The Eighteenth International Symposium on Artificial Life and Robotics 2013 (AROB 18th '13),
Daejeon Convention Center, Daejeon, Korea, January 30-February 1, 2013

Hand gesture recognition using subunit-based dynamic time warping

Yanrung Wang[1], Atsushi Shimada[1], Takayoshi Yamashita[2], and Rin-ichiro Taniguchi[1]

[1] Kyushu University, Japan
[2] OMRON Corporation, Japan

kenyou@ limu.ait.kyushu-u.ac.jp

Abstract: A subunit-based Dynamic Time Warping (DTW) approach is introduced for hand movement recognition. Two major contributions distinguish the proposed approach from conventional DTW. (1) A set of hand movement subunits is constructed using a data-driven method. The learning is based on subunits instead of the whole hand movement for more efficient learning. (2) A more accurate similarity measure is offered using subunit-to-subunit matching to absorb the difference between two similar sub-sequences belonging to the same subunit, and only keeping the distances between sub-sequences that relate to different subunits. Compared with the conventional DTW approach, the proposed approach is experimentally demonstrated to be both accurate and efficient for locally collected datasets.

Keywords: hand gesture, gesture recognition, subunit

1 INTRODUCTION

Vision-based hand gesture recognition has attracted considerable attention because of its new and fascinating applications such as interactive human-machine interfaces, sign language interpretation, and virtual environments [1]. Features such as appearance, shape, and orientation often play an important role in hand gesture recognition. In this paper, we consider hand gestures as movement trajectories and focus on recognition of the movement trajectories.

Dynamic time warping (DTW) [2] is widely used to recognize movement trajectories, because it simultaneously aligns time-variable data and computes a likelihood of similarity. Generally speaking, there are two major limitations to the use of DTW in hand movement recognition. (1) DTW matching uses information about individual training examples that it is sensitive to variations in training data. Hence, it is difficult to support efficient personalized gesture recognition. (2) DTW is sensitive to noise and unable to distinguish movement trajectories that have similar subsequences, as it requires continuity along the warping path. The use of DTW consequently requires the development of many prototypes to achieve proper performance, leading to an expensive computational load.

To address these issues, we develop an effective recognition approach that combines the use of the DTW distance metric and subunits, widely investigated in the field of sign language [3][4]. Subunits are elementary units in a language and there are far fewer subunits than words in the vocabulary of the language, which is expected to lead to smaller data size in training and a smaller search space in recognition.

2 OVERVIEW OF THE PROPOSED APPROACH

Our system handles color image sequences in real time to recognize numbers from 0 to 9 by the hand movement trajectories. In the training phase, all training data are mapped to sequences of digits between 0 and 7 according to their orientation feature and then segmented into the set of basic motion units according to changes in orientation. Next, subunits are selected via clustering and set as the yielded cluster centers. In this case, each training sequence is mapped to a sequence of subunits. In the testing phase, the test sequence is also represented as a sequence of subunits and then classified according to DP matching between the test sequence and training sequences. Specifically, DTW distance is measured by subunit-to-subunit matching to improve recognition accuracy and online learning is used to adapt the training set to the user's individual habits.

3 HAND MOVEMENT REPRESENTATION

Hand movement trajectories are obtained by detecting the top most point of the hand skin region as the fingertip. To represent these trajectories, we use the orientation feature, which has been shown to provide high accuracy in hand movement recognition in previous work [5]. A hand movement is a spatio-temporal trajectory that consists of fingertip positions $(x_t, y_t), \mathrm{t} = 1,2,\dots,T-1$, where T indicates the length of trajectory. Similar to [5], we calculate the orientation feature according to the positions of fingertips between consecutive frames as follows.

$$\theta_t = arctan\left(\frac{y_{t+1} - y_t}{x_{t+1} - x_t}\right); t = 1,2,\dots,T-1 \quad (1)$$

The orientation θ_t is quantized into a set of codewords from 0 to 7 by dividing it by 45◦. Therefore, a hand movement can be represented by a sequence of digits according to the yielded codewords as shown in Fig. 1 and the similarity between two movements can be measured by the DTW distance [6]. The Examples of the DTW distance between movements are illustrated in Fig. 2. The examples show that the measurement procedure of adding the current distance to the smallest one can result in similar trajectories being treated as dissimilar, which leads to inaccuracies in hand movement recognition.

4 HAND MOVEMENT SUBUNIT CONSTRUCTION

Motivated by [3] and [4], we consider a hand movement as a sequence of basic motion units, referred to as the common pattern of hand movements, and carry out a self-organization process to select a representative set of motion units from the training set as subunits. All training data are segmented into a set of sub-movements. Clustering is performed for these sub-movements (basic motion units) to find the common pattern of hand movements.

4.1 Motion unit segmentation

Motion unit segmentation can be thought of as a boundary detection problem. The use of trajectory discontinuity and motion speed discontinuity has been shown to be effective in detecting the subunit boundary in sign language recognition [7]. We employ changes in orientation as trajectory discontinuity metrics to detect unit boundaries when the current orientation is very different from that in a neighboring frame or that in the starting frame of the motion unit.

4.2 Subunit clustering

To select a set of representative subunits from all submovements of the training set, we perform k-medoids clustering using the DTW distance metric. The k-medoids algorithm, a variant of k-means clustering, computes medoids instead of centroids as cluster centers to minimize the sum of intra-class distances. To determine the number of clusters, we employ an iterative clustering that selects all sub-movements of the training set as the initial cluster centers and iteratively merges similar clusters until convergence to obtain the "optimal" number of clusters. Furthermore, we use a kmeans-like algorithm for k-medoids clustering [8] to overcome the drawback that partitioning around medoids (PAM, K-medoids) works inefficiently for large data sets because of the complexity.

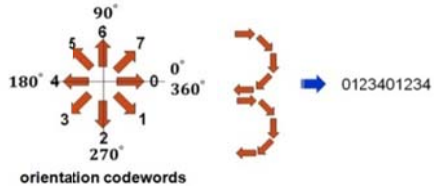
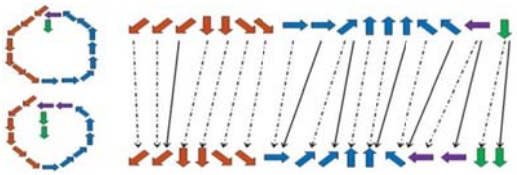
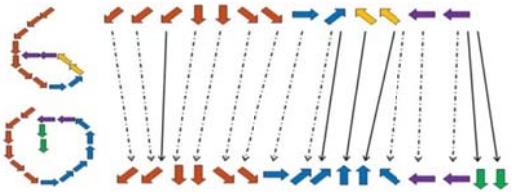


Fig. 1. Orientation codewords and an example of movement representation using orientation codewords



(a) distance between '0' and '0' is 7



(b) distance between '0' and '6' is 6

Fig. 2. Examples of the DTW distance between movements

The yielded cluster centers are then used as subunits to map the training data to subunit sequences. We build a subunit-to-subunit distance matrix during construction of subunits and use it as a look-up table to speed up the recognition procedure.

5 SUBUNIT-BASED RECOGNITION

We propose two-step submovement-to-subunit and subunit-to-subunit matching in the recognition process to improve the performance of recognition and employ subunit-based learning to overcome sensitivity to variations in the training data.

5.1 Subunit-based learning

Instead of training entire hand movements composed of orientation codewords, we train each movement as a concatenation of subunits. The advantages are as follows. (1) The amount of training materials needed is reduced as all training data are composed of a limited set of subunits. (2) A simplified enlargement of training data is achieved by composing new training data using the existing subunits.

5.2. Submovement-to-subunit matching

Let P_x be a testing sequence and $S = \{s_1, s_2, \ldots, s_{|S|}\}$ be a set of $|S|$ subunits constructed from training data. Similar to training sequences, the test sequence P_x is also mapped to a sequence of digits according to changes in

orientation and then segmented into m submovements u_{xi}. We calculate DTW distances between submovement u_{xi} and all subunits to find the nearest subunit s_{xi} and then use these subunits to recompose the testing sequence P_x. The yielded testing sequence $P_x = \{s_{x1}, \dots, s_{xi}, \dots, s_{xm}\}$ is used to perform subunit-to-subunit matching with training data.

5.3. Submovement-to-subunit matching

Hand movement trajectories are recognized through dynamic subunit sequence matching. Let $P_y = \{s_{y1}, \dots, s_{yj}, \dots, s_{yn}\}$ be a training sequence consisting of n subunits. The distance $DTW(P_x, P_y) = D(s_{xm}, s_{yn})$ is calculated as follows.

$$D(s_{xm}, s_{yn}) = \min \begin{cases} D(s_{xi-1}, s_{yj}) + cost \\ D(s_{xi}, s_{yj-1}) + cost \\ D(s_{xi-1}, s_{yj-1}) + cost \end{cases} \quad (2)$$

$$cost = \min \begin{cases} 0 & if\ s_{xi} = s_{yj} \\ dist(s_{xi}, s_{yj}) & if\ s_{xi} \neq s_{yj} \end{cases} \quad (3)$$

Here, $dist(s_{xi}, s_{yj})$ is obtained using the look-up table generated during the construction of subunits.

6 EXPERIMENTS

To test the proposed approach for hand movement recognition and to compare with conventional DTW, we perform evaluations in terms of the recognition rate and average computational time for a locally collected hand movement corpus. Here, the average computation time is the average time taken to calculate the distance.

The constructed corpus contains 10 different classes of hand movement trajectories from 0 to 9, performed by seven subjects in our laboratory environment. Each of the 10 classes of trajectories is repeated 25 times by each subject. To evaluate the performance for datasets of different size, we randomly select 9, 15, and 30 training samples from each class, performed by three subjects, to construct the training set. The other data corresponding to the other four subjects are used as a test set. To obtain results that are more reliable, the construction of subunits and evaluation of recognition performance were repeated five times using different datasets constructed relating to different subjects.

6.1 Evaluation of the recognition rate

Recognition rates classified according to three different sizes of training set are compared in Fig. 3. Compared with conventional DTW, the proposed approach showed a significant improvement when there were only nine training data. The findings indicate that the proposed approach is able to overcome the sensitivity to training data of conventional DTW to offer high recognition accuracy even when there are few training data. The two main reasons for the improvement are as follows.

6.1.1 Increase in the variety of training data

To train each movement as a concatenation of subunits increases the variety of training data such that it is possible to recognize new training patterns not seen in training.

For instance, we might have a training set of three training data $P_y = \{u_{y1}, u_{y2}, u_{y3}\}$, $P_z = \{u_{z1}, u_{z2}\}$, and $P_w = \{u_{w1}, u_{w2}\}$, where u_{yj}, u_{zk}, and u_{wl} are segmented submovements and are clustered into three subunits $s_1 = \{u_{y1}\}$, $s_2 = \{u_{y2}, u_{z1}, u_{w1}\}$ and $s_3 = \{u_{z2}, u_{w2}\}$. According to the yielded subunit set, training data are mapped to sequences of subunits $P_y = \{s_1, s_2, s_3\}$, $P_z = \{s_2, s_3\}$ and $P_w = \{s_2, s_3\}$.

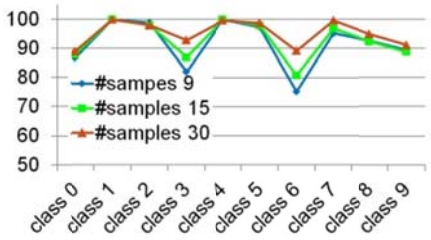
In the example, we only train two training prototypes s_2s_3 and $s_1s_2s_3$ for three training data because P_z and P_w are mapped to the same prototype s_2s_3. The reduction of training prototypes improves learning efficiency while maintaining the variety of training data to avoid loss of recognition accuracy. In addition, training patterns that can be represented by the training prototype s_2s_3 are not only P_z and P_w but also $u_{w1}u_{z2}$, $u_{w1}u_{y3}$, and so on. That is, the variety of P_z and P_w is increased to $|s_2||s_3|$ training patterns because of the use of existing subunits that include motion units from the other training data. It is thus also possible to recognize new patterns, even though they are not seen in the training. These merits achieve an improvement of the recognition rate without requiring high computational complexity.

6.1.2 A more accurate similarity measure

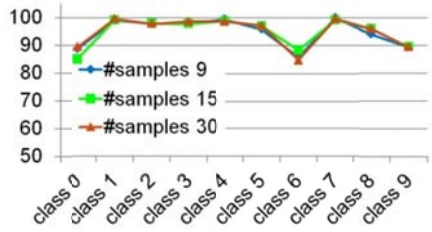
The conventional DTW distance metric is sensitive to noise and unable to find movement trajectories that have similar sub-sequences. Therefore, similar trajectories may be treated as dissimilar, leading to inaccurate recognition. As illustrated in Fig. 4, the proposed approach offers a more accurate similarity measure because it absorbs the difference between two similar sub-sequences belonging to the same subunit and only keeps the distances between sub-sequences that relate to different subunits.

6.2 Evaluation of average computation time

The average computation time and the number of training prototypes when using subunit-based learning are given in Fig. 5. The results indicate that a significant improvement in computational complexity was obtained.



(a) average recognition rate using conventional DTW



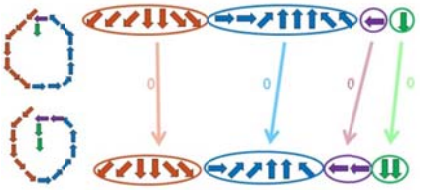
(b) average recognition rate using subunit-based DTW

Fig. 3. Comparison of the recognition rate

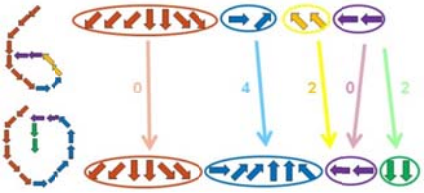
The reduction of the number of training prototypes, due to the fact that multiple training data were mapped to single training prototype, was one of the causes of the improvement in computational cost. The major reason for the improvement is that the distance between subunits was rapidly obtained using the lookup table in the procedure of subunit-to-subunit matching. These findings shown in Fig. 3 and Fig. 5 support the claim that the proposed approach improves recognition accuracy while not increasing the computational load.

7 CONCLUSION

This paper proposes a subunit-based approach to hand movement recognition. In contrast to conventional DTW approaches, we share subunits across hand movements to obtain a smaller training data size and search space to improve recognition performance. In addition, a more robust similarity measure, using subunit-to-subunit matching, is offered. The experimental results demonstrate that the proposed approach is both accurate and efficient for hand movement recognition. Our future research will focus on incremental learning for the subunit itself to support efficient personalized recognition.



(a) distance between '0' and '0' is 0



(a) distance between '0' and '0' is 8

Fig. 4. Examples of the subunit-based DTW distance between movements

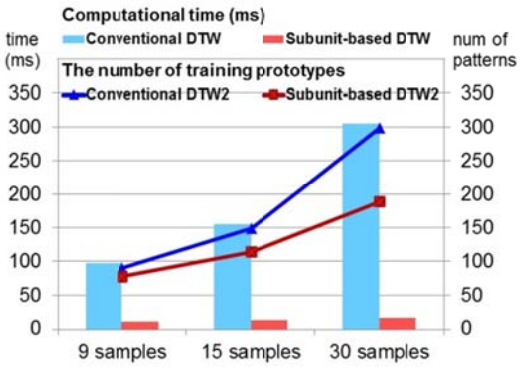


Fig. 5. Average computation time of recognition for different sizes of datasets

REFERENCES

[1] Wachs JP, Kölsch M Stern, et al (2011), Vision-based hand-gesture applications. Commun. ACM 54:60-71

[2] Okada S, Hasegawa O (2008), Motion recognition based on dynamic-time warping method with self-organizing incremental neural network. Proceedings of ICPR'08, pp.1-4

[3] Roussos A, Theodorakis S, Pitsikalis V, et al (2010), Hand tracking and affine shape-appearance handshape subunits in continuous sign language recognition. Proceedings of International Workshop on Sign, Gesture and Activity, ECCV 2010, pp.258-272

[4] Bauer B, Kraiss KF (2002), Towards an automatic sign language recognition system using subunits. In Revised Papers from the International Gesture Workshop on Gesture and Sign Languages in Human-Computer Interaction, vol.14, pp.64-75

[5] Elmezain M, Al-Hamadi A, Michaelis B (2008), Realtime capable system for hand gesture recognition using hidden markov models in stereo color image sequences. The Jour al of WSCG'08, vol.16, no.1, pp.65-72

[6] Cha SH, Shin YC, Shihari SN (1999), Approximate stroke sequence string matching algorithm for character recognition and analysis. In Proceedings of the ICDAR 1999, pp.53-56

[7] Han JW, Awad G, Sutherland A (2009), Modelling and segmenting subunits for sign language recognition based on hand motion analysis. Pattern Recognition Letters 30(6): 623-633

[8] Park HS, Jun CH (2009), A simple and fast algorithm for k-medoids clustering. Expert Systems with Applications 36:3336-3341