Hand Gesture Recognition Using SIFT Features on Depth Image

Hasan Mahmud, Md. Kamrul Hasan, Abdullah-Al-Tariq, M. A. Mottalib
Systems and Software Lab (SSL), Department of Computer Science and Engineering
Islamic University of Technology (IUT)
Dhaka, Bangladesh
e-mail: {hasan, hasank, tariq93, mottalib}@iut-dhaka.edu

***Abstract*—In this paper, we present a hand gesture recognition system using Scale Invariant Feature Transform (SIFT) on depth images. Due to SIFT features and depth information, our approach is robust against rotations, scaling, illumination conditions, cluttered background, and occlusions. Previously, SIFT features were applied on binary images which do not provide enough discriminating key points to classify close gestures. We have extracted SIFT keypoints from each depth silhouette and applied k-means clustering to reduce feature dimensions. Bag-of-word features were generated using vector quantization technique, which maps keypoints from each training image into a unified dimensional histogram. These bag-of-word features were fed into multiclass Support Vector Machine (SVM) classifier for training. We have tested our results for five close symbolic gestures, compared the results for both binary and depth images and found higher accuracy for depth images. The proposed recognition scheme can be used to develop human gesture based interactive Human-Computer Interaction (HCI) applications.**

Keywords-Hand gesture recognition, SIFT features, depth image, HCI, SVM

I. INTRODUCTION

A gesture can be defined as a physical movement of the hands, arms, face and body with the intent to convey information or meaning [28]. Computer based gesture recognition technique has been an important area of research in HCI. Gesture based interaction works as a communication tool for people with cognitive or physical impairments. Hand gestures provide a complementary modality to speech for expressing one's ideas. Information associated with hand gestures in a conversation conveys information in degree, discourse structure, spatial and temporal structure.

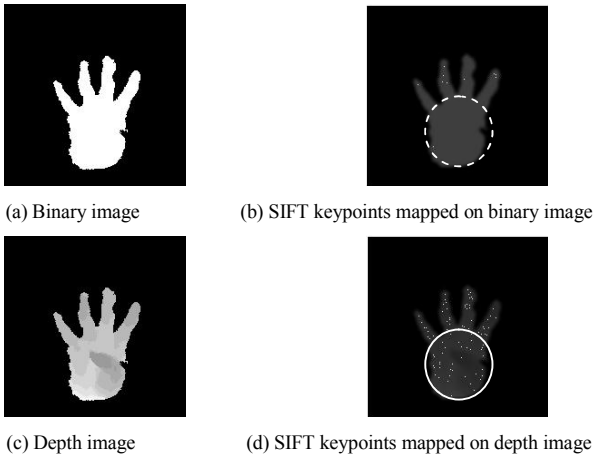
There are two approaches to recognize human to computer interactions, one is sensor-based and the other is vision-based approach. Both of them have some advantages and disadvantages. In sensor-based approaches, the user needs to wear sensor enabled gloves, which limits the naturalness in interactions with computers. In vision-based approaches, the problem of object segmentation from the occluded background containing various illumination conditions can reduce recognition rate [32]. An ideal appearance based hand gesture recognition technique should fulfill the requirements in terms of real time performance, recognition accuracy with high degree of freedom (DoF), robustness against transformations, cluttered background, and with different hands shapes of different people. We have used SIFT features as local oriented features on depth silhouettes captured by Microsoft Kinect. Previously, SIFT feature based gesture recognition [1] was implemented using binary images captured from webcam. Binary images do not contain context information of the hand fingers and also images captured from webcams will not give better accuracy for occluded background. SIFT features are used in many other computer vision applications like robot localization and navigation, object recognition, image stitching etc. [8]. For any object in an image, interesting points or key points on the object can be extracted to provide a "feature description" of that object. This description, extracted from a training image, can then be used to identify the object when attempting to locate the object in a test image containing other objects. To perform reliable recognition, it is important that the features extracted from the training image are recognizable even under changes in image scale, noise and illumination. Such points usually lie on high-contrast regions of the image, such as object edges and are invariant to translation, rotation and scale. SIFT algorithm consider these points as key points naturally.

Object recognition based on the depth information does not have the difficulty of light variations, color distortion as they use infrared lights. Moreover, segmentation of objects from images is faster and easier using depth data captured using Kinect. Formerly, Kinect was developed for tracking full body motion. Many researchers proposed diversified applications like, interactive visual display [2], system for physical therapy [3], gesture based robot navigation [4]-[6], sign language recognition [7], etc.

In this paper, we develop a static hand gesture recognition system using SIFT features so that we can recognize gestures representing one, two, three, four and five numeric symbols. These are called symbolic gestures because they depict the numerical symbols from 1-5. These are close gestures and need to be recognized precisely. To do that, we have captured depth silhouettes after segmentation using depth data stream. We have extracted the SIFT features from those depth images for individual gestures. Individual gesture contains 100 samples and key points are extracted to generate bag-of-word features. Dimensionality of feature space has been reduced using k-means clustering because SIFT algorithm [8], produces too many local features as keypoints. Moreover, from binary images it is not possible to extract

 ISBN: 978-1-61208-468-8

local finger context information (difference on information regarding folding of two fingers and three fingers to produce gestures for numeric symbol 3 and 2 respectively). Figure 1, shows the difference of SIFT keypoints between binary and depth images. The marked region by solid circle in the Figure 1(d) contains discriminative keypoints in depth image but those are missing in binary image, marked by dashed circle in Figure 1(b). Depth images contain discriminating keypoints, which are important to distinguish gestures precisely. In binary images, the keypoints extracted from the contours are not sufficient to get more accuracy.



(a) Binary image (b) SIFT keypoints mapped on binary image

(c) Depth image (d) SIFT keypoints mapped on depth image

Figure 1. Comparison of discriminating keypoints mapped into binary (a) (b) and depth (c) (d) images

We tested our results and got 99% accuracy for depth images whereas binary images gave 96% correct recognition of gestures for cluster size 1600.

This paper is organized as follows: Section II briefly describes the related work. Section III explains the proposed system. Section IV shows our experimental result. Section V discusses the conclusion and future work.

II. RELATED WORK

Appearance based hand gesture recognition using depth images captured from Microsoft Kinect device [9] does not have the difficulty of light, color distortion. Kinect sensors project an infrared pattern of 307, 200 dots in a 640×480 mesh and receive the reflected pattern through a Complementary Metal-Oxide Semiconductor (CMOS) monochrome sensor. Using the triangulation method, Kinect measures the depth value of every point presented in millimeters. This depth stream can be used to distinguish the hand region from the background image, a process called hand segmentation. Many authors have developed applications of gesture recognition using depth sensors in different fields such as interactive displays [2], physical rehabilitation [3], and robot guidance [6], [10], [11] or sign language recognition [7]. Additionally, completely different applications for Kinect have been developed such as in [12], [13]. Hand segmentation is challenging due to occlusions, different lighting conditions or appearance of skin-colored object apart from hand [14]. Some of the solutions to these problems are the assumption of some particular situations like hand being the front-most object or use of full-body tracking algorithms. There are various feature extraction techniques used for hand gesture recognition like, scale-space color feature [15], hand contour feature [16], eigenspace technique but they are not invariant to translation, rotation and scaling.

Different people have a variety of hand shapes and sizes performing the same gesture with slight variation in translation and rotation. An efficient hand gesture recognition method should be invariant to these changes. We have used SIFT algorithm that extracts local invariant features as keypoints preserving hand shape information. Previous research works have addressed local invariant features. In [17], SIFT features and Adaboost learning algorithm was applied to achieve in-plane rotation invariant property but they need to work with some other features like contrast context histogram. Haar-like features as described in [18], focus on the information of a certain area of an image rather than each important pixel or pixel containing key information. Viola and Jones [19] used learning-based object detection technique, which requires large training time, large training images and more computational power for both testing and training.

Features are interesting keypoints or salient points (corner, edges etc.) containing important local information of an image. Identifying the keypoints (detection), extracting vector feature descriptor in the neighborhood of each keypoint (description), and finding the correspondence between descriptors in multiple views are the main components of local features in order to classify images for object recognition. In computer vision, SIFT algorithm is a robust feature detection and description method for local features in images. SIFT extracts distinctive invariant features from images that are invariant to image scale, rotation and translation. SIFT has been used in different applications like image mosaic, stitching, recognition, and retrieval. Later on, other variations of SIFT were introduced like Principal Component Analysis (PCA)-SIFT [20], Gradient Location-Oriented Histogram (GLOH) [21], Speeded up Robust Features (SURF) [22], etc. SIFT is faster for lower resolution images. As the Difference of Gaussian (DOG) representation of hand gesture images are not explicitly affine invariant, we are using SIFT features instead of other variants of SIFT. Moreover, the training images from Kinect were captured in real-time under different lighting conditions so feature vector is robust in terms of scale, rotation, illumination changes, and cluttered background. The number of keypoints from each training image is different and training requires unified dimensional feature vector. Bag-of-Feature (BoF) representation was used in [23] to obtain a global information of visual data out of arrays of local point descriptors generated by SIFT algorithm.. BoF is a visual descriptor used for visual data classification. BoF was inspired by the concept of Bag of Words that is used in document classification. A bag of words is a sparse vector of occurrence counts of words; that is, a sparse histogram over the vocabulary. In computer vision, a bag of visual words of features is a sparse vector of occurrence counts of a vocabulary of local image features. BoF feature depends on dense local motion features and they do not have any relation between features on spatial domain and temporal domain. As we are recognizing static gestures,

 ISBN: 978-1-61208-468-8

we have used BoF representations of images for training and testing. Our main contributions in this research works are: applying SIFT over depth images (extracted after segmentation using depth stream) which was not previously considered; obtaining a better accuracy by understanding the significance of BoF vector dimensions pertaining to the important keypoint descriptors found in gray scale depth image. Finally, the proposed system is robust against rotation, scaling, occlusions unlike state-of-the-art template matching approaches.

III. PROPOSED SYSTEM

Our proposed system uses SIFT features on depth images instead of binary images. Depth images of hand gestures were captured using Microsoft Kinect following hand segmentation using depth map. The proposed system consists of (1) Hand segmentation, (2) Feature extraction using SIFT, (3) Feature dimension reduction using k-means clustering and vector quantization, (4) Gesture classification using multiclass SVM. The overall architecture is presented in Figure 2.

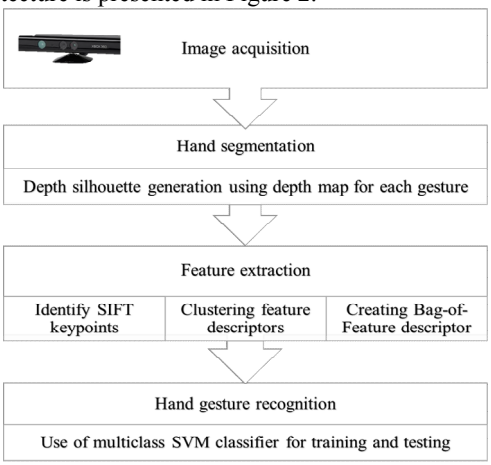


Figure 2. Hand gesture recognition architecture

A. Hand Segmentation

Segmentation is the pre-processing step in which we have used depth information to segment the hand gesture region. Traditionally, color based hand segmentation is affected by the appearance of skin color-like objects, lighting conditions and occluded scenes.

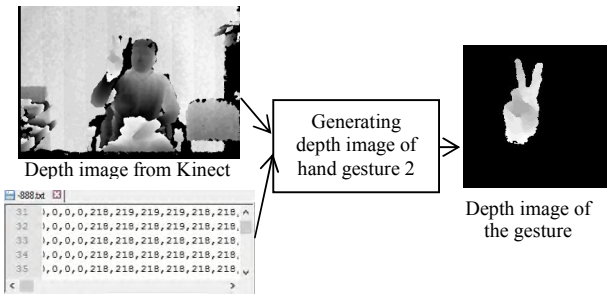


Figure 3. Hand gesture segmentation

We assume that the hand is the closest object in front of Kinect and the gesture information is contained within the 200×200 image. Using a depth threshold we eliminate all sorts of confusion related to background.

This also filters the cluttered background or overlapped images (i.e. hand overlapped with face) and illumination changes, which justifies the use of depth sensor. In the images from Kinect depth stream the hand region does not show enough contrast variations so that we can get more discriminative keypoints to differentiate close gestures. So we have assigned different contrast levels from 155-255 into 10 levels in gray scale to achieve good depth silhouettes. First, we have taken the depth map values of each gesture image of size 640×480 and extracted depth values within 200×200 region containing gesture information and generated the depth silhouette for the corresponding gesture. The segmentation process is shown in Figure 3. The binary images were generated without any gray level (i.e. just assigning 0 (background) and 255 (foreground)) values to compare our result with depth images.

B. Feature Extraction

We have extracted local discriminative features from gesture images using SIFT algorithm, which produces rotation and scale invariant 128-dimensional feature descriptors.

1) Generating SIFT feature descriptor: SIFT algorithm works in four steps.

a) Detecting scale space extrema: We calculate the Laplacian of Gaussian (LoG) for the image $I(x, y)$, with diffetent sigma (σ) values which represents the scale parameter. We convolute the image with gaussian filter to produce a blurred image, L, using "(1)".

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y) \quad (1)$$

where, $G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2/2\sigma^2)}$

SIFT algorithm uses Difference of Gaussian (DoG), using "(2)", by taking the difference of blurred images for two different σ values (σ and $k\sigma$), i.e.

$$DoG(x, y, \sigma) = L(x, y, k\sigma) - L(x, y, \sigma) \quad (2)$$

This process is done for different octaves of the image in Gaussian Pyramid. For our case, we got optimal values with initial $\sigma = 1.275$ and $k = \sqrt{2}$. With this process we get scale invariant representations of hand gesture features.

b) Keypoint localization and filtering: Once DoG images are found, they are searched for local extrema over scale and space. One pixel in the image is compared with its 8 neighbours as well as 9 pixels in next scale and 9 pixels in previous scale. Then, a pixel is selected if it is larger or smaller (extrema) than all 26 neighbours. This is a keypoint best representing in a scale. We remove the edge keypoints which are subject to aperture problem using Harris corner detection and reject points with low contrast using thresholding in the DoG images.

c) Orientation assignment: We have found that keypoints stable after localization and filtering. We assign orientatoin to each keypoint to get rotation invariant property. Since we already know the scale and location of

the extrema, we calculate the gradient direction and magnitute of each pixel around the keyoint. Gradient magnitude and orientation are determined using "(3)" and "(4)".:

$$m(x,y)=\sqrt{(L(x+1,y)-L(x-1,y))^2+(L(x,y+1)-L(x,y-1))^2} \quad (3)$$

$$\theta(x,y)=\tan^{-1}((L(x,y+1)-L(x,y-1))/(L(x+1,y)-L(x-1,y))) \quad (4)$$

A gradient histogram (orientation histogram) is created with 36 bins covering 360 degrees for each keypoint and 80% of points represent the directions as a keypoint direction.

d) Keypoint descriptor: Each keypoint has x, y, σ, m, θ . We create the keypoint descriptor by taking a 16×16 window of neighbourhood around the keypoint. It is devided into 16 sub-blocks of 4×4 size. For each sub-block, 8 bin orientation histogram is created. So a total of 4×4×8=128 bin values are available as a vector to form the keypoint descriptor.

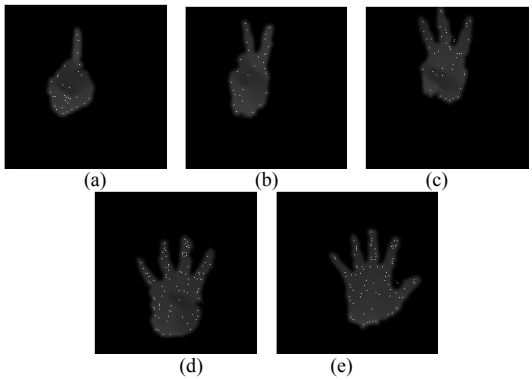


Figure 4. 200×200 training images with SIFT keypoints mapped in to depth silhouettes

We get different number of keypoints (vectors) for different hand gesture images. Figure 4 shows some training images with keypoints mapped in to them. Figures 4(a)-4(e), represent hand gestures one to five with 33, 37, 43, 75, and 87 SIFT keypoints repectively. We noticed that, as the number of fingers increases, keypoints also increases because the area of hand increases. We have used 100 images per gesture for which we extracted the keypoint descriptors, as a total we got 27389 keypoints for 500 training images. These 128 dimensional keypoints need to be distributed into suitable number of clusters so that all the gesture information produced by hand fingers is retained. So it is intuitive that, gray images will require more space to be distributed in to clusters than binary images for better accuracy.

2) Clustering feature descriptors: The number of keypoints for each gesture image is different and each keypoint is a 128 dimensional vector representing the feature descriptor. This creates difficulties because if we take a gesture image containing at least 35 keypoints the image dimension becomes 35×128 = 4480 and for 87 keypoints, 87×128 = 11136, which need to be reduced. Moreover, each training image should be represented using unified dimensional feature vector for training using a multiclass classifier such as SVM [26]. So there is a need of clustering the SIFT feature descriptors and generating unified dimensional image representation. We have used k-means clustering [24] from all the availabe state-of-the-art clustering approaches because it is faster for large datasets than hierarchical clustering. If we observe the keypoints in Figure 4, we see the keypoints are well distributed and distinctive. Morever, there were no outliers in the depth gesture images so k-means clustering should give better result with suitably chosen cluster size or codebook size. If the size of the codebook is too small, then each bag-of-words vector will not represent all the keypoints and large codebook size may increase the risk of overfitting. As the number of keypoints in the depth image is greater than the binary image, intuitively, we will get better accuracy for higher cluster size than for a binay image. We choose our cluster size 1600, the size of the codebook or visual vocabularies to build our cluster model.

K-means clustering partitions the vector space (128-dimensional feature vector) into k clusters where each feature point belongs to the cluster with nearest mean. Then the centroids (codevectors) are moved to the average location of all the keypoints assigned to them, and the assignments are redone until the assignments stop changing. The objective of k-means clustering is to minimize total intra-cluster variance or the squared error function using "(5)".

$$J=\sum_{j=1}^{k}\sum_{i=1}^{n}\left\|x_i^{(j)}-c_j\right\|^2 \quad (5)$$

where k is the number of clusters, n is number of samples, c is the centroid of clusster j.

Figure 5 shows the demonstration of k-means clustering for five keypoints: A, B, C, D, and E to form two clusters.

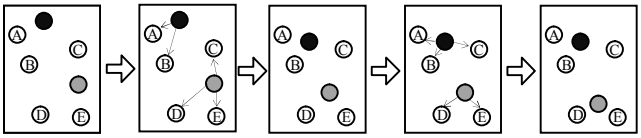


Figure 5. Demonstration of k-means clustering

The keypoint vectors for each training image are employed to build cluster model encoding each keypoint by the index of the cluster (codevector) to which it belongs. So, each feature vector (keypoint) is assigned to one cluster center which has minimum Euclidean distance in 128-dimensional feature vectors. Keypoints assigned to the same cluster center will be in the same subgroup to get k disjoint subgroups of keypoints. So, k-means clustering decreases the dimension of each training image with n keypoints ($n\times128$) to $1\times k$, where k is the number of clusters.

3) Generating bag-of-features: After extracting the SIFT features from depth images and building k-means clustering model, we generate bag-of-features for training. We apply vector quantization (VQ) process [25] to map keypoints of every training image into a unified dimensional histogram vector after k-means clustering.

That means, each keypoint, extracted from training image, will be represented by one component in the generated bag-of-feature vector with value equal to the index of the centroid in the cluster model with nearest Euclidean distance. Figure 6 shows the process of generating the bag-of-features.

4) Gesture recognition using SVM classifier: The generated bag-of-feature vectors with their related class label numbers are fed into the multiclass SVM training model. SVM is a supervised learning method that uses a non-linear mapping to transform original training dataset into higher dimension and searches for a linear optimal separating hyperplane. SVM finds the hyperplane using support vectors (datapoints closest to the separating hyperplane) and margins defined by the support vectors. Maximum margin hyperplane provides maximum separation between the classes.

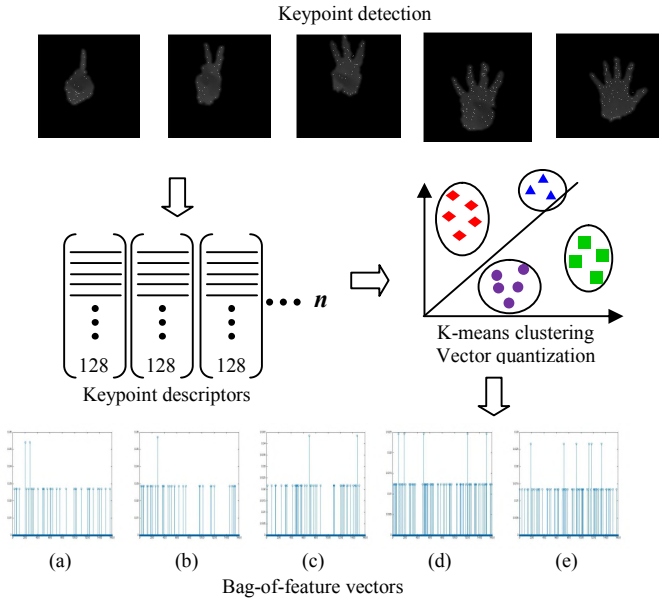


Figure 6. Generating bag-of-feature for training. (a)-(e): Bag-of-feature generated for gesture 1-5 from individual depth silhouette for 1600 clusters

There are different approaches to decompose the multiclass problem into several binary problems using SVMs as binary classifier. We have used multiclass SVM classifier that models a given training set with the corresponding group vector (representing class labels from 1 to 5) and classifies a test set using one against all approach. We have used the SVM tool found at [26]. For testing we have used 5-fold cross validation for 100 samples per gesture.

IV. Experimental Result

We have created our dataset, which contains 100 samples of size 640×480 depth images containing five gestures (representing symbolic gestures: one, two, three, four, and five). These samples are all taken in complex scenes with cluttered background. Experimental images were collected using Microsoft Kinect for Windows SDK version 1.8. The dataset can be found in [27]. Next, we cropped the 200×200 portion from the middle of every image and we took the closest 10 millimeters of each image which represent the hand of the subject. We give gray levels from 155-255 to each pixel of the hands depending on their depth and get a good contrast ratio. We collected the 128 dimensional SIFT features for each key points from each image and we divided the whole data set into 5 sections each containing 20 images. As we followed 5-fold cross validation process we used 4 of these image groups and one of them to test. We repeated the process to test each group of images. We also tested the binary images in the same process and varied the number of clusters in doing so. Then, we compiled the results for each test and took the average.

At first, we generated the confusion matrices using both binary and depth images for five gestures for different cluster size.

Table I. Confusion Matrix on Binary Image

	1	2	3	4	5
1	19	0	0	0	1
2	0	18	0	0	2
3	0	0	20	0	0
4	0	1	0	19	0
5	0	0	0	0	20

Table II. Confusion Matrix on Depth Image

	1	2	3	4	5
1	20	0	0	0	0
2	0	20	0	0	0
3	0	0	19	0	1
4	0	0	0	20	0
5	0	0	0	0	20

We took 100, 200, 400, 800, 1200 and 1600 clusters to test our proposed method and compared out results with the ones we get from binary images.

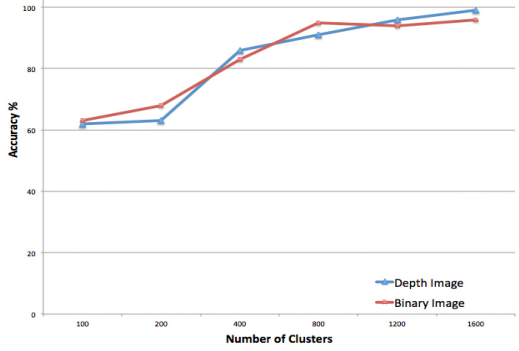


Figure 7. Overall accuracy comparison between proposed method and existing method.

According to our observation, our proposed method using depth image performs better than the binary image based approach if we increase the number of clusters. The confusion matrices generated by 5 gestures (1-5) for both binary and depth images using 1600 clusters are given in Table I and Table II. We are getting better results from our proposed method for higher number of clusters because we are using depth information and it preserves edge information more significantly than binary images.

 ISBN: 978-1-61208-468-8

For our given gestures, we can distinguish each position of fingers from the depth image, which is not possible for binary image. But, this can only be achieved if the cluster numbers are high enough. This is because the number of keypoints generated using the gray scale images is greater than the number of keypoints found using binary images. The overall result is shown in Figure 7.

V. CONCLUSION AND FUTURE WORK

This paper presents an effective way of exploring depth information for hand gesture recognition. We use SIFT keypoint descriptor to extract local keypoint features over depth images which is attempted for the first time in gesture recognition research. The recognition technique is inherently robust against scaling, rotation, and illumination conditions. As we have used depth images, the system is also robust against cluttered background and overlapped objects. By comparing our results with binary images, we found higher accuracy for gray-scale depth images because we get more discriminative keypoint features that preserve the hand finger shape information. Experimental results show that our system is able to achieve up to 99% accuracy over binary image based gesture recognition. We have also created our own depth gesture dataset for these gestures which are made publicly available [31]. There are three important factors that affect the system accuracy. The First factor is correctly preparing the depth images containing gestures. To produce depth silhouettes, our segmentation process is faster and effective with the help of depth map information. Secondly, the chosen number of clusters is justified because the number of keypoints in depth images is greater than binary images. These keypoints have contributed significantly to increase the recognition accuracy. We need the cluster size more than that required for binary images and our experimental results also prove that. Third, the number of training images is important and we took sufficient amount of training images to develop cluster model and later on to classify using SVM classifier. In future, we will try to analyze how the variation of clustering affects the accuracy using the PCA [20].

REFERENCES

[1] W. Lin, Y. Wu, W. Hung, and C. Tang, "A Study of Real-Time Hand Gesture Recognition Using SIFT on Binary Images," Advances in Intelligent Systems & Applications, SIST 21, Springer-Verlag Berlin Heidelberg, 2013, pp. 235–246..

[2] S. Zhang, W. He, Q. Yu, X. Zheng, "Low-Cost Interactive Whiteboard Using the Kinect," In Proceedings of the International Conference on Image Analysis and Signal Processing (IASP), Huangzhou, China, 9–11 November 2012; pp. 1–5.

[3] Y. J. Chang, S. F. Chen, J. D. Huang, "A Kinect-based system for physical rehabilitation: A pilot study for young adults with motor disabilities," Res. Dev. Disabil. 2011, 32, 2566–2570.

[4] A. Ramey, V. Gonzalez-Pacheco, M .A. Salichs, "Integration of a Low-Cost rgb-d Sensor in a Social Robot for Gesture Recognition," In Proceedings of the 6th ACM/IEEE International Conference on Human-Robot Interaction (HRI), Lausanne, Switzerland, 6–9 March 2011, pp. 229–230.

[5] M. V. D. Bergh, D. Carton, R. D. Nijs, N. Mitsou, C. Landsiedel, K. Kuehnlenz, D. Wollherr, L. Van Gool, M. Buss, "Real-time 3D hand gesture interaction with a robot for understanding directions from humans", In Proceedings of the IEEE RO-MAN, Atlanta, GA, USA, 31 July–3 August 2011, pp. 357–362.

[6] D. Xu, Y. L. Chen, C. Lin, X. Kong, X. Wu, "Real-Time Dynamic Gesture Recognition System Based on Depth Perception for Robot Navigation," In Proceedings of the IEEE International Conference on Robotics and Biomimetics, Guangzhou, China, 11–14 December 2012; pp. 689–694. Sensors 2013, 13 11860.

[7] Z. Zafrulla, H. Brashear, T. Starner, H. Hamilton, P. Presti, "American sign language recognition with the kinect," In Proceedings of the 13th International Conference on Multimodal Interfaces, Alicante, Spain, 14–18 November 2011; pp. 279–286.

[8] D. G Lowe, "Distinctive image feature from scale-invariant keypoints," Int. J. Comput. Vis., vol. 60, no. 2,pp. 91-110, Nov. 2004.

[9] Microsoft Corp. Redmond WA. Kinect for Xbox 360. 2010.

[10] A. Ramey, V. Gonzalez-Pacheco,; M. A. Salichs, "Integration of a Low-Cost rgb-d Sensor in a Social Robot for Gesture Recognition," In Proceedings of the 6th ACM/IEEE International Conference on Human-Robot Interaction (HRI), Lausanne, Switzerland, 6–9 March 2011; pp. 229–230.

[11] M. Van den Bergh, D. Carton, R. De Nijs, N. Mitsou, C. Landsiedel, K. Kuehnlenz, D. Wollherr, L. Van Gool, M. Buss, "Real-time 3D hand gesture interaction with a robot for understanding directions from humans," In Proceedings of the IEEE RO-MAN, Atlanta, GA, USA, 31 July–3 August 2011; pp. 357–362.

[12] O. M. Mozos, H. Mizutani, R. Kurazume, T. Hasegawa, "Categorization of indoor places using the Kinect sensor," Sensors 2012, 12, 6695–6711.

[13] G. Azzari, M. L. Goulden, R. B. Rusu, "Rapid characterization of vegetation structure with a Microsoft Kinect sensor," Sensors 2013, 13, 2384–2398.

[14] J. Suarez, R. R. Murphy, "Hand Gesture Recognition with Depth Images: A Review," In Proceedings of the IEEE RO-MAN, Paris, France, 9–13 September 2012, pp. 411–417.

[15] L. Bretzner, I. Laptev, and T. Lindeberg, "Hand gesture recognition using multiscale color features, hieracrchichal models and particle filtering," in Proc. Int. Conf. Autom. Face Gesture Recog., Washington, DC, May 2002.

[16] A. Argyros and M. Lourakis, "Vision-based interpretation of hand gestures for remote control of a computer mouse," in Proc. Workshop Comput. Human Interact., 2006, pp. 40–51.

[17] C. Wang and K. Wang, "Hand Gesture Recognition Using Adaboost With SIFT for Human Robot Interaction," vol. 370. Berlin, Germany: SpringerVerlag, 2008.

[18] Q. Chen, N. Georganas, and E. Petriu, "Real-time vision-based hand gesture recognition using Haar-like features," inProc. IEEE IMTC, 2007, pp. 1–6.

[19] P. Viola and M. Jones, "Robust real-time object detection," Int. J. Comput. Vis., vol. 2, no. 57, pp. 137–154, 2004.

[20] Y. Ke and R. Sukthankar, "PCA-SIFT: A more distinctive representation for local image descriptors," in Proc. IEEE Conf. Comput. Vis. Pattern Recog., 2004, pp. II-506–II-513.

[21] K. Mikolajczyk, and C. Schmid, , "A performance evaluation of local descriptors", IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(10):1615–1630.

[22] H. Bay, A. Ess, T. Tuytelaars, and L. Gool, "SURF: Speeded up robust features," Comput. Vis. Image Understand. (CVIU), vol. 110, no. 3, 2008, pp. 346–359..

[23] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," inProc. IEEE Conf. Comput. Vis. Pattern Recog., 2006, pp. 2169–2178.

[24] D. J. C. MacKay, Information Theory, Inference, and Learning Algorithms. Cambridge, U.K.: Cambridge Univ. Press, 2003.

[25] A. Bosch, X. Munoz, and R. Marti, "Which is the best way to organize/ classify images by content?" Image Vis. Comput., vol. 25, no. 6, Jun. 2007, pp. 778–791.

[26] Multiclass Support Vector Macine Classifier, http://www.mathworks.com/matlabcentral/fileexchange/33170-multi-class-support-vector-machine/

 ISBN: 978-1-61208-468-8

[27] SSL hand gesture dataset 2 [http://cse.iutoic-dhaka.edu/group/ssl/SSL_HandGestureDataset_2.rar].

[28] H. Haitham, A. K. Sameem, "Human–computer interaction using vision-based hand gesture recognition systems: a survey", Journal of Neural Computing and Applications, vol. 25, issue 2, August 2014, pages 251-261.

 ISBN: 978-1-61208-468-8