Hand Gesture Recognition using Depth Information and Dynamic Time Warping

Aboubakar Mountapmbeme, Hasan Mahmud*, and Md. Kamrul Hasan

***Abstract*—Human computer interaction has introduced a new paradigm for the construction of computer interfaces. This requires the construction of easy to use, natural and intuitive computer interfaces. One of the best techniques employed to achieve this goal is interaction using hand gestures. As a results, the design of hand gesture recognition system has been a concern for the past years. Making a vision based hand gesture recognition system natural and easy to use often requires designing the system to run efficiently in constrained environments with cluttered background. However, this constraints limit the accuracy and efficiency of hand gesture recognition systems, which is a major concern nowadays. The introduction of depth sensing devices such as the Microsoft Kinect has boosted research activities in this area in the past years. It has also led to the design of a good recognition system with better accuracy and high efficiency. In this paper, we try to implement a hand gesture recognition system that makes use of the time-series representation of the contour points of the hand shape and uses Dynamic Time Warping (DTW) for classification of hand gestures. DTW is well known for its accuracy and effectiveness in matching time-series representations. Our proposed system makes use of the depth information of the scene provided by the Microsoft Kinect for hand segmentation. This allows our system to be used in challenging backgrounds. We evaluate our system and compare the result with other existing system. The results of evaluation shows that our hand gesture recognition system is accurate and efficient with a mean accuracy of 94.6% and mean running time of 0.5179s. Our system is also invariant to scaling, rotation and translation and runs effectively in complex background settings.**

***Keywords*—Human computer Interaction; hand gesture recognition; Microsoft Kinect; Dynamic Time Warping.**

* Corresponding author.

Aboubakar Mountapmbeme, Hasan Mahmud and Md. Kamrul Hasan are with the Systems and Software Lab, Department of Computer Science and Engineering (CSE), Islamic University of Technology (IUT), Dhaka. e-mail: {sidick, hasan, hasank}@iut-dhaka.edu,.

Manuscript received April 22, 2015; revised December 05, 2015.

I. INTRODUCTION

The recent development of depth sensors such as the Microsoft Kinect has led to a boost in research activities in the field of gesture recognition. This has also been geared by the need to attain the goal of Human computer Interaction (HCI), which is to bring the interaction between humans and machines to the human level. This mainly involves designing and implementing easy-to-use, intuitive and natural user interfaces. Interaction using gestures and hand gesture in particular represents one of the ways in which we can attain this goal. As stated in [1], gestures are expressive, meaningful body motions involving physical movements of the fingers, hands, arms, head, face, or body with the intent of conveying meaningful information or interacting with the environment. As we can see from this definition, gesturing is a natural part of human communication and as such, gesture based interactive interfaces represents one of the best techniques or methods that can provide the naturalness and intuitiveness sort by HCI, for interacting with computers. The use of the depth information has led to the design of better hand gesture recognition frameworks in terms of accuracy and running time. However, a huge work needs to be done in order to achieve better accuracy and running time without any restrictions to the user environment. Depth cameras are an improvement over the previous techniques [1] used for capturing hand gestures. Before depth cameras, gestures and hand gestures in particular had been captured using electromechanical devices and vision-based (colour) cameras as described in [1] [2]. These techniques have many constraints and in most cases naturalness cannot be achieved. As described in [1], vision-based techniques impose restrictions on the gesturing environment, such as special lighting conditions and simple and uncluttered background. Electromechanical devices on the other hand are accurate in capturing hand gestures, however, naturalness is usually eliminated since these devices worn by users

are usually heavy and cumbersome [1]–[3].

Another problem often encountered when designing hand gesture recognition system is the choice of the algorithm to use to subsequently recognise the hand gestures. The choice of this algorithm often depends on the features used to represent the hand gestures. With respect to feature extraction, we usually have high-level feature based approaches, 3D feature based approaches and low-level feature based approaches [3]. Algorithms such as Hidden Markovs Model (HMM), Dynamic Time Warping (DTW), Principal Component Analysis (PCA), template matching and many others having been used in dynamic hand gesture recognition systems [1]. Recently, Ren et al. [3]–[5] introduced a novel method called Finger Earth Movers Distance (FEMD) to recognize static hand gestures. This framework has demonstrated better accuracy and performance compared to previous frameworks [3], [5].

In this paper, we present a static hand gesture recognition framework based on the depth information provided by the Microsoft Kinect and template matching through DTW to recognize hand gestures. Our feature vector is a 2D feature vector based on the 2D time-series representation of the hand shape. This time-series representation enables us to exploit the efficiency and accuracy of DTW in matching time-series representations. The time series of static hand gestures is rotation invariant. Moreover, DTW algorithm provides the mechanism to match templates even if some degree of scale, rotation or translation is needed to match. We demonstrate the efficiency and performance of our framework by comparing it with the FEMD framework introduced in [5].

II. RELATED WORK

Here, we give a brief overview of some of the current gesture recognition systems that are based on depth information. Raheja et al. [6] introduced a method for tracking fingertips and centre of palm using Microsoft Kinect. However, this method only detects the fingertips and centre of palm. It does not recognize hand poses as it assumes that the fingertips are always the closest part of the hand to the Kinect sensor. This method is designed mainly for tracking the fingertips and centre of palm.

Kulshreshth et al. [7] proposed a hand gesture recognition system using Microsoft Kinect. Here, they represent the hand shape using a Fourier descriptor based on a centroid distance function of equidistant contour points of the hand shape. A feature vector of the hand shape is extracted based on the Fourier descriptor. Gesture recognition is done by template matching. This method however imposes some constraints on the gesturer. It requires the fingers to be widely opened and straight, it requires the gesturing hand to be parallel with the camera plane and also that the hand shape should appear approximately in the middle of the captured depth image. This constraints reduces the degree of naturalness in gesturing.

Ren et al. [3]–[5] developed a hand gesture recognition framework based on the Microsoft Kinect. They introduced a metric called Finger-Earth Movers Distance (FEMD), as a dissimilarity measure between two hand shapes. This framework represents the input hand shape by global features. That is each finger in the hand shape represents a cluster in the signature. The hand signature is obtained from the time-series representation of the contour points of the hand shape. From this time-series, fingers are identified and used as clusters in the hand signature. Hand gesture recognition is done through template matching by measuring the FEMD distance between two hand signatures. This method produced better accuracy and robustness to cluttered background. However, as this method relies on the fingers to represent the hand shape, there is a problem in accurately segmenting the fingers from the time-series representation of the hand shape. Also, as will be described later, this technique usually has a high confusion rate between hand gestures.

DTW has been used as a similarity measure to match hand shapes in [8]. A 1D feature vector of Euclidean distances of contour points with centre of hand shape is defined and used as input to DTW. However, this 1D description does not accurately depict the topology of the hand shape. Also, the system is based on RGB camera and the background of the hand shape is restricted to white background only. Thus this system cannot be used in real world scenarios with complex backgrounds and lighting conditions.

Doliotis et al. [2] have demonstrated the efficiency of DTW in matching time-series representations of hand gestures. They have proposed a method for tracking dynamic hand gestures using Microsoft Kinect and subsequently employing DTW to recognize the tracked gestures. Converting dynamic hand gestures into time series is intuitive hence the application of DTW in recognizing dynamic hand gesture is natural. However,

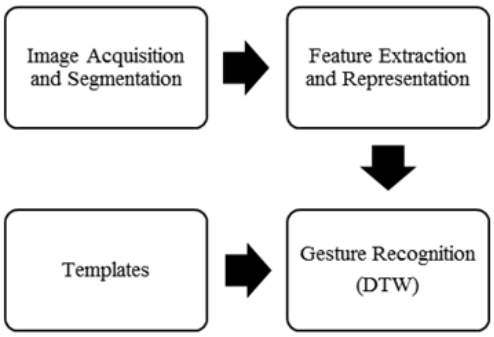

Fig. 1: Proposed Hand Gesture Recognition

representation of static hand gesture into contour information as time series is a research challenge of using DTW.

III. HAND GESTURE RECOGNITION SYSTEM

In this section, we describe our proposed Hand gesture recognition system in detail. The system is depicted in Figure 1.

The system consists of three modules namely, image acquisition and segmentation, feature extraction and representation, and gesture recognition.

A. Image acquisition, segmentation, feature extraction and representation

We follow the same approach as in [5] for image acquisition, segmentation and feature extraction. This approach has proven to be robust and more accurate. Using the Kinect sensor, we capture the RGB image and the depth image of the gesturer. The depth values are stored in millimetres. The Kinect sensor has proven to be robust in capturing images for used in hand gesture recognition systems [5], [9], [10] . Before segmenting the hand shape or region of interest (ROI), some pre-processing is performed. This involves calibrating the RGB and Depth Images, such that, pixel (i, j) of the depth image corresponds or maps to pixel (i, j) of the RGB image, where i and j are pixel coordinates. The RGB image is also converted into grey-scale.

To extract the region of interest, first, we locate the smallest depth value from the depth image. This corresponds to the closest point of the hand from the camera plane. We call this value minimum-distance. Next, an emperical threshold value is added to the minimum-distance to give the segmentation threshold. This segmentation threshold is then used to segment the hand region from the rest of the image. This approach has proven to be robust in cluttered and noisy environments [9]. It is important to note that the hand should be the closest object to the camera for proper segmentation.

As described in [5], the user needs to wear a black belt on the wrist of the gesturing arm. This allows flexibility in gesturing as well as facilitates segmentation. As shown in figure 3(a, b), the segmentation threshold is determined empirically. The segmentation threshold is the sum of a minimum distance (MD) and a depth threshold (DT). The minimum distance (MD) is easily obtained from the depth image as the minimum value in the depth matrix. The depth threshold is estimated based on different possible orientations of the hand shape. After multiple measurements and testing, an upper value is chosen as the depth threshold (DT), such that, the sum of the depth threshold and the minimum distance will allow us to isolate or segment the hand shape including the black belt from the rest of the image. This sum, (DT + MD) is referred to as the segmentation threshold. In our scenario, the depth threshold was estimated at 200mm. After segmentation, the hand shape is ready as input for the next module. Figure 2 depicts the steps of image acquisition and segmentation.

In feature extraction, the choice of the features should be such that the system will present a high degree of invariance to scaling, translation, and rotation. The representation of the features depends on the algorithm to be used to recognise the gestures. We choose the contour of the hand shape as our features as shown in Figure 4. We then represent these features as a time-series [5]. Contour information has been used successfully by [5] and [7]. In [5], the information is represented as a time-series. Whereas in [7], representation is based on a discrete Fourier transform of the contour information. As described in [5], the time-series representation of contour points records the relative distance between each contour vertex and the centre point. See Figure 5.

A plot of the time-series curve for the corresponding contour points shown in figure 5 is given in Figure 6. Figure 6 clearly shows that this time-series representation of the hand shape preserves all the contour information, making it suitable for use as a discriminant feature for classification. As opposed to [5], who uses the time-series curve to extract finger

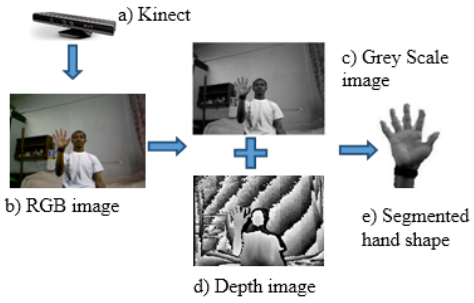

Fig. 2: Image acquisition steps (a-e)

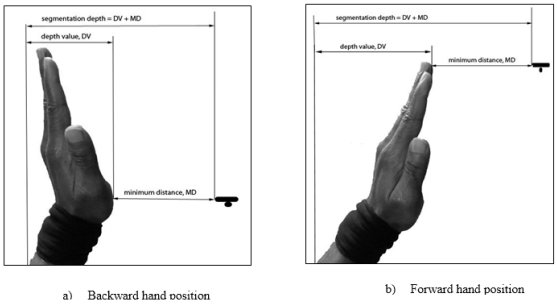

Fig. 3: Determining the depth threshold

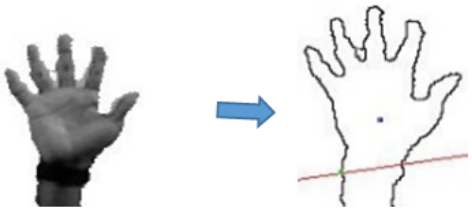

Fig. 4: Extracting the contour points from the hand shape. Green dot represents the initial points and the blue dot represents the centre of hand shape

clusters used to represent a hand signature, we use the entire time-series curve to define our feature vector. In [5], the intersection points of the finger clusters with a horizontal line drawn across the time-series is used to define the feature vector. However, we use the entire time-series as a 2D feature vector, consisting of the Euclidean distances of each contour point in one dimension and the angle this contour point makes with the initial point relative to the centre as the second dimension.

We convert the time-series into a 2D feature vector, f. That is, f is defined as follows:

$$f = \{(d_i, a_i), ..., (d_n, a_n)\},$$

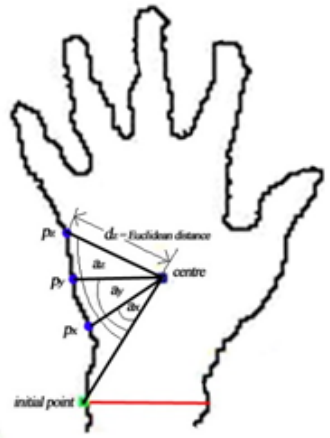

Fig. 5: Extraction of Euclidean distance of each contour point with centre of hand and angle made by each point with the initial position.p_x, p_y and p_z are contour points.

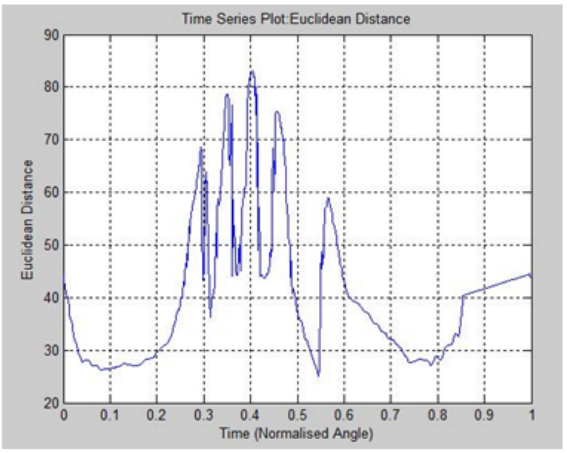

Fig. 6: Time-Series plot of hand shape in Fig. 4(b)

where d_i and a_i are respectively the Euclidean distance and angle of vertex i and n is the number of vertices in the boundary of the extracted hand shape. After this step, the feature vector representation f is ready to be passed as input to DTW in the next module. As mentioned in the related works, [8] has used contour information to define a 1D time-series that consist of the Euclidean distance of each contour point with the centre of hand shape . DTW is then used with this 1D time-series to classify gestures. However, using a 1D time-series does not preserve much of the discriminatory information or topology of the hand shape as

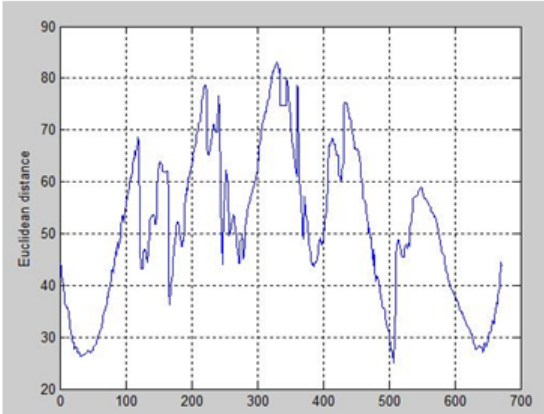

Fig. 7: 1D time-series of hand shape of Fig. 4(b)

Fig. 8: Representative of each class of gestures in our data set. Classes are numbered 1 to 10

shown in Figure 7. Figure 7 is a 1D time-series of the corresponding hand shape of Figure 5. As we can see, the topology of the hand shape is not clearly depicted as opposed to the one in Figure 6. Therefore, based on preserving discriminatory information as required in classification problems, the 2D time-series, consisting of the Euclidean distances of each contour point and the corresponding angle this points makes with the initial point, preserves discriminatory information better than the 1D definition. Thus it should produce better classification time-series than its 1D counterpart. This is clearly demonstrated in our classification results as compared to that in [8].

B. Gesture Recognition (Template Matching using DTW)

Our gesture recognition step involves a similarity measure through template matching using DTW. DTW has been used for recognising dynamic hand gestures. In this step, DTW is applied between the feature vector representation of the incoming unknown hand shape and a set of predefined templates stored in our database. The unknown hand shape is classified as belonging to the class of the template with which it has the minimum DTW value. That is the unknown gesture is classified under class c, such that

$$c = arg \ \ minDTW(U, T_c)$$

Where U is an unknown input gesture representation, T is a template from the database of template and c is the class label of T.

A nice and brief description of DTW algorithm is given in [2]. The effectiveness of DTW in this case can be explained by the fact that, the more points there are in the feature vector, the more accurate DTW is in matching patterns. Thus, because we use the entire contour of the hand shape, coupled with the fact that our feature representation in a 2D time-series preserves most of the discriminatory information, we expect to get better results. In [2], DTW has been used to classify dynamic hand gestures instead of static hand gestures as is our case. The feature vector in [2] consists of the centroid position of each frame (a static image) in the path of the dynamic hand gesture. Thus, multiple static images (frames) are used to form a feature vector in [2] without making use of the contour information.

IV. EXPERIMENTAL RESULTS

We created a new dataset which contains 200 samples (200 pairs of RGB and depth data). Our dataset can be found in [11]. We had two people perform each gesture type 10 times. We combined our dataset with the dataset provided in [5] which contains 1000 samples. So in total we used 1200 samples to evaluate our system. These samples are all taken in complex scenes with cluttered background.

This dataset defines 10 different classes of hand gestures as shown in Figure 8. We used this dataset to create a database of template against which our system was evaluated. The database contains 200 templates, created from 200 samples. Each class of gesture above is represented by 20 templates in the database.

All the experiments were conducted on an Intel Core i3-2120 CPU @ 3.30 GHz with 4GB of RAM. This was running a Window 7 32-bit operating system.

TABLE 1: Experimental results per class

Class Label	True	False
1	97	3
2	92	8
3	95	5
4	89	11
5	88	12
6	96	4
7	95	5
8	98	2
9	96	4
10	100	0

Our images were captured using Microsoft Kinect for Windows SDK version 1.8.

We experimented on a series of test cases with unclassified gestures from each class in order to determine the accuracy and running time of our system. Specifically, we tested 100 unclassified gestures from each class. The results are shown in the Table 1. From the results, we obtained a mean accuracy of **94.6%** and a mean running time of **0.5179s**. In addition to a better performance of our system in terms of accuracy and running time, our system also demonstrated great robustness to cluttered background. The images in our dataset as well as those provided in [5] were taken in challenging environments with different lighting conditions. This robustness to cluttered background is because the extraction of the hand shape or region of interest is based on depth information from the depth data provided by the Microsoft Kinect. Thus the system is unaffected by any configuration of the background.

Also, our system has proven to be invariant to scale, translation and rotation. The user is free to gesture in any position and orientation. This is because our feature vector is based on contour information. Because the initial point as shown in Figure 5 is always fixed, the contour points of the hand shape represented as a time-series curve as described above remains unchanged no matter the orientation of the hand. We have intentionally taken hand gesture samples with different hand orientation from our — hand gesture dataset1 [11]. Due to the efficiency of DTW in matching time-series curves, the system is also invariant to scaling.

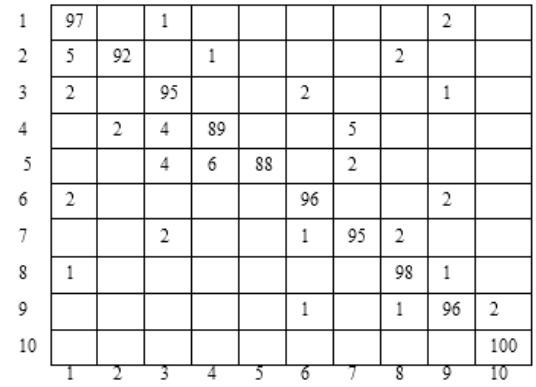

Fig. 9: Confusion matrix of our system

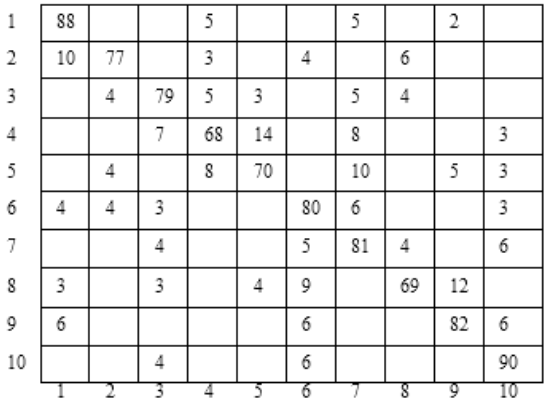

Fig. 10: Confusion matrix of our system with 1D feature vector

Also, the hand shape can be located on any part of the image as long as it is the closest object to the Kinect sensor. This is because segmentation is based on depth data, making the system translation invariant.

V. COMPARISON WITH THE FEMD FRAMEWORK

To further demonstrate the accuracy and efficiency of our system, we compare it to the FEMD framework [3]–[5], by developing the confusion matrix in Figure 9 and compare it to the one provided in [4].

There is a high confusion rate among all the classes of the FEMD system [4], particularly class 2 (which is the most confused class by 18%) is confused with class 1, 3, 4, 7, and 8. Whereas confusion rate of class 2 in our system is only 8% and the number of classes confused for class 2 is relatively small, mainly 3 classes (see figure 9).

TABLE 2: Comparison of Mean Accuracy and running time of our system with that of FEMD system [4]

	Mean Accuracy	Mean Running Time
Our System	94.6%	0.5179 s
Thresholding decomposition + FEMD	90.6%	0.5004 s

Figure 10 shows the confusion matrix of our system when implemented using a 1D feature vector, consisting of only Euclidean distances of contour points relative to the center point as in [8].

The average accuracy in this case is 78.4%, relatively lower compared to the one with a 2D feature vector.

However, this accuracy falls in the upper range of that given in [8]. Also, we can notice the high degree of confusion between the classes. The confusion rate is high and the highest confusion rate is registered between classes 4 and 5, and classes 8 and 9. Table 2 compares our mean accuracy and mean running time to that of FEMD system.

VI. CONCLUSIONS

We have designed and developed an accurate and efficient static hand gesture recognition system that performs well in uncontrolled environments. After evaluating our system with a gesture set of 10 classes, we obtained a mean accuracy of 94.6% and a mean running time of 0.5179s. Our system has also demonstrated better accuracy compared to FEMD system [4] (90.6%). The system has proven to be robust to cluttered background as well as insensitive to lighting conditions. The system is also invariant to scaling, translation and rotation. Our system has also demonstrated the power of DTW in matching time-series representations.

As our future work, we plan in studying the feasibility of upgrading our system to recognise dynamic hand gestures as well. Optimising the system for better accuracy and performance is also part of our future work. Finally, we intend to develop some real-life applications to demonstrate the efficiency and accuracy of our system.

REFERENCES

[1] S. Mitra and T. Ach, "Gesture recognition: A survey," *IEEE Transactions On Systems, Man, And Cybernetics*, vol. 37, no. Part C: Applications And Reviews, pp. 311–324, 2007.

[2] P. Doliotis, A. Stefan, C. Mcmurrough, D. Eckhard, and V. Athitsos, "Comparing gesture recognition accuracy using color and depth information," in *Proceedings Of The 4th International Conference On Pervasive Technologies Related To Assistive Environments*, 2011.

[3] Z. Ren, J. Meng, and J. Yuan, "Depth camera based hand gesture recognition and its applications in human-computer-interaction," in *Information, Communications And Signal Processing (ICICS)*, 2011.

[4] Z. Ren, J. Yuan, and 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*, 2011.

[5] Z. Ren, J. M. J. Yuan, and Z. Zhang, "Robust part-based hand gesture recognition using kinect sensor," *IEEE Transactions On Multimedia*, vol. 15, pp. 1110–1120, 2013.

[6] J. L. Raheja, A. Chaudhary, and S. K, "Tracking of fingertips and centre of palm using kinect," in *Proceedings Of The 3rd IEEE International Conference On Computational Intelligence, Modelling And Simulation*, 2011.

[7] A. Kulshreshth, C. Zorn, and J. J. L. Jr, "Poster: Real-time markerless kinect based finger tracking and hand gesture recognition for hci," in *3D User Interfaces (3DUI), 2013 IEEE Symposium On*, 2013.

[8] S. S. Jambhale and A. Khaparde, "Gesture recognition using dtw and piecewise dtw," in *Electronics And Communication Systems (ICECS), 2014 International Conference*, 2014.

[9] K. K. Biswas and S. K. Basu, "Gesture recognition using microsoft kinect," in *Automation, Robotics And Applications (ICARA), 2011 5th International Conference*, 2011.

[10] Y. Li, "Hand gesture recognition using kinect," in *Department Of Computer Engineering And Computer Sciences*, 2012.

[11] "Hand gesture dataset 1: [link will be provided upon request]."

Aboubakar Mountapmbeme is currently working as Assistant Programmer in the Computer Centre of the Islamic University of Technology (IUT), Dhaka, Bangladesh. He received his Bachelor degree in the department of Computer Science and Engineering (CSE) of IUT in 2014. He is an experienced software engineer with a lot of interest in research. His research interest lies within Software Engineering, Databases, Big Data analysis and Human-Computer Interaction. He is currently pursuing his M.Sc. in Computer Science and Engineering in the Department of CSE of IUT.

Hasan Mahmud has received his Bachelor degree in Computer Science and Information Technology (CIT) from Islamic University of Technology (IUT), Bangladesh in 2004. Then he joined as a faculty member in Computer Science and Engineering (CSE) department at Stamford University Bangladesh. He did his Master of Science degree in Computer Science from University of Trento (UniTN), Italy in 2009. He had received University Guild Grant Scholarship for the two years (2007-2009) Masters study and also awarded with early degree scholarship. He has different research articles published in several international journals and conferences. From 2009 he is working as an Assistant Professor in the department of Computer Science and Engineering (CSE) of Islamic University of Technology (IUT), Bangladesh. He is now pursuing his PhD degree in CSE under the guidance of Prof. Dr. M. A. Mottalib and Dr. Md. Kamrul Hasan at IUT. His research interest focuses on Human-Computer Interaction, Gesture based Interaction, Machine learning.

Md. Kamrul Hasan Md. Kamrul Hasan has received his PhD from Kyung Hee University, South Korea. Currently he is working as an Associate Professor of CSE Department in Islamic University of Technology (IUT), Gazipur, Bangladesh where he has been serving for ten years. Previously, He obtained a B.Sc. in CIT degree from IUT. He has long experience in software as a developer and consultant. His current research interest is in intelligent systems and AI, software engineering, cloud computing, data mining applications and social networking.