A Subunit-Based Dynamic Time Warping Approach for Hand Movement Recognition

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

[1] Graduate School of Information Science and Electrical Engineering
Kyushu University, Japan
{kenyou,atsushi,rin}@limu.ait.kyushu-u.ac.jp
[2] OMRON Corporation, Japan
takayosi@omm.ncl.omron.co.jp

Abstract. A subunit-based Dynamic Time Warping (DTW) approach is proposed 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 common sub-movements (subunits) are shared across hand gestures to obtain a smaller training data size and search space to improve recognition performance. (2) A similarity measure robust to variability 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. Our experimental results demonstrate the efficiency and accuracy of the proposed approach.

Keywords: hand movement, 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 hand location, appearance, motion, shape, and orientation often play an important role in hand gesture recognition. In this paper, we consider hand gestures as hand 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. However, 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 sub-sequences, as it requires continuity along the warping path.

A. Petrosino (Ed.): ICIAP 2013, Part I, LNCS 8156, pp. 672–681, 2013.

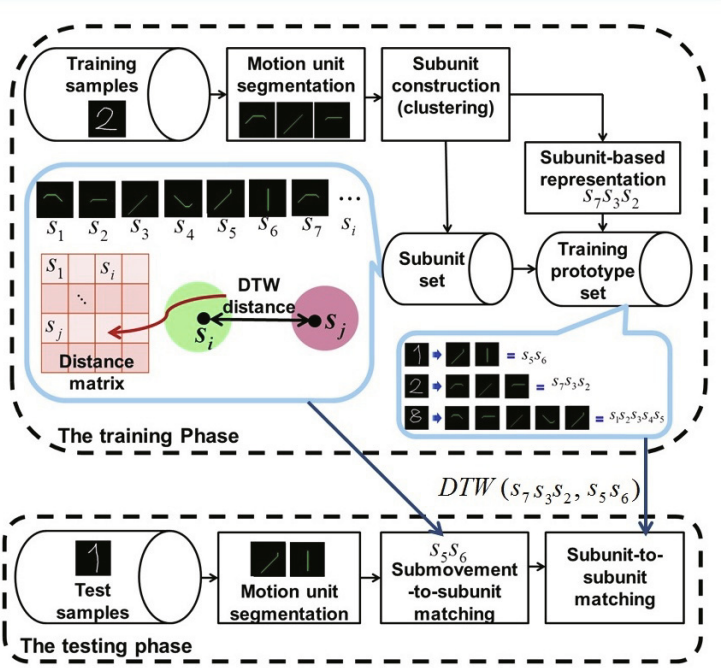


Fig. 1. Flowchart of the proposed approach

The conventional 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. Figure 1 is a block diagram of our hand movement recognition system. 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 k-medoids 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

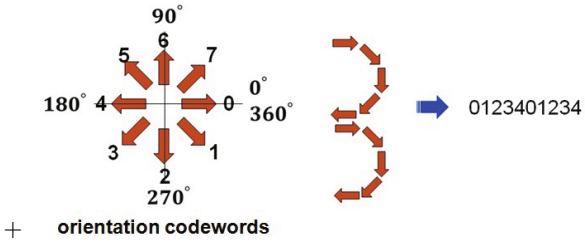


Fig. 2. Orientation codewords and an example of movement representation

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 and Learning

3.1 Hand Movement Representation

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

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

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 between 0 and 7 according to the yielded codewords as shown in Figure 2.

3.2 Hand Movement Subunit Construction

Currently, subunit-based recognition is a main focus of sign language research. There are two methods with which to perform recognition by means of subunits. The first method is based on linguistic analysis to determine subunits [6][7]. The second method segments signs into subunits employing a data-driven process without any linguistic knowledge about sign languages, and all subunits are self-organized from the data themselves [4].

(a) current orientation differing from the starting orientation

(b) significant changes between consecutive frames

Fig. 3. Segment motion unit according to changes in orientation

Motivated by [4] and [3], 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 that are considered in a common pool of features. Clustering is performed for these sub-movements (basic motion units) to find the common pattern of hand movements.

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 [8]. 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. As shown in Figure 3, all training data can be broken into sub-movements according to the characteristic.

Subunit Clustering. To select a set of representative subunits from all sub-movements 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. As opposed to k-means clustering, k-medoids clustering is more robust to outliers and avoids the difficulty of the averaging step of k-means clustering with DTW distance metric [9].

To determine the number of clusters, we employ an iterative k-medoids algorithm 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 k-means-like algorithm for k-medoids clustering [10] to overcome the drawback that partitioning around medoids (PAM, K-medoids) works inefficiently for large data sets because of the complexity.

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 computed off-line during construction of subunits and use it as a look-up table to speed up the recognition procedure.

3.3 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.

4 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 online learning to overcome sensitivity to variations in the training data.

4.1 Submovement-to-Subunit Matching

Let P_x be a testing sequence and $S = \{s_1, s_2, ..., s_{|S|}\}$ 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 between 0 and 7 according to changes in orientation and then segmented into m submovements u_{xi} in the same way as described in Section 3.2. We calculate DTW distances between submovement u_{xi} of the index i 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 $P_x = \{s_{x1}, s_{x2}, ..., s_{xi}, ..., s_{xm}\}$ is used to perform subunit-to-subunit matching with training data.

4.2 Subunit-to-Subunit Matching

Hand movement trajectories are recognized through dynamic subunit sequence matching. Let $P_y = \{s_{y1}, s_{y2}, ..., s_{yj}, ..., s_{yn}\}$ be a training sequence consisting of n subunits. The DTW distance $DTW(P_x, P_y) = D(s_{xm}, s_{yn})$ is calculated as follows.

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

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

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

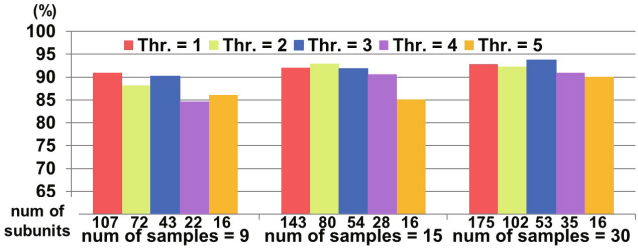


Fig. 4. Average recognition rate using different numbers of subunits

5 Experiments and Discussion

To test the proposed subunit-based DTW 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 two locally collected hand movement corpus. Here, the average computation time is the average time taken to calculate the distance.

5.1 Construction of Hand Movement Subunits

As illustrated in Section 3.2, we merge similar clusters iteratively to select a set of representative subunits. The similarity between clusters is measured as the DTW distance between medoids of clusters. Two clusters are merged if their distance is smaller than a threshold Th.

To evaluate the recognition performance with different thresholds Th, we performed experiments on subunit-based DTW with Th ranging from 1 to 5. The used 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. 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.

Figures 4 and 5 present the average recognition rate and computational time using different numbers of subunits, with Th ranging from 1 to 5. In general, the greater the number of subunits, the more the different sub-movements can be distinguished, contributing to more discriminative representation of the training data. This yields better recognition results. At the same time, the computational cost is high due to the lager search space as shown in Figure 5.

5.2 Recognition in Different Backgrounds

These experiments are aimed at demonstrating the robustness of our approach with respect to variability. We use a locally collected corpus consisting of 10 different classes of hand movement trajectories from 0 to 9. performed by 16

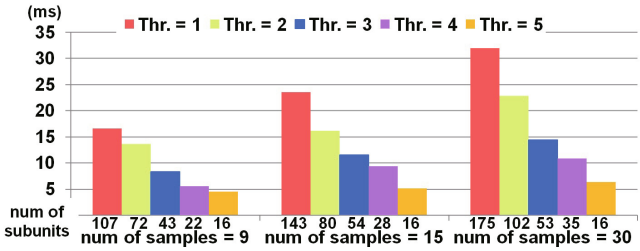


Fig. 5. Computational time using different numbers of subunits, constructed with Th ranging from 1 to 5

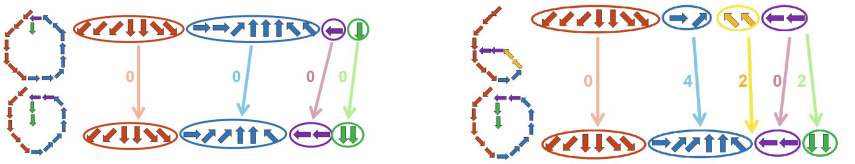
Table 1. Comparison of recognition rate in two different backgrounds

method \ background	blue	window
Conventional DTW	93.08%	71.70%
Subunit-based DTW	90%	82.5%

subjects in two different backgrounds: 1) the blue background easy to detect the position of fingertip and 2) the window background hard to to detect the position of fingertip. Each of the 10 classes of trajectories is performed one time in each background. We select 15 samples performed by 15 subjects in the blue background, to construct the training set. The trajectories performed by the other one subject in the blue and window background are used as test data.

Evaluation of the Recognition Rate. Recognition rates classified according to two different background are listed in Table 1. When using test data performed in the blue background, the subunit-based DTW achieves a recognition rate of 90%, which was 3.08% lower than that of conventional DTW. One possible reason could be that feature details is omitted due to the subunit-based movement representation. In contrast to recognition in the blue background, the subunit-based DTW approach shows a significant improvement of 10.8% when using test data performed in the window background. The two main reasons for the improvement are as follows.

A Similarity Measure Robust to Variability: 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 Figure 6, the proposed subunit-based DTW 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.



(a) distance between '0' and '0' is zero (b) distance between '0' and '6' is 8

Fig. 6. Examples of the subunit-based DTW distance between movements

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 $s_2 s_3$ and $s_1 s_2 s_3$ for three training data because P_z and P_w are mapped to the same prototype $s_2 s_3$. 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 $s_2 s_3$ 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 overcome the shortcoming of the expensive computational load resulting from a large number of training data.

These findings support the claim, made above, that the subunit-based DTW distance is able to overcome the sensitivity to training data of conventional DTW. This is expected to solve difficulties in hand movement recognition such as the larger variation within a class due to a userfs individual habits and noises.

Evaluation of Average Computation Time. Figure 7 presents the average computation time and the number of training prototypes when using test data performed in different backgrounds . The results indicate that a significant improvement in computational complexity was obtained. 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 complexity. 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.

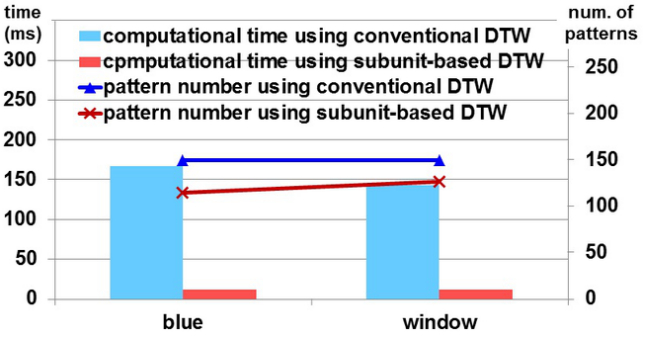


Fig. 7. Average computation time (ms) of recognition using test data performed in different backgrounds

Table 2. Comparison of average recognition rate on different size of training dataset

# training samples / method	9 samples	15 samples	30 samples
conventional DTW	95.46%	94.43%	95.24%
subunit-based DTW	90.54%	90.98%	93.3%
subunit-based DTW using online learning	95.06%	93.82%	94.5%

5.3 Training Datasets of Different Size

To evaluate performance changes resulting from differing sizes of training data, we use the same corpus described in Section 5.1. To obtain results that are more reliable, the construction of subunits and evaluationof recognition performance were repeated five times using different datasets constructed relating to different subjects.

Table 3. Comparison of average computation time on different size of training dataset

computation time(ms) / method	9 samples	15 samples	30 samples
conventional DTW	97.7	157.1	304.9
subunit-based DTW	11.6	14.425	16.2
subunit-based DTW using online learning	12.6	14.3	16.9

Recognition rates classified according to three different sizes of training set are compared in Table 2. The subunit-based DTW achieves recognition rates of 90.54%, 90.98%, and 93.3%, which are slightly lower than those of conventional DTW due to the loss of feature details. However, the proposed approach is easily applicable to incremental learning without a time consuming problem. Table 2 and Table 3 prove that a subunit-based online learning is able to address the problem and keep low computation cost.

6 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. Although a similar approach has been also proposed by [11], our proposed method realized much faster matching by introducing two-step DTWs and the efficient lookup table. Our future research will focus on incremental learning for the subunit itself to support efficient personalized recognition.

References

1. Wachs, J.P., Kolsch, M., Stern, H., Edan, Y.: Vision-based hand-gesture applications. Commun. ACM 54(2), 60–71 (2011)
2. Okada, S., Hasegawa, O.: Motion recognition based on dynamic-time warping method with self-organizing incremental neural network. In: 19th International Conference on Pattern Recognition, pp. 1–4 (2008)
3. Roussos, A., Theodorakis, S., Pitsikalis, V., Maragos, P.: Hand tracking and affine shape-appearance handshape subunits in continuous sign language recognition. In: Int. Conf. ECCV Wkshp: SGA (2010)
4. Bauer, B., Kraiss, K.-F.: Towards an automatic sign language recognition system using subunits. In: Wachsmuth, I., Sowa, T. (eds.) GW 2001. LNCS (LNAI), vol. 2298, pp. 64–75. Springer, Heidelberg (2002)
5. Elmezain, M., Al-Hamadi, A., Michaelis, B.: Real-time capable system for hand gesture recognition using hidden markov models in stereo color image sequences. Journal of WSCG, 65–72 (2008)
6. Liddell, S.K., Johnson, R.E.: American sign language: the phonological base. Sign Language Studies 64, 197–277 (1989)
7. Stokoe, W.: Sign language structure: an outline of the visual communication systems of the American deaf. Studies in Linguistics: Occasional Papers 8 (1960)
8. Han, J., Awad, G., Sutherland, A.: Modelling and segmenting subunits for sign language recognition based on hand motion analysis. Pattern Recognition Letters, 623–633 (2009)
9. Niennattrakul, V., Ratanamahatana, C.A.: On clustering multimedia time series data using k-means and dynamic time warping. In: International Conference on Multimedia and Ubiquitous Engineering, pp. 733–738 (2007)
10. Park, H.S., Jun, C.H.: A simple and fast algorithm for k-medoids clustering. Expert Systems with Applications: An International Journal, 3336–3341 (2009)
11. Oszust, M., Wysocki, M.: Modelling and Recognition of Signed Expressions Using Subunits Obtained by Data-Driven Approach. In: Ramsay, A., Agre, G. (eds.) AIMSA 2012. LNCS (LNAI), vol. 7557, pp. 315–324. Springer, Heidelberg (2012)