual students, present learning objects according to the preferences of individual tutor styles, allows automatic self-exploration at the level of student maturity and encodes the curriculum in a form that is accessible to the adaptive engine. Our model accurately represents both the structure and content of learning objects in contrast with less structured data models implicit in ontological hierarchies.**

Keywords–e-learning; adaptive; metadata; semantic; ontology.

I. INTRODUCTION

In previous work [1], we proposed an Adaptive Multimedia Presentation System (AMPS) to provide a semi-automated tool for learning that adapts to students' needs. A prototype was constructed and evaluated in a real class environment in the Cisco Academy at Bournemouth University [2]. This showed that undergraduate students benefited from using the AMPS, but preferred a more 'adaptive' system – one that met their individual needs better with less tutor intervention. These results led the writers to consider how this might be undertaken in a systematic way. The principal aim of this paper is to look further at the conceptual, semantic, and ontological modelling issues involved in making a more rigorous adaptive learning system.

In section II, we set out our overview of the Adaptive Learning System and indicate the relation of its component parts. In section III, we look at the student model and indicate its possible structure. In section IV, we look at the tutor model and the demands placed upon the system by allowing tutors to teach in their own idiosyncratic ways. In section V, we discuss the knowledge model which we use to hold both the knowledge structure in a multi connected ontology as well as the learning objects themselves. In section VI we discuss the adaptive engine which links together all these components, while in section VII we conclude by reflecting on the limitations of the model and the role of adaptation in learning.

II. THE ADAPTIVE LEARNING SYSTEM

As in nature, so in computing adaptation can take many forms. But it is important to realise that adaption is always in response to a particular stimulus. As the external factors change so the system adapts its response. This is no less true in education in the case of a learning environment; a student is presented with a range of stimuli and a range of responses are observed. Table 1. Students may be presented with learning materials which are too hard or too easy, students may learn from the learning object and accommodate the new learning as new knowledge which is incorporated into their own knowledge or they may not. Any learning system needs to adapt to these responses of the student.

TABLE 1 ADAPTION METHODS

Stage	Stimulus/state	Adaption	Method
1	Student learns from new material	Next stage of material presented	Automatically determined from subject ontology
2	Student fails to learn from new material	Reinforcement material presented	Automatically determined from subject ontology
3	Student ability tested with new material	General IQ test	Real-time response
4	Student pre-knowledge	Subject knowledge test	pre-lesson test
5	Student learning styles	selection of appropriate formats	Learning style analysis

In normal education systems the adaption is performed with varying degrees of success by the tutor. Possible contributions to the student state will include student prior knowledge, student ability and student learning styles, which we call the basic "student signature".

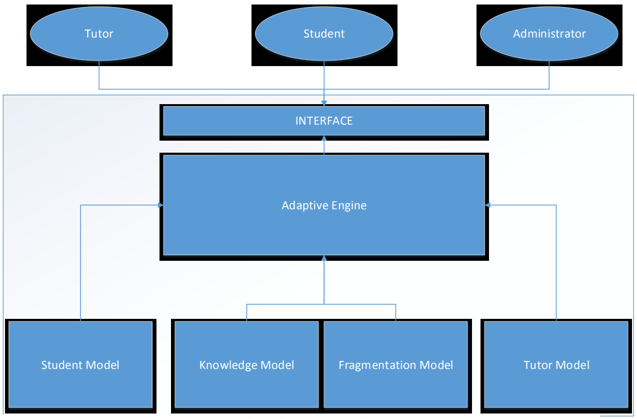


Figure 1 Adaptive Learning System

The system needs to be able to mimic the adaption of the tutor and response of the student as well as contain all the structure of the knowledge system in the form of an ontology together with all the teaching material in different forms to match appropriate learning styles. The structure of the adaptive system with these features is shown in Figure 1.

III. The Student Model

The student is the course subscriber, or person learning the course content and committed to completing a course. Once all courses to which the student has subscribed are complete the student ceases to be a student. The level of knowledge attainment that the student has reached during any point in the course has to be recorded and tracked. This means linking the attainment level to the subject ontology. The determination of whether a particular subject node has been assimilated is through the answering of test questions. The successful answering of these will update the student signature to record which subject nodes have been accessed and mastered. The component of the student model are shown in Figure 2 and will be elaborated below.

Part of the initial processing of the student will require an assessment of the pre-knowledge that the student comes to the course with and this will involve initial testing. The results of this will indicate the present level of knowledge of the student and this will be entered into the student profile or “student signature” as we call it here.

Other factors which determine the way learning is adapted to individual student needs will be include the motivation level of the student which will affect the degree of independence the student is given and the amount of reinforcement and checking on the student activity. Student ability will also be assessed to measure the speed and intelligence level at which a student is able to work.

These and other factors will be incorporated into the student signature which will be assigned to each student and which forms a central part of the Student model. The student signature proposed here is summarised in Table 2 which lists the parameters of the signature in the form of a data model structure.

TABLE 2 STUDENT DATA MODEL

Data Element	Data Type
Present Knowledge Status	Number
Ability Level	Number
Independence Level	Text
Student Signature Level	Text
Motivation Level	Free
Pre-knowledge Level	Number
Test Results	Text
Subject nodes accessed	Number
Subject nodes mastered	Number
Preferred Learning styles	Text
Student number ID	Text

IV. The Tutor Model

The tutor determines the intended delivery, format and content of courses, lessons and learning objects. As such the tutor is responsible for mapping out the ontology structure and knowledge learning map that shows what is to be learned and the relationship of the items being learned.

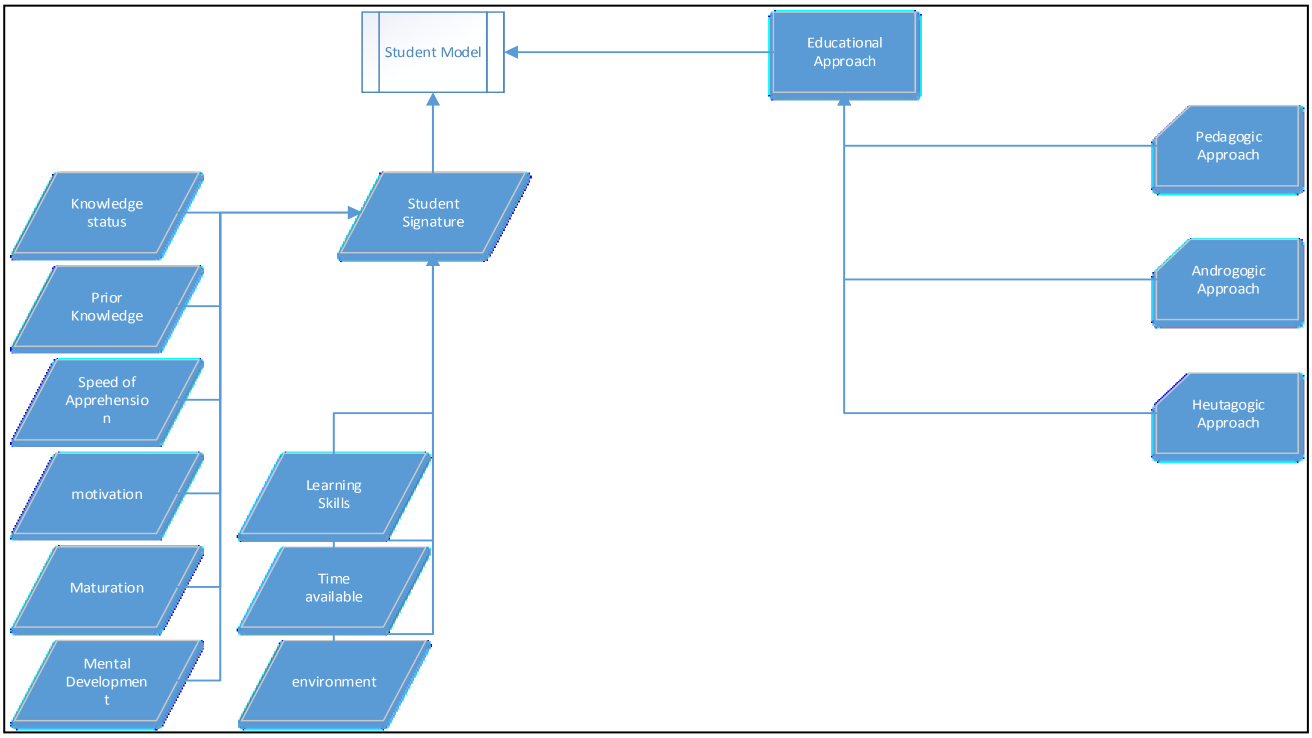


Figure 2 Student Model

The tutor is also largely responsible for determining the direction that learning should take through the knowledge field. The general educational approach that is taken with any student also depends on more general educational factors. Our tutor model is shown in Figure 3 and is designed to encompass the three educational approaches known as pedagogy, andragogy and heutagogy.

A. Pedagogy

Pedagogy is the usual approach adopted in learning institutions in which adults teach children. In this environment, it is recognized that the student has limited critical skills and even less experience. In this circumstance, the flow of knowledge is almost exclusively one way, from the teacher to the student.

B. Andragogy

As the student starts to take more responsibility for their learning [2] the teacher moves to a supportive role in assisting the student with their own learning. In the andragogical approach, learners are actively involved in identifying their needs and planning how they will be met [3].

C. Heutagogy

Heutagogy (from the Greek for “self”) was defined by Hase and Kenyon in 2000, as the study of self-determined learning [5]. Heutagogy extends learning to allow the student to dictate where and when the learning takes place and to choose the path to the learning objectives within the learning environment.

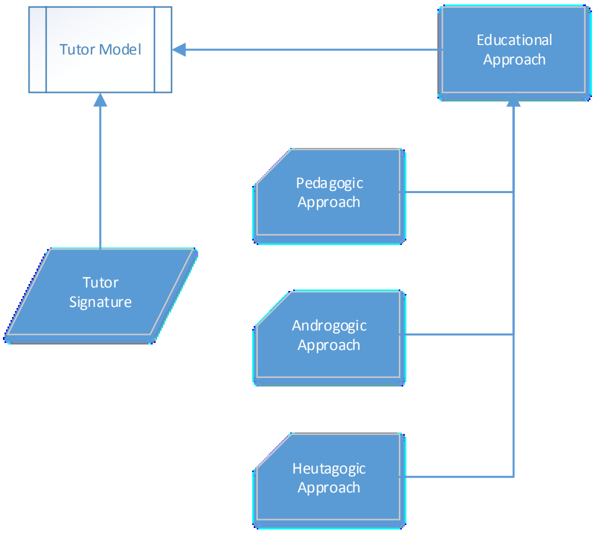


Figure 3 The Tutor Model

The student has the autonomy to choose not only content of the learning but also the order and format of the learning too. the way these ideas are incorporated into the adaptive model is to give the student a high degree of autonomy. This is at the discretion of the tutor who determines the amount of self-learning that would benefit a particular student. At the lowest level (pedagogical), the student has no say in what is learned. In the next level (andragogical), the student has the autonomy to choose which area to study next. In the final level (heutagogical) not only content is chosen but the form of learning object is chosen too.

The tutor model must contain mechanism then to 1. Determine the order in which the content is delivered to the student and it what format and 2. The degree of autonomy allowed to the student in choosing the learning direction. Different tutors may arrive at different assessments of students’ needs and different directions through the knowledge map.

The tutor signature is summarised in in the tutor data model Table 3 which indicates the basic parameters which define each individual tutor and their style of teaching.

TABLE 3 TUTOR DATA MODEL

Date element	**Data Type**
Present Knowledge Status	Type
Pedagogy	Number
Androgogy	Text
Heutagogy	Text
Knowledge presentation order	Free
Tutor Number ID	Text

V. The Knowledge Model

The curriculum to be delivered is to be stored in the adaptive system. The curriculum comprises three parts. First the structure of the knowledge shown in Figure 4 and its related parts which is contained in an ontology. Second the content of the learning which is the knowledge to be learned. Third, the different containers which hold the content. This is the form in which the knowledge is supplied and may be in text form, audio, video, PowerPoint, etc. The same knowledge may be presented in different formats to suit the student. In addition there is a requirement for test questions related to the curriculum.

It is necessary to define ontology metrics to provide measures of attributes such as complexity, level of detail or closeness of subject areas. The first step to defining these metrics is to provide each node with a unique address which defines its location on the ordered tree. Thus a body of knowledge is divided into section, sub-section, sub-sub-section etc. and so we adopt an addressing system which corresponds to this knowledge hierarchy where each address is correspondingly specified by sections, sub-sections, sub-sub-sections etc.

We use an ordered tree for this description where the branches from each node are ordered so that the sub-nodes have an order of preference. [8] This structure is then used to label an ontology where fragments of knowledge have an order determined by their pre-requisites. This model distinguishes between a taxonomy, ontology and what we call an anthology.

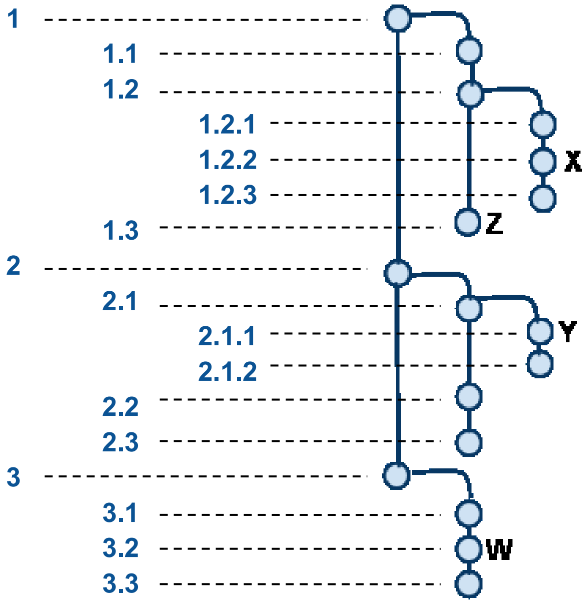


Figure 4: Knowledge hierarchy corresponding to an ordered tree

A. *Taxonomy, Ontology and Anthology*

Taxonomies specify the hierarchical relationships between concepts. Ontologies add to this attributes, properties and methods of the concepts. Anthologies take this and add to it the content of the information that the concepts specify. Ontologies are a way of sharing a common understanding of the structure of information. What Anthologies add to this is the content of the information itself.

B. *Anthology Formats*

We define Anthologies to be collections of information arranged in a hierarchical order. Where the taxonomy may be likened to the contents page of a book, the ontology is a detailed breakdown of the contents and the anthology would be likened to the whole book itself. The anthology should be understood as also containing the information for each section along with the headings. The data in each section can take the form of text, as may be found in a textbook, or a media file, video presentation etc. where the content that is stored is useful for teaching purposes. Thus we see taxonomies as a subset of ontologies and ontologies a subset of anthologies.

C. *Test Questions*

Test questions are used to test the students' knowledge of the content of a section of the curriculum. That section may be based on the course or lesson level. There may be many questions and many answers for each subject area using a Multiple Choice Question format. A dynamic set of questions is formed to become a student specific test for progress in a lesson or course. Questions and answers are determined and designed by the tutor.

TABLE 4 DATA MODEL FOR THE KNOWLEDGE SYSTEM

Data Element	**Data Type**
Node address	Number
Title	Text
Subtitle	Text
Content/Link	Free
Format	Video/audio/text/PP/other
Questions	Text
Tutor ID	Text

However it should be noted that this data model requires that the knowledge tree (or subject ontology) is contained within a relational database structure along the content-backbone where a unit is part of a course, and a lecture part of a unit and a segment part of a lecture. Segments can include learning objects.

COURSE-UNIT-LECTURE-SEGMENT

D. *Linking of Learning Objects*

Breaking up knowledge into learning objects based on the content structure highlights the importance of two aspects of the presentation of materials. Boyle [4], describes the learning object as a wrapper around content. The wrapper describes the structure of the object and includes the metadata about the object. The learning object is packaged in a standard container format which can be stored in a database. The included metadata permits fast effective searches to retrieve learning objects suitable for a particular purpose. Other data elements associated with the knowledge system are as follows.

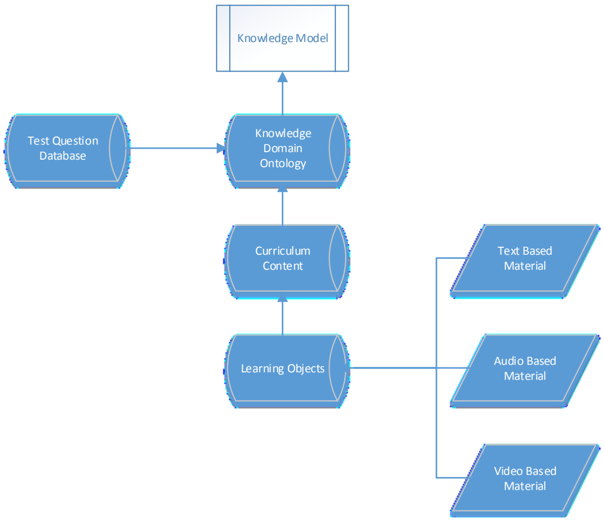


Figure 5. The Knowledge Model

E. *Segment*

A segment is defined as a learning node together with all its sub-nodes. The total number of nodes in a segment is a measure of the amount of detail contained within a segment of knowledge and can be associated with a node in the subject ontology.

ISBN: 978-1-61208-400-8

F. Complexity

We define complexity of a knowledge node to be equal to the degree centrality minus 1which is the measure of the number of sub-nodes that are connected to a given node. Thus a knowledge node composed of many sub-nodes or subdivisions is deemed to be more complex than one with fewer subdivisions and is defined as a measure of **difficulty** of the knowledge node.

G. Level

We designate the term level applied to each node by the position it occupies in the representation. We say that the **level** of a knowledge node is equal to its **importance** and represents the level of detail that a knowledge node contains.

H. Distance

The distance or separation of one node from another is a measure of how close two knowledge segments are related to the subject ontology. For a tree network this is a unique value determined by the number of steps between the nodes. Distance is a measure of the **strength of connection** between two nodes. The knowledge model structure is seen in Figure 5.

VI. The Adaptive Engine

The purpose of the adaptive engine is to choose the next node of learning for the student and the way it is presented. The way the adaptive engine works is by using the student signature and the tutor model to determine the next learning object, present it to the student in the appropriate form and to test its effectiveness. This is performed by reference to metrics attributed to the student signature with direction indicated by the tutor model. The student is guided to the next knowledge node on the subject ontology and is provide with subject content in a form which is most suitable to the individual student. The student signature will contain a measure of the prior knowledge of the student to enable adaption of content, form, independence of choice, test questions etc.

If pre-assessment shows that a degree of independent learning is appropriate for the student then a range of choices will be available to the student for them to make a choice themselves as the direction they can go in their learning within bounds set by the tutor. The adaptive engine is shown in Figure 6.

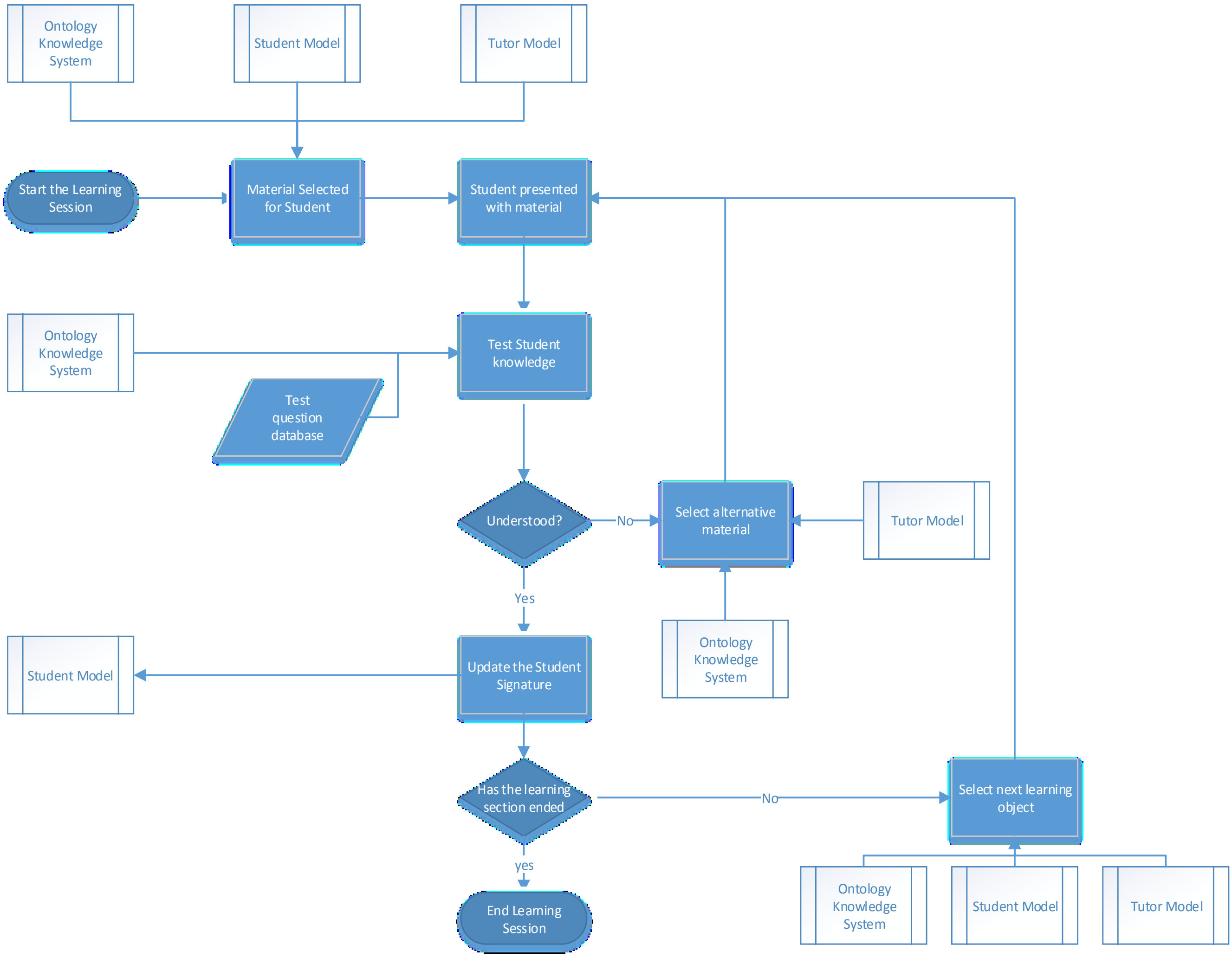

Ontology Knowledge System
Student Model
Tutor Model
Start the Learning Session
Material Selected for Student
Student presented with material
Ontology Knowledge System
Test Student knowledge
Test question database
Understood?
No
Select alternative material
Tutor Model
Yes
Ontology Knowledge System
Student Model
Update the Student Signature
Has the learning section ended
No
Select next learning object
yes
End Learning Session
Ontology Knowledge System
Student Model
Tutor Model

Figure 6 The Adaptive Engine

The adaptive engine will use the student signature to determine what has already been learned and what is still left to learn, It will use the tutor model to determine which elements need to be presented to the student to study next. We expect the segment entity to hold such attributes as **Level** (a measure of the importance of the segment) and **Complexity** (a measure of the difficulty of a knowledge node) as well as **Strength** of nodal links (a measure of the ontological proximity of the knowledge areas). Figure 1 depicts the rudimentary model of the Adaptive Learning System. Each of these three metrics are determined through an ontology calculus discussed in a previous paper. [2]

VII. Conclusion And Future Work

Investigations into semantic models and semantic modelling should be strictly logical explorations into how data models and integrity constraints can be modified without rendering the database contents (facts, meanings, and intelligent interpretations) uncertain or meaningless.

Meta-learning by the Adaptive Learning System requires awareness that it is participating in a learning process and therefore needs an explicit, built in ‘tutor model’. The Adaptive Learning System presented here implicitly assumes there is a real-life tutor who will perform the role of the tutor model, which involves intelligent and experienced selection of learning objects appropriate to the student.

In future, we need to construct a full, robust tutor model to automate the segmentation process, which needs detailed investigation of the nature of meta-learning [14] [15]. Our vision is to build this into a novel abstract conceptual data model encompassing all the properties that are needed to make explicit the qualities of an effective ‘tutor model’.

Finally, although work discussed in this paper answered research questions posed in previous papers, it has indicated further questions. In particular we would ask what further adaptation features are required and how are they to be evaluated? Also we need to further consider how should the adaptive engine structure be modelled and evaluated? Can fuzzy logic or data mining techniques be candidates for a useful algorithm? And finally we continue to explore how we determine the appropriate definition of an API, possibly by means of an IDL, between the ontology, the adaptation engine and the system’s user interface? We leave these questions to further papers.

References

[1] Cutts, S., Davies, P., Newell, D. and Rowe, N., 2009. *Requirements for an Adaptive Multimedia Presentation System with Contextual Supplemental Support Media*, Proceedings of the MMEDIA 2009 Conference, Colmar, France.

[2] Rowe, N., Cutts, S., Davies, P., and Newell, D. 2010 *Implementation and Evaluation of an Adaptive Multimedia Presentation System (AMPS) with Contextual Supplemental Support Media.* Proceedings of the MMEDIA 2010 Conference, Athens, Greece.

[3] IEEE. 2001. *IEEE Learning Technology Standards Committee* (LTSC) IEEE P1484.12 Learning Object Metadata Working Group; WG12 Home page.

[4] Boyle, T., 2003. Design Principles for Authoring Dynamic, Reusable Learning Objects. *Australian Journal of Educational Technology*.

[5] McGreal, R. (Ed.), 2004. *Online Education Using Learning Objects*. London:Routledge, 59-70.

[6] Protégé (2009) Protégé Ontology Editor, Stanford University California, USA. http://protege.stanford.edu/ [Accessed online 28 January 2010]

[7] Gruber, T., “A Translation Approach to Portable Ontology Specifications”, Knowledge Acquisition, 5(2), 199-220, 1993.

[8] Newman, M., E., J., “Networks, An Introduction”, Oxford University Press, 2010

[9] Codd, E.(1970). ‘Data Models in Database Management, ’ACM SIGMOD Record 11, No. 2

[10] Date C.J. (2000). ‘WHAT not HOW: The Business Rules Approach to Application Development’ Addison-Wesley. And Date, C. (2004). ‘Introduction to Database Systems’, 8th Ed., Pearson.

[11] Progress (2010) Objectstore, http://documentation.progress.com/output/ostore/7.2.0/pdf/user1/basicug.pdf (Last Accessed Dec 2010)

[12] Lamb, Charles, Landis, Gordon, Orenstein, Jack, Weinreb, Dan., (1991). ‘The Objectstore Database System’, *Communications of the ACM* 34 (10): 50–63.

[13] Date, C., Darwen, H. & Mcgoveran, D. (1998). ‘Relational Database Writings 1994-1997’, Addison Wesley.

[14] Chen, P. ‘The Entity-Relationship Model-Toward a Unified View of Data’ (1976), *ACM Transactions on Database Systems* 1/1/1976, ACM-Press.

[15] Chen, P. (2007). ‘Active Conceptual Modeling of Learning: Next Generation Learning-Base System Development’, with Leah Y. Wong (Eds.). Springer.

Empowering Semantic Indexing with Focus of Attention

Kimiaki Shirahama*, Tadashi Matsumura†, Marcin Grzegorzek*, and Kuniaki Uehara†

* Pattern Recognition Group, University of Siegen
† Graduate School of System Informatics, Kobe University
Email: kimiaki.shirahama@uni-siegen.de, tadashi@ai.cs.kobe-u.ac.jp,
marcin.grzegorzek@uni-siegen.de, uehara@kobe-u.ac.jp

***Abstract*—This paper addresses Semantic INdexing (SIN) to detect concepts like *Person* and *Car* in video shots. One main obstacle is the abundant information contained in a shot where multiple concepts are displayed at the same time. In other words, the detection of a target concept is adversely affected by other concepts which are incidentally shown in the same shot. We assume that a user can recognise the target concept when it appears in a salient region which attracts his/her attention. Based on this, we introduce a SIN method which utilises Focus of Attention (FoA) to extract a salient region in a shot, and constructs a feature emphasising that salient region. In addition, we develop a Weakly Supervised Learning (WSL) method to efficiently create training shots for FoA, and a shot filtering method to examine the usefulness of salient regions. Experimental results show the effectiveness of our SIN method using FoA.**

***Keywords*–Semantic indexing; Focus of attention; Weakly supervised learning; Usefulness of salient regions.**

I. Introduction

For effective filtering, categorisation, browsing and retrieval of large-scale video data, one key technology is ‘Semantic INdexing’ (SIN) to detect human-perceivable concepts (e.g., *Person*, *Car* and *Building*) in shots [1]. SIN is formulated as a binary classification problem where shots displaying a target concept are distinguished from the rest of the shots. The currently most popular approach is to represent a shot as a collection of *descriptors* which characterise patches (small regions) in the shot [2][3]. To make the following discussion clear, we define a descriptor as the representation of a patch, and a feature as the representation of a shot based on a set of descriptors. The effectiveness of this feature is attributed to considering many descriptors extracted from various patches. Even if the target concept is partially invisible due to the occlusion by other concepts or the camera setting, the feature includes descriptors extracted from patches corresponding to the visible part of the target. However, many concepts other than the target are displayed in a shot. For example, the top-left shot in Figure 1 includes the target concept *Car* and many others like *Building*, *Road* and *Sky*. Nonetheless, most of the existing methods [2][3] do not consider whether each patch belongs to the target concept or not. The resulting feature is affected by patches of other concepts, and the detection performance of the target concept is degraded.

To effectively spotlight a target concept, we propose a SIN method using ‘Focus of Attention’ (FoA). FoA implements selective attention that is a brain mechanism to determine which region in a video frame (or image) attracts the user [4]. Such an attractive region is called a *salient region*. FoA is beneficial to develop a system which can sort out visual information according to human perception. In our case, we assume that the target concept can be recognised by a user when it appears in salient regions. In other words, the user is unlikely to realise appearances of the target concept in non-salient regions. These appearances are trivial and useless for subsequent processes like video categorisation, browsing and retrieval. Therefore, we use FoA to increase priorities of salient regions and decrease those of non-salient regions. This prioritisation enables us to construct a feature which emphasises an appearance of the target concept, so that its detection performance can be improved.

FoA consists of two main processes, *bottom-up* and *top-down*. The former implements human attention driven by stimuli acquired from the external environment, where salient regions are detected based only on features. However, these salient regions are not so accurate because of the *semantic gap*, which is the lack of agreement between automatically extractable features and human-perceived semantics [5]. Thus, the top-down process implements attention which is driven by prior knowledge and expectation in the internal human mind. This is typically formulated in the framework of machine learning, where salient regions in test shots are detected by referring to training shots in which salient regions are annotated in advance. More concretely, inaccurate salient regions obtained by the bottom-up process are refined based on salient regions in training shots.

To incorporate FoA into SIN, we address the following two issues: The first is that salient regions significantly vary depending on camera techniques and shooting environments. A large number of training shots is needed to accurately detect diverse salient regions. However, due to a tremendous number of video frames in shots, it requires prohibitive cost to manually prepare many training shots. Thus, we develop an FoA method using ‘Weakly Supervised Learning’ (WSL) where a classifier to predict precise labels is constructed only using loosely labelled training data [6]. In our case, this kind of training data are shots that are annotated only with the presence or absence of a target concept. Using these training shots, we build a classifier which can identify the region of the target concept in a shot. In the top-down process, regions of the target concept in training shots are identified by the classifier and regarded as annotated salient regions.

The second issue is the discrepancy that salient regions do not necessarily contain a target concept. The reason is twofold: Firstly, there is the difficulty of objectively judging whether the target concept appears in a salient region or not. In other words, we can only use training shots where the presence of the target concept is annotated without considering its saliency.

For example, in Figure 1, training shots annotated with the presence of the target concept *Car* include the bottom-right shot where the car is shown in the small background region. It is impossible or unreasonable to regard this region as salient. The second reason for the discrepancy is possibly occurring errors in FoA. Even if the region of the target concept is salient for humans, another region may be falsely regarded as salient. A feature based on such a salient region which is discrepant with the region of the target concept incorrectly emphasises a non-target concept. To alleviate this, we develop a method which filters out shots where the target concept is unlikely to appear in salient regions, using regions predicted by the classifier in WSL. This enables SIN integrated with FoA to appropriately capture a characteristic feature for the target concept appearing in salient regions.

This paper is organised as follows: In Section II, we describe the novelties of our method by discussing insufficiencies of existing FoA methods with regard to SIN. Section III presents our method by sequentially explaining the FoA (consisting of the bottom-up and top-down processes), WSL and SIN modules. In Section IV, we evaluate our method using large-scale video data. Section V concludes this paper by providing a future extension of our method.

II. Related Work

Existing machine learning approaches for the top-down process are typically based on the *contextual cueing* which means "a user can easily search a particular object among many objects, if he/she saw the same or similar spatial layout of objects in the past". These approaches construct a model which properly integrates features extracted in the bottom-up process, using recorded eye-fixations or labelled salient regions as training data. Itti and Koch proposed an approach to compute the optimal weights to integrate features [7]. Kienzle *et al.* proposed a non-parametric approach to build a model using recorded eye-fixations [8]. Furthermore, Li *et al.* proposed an approach which simultaneously constructs a set of models to integrate features using multi-task learning [9]. However, these approaches assume the availability of training videos where salient regions are labelled. Compared to this, we incorporate WSL into FoA so as to only require training shots which are annotated just with the presence or absence of a target concept. In addition, the approaches described above are only evaluated on shots which necessarily contain some salient regions. In other words, they do not consider what is shown in a detected salient region, or how to use it in a subsequent application. This paper explores how to utilise salient regions for SIN and presents a method for filtering shots where salient regions show non-target concepts.

III. Concept Detection Using FoA

Figure 1 illustrates an overview of our SIN method where the target concept is *Car*. We call training shots annotated with the presence and absence of the target concept *positive shots* and *negative shots*, respectively. The bold arrows in Figure 1 show the dominant flow where the FoA module creates a *saliency map* for each training shot. This map is an image which represents the saliency of each pixel. Figure 1 shows saliency maps obtained for the positive and negative shots at the top. As pixels have higher saliencies, they are depicted as brighter. The positive and negative shots in Figure 1 are appropriately associated with salient regions where *Car* and *Person* are shown, respectively. The SIN module uses saliency maps to consider the saliency of each patch from which a descriptor is extracted. Thereby, the feature for each training shot is constructed by weighting descriptors based on saliencies of their patches. Then, a *detector* is constructed to identify the target concept in test shots.

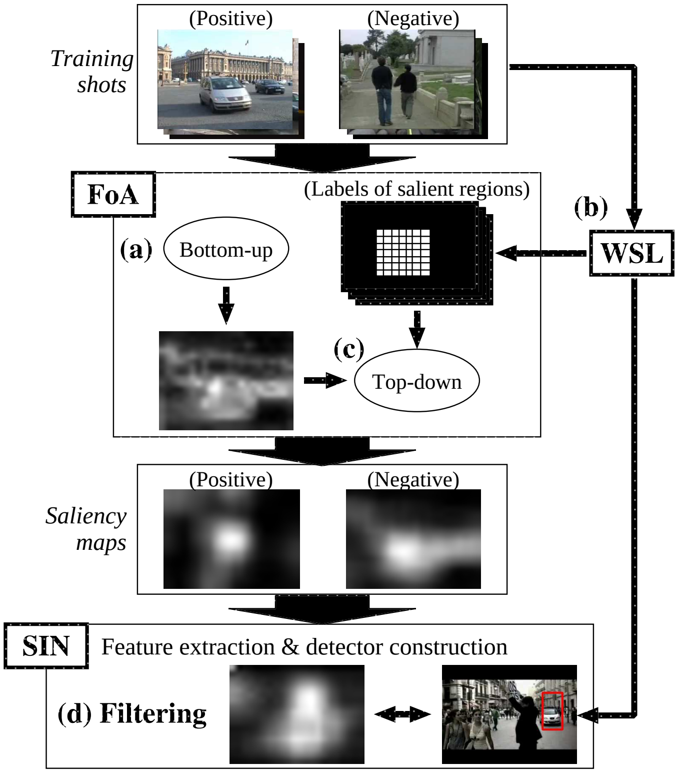


Figure 1. An overview of our SIN method using FoA with WSL.

The FoA module works as follows: First, the bottom-up process in Figure 1 (a) computes a saliency map for each training shot (for short, 'bottom-up saliency map'). Meanwhile, WSL in Figure 1 (b) builds a classifier using training shots, in order to identify regions of the target concept in positive shots. By regarding these regions as salient, the top-down process is performed as depicted in Figure 1 (c). This refines the bottom-up saliency map for each training shot into the final saliency map. Also, the filtering process in Figure 1 (d) examines whether each positive shot should be used to build a detector, by comparing its saliency map to the region detected by WSL (red rectangle). If the region is not salient, the positive shot is excluded from the training shots. This is because such positive shots mislead the detector to favour non-target concepts. Below, we describe the bottom-up and top-down processes, WSL method, and SIN method.

Bottom-up process: This simulates a retina model of a human to detect salient regions as the ones where features are different from those of surrounding regions [9]. For each video frame in a shot, the bottom-up process creates six 'feature maps', each of which describes every pixel using a different feature, that is, luminance, red-green contrast, blue-yellow contrast, flicker, motion direction, or motion strength [9]. Then, to simulate the mechanism of the horizontal cells of a retina, wavelet transform is performed per feature map so that three 'wavelet images' with different resolutions are created. Afterwards, the mechanism of the bipolar cells is simulated where high-pass filters are used to extract high-frequency components in three directions for each wavelet image. As a result, three 'edge images' are created where the difference of a region from surroundings is represented as edges. For noise reduction and computational efficiency, each edge image is smoothed by a Gaussian filter and divided into 18×22 macro-blocks, where each block is represented as the average of included pixel values. Finally, a bottom-up saliency map is created by averaging values of the same macro-block in all of 54 edge images (6 feature maps $\times$ 3 wavelet images $\times$ 3 edge images). The saliency of each macro-block is represented by a real number between 0 and 1 (see [9] for more detail).

Top-down process: Assuming that salient regions in positive shots are labelled by WSL described below, the top-down process performs *multi-task learning* which effectively solves multiple 'related tasks' at the same time by extracting the information shared among them [9]. We define a task as the refinement of the bottom-up saliency map of each positive shot. This task is related to tasks for other positive shots in the sense that they are taken by similar camera techniques and in similar shooting environments. Thus, we refine their bottom-up saliency maps using the same set of functions, which individually represent a different linear combination of features used in the bottom-up process. The bottom-up saliency map of each positive shot is refined by the weighted fusion of these functions' outputs, so that the refined saliency map matches the labelled salient region. We use an EM-like algorithm to optimise functions, and weights of their outputs for each positive shot [9].

Based on the contextual cueing described in Section II, the top-down process for a test shot begins with selecting the positive shot which has the most similar spatial layout to that of the test shot. The weighted fusion used for this positive shot is re-used to refine the bottom-up saliency map of the test shot. The similarity between two shots in terms of spatial layouts is computed as their cosine similarity based on features used in the bottom-up process.

Weakly Supervised Learning: Given training shots annotated only with the presence or absence of a target concept, we construct a classifier which can identify its region in a shot. This WSL is achieved so that the classifier characterises regions which are contained in positive shots, but are not contained in negative shots [6]. The target concept is considered to appear in these regions. The classifier is optimised by iterating the following two processes: The first examines regions in each training shot to find the 'best region' which maximises the output of the current classifier. The other process updates the classifier using the newly found best regions. As a result, the classifier outputs high values for the best regions in positive shots, while low values are assigned to all regions including the best ones in negative shots. The best regions in positive shots are regarded as the labelled salient regions.

Since a video frame in a shot contains a huge number of possible regions, the aforementioned WSL needs to efficiently find the best region. To this end, we implement the classifier as a linear SVM based on quantised SIFT descriptors, called 'Visual Words' (VWs) [10]. First, we extract SIFT descriptors from patches which have the radius of 10 pixels and are located at every sixth pixel in each video frame. Then, randomly sampled one million SIFT descriptors are grouped into $1,000$ clusters where each cluster centre is a VW. Every SIFT descriptor is quantised into the most similar VW. Based on this, a region is represented by a feature (histogram) where each dimension represents the frequency of a VW in this region. In particular, for any region, the output of the linear SVM can be computed by simply counting the frequency of each VW. This enables us to estimate the 'upper bound' for a set of regions [10]. Here, no region in the set takes the output larger than the upper bound. The best region can be efficiently found by discarding many sets of regions for which upper bounds are small. With this efficient search, we can refine the classifier (linear SVM) through many iterations.

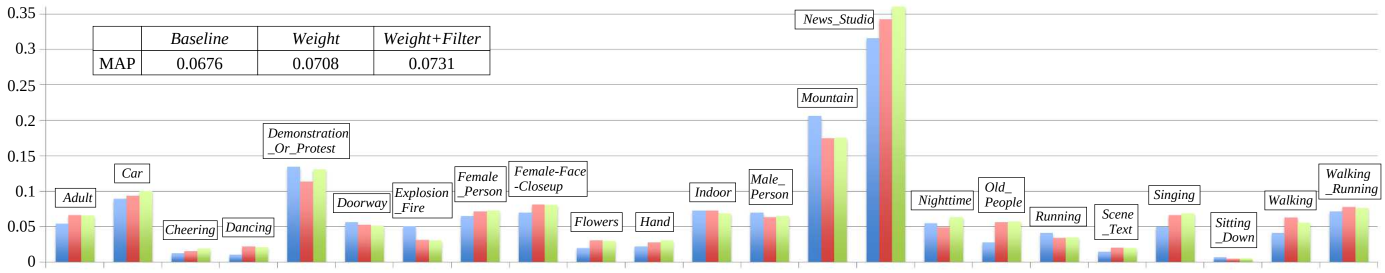
Semantic Indexing: Given a target concept, we create the feature of each training shot as a histogram where each bin represents the 'weighted' frequency of a VW. We check the saliency map to obtain the saliency of each patch from which a SIFT descriptor is extracted. This saliency is used to weight the frequency of the VW associated with the SIFT descriptor. As a result, the feature emphasises frequencies of VWs in the salient region where the target concept probably appears. Using such features, a detector is constructed as a non-linear SVM with Radial Basis Function (RBF) kernel.

Before constructing the detector, we use the classifier in WSL to examine whether salient regions in positive shots include the target concept or not. We assume that the target concept is salient if its region is large. Hence, we filter out a positive shot if the best region is very small, or the classifier's output for this region is very small. This filtering is also applied to test shots. For test shots where salient regions fail to cover the target concept, features obtained by weighting VWs' frequencies undesirably emphasise non-target concepts. To avoid this, the filtering aims to distinguish test shots where salient regions certainly include the target concept from the others. For the former, we extract features by weighting VWs' frequencies, while features for the latter are extracted without weighting. Finally, the list of sorted test shots in terms of outputs by the detector is returned as the SIN result.

IV. Experimental Results

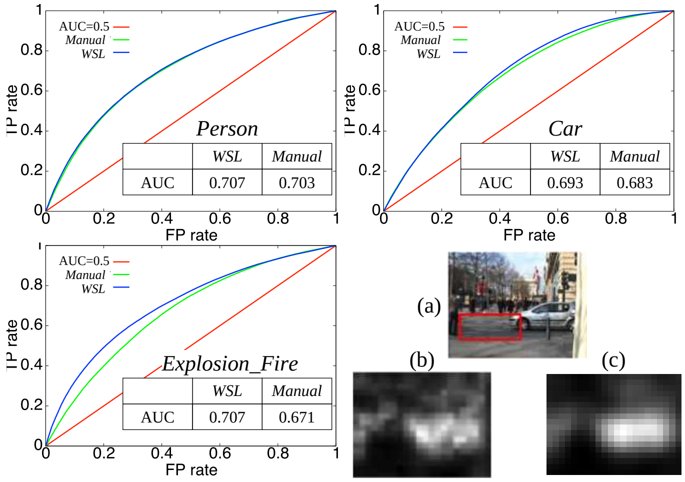
We first examine the effectiveness of FoA using WSL by targeting three concepts *Person*, *Car* and *Explosion_Fire*. For each concept, we use $1,000$ positive shots and $5,000$ negative shots in TRECVID 2009 video data [1]. The performance is evaluated on $1,000$ test shots where the ground truth of salient regions is manually provided. We compare two FoA methods, *WSL* and *Manual*, which use positive shots where salient regions are labelled by WSL and by the manual method, respectively. Figure 2 shows ROC curves for *WSL* and *Manual*. Each curve is created by calculating true positive (TP) and false positive (FP) rates using different thresholds. Here, a pixel in a saliency map is regarded as salient if its saliency is larger than a threshold. A TP is the number of pixels which are correctly detected as salient, and a FP is the number of pixels falsely detected as salient. Figure 2 shows that, for all concepts, ROC curves of *WSL* and *Manual* are nearly the same. This means that FoA can be appropriately performed even using salient regions labelled by WSL.

As another evaluation measure, an AUC represents the area under an ROC curve. A larger AUC indicates superior performance where a high TP is achieved for a small FP. Figure 2 presents that *WSL*'s AUCs are nearly the same or even larger than those of *Manual*. Note that several regions where a target concept does not appear are falsely detected by WSL, and used as labelled salient regions in the top-down



	Baseline	*Weight*	*Weight+Filter*
MAP	0.0676	0.0708	0.0731

Figure 3. Performance comparison among *Baseline*, *Weight* and *Weight+Filter*.



Person	*WSL*	*Manual*
AUC	0.707	0.703

Car	*WSL*	*Manual*
AUC	0.693	0.683

Explosion_Fire	*WSL*	*Manual*
AUC	0.707	0.671

Figure 2. Performance comparison between *WSL* and *Manual*.

process. For example, at the bottom-right of Figure 2, the red rectangular region in the image (a) is falsely regarded as showing a car. However, as seen from the bottom-up saliency map marked by (b), the saliency of this region is very low, so it cannot be a salient region even with the refinement by the top-down process (see the image (c)). Like this, errors in WSL are alleviated based on saliencies obtained by the bottom-up process. In other words, FoA works appropriately as long as regions obtained by WSL are mostly correct.

Next, we evaluate the performance of our SIN method using FoA. According to the official instruction of TRECVID 2011 SIN light task [1], we select 23 target concepts shown in Figure 3. For each target, a detector is constructed with $30,000$ training shots collected from $240,918$ shots in $11,485$ development videos, and tested on $125,880$ shots in $8,215$ test videos. To examine the effectiveness of weighting descriptors based on FoA and that of shot filtering, we compare three methods *Baseline*, *Weight* and *Weight+Filter*. *Baseline* and *Weight* use features defined by original frequencies and weighted frequencies of VWs, respectively. *Weight+Filter* extends *Weight* by adding the shot filtering process.

Figure 3 shows the performance comparison in form of a bar graph. For each concept, the left, centre and right bars represent Average Precisions (APs) of *Baseline*, *Weight* and *Weight+Filter*, respectively. A larger AP indicates a higher performance. For each method, we also exhibit the Mean AP (MAP) which is the mean of APs over 23 concepts. Figure 3 illustrates that *Weight* outperforms *Baseline* for many concepts. The MAP of the former is about 5% higher than that of the latter. This validates the effectiveness of FoA for SIN. In addition, *Weight+Filter*'s MAP indicates that adding shot filtering improves *Weight*'s MAP by about 3%. This verifies the effectiveness of shot filtering.

V. Conclusion and Future Work

In this paper, we introduced a SIN method using FoA with WSL. Experimental results showed both the validity of incorporating WSL into FoA and the effectiveness of FoA for SIN. Figure 3 shows that FoA causes the performance degradation for some concepts such as *Explosion_Fire* and *Mountain*. One main reason is non-rectangular shapes of these concepts, while our WSL method can only identify rectangular regions. In other words, rectangular regions are too coarse to precisely localise non-rectangular concepts, and inevitably include other concepts. As a result, the top-down process does not work well. Hence, we will extend our WSL method by adopting the efficient search algorithm for regions with arbitrary shapes [11].

Acknowledgment

This work was carried out when Tadashi Matsumura belonged to Graduate School of System Informatics, Kobe University.

References

[1] A. Smeaton, P. Over, and W. Kraaij, “Evaluation campaigns and trecvid,” in Proc. of MIR 2006, 2006, pp. 321–330.

[2] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek, “Evaluating color descriptors for object and scene recognition,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 32, no. 9, 2010, pp. 1582–1596.

[3] N. Inoue and K. Shinoda, “A fast and accurate video semantic-indexing system using fast map adaptation and gmm supervectors,” IEEE Trans. Multimed., vol. 14, no. 4, 2012, pp. 1196–1205.

[4] S. Frintrop, E. Rome, and H. Christensen, “Computational visual attention systems and their cognitive foundations: A survey,” ACM Trans. Appl. Percept., vol. 7, no. 1, 2010, pp. 6:1–6:39.

[5] A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, “Content-based image retrieval at the end of the early years,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 22, no. 12, 2000, pp. 1349–1380.

[6] M. Nguyen, L. Torresani, F. De la Torre, and C. Rother, “Weakly supervised discriminative localization and classification: a joint learning process,” in Proc. of ICCV 2009, 2009, pp. 1925–1932.

[7] L. Itti and C. Koch, “Feature combination strategies for saliency-based visual attention systems,” J. Electron. Imaging, vol. 10, no. 1, 2001, pp. 161–169.

[8] W. Kienzle, F. Wichmann, B. Schoelkopf, and M. Franz, “A nonparametric approach to bottom-up visual saliency,” in Proc. of NIPS 2006, 2006, pp. 689–696.

[9] J. Li, Y. Tian, T. Huang, and W. Gao, “Probabilistic multi-task learning for visual saliency estimation in video,” Int. J. Comput. Vis., vol. 90, no. 2, 2010, pp. 150–165.

[10] C. Lampert, M. Blaschko, and T. Hofmann, “Beyond sliding windows: Object localization by efficient subwindow search,” in Proc. CVPR 2008, 2008, pp. 1–8.

[11] S. Vijayanarasimhan and K. Grauman, “Efficient region search for object detection,” in Proc. of CVPR 2011, 2011, pp. 1401–1408.

Image Retrieval System Based on Combination of Color, Texture and Shape Features

Atoany N. Fierro-Radilla, Mariko Nakano-Miyatake,
Karina Perez-Daniel, Hector Perez-Meana
Escuela Superior de Ingeniería Mecánica y Eléctrica UC
Instituto Politécnico Nacional
Mexico City, Mexico
Email: mnakano@ipn.mx, afierror@hotmail.com.

Francisco Garcia-Ugalde, Manuel Cedillo-Hernandez
Facultad de Ingeniería, Universidad Nacional Autonoma de Mexico UNAM
Mexico City, Mexico
Email: fgugalde@gmail.com, mcedillohdz@hotmail.com

Abstract—**In the Content-Based Image Retrieval (CBIR) system, an effectiveness of the visual descriptors, such as color, texture and shape, determines a good retrieval performance. Recently, more than two types of visual descriptors are combined to improve the performance of CBIR. The combination method used to combine different types of visual descriptors also plays an important role to obtain a good performance. However, we are aware that researchers have not paid sufficient attention to the combination methods, so in this paper, we focused on combination of schemes of three types of visual features to obtain higher improvement of the retrieval performance. Firstly, several visual descriptors that belong to three types of visual features (color, texture and shaper) are analyzed individually to select a better descriptor from each category. Second, the several combination methods are analyzed to determine a best combination method. The performance of the proposed scheme is compared with some CBIR systems in which more than two different types of descriptors are combined.**

Keywords-CBIR system; color descriptor; texture descriptor; shape descriptor; combination methd of descriptors

I. INTRODUCTION

Multimedia data is very common in daily activities, because nowadays it is easy and inexpensive to take pictures and videos. Digital images and videos are not only important in common activities, but also in areas such as medicine, biology, astronomy, commerce, tourism, etc. Due to the importance of multimedia information, the facility of sharing these data trough high-speed Internet connections and the high storage capacities, the size of databases has been increasing considerably. As a consequence of this situation, an efficient classification, indexing and retrieval of digital images stored in a huge database have been challengeable tasks. Therefore, the Content-Based Image Retrieval (CBIR) has become an urgent research topic because the traditional retrieval method, that is manual and subjective process, has become time consuming operation with ambiguity results in a huge database.

The CBIR systems describe multimedia content using visual features, such as color, texture, shape, etc. These features are low-level visual features, which describe images making the information retrieval be fast, objective and automatic. In general, color is one of the most dominant and distinguishable features in describing image [1]. Therefore, until now, several color-based descriptors have been proposed in the literature [2]-[6]. The Histogram Intersection (HI) [2], the Color Correlogram (CC) [3], the Dominant Color Descriptor (DCD) [4], the Color Layout Descriptor (CLD) [5] and the Color Structure Descriptor (CSD) [6] are widely used as color-based descriptors in the CBIR. To improve the retrieval performance, Atoany et al. [7] proposed the Dominant Color Correlogram Descriptor (DCCD), which optimizes the CC using only eight dominant colors. Also, some texture-based descriptors have been proposed in the literature to describe the image using the texture patterns [1][8]-[12]. The steerable filters [1], The Edge Histogram Descriptor (EHD) [8], the Texture Browsing Descriptor [9], the co-occurrence matrix-based descriptors [10][11] and Local Binary Pattern (LBP) [12] are some of the most widely used texture-based descriptors. The shape feature is another important factor that can be used to identify objects and classify the image context. As the shape-based descriptors, the Fourier Descriptor (FD) [13], the moment-based descriptors, such as Pseudo-Zernike Moments (PZM) [1] and Polar Harmonic Transforms (PHT) [14], and Pyramidal Histogram of Oriented Gradients (PHOG) [15] have been used in the CBIR systems.

There are some algorithms that do not use the above mentioned visual features to characterize the images and retrieve desired images. The Scale-Invariant Feature Transform (SIFT) and Speed Up Robust Feature (SURF) are examples of these types of algorithm. In the CBIR system, the SIFT and the SURF obtain some robust interest points, and using these points together with their neighborhood regions, the relevant images are retrieved [16][17]. Recently, the learning-based approaches, such as bag-of-visual-word [18] and the deep learning [19] are used to solve the practical problem in the CBIR.

In order to improve the image retrieval performance, it is necessary to combine more than one visual feature. Several combination or fusion methods are proposed in the literature [20]-[23]. The simplest method is concatenation of two or three descriptors to generate one descriptor with large number of elements [20], while in [21], the descriptors related to color, texture and shape features are combined to generate a single completed binary region descriptor

(CBRD); in this case the combination is not simple concatenation. In another combination scheme, different types of descriptors are applied to image database in the cascade or the parallel manner [22][23]. In [22], firstly, color-based descriptor is applied to retrieve a sub-set of relevant images, and then, shape-based descriptor is applied to images belonging to the previously retrieved sub-set only. In the parallel structure of [23], several sub-sets of relevant images are extracted independently using different types of descriptor, and then, the decision stage makes a final set of the relevant images from all extracted sub-sets of images.

In [1][24]-[27], the visual features are lineally combined to improve the retrieval performance, in which the similarity of each feature is independently computed using a distance metric. And then, an adequate weight for each feature is defined to generate a weighted lineal combination of the similarity scores of the different features. This combination scheme provides us a construction of a flexible CBIR scheme depending on the application. For example in the medical image retrieval system, in which the texture feature may be more relevant than other features, the weight value assigned to texture feature can be more important than color and shape features. In [24] and [25], the color feature and the texture feature are combined. In [24], the CLD and the Texture Browsing Descriptor are used, while in [25], the EHD and several color-based descriptors, such as the CLD and the CSD, are used. In [26], the DCCD as color feature and the PHOG as shape feature are combined lineally. In [1] and [27], three features are used. In [1], the DCD, the steerable filter and the PZM are used as color, texture and shape features, respectively. The color histogram and moments as color feature, texture feature based on Gabor filter and the PZM as shape feature are combined lineally in [27].

In this paper, we propose a CBIR system, in which color, texture and shape-based descriptors are obtained and three distances between query image and database image using these three descriptors are calculated. And then, three distances are lineally combined using adequate weights. Firstly, we analyzed individually several visual descriptors, which belong to one of three types of descriptors in order to select the most adequate descriptor from each category. As color descriptor, we select the DCCD, which was proposed by Atoany et al. [7]. As texture-based descriptor, we selected the directional local motif XoR patterns (DLMXoRP) [11] and, finally, the shape feature is extracted using the PHOG descriptor [15]. Next, the weights for three features are adapted, varying their values, to improve the retrieval performance. The proposed scheme was evaluated using some common metrics used in the CBIR systems such as Average Normalized Modified Retrieval Rank (ANMRR), Average Retrieval Rate (ARR) and Average Retrieval Precision (ARP), and it was compared with some CBIR systems which combined two or more visual features.

The rest of this paper is organized as follows: In Section II, we explain the color-based descriptor used in our CBIR system, and its performance is compared with other color descriptors. In Section III, we present the texture-based descriptor used in our CBIR system together with the performance of this descriptor, and in Section IV, the selected shape-based descriptor and its performance are provided. In Section V, we provide an analysis of the combination methods of the selected three types of descriptors and the global performance of the proposed CBIR system. Finally, in Section VI, we conclude this work.

II. COLOR-BASED DESCRIPTOR

Color descriptors are divided into two categories, i) global color descriptors take into account the whole image in order to extract color information, this process does not include pre-processing or image segmentation; ii) local color descriptors extract spatial information on how pixels are distributed in certain region, and this is done using pre-processing or image segmentation [28]. Several color-based descriptors were proposed in literature, and some of them were adopted by the MPEG-7.

In the proposed CBIR system, we selected DCCD [7] due to its better performance and compact representation.

A. *Dominant Color Correlogram Descriptor (DCCD)*

First, the image is converted from RGB to HSV color space, because this color space presents more similarity to the human color perception. Then, the HSV image is quantized [4], in order to reduce the computational cost. This color quantization is done as:

$$H = \begin{cases} 0\; if\; h & \in [316,20) \\ 1\; if\; h & \in [20,40) \\ 2\; if\; h & \in [40,75) \\ 3\; if\; h & \in [75,155) \\ 4\; if\; h & \in [155,190) \\ 5\; if\; h & \in [190,270) \\ 6\; if\; h & \in [270,295) \\ 7\; if\; h & \in [295,316) \end{cases} \qquad (1)$$

$$S = \begin{cases} 0, & \text{if} \quad s \in [0,0.2] \\ 1, & \text{if} \quad s \in (0.2,0.7] \\ 2, & \text{if} \quad s \in (0.7,1] \end{cases} \qquad V = \begin{cases} 0, & \text{if} \quad v \in [0,0.2] \\ 1, & \text{if} \quad v \in (0.2,0.7] \\ 2, & \text{if} \quad v \in (0.7,1] \end{cases} \qquad (2)$$

We only consider the 8 more representative hues (red, orange, yellow, green, blue, dark blue, purple, violet), and three levels for saturation (S) and value (V). It is important to mention that Human Visual System (HVS) is irregular, that is the reason why we are using this method of color quantification. The dominant colors are determined from the quantized image with 72 colors, which is given by:

$$F = \{\{c_i, P_i\}, i = 1, \ldots, M\} \qquad (3)$$

where M<72 is the numbers of dominant colors of a quantized image, c_i is *i-th* dominant color with three components (H,S,V) and P_i is the percentage of the dominant color c_i. Firstly, the percentages Q_j, j=0,..., 71, of all existent colors are calculated, and then, M colors c_i, *i*=0,…, *M*-1 with the first *M* largest percentage are extracted as dominant colors, in this paper, we use *M*=8 dominant colors. Then, the percentages of each dominant color c_i is adjusted as

$$P_i = \frac{\overline{p_i}}{\sum_{j=0}^{M-1} Q_j} \qquad (4)$$

$$i = 0,1,\dots,M-1, \quad j = 0,1,\dots,71$$

where $\bar{p}_i = Q_i$ if c_i is a dominant color, otherwise $\overline{p_i} = 0$. Once dominant colors are obtained, the correlation of pair pixels of the same dominant color is calculated using color correlogram [3], and it is defines as:

$$\gamma_{c_ic_i}(I) \triangleq Pr_{p_1 \in I_{c_i}, p_2 \in I_{c_i}}[p_2 \in I_{c_i} ||p_1 - p_2| = 1] \qquad (5)$$

where $\gamma_{c_ic_i}(I)$, is the probability of finding a pixel p_1 of color c_i away from another pixel p_2 of the same color. Obtaining this correlogram for all dominant color c_i with $i = 0, 1, \dots, M-1$, we get the DCCD, which is given by

$$DCCD = \{c_i, CC_i\} \qquad (6)$$

where CC_i is the color correlogram of *i-th* dominant color c_i.

B. *Performace comparison of color-based descriptors*

The performance of the DCCD is compared with other conventional color-based descriptors. Table I shows experimental results of the DCCD together with six color-based descriptors using the *Database 2,* composed by 1000 Corel images, divided into 10 categories with 100 ground truth images per category. From the Table I, the DCCD and the conventional CC descriptors show better performance, although DCCD is 8 times more compact than the CC [3].

TABLE I. COMPARISON RESULTS OF TEXTURE-BASED DESCRIPTORS

Method	**ANMRR**	**ARR α 2**	**ARR & ARP α 1**	**ARR α $\frac{1}{2}$**	**ARP α $\frac{1}{4}$**
DCCD	**0.3086**	0.7590	0.**5960**	0.7560	0.8840
CC	0.3228	0.7200	0.5870	**0.7620**	**0.8880**
IH	0.3174	0.7610	0.5760	0.7380	0.8640
DCD	0.3384	0.7420	0.5590	0.6920	0.8480
LBA	0.3478	0.7320	0.5500	0.7040	0.8000
CLD	0.3194	**0.7620**	0.5740	0.7280	0.8360
CSD	0.4431	0.6190	0.4630	0.6200	0.7680

Taking into account the good performance and compactness of the DCCD, we select it as color-based descriptor.

III. TEXTURE-BASED DESCRIPTOR

Texture is an important property for characterization and recognition of image [8][10]. We analyzed the performance of several texture-based descriptors in order to select the most efficient one.

A. *Directional Local Motif Xor Pattern (DLMXoRP)*

The DLMXoRP [11] is one of the high-performance texture-based descriptors, in which an input image is divided into 3x3 overlapping blocks and a vector at a specific direction is extracted as shown by Figure 1.

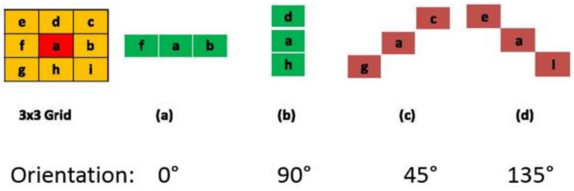


Figure 1. 3-element vector extraction [11]

In Figure 2 we can observe that a number of motif (1,2…,7) is assigned depending on the relation between the three pixel values of the extracted vector as follows [11]:

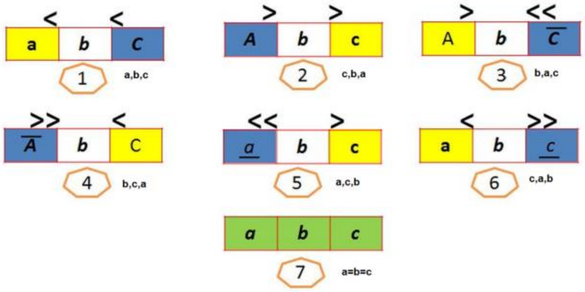


Figure 2. Seven motif asigment [11]

In order to extract the texture features, the following equations are used

$$DLMXoR_{N,R} = \sum_{i=0}^{N-1} T_{im}(p_i, p_c) \times 2^i \qquad (7)$$

where:

$$T_{im} = \begin{cases} 1 & p_i \neq p_c \\ 0 & p_i = p_c \end{cases} \qquad (8)$$

And *N* is the number of neighbor pixels, *R* is the radio of the 3x3 overlapping block, p_c is the central pixel and p_i , *i*=0,.. 7, are the neighbor pixels. Using this information, a histogram is computed by:

$$H_{DLMXoRp}^{\theta}(l) = \sum_{i=1}^{N_1}\sum_{j=1}^{N_2} f_2(DLMXoRP^{\theta}(i,j), l) \qquad (9)$$

$$l \in [0, 2^N - 1]$$

where $\theta = [0°, 45°, 90°, 135°]$ and:

$$f_2(x, y) = \begin{cases} 1 & x = y \\ 0 & x \neq y \end{cases} \qquad (10)$$

Finally, the obtained histograms are concatenated as:

$$H_{DLMXoRP} = [H_{DLMXoRP}^{0°}, H_{DLMXoRP}^{45°}, H_{DLMXoRP}^{90°}, H_{DLMXoRP}^{135°}] \qquad (11)$$

B. *Performance comparison of texture-based descriptors*

To select the most efficient texture-based descriptor, we carried out a performance comparison of several texture-based descriptors, such as MCM [10] and DLMXoRP [11]. Vipparthi and Kagar [11] compared their DLMXoRP with several LBP-based descriptors, showing superiority of this descriptor. We also compared Steerable Filters [1] and the EHD [8]. In this evaluation, we used some metrics commonly used in image retrieval evaluation, such as ANMMR, ARR and ARP, using the Corel Dataset 1k. The evaluation results of four texture-based descriptors are shown in the Table II.

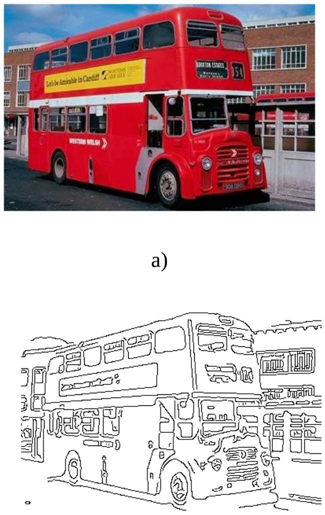
TABLE II. COMPARISON RESULTS OF TEXTURE-BASED DESCRIPTORS

Metric	Steerable Filter	MCM	DLMXoRP	EHD
ANMRR	0.5144	0.5404	**0.4460**	0.5401
ARR $\alpha = 2$	0.5820	0.5410	**0.6140**	0.5550
ARR $\alpha = 1$	0.4010	0.3810	0.**4680**	0.3790
ARP $\alpha = 1$	0.4010	0.3810	**0.4680**	0.3790
ARP $\alpha = 0.5$	0.4620	0.4660	**0.5960**	0.4400
ARP $\alpha = 0.25$	0.5120	0.5560	**0.6720**	0.5000

As shown in the Table II, the DLMXoRP provides a better performance in the CBIR task, therefore we selected this descriptor as the texture-based descriptor in our proposed CBIR system.

IV. SHAPE-BASED DESCRIPTOR

Shape is known to play an important role in human recognition and perception, providing a powerful clue to object identity [1].


a)

b)

Figure 3. Edge information extraction. a) RGB image, b)edge information

Shape-based descriptors can be categorized into two classes [13]: i) contour-based descriptors, which use the boundary information only, ignoring important information in the interior of the objects, ii) region-based descriptors, which use both, boundary and the interior information of objects.

A. *Pyramid Histogram of Oriented Gradients (PHOG)*

In the proposed CBIR system, we selected the PHOG descriptor [15], which extracts boundary information from the object.

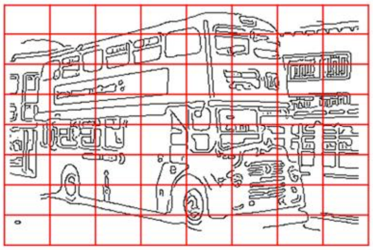


Figure 4. Segmentation of the edge image

This descriptor extracts the edges from a gray-scale image using Canny edge detector, as shown in Figure 3. Then, the image is divided into several blocks (Figure 4) in hierarchical manner to generate several pyramid levels, and in each pyramid level, a histogram of oriented gradients is computed. Finally, the PHOG descriptor is obtained concatenating all these histograms [15].

B. *Performance comparison of shape-based descriptors*

To select the most powerful shape-based descriptor, some conventional shape-based descriptors, such as PZM

[1], two representations of the PHT, which are Polar Complex Exponential Transform (PCET) and Polar Cosine Transform (PCT) [14], and the PHOG, are evaluated. The comparison results are shown by Table III.

TABLE III. COMPARISON RESULTS OF SHAPE-BASED DESCRIPTORS

Metric	PZM	PCET	PCT	PHOG
ANMRR	0.7177	0.7212	0.7066	**0.5825**
ARR $\alpha = 2$	0.3808	0.3642	0.3554	**0.4462**
ARR $\alpha = 1$	0.2000	0.2231	0.2423	**0.3538**
ARP $\alpha = 1$	0.2000	0.2331	0.2423	**0.3538**
ARP $\alpha = 0.5$	0.2308	0.2615	0.2538	**0.4692**
ARP $\alpha = 0.25$	0.3692	0.3692	0.3692	**0.6308**

From the Table III, we can observe that the PHOG outperforms other shape-based descriptors, so we decided to incorporate it as the shape-based descriptor in our CBIR system.

V. EXPERIMENTAL RESULTS

The most adequate three descriptors, which are the DCCD as color-based descriptor, the DLMXoRP as texture-based descriptor and the PHOG as shape-based descriptor, were selected through the performance comparison in the CBIR system, we analyzed several combination methods of these three descriptors. The combination of the three visual descriptors is done using a weighted linear combination given by

$$S(I,Q) = \omega_c S_{color}(I,Q) + \omega_t S_{tex}(I,Q) + \omega_s S_{shape}(I,Q) \qquad (12)$$

where I and Q are an image extracted from database and a given query image, respectively, and $S_{color}(I,Q)$, $S_{tex}(I,Q)$ and $S_{shape}(I,Q)$ are individual scores of image I respect to the query image Q in color, texture and shape aspects, respectively, and $S(I,Q)$ is the global score of I respect to Q. The weight values: ω_c, ω_t and ω_s present the grades of importance of each visual feature, and $\omega_c + \omega_t + \omega_s = 1$ must be satisfied.

To determine the most adequate combination of three weight values, the performance of proposed CBIR system is evaluated varying these three values using Corel Dataset 1K, which are shown by Table IV. In Figure 5, we present the image retrieval behavior for each class at a specific weight combination. From Table IV and Figure 5, we can observe that using combination number 7, which presents $\omega_C = 0.2, \omega_T = 0.3, \omega_S = 0.5$, the performance of the image retrieval task is improved.

TABLE IV. CBIR PERFORMANCE WITH DIFFERENT COMBINATIONS OF THREE WEIGHT VALUES

Weights	ANMRR	ARR α 2	ARR & ARP α 1	ARR α $\frac{1}{2}$	ARP α $\frac{1}{4}$
$\omega_C = 0.3$ $\omega_T = 0.3$ $\omega_S = 0.3$	0.3387	0.7520	0.5610	0.6980	0.7880
$\omega_C = 0.5$ $\omega_T = 0.3$ $\omega_S = 0.2$	0.3510	0.7390	0.5470	0.6820	0.7600
$\omega_C = 0.5$ $\omega_T = 0.2$ $\omega_S = 0.3$	0.3584	0.7290	0.5410	0.6700	0.7520
$\omega_C = 0.2$ $\omega_T = 0.5$ $\omega_S = 0.3$	0.3353	0.7360	0.5660	**0.7200**	**0.8200**
$\omega_C = 0.3$ $\omega_T = 0.5$ $\omega_S = 0.2$	0.3346	0.7390	0.5670	0.7040	0.8120
$\omega_C = 0.3$ $\omega_T = 0.2$ $\omega_S = 0.5$	0.3421	0.7550	0.5570	0.6780	0.7800
$\omega_C = 0.2$ $\omega_T = 0.3$ $\omega_S = 0.5$	**0.3306**	**0.7600**	**0.5760**	0.7060	0.8080
$\omega_C = 0.4$ $\omega_T = 0.4$ $\omega_S = 0.2$	0.3413	0.7490	0.5570	0.6960	0.7760
$\omega_C = 0.2$ $\omega_T = 0.4$ $\omega_S = 0.4$	0.3321	0.7410	0.5690	0.7180	0.8160
$\omega_C = 0.4$ $\omega_T = 0.2$ $\omega_S = 0.4$	0.3511	0.7390	0.5490	0.6840	0.7600

The proposed CBIR system with optimum weight values, obtained through above mentioned observation, is evaluated comparing with other CBIR schemes [1][26]. Both CBIR schemes use more than two visual descriptors, in [1], the DCD based on LBA algorithm is used as color-based descriptor, the PZM and Steerable filter are used as shape-based descriptor and texture-based descriptor, respectively. While in [26], the DCCD and the PHOG are used as color-based and shape-based descriptors, respectively. The comparison results are shown by Table V.

TABLE V. PERFORMANCE COMPARISON

Metric	LBA+PZM + Steerable Filters [1]	DCCD +PHOG [26]	Proposed
ANMRR	0.3672	0.2698	**0.1821**
ARR $\alpha = 2$	0.6750	0.7800	**0.8560**
ARR $\alpha = 1$	0.5420	0.6550	**0.7370**
ARP $\alpha = 1$	0.5420	0.6550	**0.7370**
ARP $\alpha = 0.5$	0.7020	0.8120	**0.8940**
ARP $\alpha = 0.25$	0.8400	0.9320	**0.9577**

The comparisons show that the proposed CBIR scheme, which combines color, texture and shape based descriptors with optimum weight values, provides much better performance compared with other CBIR systems previously proposed.

VI. Conclusions

In this paper, we analyzed several visual descriptors that belong to color-based, texture-based and shape-based descriptors. Through the performance analysis of color-based descriptors in the CBIR task, we determined that the DCCD is the most efficient color-based descriptor from its retrieval performance and compact representation. As texture-based descriptor, we selected the DLMXoRP considering its higher performance compared with the conventional texture-based descriptors, while the PHOG descriptor shows much better performance compared with other shape-based descriptors, such as the PZM and the PHT descriptors, so this descriptor is selected as shape-based descriptor in the proposed CBIR.

The three descriptors are combined lineally, and the weight values are determined after exhaustive evaluations of proposed CBIR system using Corel Dataset 1k. The determined weight values are $\omega_C = 0.2, \omega_T = 0.3, \omega_S = 0.5$, respectively, which means that the shape feature is most important compared with other two features to retrieve desired images respect to a given query image. The comparison results show that the proposed CBIR scheme outperforms considerably other CBIR schemes.

The optimum weight values are varied depending on the given query image, so adaptive process according to a given query image to determine optimum weight values can be used, which is our feature work.

Acknowledgement

We thank to the National Council of Science and Technology of Mexico (CONACyT) for the financial support during the realization of this research.

References

[1] X. Y. Wang, Y. J. Yu, and H. Y. Yang, “An effective image retrieval scheme using color, texture and shape features,” Computer Standards & Interfaces, Vol 33, Mar. 2010, pp.59-68, doi:10.1016/j.csi.2010.03.004

[2] D. Zhang and G. Lu, “Evaluation of similarity measurement for image retrieval”, International Conference on Neural Networks and Signal Processing, Dec. 2003, pp. 928-931, ISBN: 0-7803-7702-8.

[3] J. Huang, S. R. Kumar, M. Mitra, W. J. Zhu, and R. Zabih, “Image indexing using color correlograms”, International Conference on Computer Vision and Pattern Recognition, Jun. 1997, pp. 762-768, ISBN: 0.8186-7822-4.

[4] H. Shao, Y. Wu, W. Cui, and J. Zhang, “Image retrieval based on MPEG-7 dominant color descriptor”, International Conference for Young Computer Scientist, Nov. 2008, pp. 753-757, ISBN: 978-0-7695-3398-8.

[5] E. Kasutani and A. Yamada, “The MPEG-7 color layout descriptor: a compact image feature description for high-speed image/video segment retrieval”, International Conference on Image Processing, Oct. 2001, pp. 674-677, ISBN: 0-7803-6725-1.

[6] K. M. Wong, L. M. Po, and K. W. Cheung, “Dominant color structure descriptor for image retrieval”, IEEE International Conference on Image Processing, Sept. 2007, pp. 365-368, ISBN: 978-1-4244-1437-6.

[7] F. R. Atoany, P. D. Karina, M. N. Mariko, and B. Jenny, “Dominant color correlogram descriptor for content-based image retrieval”, International Conference on Image Vision and Computing (ICIVC 2014), Sept. 2014.

[8] D. K. Park, Y. S. Jeon, and C. S. Won, “Efficient use of local edge histogram descriptor”, ETRI, vol. 24, Feb. 2002, pp. 23-30, doi:10.1145/357744.

[9] B. S. Manjunath and W. Y. Ma, “Texture features for browsing and retrieval of image data”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 18, no. 8, Aug. 1996, pp. 837-842, doi: 10.1109/34.531803.

[10] N. Jhanwar, S. Chaudhuri, G. Seetharaman, and B. Zavidovique, “Content-based image retrieval using motif cooccurrence matrix”, Image and Vision Computing, vol. 22, Mar. 2004, pp. 1211-1220, doi: 10.1016/j.imavis.2004.03.026.

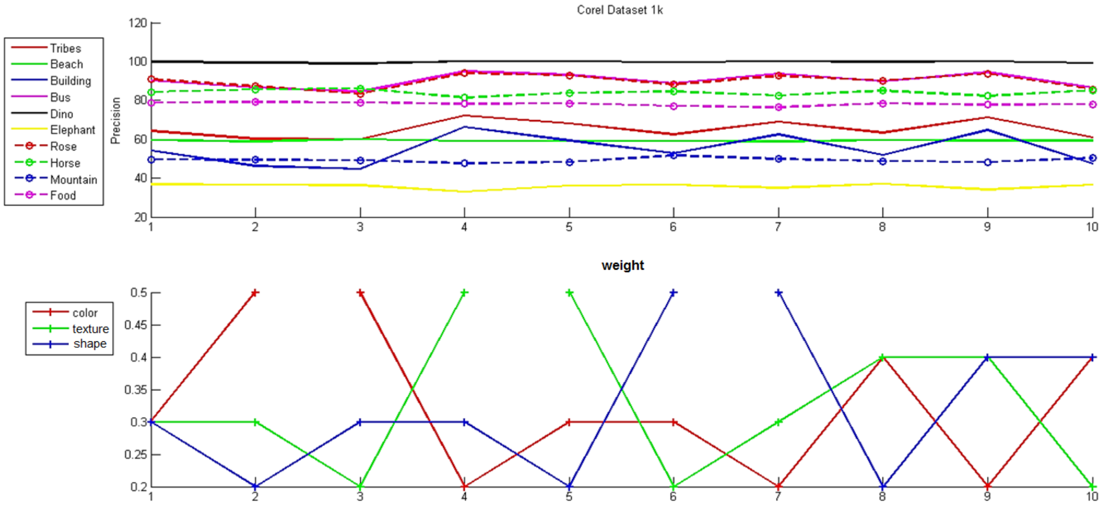


Figure 5. Average precision usgin Corel Dataset 1k

[11] S. K. Vipparthi and S. K. Nagar, “Expert image retrieval system using directional local motif XoR patterns”, Expert System with Applications, vol. 41, Dec. 2014, pp. 8016-8026, doi: 10.1016/j.eswa.2014.07.001.

[12] O. Timo, P. Matti, and H. David, “A comparative study of texture measures with classifcation based on featured distributions”, Pattern Recognition, vol. 29, no. 1, Jan. 1996, pp. 51-59, doi: 10.1016/0031-3203(95)00067-4.

[13] D. Zhang and G. Lu, “Evaluation of MPEG-7 shape descriptors against other shape descriptors”, Multimedia Systems, vol. 9, no. 1, Jul. 2003, pp. 15-30, doi: 10.1007/s00530-002-0075-y.

[14] P. T. Yap, X. Jiang, and A. C. Kot, “Two dimensional polar harmonic transforms for invariant image representation”, IEEE Transsactions on Pattern Analysis and Machine Intelligence, vol. 32, Jul. 2010, pp. 1259-1270, , doi: 10.1109/TPAMI.2009.119.

[15] Y. Bai, L. Guo, L. Jin, and Q. Huang, “A novel feature extraction method using pyramid histogram of orientation gradients for smile recognition”, IEEE International Conference on Image Processing, Nov. 2009, pp. 3305-3308, ISBN: 978-1-4244-5455-0.

[16] W. Xu, J. Wu, X. Liu, L. Zhu, and G. Shi, “Application of image SIFT features to the context of CBIR”, International Conference on Computer Science and Software Engineering, 2008, pp. 552-555, ISBN: 978-0-7695-3336-0.

[17] C. H. Manuel, G. U. Francisco, C. H. Antonio, N. M- Mariko, and P. M. Hector, “Mexican archaeological image retrieval based on object matching and a local descriptor”, International Conference on Computer Communications and Informatics (ICCCI 2015), Jan. 2015, ISBN: 978-1-4799-6805-3.

[18] Y. Jun, G. J. Yu, G. H. Alexander, and W. N. Chong, “Evaluating bag-of-visual-words representationsin scene classification”, International Workshop on Multimedia Information Retrieval, 2007, pp197-206, doi: 10.1145/1290082.1290111.

[19] W. Ji, W. Dayong, C. H. H. Steven, W. Pengcheng, Z. Jianke, Z. Yongdong, and L. Jintao, “Deep learning for content-based image retrieval: a comprenhensive study”, International Conference on Multimedia, 2014, pp. 157-166, ISBN: 978-1-4503-3063-3.

[20] V. L. Milind, B. Praveen, and J. Pritesh, “An effective content-based image retrieval using color, texture and shape feature”, Intelligent Computing, Networking, and Informatics, vol. 243, Dec. 2013, pp. 1163-1170, ISBN: 978-81.322-1664-3.

[21] S. Nishant and T. Vipin, “Region based image retrieval using integrated color, texture and shape features”, Information Systems Design and Intelligent Applications, vol. 340, Jan. 2015, pp. 309-316, ISBN: 978-81-322-2246-0.

[22] B. G. Prasad, S. K. Gupta, and K. K. Biswas, “Color and shape index for region-based image retrieval”, vol. 2059, May. 2001, pp. 716-725. ISBN: 978-3-540-42120-7.

[23] F. Pawel and F. Dariusz, “Strategies of shape and color fusions for content based image retrieval”, Computer Recognition Systems 2, vol. 45, 2007, pp. 3-10, ISBN: 978-3-540-75174-8.

[24] H. A. Jalab, “Image retrieval system based on color layout descriptor and gabor filters” International Conference on Open Systems (ICOS201), Sept. 2011, pp. 32-36, ISBN: 978-1-61284-931-7.

[25] M. Bleschke, R. Madonski, and R. Rudnicki, “Image retrieval system based on combined MPEG-7 texture and colour descriptors”, International Conference Mixed Design of Integrated Circuits and Systems, Jun. 2009, pp. 635-639, ISBN: 978-1-4244-4798-5.

[26] F. R. Atoany, P. D. Karina, N. M. Mariko, P. M. Héctor Pérez, and B. P. Jenny, “An effective visual descriptor based on color and shape features for image retrieval”, Mexican International Conference on Artificial Intelligence (MICAI 2014), Nov. 2014, pp. 336-348, ISSN: 0302-9743, ISBN: 978-3-319-13646-2.

[27] S. Ch. Ryszard, A. Tomasz, and Ch. Michal, “Integrated color, texture and shape information for content-based image retrieval”, Pattern Analysis and Applications, vol. 10, Apr. 2007, pp. 333-343. ISSN: 1433-7541.

[28] A. Talib, M. Mahmuddin, H. Husni, and L. E. George, “A weighted dominant color descriptor for content-based image retrieval”, Journal of Visual Communication & Image Representation, vol. 24, Jan. 2013, pp. 345-360, doi: 10.1016/j.jcvir.2013.01.007.

[29] N. Jhanwar, S. Chaudhuri, G. Seetharaman, and B. Zavidovique, “Content-based image retrieval using motif cooccurrence matrix”, Image and Vision Computing, vol. 22, Mar. 2004, pp. 1211-1220, doi: 10.1016/j.imavis.2004.03.026.

Audio Event Detection Using Adaptive Feature Extraction Scheme

Selver Ezgi Küçükbay[§] and Mustafa Sert[*]
Department of Computer Engineering
Başkent University
Ankara 06810 Turkey
Email: [§]seyalniz@baskent.edu.tr, [*]msert@baskent.edu.tr

***Abstract*—Audio event detection is one of the important tasks of multimedia content analysis. The noise like characteristics and the diversity of audio events make the recognition task difficult when compared with music and speech sounds. Therefore, proper application of feature extraction methods is very crucial, as well as feature selection and machine learning algorithms. Here, we propose a novel adaptive feature extraction scheme along with Support Vector Machine (SVM) learner in recognizing audio events. In our scheme, we propose to apply the widely used Mel frequency cepstral coefficients (MFCCs) feature to the problem in an adaptive way. To this end, we analyze each audio event in its frequency space to obtain a dominant frequency and then make use of the determined dominant frequency in the feature extraction phase. Extensive experiments have been conducted on sixteen (16) different audio events namely *alert*, *clear throat*, *cough*, *door slam*, *drawer*, *keyboard*, *keys*, *knock*, *laughter*, *mouse*, *page turn*, *pen drop*, *phone*, *printer*, *speech*, and *switch* using the IEEE AASP CASA Challenge Dataset to demonstrate the performance of the proposed scheme. The results show that our adaptive feature extraction scheme achieves significantly higher recognition accuracy than traditional feature extraction method with an average F-measure value of 72%.**

Keywords*–Audio event detection; Audio content analysis; Environmental sound detection; MFCC; SVM;

I. Introduction

Over the last decade, there has been an increased interest in the audio community for detecting acoustic events (also called as audio events) in audio signals. The main motivation is to develop automatic methods for recognizing sounds of particular events in any environment. However, the problem is challenging for two reasons when compared with speech and music sounds: (a) the variability and (b) the diversity of audio events (AEs). The former describes the dynamic nature of AEs and may lead to the perception of an AE as a different sound at distinct location/times; the latter is about the diversity of these sounds in the environments [12]. As a result, studies in AE recognition have received some interests in the last few years [1]–[4].

Cai *et al.* [1] work on the problem of highlight sound effects detection. They used Hidden Markov Models (HMM) with different feature extractors such as Mel Frequency Cepstral Coefficient (MFCC), Zero Crossing Rate (ZCR), sub band energies, brightness and bandwidth features in their study. They combined all features in one feature vector to achieve better results during the experiments. Their system gives Precision and Recall values of 90%. Wang *et al.* [2] present an audio event sound classification system to recognize 12 different audio events. In their study they combine Support Vector Machine (SVM) and k- Nearest Neighbor (kNN) classifier. In feature selection, they use MPEG-7 audio low level descriptors, spectrum centroid (SC), spectrum spread (SS) and spectrum flatness (SF). The classification accuracy is 85.1%. Chu *et al.* [3] propose a new method based on matching pursuit (MP) algorithm for analyzing audio events. They use 14 different audio scenes. The tests are applied through using 4 fold cross validation. Their overall accuracy is 72% for MP-based feature.

Lee *et al.* [4] present a method in order to identify and segment the frames in to regions. They used Markov model based clustering algorithm. For the dataset they download 1873 video for 25 different concepts from YouTube. They evaluate their study using average precision for each class. They yield best result for cheering segments. Beritelli et al. [15] work on a pattern recognition system for background sounds such as bus, car, construction, dump, factory, office and pool. Their classifier is Neural Networks (NN) and feature extractor is MFCC. They evaluate their systems in terms of percent misclassification and indicate accuracy between 73% and 95% depending on the duration of decision window. Muhammad et al. [8] studied on an environment recognition system. They use selected MPEG-7 audio low level descriptors and MFCC feature. In their method they eliminate MPEG-7 descriptor using Principal Component Analysis (PCA) and combine with MFCC feature. In this work, restaurant, crowded street, quiet street, shopping mall, car with open window, car with closed window, corridor of university campus, office room, desert and park are used for evaluation. For only MFCC, full MPEG-7, selected MPEG-7 and their method, the system gives accuracies of 85%, 89%, 91% and 93%, respectively. Schrder et al. [7] propose an audio-event detection system. Their system consists of two–layered hidden Markov Model as backend classifier. The system is evaluated with the materials provided in the AASP Challenge on Detection and Classification of Acoustic Scenes and Events [5]. For event-based results, the optimization applied on the dataset returns Precision, Recall and F-measure value of 66%, 58% and 62% respectively. Vuegen et al. [9] design a system based on MFCCs to train a Gaussian Mixture Models (GMM) classifier and make use of the same AASP Challenge dataset for the evaluations. The reported event-based performances for precision, recall, and F-measure are 68%, 33%, and 43%, respectively. Kucukbay et al. [10] propose a system for detection the audio events in office live environment. They propose efficient representation of MFCC features using different window and hop sizes by changing the number of Mel coefficient and also they optimize

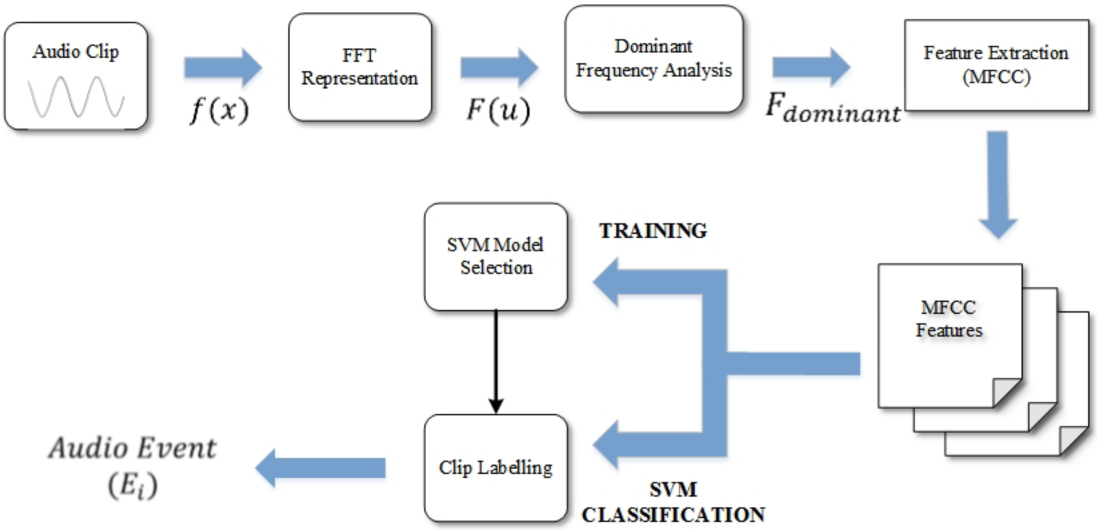


Figure 1. Block diagram of the proposed framework.

SVM parameters. The dataset are provided from subtask Office Live Environment of AASP Challenge. In the work, they use 16 distinct audio events. The tests conduct through using 5-fold cross validation gives the result of 62%, 58%, and 55% for Precision, Recall and F-measure.

Recent research shows that the performance of audio event recognition can be enhanced using suitable machine learning algorithms along with robust features [7], [11]. However, most of these studies make use of standard procedures during the feature extraction phase. For instance, in MFCC feature extraction, we obtain the coefficients from a given frequency interval, namely low- and high-frequency bounds. Using the fixed frequency bounds in the analyses of different types of sounds may lead to miss some important frequency components, since each sound source may have different bounds.

In this paper, we present a novel adaptive feature extraction scheme to recognize audio events by capturing each sound by its own frequency bounds along with SVM classifier. We consider sixteen distinct audio events from IEEE AASP CASA Challenge, namely alert, clear throat, cough, door slam, drawer, keyboard, keys, knock, laughter, mouse, page turn, pen drop, phone, printer, speech, and switch [5].

The paper is outlined as follows: In Section 2, the proposed recognition system is introduced. Empirical analysis and recognition performance are presented in Section 3 and finally, concluding remarks are given in Section 4.

II. The Proposed Scheme

The presented system consist of 7 main blocks, namely Audio Clip, FFT Representation, Frequency Analysis, Feature Extraction, MFCC Audio Features, SVM Model Training and SVM Classification. The block diagram of the proposed system is depicted in Figure 1.

A. *Adaptive Feature Extraction Scheme*

Each audio event conveys different information, i.e., comprise of different frequency components. Although the sampling rate of a signal defines the upper frequency bound of a signal in the analyses, each audio event may have its dominant frequency. In our proposed scheme, we aim to analyze each audio clip in its own frequency range during the MFCC feature extraction. Thus, we intend to capture the specific frequency range of each sound.

Specific frequency range (also referred to as dominant frequency) can be analyzed through complex methods but we verify our consideration using a fast, yet simple algorithm. In order to capture the characteristics of AEs in audio signals and to prove the effectiveness of our adaptive scheme, we use the MFCC feature due to its success in speech recognition applications. On the other hand, our proposed scheme is flexible and hence can be applied to other frequency-domain audio features. We used the standard MFCC feature extraction algorithm in [14]. In order to extract the MFCC features, we need to know the lower- and upper-frequency bounds. If we use the default values in the standard, which are defined as $300Hz$ for the lower-bound (LF) and $3700Hz$ for the upper-bound (HF), we can miss some important frequency components of an audio clip having different frequency bounds.

To solve this problem, we analyze the signal to determine its dominant frequency component. Let E denotes an audio event class (e.g., alert), then our scheme for determining the dominant frequency is given in (1).

$$f_{dominant}(E_i) = \frac{1}{N} \sum_{(k=0)\in E_i}^{N-1} f_k(idx(\max(|F_k|))) \quad (1)$$

where $1 \leq i \leq \sharp of audioevents$, E_i is the ith audio event, N is the number of audio clips in E_i, and F_k is the Fourier transformation of the k^{th} audio clip, $idx(y)$ represents the index number of y, and $f_k(z)$ denotes the frequency value of the k_{th} audio clip at index z. The main idea behind this formula is to define a dominant frequency for each AE class and make use of it in the feature extraction phase. We assume that, the most frequent frequency appeared in the signal is the dominant frequency of this clip.

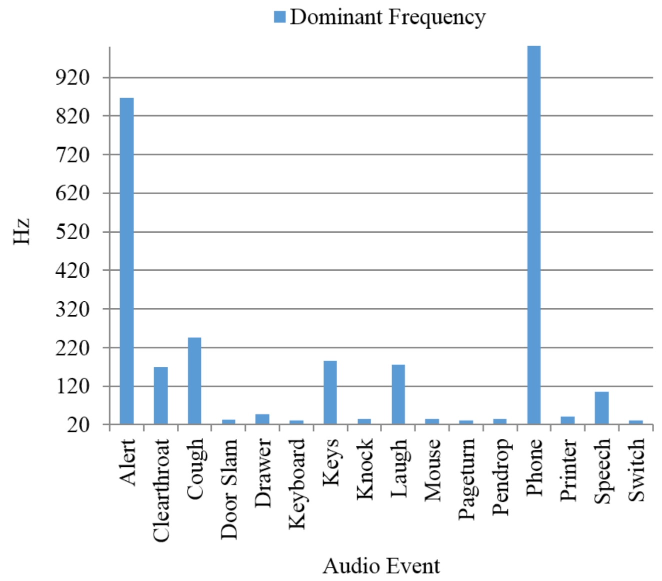


Figure 2. Dominant frequency of each audio event (AE).

TABLE I. Structure of the utilized dataset.

Audio Event Name	Duration
Alert	40 sec
Clearthroat	23 sec
Cough	23 sec
Door Slam	44 sec
Drawer	33 sec
Keyboard	1 min 16 sec
Keys	41 sec
Knock	26 sec
Laugh	30 sec
Mouse	29 sec
Pageturn	1 min 03 sec
Pendrop	16 sec
Phone	3 min 05 sec
Printer	7 min 01 sec
Speech	1 min
Switch	10 sec
TOTAL	**18 min 49 sec**

In this formula, following the Fourier transformation of the signal, which is denoted by F, we pick the frequency with the maximum magnitude as the dominant frequency and calculate the dominant frequency of each class by calculating the mean of dominant frequencies of the clips occurred in that class. Eventually, we obtain sixteen distinct dominant frequency corresponding to each AE class. We used dominant frequencies for the LF value in the feature extraction process. For the value of HF, we specified $22050Hz$ according to Nyquist theorem since the sampling rate of the clips in the dataset is $44100Hz$. Each class and their dominant frequencies are given in Figure 2. Once we find the proper frequency bounds of each class, we extract MFCC feature of each audio clip in the dataset using these frequency bounds. In our study, we choose a clip-based decision strategy for evaluating the results. When a clip is assigned to a particular class tag during the testing phase, the system selects a distinct class out of 16 different options for each frame of MFFCs that has been extracted for this particular clip.

B. *Classifier Design*

For audio event detection, we classify sounds with SVM classifier with radial basis function (RBF). This method is selected owing to achievement results in pattern recognition applications. We use LIBSVM library for the implementation [13]. Our multiclass evaluation strategy is defined as the one-versus-all approach. For each class, a separate SVM model is built such that every single SVM are trained to detect the features of particular classes and distinguish them from the others. In order to optimize the SVM parameters γ and C, we performed the grid search algorithm. Consequently, 16 different model files belonging to a particular class are created. In testing phase, the experiments conducted through using 5–fold cross validation.

III. Experimental Results

In model training and testing, we use audio event clips that are collected from the publicly available dataset of the sub-task Event Detection Office Live of the IEEE AASP Challenge Detection and Classification of Acoustic Scenes and Events [5]. These 16 distinct audio events include *short alert-beeping*, *clearing throat*, *cough*, *door slam*, *drawer*, *keyboard clicks*, *keys clinging*, *door knock*, *laughter*, *mouse click*, *turning page*, *object hitting table*, *phone ringing*, *speech*, *printer*, and *switches*.

Each class contains 20 recordings. Durations of recordings are changing because recording are collected from real-world environment. The dataset contains non-overlapping events from the office live environments. Class durations are presented in Table I.

In order to evaluate the proposed scheme, we prepared three scenarios. In the first scenario, we tested the standard MFCC implementation along with the SVM classifier. In the second one, we considered the standard MFCC feature extraction along with *optimized* SVM, and in the last scenario, we applied the proposed adaptive feature extraction scheme to the MFCC feature along with the optimized SVM. In the evaluations, we used 5–fold cross validation method and never mixed the train and the test datasets.

When we applied the proposed scheme, which considers adaptive feature extraction scheme using the dominant frequency for each class, we obtain an F-measure value of 72%.

In the second scenario in which we use fixed frequency bounds during the MFCC implementation, the recognition performance decreases to an F-measure value of 55%. And lastly, the first scenario that uses standard methods, we note an F-measure value of 48%. Our empirical results clearly show that, the proposed adaptive feature extraction scheme is superior to the standard methods and yields 17% increase in the recognition performance. Figure 3 provides a comparison for these three approaches. In addition, the proposed scheme also improves the confusion of similar AEs, such as pen drop and page turn sounds. The confusion matrices of the proposed scheme (the 3rd scenario) and the 2nd scenario are depicted in Table II and Table III, respectively. In both tables, each column of the matrix represents the instances in a predicted

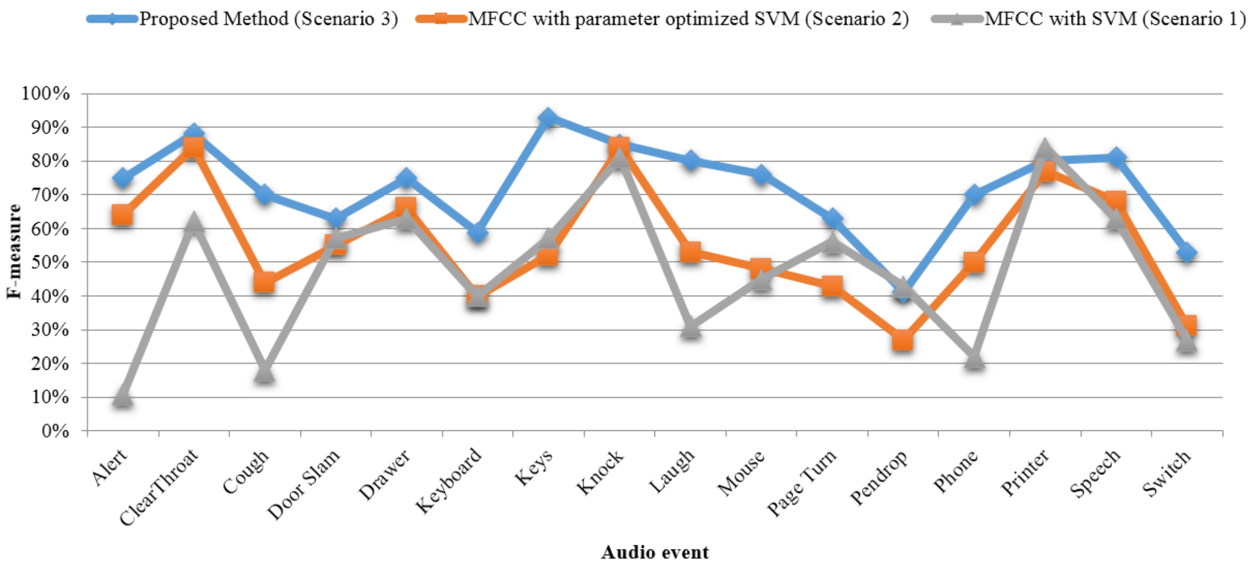


Figure 3. Recognition performances of the proposed scheme and the others.

TABLE II. Confusion matrix for 16–class classification using the proposed method (5–fold)

	Alert	Clear Throat	Cough	Door Slam	Drawer	Keyboard	Keys	Knock	Laughter	Mouse	Page Turn	Pen Drop	Phone	Printer	Speech	Switch
Alert	**13**	0	0	0	0	0	0	0	1	0	0	0	4	0	2	0
Clear Throat	0	**18**	1	0	0	0	0	0	0	0	0	0	0	0	1	0
Cough	0	0	**13**	1	0	0	2	0	0	0	0	2	1	1	0	0
Door Slam	0	0	0	**12**	4	0	0	0	0	0	1	0	0	3	0	0
Drawer	0	0	0	2	**17**	0	0	0	0	0	0	0	0	0	1	0
Keyboard	0	0	0	1	0	**11**	0	0	0	1	4	1	0	0	0	2
Keys	0	0	0	0	0	0	**19**	0	0	0	0	1	0	0	0	0
Knock	0	0	0	1	0	0	0	**17**	0	2	0	0	0	0	0	0
Laughter	0	1	1	0	0	0	0	0	**15**	0	0	0	1	1	1	0
Mouse	0	0	0	1	0	0	0	1	0	**15**	2	0	0	0	0	1
Page Turn	0	0	0	0	1	3	0	0	0	0	**15**	0	0	0	0	1
Pen Drop	0	0	1	0	1	1	0	0	0	1	4	**7**	0	3	1	1
Phone	1	1	0	0	1	0	0	1	0	0	0	0	**14**	0	2	0
Printer	0	0	0	0	1	0	0	1	0	0	0	0	0	**18**	0	0
Speech	0	0	0	0	1	0	0	0	0	0	0	0	0	0	**19**	0
Switch	0	1	0	0	0	3	0	0	1	1	2	3	0	0	0	**9**

class, whereas the rows represent the instances in an actual class. In Table III, we note that some of the sound classes such as *pen drop*, *switch*, *cough*, and *phone*, are mixing with other classes and decreases the overall performance. When used our method, we can read from Table II that the correct hit of *pen drop* increases by 3, *switch* and *phone* sounds increase by 5, *cough* sound increases by 6, and like many others increase the overall recognition performance dramatically.

This improvement can be described as using the own frequency spectrum of each sound provides the utilized frequency-spectrum feature to capture the characteristics of sounds better than using a fixed frequency range. Specifically,

TABLE III. Confusion matrix for 16–class classification using the MFCC with parameter optimized SVM (5–fold)

	Alert	Clear Throat	Cough	Door Slam	Drawer	Keyboard	Keys	Knock	Laughter	Mouse	Page Turn	Pen Drop	Phone	Printer	Speech	Switch
Alert	**15**	0	0	0	0	0	0	0	0	0	0	0	3	0	1	1
Clear Throat	0	**17**	0	0	1	0	0	0	0	1	0	0	0	0	1	0
Cough	1	1	**7**	0	2	1	2	0	4	0	2	0	0	0	0	0
Door Slam	0	0	0	**12**	3	0	0	1	0	0	1	0	0	3	0	0
Drawer	0	0	0	3	**15**	0	0	0	0	0	1	0	0	0	1	0
Keyboard	0	0	0	1	0	**9**	3	0	0	2	5	0	0	0	0	0
Keys	0	0	0	1	0	3	**12**	0	0	0	3	1	0	0	0	0
Knock	0	0	0	0	0	0	0	**18**	0	2	0	0	0	0	0	0
Laughter	0	0	3	0	1	0	1	0	**10**	1	0	0	1	1	2	0
Mouse	1	0	0	0	0	1	0	2	0	**10**	3	0	0	0	3	0
Page Turn	0	0	0	0	1	5	2	0	0	1	**11**	0	0	0	0	0
Pen Drop	0	0	1	3	1	2	3	0	0	1	4	**4**	0	1	0	0
Phone	6	0	0	0	0	0	0	1	1	1	0	1	**9**	0	1	0
Printer	0	0	0	2	1	0	0	1	0	0	0	0	0	**16**	0	0
Speech	1	1	0	0	0	0	1	0	2	0	0	0	0	0	**15**	0
Switch	2	1	0	0	0	3	3	0	0	2	3	2	0	0	0	**4**

we assign the computed dominant frequency as the low frequency bound and perform the frequency analyses using the range of $[f_{dominant}, 22050Hz]$. Another option might be to use the computed dominant frequency as the high frequency, but in this case we have to compute the low frequency bound by introducing additional computation cost, since we do not know it in advance. In our case, we know the high frequency bound in advance (i.e., 22050 *Hz* by the *Nyquist* theorem). We can read from the Figure 3 that in some cases (e.g., *printer*, *knock*, and *pendrop*), the proposed method performs similar success rates as the standard methods. To the best of our knowledge, this is because of the short durations used in the training and/or the characteristics of these sounds are quite hard to capture for the MFCC.

IV. Conclusion

This paper introduce a novel adaptive feature extraction scheme for the recognition of sixteen distinct audio events namely *alert*, *clear throat*, *cough*, *door slam*, *drawer*, *keyboard*, *keys*, *knock*, *laughter*, *mouse*, *page turn*, *pen drop*, *phone*, *printer*, *speech*, and *switch* from audio clips. In the experiments, clips are recognized and tested using the proposed scheme based on the MFCC feature and the SVM classifier.

Our study shows that, when we apply specific frequency limits for each class, we attain 72% F-measure score, which is better than both the standard methods (F-measure value of 48% and 55%) and the event-based results of the IEEE AASP Challenge (61.52% F-measure value) [8]. Based on the experiments, the proposed scheme outperforms the standard methods by 17% and the IEEE AASP Challenge results by 10.48%.

Our feature work lies on the detection of audio scenes using the audio events.

References

[1] R. Cai, L. Lie, Z. Hong-Jiang, and C. Lian-Hong, “Highlight sound effects detection in audio stream,” Multimedia and Expo (ICME’03), International Conference on, 2003, pp.37–40.

[2] W. Jia-Ching, W. Jhing-Fa, H. Kuok Wai, and H. Cheng-Shu, “Environmental Sound Classification using Hybrid SVM/KNN Classifier and MPEG-7 Audio Low-Level Descriptor,” Neural Networks, IJCNN ’06. International Joint Conference on, 2006, pp.1731–1735.

[3] S. Chu, S. Narayanan, and C.-C.J. Kuo, “Environmental sound recognition using MP-based features,” Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on, 2008, pp.1–4.

[4] K. Lee, D. Ellis, and A. Loui, “Detecting Local Semantic Concepts in Environmental Sounds using Markov Model based Clustering,” Proc. IEEE ICASSP, 2010, pp.2278–2281.

[5] D. Giannoulis, E. Benetos, D. Stowell, and M. D. Plumbley, “IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events - Development Dataset for Event Detection Task, subtask 1 - OL,” Applications of Signal Processing to Audio and Acoustics (WASPAA), 2013 IEEE Workshop on, 2013, pp.1–4.

[6] C.C. Chang and C.J. Lin, “LIBSVM: A library for support vector machines,” ACM Transactions on Intelligent Systems and Technology, Vol. 2, No. 3, 2011, pp.1–27.

[7] J. Schroder et al., “On the use of spectro-temporal features for the IEEE AASP challenge detection and classification of acoustic scenes and events,” Applications of Signal Processing to Audio and Acoustics (WASPAA), 2013 IEEE Workshop on, 2013, pp.1–4.

[8] G. Muhammad, Y. A. Alotaibi, M. AlSulaiman, and M.N. Huda, “Environment Recognition Using Selected MPEG-7 Audio Features and Mel-Frequency Cepstral Coefficients,” Digital Telecommunications (ICDT), 2010 Fifth International Conference on, pp.11–16, 2010.

[9] L. Vuegen, B. Van Den Broeck, P. Karsmakers, J. F. Gemmeke, B. Vanrumste, and H. Van hamme, “An mfcc-gmm approach for event detection and classification,” IEEE AASP Challenge on Detection and Classification Acoustic Scenes and Events, 2013.

[10] S. E. Kucukbay and M. Sert, “Audio-based event detection in office live environments using optimized mfcc-svm approach,” IEEE International Conference on Semantic Computing (ICSC’15), 2015, pp. 475–480.

[11] L. Chen, S. Gunduz, and M. T., “Mixed Type Audio Classification with Support Vector Machine,” Multimedia and Expo, 2006 IEEE International Conference on, 2006, pp.781–784.

[12] C. Okuyucu, M. Sert, and A. Yazici, “Audio Feature and Classifier Analysis for Efficient Recognition of Environmental Sounds,” Multimedia (ISM), 2013 IEEE International Symposium on, 2013, pp.125–132.

[13] L. Rabiner and J. Biing-Hwang, “Fundamentals of Speech Recognition.” Alan V. Oppenheim, Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1993.

[14] C. C. Chang and C.J. Lin, “LIBSVM: A library for support vector machines,” ACM Transactions on Intelligent Systems and Technology, Vol. 2, No. 3, 2011, pp. 1–27.

[15] F. Beritelli and R. Grasso, “A pattern recognition system for environmental sound classification based on mfccs and neural networks,” IEEE International Conference on Signal Processing and Communication Systems (ICSPCS 2008), 2008, pp. 1-4.

Multimedia and Serious Games in Education

Carlos Oliveira
Department of Informatics
IFRJ
Rio de Janeiro, Brazil
carlos.roberto@ifrj.edu.br

***Abstract*—According to recently studies, people spend many hours a day using entertainment applications on the Internet. It poses a great opportunity for serious games that can be used on education. This is a recent research area that needs to be addressed. This paper presents an educational game and the tools that are being used to developed this game. The idea is to move away from the traditional textbook approach and make learning pleasurable. In a near future, this game will be evaluated by history and geography teachers and students.**

Keywords-serious game; history and geography learning; online game; development tools.

I. INTRODUCTION

We live in an era in which we are surrounded by information. Each day more information is produced and disseminated. This information is from different subjects and come from different sources, such as TV, newspapers and Internet. Nevertheless, Internet is the growing media in terms of user's interesting and access. If well used, information can aggregate value to products and also to people who has specific knowledge. Despite of the large amount of information received, people are quite poor at understanding and remembering information they have received out of context or too long before they can make use of it [1][2][3]. Thus, it poses a problem. How to assure the information will be remembered later?

Different from what happens in schools, games give information on demand, just in time and not out of the context of the game purposes. According to Gee [4], good games find ways to put information inside the worlds the players move through, and make clear the meaning of such information and how it applies to the world.

It is possible to use games to enhance learning at schools. Young people stay plugged on computers and other devices many hours a day. This public is very enthusiastic with technologies. They are also heavy information consumers. It is also important to consider that, in general, players read about a given subject not only inside the game, but also outside the game on websites, books, etc.

The goal of this paper is to present a game that is being developed to be used on schools. The game aim to help teachers on teaching subjects concerned to geography and history. Thus, it is presented the game's story and the tools used on the game development.

The rest of the paper is organized as follows. The game is presented in Section II. In Section III, it is presented the tools used to develop the game. Aiming to give a general idea of the game codification, it is also presented part of the Java script that contains the code to control the character of the game. Section IV concludes the paper and outlines future directions.

II. GAME

The game's story takes place on Earth after an alien has an accident with his ship and fall down on our planet. In their search for the key parts of his vehicle, the character passes through well-known touristic points of our planet and knows a bit of our culture. Therefore, the aim of the game is to find the parts of the ship. During this quest, the player will go through different cities on the planet. When the character goes through Rio de Janeiro, for example, the player gets information about the city's history, climate, traditional festivals, among other information that helps the player to know the city. In this game the player also listen to popular songs on the cities where the character is.

It would be impossible to show graphics and information on all touristic cities on Earth. Thus, the game focus on major world cities, such as New York, Rio de Janeiro, London, Paris, Berlin, Rome, Jerusalem, Moscow, Tokyo, Beijing and Sydney.

We believe that the game may be used to facilitate the approach of some subjects in disciplines such as history and geography. Furthermore, the development of the game back interesting with regard to computer programming, graphs for the development of the game and the game logic creation of challenges.

III. TOOLS

The game is being developed as presented in Figure 1. The game has a component that stores several songs. The second component stores scenarios and their objects. The third component gives intelligence to the game. Because of this component, songs are played according to the visited city. Also, it gives intelligence to the characters.

To develop the game songs, we used the program Fruity Loops [5]. The program Blender [6] was used to create the character and objects of the game. The Unity3D [7] was used to render the scenarios, to put the objects on the scenario, and to make the logic of the game.

In this section, these tools and screenshots of them are presented.

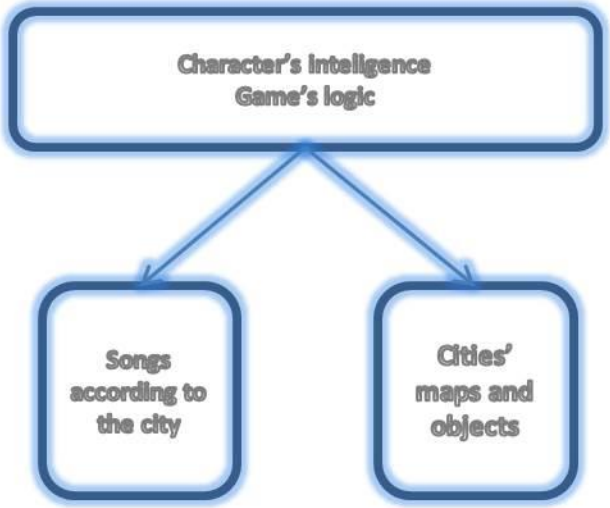


Figure 1. Game architecture.

A. Fruity Loops

Fruity Loops Studio resembles a complete recording studio for multimedia projects which is ideal for music demos, songs and any means of audio production. This virtual studio processes audio using an internal 32-bit floating-point engine. It can support sampling rates up to 192 khz either WDM or ASIO enabled drivers. The Mixer interface allows users for channel configurations and mixing 2.5, 5.1, or 7.1 surround sound possible. In addition, Fruity Loops studio also comes with a variety of plug-ins and generators (synthesizers) written in the program's own native plug-in architecture. However, first time users might be overwhelmed with its unlabeled icons and confusing file browser which makes the learning curve steeper. Once the users are already familiar with the program, they can make music in no time. A Fruity Loops screenshot is presented in Figure 2.

Figure 2. Fruity Loops Screenshot.

B. Blender

Blender is a professional free and open-source 3D computer graphics software product used for creating animated films, visual effects, art, 3D printed models, interactive 3D applications and video games. Blender's features include 3D modeling, texturing, raster graphics editing, rigging and skinning, fluid and smoke simulation, particle simulation, soft body simulation, sculpting, animating, match moving, camera tracking, rendering, video editing and composing. Alongside the modeling features it also has an integrated game engine. A Blender screenshot is presented in Figure 3.

Copyright (c) IARIA, 2015. ISBN: 978-1-61208-400-8

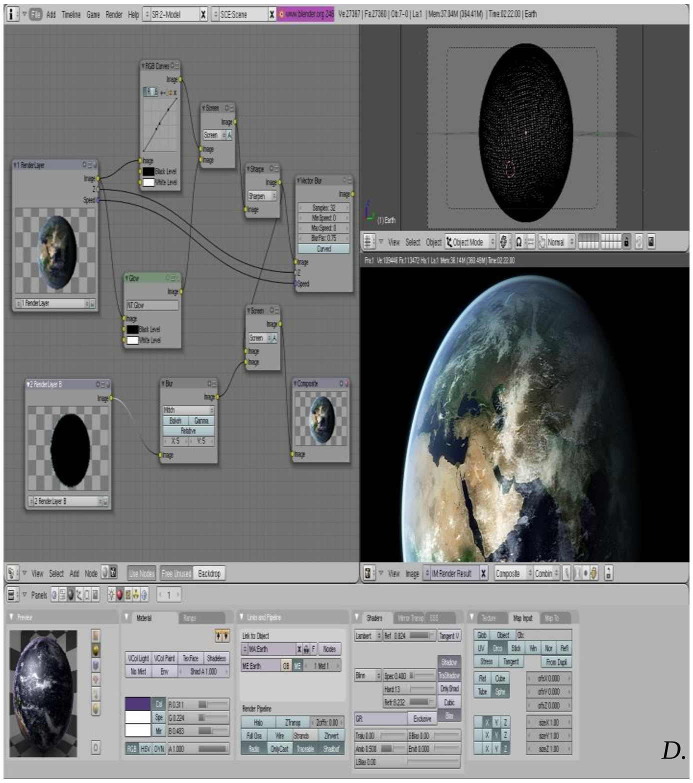
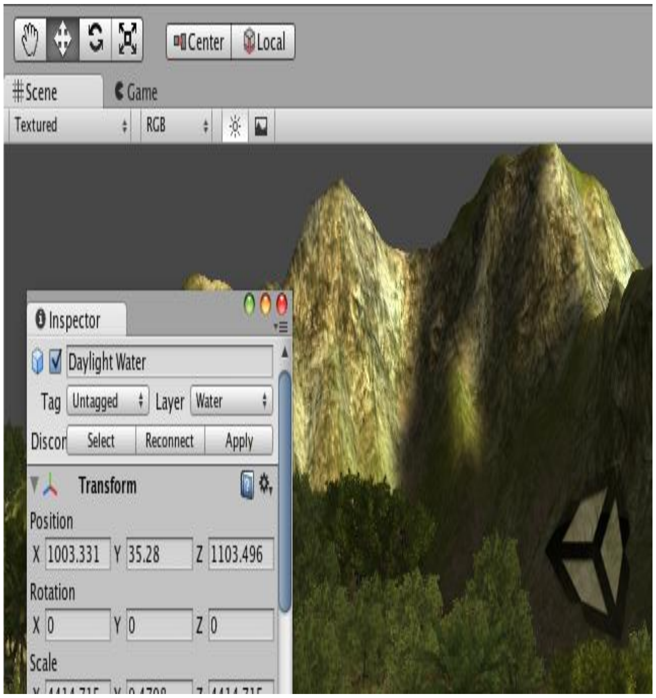


Figure 3. Blender Screenshot.

C. Unity3D

Unity is a cross-platform game creation system including a game engine and integrated development environment. It is used to develop video games for web sites, desktop platforms, consoles, and mobile devices. With an emphasis on portability, the graphics engine targets the following APIs: Direct3D on Windows and Xbox 360; OpenGL on Mac, Windows, and Linux; OpenGL ES on Android and iOS; and proprietary APIs on video game consoles. The game engine's scripting is built on MonoDevelop \cite{mono}, the open-source implementation of the .NET Framework. Programmers can use UnityScript (a custom language with ECMAScript-inspired syntax, referred to as JavaScript by the software).

Uniy3D is used to deal with the codification of the game. Unity combines the character, songs, objects and the codification of the game. A Unity3D screenshot is presented in Figure 4. The variables declaration that permits all character movements on the game are shown below.



Daylight Water

Tag Untagged Layer Water

Position

X 1003.331 Y 35.28 Z 1103.496

Rotation

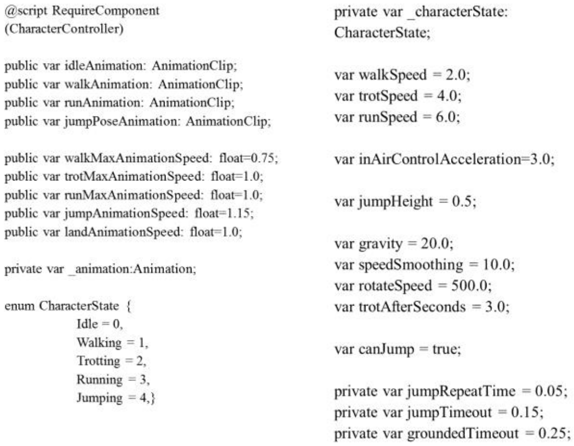
X 0 Y 0 Z 0

Scale

Figure 4. Unity3D Screenshot.

D. Scripts

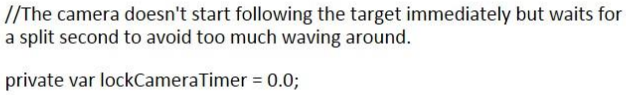
In this subsection, parts of the game code are presented. Figure 5 presents initial values that defines the speed of the character when walking, running or jumping. The variable presented in Figure 6 is configured to follow the character during the game. Thus, the player controllers the character as a third-person game.



```
@script RequireComponent
(CharacterController)

public var idleAnimation: AnimationClip;
public var walkAnimation: AnimationClip;
public var runAnimation: AnimationClip;
public var jumpPoseAnimation: AnimationClip;

public var walkMaxAnimationSpeed: float=0.75;
public var trotMaxAnimationSpeed: float=1.0;
public var runMaxAnimationSpeed: float=1.0;
public var jumpAnimationSpeed: float=1.15;
public var landAnimationSpeed: float=1.0;

private var _animation:Animation;

enum CharacterState {
        Idle = 0,
        Walking = 1,
        Trotting = 2,
        Running = 3,
        Jumping = 4,}

private var _characterState:
CharacterState;

var walkSpeed = 2.0;
var trotSpeed = 4.0;
var runSpeed = 6.0;

var inAirControlAcceleration=3.0;

var jumpHeight = 0.5;

var gravity = 20.0;
var speedSmoothing = 10.0;
var rotateSpeed = 500.0;
var trotAfterSeconds = 3.0;

var canJump = true;

private var jumpRepeatTime = 0.05;
private var jumpTimeout = 0.15;
private var groundedTimeout = 0.25;
```

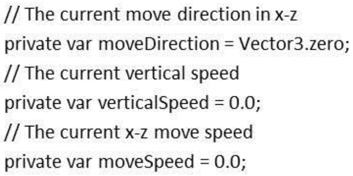
Figure 5. Variables.

Figure 7 presents the variables that allow the character to make two axis movements on the maps. All maps have collision objects that the player needs to detour. Collisions are allowed because of the variable presented in Figure 8.



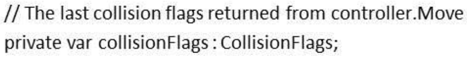
```
//The camera doesn't start following the target immediately but waits for
a split second to avoid too much waving around.

private var lockCameraTimer = 0.0;
```

Figure 6. Camera's code.



```
// The current move direction in x-z
private var moveDirection = Vector3.zero;
// The current vertical speed
private var verticalSpeed = 0.0;
// The current x-z move speed
private var moveSpeed = 0.0;
```

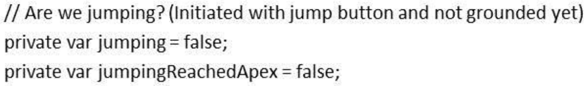
Figure 7. Character's movement.



```
// The last collision flags returned from controller.Move
private var collisionFlags : CollisionFlags;
```

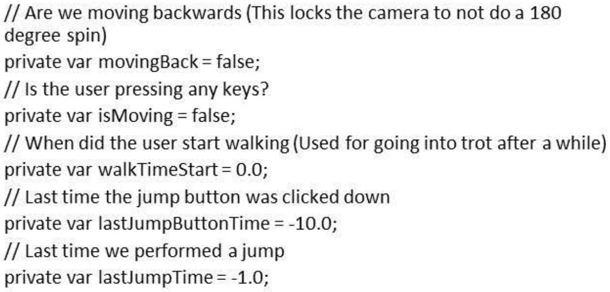
Figure 8. Collision flags.

Jumping makes possible to detour from collision objects. Initially the character is not jumping. Jumping is allowed by the variables presented in Figure 9. The character can jump on different heights. The height of jumping is defined on the variable presented in Figure 11. The height of jumping is calculated taking into account the last jump. Thus, the value of the last jump is stored on a variable presented in Figure 10. Figure 10 also presents variables that permits to show the character when jumping or making other movements.



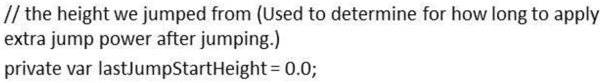
```
// Are we jumping? (Initiated with jump button and not grounded yet)
private var jumping = false;
private var jumpingReachedApex = false;
```

Figure 9. Jump button.



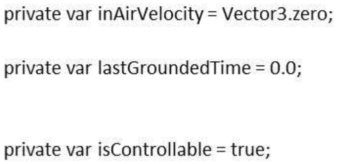
```
// Are we moving backwards (This locks the camera to not do a 180
degree spin)
private var movingBack = false;
// Is the user pressing any keys?
private var isMoving = false;
// When did the user start walking (Used for going into trot after a while)
private var walkTimeStart = 0.0;
// Last time the jump button was clicked down
private var lastJumpButtonTime = -10.0;
// Last time we performed a jump
private var lastJumpTime = -1.0;
```

Figure 10. Camera's movement.



```
// the height we jumped from (Used to determine for how long to apply
extra jump power after jumping.)
private var lastJumpStartHeight = 0.0;
```

Figure 11. Jumping.



```
private var inAirVelocity = Vector3.zero;

private var lastGroundedTime = 0.0;


private var isControllable = true;
```

Figure 12: Speed

IV. CONCLUSION AND FUTURE WORK

We are now finishing the game development. The game aims to facilitate the teaching of different subjects, such as geography and history. We argue that it is possible to use games to enhance learning at schools because young people stay plugged on computers and other devices that run games. In this game, we also aim to allow users to modify existing maps and create others. This will permit players to produce information (knowledge) and not only consume. Our goal is to allow users to change the game with no programming knowledge.

ACKNOWLEDGMENT

The author acknowledges financial support from Federal Institute of Rio de Janeiro during the work presented in this paper.

REFERENCES

[1] L. Barsalou: Language Comprehension: Archival Memory or Preparation for Situated Action? In: Discourse Processes, 1999, pp. 61–81.

[2] A. L. Brown: The advancement of learning. Educational researcher, 1994, pp. 4–12.

[3] A. M. Glenberg, D. A. Robertson: Indexical understanding of instructions. Discourse Processes, vol. 28, no. 1, 1999, pp. 1–26.

[4] J. P. Gee: What video games have to teach us about learning and literacy. Computers in Entertainment (CIE), v. 1, n.1, 2003, p.20-20.

[5] Introducing FL Studio 11, http://www.image-line.com/flstudio/ (Accessed 03/05/2015)

[6] Blender, http://www.blender.org/ (Accessed 03/05/2015)

[7] Unity3D, https://unity3d.com/ (Accessed 03/05/2015)

[8] MonoDevelop, http://monodevelop.com/ (Accessed 03/05/2015)

An Interaction Model Based on Narrative Programs to Improve Understanding and Contribution to Non-Linear Narratives.

Joaquim Colás, Alan Tapscott, Ayman Moghnieh, and Josep Blat
Universitat Pompeu Fabra
Email:{Joaquim.colas, alan.tapscott, ayman.moghnie, josep.blat} @upf.edu

Abstract—**Collaborative creation of stories poses new challenges to the authoring task. Being able to comprehend a large non-linear information space and to take into account inputs from other creators are important to contribute meaningfully and consistently. This work presents a model based on the classic semiotics concept of "narrative programs" to structure and present the information with the purpose of making non-linearity more accessible, facilitating contribution, and inspiring creative opportunities. We introduce a prototype that implements this model, and use it in an experiment to explore how users read a non-linear story, understand it, and contribute to it. Results show how users identified the main characters and related them to their narrative programs achieving high levels of comprehension, which the correspondence between comprehension and contribution consistency was, and that the users expanded the narrative from multiple points of view.**

Keywords-storytelling; comprehension; interaction models; authoring; collaborative creation.

I. INTRODUCTION

The traditional mono-directionality of storytelling is challenged by media concepts such as Transmedia (to combine different channels to create different narrative experiences in the same narrative universe, understood as the self-consistent fictional setting where the stories take place [1]), or by narrative "prosumers" (proactive consumers) who demand to actively participate in the development of those narrative universes (for instance in fan-fiction websites [2], where followers of a movie, TV series, novel series or other fiction franchises share their own stories taking place in their original universes). Nevertheless, the new types of narratives can grow into larger and more complex structures and pose new challenges to the creative authors, whose different contributions must deal with the specific requirements of the narrative genre, such as thematic and logic coherence and cause-to-effect connectivity [3].

Using a Research Through Design [4] approach, in previous works [5, 6] we identified that tools to support collaborative authoring require 1) providing the interaction mechanisms that allow the user to expand a story at any point of space and time, and 2) to empower the user to have a comprehensive view of all the large narrative space.

Comprehension (being able to understand the narrative content and to establish logical relations) can be a key factor for the creator to contribute meaningfully and consistently, as its lack when multiple users collaborate in the same space and not take into account the other contributions leads to narrative inconsistencies [6] (i.e., parts of the story contradict other parts), while psychological studies have highlighted comprehension as a factor for good authoring performance in terms of structure and consistency [7].

On the other hand, authoring in digital storytelling has been approached from diverse angles: some works close to automatic generation, as the ones by Pizzi and Cavazza [8] or Swatjes and Theune [9] propose authoring as a co-creation between generative Artificial Intelligences (AIs), which will grant the correctness of the information, and humans. Some researchers have worked with children and tangible interfaces for the creation of emergent fairytales [10, 11], where the systems try to respond consistently to the improvised actions of the kids. Most of the state of the art of interactive storytelling presents authoring tools that use graphs for organizing the non-linear narrative structures [12, 13, 14]. The collaborative online experiment by Likarish [15] pointed out the need of tools that provide the authors with the necessary information when contributing to multi-authored spaces.

In this paper, we propose an interaction model to facilitate the navigation of non-linear narrative spaces and to increase the contributors' awareness of the other authors input. Our model uses the "narrative programs" concept [16, 17] from narrative semiotics (which studies the creation of meaning in narratives) to structure the narration in character storylines and to present a way to connect them meaningfully. We turned this model into a prototype, *Proppulsion*, which is used in an experiment to test the readers' comprehension of the story, and to analyze the contributions of those who expand it.

This paper is structured as follows. In Section II, we review the related work on information models for storytelling systems. Section III introduces our model based on Narrative Programs for presenting and exploring narrative spaces from the perspective of the character roles and their relations towards other characters. In Section IV, we introduce Proppulsion and explain the setting and development of the experiment, followed by the presentation of the results in Section V. In Section VI, we discuss our findings: we point at how users identified the main characters relating them to their narrative programs and used their storylines as a backbone for exploring the whole narrative; how users who achieved greater comprehension also seemed to achieve greater consistency in their contributions; and how the system encouraged them to expand the story from multiple points of view. Finally, Section VII briefly summarizes our main conclusions and indicates some future work.

II. Information Models for Storytelling

The study of narrative information models has been usually approached with the goal of building intelligent *generative systems* that automatically produce narratives. Computational models to be processed through AI are far from our goal of interaction models aimed at being understood by authors, but it is convenient to indicate some of their aspects that are relevant for our approach.

Bailey [18] divides automatic story generation models into author models (imitating the human processes of authoring), story models (following a structural grammar) and world models (populating a setting with agents whose interactions result in a story) and proposes a model based on the reader's perspective. For Riedl and Young [19], generative systems can be categorized within a framework that balances plot coherence (author-centric systems) with character believability (character-centric systems). Mateas and Sengers [20] define story-understanding systems as those which "seek to model the processes by which a human understands a story".

From our perspective of narrative information models intended to support the interaction of human authors, we distinguish two types of models, depending on whether the story content is produced automatically or by an author.

Among the models for automatic *generative systems*, some are plot-based, when the system follows a set of rules to generate the story that has a certain semiotic structure; others are character-based, when the model is used to generate the actions of a set of characters and the narrative emerges from those actions, as in Cavazza's work [21]. This vision of the narrative, as the result of multiple characters each following his/her own narrative programs, helps to form our vision of a multi-linear story. Gervás [22] uses an implementation of the formal model of Propp's morphology of folk-tales, from which we draw some basic concepts in the next section. Some systems using generative models can be interactive as well, as Mateas and Stern Façade [23], where a user takes part in the story as a character and the system has to generate storyworld events and respond to his/her actions.

Other models support *authoring systems,* where one or more users perform the role of author. A lot of examples come from the field of authoring systems for interactive narratives, as Storytec [12], Scenejo [13] or Narrative Threads [14]. Those systems present the users tools to produce narratives and, as in classical hypertext narratives, they have to deal with non-linearity, since the author needs to build a changing structure that varies depending on the choices of the player. Quite a few of them (including [12, 13, and 14]) use graphs to represent those configurations. Hartman et al. [24] use Propp's structures to build those graphs.

How readers understand a narrative is useful not only for AI systems, as Matheas et al suggest for "story-understanding systems", but for the design of authoring systems as well. Also, classic semiotic models reflect how stories are understood from a human perspective, and this has been used for generative systems to build stories, but not so frequently for helping humans to deal with them. In this paper we adopt some of their notions.

In the context of collaborative non-linear storytelling, the distinction between author and consumer profiles is less clear. Authors do not prepare non-linear structures that will be experienced linearly by a reader, but read and then contribute to a global, multi-storyline structure that can be explored in many ways. We discuss next how we apply ideas from classic semiotics models, which help to understand and conform linear narrative structures, to this non-linear potentially ever-growing information space, in order to facilitate the authors to comprehend it and fit in it their contributions.

III. An interaction model based on "Character Narrative Programs"

In Propp's morphology of folk-tales [16], the story is driven by a concatenation of actions (called *functions*) of the main protagonist to reach his/her goal. The other characters perform simple functions within this chain depending on their roles in the story (rewarding the protagonist for accomplishing his/her goal, helping the protagonist in his/her quest, being an antagonist trying to defeat the protagonist plans, etc.). Greimas revised these concepts in his semiotics theory, where he defined *Narrative Programs* as the selection of events linked together revealing a direction or an intentionality to form a coherent narrative, thereby providing the narrative with meaning [17].

This resounds with findings of our previous work [5], where users of the CrossTale interface found useful exploring and creating collaborative stories through linear paths, which we call storylines. We saw that users mainly perceive storylines as character-driven, and that plots that follow the development of a character were preferred.

In this paper, we reinforce our approach by adapting the Narrative Program concept. Each character has his/her own narrative program, i.e., his/her own goal and associated storyline. When a character has a role in another character's storyline, the two storylines cross. For instance, in a classical tale, from the protagonist perspective (the prince), a wizard can be a "helper character" in his mission to save the princess, but in a multi-storyline narrative, the wizard is also the protagonist of his own storyline, and he helps the prince as part of his own narrative program.

Readers/authors can re-arrange the narrative space around a selected character storyline to explore and understand how the existent narrative programs connect, getting a consistent "bigger picture". On the other hand, this multiple-points-of-view approach to the narrative space could encourage creators to develop different character storylines, generating opportunities for rich contributions.

Next, we define each classical semiotics concept we use in our approach, explain how it relates to previous computational and interaction models for narratives and how we apply it in our proposed model.

A. Main and Secondary Characters

In Propp's approach, the main character's narrative program is the leitmotif of the story, while multiple secondary characters appear within this storyline. Plot-based systems built on classic semiotic models follow this. Character-based systems can have multiple protagonists depending on the complexity of the agents' (characters') actions. Authoring focused on reader's interaction tends to put the reader/player in the place of the main character, while multi-author systems let authors control one or more characters [11], without distinguishing between main and secondary ones. Our approach presents the user (both reader and author) an explicit multiple-points-of-view exploration through the use of character-driven storylines. Each character performs as the main protagonist of his/her storyline, while the others are presented as secondary and defined by their relation with the protagonist's narrative program, described by the secondary character's role on it.

B. Narrative Programs

The main character undertakes multiple sub-tasks to accomplish his/her goal, creating a chain of events. Secondary characters' narrative programs usually refer only to their roles in the main story. Some plot-based systems also use the protagonist's narrative program as the story central structure. Character-based generative systems use narrative programs as agents' goals, and their planning steps become action sequences. In authoring systems the narrative program tends to be implicit, as it is developed by the authors' decisions. In our interaction model based on narrative programs, when focusing on a single character, his/her actions in the overall narrative space are presented in a linear and coherent sequence as the main plot of that sub-story.

C. Character Roles

Each character has a role or a small set of roles. Traditionally, they are always defined in relation with the protagonist (helper, antagonist, quest-giver, etc.), so that one could talk about "absolute" roles. In authoring systems, the roles of the characters are implicit in the story description. In character-based generative systems, roles are implicit in the character's goals through their relation with those of the other agents; thus roles are "relative" to those of other characters, as each character is the protagonist in his/her storyline and plays different roles in the others' storylines. Our approach makes explicit this notion of relative role.

D. Time and Space

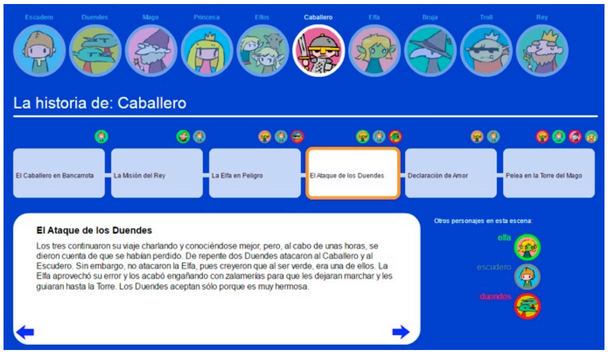
In classical tales morphology, time is relative to the development of the main character story, while space is lightly considered. Some systems use a discretization of time (e.g., character-based systems using planning perform cycles of actions) or discretize space in finite "places" (e.g., [11]). Previously [5], we used a loose discretization of time in frames, while places were a list of settings. Users understood time in a vague way, contextualizing each scene depending on the semantic relation with the nearby ones, while place was just considered as an ambient accessory. In this paper, each scene has a global reading order, so that there is an implicit global sequence of scenes when a sub-set is chosen to read. Time is, and implicitly put, in relation between storylines when they cross. Space is not considered as a specific object but implicit in each scene description.

To sum it up, our model draws from the classic semiotic elements of character narrative programs and roles but puts them in a multi-linear context, where each character can work as the protagonist of his/her own tale. It uses this structure to present the non-linear information to the reader so that s/he can explore and understand it in terms of the relations between the multiple stories. We aim at helping the readers achieve a better comprehension and suggest them new ways of contribution as authors.

In the next section, we present a small first experiment with this model to observe the kind of exploration encouraged by its use, to determine if readers can get a good comprehension of a non-linear story that has to be read in a fragmented manner, to test how comprehension helps them to achieve more consistent contributions, and to observe the kind of contributions elicited.

IV. EXPERIMENTAL SETTING

The interaction model we propose was implemented into a basic prototype we named "Proppulsion" (Figure 1). It reads a JSON (JavaScript Object Notation) file containing the story (a set of ordered unitary scenes, characters, and the definition of relations between them in each scene) and presents it through an interactive interface. There is a row of characters' icons at the top of the interface (in randomized order so that a hierarchy among them cannot be presumed). By clicking on one of them, the character's storyline (i.e., narrative program) is shown, as the series of scenes where s/he has a role presented in temporal order. The user can read it sequentially by using the "previous" and "next" buttons or in a desired order by selecting the titles of the scenes. In each scene, the interface shows a list of the secondary characters and their role with respect to the narrative program of the current protagonist's (i.e., the character chosen) in that scene, defined by a colour code as "helper", "opponent" or "other". At any moment, the user can switch to another character.



El Ataque de los Duendes

Los tres continuaron su viaje charlando y conociéndose mejor, pero, al cabo de unas horas, se dieron cuenta de que se habían perdido. De repente dos Duendes atacaron al Caballero y al Escudero. Sin embargo, no atacaron la Elfa, pues creyeron que al ser verde, era una de ellos. La Elfa aprovechó su error y los acabó engañando con zalamerías para que les dejaran marchar y les guiaran hasta la Torre. Los Duendes aceptan sólo porque es muy hermosa.

Figure 1. The *Proppulsion* interface.

The experiment with Proppulsion was double blind: an external author created the story, a fairy-tale with 10 typical characters, each having different objectives, and 13 scenes. The story was written from a third person, omniscient point of view, and revolved around the kidnap of a Princess by an evil Wizard who wanted to seduce her. The Wizard's wife, a Witch, wanted to recover her husband with a love potion, but her plan backfires. The King offered a reward to recover the princess, and a Knight and his Squire volunteered. An Elf maiden also wanted to find the Princess to kill her, tricked by the Witch, and she needed a dagger from the Troll. The Squire, the Knight and the Elf, who were in most of the scenes, met halfway the adventure and helped each other, but the conflict arose when the Elf threatened to kill the Princess. The Troll, the King, a group of Elves, and a group of Goblins appeared only briefly. In the end, each character had his/her goal, and each character sub-story crossed at some point of his/her line with some of the other ones. 17 subjects of diverse ages and backgrounds took part in the experiment. They did not know precisely its goal. It was conducted individually in two phases.

The first phase focused on *reading / understanding*. After signing a consent form, the subject received a brief introduction to Proppulsion interface and content. Then, s/he was asked to take as much time as s/he wanted to read, in any desired order. During this phase, we measured the reading time, kept a log of the characters and scenes selected, and mouse-tracked subjects' navigation. At the end of this phase we asked a series of questions discussed later.

The focus of the second phase was *authoring / contributing*. Subjects were offered to freely write more scenes for the story, indicating at which point of the narrative the scene was placed. The time taken for contributing was measured and a shorter questionnaire was asked at the end.

In the first phase, we asked subjects about "perceived easiness of reading", "perceived comprehension" and "perceived enjoyment" through some Likert scaled questions. We also asked the reader some questions to test his/her understanding of the story (such as who was the protagonist/s? or the main plot/s), and his/her method for reading (How did you choose what to read?).

We measured the reader's comprehension quantitatively, borrowing Tanenbaum's strategy [25], where it was tested through a questionnaire after users had read a non-linear story in a partial, non-chronological way. The external author prepared a set of questions on her story asking the subject to relate different events. A panel of judges who had read the story selected a test from them. When answering the test, subjects were allowed to return to read the story. The same panel of judges scored the answers, and we tested the agreement of the judges on the resulting scores by measuring the Cohen-Kappa coefficient of inter-rater reliability [26]. We also measured the time taken to answer those questions, and the time employed to read when answering.

In the second phase, the judges rated the contributions in terms of consistency (if the events fitted with the rest of the story), and the agreement of the judges was also tested. The perceived ease of contribution was measured with a questionnaire using Likert scales too.

V. RESULTS

Two subjects of the 17 took too long to complete the experiment (+ two times the standard deviation) and their results were excluded from further analysis. The time results were normally distributed with a confidence level of 84%. Table I summarizes the quantitative results of both phases.

TABLE I. QUANTITATIVE RESULTS

	Exp. total time (sec.)	**Initial reading time (sec.)**	**Compr. test time (sec.)**	**Reading time during compr. test (sec.)**
Mean /sd	1373.60 / 341.06	457.20 / 102.34	253.67 / 97.75	66.13 sec / 71.38
	Total time contrib. (sec.)	**Time writing (sec.)**	**Time reading when contrib. (sec.)**	**Total reading time (sec.)**
Mean /sd	257.00 / 141.81	328.67 / 242.49	72.33 / 57.43	552.27 / 90.69
	Perc. ease of reading (/4)	**Perceived compr. (/4)**	**Enjoyment (/4)**	**Compr. test result (/4)**
Mean /sd	3.11 / 0.53	3.07 / 0.36	3.49 / 0.49	3.32 /0.39
	Consist. of contrib. (/4)		**Perc. ease of contrib. (/4)**	
Mean /sd	3.67 / 0.30		2.83 / 0.43	

For the two items rated by the panel of judges (comprehension and consistency of contribution), we excluded the judge with the lowest item-total correlation and achieved a moderate agreement in the scores (For the compr. test, percentage of overall agreement Po: 0.583332, free-marginal kappa: 0.444443; for the consist. evaluation Po: 0.619047, Free-marginal kappa: 0.492063).

A. Navigation and story/character perception

People understood the story from a character-centric point of view, and viewed it as a multi-character tale. When asked about the plot, all subjects referred to specific characters and their goals, and 14 out of 15 pointed out that there were multiple stories in one. Plots are regarded as implicit in the character storylines.

When asked about who was/were the main character/s, people chose those characters with long and defined narrative programs. Table II shows that the characters appearing in more scenes are those more often chosen as protagonists by the readers. Characters who do not appear on the table were not mentioned by any subject and appear only in one or two scenes. The number of scenes is not the only factor for relevance. While the Knight and the Squire appear in the same number of scenes, subjects mentioned the Knight twice than the Squire. This could be due to the Knight having a mission, as defined in Proppean terms (a quest giver, the King, gives him a quest, to rescue the Princess, in order to obtain a reward), while the Squire acts as his helper.

TABLE II. CHARACTERS BY MENTIONS , SCENES AND USES

Character	*Mentions as Protagonist*	*Scenes in the Story*	*Times used in Contributions*
Elf maiden	*12*	*7*	*5*
Knight	*10*	*6*	*3*
Squire	*5*	*6*	*0*
Princess	*3*	*4*	*5*
Wizard	*2*	*2*	*2*
Witch	*1*	*3*	*4*

B. *Reading patterns and story comprehension*

The analysis of the logs shows that subjects quickly identified the main (longer, protagonist-based) storylines, focused on reading them linearly, and then backtracked to read the secondary character's stories non-linearly, despite the random order of the icons. In the questionnaires, subjects explained that they liked to read this way: first understand a single story and then read the related characters stories to understand their relationships with the main plot(s).

40% of the subjects selected all the characters, and read all the scenes of the storyline of each character: in the end they read the whole story. The other 60% only chose part of characters; 75% of them read all the scenes of the characters they selected, while the remainder 25% only read some scenes of each character they had selected.

Subjects achieved a high degree of comprehension (3.32 points out of 4). The comprehension of those who read it entirely was slightly better (avg. 3.533, sd. 0.1902) than those who did not (avg. 3.1844, sd. 0.4275) but this difference was not significant (T-Test t(13)=1.8614, p=0.0854). The direct observations seem to indicate that reading one storyline gives enough information about the related storylines to be able to understand them without exhaustive reading.

People taking longer to read at the beginning of the experiment seemed to need less time reading when answering the comprehension tests (Pearson's correl. coef.: 0.4544), while the reading time did not seem related to the comprehension achieved (correl. 0.1311).

C. *Reading impact on contribution*

Half of the subjects contributed to the story. People with better comprehension did not perceive the contribution task as easier, quite the opposite (Correl. -0.5633 between comprehension and perceived ease of contribution); neither did they contribute more quickly than others (the correlation between comprehension and contribution time is a weak 0.2495). It seems that people with higher comprehension are more concerned about the complexity of the story they have to contribute to. On the other hand, those with better comprehension needed to read a lot less when contributing (correl. -0.9094 between comprehension and reading time during contribution).

The judge-rated consistency of contributions was high (3.67 points out of 4). It is quite remarkable that there is a strong correlation (0.8120) between comprehension and consistency of contribution, which seems to indicate that people with better understanding of the story create scenes that fit better with the existing events.

D. *Interest of the contributors*

An analysis of the contributions indicates that there is an interest in expanding the stories of the characters considered "main characters", but the authors also expand the stories of the characters regarded as "secondary" (see Table II).

VI. DISCUSSION

In some way, our proposal relates to the traditional hypertext storytelling, as it challenges the reader to navigate a non-linear story and the author to build its structure. Proppulsion readers interact with the narrative on *interpretative* (understanding the story) and *functional* (manipulating the interface) levels, but not on an *explicit* one as hypertext readers do when their elections alter the story (using the interactivity levels of Salen and Zimmerman [27]).

Pope [28] discusses how hypertext fiction, although still commercially produced (e.g., Storyspace [29]), does not appeal to a wider audience, pointing as problems unsatisfying hyper-linking, random plot structures, and lack of closure, while Berstein [30] described similar problems as lack of coherence, causality, and closure. Pope highlights the interface as an influential factor in reading enjoyment, and fulfilling the reader's expectations to add purposefully to what has already been read.

Unlike this perception of hypertext fiction reading as hard, our experiment revealed that the subjects perceived the non-liner story as easy to read and understand. In consonance with our previous CrossTale experiments [5, 6], following storylines proves useful for reading the nonlinear narrative space. Associating storylines with the character narrative programs resulted in a quite natural way to comprehend the story, with readers characterizing them as having one protagonist accomplishing one goal. The temporal, thematic, and cause-to-effect qualities the Narrative Program seem to be a useful tool to achieve this "meaningfulness" that Pope and others demand for hypertext narrative links.

The reader-perceived "main characters" of the story are those with longer and more defined narrative programs in terms of classic semiotics: characters that receive a mission and follow a series of events to accomplish their goals, finding helpers and opponents on their path. The classical narrative roles still apply to the protagonist perception, but the "multiple-points-of-view" perception prevails: in our story, the elf is an anti-hero character that acts as an antagonist of the knight, since their missions are opposed and she becomes a traitor, but she is regarded as the main protagonist along the knight since she has also a defined goal and takes lots of steps towards its accomplishment. It would be interesting to experiment with different stories combining different characters and roles in unexpected ways, to deepen on this understanding of how readers recognize main characters. People identify those characters quickly, and they use those main storylines as the backbone of their navigation.

Reading all the scenes of all characters, or spending more time reading, were not decisive factors to raise the comprehension level. Comprehension seems to be achieved through the ability to identify key characters and scenes and to understand their relations with other storylines, rather than through an exhaustive processing of all the information in the narrative space. With those key events the reader's mind can establish connections and fill the gaps in the story, as in Tanenmbaum's experiment involving non-linear stories [25]. Then, making explicit the relations between characters in each part of the story (i.e., their roles in the main character narrative program) empowers the subjects understanding.

Subjects with higher levels of comprehension needed less time for contributing and achieved higher levels of consistency with the previous story, which is consistent with psychological studies on the effect of comprehension in authoring tasks [7] and reinforces our hypothesis that, in a collaborative context, enhancing the comprehension of the readers will enhance their ability to contribute.

Non-linearity seems to encourage expanding the story from different character's points-of-view. Although the "main characters" are used regularly, people also expand the stories of characters regarded as "secondary". We hypothesize that those "secondary" characters can become "main characters" for the future readers, encouraging participation. Berstein's Thespis [30] proposed a theatre-inspired system in which each author acts as an autonomous character. Some multi-user tangible interfaces [11] also take this approximation, each author developing one character in the story. Our proposal differs in that any number of writers can develop any number of characters, but in this experiment, as in previous ones [5], it seems that it is usual to concentrate on one storyline at a time.

Finally, compared with Crosstale [5, 6], the proportion of subjects who became contributors after reading was smaller. The experiment demanded subjects to complete a long series of tests after reading and this might have disrupted a possible creative task. Also, Crosstale presented a visual scene editor that might have made the contribution task more appealing.

VII. CONCLUSIONS AND FUTURE WORK

This model to represent and interact with non-linear stories based on the classic semiotics concept of Narrative Programs, focused on human authoring, represents a quite different approach from most current models based on semiotics, which are oriented towards automatic generation, although it shares some aspects of those which pay attention to readership.

The resulting exploration and development of multiple point-of-view storylines within a larger narrative space resounds with traditional hypertext fiction, plagued with reading issues. The experiment with our small prototype shows that we seem to have avoided the issues, with pleasurable reading and proficient comprehension, based on reading through connected storylines. This understanding led to contributions with a good level of consistency, featuring largely, but not exclusively, the main characters.

We intend to use larger narrative spaces to determine how comprehension (and engagement) scales as a massively-authored narrative grows, and whether contributions preserve consistency and the overall meaning. We also intend to see which non-obtrusive support can be automatically provided to authors, in addition to more visual means of contributing.

REFERENCES

[1] CA. Scolari, "Transmedia storytelling: Implicit consumers, narrative worlds, and branding in contemporary media production", International Journal of Communication, vol. 3, 2009, pp. 586-606.

[2] "FanFiction", http://fanfiction.net, Web, retrieved: 9 Feb. 2015.

[3] S. Chatman, "Story and Discourse: Narrative Structure in Fiction and Film", Cornwell University Press, 1990.

[4] J. Zimmerman, J. Forlizzi, S. Evenson, "Research through design as a method for interaction design research in HCI", Proceedings of the SIGCHI conference on Human factors in computing systems, 2007, pp. 493-502.

[5] J. Colàs, A. Tapscott, A. Moghnieh, J. Blat, "Shared Narratives as a New Interactive Medium: CrossTale as a prototype for Collaborative Storytelling", International Journal On Advances in Telecommunications, vol. 6, no 1 and 2, 2013, pp. 12-23.

[6] A. Tapscott, J. Colàs, A. Moghnieh, J. Blat, "Writing Consistent Stories based on Structured Multi-Authored Narrative Spaces", OASIcs-OpenAccess Series in Informatics, vol. 32, Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2013, pp. 277-292.

[7] K. Cain, "Text comprehension and its relation to coherence and cohesion in children's fictional narratives", British Journal of Developmental Psychology, vol. 21, no 3, 2003, pp. 335-351.

[8] P. David, and M. Cavazza, "From debugging to authoring: Adapting productivity tools to narrative content description", Interactive Storytelling, Springer Berlin Heidelberg, 2008, pp. 285-296.

[9] S. Ivo, and M. Theune, "Iterative authoring using story generation feedback: debugging or co-creation?", Interactive Storytelling, Springer Berlin Heidelberg, 2009, pp. 62-73.

[10] K. Ryokai, and J. Cassell, "StoryMat: a play space with narrative memories", Proceedings of the 4th international conference on Intelligent user interfaces, ACM, 1998, p. 201.

[11] M. Theune, T. Alofs, J. Linssen, I. Swartjes, "Having one's cake and eating it too: coherence of children's emergent narratives", OASIcs-OpenAccess Series in Informatics, vol. 32, Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2013, pp. 293-309.

[12] S. Gobel, L. Salvatore, and R. Konrad, "StoryTec: A digital storytelling platform for the authoring and experiencing of interactive and non-linear stories", Automated solutions for Cross Media Content and Multi-channel Distribution, AXMEDIS'08 , IEEE, 2008, pp. 103-110.

[13] U. Spierling, S. A. Weiß, W. Müller, "Towards accessible authoring tools for interactive storytelling", Technologies for Interactive Digital Storytelling and Entertainment, Springer Berlin Heidelberg, 2006, pp. 169-180.

[14] K. Howland, J. Good, B. du Boulay, "Narrative threads: A tool to support young people in creating their own narrative-based computer games", Transactions on Edutainment X, Springer Berlin Heidelberg, 2013, pp. 122-145.

[15] P. Likarish, and J. Winet, "Exquisite Corpse 2.0: qualitative analysis of a community-based fiction project", Proceedings of the Designing Interactive Systems Conference, ACM, 2012, pp. 564-567.

[16] V. Propp, "Morphology of the Folktale", University of Texas Press, 1968.

[17] A. J. Greimas, "Structural semantics: An attempt at a method", Lincoln: University of Nebraska Press, 1983.

[18] P. Bailey, "Searching for storiness: Story-generation from a reader's perspective", Working notes of the Narrative Intelligence Symposium, 1999, pp. 157-164.

[19] M. O. Riedl, and R. M. Young, "Character-focused narrative generation for execution in virtual worlds", Virtual Storytelling. Using Virtual Reality Technologies for Storytelling , Springer Berlin Heidelberg, 2003, pp. 47-56.

[20] M. Mateas, and P. Sengers, "Narrative intelligence", Proceedings AAAI Fall Symposium on Narrative Intelligence, 1999 , pp. 1-10.

[21] M. O. Cavazza, F. Charles, S. J. Mead, " Character-based interactive storytelling", IEEE Intelligent systems, 2002.

[22] P. Gervás, “Propp's Morphology of the Folk Tale as a Grammar for Generation”, OASIcs-OpenAccess Series in Informatics, vol. 32, Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2013, pp. 106-122.

[23] M. Mateas, and A. Stern, “Façade: An experiment in building a fully-realized interactive drama”, Game Developers Conference, 2003, pp. 4-8.

[24] K. Hartmann, S. Hartmann, M. Feustel, “Motif definition and classification to structure non-linear plots and to control the narrative flow in interactive dramas”, Virtual Storytelling. Using Virtual Reality Technologies for Storytelling, Springer Berlin Heidelberg, 2005, pp. 158-167.

[25] J. Tanenbaum, K. Tanenbaum, M. S. El-Nasr, M. Hatala, “Authoring tangible interactive narratives using cognitive hyperlinks”, Proceedings of the Intelligent Narrative Technologies III Workshop, ACM, 2010, p. 6.

[26] J. Cohen, “Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit”, Psychological bulletin, vol. 70, no 4, 1968, p. 213.

[27] K. Salen, and E. Zimmerman, “Rules of play: Game design fundamentals”, MIT press, 2004.

[28] J. Pope, "A Future for Hypertext Fiction", Convergence: The International Journal of Research into New Media Technologies , vol. 12, no 4, 2006, pp. 447-465.

[29] “Eastgate Storyspace”, http://www.eastgate.com/storyspace, Eastgate Systems Inc., Web, retrieved: 9 Feb. 2015.

[30] M. Bernstein, “Card shark and thespis: exotic tools for hypertext narrative”, Proceedings of the 12th ACM conference on Hypertext and Hypermedia, 2001, pp. 41-50.

On the Estimation-Based Closed-Loop Power Consumption Control in Multimedia Mobile Devices

Qiong Tang, Ángel M. Groba, Eduardo Juárez and César Sanz
Centro de Investigación en Tecnologías del Software y Sistemas Multimedia para la Sostenibilidad (CITSEM)
Universidad Politécnica de Madrid (UPM)
Madrid, Spain
e-mail: qiong.tang@alumnos.upm.es; {angelmanuel.groba, eduardo.juarez, cesar.sanz}@upm.es

***Abstract*—In this paper, a closed-loop approach to control the power consumption of multimedia mobile devices is presented, such that the feedback signal is an estimation based on monitored system events. First, the power estimation method is presented and validated. Afterwards, prior to the implementation of the real-time control system, off-line estimation data are used to get a system model, which enables the application of classic control-theory methods to analyze and design an integral controller whose behavior is then simulated. The target system is a video decoder running in an embedded development platform. The simulation results show how the controller achieves null average steady-state error with short settling times, even in the presence of estimation noise or disturbance, thus predicting promising results for the closed-loop approach to the final real-time system implementation.**

Keywords—multimedia; embedded; power estimation; modeling; PMC; DVFS; closed-loop control.

I. INTRODUCTION

Embedded and mobile multimedia systems require, like others, the optimization of the quality of experience (QoE) they offer to the user. However, their common battery dependency makes also necessary the optimization of their energy consumption. Indeed, for example, the wide spectrum of usual available applications for current smart phones make them to have quite limited operating times, especially when they execute common video encoding, decoding and/or presentation applications. Furthermore, the increasing complexity of emerging video standards, such as High Efficiency Video Coding (HEVC) [1], will probably increase this limitation with respect to other previous ones, like H.264/AVC [2]. Therefore, there is an increasing effort into trying to reduce the energy consumption of this kind of systems from different points of view. Particularly, we are interested in optimizing their energy consumption in relation with applications of video decoding, obviously keeping a reasonable trade-off with QoE. Although our research group has been already working on these issues since several years ago [3][4][5], we are starting now a new research branch based on a less heuristic and more systematic approach, whose validity and efficiency is wanted to be tested.

As one of the first-stage results of this new approach, in this paper, we present the formal application of classic closed-loop control techniques to the power-consumption regulation of a video decoding application running in an embedded multimedia platform. This will be based on power estimations from available system indicators in order to avoid the need of a power monitor subsystem.

For this purpose, in general terms, the system should be modeled as a real-time closed-loop control system (see Figure 1), in which the *controlled output* follows the *target* signal (often called *set-point* in the Control jargon). This is achieved by a *controller*, which processes the system *error* between the target and the *feedback* information coming from a sensor and generates the *action* signal to the device under control (often called *plant* in the Control jargon). In a typical industrial process control, the plant is normally designed to be controlled in this way, so it usually offers action inputs able to vary the plant outputs and even sensors for feeding the output values back. In our case, we face the previous problem of adapting our plant (the embedded multimedia platform) to this topology, because it is not initially thought to be controlled in this way. Therefore, the first task is to identify and set up both an action and a feedback signal in the plant. For the former, we have identified and used the dynamic voltage and frequency scaling (DVFS) mechanism, present in many commercially available platforms and able to act on the system consumption by varying the operating performance point (OPP) of the microprocessor unit (MPU). For the later, what we have identified is a lack of direct consumption sensors in the majority of present commercial embedded platforms. Hence, we have decided to adopt an intermediate solution, which is to estimate the power consumption from other available signals in the plant, i.e., event counts. This leads to a structure like the one shown in Figure 1, which decouples the consumption optimization infrastructure from specific instrumentation needed to monitor actual power consumption, thus increasing the platform autonomy and the control-system applicability.

The power consumption estimation we present is based on previous work [3][4] in which static energy estimations are mathematically calculated for video decoding tasks for a fixed OPP by off-line correlation between actual energy measurements and significant-events counts taken from the processor performance monitoring counters (PMCs). Now, this estimation method is extended to a system in which the OPP is variable. In a next step, the estimations will be periodically calculated in real time and the estimator will feed consumption samples back to the controller, which, in turn, will drive the DVFS system to set the suitable OPP.

ISBN: 978-1-61208-400-8

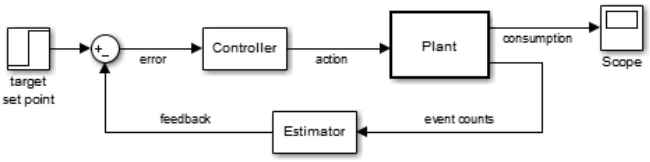


Figure 1. Diagram of a closed-loop consumption control system based on estimation feedback.

But before the control system is finally implemented to work in real time, a previous work of modeling and analysis is done to design and validate a first controller based on simulation results. This is the work presented in this paper, with the following structure: in Section II previous related work is presented; Section III describes the test bed used; in Section IV the power estimation process is sketched; in Section V classical control theory is applied to model the system, design the controller and get a system simulator; Section VI outlines the main results; and Section VII presents conclusion and future work.

II. Related Work

Given the increasing concern about saving energy whenever is possible, a great number of research developments can be found related to energy consumption optimization in microprocessor-based systems. They range from small battery-operated systems [4][6] to larger data centers or web servers [7][8], focusing also in multimedia applications [3][5][9]. Furthermore, the DVFS method is being used to act on the system consumption since a number of years ago [10][11][12]. On the other hand, the application of closed-loop techniques also appears in the literature of all these fields [6][7][9], also with widespread use of DVFS. However, among this group of solutions, there is not a clear proposal of how to feed back the closed-loop system, mainly because there is not an obvious feedback signal available in most conventional platforms, as mentioned above.

For example, in [13], [14] and [15], the controlled variable is the processor utilization factor (U), which is modulated through the DVFS system by means of P [13], PI [14] and PID [15] controllers. The energy savings increase as U approaches 100%, while meeting the task deadlines.

Another set of approaches are found in which the controlled variable is the occupancy level of certain system queues, given that keeping it constant implies that just the needed energy is being consumed. A couple of examples based also on DVFS are [16], in which authors explore the benefits of a control architecture on the throughput and energy consumption of a video encoder executed on a multiple-processor system on chip (MPSoC); and [17], where a nonlinear controller is used in queue-based streaming applications.

Also, there are approaches in which the control loop adapts the OPP to the just needed frequency by estimating the processor workload, like [18], where a Kalman filter estimates the computation time needed by MPEG-2 decoded frames.

In some specific cases, the target system includes a power monitor unit able to feed actual consumption data back to the closed-loop controller, as in [19] for a chip multiprocessor with an optimal controller. However, our aim is to reach a control system which can regulate the power consumption of an embedded multimedia system without the need of added power monitors but basing it on power estimations derived from commonly available information. It is an approach similar to the one used in [20], where the use of PMCs is proposed for estimating L2 cache consumption, but this is extended in our case to the processor and based on its DVFS mechanism.

Combining PMCs with multivariate adaptive regression splines (MARS) method to build an energy model has been successfully used in different cases [21]-[24]. Now, we apply this methodology to control video decoder consumption in closed-loop.

III. Test Bed

Our test bed is based on a single-core hardware development platform for multimedia embedded systems: BeagleBoard [25]. It features, apart from a number of peripherals, 2Gb NAND and 2Gb SDRAM of memory and an OMAP3530 processor system [26]. This system includes a MPU based on an up to 720-MHz ARM Cortex-A8 core, a digital signal processor (DSP) core and other coprocessors. The Cortex-A8 architecture includes 4 PMCs and one specific counter for CPU clock cycles. Related to the BeagleBoard peripherals, it is worth mentioning the possibility a subset of them offers for changing the MPU supply voltage and clock frequency (DVFS subsystem).

This development platform allows us to execute video decoding applications while monitoring their power consumption in order to tune the power estimator. An Agilent set of programmable power supply, battery emulator and PC-based GPIB-linked acquisition system [27] has been used to supply the board and acquire records of its current consumption (see Figure 2). To simplify the work and focus on the energy consumption caused by the MPU, the memory subsystem and the related I/O buses, the board has been configured as a minimal system that disables the unnecessary components.

With respect to the software part, a Linux 3.8.0 kernel, patched to support the platform DVFS mechanism, is running in the processor. On the other hand, taking advantage of our expertise in the Open RVC-CAL Compiler (Orcc) [3], a MPEG4-Part2 decoder is built from [28] to be used as the power-consuming video application. This user-level application is suitably modified to include the estimation module (see Figure 2). Besides, several video sequences [29], accessed through the platform SD card, are used to test the system.

The DVFS subsystem is managed through the *cpufreq* Linux driver. This driver includes four predefined governors to fix the MPU OPP, two static and two dynamic, which react to the system load. This is achieved by a function called *cpufreq_driver_target*, one of whose input parameters is the target frequency of the desired OPP to switch to. This function searches the target frequency among the ones of the OPPs defined in an internal table and selects the appropriate one by applying a ceil- or a floor-rounding algorithm, depending on another input parameter. The function then sets the frequency and the voltage corresponding to the selected OPP. The default *cpufreq* definitions for the BeagleBoard only consider 6 OPPs.

In order to decrease this strong nonlinearity in the DVFS-based plant input, additional valid OPPs were searched, verified and included into the *cpufreq* table, reaching a total of 27.

In this first stage in which the power estimator has to be validated prior to the final control-system implementation, a third-party tool, Performance Application Programming Interface (PAPI) [30], is used to access PMCs from the application level, offering both a high- and a low-level interface. The high-level interface includes start, stop and read sets of PMCs and other simple operations which can obtain accurate measurements of applications. The fully programmable low-level interface provides the possibility to control the counters. PAPI has been implemented on a number of Linux platforms and the latest release now provides support for ARM Cortex A8. In our work, PAPI is employed as the PMC driver, which is used from the application to take PMC event samples after decoding every video frame.

During decoding, the application sends signals to a modified *cpufreq* governor into the operating-system kernel to set the OPP, thus achieving test-oriented OPP changes, such as increase, decrease or jump among OPP table values. Meanwhile, the application also selects the suitable power estimation model depending on the active OPP. After that, the estimation procedure calculates a power estimation sample and then writes it into a SD-card file for off-line validation purposes. Figure 2 shows an explanatory block diagram of the test bed.

IV. Power Consumption Estimation

Our work is based on that presented in [3] and [4], where static estimations of CPU and memory energy consumption are mathematically calculated following the MARS methodology while executing video decoding tasks in a non-DVFS scenario. This is achieved by correlating actual energy measurements with significant-events counts taken from the processor PMCs. These previous approaches are enhanced in this paper to work in a scenario in which the OPP is variable and power estimation samples are successively required in execution time.

The first phase of the estimator implementation leads to identify the set of events which are most significant with respect to the power estimation. This is achieved by a filtering procedure. Thus, the Spearman's rank correlation ρSi is computed between each event and power consumption. After this step, a threshold α is established to eliminate any events below this threshold. On the other hand, to reduce event redundancy, correlations ρ(i,j) between each pair of events 'i' and 'j' are computed to identify the event relationship. The purpose is to eliminate those events whose information can be obtained from other events. Hence, starting from the event 'a' with the largest correlation ρSa, those events 'j' whose correlation ρ(a,j) exceeds certain threshold β are eliminated. Then, the procedure continues with event 'b', with correlation value ρSb, to eliminate the events 'j' whose ρ(b,j) exceeds β. This process is repeated until there are no more events to eliminate. Finally, the remaining events are highly related with power consumption and orthogonal among them. In this work, α is set to 0.5 and β is set to 0.9. TABLE I lists the events resulting from the filter procedure.

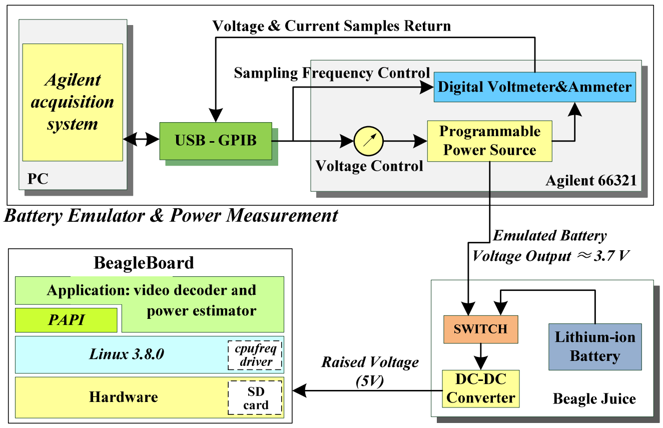

Voltage & Current Samples Return
Agilent acquisition system
PC
Sampling Frequency Control
Digital Voltmeter&Ammeter
USB - GPIB
Programmable Power Source
Voltage Control
Agilent 66321
Battery Emulator & Power Measurement
Emulated Battery Voltage Output ≈ 3.7 V
BeagleBoard
Application: video decoder and power estimator
PAPI
Linux 3.8.0
cpufreq driver
Hardware
SD card
Raised Voltage (5V)
SWITCH
Lithium-ion Battery
DC-DC Converter
Beagle Juice

Figure 2. Block diagram of the test bed.

Once the list of significant events is obtained, the MARS method is applied to get an estimator able to estimate power consumption from event counts read from PMCs. Combining PMCs with MARS method to build an energy model has been successfully used in different cases [21]-[24]. Previous work proves that this methodology is also suitable for video decoders [22]. In our work, power estimation models have been tuned by combining 78 video sequences belonging to two resolution groups (CIF and QCIF) and by off-line correlation between event counts and power measurements. Since the measurement system captures current values of the BeagleBoard supply, the power is calculated multiplying them by the board supply voltage. Anyway, the acquired current signal is filtered to remove sporadic consumption spikes not due to the video decoding task itself. Models have been tested in this way with all the available test sequences against 27 OPPs and approximately 95% of the models obtained from a combination of training sequences have an average relative estimation error less than 5%.

Once implemented the estimator as explained above and in order to model the plant for designing the closed-loop controller, a subset of 8 sequences has been used to average the estimated power consumption for each OPP. TABLE II shows average values of estimated power, measured power consumption and relative error between them for all OPPs. The total mean relative estimation error is -0.37%.

V. System Modeling, Analysis and Design

One means of designing the system controller is to base it on a suitable model of the plant. As a first approach to the problem, a simplified model is used to facilitate the application of the classic control theory. Later, those models could be refined and sophisticated and different advanced closed-loop control strategies could be applied.

The records of consumption estimation used to validate the estimator module are also useful for modeling purposes. For example, if the estimator is considered to be included into the plant itself, the analysis of how it responds to a change in the OPP enables the plant dynamics modeling. Thus, for example, Figure 3 shows the estimator output for the OPP changing from number 26 to number 27 during a certain video sequence decoding. Since the estimation period is the video-frame

TABLE I. SELECTED EVENTS AND FUNCTIONALITY

PAPI events	Description
L2_TCM	Level2 total cache misses
TLB_IM	Instruction translation look aside buffer misses
BR_TKN	Conditional branch instruction taken
SR_INS	Store instructions executed
TOT_CYC	Total cycles

TABLE II. POWER ESTIMATION DATA FOR ALL OPPS

No.[a]	C[b]	E[c]	e[d]
1	0.930	0.926	-0.34
2	0.985	0.982	-0.26
3	0.995	0.990	-0.46
4	1.005	1.005	0.05
5	1.025	1.020	-0.46
6	1.035	1.031	-0.40
7	1.055	1.046	-0.78
8	1.075	1.074	-0.05
9	1.095	1.092	-0.20
10	1.120	1.112	-0.70
11	1.140	1.137	-0.20
12	1.165	1.160	-0.36
13	1.180	1.173	-0.60
14	1.205	1.197	-0.63
15	1.225	1.222	-0.25
16	1.295	1.288	-0.47
17	1.315	1.313	-0.10
18	1.330	1.326	-0.26
19	1.350	1.347	-0.16
20	1.370	1.361	-0.66
21	1.389	1.381	-0.60
22	1.409	1.400	-0.70
23	1.439	1.438	-0.13
24	1.449	1.447	-0.16
25	1.484	1.479	-0.38
26	1.504	1.502	-0.16
27	1.634	1.625	-0.58

[a] OPP number; [b] Average consumption (W)
[c] Average estimation (W); [d] Relative error (%)

decoding period, it is long enough as to allow the estimator to complete its OPP switch from one sample to the next, as it can be seen at t=260s in Figure 3. Hence, a simplified discrete transfer function model of the plant could be G(z)=1/z [31], which relates the power estimation with OPP average power level. This simple model is valid while the system sample period is much longer than the settling time of the analog power consumption process or, correspondingly, than the time resolution of the PMCs used to estimate the power. The frame period applied to the estimations for open-loop tuning and modeling purposes is obviously long enough. Furthermore, although the sample period to be used in the real-time closed-loop control system will be shorter than the frame period, it will still be much longer than consumption settling time, otherwise the system overhead would be unbearable.

Thinking on implementing a closed-loop automatic power regulation system, its stability is one of the characteristics that must be ensured apart from other technological issues. As a first approach, the simplest controller that can be used in closed loop is a P one [32]. If we call K the gain of this controller, the transfer function of the closed-loop system is M(z)=K/(z+K). Therefore, the critical gain which leads the system to instability is K_c=1. Then, the lower bound for the closed-loop system error in steady state is min(e_{ss})=100/(1+ K_c)=50% [33], which is too high. In order to avoid this limitation and still keeping a classic linear controller, an integral action can be added to it.

As a first and simple approach to the integral action, one can choose between a forward and a backward rectangular rule (FRR and BRR, respectively) [31]. Their corresponding Z transfer functions differ only on a zero in z=0, which appears in the second case. Hence, the zero-pole cancellation in a series of a BRR I controller and G(z) will enable shorter settling times than with a FRR I because the system dominant pole can be closer to z=0.

Thus, considering the BRR option, the closed-loop pole of the system is p_{CL}=1-KT, being again K the controller gain. Let us consider a sample period T of 100ms, which seems to be a good trade-off value for keeping reasonable overhead, immunity to jitter effects and frequency of control actions. For this period, the critical gain which leads the system to instability (p_{CL}=-1) is K_c=20, whereas the gain for the shortest settling time (p_{CL}=0) is K=10. This is the gain used for the I controller, which can be seen included into the *Simulink* block diagram of Figure 4.

The block in Figure 4 between the controller and the plant is modeling the nonlinearity implied by the plant interface, which only admits 27 different levels, i.e., the 27 available OPPs. It is modeled as a quantization process, whose steps are defined by the average power estimation values shown in TABLE II. In order to limit the maximum quantization error to ±step/2, the breakpoints are set in the midpoint between estimation levels. Since, for simplifying purposes, the plant model expects power levels as input values, the final implementation will include a module which translates each of the 27 average power estimation values into its corresponding OPP frequency as the input parameter for the *cpufreq* interface.

The closed-loop diagram of Figure 4 also includes a disturbance input in order to test the capability of the system to react to disturbances on its controlled output. The disturbance input would simulate the effect of a consumption variation when the system is following the set point in steady state, due, for example, to a variation in the processor load.

The disturbance input of Figure 4 can also be used to inject real *estimation noise* into the ideal signals of the *Simulink* model. For example, the third column of TABLE I represents average values of power estimation for each OPP but the actual estimation values do fluctuate around those averages, as can be distinguished in Figure 3. The estimation signal can be modeled in each OPP as a constant (its corresponding average value) plus a random-like "noise" with a maximum peak close to 9mW and zero mean. Figure 5 shows an example of this noise.

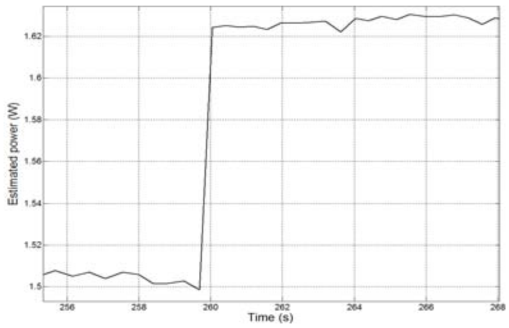


Figure 3. Estimated consumption for an OPP26 to OPP27 step.

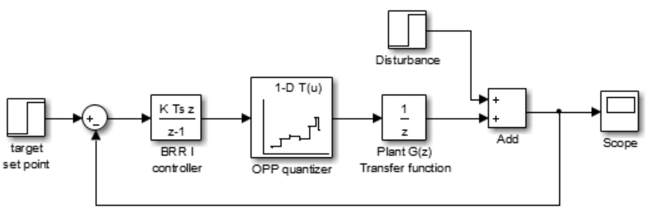


Figure 4. Diagram of the system simulator with disturbance input.

VI. RESULTS

The system simulator of Figure 4 has been tested for a number of set points and disturbances, mainly step shaped. It is worth noting that a step-shaped input would simulate a constant power desired for the system consumption (set point) or a sudden change in the system power consumption (disturbance). Therefore, the following results must be understood within a scenario in which the system is continuously consuming power, for instance in a video-streaming application, which in turn is desired to be constant, regardless now of energy or QoE issues.

As a summary, Figure 6 shows the system time response for a set-point step from OPP1 level to OPP15 level in t=0 and a disturbance step (undesirable and unexpected increase of consumption) of a 40% of the input step in t=0.5s. In that figure, it can be seen, on one hand, how if the OPP is changed at t=0, the power estimator would not reflect the corresponding consumption change until next sample period (t=0.1s). On the other hand, with the system consumption stable at OPP15 level, the power estimator outputs a sudden consumption increase at t=0.5s, which keeps until the next sample time at t=0.6s. At that moment, the I controller detects the anomaly, i.e., a power consumption higher than desired, and corrects it immediately by decreasing its output to the plant (i.e., by setting a lower OPP). However, since the disturbance value implies that there is not any OPP which cancels exactly its effect, i.e., none OPP applied to the plant reaches a consumption equal to the target, the I controller makes the response oscillate. This oscillation in the system power consumption can be seen, on one hand, as an undesirable behavior of the system, given that the set point is not oscillating, or, on the other hand, as the only way the nonlinear control system can satisfy the set-point requirement, in average.

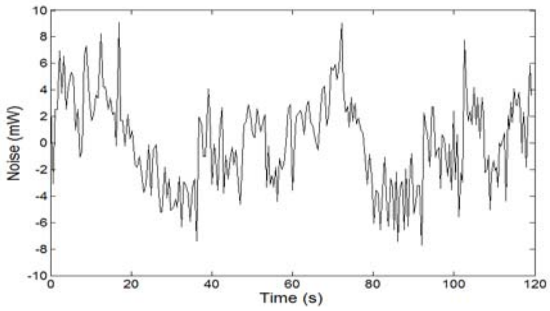


Figure 5. Simulated noise to be added to the estimation signal.

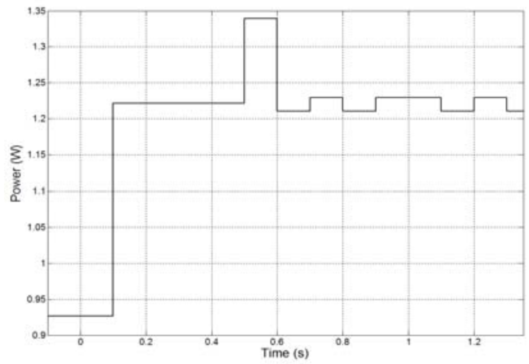


Figure 6. Closed-loop time reponse for an OPP1 to OPP15 input step and disturbance of 40% at t=0.5s.

If the disturbance input of Figure 4 is used to inject the characteristic noise of the power estimation (see Figure 5) and we let enough response time, the system output for an example set point equivalent to OPP15 level is the one shown in Figure 7. In that figure, it can be seen how the system reaches a consumption with the average value of OPP15 (see TABLE II), as desired, but some glitches appear occasionally. What happens is that, due to the quantized input to the plant, the noise deviations cannot be corrected until the accumulated error in the I controller reaches an OPP breakpoint threshold. In that moment, perhaps the minimum action change, i.e., the next OPP level, is greater than needed, thus triggering the aforementioned glitches, whose levels correspond to the neighboring OPP values. This is one of the disadvantages of the I controller, which should be solved or minimized in the implemented system by applying specific corrective techniques or other types of controllers.

VII. CONCLUSION AND FUTURE WORK

A power consumption estimator, which estimates the power consumption of a video decoding application running in a commercial embedded development platform has been used to get a mathematical model of the consumption process. The estimator, which is based on PMC values and is sensitive to the active OPP, could be deployed in a wide range of systems without needing specific power monitoring hardware. From the mathematical model, classic analysis and design techniques have been applied to get a suitable controller able to keep track of the power consumption in closed loop. Prior to the implementation of the real-time control system, simulation results have proved the system stability and average regulation of power consumption according to the set point and in the presence of consumption variations and estimation noise. This paves the way for applying such closed-loop techniques to optimize the power consumption of multimedia hand-held devices.

From now on, the real-time control system has to be implemented by feeding power estimations back. Then, its response will be contrasted with the simulation results presented in this paper. This will open the door to further

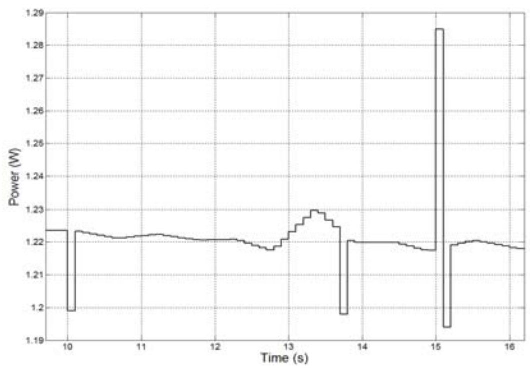


Figure 7. Closed-loop reponse for OPP15 target and noise disturbance.

improvements, such as the test of different controllers and design methods, or the suitable programming of a variable set point to achieve different objectives involving battery life time, QoE or performance parameters, among others, all related to the power optimization problem. Also, the test bed can be moved from uniprocessor to multiprocessor platforms, managing issues like load distribution or multiple power sinks.

ACKNOWLEDGMENT

This work has been supported by the Spanish Ministry of Economy and Competitiveness under grant TEC2013-48453-C02-2-R. Authors wish also to acknowledge the support of the UPM Department of Telematic Engineering and Electronics.

REFERENCES

[1] G. J. Sullivan, J. R. Ohm, W. J. Han, and T. Wiegand, "Overview of the High Efficiency Video Coding (HEVC) Standard," IEEE Trans. on Circuits and Systems for Video Technology, vol. 22, issue 12, Dec. 2012, pp.1649-1668.

[2] ITU-T and ISO/IEC JTC 1, “Advanced Video Coding for Generic Audiovisual Services,” ITU-T Rec. H.264 & ISO/IEC 14496-10.

[3] R. Ren et al., “A PMC-driven methodology for energy estimation in RVC-CAL video codec specifications,” Signal Processing: Image Communication, vol. 28, issue 10, Nov. 2013, pp. 1303–1314.

[4] R. Ren, E. Juárez, F. Pescador, and C. Sanz, “A stable high-level energy estimation methodology for battery-powered embedded systems,” Procs. of 16th IEEE Intl. Symposium on Consumer Electronics (ISCE 2012), June 2012, pp. 1-3.

[5] E. Juárez, F. Pescador, P. Lobo, A. Groba, and C. Sanz, “Distortion-energy analysis of an OMAP-based H.264/SVC decoder,” Mobile Multimedia Communications, Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, Vol. 77, 2012, pp 544-559.

[6] D. Le and H. Wang, “An effective feedback-driven approach for energy saving in battery powered systems,” Procs. of 18th IEEE Intl. Workshop on Quality of Service, June 2010, pp. 1-9.

[7] X. Wang and Y. Wang, “Co-Con: coordinated control of power and application performance for virtualized server clusters,” Procs. of Intl. Workshop on Quality of Service, 2009, pp. 1-9.

[8] V. Petrucci, E. V. Carrera, O. Loques, J. C. B. Leite and D. Mossé, “Optimized management of power and performance for virtualized heterogeneous server clusters,” Procs. of Intl. Symp. on Cluster, Cloud and Grid Computing, 2011, pp. 23-32.

[9] G. Cao, A. Ravindran, S. Kamalasadan, B. Joshi and A. Mukherjee, “A cross-stack predictive control framework for multimedia applications,” Procs. of Intl. Symposium on Multimedia, 2013, pp. 403-404.

[10] D. C. Snowdon, S. M. Petters, and G. Heiser, “Accurate on-line prediction of processor and memory energy usage under voltage scaling,” Procs. of 7th Intl. Conf. on Embedded Software, Sep. 2007, pp. 84-93.

[11] M. E. Salehi, M. Samadi, M. Najibi and A. Afzali-Kusha,, “Dynamic voltage and frequency scheduling for embedded processors considering power/performance tradeoffs,” IEEE Transactions on Very Large Scale Integration Systems, vol. 19, issue 10, Aug. 2010, pp. 1931–1935.

[12] X. Lin, Y. Wang, Q. Xie and M. Pedram, “Task scheduling with dynamic voltage and frequency scaling for energy minimization in the mobile cloud computing environment,” IEEE Transactions on Services Computing, in press.

[13] X. Wang, X. Fu, X. Liu and Z. Gu, “PAUC: power-aware utilization control in distributed real-time systems,” IEEE Trans. on Industrial Informatics, vol. 6, no. 3, Aug. 2010, pp. 302–315.

[14] A. S. Ahmadian, M. Hosseingholi, and A. Ejlali, “A control-theoretic energy management for fault-tolerant hard real-time systems,” Procs. of IEEE International Conference on Computer Design (ICCD), Oct. 2010, pp. 173-178.

[15] A.K. Mishra, S. Srikantaiah, M. Kandemir and C.R. Das, “CPM in CMPs: coordinated power management in chip-multiprocessors,” Procs. of Intl. Conf. for High Performance Computing, Networking, Storage and Analysis, 2010, pp. 1-12.

[16] S. Garg, D. Marculescu and R. Marculescu, “Custom feedback control: enabling truly scalable on-chip power management for MPSoCs,” Procs. of International Symposium on Low-Power Electronics and Design, 2010, pp. 425 - 430.

[17] A. Alimonda, S. Carta, A. Acquaviva, A. Pisano, and L. Benini, “A feedback-based approach to DVFS in data-flow applications,” IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol. 28 , issue 11, Nov. 2009, pp. 1691–1704.

[18] S. Y. Bang, K. Bang, S. Yoon, and E. Y. Chung, “Run-time adaptive workload estimation for dynamic voltage scaling,” IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems, vol. 28, issue 9, Sep. 2009, pp. 1334–1347.

[19] Y. Wang, K. Ma, and X. Wang, “Temperature-constrained power control for chip multiprocessors with online model estimation,” Procs. of 36th International Symposium on Computer Architecture (ISCA), June 2009, pp. 314-324.

[20] X. Wang, K. Ma, and Y. Wang, “Achieving fair or differentiated cache sharing in power-constrained chip multiprocessors,” Procs. of 39th Intl. Conf. on Parallel Processing, 2010, pp. 1-10.

[21] B. Goel et al., “Portable, scalable, per-core power estimation for intelligent resource management,” Procs. of International Green Computing Conference, Aug. 2010, pp. 135-146.

[22] Y. Xiao et al., “A system-level model for runtime power estimation on mobile devices,” Procs. of IEEE/ACM Intl. Conf. on Green Computing and Communications, 2010, pp. 27-34.

[23] K. Singh, M. Bhadauria, and S. A. McKee, “Real time power estimation and thread scheduling via performance counters”, ACM SIGARCH Computer Architecture News, vol. 37, issue 2, July 2009, pp. 46-55.

[24] C. Lively et al., “Power-aware predictive models of hybrid (MPI/OpenMP) scientific applications on multicore systems”, Computer Science-Research and Development, vol. 27, issue 4, Nov. 2012, pp. 245-253.

[25] BeagleBoard.org Foundation. *BeagleBoard*. [Online]. Available from: http://beagleboard.org/beagleboard, 2015.02.20.

[26] Texas Instruments. *OMAP3530 Applications Processor*. [Online]. Available from: http://www.ti.com/product/omap3530, 2015.02.20.

[27] Agilent Technologies. *Agilent 14565B Software and 66319B/D and 66321B/D Mobile Communications DC Sources*. [Online]. Available from: http://cp.literature.agilent.com/litweb/pdf/5990-3503EN.pdf, 2015.02.20.

[28] GitHub ORCC. *Orc-apps, Mpeg4 Part2*. [Online]. Available from:https://github.com/orcc/orc-apps/tree/master/RVC/src/org /sc29/wg11/mpeg4/part2, 2015/02/23.

[29] SourceForge. *Open RVC-CAL Compiler, sequences*. [Online]. Available from: http://sourceforge.net/projects/orcc/files /Sequences, 2015.02.23.

[30] PAPI. *Performance Application Programming Interface*. [Online]. Available from: http://icl.cs.utk.edu/papi/, 2015.02.23.

[31] G. F. Franklin and J. D. Powell, “Digital control of dynamic systems,” 3rd edition, Addison Wesley, 1997.

[32] C. L. Phillips and J. M. Parr, “Feedback control systems,” 5th edition, Prentice Hall, 2010.

[33] K. Ogata, “Modern control engineering,” 5th ed., P. Hall, 2010.