World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:6, No:8, 2012

FSM-based recognition of dynamic hand gestures via gesture summarization using key video object planes

M.K. Bhuyan

International Science Index, Computer and Information Engineering Vol:6, No:8, 2012 waset.org/Publication/11224

Abstract—The use of human hand as a natural interface for human-computer interaction (HCI) serves as the motivation for research in hand gesture recognition. Vision-based hand gesture recognition involves visual analysis of hand shape, position and/or movement. In this paper, we use the concept of object-based video abstraction for segmenting the frames into video object planes (VOPs), as used in MPEG-4, with each VOP corresponding to one semantically meaningful hand position. Next, the key VOPs are selected on the basis of the amount of change in hand shape – for a given key frame in the sequence the next key frame is the one in which the hand changes its shape significantly. Thus, an entire video clip is transformed into a small number of representative frames that are sufficient to represent a gesture sequence. Subsequently, we model a particular gesture as a sequence of key frames each bearing information about its duration. These constitute a finite state machine. For recognition, the states of the incoming gesture sequence are matched with the states of all different FSMs contained in the database of gesture vocabulary. The core idea of our proposed representation is that redundant frames of the gesture video sequence bear only the temporal information of a gesture and hence discarded for computational efficiency. Experimental results obtained demonstrate the effectiveness of our proposed scheme for key frame extraction, subsequent gesture summarization and finally gesture recognition.

Keywords—Hand gesture, MPEG-4, Hausdorff distance, Finite state machine.

I. Introduction

One very interesting field of research in Pattern Recognition that has gained much attention in recent times is Gesture Recognition. Gesture may be described as the manner in which a person moves his body and limbs to express an idea or sentiment. People frequently use gestures to communicate in their day-to-day life. Therefore, gestures are a natural means of conveying information. This has motivated to use gestures for communicating with computers. Thus, gestures provide an attractive and user-friendly alternative to interface devices like keyboard, mouse and joysticks in human-computer interaction (HCI). Accordingly, the basic aim of gesture recognition research is to build a system which can identify/interprete specific human gestures automatically and use them to convey information (*i.e.*, communicative as in sign-language communication) or for device control (*i.e.*, manipulative as in controlling robots without any physical contact between human and computer). One type of human gesture of particular interest is hand gesture where the position, shape and motion of the hand convey information. But, since the meanings of hand gestures depend on people and their culture, it is generally not possible to derive a universal model necessary for hand gesture recognition. However, it is possible to define a set of application specific hand gestures that can be modelled appropriately thereby reducing ambiguity in recognition.

M.K. Bhuyan is with the Department of Electronics and Electrical Engineering, Indian Institute of Technology Guwahati, Guwahati, India 781039 E-mail: mkb@ iitg.ernet.in .

The task of locating meaningful patterns from a stream of input signal is called pattern spotting [1]. Gesture spotting is one of the challenging aspect in the field of gesture recognition, where it is required to detect the start point and the end point of a gesture pattern. This difficulty is due to two aspects of signal characteristics: Segmentation ambiguity [2] and Spatio-temporal variability [3]. The segmentation ambiguity may be attributed to the unintentional hand movements between different gestures resulting in co-articulation. The other difficulty in gesture spotting comes from the fact that the same gesture varies in shape and duration from one signer to another. The variation also exists even when the same gesture is made by a particular signer at different times. An ideal gesture recognizer should be capable of extracting gesture segments correctly from a continuous input video sequence and matching them with reference patterns or templates allowing a wide range of spatio-temporal variability.

As with any Pattern Recognition problem, the task of gesture recognition also consists of two major components: Feature extraction and Classification. Accordingly, in hand gesture recognition it is necessary that the static and/or dynamic configuration of the human hand be measurable by the machine, thereby forming gesture features to be used in classification. Initial attempts to measure hand configuration resulted in mechanical devices, *e.g.*, glove-based devices, that directly measure hand/arm joint angles and their spatial position. For example, Fels and Hinton used data gloves and Polhemus sensors to extract 3D hand location, velocity, and orientation. Feature vectors were then formed to represent hand gestures and were used to train a multilayer neural network for translating hand gestures to synthesized speech in [4] and [5]. Some other Polhemus/Glove-based hand gesture recognition techniques include [6], [7], [8] and [9]. But, glove-based gestural interface requires the user to wear a cumbersome glove that carries a load of cables connecting it to the computer. This hinders the naturalness with which the user can interact with the computer.

Awkwardness in using gloves is overcome by using vision-based non-contact interaction techniques. They use color-based vision segmentation, silhouettes or edges to track the hand and fingers. Unfortunately, most of the works on vision-based

International Science Index, Computer and Information Engineering Vol:6, No:8, 2012 waset.org/Publication/11224

gestural HCI have mainly been focused on the recognition of static hand gestures or postures. But, hand gestures are in general dynamic actions where the motion of the hands conveys as much information as their posture does. So, appropriate interpretation of dynamic gestures on the basis of hand movement in addition to shape and position is necessary.

All vision based approaches to hand/finger tracking require to locate hand regions in video sequences. An approach based on the 2D locations of fingertips and palms was used by Davis and Shah as early as in 1994 [10]. Skin color offers an effective and efficient way to segment out hand regions. However, many of these techniques are plagued by some difficulties such as large variation in skin tone, unknown lighting conditions and dynamic scenes. A solution to this is the 3D model-based approach, in which the hand configuration is estimated by taking advantage of 3D hand models [11]. However, they lack the simplicity and computational efficiency. An alternative approach is the appearance-based modeling [12]. Although it is easier for the appearance-based approach to achieve user-independence than the 3D model-based approach, there are two major difficulties associated with this approach, *viz.*, automatic feature selection and training data collections. Another important approach for hand tracking is the prediction algorithm combined with Kalman tracking and application of probabilistic reasoning for final inference [13], [14]. But, this approach requires fine initial model and also demands additional computations whenever there is any rotation and change in shape of the model. Moreover, it is very much sensitive to the background noise and is based on small motion assumption that often fails to hold in hand motion.

An HMM-based hand gesture recognition system was developed by Yamato *et al.* in [15] in which he made use of the mesh features of 2D moving human blobs such as motion, color and texture, to identify human behavior. In the learning stage, HMMs were trained to generate symbolic patterns for each action class, and the optimization of the model parameters was achieved by forward-backward algorithm. In the recognition process, given an image sequence, the output result of forward calculation was used to guide action identification. However, the selected mesh features do not match the human posture space well because it is sensitive to position displacement. Later several other HMM-based hand gesture recognizers were developed in [16] and [17].

A conditional density propagation (CONDENSATION) algorithm was proposed by Isard and Blake in [18], where they adopted the stochastic differential equation to describe complex motion model, and combined this approach with deformable templates to cope with people tracking. This idea was extended by Black and Jepson to recognize gestures [19]. However, modeling the dynamics alone may not be sufficient for dynamic hand gesture recognition. The Finite State Machine is a usually employed technique to handle this situation. Some of the state-based approaches had been proposed in [20], [21] and [22]. But, all these algorithms use each and every frame in the gesture sequence to build up the gesture model. This seems to be computationally inefficient. Since there exists large amount of temporal redundancy between frames in a video sequence it is quite possible to obtain the gesture model from only a selected number of frames in the sequence.

In view of the various problems posed by one or the other algorithms mentioned here, in this paper, we propose an algorithm for dynamic hand gesture recognition that is based on finite state representation and summarization of gestures using key video object planes (VOPs). In our proposed technique, we use the concept of object-based video abstraction for segmenting the frames into VOPs, as used in MPEG-4, where hand is considered as a video object (VO). A binary model for the moving hand is derived and is used for tracking in subsequent frames. The Hausdorff tracker [23], [24] is used for the purpose.

Next, the key VOPs are selected on the basis of Hausdorff distance measure, thereby transforming an entire video clip into a small number of representative frames that are sufficient to represent a particular gesture sequence [25] and [26]. These frames are the key frames that best represent the content of the sequence in an abstracted manner. These key VOPs are the input to the gesture classification system that uses state based approach for representation and recognition of gestures.

Since a gesture can be defined as an ordered sequence of states in the spatial-temporal space, we represent a particular gesture as a sequence of key frames and the corresponding key frame duration, which constitute a finite state machine (FSM). For recognition, the shape similarity between the shapes of the incoming data sequence and the states of the FSM is measured by Hausdorff distance measure. We use Angular Radial Transformation (ART) based shape descriptor, as used for shape description in MPEG-7 multimedia content description interface [27], for indexing different FSMs during recognition process.

One notable advantage of our proposed scheme is that, it is robust to background noise. The tracker can track the hand as an object very efficiently even without adopting any kind of background filtering. Moreover, in the algorithm for VOP generation no extra computation on account of scaling and rotation is required, as is essentially required in tracking algorithms like Kalman filter based tracking. In our algorithm, the concept of " shape change" of the tracked object can accommodate scaling and rotation of the video object in successive frames of the gesture video sequence. The only computation required for the shape change is the model update in each and every frame of the video sequence. The model update computation using motion vector is much more computationally simpler than the other computations, *viz.*, affine transformation required for scaling and rotation. Compared to the HMM based systems, while the number of states and the structure of the HMM must be predefined, in our proposed approach, a gesture model is available immediately. The statistical nature of an HMM precludes a rapid training phase as pointed out in [21]. To train an HMM, well-aligned data segments are required, whereas in the FSM representation the training data is segmented and aligned simultaneously to produce a gesture model. Another advantage of using FSM is that it can handle gestures with different lengths/states. The only one input to the FSM is the spatio-temporal variance, which produces the recognizer after some training sessions. Our proposed FSM is more robust

International Science Index, Computer and Information Engineering Vol:6, No:8, 2012 waset.org/Publication/11224

to spatio-temporal variability of incoming gesture sequence, where FSM adapts quickly to accommodate this variability during the training phase. Unlike other template matching algorithms [28], [29] and [30], where an image sequence is first converted into a static shape pattern, and then compares it to prestored action prototypes during recognition, our proposed recognizer only compares the selected key frames of the incoming video sequence with the existing key frames in FSMs. The key frame based shape comparison greatly enhances the recognition speed.

The organization of the rest of the paper is as follows. In Section II we present our proposed recognition method. Subsection II-C shows the approach of key VOP extraction. Experimental results are shown in Section III. Finally, we draw our conclusion in Section IV.

II. Proposed Scheme for Hand Gesture Recognition

Fig. 1 shows the basic block diagram for the proposed hand gesture recognition system. From the input gesture video sequence VOPs for different hand positions are obtained. During this phase Hausdorff tracker tracks the change in hand positions from one frame to the next in the incoming gesture video sequence. Next, key VOPs are extracted by measuring shape similarity using Hausdorff distance measure. The key frame selection eliminates redundant frames. These representative frames are the inputs to our proposed gesture recognition module. Recognition is accomplished by matching the input state sequence with the states of different FSMs obtained during training – each FSM constructed through training corresponds to a particular gesture – the gesture corresponding to the matched FSM is the recognized gesture. However, instead of comparing the input to all the representative FSMs we propose to compare only the first state of the input to that of all the FSMs in the gesture vocabulary using ART shape descriptor. By doing so we can select one or few likely FSMs and discard all others from consideration in the recognition process. This greatly speeds up the recognition task.

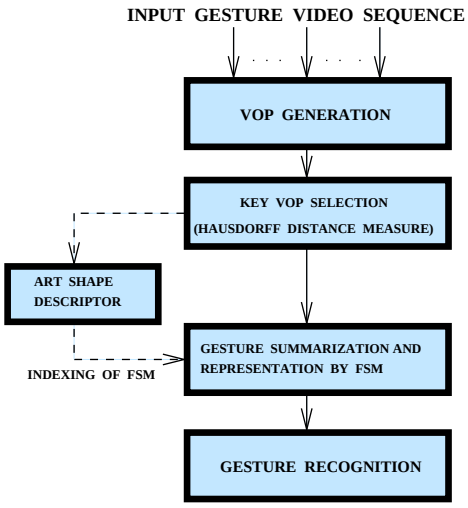

Fig. 1. Block diagram for the proposed scheme

A. Hausdorff distance

The Hausdorff distance measure can be used to measure the similarity between two shapes. It is defined as the maximum function between two sets of points O and I, as given below [24].

$$H(O,I) = max\{h(O,I), h(I,O)\} \quad (1)$$

where $h(O,I) = max_{o\in O} min_{i\in I} \parallel o - i \parallel$

and $h(I,O) = max_{i\in I} min_{o\in O} \parallel i - o \parallel$

Feature points are denoted by $o_1, ..., o_m$ for the set O and $i_1, ..., i_n$ for the set I. The computation of Hausdorff distance is done by distance transform algorithm, which is discussed in Appendix A.

B. Hand image segmentation and VOP generation

From the input video sequence VOPs for different hand positions are obtained. The VOP generation program may be divided into four stages as follow and as depicted in Fig. 2.

- Stage 1: Initial hand model extraction.
- Stage 2: Edge detection of input video sequences.
- Stage 3: Object tracking and model update.
- Stage 4: VOP extraction.

The initial model image is generated from the first two gray-scale images as shown in Stage 1 block of Fig. 2. This model is continuously updated in Stage 3 and is used for object tracking. The edge image for each frame is generated in Stage 2 and is also used in Stage 3. In Stage 4, VOPs are extracted from the corresponding updated model.

The core of this algorithm is an object tracker that matches a two-dimensional (2D) binary model of the video object against subsequent frames using the Hausdorff distance. The best match found indicates the translation the object has undergone, and the model is updated in every frame to accommodate for rotation and change in shape. However, the method will be effective only if the video object changes slowly from one frame to the next which we assume to be true in the present case.

Video object extraction and VOP generation algorithm was originally developed for use in object based video coding as in MPEG-4 [23], object based video abstraction for surveillance [31], etc. However, our method differs from the method in [23] and [31] in the sense that we propose to use median filtering and logical AND operation suiting our purpose of application, as explained below.

1) Initial hand model generation: Initial model is necessary for tracking of video object in successive video frames of the incoming video sequence. This is accomplished through the following steps.

Change detection: Generates change detection mask via thresholding of difference image formed from the two initial frames in a gesture sequence.

Median filtering: Removes noises in the threshold difference image thereby generating finer initial model.

Thinning: Reduces the width of the model by morphological thinning process.

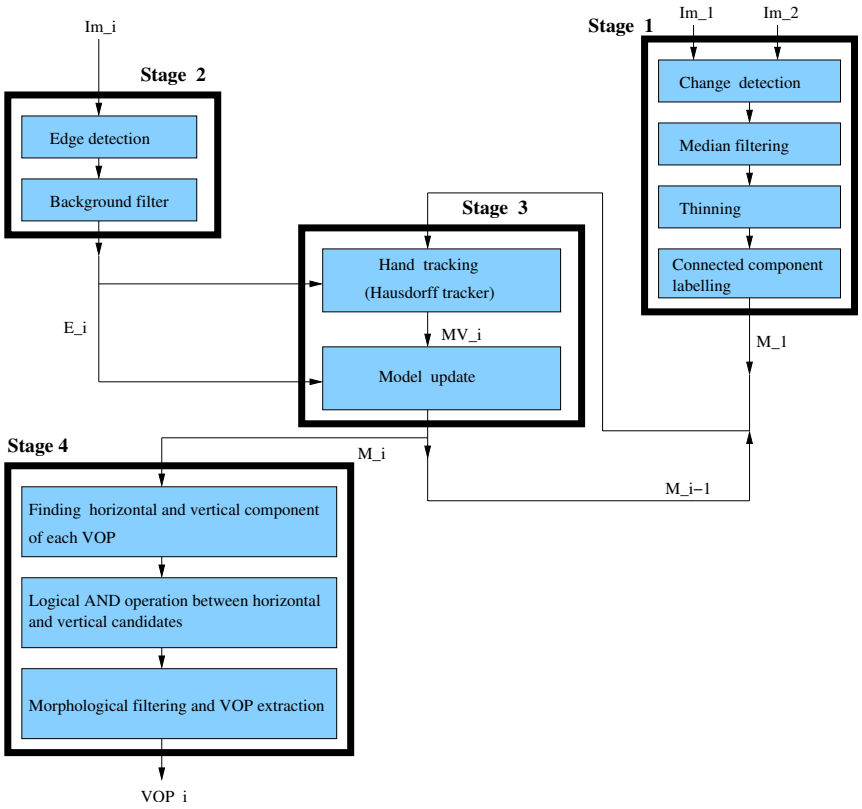

Fig. 2. Block diagram for the VOP generation algorithm

Connected component labelling: Eliminates short edges of the model.

An example of initial model generation using all these steps is demonstrated in Fig. 3.

2) Edge detection and background filtering: Once the initial hand model is obtained, the task is to determine in subsequent frames any change in the hand shape with respect to the initial model. Now, in order to reduce the computational cost, we propose to use edge images instead of the complete frames.

Background filtering is one desired step when VOPs are extracted from cluttered background. Removal of background edges reduces the number of points to be considered during the process of shape change detection and thus speeds up the tracking process. However, our algorithm is capable of functioning even without background removal as long as the background does not change significantly. This is because our algorithm uses large variation in hand shape for gesture modeling. Therefore, as long as the background edges are more or less stationary in a sequence the performance of our algorithm is not affected.

3) Hand tracking and model update: The Hausdorff object tracker finds the position where the input hand model best matches the next edge image and returns the motion vector MV_i that represents the best translation. As the tracked object moves through a video sequence, it may rotate or change its shape. To allow for this, the model must be updated every frame. The model image is updated by using the motion vector. First, the current model M_{i-1} is dilated and then shifted by motion vector MV_i. Then, the portion of edge image (E_i) that overlaps this shifted model is selected as the new model image M_i as shown in the block diagram given in Fig. 2.

4) VOP extraction: The region between the first and the last edge points in a row is a horizontal candidate for the object in a frame while that in each column is the vertical candidate. After finding all the horizontal and the vertical candidates in a frame, the VOP is generated by logical AND operation and further processed by alternative use of morphological operations like closing and filling. It is to be noted that in contrary with the existing VOP generation algorithm given in [23], we propose the logical AND operation between horizontal and vertical VOP candidates instead of logical OR operation. This is because, individual finger information of hand gestures is not obtained by using OR operation but is possible with AND operation. This is illustrated in Fig. 4. It is seen that for hand images, logical OR operation between vertical and horizontal candidates results in a blob with all fingers joined together and hence resulting in loss of finger information in the hand image as shown in Fig. 4(d). But, on the other hand, AND operation results in a hand figure as shown in Fig. 4(e), in which individual fingers are distinguishable as desired for our purpose. Finally, Fig. 5 shows the extracted VOPs from a long gesture video sequence.

C. Key VOP selection

After the VOPs are extracted, binary alpha planes are generated. A binary alpha plane indicates whether or not a

International Science Index, Computer and Information Engineering Vol:6, No:8, 2012 waset.org/Publication/11224

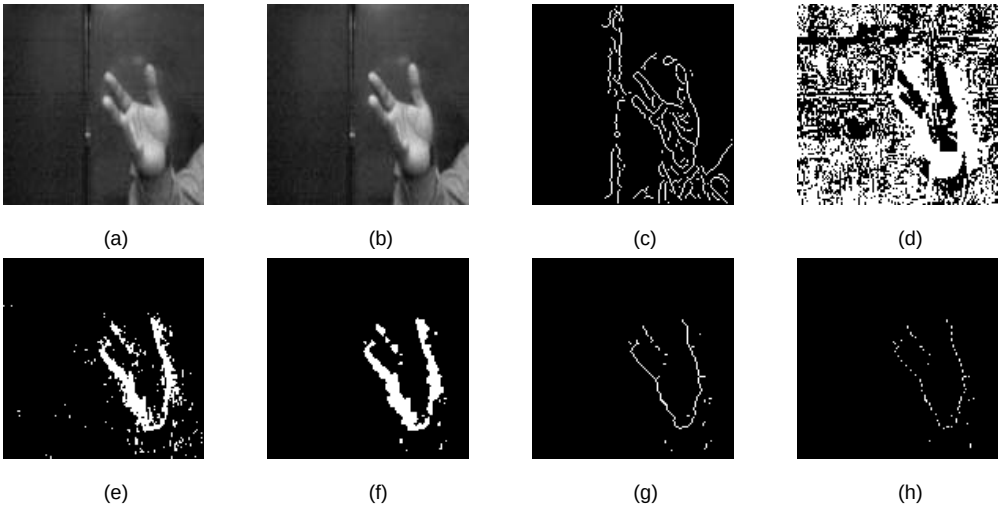

Fig. 3. Initial model generation, (a)-(b) Two initial video frames (c) Edge image of the first frame (d) Difference image (e) Threshold difference image (f) Median filtering of threshold difference image (g) Thinning (h) Connected component labelling

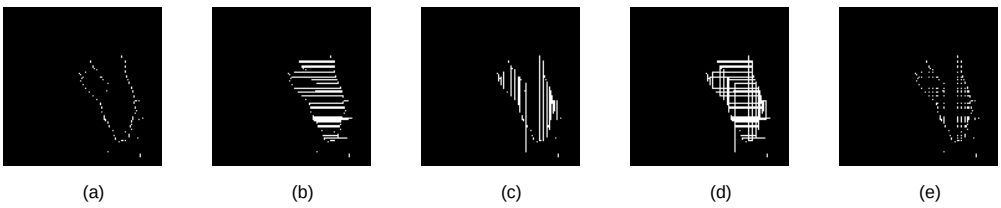

Fig. 4. Initial model generation, (a)-(b) Two initial video frames (c) Edge image of the first frame (d) Difference image (e) Threshold difference image (f) Median filtering of threshold difference image (g) Thinning (h) Connected component labelling

Fig. 5. Extracted VOPs from the hand video sequences

International Science Index, Computer and Information Engineering Vol:6, No:8, 2012 waset.org/Publication/11224

International Science Index, Computer and Information Engineering Vol:6, No:8, 2012 waset.org/Publication/11224

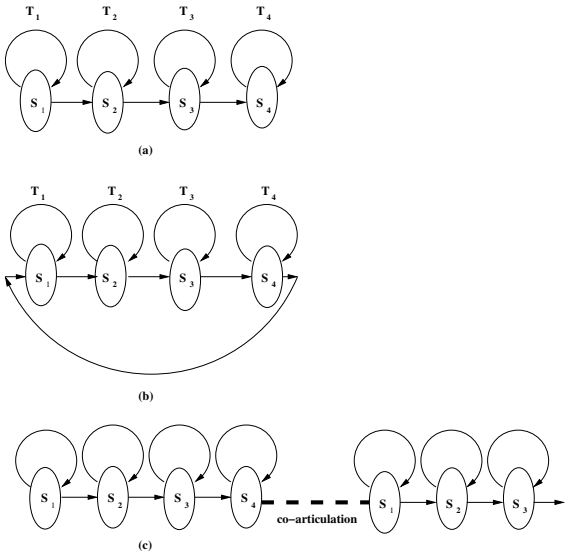

Fig. 6. (a) Finite state representation and summarization of a gesture. (b) FSM for the same gesture repeated again and again. (c) FSM for the gestures connected sequentially with co-articulation.

pixel belongs to a VOP. After getting the binary alpha planes, the key VOPs are selected on the basis of shape similarity measure using Hausdorff distance measure.

For key VOP selection, the first VOP of a video sequence is declared as a key VOP, and whenever the Hausdorff distance between the mass center aligned contours of a key VOP candidate and its temporally closest key VOP is larger than an adaptive threshold, the key VOP candidate is selected as a new key VOP [32]. The threshold is given by

$$T = \lambda\sqrt{min(M_1, M_2)^2 + min(N_1, N_2)^2} \qquad (2)$$

where λ is a predefined scale factor that is constant for all VOPs, M_1 and N_1 are the width and height of the key VOP respectively, and M_2 and N_2 are the width and height of the candidate VOP, respectively.

D. Finite states representation of gestures

The proposed FSM for gesture recognition consists of a finite number of key frames and the corresponding key frame durations as shown in Fig. 6. The key frame duration is measured in terms of the number of video frames between the concerned key frame and the next key frame. The state transition occurs only when the shape similarity of key VOPs and duration criteria are met. A threshold T_s is predefined to allow a certain degree of shape variance in each state and a counter C_{dur} is used to judge the key frame duration criterion of the FSM.

During training, the input gesture sequence is transformed into a state sequence that is composed of key frames along with information about their durations. This gives the representation of the particular gesture in a summarized format. Thus, an FSM for gesture summarization is constructed. Algorithm 1 gives the pseudo-code for the state representation algorithm.

Algorithm 1 : Training of FSM for Gesture Recognition

begin
initialize $\mathbf{C_{max}}$, $\mathbf{C_{min}}$, $\mathbf{T_s}$.
$m \leftarrow 0$
do $m \leftarrow m + 1$
read the training data sequence d_m
generate VOPs $(VOP_1, VOP_2....VOP_{n_{max}})$
read n_{max}
assign VOP_1 as key VOP
$n \leftarrow 1$
do $n \leftarrow n + 1$
if shape of VOP_{n-1} and VOP_n are not similar
assign VOP_n as key VOP ($KVOP_n$)
count and store C_{dur_n} for $KVOP_n$
determine and update C_{min_n} and C_{max_n} for $KVOP_n$
until $n = n_{max}$
until $C_{min_n} \leq C_{dur_n} \leq C_{max_n}$ for $KVOP_n$
and shape convergence criterion met for $KVOP_n$
return **KVOP**, $\mathbf{C_{dur}}$
end

The counter value $\mathbf{C_{max}}$ and $\mathbf{C_{min}}$ assign the allowable ranges of duration for each KVOP. The values of $\mathbf{C_{max}}$, $\mathbf{C_{min}}$ and $\mathbf{T_s}$ for a particular gesture are obtained by running the algorithm for several times.

E. Recognition and co-articulation detection

As mentioned earlier, when different gestures are occurring sequentially co-articulation problem arises, as shown in Fig. 6(c). For recognition purpose, the incoming gesture states are matched with the states of all different FSMs obtained through training. Recognition is nothing but a string matching between the input gesture sequence sequence and the state sequence of an FSM. For an input sequence, the gesture recognizer decides whether to stay at the current state or to jump to the next state based on the shape similarity and duration criteria. If all the states of the FSM are passed successfully then a gesture is recognized. Else, co-articulation is detected.

F. FSM indexing for fast recognition

The searching of a suitable FSM in response to a particular gesture can be accelerated by indexing different FSMs by Angular Radial Transformation (ART) shape descriptor. As shown in the Fig. 7, the proposed indexing method searches only the first state of different FSMs, where ART shape descriptor coefficients preserve the shape information of the first state of all the FSMs in our gesture vocabulary. The shape similarity between first state of the incoming gesture sequence and the first state of all the FSMs are compared by ART shape descriptor. The best match indicates the appropriate FSM to be investigated during recognition. If during this matching process two or more FSMs are activated then all these FSMs are considered for investigation during recognition. The shape

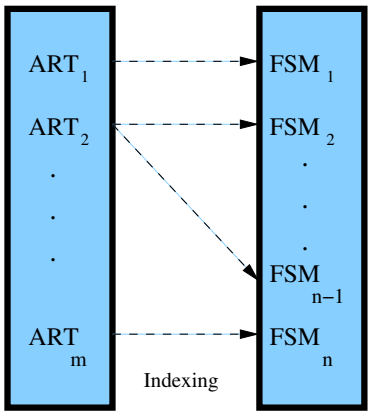

Fig. 7. Indexing of FSMs by ART shape descriptor, where $(n \geq m)$.

similarity measurement by ART shape descriptor is discussed in Appendix B.

G. Static hand shape recognition

Our FSM representation is extendable to static hand shape recognition, where no temporal information is required for modeling the FSMs. In this case, FSMs store only the shape information of gestures to be identified and consists of only one state occurring for indefinite time.

III. Experimental results

A. Experimental gesture sequences

In our experiment, we have selected ten different gesture sequences in view of realization of two practical scenarios *viz.*, remote robot control and hand gesture based menu activations in window based softwares.

1. Stop grasp OK sequence (**OK**): At starting time hand is closed and then gradually goes to complete open position and then goes again to closed stop position.

2. Clic sequence (**CLIC**): At start, one finger is up and then gradually hand goes to close position and then again goes to one indicative position.

3. No sequence (**NO**): At start, hand shows one by one finger in the up position with left sided inclination and gradually goes to the right inclination and finally to the left original inclination with hand making the same sign.

4. Rotate sequence (**ROTATE**): Here palm of the hand is rotated very slowly from left to right and then to left in the open handed position.

5. Close to Open (**CO**): At start hand is fully in the close position and gradually goes to complete open position, as shown in Fig. 8(a).

6. Open to Close (**OC**): It is just the opposite of Close to Open sequence, as depicted in Fig. 8(b).

7. Close to One (**CONE**): At start hand is in the complete close position and slowly it moves to one sign indicative position. Fig. 8(c) shows this sequence.

8. Close to Two (**CTWO**): At start, hand is in the complete close position and slowly it moves to two sign indicative position with two fingers up, as illustrated in Fig. 8(d).

9. Close to One and then Two (**COT**): At the starting time hand is in the complete close position and slowly it shows one indicative position and then gradually it moves to two sign indicative position and then wait in this position for a considerable period of time to indicate the end of the gesture.

10. Open to One (**OO**): The hand is completely open at start and slowly it turns into one sign indicative position.

B. Experimental conditions

1) Experimental Constraints: In our gesture recognition system, the user is constrained to the following four phases for making a gesture.

1) Insert the hand after some time within the capture range of the camera so that the background image can be read. We use 640×480 image for this purpose.
2) Keep hand still (fixed) in start position until gesture motion begins.
3) Move fingers and hand smoothly and slowly to different most prominent gesture positions.
4) Complete the gesture by keeping the fingers and hand in the final gesture position for relatively longer period of time.

2) Choice of background: During our experiment, we use comparatively simple background as seen in Fig. 8. This simple background enhances the tracking performance. Moreover, all the experiments are done under nearly constant uniform illumination condition.

3) Delay in start of tracking: During our recognition, we start tracking of hand movements after certain amount of time delay. This delay is required for stabilization of hand position in the first start position while making a gesture. So, in our experiment we read the hand positions after 60 initial frames.

C. Gesture summarization

Table 1 gives the different parameters used in our experiment for VOP generation. The results for KVOP selection for four sequences, viz.,"Rotate" sequence, "No" sequence, "Clic" sequence and "Stop Grasp OK" sequence, taken from Sebastien Marcel's gesture database are shown in Figs. 9(a)-10(b). First row of each figure shows the original video sequences, second row shows the difference edge images, third and fourth rows show horizontal and vertical candidates of a particular VOP respectively. The AND combination of vertical and horizontal candidates of each VOP is shown in the fifth row. After morphological filtering we get the binary alpha plane corresponding to each VOP, as shown in the sixth row. The last row of each figure shows the key binary alpha planes with all redundant frames discarded. From this we observe that redundant frames in a gesture video sequence bear only the temporal information of a gesture, whereas the most prominent and significant frames, *i.e.*, the key frames hold the actual information identifying a particular gesture. Figure 11 demonstrates this concept. Thus, we conclude that the key frames in a sequence are sufficient to represent a gesture in a summarized format.

Table 2 shows the number of key frames obtained in some of the test sequences. We observe that the number of key frames

International Science Index, Computer and Information Engineering Vol:6, No:8, 2012 waset.org/Publication/11224

International Science Index, Computer and Information Engineering Vol:6, No:8, 2012 waset.org/Publication/11224

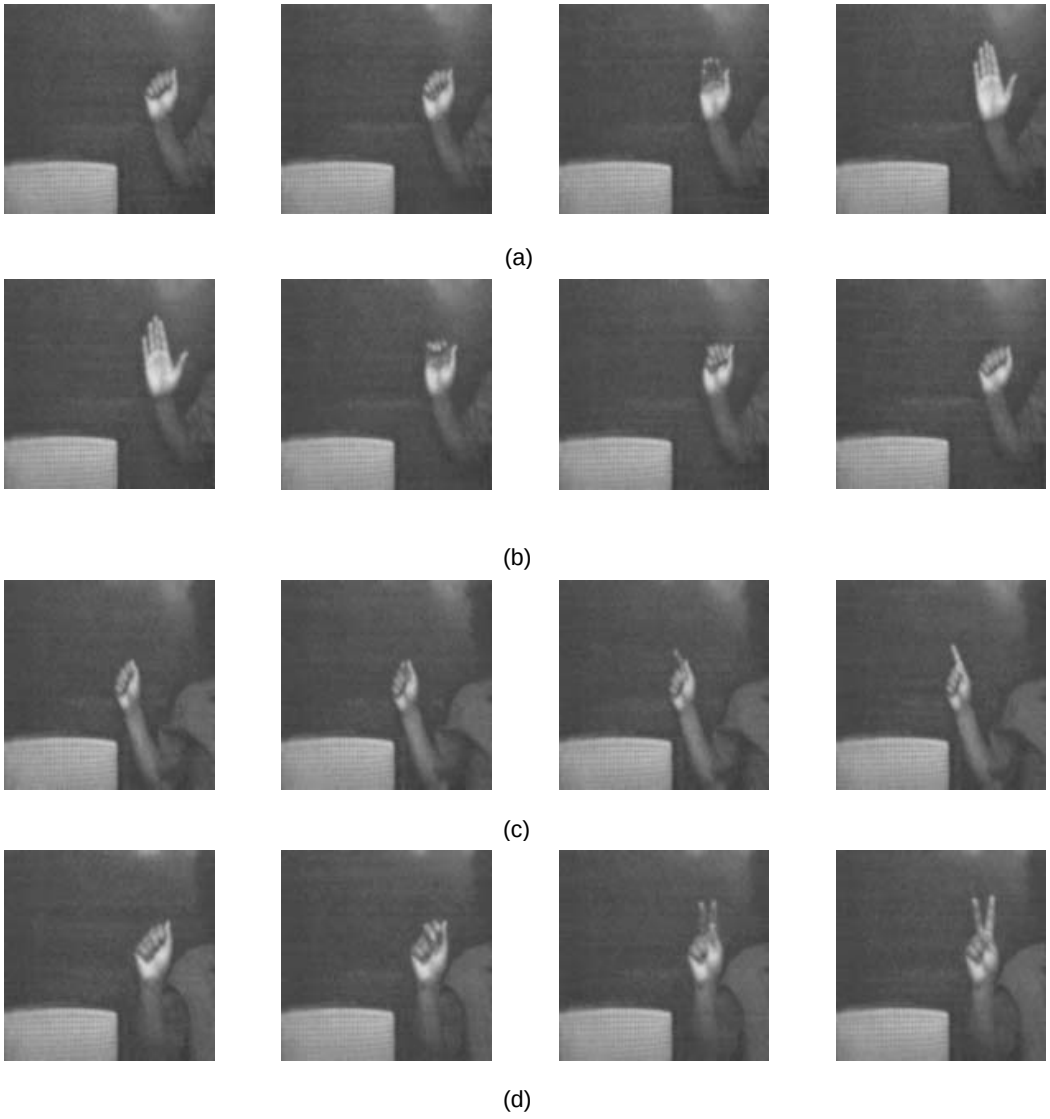

Fig. 8. Some of the experimental dynamic gesture sequences (a) Close to open (b) Open to close (c) Close to one (d) Close to two

obtained is considerably less than the total number of frames in the sequence. So, gesture summarization will indeed lessen the computational burden in the recognition process.

D. Dynamic gesture classification results

The classification results corresponding to different gesture sequences are shown in Table III. Here we label these gesture sequences as $1-10$ in the same order as mentioned in Section III-A. It is seen that gesture recognition and classification accuracy rate of our proposed method is comparable to other existing recognition methods although the computational cost is reduced significantly. However, we note that, our recognizer makes false judgement in classifying **CO** and **CT** sequences in some critical conditions. The ambiguity in recognition occurs if the hand makes very slow transition from one state to the next. The waiting state for one should not be long enough and transition from one to two should be very gradual and steady to avoid this ambiguity.

E. Static hand shape recognition

We use four hand shapes as shown in the Fig. 12 for static hand shape recognition. In all these cases, our recognizer gives an average accuracy of more than 97%, which is a very good recognition rate particularly for any Human Computer Intelligent Interaction (HCII) system.

IV. Conclusions

In our proposed method, MPEG-4 based video object extraction is used for finding different hand positions and shapes from the gesture video sequence. The VOP based method of segmentation requires no computational work for rotation and scaling of the object to be segmented out, where the shape change is represented explicitly by a sequence of two dimensional models, one corresponding to each image frame. Incorporation of median filtering in the model formation stage greatly enhances the tracking performance. Moreover, intro-

International Science Index, Computer and Information Engineering Vol:6, No:8, 2012 waset.org/Publication/11224

TABLE I
DIFFERENT PARAMETERS USED FOR VOP GENERATION

Stages	Parameter	Comment
Change detection	Threshold = 8	Depends on the visual characteristics of the gesture video sequence
Connected component labelling	Filter Threshold = 8	Edges shorter than 8 are eliminated
Edge detection (Canny detector)	SIGMA= 0.25 Low Threshold= 0.08 High Threshold= 0.2	Standard deviation of Gaussian Hysterisis thresholding
Median Filtering	3×3 neighbourhood	Removes salt and peeper noise
Background filtering	BGS= 1, BGE=10	Starting and Ending frame of background reference.
Hand tracking	RANK = 10000 X_SEARCH= 16 Y_SEARCH= 16	Generalized Hausdorff distance of this rank Object searching range
Model update	DIST_THRES= 5	Within this distance pixels of old model are updated.
VOP extraction	Filter Threshold = 2 AND= 1	Short undesired edges are removed Pixel by pixel logical AND operation.

TABLE II
TABLE SHOWING SUMMARIZATION OF SOME GESTURE SEQUENCES

Gesture sequences	Total no. of frames in actual sequence	Total no. of frames in summarized sequence	Comment
Open to close	270-320	7-9 *	* No. of key frames depends on the value of the threshold
Close to open	250-305	5-7	
Rotate	70-78	5-6	

(a) (b)

Fig. 9. Test results for (a) “Rotate” sequence, (b) “No” sequence

(a) (b)

Fig. 10. Test results for , (a) “Clic” sequence, (b) “Stop Grasp OK” sequence

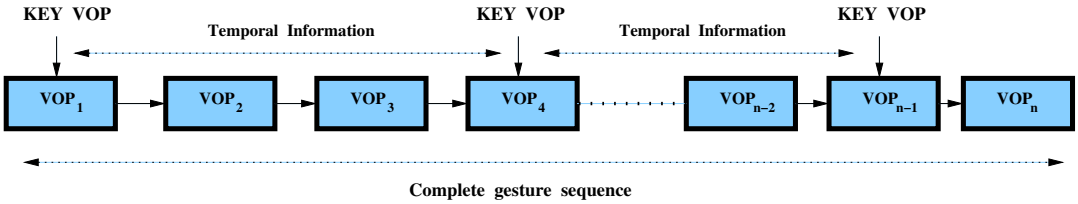

Fig. 11. Summarization of long gesture video sequence.

TABLE III
EXPERIMENTAL RESULTS: GESTURE RECOGNITION RATE
No. of training samples per class : 50 No. of test samples per class : 25

Actual class label	No. of test pattern assigned to predefined class											Acc rate	Err rate	Rej rate
	1	2	3	4	5	6	7	8	9	10	Reject			
1	23	0	0	0	1	1	0	0	0	0	0	92.0	8.0	0.0
2	0	20	0	0	0	0	3	1	0	1	0	80.0	20.0	0.0
3	0	0	21	0	0	0	2	0	0	2	0	84.0	12.0	4.0
4	0	0	0	22	0	1	0	0	0	1	1	88.0	8.0	4.0
5	2	0	0	0	23	0	0	0	0	0	0	92.0	0.0	8.0
6	1	0	0	0	0	24	0	0	0	0	0	96.0	4.0	0.0
7	0	1	3	0	0	0	20	0	0	1	0	80.0	20.0	0.0
8	0	1	0	0	0	0	0	20	3	0	1	80.0	16.0	4.0
9	0	0	0	0	0	0	3	2	19	0	1	76.0	12.0	12.0
10	0	0	3	0	0	0	1	0	0	20	1	80.0	16.0	4.0
Average												**84.8**	**11.6**	**3.6**

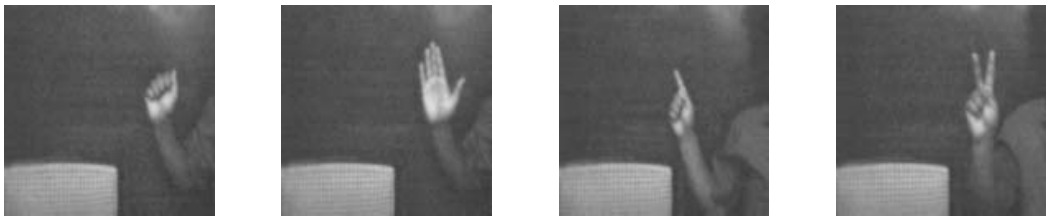

Fig. 12. Some of the experimental static hand shapes (a) Close (b) Open (c) One (d) Two

duction of logical AND operation in the final VOP generation step preserves the finger information of a hand.

From the test results it is concluded that by using key frames, a particular gesture can be uniquely determined and can be represented in terms of a finite state machine with key frames and corresponding frame duration as states. One notable advantage of finite state representation of gesture is that it handles different gestures consisting of different number of states. The key frame based gesture representation is nothing but the summarization of the gesture with finite number of unique states. The advantage of key frame based state representation is that only the shape similarity measurement for key frames are required instead of all frames of the video sequence. Moreover, key frame based gesture classification can solve the co-articulation problem to some extent. The key frame based gesture representation is also useful for both gesture recognition and coding of video frames in compressed domain.

APPENDIX

We have used distance transform algorithm [33] for computation of Hausdorff distances in different steps of our proposed algorithm. Computation of distance transformation (DT) is accomplished using Chamfer 5-7-11 mask shown in Fig 13. The DT is initialized by assigning zero to edge pixels. In two iterations, the distances are calculated by centering the mask at each pixel in turn and updating the distance of this pixel.

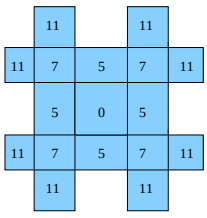

Fig. 13. Mask Chamfer 5-7-11

The region-based shape descriptor expresses pixel distribution within a 2-D object region; it can describe complex objects consisting of multiple disconnected regions as well as simple objects with or without holes. Consequently, an ART based descriptor was recently adopted by MPEG-7 [34]. Conceptually, the descriptor works by decomposing the shape into a number of orthogonal 2-D basis functions

International Science Index, Computer and Information Engineering Vol:6, No:8, 2012 waset.org/Publication/11224

(complex-valued), defined by the Angular Radial Transform (ART) [27], [35]. The normalized and quantized magnitudes of the ART coefficients are used to describe the shape.

From each shape, a set of ART coefficients F_{nm} is extracted, using the following formula:

$$F_{nm} = < V_{nm}(\rho,\theta), f(\rho,\theta) >$$

i.e.,

$$F_{nm} = \int_0^{2\pi} \int_0^1 V^*_{nm}(\rho,\theta), f(\rho,\theta)\rho d\rho d\theta \quad (3)$$

where $f(\rho,\theta)$ is an image function in polar coordinates and V_{nm} is the ART basis function that are separable along the angular and radial directions, that is,

$$V_{nm}(\rho,\theta) = \frac{1}{2\pi}\exp(jm\theta)R_n(\rho) \quad (4)$$

$$R_n(\rho) = \begin{cases} 1 & \text{if } n = 0 \\ 2\cos(\pi n\rho) & \text{if } n \neq 0 \end{cases} \quad (5)$$

Descriptor representation

The ART descriptor is defined as a set of normalized magnitudes of complex ART coefficients. Twelve angular and three radial functions are used $(n < 3, m < 12)$. For scale normalization, ART coefficients are divided by the magnitude of ART coefficient of order $n = 0, m = 0$. Therefore, discarding the normalized ART co-efficient of order $n = 0, m = 0$, which is unity, we have 35 coefficients in all. To keep the descriptor size to a minimum, quantization is applied to each coefficient using four bits per coefficient. Hence, the default region-based shape descriptor has total 140 bits.

Shape similarity measurement by ART shape descriptor

The distance (or dissimilarity) between two shapes described by the ART descriptor is calculated using an $L-1$ norm, for example, by summing up the absolute differences between ART coefficients of equivalent order $(L = 2)$.

$$Dissimilarity = \sum_i \parallel M_d[i] - M_q[i] \parallel \quad (6)$$

Here, the subscript d and q represent image in the database and query image, respectively and M is the array of ART descriptor values.

REFERENCES

[1] R.C. Rose, Discriminant Wordspotting Techniques for Rejection Non-Vocabulary Utterances in Unconstrained Speech, Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, San Francisco, vol.II (1992) 105-108.
[2] K. Takahashi, S. Seki, R. Oka, Spotting Recognition of Human Gestures from Motion Images, Technical Report IE92-134, The Insttitute of Electronics, Information, and Communication Engineers, Japan, (1992) (in Japanese) 9-16.
[3] T. Baudel, M. Beaudouin-Lafon, CHARADE: Remote Control of Objects Using Free-Hand Gestures, Communication ACM, 36 (7) (1993) 28-35.
[4] S.S. Fels, G.E. Hinton, Glove-Talk: A Neural Network Interface between a Data-Glove and a Speech Synthesizer, IEEE Transaction on Neural Networks, 4(1) (1993) 2-8.
[5] S.S. Fels, G.E. Hinton, Glove-Talk II: A Neural Network Interface which Maps Gestures to Parallel Format Speech Synthesizer Controls, IEEE Transaction on Neural Networks, 9(1) (1997) 205-212.
[6] D.L. Quam, Gesture Recognition With a Data Glove, Proceedings of the IEEE Conference on National Aerospace and Electronics, vol. 2 (1990).
[7] D.J. Sturman, D. Zeltzer, A Survey of Glove-Based Input, IEEE Computer Graphics and Applications, vol. 14 (1994) 30-39.
[8] W.W. Kong, S. Ranganath, 3-D Hand Trajectory Recognition for Signing Exact English, Proceedings of the 6th IEEE International Conference on Automatic Face and Gesture Recognition, (2004) 535-540.
[9] Ng.Y.Y. Kevin, S. Ranganath, D. Ghosh, Trajectory Modeling in Gesture Recognition using Cybergloves and Magnetic trackers, Proceedings of the IEEE TENCON 2004, Chiang Mai, Thailand, vol.A (2004) A.571-A.574.
[10] J. Davis, M. Shah, Recognizing Hand Gestures, Proceedings of the European Conference on Computer Vision, ECCV, (1994) 331-340.
[11] V.I. Pavlovic, R. Sharma, T.S. Huang, Visual Interpretation of Hand Gestures for Human-Computer Interaction: A Review, IEEE Transaction on Pattern Analysis and Machine Intelligence, 19 (7) (1997) 677-695.
[12] Y. Wu, T.S. Huang, Self-supervised Learning for Visual Tracking and Recognition of Human Hand, Proceedings of 17th National Conference on Artificial Intelligence, (AAAI'2000), (2000) 243-248.
[13] J. Zieren, N. Unger, S. Akyol, Hands Tracking from Frontal View for Vision-Based Gesture Recognition, Proceedings of DAGM Symposium, (2002) 531-539.
[14] Y. Wu, T.S. Huang, Hand Modelling, Analysis, and Recognition for Vision-Based Human Computer Interaction, IEEE Signal Processing Magazine, (2001) 51-60.
[15] J. Yamato, J. Ohya, K. Ishii, Recognizing Human Action in Time-Sequential Images Using Hidden Markov Model, Proceeding of the IEEE Conference on Computer Vision and Pattern recognition, (1992) 379-385.
[16] C. Vogler, D. Metaxas, ASL Recognition Based on a Coupling Between HMM and 3D Motion Analysis, Proceedings of the International Conference on Computer Vision, (1998) 363-369.
[17] A. Ramamoorthy, N. Vaswani, S. Chaudhury, S. Banerjee, Recognition of Dynamic Hand Gestures, Pattern Recognition, 36 (2003) 2069-2081.
[18] M. Isard, A. Blake, Contour Tracking by Stochastic Propagation of Conditional density, Proceedings of the European Conference on Computer Vision, (1996) 343-356.
[19] M. Black, A Jepson, Recognition Temporal Trajectories using Condensation Algorithm, Proceedings of the International Conference on Automatic Face and Gesture Recognition, Japan, (1998) 16-21.
[20] J. Davis, M. Shah, Visual Gesture Recognition, Vision, Image and Signal Processing, 141 (2) (1994) 101-106.
[21] A.F. Bobick, A.D. Wilson, A State-Based Approach to the Representation and Recognition of Gesture, IEEE Transaction on Pattern Analysis and Machine Intelligence, 19 (12) (1997) 1325-1337.
[22] P. Hong, M. Turk, T.S. Huang, Gesture Modelling and Recognition using Finite State Machines, Proceeding of the IEEE Conference on Face and Gesture Recognition, (2000) 410-415.
[23] T. Meier, K.N. Nagan, Automatic Segmentation of Moving Objects for Video Object Plane Generation, IEEE Transaction on Circuits and Systems for Video Technology, 8 (5) (1998) 525-538.
[24] D.P. Huttenlocher, G.A. Klanderman, W.J. Rucklidge, Comparing Images using the Hausdorff Distance, IEEE Transaction on Pattern Analysis and Machine Intelligence, 15 (9) (1993) 850-863.
[25] M.K. Bhuyan, D. Ghosh, P.K. Bora, Finite State Representation of Hand Gestures using Key Video Object Plane, Proceedings of the IEEE TENCON 2004, Chiang Mai, Thailand, vol.A (2004) A.579-A.582.
[26] M.K. Bhuyan, D. Ghosh, P.K. Bora, Key Video Object Plane Selection by MPEG-7 Visual Shape Descriptor for Summarization and Recognition of Hand Gestures, Proceedings of the 4th Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP 2004), Kolkata, India (2004) 638-643.
[27] B.S. Manjunath, P. Salembier, T. Sikora, ed., Intoduction to MPEG-7, Multimedia Content Description Interface, John Wiley and Sons, Ltd, (2002).
[28] Y. Cui, J.J. Weng, Hand Segmentation using Learning-Based Prediction and Verification for Hand Sign Recognition, Proceedings of the IEEE CS Conference on Computer Vision and Pattern Recognition, (1997) 88-93.
[29] R. Polana, R. Nelson, Low Level Recognition of Human Motion, Proceeding of the IEEE CS Workshop on Motion of Non-Rigid and Articulated Objects, Austin, (1994) 77-82.
[30] A.F. Bobick, J. Davis, Real-Time Recognition of Activity using Temporal Templates, Proceedings of the IEEE CS Workshop on Applications of Computer Vision, (1996) 39-42.
[31] C. Kim, J.N. Hwang, Object-Based Video Abstraction for Video Surveillance Systems, IEEE Transaction on Circuits and Systems for Video Technology, 12 (12) (2002) 1128-1138.

International Science Index, Computer and Information Engineering Vol:6, No:8, 2012 waset.org/Publication/11224

[32] B. Erol, F. Kossentini, Automatic Key Video Object Plane Selection using the Shape Information in the MPEG-4 Compressed Domain, IEEE Transaction on Multimedia, 2 (2) (2000) 129-138.
[33] G. Borgefors, Distance Transformations in Digital Images, Computer Vision, Graphics and Image Processing, 34 (1986) 344-371.
[34] B. Erol, F. Kossentini, Local Motion Descriptors, Proceedings of the IEEE 4th Workshop on Multimedia Signal Processing, (2001) 467-472.
[35] M. Bober, MPEG-7 Visual Shape Descriptors, IEEE Transaction on Circuits and Systems for Video Technology, 11 (6) (2001) 716-719.

International Science Index, Computer and Information Engineering Vol:6, No:8, 2012 waset.org/Publication/11224