Continuous Hand Gesture Segmentation and Co-articulation Detection

M.K. Bhuyan[1], D. Ghosh[2], and P.K. Bora[1]

[1] Department of Electronics and Communication Engineering,
Indian Institute of Technology, Guwahati, India
manas_kb@iitg.ernet.in, prabin@iitg.ernet.in
[2] Faculty of Engineering and Technology,
Multimedia University, Melaka Campus, Malaysia
dghosh_iitg@hotmail.com

Abstract. Gesture segmentation is an extremely difficult task due to both the multitude of possible gesture variations in spatio-temporal space and the co-articulation of successive gestures. In this paper, a robust framework for this problem is proposed which has been used to segment out component gestures from a continuous stream of gestures using finite state machine and motion features in a vision based platform.

1 Introduction

One very interesting field of research in Pattern Recognition that has gained much attention in recent times is Gesture Recognition. Hand gesture recognition from visual images finds applications in areas like human computer interaction, machine vision, virtual reality and so on. Many vision-based gesture recognition systems assume that the input gestures are isolated or segmented. This assumption makes the recognition task easier, but at the same time it limits the naturalness of the interaction between the user and the system, and therefore negatively affects the user's experience. In more natural settings, the gestures of interest are embedded in a continuous stream of motion, and their occurrence has to be detected as part of recognition. This is precisely the goal of gesture spotting *i.e.*, to locate the start point and end point of a gesture pattern, and to classify the gesture as belonging to one of predetermined gesture classes. Another important issue of gesture recognition is co-articulation, which makes the extraction and segmentation of gesture commands even harder in continuous hand movements. Co-articulation is a phenomenon in which one gesture influences the next in a temporal sequence [1]. This happens due to hand movement during transition from one gesture to the next. The problem is very significant in case of fluent sign language. Recognition of co-articulated gestures is one of the difficult tasks in gesture recognition.

Zhao *et al.* calculates velocity, and treats local minima in the velocity as gesture boundaries [2]. However, this method does not produce gestures that are consistent with human perception. The indirect approach uses a state space

P. Kalra and S. Peleg (Eds.): ICVGIP 2006, LNCS 4338, pp. 564–575, 2006.

model of gestures [3]. Lee and Kim proposed a gesture spotting system using a threshold model that calculates the threshold likelihood of a given input sequence as the basis of approving or rejecting the input pattern as a gesture. For gesture spotting, this system detects the end point of a gesture and finds a corresponding start point by searching the Viterbi path from the end point [4]. However, the method has a problem in that the system cannot report the detection of a gesture immediately after the system reaches its end point. Moreover they used heuristic information, such as moving the hand out of the camera range. Such a heuristic is not very natural to humans. Nishimura and Oka also proposed a gesture spotting method using continuous dynamic programming (CDP), which worked on a frame-by-frame basis and matched a sequence of input feature vectors and standard patterns corresponding to each gesture [5]. The matching result is the cumulative distance between cumulative frames and a gesture, and the best match is the result of spotting and recognition. However, this approach is limited to sets of gestures that do not contain any intermediate poses that resemble the start or end poses. In practical applications, this approach of using state-space models for achieving gesture segmentation and recognition severely limits the number of gestures that can be segmented. Vogler and Metaxas used context-dependent HMMs for recognition of continuous gestures [6]. However the context-dependent modelling has some inherent problems. First, it is linguistically implausible, because it fails to model movement epenthesis properly. Second, by using signs as the basic phonetic unit, the number of states used in the HMM recognition network grows roughly with order of $O(W^2)$, where W is the number of signs in the vocabulary, as the number of possible contexts itself grows with order $O(W^2)$.

Though the "segmentation" as well as "co-articulation detection" are the the most important and open research issues for continuous hand gesture recognition, not much vision based approaches are reported till date of this research work. The techniques developed so far for co-articulation detection are not always successful for wide range of gesture vocabulary. Moreover, these algorithms do not address the problems associated with the recognition of continuous hand gestures of different spatio-temporal behavior *viz.,* gestures having only local motions, gestures having only global motions, gestures having both local and global motions, and also fluent finger spelling. Motivating by these facts, we propose a more general and relatively simple model for continuous gesture segmentation by combining good features of state based and motion based approaches. In our method, we first segment the input video stream by detecting gesture boundaries at which the hand pauses for a while during gesticulation. Next, every segment is checked for co-articulation via finite state machine (FSM) matching or by using hand motion information. Thus, co-articulation phases are detected and eliminated from the sequence and we are left with a set of isolated gestures. We proposed to use FSM for segmentation of gestures having only local hand motions, where we used only some selected frames for building up a gesture model. For gestures having both local and global motions, we first determine the co-articulated strokes and subsequently used it for determining gesture boundary.

2 Proposed Scheme for Gesture Spotting and Co-articulation Detection

2.1 Gesture Boundary Detection

Generally a gesture starts and ends with the hand staying in a standstill position for a while. That is, a signer generally starts making a sign from a "pause" state and ends in a "pause" state in case of continuous gesturing. Based on this idea, we propose to use the hand motion information for locating the boundary points of each individual gesture in a continuous stream of gestures. A boundary point is detected whenever the hand pauses during gesturing.

The first step in the proposed method for gesture spotting involves generating the video object plane (VOP) from each input frame in the continuous stream of video. A series of key VOPs are extracted from the generated VOPs, which also gives the duration of each key VOP in terms of the number of frames between each pair of key VOPs [7]. A VOP model diagram showing a portion of a continuous gesture sequence is given in Fig. 1. The diagram shows two gestures in the sequence connected with a co-articulation phase in between them. In the figure, $\mathrm{KVOP}_{m,n}$ represents the n^{th} key VOP in gesture number m and $\mathrm{T}_{m,n}$ is the corresponding time duration, expressed in terms of the number of video frames between $\mathrm{KVOP}_{m,n}$ and $\mathrm{KVOP}_{m,n+1}$

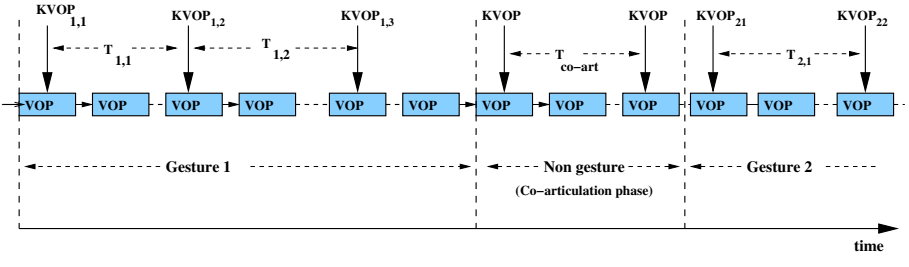

Fig. 1. VOP model for a portion of a continuous gesture sequence showing co-articulation

A key VOP in the sequence corresponds to a particular hand shape and/or position and the key VOP duration indicates the time for which the hand remains more or less fixed in that pose and position. This key VOP information is used in the proposed gesture spotting algorithm as well as for gesture recognition. For gesture spotting, a "pause" in the sequence is detected whenever the duration of a key VOP exceeds or at least equal to the minimum time for which the hand pauses at the starting or ending of a single gesture, as determined during the training session.

In the proposed algorithm, we assume that the camera starts capturing the hand image sometime before the signer starts gesturing. Therefore, the first

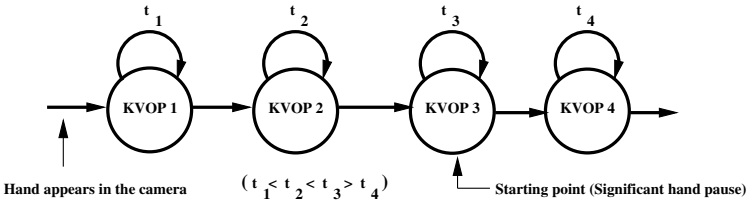

Fig. 2. Gesture sequence starting point detection

"pause" in the video indicates the starting of the first gesture in the sequence, as illustrated in Fig. 2. The gesture may end when the next "pause" in the video stream is spotted. Assuming that there is always some non-gestural movement in between two gestures, the third and fourth "pauses" in the input video will indicate the starting and ending of the second gesture in the sequence respectively and so on. Thus, all the gestures in the input sequence can be conveniently selected out. However, the scheme will fail under certain situations, as listed below, and will result in incorrect spotting of gestures.

1. The assumption that there is always some non-gestural movement of hand in between two gestures may not be always true. For example, if the end position or pose of a gesture is same as the start position or pose of the next gesture then there is generally no extra movement in between these two gestures. In that case, the two gestures are adjoined to each other in the sequence with a common "pause" indicating end of the first gesture and start of the next gesture.
2. In the case of fluent finger spelling, there is generally no motion during the gesturing period while the hand may move in between two gesture poses due to co-articulation. That means, here a "pause" itself in the sequence corresponds to a gesture sign as if the start-point and end-point of the gesture have merged together.
3. In the case of gestures involving global hand motion, there may be some "pauses" within a single gesture. When the hand traverses in space, it makes one or more hand strokes to build up a complete gesture trajectory. Since a hand stroke generally starts from a "pause" and ends in a "pause", a multi-stroke gesture will contain some extra pauses in between. Some examples of ideal gesture trajectories that are made up of one or more gesture strokes are shown in Fig. 3. The first two examples are gestures having global motion only representing "One" and "Square", respectively. As we see in the figure, "One" is a single stroke gesture while "Square" is made up of four strokes. Therefore, the gesture "One" will start from a pause and will end in the next "pause". But, there will be three intermediate "pauses" in the "Square" indicating gesture. The last example is the trajectory of a gesture composed of both local and global motions that represents a sentence in sign language. Here each word in the sentence is signed by a single stroke associated with

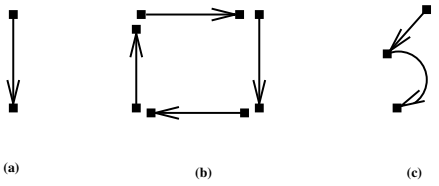

Fig. 3. (a) Single hand stroke for a gesture (b) Multiple hand strokes for a gesture (c) Multiple strokes for a sentence of sign language

changing the hand pose. Hence, the number of strokes in the gesture is equal to the number of words in the sentence. Accordingly, for this example, there will be one extra pause in between.

Hence, it is not always possible to have reliable spotting of gestures only by detecting pauses in an input video. In view of this, we propose to check the nature of hand movement in every video segment in between two "pauses" in the input stream. Assuming that there is no unintentional movement other than co-articulation in between gestures, the proposed gesture spotting method requires to determine the occurrence of co-articulation in the sequence. This is done through FSM matching and/or using motion features, as explained in the next section.

2.2 Co-articulation Detection in Continuous Gestures with Local Motion Only

Following two steps are used for co-articulation detection for gestures having only local hand motions.

Step 1: We assume that after completing a gesture, the signer holds his hand for sometime in the last signed pose in the gesture and then quickly moves it to the starting pose of the next gesture in the sequence. The signer does this by bending his fingers and/or moving his palm in a very short span of time while holding the hand more or less fixed at one position in space. That means, the co-articulation phase is also made up of local hand motions only and hence may be represented by an FSM model. Now since there is fast change in hand shape during co-articulation compared to that during a gesturing phase, the time duration associated with each state of an FSM representing co-articulation will be generally very small than that for an FSM representing a gesture. Therefore, co-articulation in continuous gesturing may be detected if the key VOP durations in between two "pauses" are below a certain threshold. The value of this threshold may be decided during the training session.

Step 2: The co-articulation detected in Step 1 can be verified by representing the KVOPs in the detected co-articulation phase by an FSM. The input video sequence is represented by an FSM and is matched to all the prototype FSMs

contained in the database, each prototype FSM representing a local motion gesture in our gesture vocabulary. If no match is obtained then co-articulation is detected.

2.3 Co-articulation Detection in Continuous Gestures Having Global Motion

In the case of gestures having global motion only or gestures having both global and local motions, the gesturing hand traverses in space to form a gesture trajectory. After a gesture trajectory is complete, the hand pauses for a while and then moves with very high velocity to the starting position of the next trajectory. After that, the hand again pauses for a while before starting the next trajectory. Based on this, we now propose to detect co-articulation by observing the motion of the hand between two "pauses" in the input hand motion video.

A gesture phase can be divided into three motion stages – preparation, stroke and retraction, in addition to the starting and ending "pauses" [8]. On the other hand, a co-articulation phase starts from a pause, makes a fast hand stroke and finally ends up in another pause. Therefore, it is possible to distinguish a co-articulation from a gesture stroke if we can determine the behavior of the gesturing hand in between two "pauses". For this, we compute two motion parameters, *viz.*, velocity and acceleration, at every key VOP instant and decide the nature of hand movement at that instant. The proposed scheme for co-articulation detection in the continuous gestures with global hand motions consists of the following two stages.

Step 1: Co-articulation detection by motion features

Acceleration feature for co-articulation detection:

The most important motion parameter that can discriminate a co-articulation phase from a gesture is the change in speed or acceleration. During co-articulation the speed of the hand increases to a very high value from almost zero value and then abruptly comes down to almost zero as illustrated in Fig. 4. That means, the hand moves with very high acceleration (positive or negative) during the co-articulation phase. On the other hand, during gesturing the speed of the hand gradually increases from a pause, may remain constant for sometime and then gradually comes down to almost zero. Therefore, acceleration feature may be a good measure to check for co-articulation.

Velocity feature for co-articulation detection:

The speed of the hand is generally very high while making a stroke. But, that during the preparation and retraction stages is generally very small. That means, the average velocity of the hand during co-articulation is generally very large compared to that during a gesturing phase. Hence, the measure of velocity may

serve as an additional feature to detect co-articulation. We now describe how these two motion parameters are used to discriminate between a co-articulation phase and a gesturing phase.

Fuzzy method for co-articulation detection:

We have formulated a scheme to measure the motion behavior in terms of some fuzzy sets and rules to determine whether a particular motion is gesture phase or co-articulation. By observing different gesture samples, we first define four fuzzy sets to denote the different ranges of speed. They are 'Zero' (ZO), 'Positive Small' (PS), 'Positive Medium' (PM) and 'Positive Large' (PL). The corresponding fuzzy membership functions are plotted in Fig. 5. Similarly, we define five fuzzy sets to quantify change in speed in terms of some fuzzy measures. They are 'Negative Medium' (NM), 'Negative Small' (NS), 'Zero' (ZO), 'Positive Small' (PS) and 'Positive Medium' (PM); the corresponding membership functions are shown in Fig. 6. Mathematically these two motion features can be expressed in terms of motion vector (MV_i) as follows.

$$\text{Speed}: \qquad S_i = \sqrt{(x_i - x_{i+1})^2 + (y_i - y_{i+1})^2} = MV_i \tag{1}$$

$$\text{Change in speed}: \qquad \Delta S_i = S_i - S_{i-1} = MV_i - MV_{i-1} \tag{2}$$

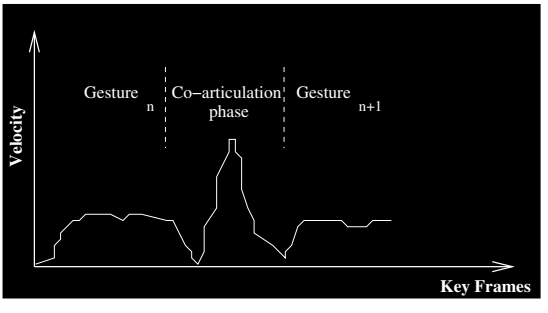

Fig. 4. Example of a typical velocity plot for connected sequentially global motion gestures in a continuous video stream

As given in Algorithm 1, motion vector MV_i for i^{th} video object plane in the gesture sequence is determined from generalized Hausdorff distance measure. Hausdorff tracker can be used to track non–rigid objects in a complex scene [9], [10]. In our algorithm, Hausdorff distance is computed using distance transform algorithm [11]. As explained earlier, a typical gesture can be divided into 5 motion phases. There are three distinct phases: preparation, stroke and end. The stroke is distinguished from the others by the speed and the change of speed. Table 1 shows these 5 motion phases. These two motion parameters are subsequently fuzzified and the motion stage through which the hand is undergoing is determined using some fuzzy rules, as stated below.

Algorithm 1. Estimation of Motion vector

Given $(i-1)^{th}$ VOP O and the i^{th} VOP I and a set of translated vectors $\mathbf{T}$
begin
  for $\boldsymbol{t} = (t_x, t_y) \in \mathbf{T}$
    Calculate distance transform of edge images O and I.
    Calculate $h_{p,t}(O, I)$.
    Calculate $h_{j,t}(I, O)$.
    Determine $H_t(O, I) = max\{h_p(O, I), h_j(I, O)\}$.
  end.
Find $min\{H_t(O, I)\}$ over $\boldsymbol{t} \in \mathbf{T}$.
    Find translation vector $t^{'} = (t_x^{'}, t_y^{'})$ corresponding to $min\{H(O, I)\}$.
    $MV_i = (t_x^{'}, t_y^{'})$.
return MV_i
end

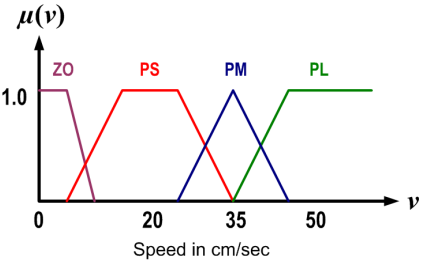

Fig. 5. Fuzzy membership functions defining different ranges of speed

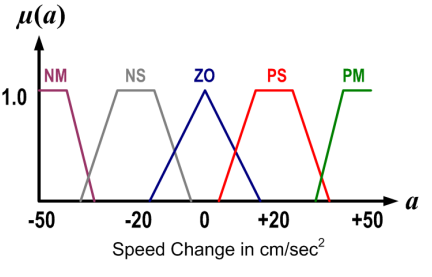

Fig. 6. Fuzzy membership functions defining different ranges of change in speed

From a large set of real gestures we observe that the speed of the hand generally lies within a certain range at every stage in a gesture or co-articulation. So, is the case for change in speed. For example, during the "Preparation" stage the speed is generally 'Positive Small' and the change in speed is either 'Positive Small' or 'Negative Small'. This we can write in the form of a fuzzy rule as

– "IF the speed is Positive Small AND the change in speed is Positive Small OR Negative Small, THEN the hand is in Preparation stage".

Accordingly, if the hand is moving at a speed v and change in speed is a then, using this fuzzy rule and applying min-max method, the degree of certainty by which we can say that the hand is in "Preparation" stage is given as

$$\mu_{\text{Prep}}(v, a) = \min\left[\mu_{\text{PS}}(v), \max\left[\mu_{\text{PS}}(a), \mu_{\text{NS}}(a)\right]\right] \tag{3}$$

The IF-THEN fuzzy rules for all the other three motion phases can be derived in a similar manner and using Table 1 that gives the fuzzy sets for the two motion parameters corresponding to the different stages of hand motion. Using all these fuzzy rules we can compute the degrees of confidence that the hand is doing "Pause", "Preparation", "Stroke" and "Retraction" at a given instant of time. Finally, we decide for the motion stage that has the maximum degree of confidence. In doing so, we are able to find the sequence of hand motion in an input stream of hand gesture video and a motion pattern. If the extracted motion pattern follows the motion phases of Table 1 in order, the video segment is classified as a gesture, otherwise it is labelled as co-articulation.

Table 1. Gesture motion phases and corresponding motion parameters

Motion phase	Speed	Change in Speed
Pause	ZO	ZO, PS, NS
Preparation	PS	PS, NS
Stroke	PL, PM	ZO, PS, NS, PM, NM
Retraction	PS	NS

Step 2: Verification of Co-articulation Using Trajectory Shape

The process of co-articulation detection can be made more reliable and accurate by considering the hand trajectory shape for verification. It is observed that during co-articulation the hand generally moves in a straight path. So, a motion phase is not a co-articulation if the trajectory is not a straight line. Note that a straight-line trajectory does not always indicate co-articulation. It may also represent a valid gesture stroke. The Step I discriminates between co-articulation and such a gesture stroke.

However, the above scheme fails to spot gestures if there is no co-articulation phase between two gestures. In such cases, gesture spotting is done along with recognition. In these types of gestures, every video segment between two "pauses" corresponds to either a gesture stroke in space or a co-articulation phase. Therefore, as a first step towards recognition, it is required to check whether an input segment is a co-articulation or a gesture stroke. If it is a co-articulation then it is discarded and we move on to the next segment. If it is a stroke then we check

whether it is a complete gesture trajectory or not by comparing it to all the prototype trajectories contained in the gesture vocabulary database. If not, we then move on to the next segment and check for gesture stroke. This we continue till a valid gesture trajectory is obtained by connecting all these individual hand strokes in a sequential manner.

3 Experimental Results

In a first set of experiments, we considered some sequences of continuous gesturing with local hand motions only. We have used five different gesture sequences taken from Sebastien Marcel's gesture database and Thomas Moeslund's gesture recognition database. The continuous gesture sequences were formed by performing different gesture signs in different orders in such a way that while some gestures were connected to each other in the sequence there were cases where two gestures were separated by a co-articulation phase in between. In our experiments, we achieved an overall recognition accuracy of 90.4%. This high recognition rate confirms that our proposed algorithm for gesture spotting and co-articulation detection was indeed effective in segmenting out meaningful individual gestures in the input sequences accurately and precisely.

In a second set of experiment, we considered some trajectory patterns indicating "One", "Two", "Five", "Seven" and "Three", as shown in the first row of Fig. 7. These gesture signs were performed one after another in different orders to build up different sequences of continuous hand motion gestures with global motion only. The second row of Fig. 7 shows another set of gesture trajectories, *viz.*, "Square", "Circle", "Diamond", "triangle" and "W". These gestures were used in our third set of experiments. As we observe, each gesture in the first row starts from some point at top of the frame and ends somewhere at the bottom. So, in our second set of experiments we always have a co-articulation phase in between two gestures in a sequence. On the other hand, in our third set of experiments the starting and ending of all gestures in a sequence are in the vicinity of each other. That means, here the gestures are generally connected to each other without any co-articulation in between. Fig. 8(a) shows a sample of the continuous gesture sequence without any co-articulation while Fig. 8(b) shows a sequence of gestures connected by co-articulation strokes in between.

In our second set of experiments, we used hand motion information to discriminate the co-articulation phases in the input gesture sequences. Subsequently, all individual gesture patterns in the sequences were segmented out and were identified with an overall accuracy of 90%. This demonstrates the efficiency of our proposed method for co-articulation detection and subsequent trajectory guided recognition. We also observed that the acceleration of the hand was significantly high during co-articulation compared to that during the gesturing phase. Finally, in our third set of experiments, individual gesture patterns in the sequences were segmented out and were identified with an overall accuracy of 94%. This shows that the proposed system is capable of identifying gesture strokes and

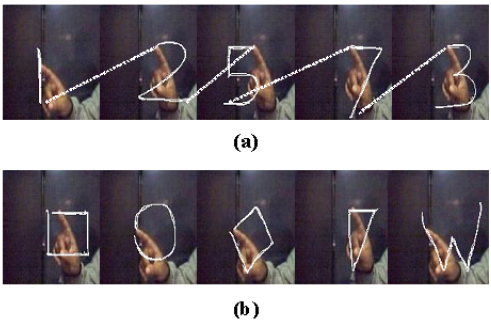

Fig. 7. Trajectories of the gestures used in our experiments: (a) Gestures that produce co-articulation, used in our 2nd set of experiments, (b) Gestures that do not produce co-articulation, used in our 3rd set of experiments

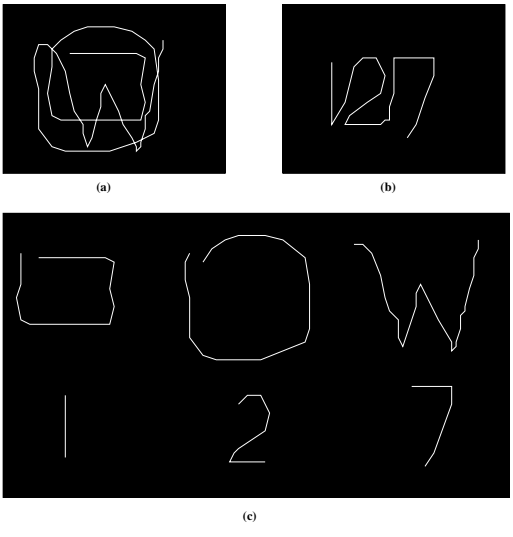

Fig. 8. (a) Continuous motion trajectory without co-articulated stroke (b) Continuous motion trajectory with co-articulated stroke (c) Segmented out trajectory

trajectory patterns with very high accuracy. The segmented out gesture trajectories from the gesture sequence samples in Fig. 8(a) and Fig. 8(b) are shown in Fig. 8(c).

4 Conclusion

Co-articulation is one of the main challenges in continuous gesture recognition. Motion interpretation is a quite ill-posed problem, in which cognitive science and psychological studies need to be combined. That is why, not many vision based

approaches for estimating co-articulation have been reported in the literature till date. Most of the proposed algorithms till now have success only for some specific gesture vocabularies, which can not be generalized for all kinds of gestures performed in different contexts. The proposed system for co-articulation detection in a continuous stream of gestures performs well for different types of gesture sequences having different spatio-temporal and motion behaviour in a common vision-based platform. One notable advantage of the proposed method is that finger motion during gesticulation is considered as the shape change of the video object, which can be efficiently quantified using FSM based representation.

References

1. Shamaie, A., Hai, W., Sutherland, A.: Hand gesture recognition for HCI, ERCIM News, http://www.ercim.org/publication/Ercim_News, **46** (2001)
2. Zhao, L.: Synthesis and acquisition of laban movement analysis qualitative parameters for communicative gestures, Ph.D Thesis, CIS, University of Pennsylvania, (2001)
3. Aggarwal, J., Cai, Q.: Human motion analysis: A review, Proc. Nonrigid and Articulated Motion Workshop, (1997) 90–102
4. Lee, H. K., Kim, J. H.: An HMM based threshold model approach for gesture recognition, IEEE Trans. Pattern Analysis and Machine Intelligence, **21(10)** (1999) 961–973
5. Nishimura, T., Oka, R.: Towards the integration of spontaneous speech and gesture based on spotting method, Proc. IEEE/SICE/RSJ International Conf. Multisensor Fusion Integration Intelligent System, (1996) 433–437
6. Vogler, C., Mextaxas, D.: Adapting hidden Markov models for ASL recognition by using three-dimensional computer vision methods, Proc. IEEE International Conf. on Systems, Man and Cybernetics, (1997) 156–161
7. Bhuyan, M.K., Ghosh, D., Bora, P.K.: Key video object plane selection by MPEG-7 visual shape descriptor for summarization and recognition of hand gestures, Proc. 4th Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP), (2004) 638–643
8. Kendon, A.: Conducting interaction, Cambridge University Press, (1990)
9. Huttenlocher, D.P., Noh, J.J., Rucklidge, W.J.: Tracking non-rigid objects in complex scene, Proc. 4th International Conf. Computer Vision, (1993) 93–101
10. Bhuyan, M.K., Ghosh, D., Bora, P.K.: Estimation of 2D motion trajectories from video object planes and its application in hand gesture recognition, Lecture Notes in Computer Science, Springer-Verlag, **(LNCS 3776)** 509–514
11. Borgefors, G.: Distance transformations in digital images, Computer Vision, Graphics and Image Processing, **34** (1986) 344–371