SICE Annual Conference 2012
August 20-23, 2012, Akita University, Akita, Japan

Gesture Recognition with Depth Images - A Simple Approach

Upal Mahbub[1], Hafiz Imtiaz[1], Tonmoy Roy[1], Md. Shafiur Rahman[1], and Md. Atiqur Rahman Ahad[2]

[1]Department of Electrical and Electronic Engineering,
Bangladesh University of Engineering and Technology, Dhaka-1000, Bangladesh
E-mail: omeecd@eee.buet.ac.bd; hafiz.imtiaz@live.com; tonmoyroy@live.com; shafeey@live.com
[2]Department of Applied Physics, Electronics & Communication Engineering
University of Dhaka, Bangladesh
Email: atiqahad@univdhaka.edu

Abstract: A novel approach for gesture recognition is developed in this paper based on template matching from motion depth image. The proposed method uses a single example of an action as a query to find similar matches from a good number of test samples. No prior knowledge about the actions, the foreground/background segmentation, or any motion estimation or tracking is required. A novel approach to separate different gestures from a single video is also introduced. The proposed method is based on the computation of space-time descriptors from the query video which measures the likeness of a gesture in a lexicon. The descriptor extraction method includes the standard deviation of the depth images of a gesture. Moreover, two dimensional discrete Fourier transform is employed to reduce the effect of camera shift. Classification is done based on correlation coefficient of the image templates and an intelligent classifier is proposed to ensure better recognition accuracy. Extensive experimentation is done on a vast and very complicated dataset to establish the effectiveness of employing the proposed method.

Keywords: gesture recognition, depth image, standard deviation, 2D Fourier transform.

1. INTRODUCTION

Hand gesture recognition and analysis is very important in computer vision and man-machine interactions for various applications [1], [2]. Wu and Huang [3] presented a survey on gesture recognition, whereas, action recognition issues are analyzed in [1]. The application arenas are gaming, computer interfaces, robotics, video surveillance, body posture analysis, action recognition, behavior analysis, sign language interpretations for deaf, facial and emotion understanding, etc. [4] [5], [6].

There are several methods proposed by a good number of researchers in the field of action and gesture recognition [4], which are already implemented in practical applications. The basis of representation of the action, which is usually different in each method, is commonly related to appearance, shape, spatio-temporal orientation, optical flow, interest-point or volume. For example, in the spatio-temporal template based approaches, an image sequence is used to prepare a motion energy image (MEI) and a motion history image (MHI), which indicate the regions of motions and preserve the time information of the motion [5], [2] as well. An extension of this approach to three dimension is proposed in [7]. Recently, some researchers explored optical flow based motion detection and localization methods and frequency domain representation of gestures [8], [9]. Among other methods, the Bag-of-words method has been used successfully for object categorization in [10]. [11] proposed a method for segmenting a periodic action into cycles and thereby classify the action.

Even though it is very important to understand various hand gestures, it is extremely difficult to do so due to various challenging aspects: dimensionality and varieties of gestures, partial occlusion, boundary selection, head movement, background, cultural variations in sign languages, etc. Usually, most of the existing works on gesture recognition and analysis cover only a few classes or types of gestures and usually, these are also not continuous and not in random orders. Existing methods can be classified into several categories, such as view/appearance-based, model-based, space-time volume-based, or direct motion-based [1] methods. Template-matching approaches [2], [12] are simpler and faster algorithms for motion analysis or recognition that can represent an entire video sequence into a single image format. Recently, approaches related to Spatio-Temporal Interest feature Points (STIP) are becoming prominent for action representation [13]. However, due to various limitations and constraints, no single approach seems to be enough for wider applications in action understanding and recognition. Hence, the challenging aspects of gesture recognition still remains in its infancy.

Moreover, though a number of the available gesture recognition methods have acquired high accuracy in different datasets, most of them depend on a good amount input to train the system and do not perform very well if the number of training data were limited. [14] proposed a method for action recognition using a patch based motion descriptor and matching scheme with a single clip as the template.

The Kinect sensor with a mounted depth sensor encouraged the researchers to develop some algorithms to take advantage of the depth information. Among such works, [15] proposed a method for hand detection by integrating RGB and depth data which involved finding possible hand pixels by skin color detection based on the fact that neck and other skin colored cloths parts of the body will be behind the hand in most of the cases.

PR0001/12/0000- 1723¥400

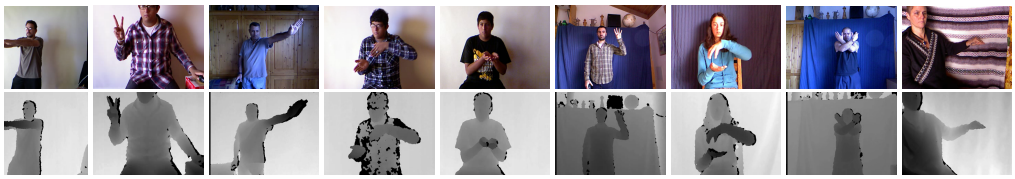

Fig. 1 Extracts from the ChaLearn Gesture Dataset : Top row (left to right) has ten RGB gesture data from devel01 to devel10 lexicons; bottom row contains the corresponding depth image data

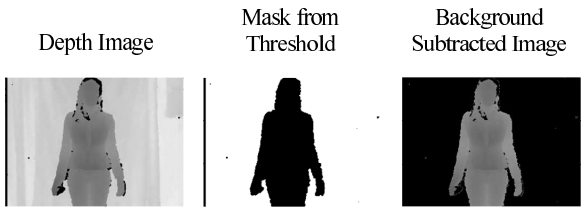

Fig. 2 Background subtracted from depth image.

Another robust method for detection of fingers to easily identify sophisticated and confusing gestures was proposed in [16], namely the Finger-Earth Mover Distance (FEMD) [17]. However gesture recognitions in these cases were simpler since a small set of gestures were used with very little variation in types.

In this paper, an efficient gesture recognition mechanism is developed based on a sequence of depth images of human gestures. The depth images are obtained by utilizing the infrared depth sensor present in a Microsoft Kinect^{TM} Sensor. Motion based template matching techniques are employed on the training videos to obtain the key features from the gesture sequences. The feature vector is formed employing statistical operations both in temporal and spectral domains. The testing phase is divided into several steps. First, different gestures are separated from the long test sequences. Then, similar to the train feature vectors, test feature vectors are generated for each gesture. Finally, every test feature vector is compared to each train feature vectors for different gestures and a classifier is employed to find the best possible match of a gesture from the given training vocabulary. The performance of the recognition algorithm is evaluated with respect to the percentage of Levenshtein distance (LD) [18]. Through extensive experimentation satisfactory recognition performance for a wide variety of gestures including sign languages, cluttered background scenarios, partially visible human figure scenarios etc. is achieved.

2. PROPOSED METHOD FOR GESTURE RECOGNITION

In this paper, a method of gesture recognition from a small vocabulary of gestures is developed. It is evident that many consumer applications of gesture recognition will become possible only if systems can be trained to recognize new gestures with very few examples, and, in the limit, just one. Here it is assumed that the depth image is available in addition to traditional RGB image. The gestures had to be recognized from a set of gestures and matched with a known vocabulary.

2.1 Gesture Dataset

There are some clearly-defined hand or body or head gesture datasets, e.g., Cambridge gesture dataset [19], Naval Air Training and Operating Procedures Standardization (NATOPS) aircraft handling signals database [20], Keck gesture dataset [21], Korea University Gesture (KUG) database [22], etc. Though all these datasets are well known for their contents and complexities, all of them addresses a particular type of gestures limited to very few classes and application domains. Therefore, for the simulation purpose of the proposed method, a very rich but extremely complicated dataset, namely, the ChaLearn Gesture Dataset (CGD2011), is considered in this paper [23]. It is a very challenging database. It has development data as well as validation data. The development data consisted of batches devel01, devel02 ... devel20, etc. For the develXX batches, all the labels are provided by the dataset creators. Each sub-dataset or lexicon consists of 47 video sequences of 1 to 5 gestures, having initials with 'K' to denote depth images from Kinect sensor and 'M' as the original RGB videos. So, there are actually 47×2 number of video files under a lexicon. Each video has 8 to 13 unique actions, and the numbers are varied from one lexicon to another. There are around $50,000$ gestures recorded with the Kinect^{TM} camera with image sizes 240×320 pixels at 10 frames per second. The videos are recorded by 20 different users and grouped in 500 batches of 100 gestures. The data are available from [23] in 2 formats: A lossy compressed AVI format and a quasi-lossless AVI format. To get a sufficient spacial resolution, only the upper body is framed.

Some more attributes about this database are: fixed camera, availability of depth data, single user within a batch, homogeneous recording conditions within a batch, small vocabulary within a batch, gestures performed mostly by arms and hands, camera framing mostly the upper body (some exceptions), only one labeled example of each unique gestures, variations in recording conditions (various backgrounds, clothing, skin colors, lighting, temperature, resolution), some parts of the body may be occluded, some users are less skilled than others, some users made errors or omissions in performing the gestures, etc. It has lexicons from nine categories corresponding to various settings or application domains; they include (1) body language gestures (like scratching head, crossing arms), (2) gesticulations performed to accompany speech, (3) illustrators (like Italian gestures), (4) emblems (like Indian Mudras), (5) signs (from sign lan-

guages for the deaf), (6) signals (like referee signals, diving signals, or marshaling signals to guide machinery or vehicle), (7) actions (like drinking or writing), (8) pantomimes (gestures made to mimic actions), and (9) dance postures. Sample frames of both RGB and depth data of the first 10 lexicons of the development data are shown in Fig. 1.

2.2 Feature Extraction and Training

The depth data provided in the dataset are in RGB format and therefore need to be converted to grayscale before processing. The grayscale depth data is a true representation of object distance from the camera by varying intensity of pixels from dark to bright for near to far away objects, respectively. Next in order to separate the person and the background of the grayscale image Otsu's method [24] is employed. In this method a threshold is chosen based on minimization of the intra-class variance of the black and white pixels for making a black and white binary image. Utilizing the binary image the background from the depth image is filtered out and thereby the human subject is separated as shown in Fig. 2.

In the proposed methods of gesture recognition two types of operations are performed for obtaining feature vectors from training samples : a. calculating standard deviation (STD) on each pixel values across all the frames and b. employing two dimensional Fourier transform (2D-FFT) on each frame. The process of feature extraction, training and classification are shown in brief in Fig. 3.

For the n-th gesture at lexicon L consisting of D frames, the standard deviation ${}_D\alpha_n^2(x,y)$ of pixel (x,y) across the frames is given by

$$ {}_D\alpha_n^2(x,y) = \frac{\sum (I_{xy}(d) - \overline{I_{xy}})^2}{D}. \tag{1} $$

Here $I_{xy}(d)$ is the pixel value of the location (x,y) of the frame d. Where $x = 1,2,3 \ldots r$, $y = 1,2,3 \ldots s$ and $d = 1,2,3 \ldots D$. r and s corresponds to the maximum pixel value along x and y axis, respectively. $\overline{I_{xy}}$ is the average of all $I_{xy}(d)$ values from frames 1 to d. Therefore, for the whole frame, the matrix obtained is defined as

$$ {}_D\Delta_n^L = \begin{pmatrix} {}_D\alpha_n^2(1,1) & {}_D\alpha_n^2(1,2) & \ldots & {}_D\alpha_n^2(1,s) \\ {}_D\alpha_n^2(2,1) & {}_D\alpha_n^2(2,2) & \ldots & {}_D\alpha_n^2(2,s) \\ \vdots & \vdots & \ddots & \vdots \\ {}_D\alpha_n^2(r,1) & {}_D\alpha_n^2(r,2) & \ldots & {}_D\alpha_n^2(r,s) \end{pmatrix} \tag{2} $$

The silhouettes obtained by performing standard deviation across frames on the training samples of development data lexicon 1 of the ChaLearn gesture dataset is shown in Fig. 4. It is evident from the figure that performing standard deviation across frames enhances the information of movement across the frames while suppresses the static parts. Therefore, information about the path of motion flow and type of the gesture is easily observable.

Another feature can be obtained from the gesture samples by taking the absolute value of the two-dimensional Discrete Fourier Transform (2D DFT) using fast Fourier Transform (FFT) on each background subtracted depth image frame and then calculating the standard deviation of each point at the spectral domain across frames. For an image of size $r \times s$ the $2D$ discrete Fourier transform is given by

$$ F_{kl}(d) = \frac{1}{rs} \sum_{x=0}^{r-1} \sum_{y=0}^{s-1} I_{xy}(d) e^{-j2\pi(\frac{kx}{r} + \frac{ly}{s})}, \tag{3} $$

where, the exponential term is the basis function corresponding to each point $F_{kl}(d)$ in the Fourier space and $I_{xy}(d)$ is the value of the (x,y) pixel of the image given in spatial domain for the d-th frame [25]. Performing Fourier transform prior to taking STD has certain advantages. When the depth values of the person are taken to the frequency domain, the position of the person becomes irrelevant, i.e. any slight movement of the camera vertically or horizontally would be nullified in the frequency response. For example, if the camera is steady in the training samples but moves a little bit in the test samples, the STD on temporal values of the test data would be difficult to project on that of the training data for finding a match. However, spectral domain transformation of the data before taking standard deviation across frames will suppress the time resolution and thereby reduce the effect of camera movement [4].

2.3 Finding Action Boundary

While testing, the test data sequences containing one one or more gestures are needed to be separated for different gestures. The gestures provided in the ChaLearn gesture dataset are separated by returning to a resting position [23]. Thus, subtracting the initial frame from each frame of the movie can be a good method for finding the gesture boundary. For a $r \times s$ depth image if $I_{xy}(d)$ is the pixel value (depth) of position (x,y) at the d-th frame, then the following formula can be used to generate a variable P_d which gives an amplitude of change from the reference frame

$$ \begin{aligned} p_1 &= (I_{11}(d) - I_{11}(1))^2 + \cdots + (I_{1s}(t) - I_{1s}(1))^2 \\ p_2 &= (I_{21}(d) - I_{21}(1))^2 + \cdots + (I_{2s}(t) - I_{2s}(1))^2 \\ p_3 &= (I_{31}(d) - I_{31}(1))^2 + \cdots + (I_{3s}(t) - I_{3s}(1))^2 \\ &\ldots \quad \ldots \quad \ldots \\ &\ldots \quad \ldots \quad \ldots \\ p_r &= (I_{r1}(d) - I_{r1}(1))^2 + \cdots + (I_{rs}(t) - I_{rs}(1))^2 \end{aligned} \tag{4} $$

Thus, the sum square difference of the current frame from the reference frame can be obtained by

$$ P_d = \sum_{k=0}^{r} p_k \tag{5} $$

The value of P_d is then normalized. The smoothed graph of normalized P_d vs. d is shown in Fig. 5.a and 5.b for two different development data samples. In the figure, both the actual boundary (determined by observing the

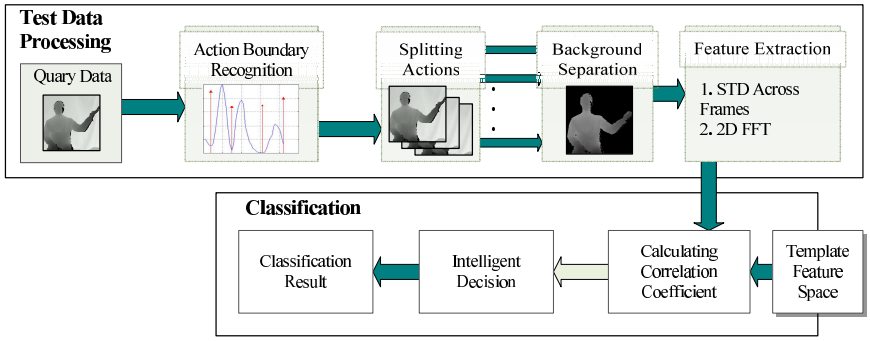

Fig. 3 Proposed Method of Gesture Recognition

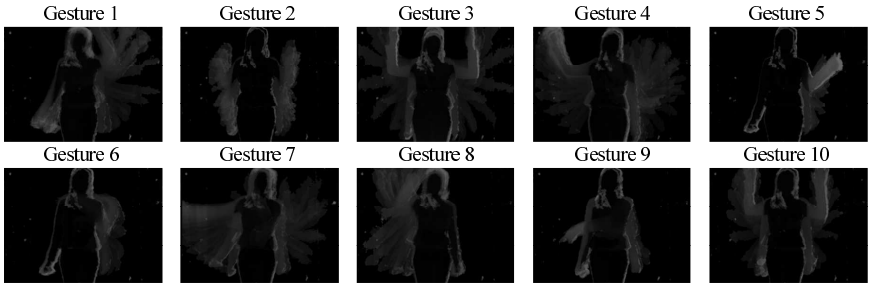

Fig. 4 STD outcome on each pixel across time

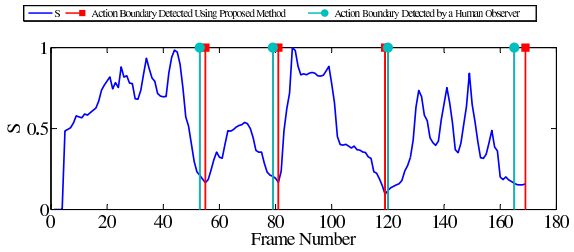

Fig. 5 Action Boundary Detection

gesture sequence) and the boundary detected by the proposed algorithm are shown. Every crest in these curves represent a gesture boundary. Using the location of these crests in time along with some intelligent decision making on the curve, the action boundaries can be found with fairy high accuracy.

2.4 Classification

After extracting the features from the gestures in the training dataset of a particular lexicon a feature vector table is formed for that lexicon as shown in Fig. 3. Then, for the test samples, first the gestures are separated if multiple gesture exists. Then features are extracted for each gesture in a manner similar to that done for training samples. For gesture sequences of frame size $r \times s$ the feature vector obtained from the test gestures are also $r \times s$ matrices for each types of features. Next the correlation coefficient is calculated between the features obtained from the test gesture to that similar feature obtained from each of the train gestures. A coefficient of correlation, τ, is a mathematical measure of how much one number can expected to be influenced by change in another. It is a widely used measure for image and gesture recognition [25]. The correlation coefficient between two images A and B is defined as

$$\tau = \frac{\sum_{k=0}^{r-1}\sum_{l=0}^{s-1}(A_{kl}-\overline{A})(B_{kl}-\overline{B})}{\sqrt{\sum_{k=0}^{r-1}\sum_{l=0}^{s-1}(A_{kl}-\overline{A})^2}\sqrt{\sum_{k=0}^{r-1}\sum_{l=0}^{s-1}(B_{kl}-\overline{B})^2}}, \quad (6)$$

where, A_{kl} is the intensity of the pixel (k,l) in image A and B_{kl} is the intensity of the pixel (k,l) in image B. $\overline{A}$ and $\overline{B}$ are the mean intensity of all the pixels of image A and B, respectively. If $\tau = \pm 1$ then there is a strong positive/negative correlation between the two images, i.e. they are identical/negative of one other. If τ is zero then there is no correlation among the matrices.

A high value of the correlation coefficient between the same feature of the test sample and that of one of the training sample indicates a higher probability of the two gestures to be identical and vice versa for a low value of correlation coefficient. In the proposed method, for using combination of more then one features similar features are matched using correlation coefficient separately and then decision is taken from the two arrays of correlation coefficients to identify the appropriate match in a way that result of both matches influence the decision making.

3. EXPERIMENTAL RESULTS

For the purpose of evaluating gesture recognition performance of the proposed method, the Levenshtein distance (LD) measure is employed. The Levenshtein distance, also known as edit distance, is number of deletions, insertions, or substitutions required to match an array with another [18]. The LD measure has a wide range of applications, such as spell checkers, correction systems for optical character recognition and other such systems [26], [27]. For the proposed gesture recognition system,

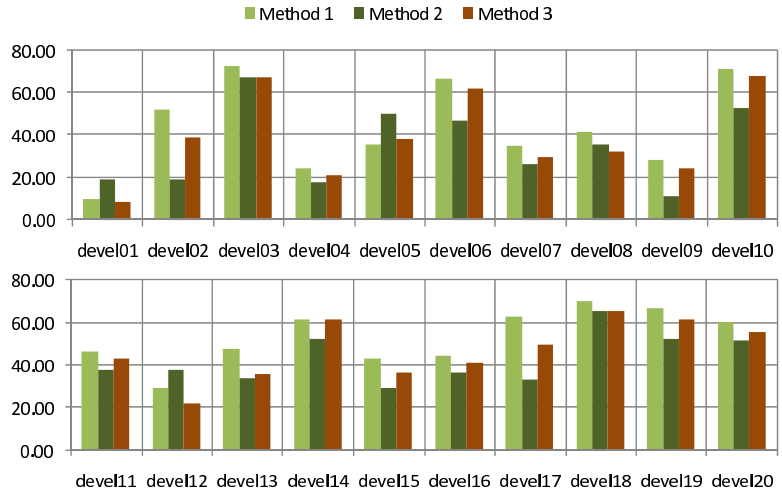

Fig. 6 Levenshtein Distances (%) for different lexicons

if the list of labels of true gesture in a test sequence is T while the labels corresponding to the recognized gestures for the same sequence if R, then the Levenshtein distance L(R, T) will represent the minimum number of edit operations that one has to perform to go from R to T (or vice versa). The evaluation criteria is the percentage LD which is the sum of all the LDs obtained from a lexicon divided by the total number of true gestures in that lexicon and multiplied by 100. It is evident that the higher the value of LD, the more is the number of wrong estimations [18].

For the purpose of analysis, the first 20 lexicons, namely $Devel01$ to $Devel20$ from $CGD2011$ [23] dataset are considered. In this dataset, different actions in a set are assigned a number as a label and a string containing the labels of actions are provided for each gesture sequence. For the experimentation purpose, three sets of features are formed by combining the two basic operations. Also, method of classification deviated a little bit with the types of features. Thus, three different methods are proposed. These methods are,

1. taking STD of the depth image and using maximum value of the correlation coefficient as best match,
2. taking STD of the absolute value of the frequency domain spectrum of the depth image and using maximum value of the correlation coefficient as best match
3. measuring STD of both time and spectral domain images, measuring (a) correlation coefficients for the spectral domain feature and (b) multiplication of correlation coefficient for both features. Then, making intelligent selection from the three highes values of (a) by observing corresponding values at (b)

Simulation are run on the dataset using these methods and the result obtained in terms of percentage of Levenshtein Distance (LD) is shown in Fig. 6 for each lexicon of the first 20 development data. It can be seen from the figure that the result varies widely from set to set. The wide variation is caused by the different types of motions in different sets. For example, in devel03, devel07, devel10, etc. percentage LD is higher because of the gestures being mostly sign languages and therefore very difficult to differentiate while other lexicons include easily differentiable gestures. The accuracy of the proposed method are also limited by the accuracy of separation of different actions from the videos and unexpected movements of the performer.

Table 1 AVERAGE LEVENSHTEIN DISTANCE (%) FOR PROPOSED METHODS

Method	Average Lev. dist %
Method1	48.73
Method2	38.95
Method3	43.43

The best results are obtained for Method 2 which is very much expected (Fig. 6) because due to the spectral domain operation the feature in this method is robust enough to perform well against camera movement. The average percentage LD is shown in Table 1.

4. CONCLUSION

A novel approach for gesture recognition employing combinations of statistical measures and frequency domain transformation obtained from depth motion images is proposed in this paper. After several pre-processing steps, features from gesture sequences are extracted based on two basic operations - calculating standard deviation across frames and taking two dimensional Fourier transform. The measures are then combined to find the most useful method for recognizing gestures. The usefulness of using intelligent classifier based on the correlation coefficient from the standard deviation of the 2 dimensional Fourier transform of the image is apparent from the results. The dataset, namely, the ChaLearn Gesture Dataset 2011, targeted for experimental evaluation is a rich but difficult dataset to handle having a lot of variation and classes. However, through extensive experimentation it is proved that even for one of the most complex dataset the proposed method can provide a satisfactory level of

recognition accuracy.

ACKNOWLEDGEMENTS

The authors would like to express their sincere gratitude towards the organizers of ChaLearn Gesture Challenge for providing the dataset and to Center for Natural Science and Engineering Research (CNSER) for providing constant support throughout this work.

REFERENCES

[1] M. A. R. Ahad, J. K. Tan, H. Kim, S. Ishikawa, Motion history image: its variants and applications, Mach. Vision Appl. 23 (2) (2012) 255–281. doi:10.1007/s00138-010-0298-4.

[2] A. F. Bobick, J. W. Davis, The recognition of human movement using temporal templates, IEEE Trans. Pattern Anal. Mach. Intell. 23 (3) (2001) 257–267. doi:10.1109/34.910878.

[3] Y. Wu, T. Huang, Vision-based gesture recognition: a review, LNCS 103–115.

[4] M. A. R. Ahad, Computer Vision and Action Recognition: A Guide for Image Processing and Computer Vision Community for Action Understanding, Atlantis Ambient and Pervasive Intelligence, Atlantis Press, 2011.

[5] M. A. R. Ahad, J. Tan, H. Kim, S. Ishikawa, Human activity recognition: Various paradigms, in: Control, Automation and Systems, 2008. ICCAS 2008. International Conference on, 2008, pp. 1896–1901. doi:10.1109/ICCAS.2008.4694407.

[6] Y. Fang, K. Wang, J. Cheng, H. Lu, The recognition of human movement using temporal templates, IEEE ICME.

[7] D. Weinland, R. Ronfard, E. Boyer, Free viewpoint action recognition using motion history volumes, Comput. Vis. Image Underst. 104 (2) (2006) 249–257. doi:10.1016/j.cviu.2006.07.013.

[8] U. Mahbub, H. Imtiaz, M. A. R. Ahad, An optical flow based approach for action recognition, in: Computer and Information Technology (ICCIT), 2011 14th International Conference on, 2011, pp. 646 –651. doi:10.1109/ICCITechn.2011.6164868.

[9] H. Imtiaz, U. Mahbub, M. Ahad, Action recognition algorithm based on optical flow and ransac in frequency domain, in: SICE Annual Conference (SICE), 2011 Proceedings of, 2011, pp. 1627 – 1631.

[10] J. Willamowski, D. Arregui, G. Csurka, C. Dance, L. Fan, Categorizing nine visual classes using local appearance descriptors, in: Workshop on Learning for Adaptable Visual Systems in IEEE Internat. Conference on Pattern Recognition, 2004.

[11] H. J. Seo, P. Milanfar, Action recognition from one example, IEEE Trans. Pattern Anal. Mach. Intell. 33 (5) (2011) 867–882. doi:10.1109/TPAMI.2010.156.

[12] M. A. R. Ahad, J. Tan, H. Kim, S. Ishikawa, Analysis of motion self-occlusion problem for human activity recognition, Vol. 5, 2010, pp. 36–46.

[13] I. Laptev, T. Lindeberg, Space-time interest points, Computer Vision, IEEE International Conference on 1 (2003) 432.

[14] W. Yang, Y. Wang, G. Mori, Human action recognition from a single clip per action, Learning (2009) 482–489.

[15] M. Tang, Recognizing hand gestures with microsoft s kinect, stanfordedu 14 (4) (2011) 303–313.

[16] Z. Ren, J. Yuan, Z. Zhang, Robust hand gesture recognition based on finger-earth mover's distance with a commodity depth camera, in: Proceedings of the 19th ACM international conference on Multimedia, MM '11, ACM, New York, NY, USA, 2011, pp. 1093–1096. doi:10.1145/2072298.2071946.

[17] R. Zhou, J. Meng, J. Yuan, in: 8th International Conference on Information, Communications and Signal Processing (ICICS), 2011, pp. 1 – 5.

[18] V. I. Levenshtein, Binary codes capable of correcting deletions, insertions, and reversals, Soviet Physics Doklady 10 (8) (1966) 707–710.

[19] T.-K. Kim, R. Cipolla, Canonical correlation analysis of video volume tensors for action categorization and detection, IEEE Trans. Pattern Anal. Mach. Intell. 31 (8) (2009) 1415–1428. doi:10.1109/TPAMI.2008.167.

[20] Y. Song, D. Demirdjian, R. Davis, Tracking body and hands for gesture recognition: Natops aircraft handling signals database., in: FG, IEEE, 2011, pp. 500–506.

[21] Z. Jiang, Z. Lin, L. Davis, Recognizing human actions by learning and matching shape-motion prototype trees, IEEE Trans. Pattern Anal. Mach. Intell. 34 (3) (2012) 533–547. doi:10.1109/TPAMI.2011.147.

[22] B.-W. Hwang, S. Kim, S.-W. Lee, A full-body gesture database for automatic gesture recognition, in: Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition, FGR '06, IEEE Computer Society, Washington, DC, USA, 2006, pp. 243–248. doi:10.1109/FGR.2006.8.

[23] Chalearn gesture dataset (CGD2011), ChaLearn, California, 2011.

[24] N. Otsu, A threshold selection method from gray-level histograms, IEEE Transactions on Systems, Man, and Cybernetics 9 (1) (1979) 62–66.

[25] R. C. Gonzalez, R. E. Woods, Digital Image Processing, 2nd Edition, Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2001.

[26] J. D. Golić, M. J. Mihaljevi, A generalized correlation attack on a class of stream ciphers based on the levenshtein distance, Journal of Cryptology 3 (1991) 201–212, 10.1007/BF00196912.

[27] A. Marzal, E. Vidal, Computation of normalized edit distance and applications, Pattern Analysis and Machine Intelligence, IEEE Transactions on 15 (9) (1993) 926–932. doi:10.1109/34.232078.