SEGMENTATION AND RECOGNITION OF CONTINUOUS GESTURES

Hong Li[1] *Michael Greenspan*[1,2]

[1] Dept. of Electrical & Computer Engineering, Queen's University, Kingston, Canada
[2]School of Computing, Queen's University, Kingston, Canada
email: 1hl5@qlink.queensu.ca, michael.greenspan@queensu.ca

ABSTRACT

A novel method is introduced to segment and recognize time-varying human gestures from continuous video streams. Motion is represented by a 3D spatio-temporal surface based upon the evolution of a contour over time. The warping paths between the input signal and a set of Gesture Models are obtained using Continuous Dynamic Programming and the boundary of a gesture is located by analyzing all possible gesture candidates during a specific period of time. Correlation and Mutual Information are employed to select the best candidate when more than one gesture is recognized at the same time period. The system has been implemented and tested on continuous gesture sequences containing 8 different gestures performed by 4 subjects. The results demonstrate that the proposed method is very effective, achieving a recognition rate of 95.9%.

Index Terms— Continuous Gesture Recognition, Dynamic Time Warping, Continuous Dynamic Programming, Motion Signature, Gesture Model.

1. INTRODUCTION

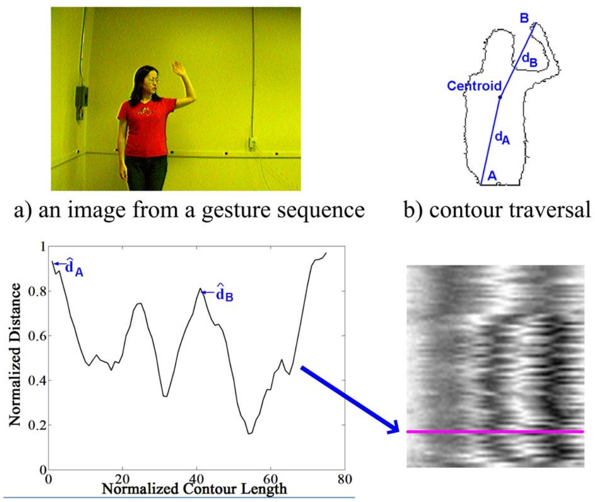
Automatic gesture recognition from continuous video streams has a variety of potential applications varying from smart surveillance, human-machine interaction to biometrics [1, 2]. In most naturally occurring scenarios, gestures are linked together in a continuous varying stream, without any obvious pause or break between individual gestures. The recognition of such gestures is therefore closely related to the segmentation of such a stream into individual gestures, i.e. determining the start and end times of individual gestures, and segmentation and recognition in this way become aspects of the same problem.

The task of gesture segmentation is to extract meaningful patterns from a stream of input signals. Due to the large inter- and intra- human gesture variations appearing both in the spatial and temporal domains, segmenting and recognizing continuous gestures is considered to be challenging. Dynamic Time Warping (DTW) and Hidden Markov Models (HMM) have been widely used in continuous gesture recognition, following the path of continuous speech recognition [3, 4, 5, 6]. Since DTW can't locate gesture boundary in a continuous stream, Darrell et al. [3] performed an exhaustive search at each time instance using DTW. Liang et al. [4] used the time-varying parameter to detect the endpoints in continuous Taiwan Sign Language, and HMM is employed for recognition. Lee et al. [5] constructed a threshold model based on HMM to segment gestures in sequences. Starner et al. [6] employed HMM to recognize structured American Sign Language sentences without explicit segmentation at the word level. However, the use of HMM recently has been criticized for its large training set requirement and complicated classification models [7, 8].

In this paper, we propose a novel approach to segment and recognize gestures in continuous streams based on a framework combining Continuous Dynamic Programming (CDP) [9] and two matching schemes, i.e. Correlation and Mutual Information (MI) [10, 11]. Both CDP and DTW are based upon dynamic programming. In CDP, the main idea is to allow each time instance to be a start time during the warping, which makes it attractive for dealing with the continuous stream segmentation problem. CDP has been previously employed for gesture recognition [12, 13]. In conventional CDP-based approaches, however, the large variation across gestures makes it difficult to find appropriate thresholds for each class on which decisions could be made [12, 13]. In contrast, our approach only employs a single global threshold to rule out obvious incorrect warping paths and the real gesture is derived by analyzing all candidates paths obtained. Correlation and MI, with a proven effectiveness to deal with multi-scale gestures [10, 11], are further employed to choose the best gesture candidate whenever more than one gesture is found in one specific period of time.

2. MOTION SIGNATURE AND GESTURE MODELS

We chose to represent motion by a 3D spatio-temporal surface, a *Motion Signature*, based on the contour evolution over time. The subject is first segmented from the background based on statistics of the color distribution [14], and the border following method proposed by Suzuki et al. [15] is then used to extract the contour of the subject. To parameterize the contour, we adopt the 1D distance-to-centroid shape de-



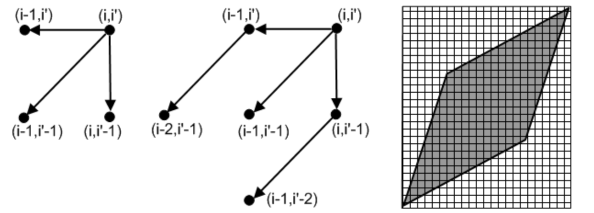
a) an image from a gesture sequence b) contour traversal

c) shape descriptor d) Motion Signature

Fig. 1. Motion Signature: a contour-based motion representation.

scriptor which clockwise unwraps the contour with respect to the centroid, similar to [16, 17]. A Motion Signature is then obtained by stacking consecutive 1D distance signals along the time axis. Fig. 1 b) and c) show the contour traversal process and the normalized 1D signal, respectively. The 3D surface of a Motion Signature can be visualized as a 2D image (Fig. 1 d)), where each horizontal line indicates one distance signal and the intensity corresponds to the distance value. To account for variations of the motion, a Gesture Model consisting of a set of *mean images* and *variance images* is then constructed using Motion Signatures at multiple time scales [10, 11].

Segmenting individual gestures from a continuous stream is difficult since we assume that gestures can be carried at any time in an arbitrary order, and that the duration of each gesture is also unknown due to speed variations. In our previous work [11], we have applied DTW to Compound Gesture Models (CGMs), which are composed of the concatenation of two Gesture Models, to approximately estimate the endpoints of a gesture in a stream. The gesture is then recognized by finding the best match using Correlation or Mutual Information over the estimated endpoints. A limitation of that algorithm is that, since the number of the CGMs is K^2 for a database of K gestures, the computational cost to warp the input to all the CGMs is $O(K^2MN)$, given that the time complexity of DTW is $O(MN)$ (M and N, the lengths of the two warped signals). It can be prohibited for large database in realtime applications. In the following, we'll show that the computational cost can be reduced to $O(KMN)$ when CDP is employed to segment the gestures.



a) Type I b) Type II c) Parallelogram Constraint

Fig. 2. Local and global constraints.

3. GESTURE SEGMENTATION AND RECOGNITION

In this section, we'll first compare DTW to CDP. Then a new gesture segmentation and recognition method based on CDP and two other matching algorithms, i.e. Correlation and Mutual Information, will be presented.

3.1. CDP vs. DTW

DTW makes use of dynamic programming to align two similar signals, and CDP can be viewed as an extension to DTW. A global optimum path is found by recursively accumulating the locally optimal paths. Given a test pattern Z_t and reference pattern $Z_{t'}$, the best time warp will minimize the accumulated distance along the path through the grid from $(0, 0)$ to (t, t'). When the local range of the path in the vicinity of the point (i, i') is restricted to its immediate three neighbors, i.e., local constraint type I shown in Fig. 2 a), DTW can be formulated as:

$$R_{i,i'} = r_{i,i'} + \min(R_{i,i'-1} + R_{i-1,i'} + R_{i-1,i'-1}) \quad (1)$$

where

$$R_{1,1} = r_{1,1}$$
$$R_{i,1} = r_{i,1} + R_{i-1,1}$$
$$R_{1,i'} = r_{1,i'} + R_{1,i'-1}$$

In Eq.(1), $R_{i,i'}$ is the partial sum cost, and $r_{i,i'}$ measures the distance of gesture signals at two temporal instances. The test pattern is generally classified as the reference pattern where the accumulated distance is minimum.

The limitation that the endpoints of the two signals have to be aligned prevents DTW from being applied directly in continuous gesture recognition since we don't know *a priori* the location of the endpoints of each gesture in the input sequence. In early work, some techniques have been proposed to overcome this problem for the application of Connected Speech Recognition (CSR), i.e. early-decision, level building, etc., and a thorough review of these techniques can be found in [18].

CDP is a variation of DTW in the sense that the global optimum path of CDP is found the same way as DTW, i.e. recursively accumulating the locally optimal paths. If we use the same local constraint as in Eq.(1), then by revising the

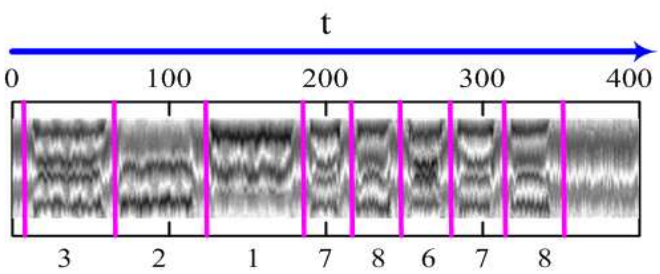


Fig. 3. A continuous gesture sequence containing 8 gestures. From left to right: 3 ("wave two hands"), 2 ("wave left hand"), 1 ("wave right hand"),7 ("left down right up"), 8 ("left up right down"), 6 ("raise two hands"), 7, 8.

initialization condition $R_{i,1} = r_{i,1} + R_{i-1,1}$ to $R_{i,1} = r_{i,1}$, we allow each frame of the input sequence to be a start time candidate. The accumulated distance of $R_{i,t'}$ represents the warping distance between the reference model $Z_{1:t'}$ and the partial input $Z_{i_0:i}$, where i_0 can be obtained by tracing back the warping path. When the value of $R_{i,t'}$ reaches a minimum at time i, the endpoint of a potential gesture has been segmented.

CDP based on Eq.(1) doesn't have any constraint on its warping path. Ratanamahatana et al. [19] point out that wider warping windows do not always give higher recognition accuracy, as generally believed. In fact, the accuracy usually peaks at certain warping window sizes, and then drops as the window size increases. In our experiment we also found that the warping window size does affect recognition rate. However, because the speed variance of gestures in our database is relatively large, the choice of the best global warping window size is difficult. We therefore adopt a moderate warping window to give a certain constraint. If the local constraint type II (as shown in Fig. 2 b)) is applied, a global parallelogram constraint is enforced automatically, which is shown in the shadowed area in Fig. 2 c).

3.2. Algorithm implementation

Based on CDP, a gesture can be segmented as long as the accumulated distance value for one pattern falls below a predefined threshold during the warping process. Usually, a threshold for each pattern has to be defined in advance based on training data. However, due to the large spatial and temporal variations among gestures, it's very difficult to find appropriate thresholds based on limited training data. Moreover, since threshold values vary among different gesture classes, a decision will be hard to make when more than one gesture have accumulated distance values below their own thresholds at roughly the same time. Therefore, applying CDP directly for continuous gesture recognition usually won't give satisfactory results.

Total #	*Subs*	*Dels*	*Ins*	*Rec. Rate*
640	5	8	13	95.9%

Table 1. Results for continuous gesture recognition.

In our approach, the endpoint detection will not depend on thresholds learned from each individual class. We use only one uniform threshold Γ to rule out those warping paths with sufficiently large accumulated distances. Correct gestures are determined first by analyzing the start and end times of all possible gesture candidates. When there is only one gesture occurring within a specific time, a decision that a gesture is segmented can be made easily. However when multiple gestures are found within the same time period, then a voting scheme incorporating correlation and MI, is used to select the most probable gesture class. The details are described as follows: The input sequence with multiple gestures is first warped to all Gesture Models at their maximum scales based on CDP, respectively. Each warping produces a list of accumulated distance values α_k, where k is the gesture class number and the element of α_k is $R_{i,t'}$ as defined in Eq.(1). A small fixed size window is then slid through each list α_k. Within the window, The smallest distance value is found and kept only if it is smaller than the threshold Γ. The time when the smallest value is observed, the gesture class, the start time (can be traced back through the warping path), together with the distance value are put into the candidate list β. In list β, those candidates sharing the same gesture class and start time, but with larger distance values will be removed, and many false local minima are ruled out.

For two adjacent candidates C_a and C_b, if the start time of C_b is later than the end time of its preceding gesture C_a, and their start times are not the same, then C_a is believed to be the true gesture. Alternately, if the start time of C_b is earlier than the end time of C_a, which indicates that the two candidates overlap, we cannot decide the correct gesture since both of the warping distances reach their local minima. Instead, we make use of a voting scheme to make the decision. That is, both C_a and C_b are matched to the Gesture Models at their corresponding scales using both Correlation and MI similar to [10, 11]. The correct gesture belongs to the one getting at least two votes among the three.

Since the input signal is only warped to the K individual Gesture Models, the time complexity of this algorithm is $O(KMN)$. Fig. 3 shows a continuous gesture sequence of 400 frames automatically segmented by the algorithm described above. The gestures are performed in a random order with speed variations, e.g. gesture No.7 ("left down right up") was slower when it was repeated the second time. Our approach successfully segmented and recognized all 8 gestures.

4. EXPERIMENTS

To characterize the performance of the method, we executed a set of experiments on a number of subjects. All data were captured with a Point Grey Firefly camera using an image resolution of 320×240, at a frame rate of 15 fps. There were a total of 8 Gesture Models in the database, each of which was learned using 30 instances performed by three subjects. To evaluate our algorithm for continuous gesture recognition, we collected 80 video clips performed by three subjects, with two of them the same as the training subjects. For each video clip, the subject was asked to arbitrarily perform any 8 gestures continuously in random order, without any obvious pauses between gestures. In total, there were 640 gestures for testing.

We manually segmented and annotated all 80 video clips as the ground truth. To evaluate the results, we used the criterion used in continuous speech recognition where the recognition rate is based on three error types: *Substitution*, where an incorrect gesture was substituted for the correct one; *Deletion*, where a correct gesture was omitted in the recognized sequence; and *Insertion*, an extra gesture was added in the recognized sequence. The recognition rate was then calculated as [6]:

$$Rec.\ Rate = 100\% \times (1 - \frac{Subs + Dels + Ins}{No.\ of\ correct\ gestures}) \quad (2)$$

Table 1 shows that our algorithm reaches an average recognition rate of 95.9% on the test data. The result demonstrates that the proposed approach was very effective at recognizing gestures in continuous sequences. Currently, the algorithm was not optimized for good time performance. But we expect that with further optimization, the algorithm could be executed in real-time.

5. CONCLUSION

We have presented a novel method for the segmentation and recognition of time-varying continuous human gestures, and experimental results have demonstrated the method to be effective. Using a framework that combines Continuous Dynamic Programming and two other matching methods, i.e. Correlation, and Mutual Information, the temporal endpoints of a gesture were segmented and a gesture was recognized. The proposed method is both computationally efficient and robust: in experiments containing 80 continuous gesture sequences, the resulting average recognition rate was 95.9%.

In future work, we plan to apply this approach to other type of gestures, e.g. sign language. We are also implementing a real-time gesture recognition system based on the described method.

6. REFERENCES

[1] J. K. Aggarwal and Q. Cai, "Human motion analysis: A review," *CVIU*, vol. 73, no. 3, pp. 428–440, 1999.

[2] T. B. Moeslund and E. Granum, "A survey of computer vision-based human motion capture," *CVIU*, vol. 81, pp. 231–268, 2001.

[3] T. Darrell and A. Pentland, "Space-time gestures," in *Proc. IEEE Conf. Comp. Vis. Pat. Rec.*, 1993, pp. 335–340.

[4] Rung-Huei Liang and Ming Ouhyoung, "A real-time continuous gesture recognition system for sign language," in *IEEE Intl. Conf. Automatic Face and Gesture Rec.*, 1998, pp. 558–567.

[5] Hyeon-Kyu Lee and Jin H. Kim, "An hmm-based threshold model approach for gesture recognition," *IEEE Trans. PAMI*, vol. 21, no. 10, pp. 961–973, 1999.

[6] T. Starner, J. Weaver, and A. Pentland, "Real-time american sign language recognition using desk and wearable computer based video," *IEEE Trans. PAMI*, vol. 20, no. 12, pp. 1371–1375, 1998.

[7] R. Bowden, D. Windridge, T. Kadir, A. Zisserman, and M. Brady, "A linguistic feature vector for the visual interpretation of sign language," in *European Conference on Computer Vision*, 2004, vol. 1, pp. 391–401.

[8] Shu-Fai Wong and R. Cipolla, "Continuous gesture recognition using a sparse bayesian classifier," in *IEEE Intl. Conf. Pat. Rec.*, 2006, vol. 1, pp. 1084–1087.

[9] R. Oka, "Spotting method for classification of real world data," *The Computer Journal*, vol. 41, no. 8, pp. 559–565, 1998.

[10] H. Li and M. Greenspan, "Multi-scale gesture recognition from time-varying contours," in *ICCV*, 2005, vol. I, pp. 236–243.

[11] H. Li and M. Greenspan, "Continuous time-varying gesture segmentation by dynamic time warping of compound gesture models," in *International Workshop on Human Activity Recognition and Modelling (HARAM)*, 2005, pp. 35–42.

[12] H. Wu, R. Kido, and T. Shioyama, "Improvement of continuous dynamic programming for human gesture recognition," in *IEEE Intl. Conf. Pat. Rec.*, 2000, vol. 2, pp. 945–948.

[13] J. Alon, V. Athitsos, and S. Sclaroff, "Accurate and efficient gesture spotting via pruning and subgesture reasoning," in *ICCV 2005 HCI Workshop*, 2005, pp. 189–198.

[14] C. R. Wren, A. Azarbayejani, T. Darrell, and A. P. Pentland, "Pfinder: Real-time tracking of the human body," *IEEE Trans. PAMI*, vol. 19, no. 7, pp. 780–785, 1997.

[15] S. Suzuki and K. Abe, "Topological structural analysis of digitized binary images by border following," *Computer Vision, Graphics, and Image Processing*, vol. 30, pp. 32–46, 1985.

[16] P. Peixoto, J. Goncalves, and H. Araujo, "Real-time gesture recognition system based on contour signatures," in *IEEE Intl. Conf. Pat. Rec.*, 2002, vol. 1, pp. 447–450.

[17] L. Wang, T. Tan, H. Ning, and W. Hu, "Silhouette analysis-based gait recognition for human identification," *IEEE Trans. PAMI*, vol. 35, no. 12, pp. 1505–1518, 2003.

[18] H.F. Silverman and D.P. Morgan, "The application of dynamic programming to connected speech recognition," *IEEE ASSP Magazine*, vol. 7, no. 3, pp. 6–25, 1990.

[19] C. A. Ratanamahatana and E. Keogh, "Everything you know about dynamic time warping is wrong," in *3rd Workshop on Mining Temporal and Sequential Data*, 2004.