Vision Based Hand Gesture Recognition Using Dynamic Time Warping for Indian Sign Language

Washef Ahmed
Advanced Signal Processing Group
Centre for Development
of Advanced Computing
Kolkata, India

Kunal Chanda
Advanced Signal Processing Group
Centre for Development
of Advanced Computing
Kolkata, India

Soma Mitra
Advanced Signal Processing Group
Centre for Development
of Advanced Computing
Kolkata, India

***Abstract* — This paper presents an algorithm of Hand Gesture Recognition by using Dynamic Time Warping methodology. The system consists of three modules: real time detection of face region and two hand regions, tracking the hands trajectory both in terms of direction among consecutive frames as well as distance from the centre of the frame and gesture recognition based on analyzing variations in the hand locations along with the centre of the face. The proposed technique of ours overcomes not only the limitations of a glove based approach but also most of the vision based approach concerning different illumination conditions, background complexity and distance from camera which is up to 2 meters. Also by using Dynamic Time Warping Algorithm which finds the optimal alignment between the stored database features and query features, improvement in recognition accuracy is observed compared to conventional methods. Experimental results show that the accuracy is 90% in recognizing 24 gestures based on Indian Sign Language.**

***Keywords*— Human Computer Interaction, Hand gesture recognition, skin color segmentation, hand tracking, Dynamic Time Warping.**

I. INTRODUCTION

Who invented sign language? No one has the answer to this question, but it is most likely that the deaf themselves were the ones who created a variety of gestures in order to communicate. Although in India there are currently 22 official languages and over 415 other living languages. It is the Indian Sign Language (ISL) which is one of the living languages in India used by the Deaf and Dumb community. The prevalence of deafness in India is fairly significant. It is the second most common cause of disability. Approximately 63 million people (6.3%) in India suffer from significant auditory loss. With the implementation of training in Indian Sign Language and interpreter course in 2001, initiated by Ali Yavar Jung National Institute for the Hearing Handicapped (AYJNIHH), the usage of ISL is increasing. It has been reported that the number of Sign Language interpreters is significantly less than the required number to create a bridge of communication between the deaf and dumb and the ordinary people.

With the advancement of technology such bridge can be reduced. However for a Human Computer Interface researcher the task of recognizing hand gestures is highly challenging due to complex background, presence of non-gesture hand motions and different illumination environments. Based on the requirement we have started work to recognize gestures based on Indian Sign Language. Our focus is on some of the important domains like Medical, Police Station, Court, Bank/Post Office, Travel etc. Although most of the work in this area of research had tried to solve the problem by using gloves which are marked, markers or maintaining a simple background [1-3]. It is found that glove-based gesture interfaces require wearing a cumbersome device by the user carrying a load of cables that connect the device to a computer. Moreover for regular users the cost of the data glove is often too expensive. In another real-time gesture recognition system which can recognize American Sign Language letter spelling alphabet and digits but the gestures are static without any motion. Instead vision based methods are more natural and useful for real-time applications. It uses image processing algorithms to detect and track hand signs as well as facial expressions of the user. This approach is easier to the user since there is no need to wear any extra hardware. However, there are accuracy problems related to image processing algorithms and these problems are yet to be modified [4].

In the proposed framework we developed a hand gesture recognition system to recognize gestures in motion. It has been addressed previously that unlike most of the existing systems, our system neither users any marker nor instrumented gloves. This new barehanded technique that is proposed uses only 2D video sequences as input. The approach is to binarize the frame based on skin color so as to segment out the face region and the two regions of the hand. This involves detecting the face and hand locations. This is followed by tracking the trajectory of the moving hand. Here at first the centre of mass of the two hands along with the centre of the face is determined. Among consecutive frames the orientation of the two hands are considered while within each of the frame, distance of the centre of mass of the face region along with the centre of mass of the hand region is determined w.r.t the centre of frame which is fixed. With this analysis of the hand-position, variations are considered. Finally the obtained motion information is been used for

978-1-5090-1987-8/16/$31.00 ©2016 IEEE

recognizing the hand gestures based on signs. At this juncture arises a pertinent question which arises for measuring similarity between two temporal sequences which varies with time. In this paper the problem is addressed using Dynamic Time Warping. Although in this paper a scheme to recognize hand gesture recognition based on a low cost simple approach based on orientation and distance measure in the trajectory of hand movement is discussed. But our main contribution is in applying Dynamic Time Warping [7], [9] for efficient computation of similarity measure.

The rest of the paper is organized as follows:

Section II gives a brief overview of the gesture recognition scheme. The detection and localization of the face and hand regions are discussed in Section III. Section IV gives an explanation of tracking the trajectory of the moving hand approach. Section V demonstrates the recognition approach using Dynamic Time Warping for similarity measurement. The conclusion is given in Section VI.

II. OVERVIEW OF THE GESTURE RECOGNITION SCHEME

In this paper our main objective is to develop a low cost hand gesture recognition system based on Indian Sign Language. As such the system is considered within a low cost vision system which is executed in a common PC equipped with USB web cam. We know that a gesture is a specific combination of hand position, orientation and flexion observation at some time instance. Our recognition engine should be able to identify under different degrees of scene background complexity and illumination conditions. The algorithm proposed in our recognition engine is shown in Fig. 1:

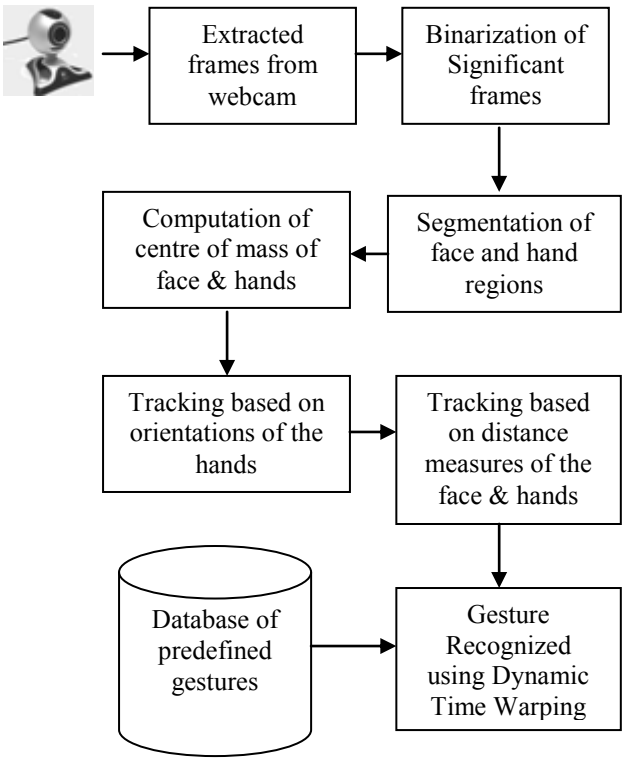

Fig. 1. Proposed System Overview of the Hand Gesture Recognition Scheme.

The proposed technique depends on the following algorithmic steps:

1. Frames from the webcam are captured and alternate frames considered as significant frames are taken as input.
2. Each of the significant frames are subjected to binarization based on skin color and pixels with motion difference between two consecutive frames.
3. Within each of the binarized frame, face region and hand regions are segmented out based on detection of the largest objects in the scene.
4. The centre of mass for the face and the two hand regions are determined.
5. For tracking, at first orientations of the centre of mass for each of the hands between consecutive frames are considered. A 8 neighborhood coded value is assigned based on the degree of orientation.
6. Next Euclidean distance measure is performed for the centre of mass of the face region and the two hands with the centre of the frame for each of the significant frames.
7. For pattern matching, a database of predefined gestures of each performed at a previous time are stored. Distance measure is performed with the feature vectors, one from the orientation and the other from distance. As numbers of frames vary with time for database and query gesture, Dynamic Time Warping algorithm is used. Same gestures show least distance.

III. DETECTION AND LOCALIZATION OF THE FACE AND HAND REGIONS

The proposed technique first considers extraction of frames from a real time video sequence for a given hand gesture. In our experiment we have considered 24 hand signs. The Fig. 2 shows some of the video sequences, one considered as pre defined gestures stored in database and the other as a query.

Fig. 2. Video sequences for the hand signs considered for the experiment for Check-In and Housekeeping.

It has been observed that frames with one interval apart are found to be quite significant. The Fig. 3 below shows some of the significant frames for a given hand sign.

Generally the frame capture rate is 25 frames/sec which consists of 1500 frames per minute. It has been observed that there is very negligible changes happening within consecutive frames starting from beginning. However significant change is realized at the feature level based on our experimental setup at an interval of 10 frames. So we have selected the frames at 10 frames apart which we have named as significant frame.

Fig. 3. Significant frames for a particular hand sign.

We next developed a real-time face localization and hand tracking method which is robust and reliable in complex background. For this the skin color detection and motion detection is jointly considered for binarization of the given frame[5].

The skin region can easily be detected using the color tone information. RGB based color is classified as skin if:

$$0.08 < \frac{3 * B * R^2}{(R + G + B)^3} < 0.12 \quad \text{AND}$$

$$1.5 < \frac{R * B * G^2}{G * B} < 1.8 \quad \text{AND}$$

$$\frac{R + G + B}{3 * R} + \frac{R + G + B}{R - G} < 1.4$$

It has been observed that the given skin color tone information may include a wide range of colors. In our system, the motion of the object provides important and useful information for object localization and extraction [6]. To find the movement information, we assume that the input gesture is non-stationary. When objects move in the spatial-time space (i.e, an image sequence), motion detector is able to track the moving objects by examining the local gray-level changes.

Let $F_i(x,y)$ be the ith frame of the sequence and $D_i(x,y)$ be the difference image between the ith and the $(i+1)$th frame defined as $D_i(x,y) = T_i\{| F_i(x,y) - F_{i+1}(x,y)|\}$

where T_i is a thresholding function which combined with skin tone pixels gives a reliable hand regions as well as the face portion.

$F_i(x,y)$ and $D_i(x,y)$ are all 640 X 480 images, and $D_i(x,y)$ is binary image defined as follows:

$$D_i(x,y) = \begin{cases} 1, & | F_i(x,y) - F_{i+1}(x,y)| \geq \text{threshold combined with } F_i(x,y) \text{ ^ } S_i(x,y) \\ & \text{where } S_i(x,y) \text{ indicates skin tone pixels} \\ 0, & \text{Otherwise} \end{cases}$$

The Fig.4 below shows the binarized image for each of the significant frames.

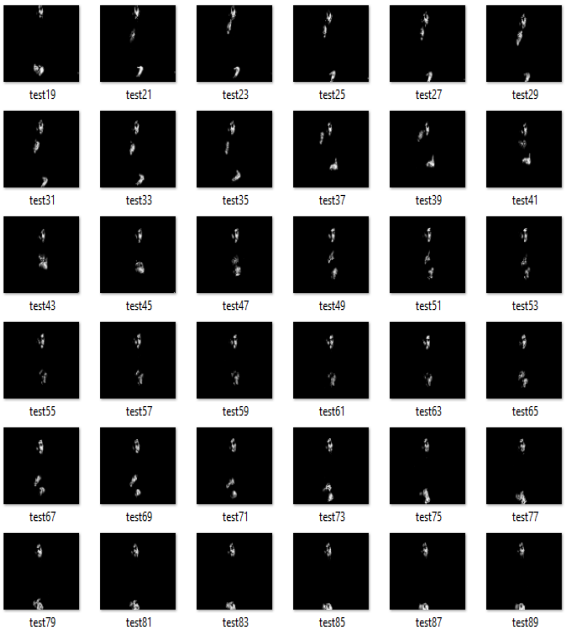

Fig. 4. Binarization of the significant frames.

This is followed by determining the centre of the hand regions and the centre of the face.

The binarized image is obtained based on adaptive skin color model. At first by gray level histogram analysis, skin region of detected face is obtained by eliminating eyes, nostrils, mouth. Color distributions in normalized red, normalized green and original red are assumed to be Gaussian distributions so that the means and standard deviations are calculated to build the adaptive skin color model[8]. Afterward, we used that skin color model to detect the other skin color regions for that person. Each of the significant frames are binarized based on skin color and pixels with motion difference between two consecutive frames. Within each of the binarized frame, face region and hand regions are segmented out based on detection of the largest three objects in the scene.

The binarized image so obtained has the face region, the two hand regions as the largest object. Although the hand shapes and the face region are not accurately binarized, but it still serves our purpose for locating the centre of mass for each.

In spite of the presence of small white regions, our algorithm is able to determine the above three regions as the largest object ignoring the others. This is due to the fact that our algorithm checks for the continuity of an object and is recursively called for that object until and unless that continuity is broken. Obviously what comes out as output are within the largest three objects in descending order of their size are the face region, the left hand and the right hand region.

This is then followed by determing the centre of mass of the three regions considering only the (x,y) co-ordinates of the pixels concerning each of the regions.

Thus it is ensured by the above steps of our algorithm that under complex background and within reasonable illumination variation, our algorithm sets up reliable information for tracking the trajectory of the moving hand. Fig. 5 shows some of the detected centre of mass for the face and hand regions, where the accuracy of determining centre of mass is 100%.

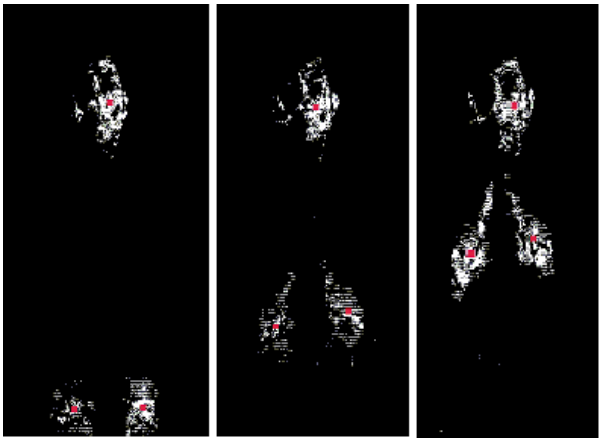

Fig. 5. Detection of the centre of mass of the face and hand regions.

IV. TRACKING THE TRAJACTORY OF THE MOVING HAND

The overall system for hand region tracking has two stages: In one stage tracking is done based on the analysis between two consecutive frames in motion, i.e inter frames and in the other analysis is done within a frame, i.e intra frame.

The first stage focuses on the motion information based on the orientation of the centre of mass of the two hands for a frame subjected to considering the previous frame as the reference frame.

If for the *i*th frame for a given hand region the centre of mass is considered as the origin of reference, then in the (*i*+1) th frame the location of the centre of mass is assigned a coded value within and on 0-7. This is based on the angle computed between the two consecutive centre of mass using the formula:

$$\theta_{(a,b)} = \tan^{-1}(\Delta y / \Delta x)$$

where Δy and Δx is the difference in the y and x coordinate between two consecutive centre of mass.

The eight neighboring regions has within it a measure of 45°. The Fig. 6 below shows demonstrates the value assigned for each of the eight neighborhood regions

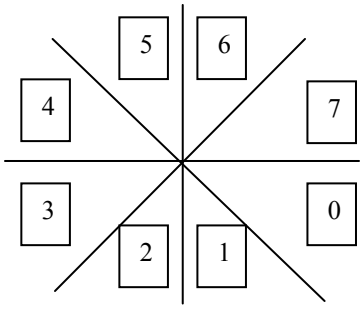

Fig. 6. Assigned values to the eight neighborhood regions.

The Fig. 7 below shows how the orientation is determined between two consecutive frames in a given sequence. For the example below the coded value of the below sequence is 6.

This is carried out between all the consecutive frames for both the left as well as the right hand.

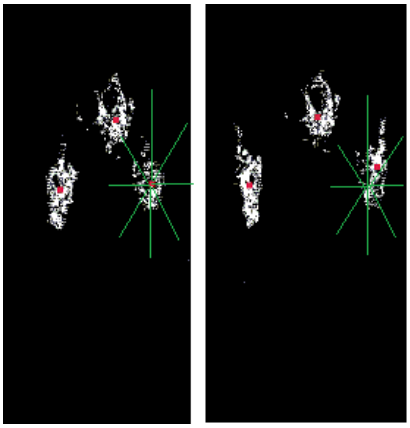

Fig. 7. Assigning coded values based on orientations of two consecutive centre of mass

The second stage focuses on the distance computed between the centre of mass of the face region as well as the centre of mass of the hand regions with the centre of the frame as it is static. We use the Pythagoras's formula as the distance metric:

$$d_{(a,b)} = \text{sqrt}(\Delta x^2 + \Delta y^2)$$

where Δy and Δx is the difference in the y and x coordinate between the centre of mass of any of the three regions w.r.t the centre of the frame.

The Fig. 8 below shows the distance measured across all the three objects w.r.t the centre of the frame.

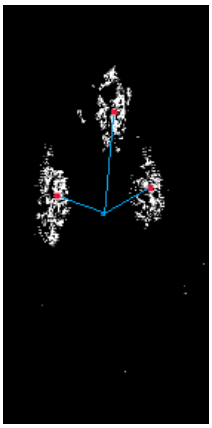

Fig. 8. Distance measured across all the three objects w.r.t centre of frame.

Here in our work we are considering tracking the hands trajectory both in terms of direction among consecutive frames as well as distance from the centre of the frame and gesture recognition based on analyzing variations in the hand locations along with the centre of the face.[10] Hence along with distance measure, orientation of the hand motion as feature vector is also considered which guarantees that two different gestures will never have more or less same distance measure.

Thus we have obtained the feature vectors having sufficient information about every movement and in which direction. From the first stage we have obtained two coded values for each of the hands and from the second stage we have obtained the distance measures of the centre of mass of the face region as well as the two hand regions with the centre of the frame.

For each frame the feature vector length is five and thus the total feature vector length for a given sequence is (N X 5) where N is the number of frames.

These feature vectors for a given gesture are computed and stored in a database as predefined gestures. For our experiment we have computed for 24 of such gestures. Now for recognition we have considered for each of the gestures hand signs taken at a different time by the same signer as query gestures. In this case feature will be of non uniform length.

V. HAND GESTURE RECOGNITION BASED ON DYNAMIC TIME WARPING

After extraction of feature vectors for a query gesture, it is compared with all the feature vectors of the predefined gestures stored in database to find the closest match.

The Fig. 9 below shows the responses due to the coded values based on the orientations for the left hand. The first two graph shows more similarity as they are from the same gesture as compared to the third graph which is from a different gesture. This holds true for right hand. It is to be noted that in each of the graphs, x-axis holds values for number of frames while y-axis holds coded values for each of the frames.

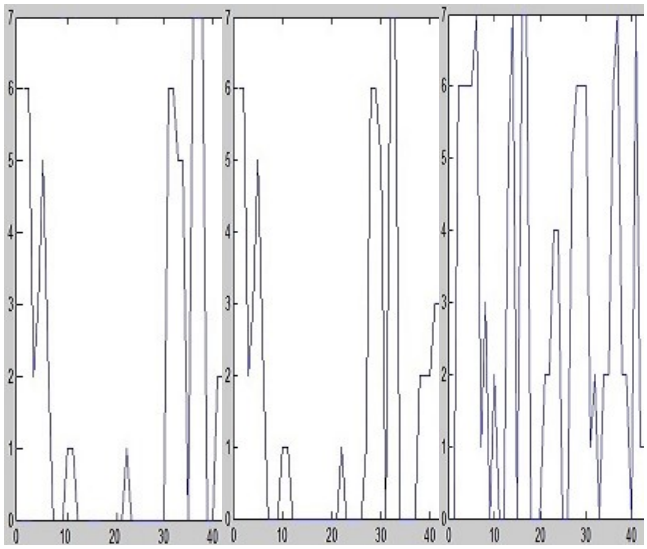

Fig. 9. Responses due to orientations for the left hand. The first two are from the same hand sign taken at a different time while the third is a different hand sign.

The Fig. 10 below shows the responses due to the magnitude of the distances based on the distance of the centre of mass of the left hand with the centre of the frame. The first two graph shows more similarity as they are from the same gesture as compared to the third graph which is from a different gesture. This holds true for the right hand as well as for the face region. It is to be noted that in each of the graphs, x-axis holds values for number of frames while y-axis holds distances for the left hand with the centre of the frame.

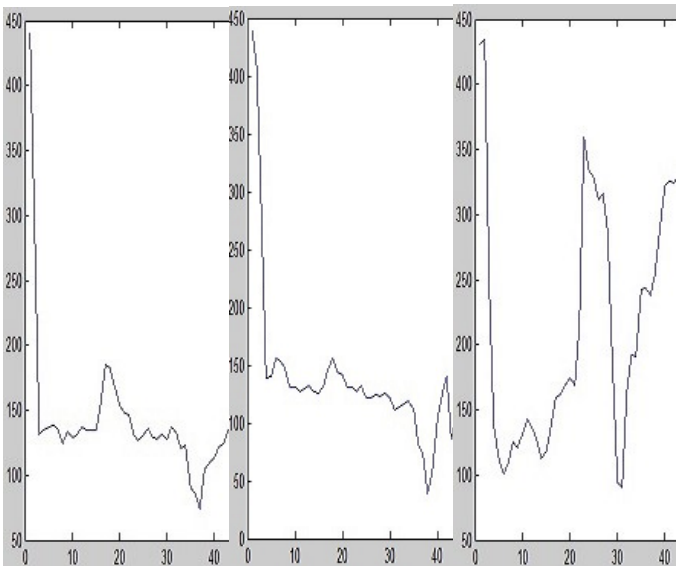

Fig. 10. Responses due to distances for the left hand with the centre of the frame. The first two are from the same hand sign taken at a different time while the third is a different hand sign.

In the present paper we have used Dynamic Time Warping (DTW) algorithm to find an optimal alignment between two given time dependent sequences, under certain restrictions of non uniform length. The sequences in our case are generated from both the database as well as query image frames of the 24 gestures used in this paper. Intuitively, the sequences are warped in a nonlinear fashion to match each other. It has been our observation that DTW has been successfully applied to automatically cope with time deformations and different speeds associated with time-dependent data generated from each of the gestures.

Our objective is to compare two time-dependent sequences $X := (x_1,x_2,x_3,.....x_N)$ of length N and $Y := (y_1,y_2,y_3,.....y_M)$ of length $M \in N$. These sequences are discrete signals (time-series) or, more generally, feature sequences sampled at equidistant points in time. DTW graph is generated between each of the sequences as shown in Fig. 11.

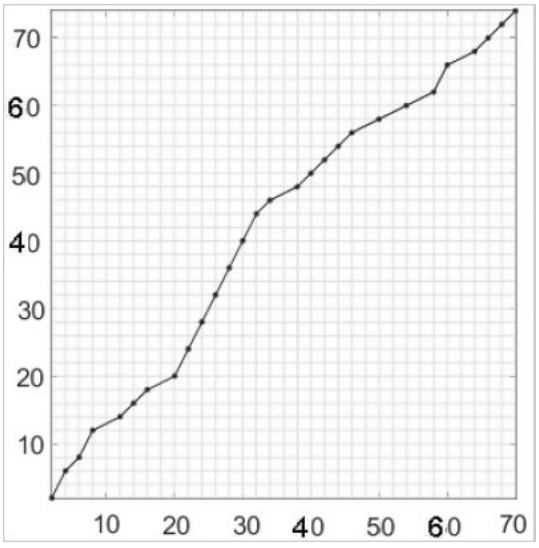

Fig. 11. Dynamic Time Warping graph generated using sequences used from our hand gesture recognition database.

In the following, we fix a feature space denoted by $\mathcal{F}$. Then

$x_N, y_M \in \mathcal{F}$ for $n \in [1:N]$ and $m \in [1:M]$.

Again, to compare two different features $x, y \in \mathcal{F}$, one needs a local cost measure, sometimes also referred to as local distance measure, which is defined to be a function

$c : \mathcal{F} \times \mathcal{F} \rightarrow \mathbb{R}_{\geq 0}$.

On evaluating the cost function c(x,y) is found to be smaller when x and y are sequences of the same gesture, otherwise c(x,y) is large. Evaluating the local cost measure for each pair of elements of the sequences X and Y, we obtain the cost matrix $C \in \mathbb{R}^{NXM}$ defined by $C(n,m) := c(x_N, y_M)$. This is represented in Table II below.

Since there are no publicly released dataset for dynamic gesture recognition, we collected a dataset of twenty-four Indian Sign Language (ISL) gestures included from domains like, Medical, Court, Hotel etc. Each of the domains have eight different signs which are performed twice by two different signers. From one, the gesture features are stored in the database as predefined gestures which are made to compare with the features from the other as query gesture.

The results are shown in following Table I. Experimental results show that the recognition rates are approximately 90% for dynamic hand gestures.

TABLE I. EXPERIMENTAL RESULTS SHOWING RECOGNITION RATE AROUND 90%

	Hand Gestures based on ISL	**Accuracy**
Dynamic Hand Gesture	Blanket	87.2%
	Buffet	91.37%
	Check In	89%
	House Keeping	85.47%
	Laundry	89.13%

In Table II below a snapshot is provided considering 5 pairs of gestures out of 24. Considering the cost measure value for each pair of gestures for right hand, we observe that for pairs of the same gesture from database and query sequences, the cost measure value is minimum. This observation is made across 21 such pairs of gestures where the cost measure value for similar gestures is less as compared to the cost measure value from other gestures. The recognition rate of this system is approximately 90%.

Most of the work is based on American Sign Language. In one of such work classification of hand gesture as a similarity measure using Dynamic Time Warping (DTW) and piecewise DTW is performed using effective contour signature achieving a success rate of 84% approximately. In another work where sign language recognition using Dynamic Time Warping for sign trajectory and Histogram of Oriented Gradient (HoG) for shape representation is performed, an 82% accuracy in ranking signs in the 10 matches is obtained.

Our work is based on Indian Sign Language using orientation and distance measures with DTW for similarity measurement among different gestures obtaining a recognition accuracy around 90%.

TABLE II. SNAPSHOT OF THE TOTAL EXPERIMENTAL RESULTS CARRIED OUT ON TWENTY-FOUR HAND SIGNS

		Database									
		Blan-ket (69)	Blan-ket (48)	Buff et (95)	Buff et (80)	Chec kin (85)	Chec kin (83)	House keep ing (131)	House keep ing (155)	Laun dry (74)	Laun dry (73)
Query	Blanket (69)	0	**3**	36	34	18	5	33	40	14	19
	Blanket (48)	**3**	0	42	40	31	7	29	25	26	18
	Buffet (95)	36	42	0	**1**	28	42	168	113	25	32
	Buffet (80)	34	40	**1**	0	26	40	160	105	25	31
	Check In(85)	18	31	28	26	0	**3**	51	47	32	33
	Check In(83)	5	7	42	40	**3**	0	47	43	28	20
	House keeping (131)	33	29	168	160	51	47	0	**10**	25	30
	House keeping (155)	40	25	113	105	47	43	**10**	0	21	26
	Laundry (74)	14	26	25	25	32	28	25	21	0	**2**
	Laundry (73)	19	18	32	31	33	20	30	26	**2**	0

VI. CONCLUSION

In this paper, a technique along with application of Dynamic Time Warping methodology is used to perform similarity measurement efficiently. The proposed technique increases the adaptability of a hand gesture recognition system. The technique works well under different degrees of scene background complexity and illumination conditions.

REFERENCES

[1] C. Maggioni, "New ways of Operating a Computer", *Proc.* Int'l Workshop Automatic Face and Gesture Recognition, 1995.

[2] E. Hunter, J. Schlenzig and R. Jain, "Posture Estimation in Reduced-Model Gesture Input Systems", Proc. Int'l Workshop Automatic Face and Gesture Recognition, pp. 296-301, 1995.

[3] A. Bobick and A. Wilson, "A state-based technique for the summarization and recognition of gesture", In Proc. IEEE Fifth Int. Conf. on Computer Vision, Cambridge, pp. 382-388, 1995.

[4] Y. Wu and T.S. Huang, "Hand Modelling, Analysis and Recognition for Vision-Based Human Computer Interaction", IEEE Signal Processing Magazine, pp. 51-60, 2001.

[5] M. J. Jones and J. M. Rehg, "Statistical color models with application to skin detection", Proceedings of the International Conference on Computer Vision and Pattern Recognition, vol 1, 1999, pp. 274-280.

[6] J. Lin, Y. Wu and T.S. Huang, "Modeling the constraints of human hand motions", Proceedings of Workshop on Hand Motion, pp. 121-126, 2000.

[7] M. Muller, "Information Retrieval for Music and Motion" Book chapter "Dynamic Time Warping", Springer ISBN: 978-3-540-74047-6, pp. 318, 2007.

[8] D. H. Liou, Chen-Chiung Hsieh and David Lee, "A real-time hand gesture recognition system by adaptive skin-color detection and motion history image", Dept. of CSE, Tatung University, Taipei, Taiwan, Reallusion Inc, Taiwan, at 2nd International Conference on Signal Processing Systems (ICSPS), Dalian, 2010.

[9] S. Sandeep and A. Khaparde, "Gesture recognition using DTW & piecewise DTW", International Conference on Electronics and Communication Systems (ICECS), India, 2014.

[10] P. Jangyodsuk, C. Conly and V. Athitsos, "Sign Language Recognition using Dynamic Time Warping and hand shape distance based on Histogram of Oriented Gradient features", 7th ACM International Conference on Pervasive Technologies Related to Assistive Environments, Greece, 2014.