Neurocomputing 175 (2016) 674–685

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Sparse Observation (SO) Alignment for Sign Language Recognition

Hanjie Wang [a], Xiujuan Chai [a,c,*], Xilin Chen [a,b,c]

[a] Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing 100190, China
[b] Department of Computer Science and Engineering, University of Oulu, Finland
[c] Cooperative Medianet Innovation Center, China

ARTICLE INFO

Article history:
Received 17 July 2014
Received in revised form
17 August 2015
Accepted 30 October 2015
Communicated by T. Heskes
Available online 10 November 2015

Keywords:
Sign Language Recognition
Hidden Markov Model
Dynamic Time Warping
Stable Marriage Problem
RGB-D data

ABSTRACT

In this paper, we propose a method for robust Sign Language Recognition from RGB-D data. A Sparse Observation (SO) description is proposed to character each sign in terms of the typical hand postures. Concretely speaking, the SOs are generated by considering the typical posture fragments, where hand motions are relatively slow and hand shapes are stable. Thus the matching between two sign words is converted to measure the similarity computing between two aligned SO sequences. The alignment is formulated as a variation of Stable Marriage Problem (SMP). The classical "propose-engage" idea is extended to get the order preserving matched SO pairs. In the training stage, the multiple instances from one sign are fused to generate single SO template. In the recognition stage, SOs of each probe sign "propose" to SOs of the templates for the purpose of reasonable similarity computing. To further speed up the SO alignment, hand posture relationship map is constructed as a strong prior to generate the distinguished low-dimensional feature of SO. Moreover, to get much better performance, the motion trajectory feature is integrated. Experiments on two large datasets and an extra Chalearn Multi-modal Gesture Dataset demonstrate that our algorithm has much higher accuracy with only 1/10 time cost compared with the HMM and DTW based methods.

1. Introduction

As a kind of visual language, Sign Language (SL) is the most important communication means to exchange information within deaf community and also between deaf and hearing societies. Automatic Sign Language Recognition (SLR) is very important in many applications, such as sign language translation, sign language tutor, and special education [1–3]. However, the SLR still remains challenging as the complexity of the sign activities from the large scale body motion to tiny finger motion and also various hand postures.

From the literatures, four kinds of devices are used for SLR. They are data glove, video camera, depth camera and accelerator sensor. Many early SLR systems were equipped with data gloves and 3D trackers to collect the various information of hand shapes and the locations of two hands [4–6]. Although the cyber-glove and the position tracker can provide accurate and robust hand data, they are very expensive and inconvenient for the wearable characteristic. Some researchers explored the SLR based on video camera. Lee et al. [7] used VICON for data acquisition. However, the dorsum of the signers right hand should be secured with miniature reflective markers. This kind of work could be regarded as the substitute for the data-glove based method. From the viewpoint of the pure vision, Wang et al. [8] presented a method to classify 1113 signs and obtained 78% recognition accuracy for the top 10 candidates on dataset collected with normal cameras. One challenge of pure vision based method is the difficulty for the accurate hand tracking and segmentation. With the emergence of the novel sensors, there also appeared some ACC and sEMG based Sign Language Recognition (SLR) works [9,10]. Of course, such devices still belong to the wearable scope. Fortunately, the release of Microsoft Kinect sensor frees the signer from wearable devices by providing depth information as well as color images simultaneously [11–14]. For instance, by comparing between prototype Kinect-based CopyCat system and their previous CopyCat system, Zafrulla et al. [15] verified that the Kinect improved user comfort as well as system robustness.

Intrinsically, SLR is similar to activity recognition and gesture recognition. However, there are obvious differences among them. Different from the human body actions [16], most signs are performed only with the upper body (especially focus on hands and arms) and occasionally with subtle facial expressions [17]. Different from the gestures, the size of vocabulary of signs is much larger. For example, there are 20 categories of gestures in Chalearn Multi-modal Gesture Dataset [18] while we collected 1000 classes of signs for evaluating the proposed method in this paper. In

[*] Corresponding author.
E-mail addresses: hanjie.wang@vipl.ict.ac.cn (H. Wang), xiujuan.chai@vipl.ict.ac.cn (X. Chai), xlchen@ict.ac.cn (X. Chen).

http://dx.doi.org/10.1016/j.neucom.2015.10.112
0925-2312/© 2015 Elsevier B.V. All rights reserved.

activity and gesture recognition, the challenging and popular works dealing with one-shot learning [19,16,20] and even zero-shot learning [21] achieved good performance. One-shot learning uses only one sample from each class for training the model and zero-shot learning selects relative attributes as a semantic-link between the missing and available class. When compared with the few categories in action recognition, the size of the sign vocabulary is usually large. Consequently, to classify hand motions and postures with subtle changes using one-shot learning and zero-shot learning remains a big challenge for SLR. On the contrary, by collecting a large database, where each class of signs has more than one repetition for training and test, this paper focuses on multiple instance learning and proposes an instance merging method.

Traditional Hidden Markov Model (HMM), which is heavily borrowed from speech recognition, is the mainstream in the previous SLR works [4,17,18,22] and also recently works such as Martinez et al. [23]. The features for most of the methods are extracted from dense frames and are required sufficient training data to learn the parameters. So did Conditional Random Fields (CRF) [24], which also used parameters to encode given observations. In recent years, inspired by the basic phoneme in speech recognition, researches explored the basic unit in SL [25] and used HMM to create classifiers [5]. Wang et al. [5] focused on large vocabulary Chinese Sign Language Recognition based on phonemes. Eng-Jon et al. [26,27] recently created discriminative, multi-class classifier based on sequential patterns for SLR. Besides the state space based method, template matching methods are another kind of method class. Among them, Dynamic Time Warping (DTW) is one of the techniques widely used in gesture recognition [28]. To speed up the DTW, Stan Salvador and Philip Chan proposed fast-DTW [29]. Lichtenauer et al. [30] proposed a Statistical DTW, which outperforms HMM based methods for SLR. To build a practical systems, Chai et al. [3] proposed the state-of-the-art isolated SLR approach that matching aligned trajectories of hands with Kinect sensor. From the methods mentioned above, elements alignment is crucial for an optimal matching between two temporal sequences. In our work, we extracted SOs from isolated signs. Therefore, the sign matching can be achieved through the SOs alignment by solving a variant Stable Marriage Problem.

While there are published datasets captured with Kinect sensor for body actions recognition [16] and for gesture recognition [18,20], there is still lacking of available large vocabulary datasets for SLR, especially those captured by Kinect sensor to the best of our knowledge. In previous literatures, some experiments were conducted on small datasets. Chalearn Multi-modal Gesture Dataset [18] contains 20 Italian gestures, which were continuously performed by multiple signers variously. Almeida et al. [31] created a database with 34 specific signs recently. There were also some other experiments conducted on large datasets and yet without depth cue. For example, Eng-Jon Ong et al. [26] achieved 74.1% correct rate on 982 signs using Sequential Pattern Trees on singer dependent test. Their work can be taken as one of the state-of-the-art methods in SLR research. They also extended the Sequential Pattern Trees, which were used to recognize continuous SLR [27]. To further promote the performance of SLR by the assistant of the depth cue, we collected two datasets using Kinect sensor for research usage. One of them has a vocabulary of 370 and the other is up to 1000 classes of signs.

From above reviews, Kinect is a good balance between the data glove and pure visual camera. Its RGB-D data can be taken as the input for a natural and facilitate SLR system. However, RGB-D data based SLR also confronted with some intrinsic problems. For example, the registration between RGB and depth cues cannot be always accurate especially for the frames with fast hand motion. In other words, the accurate hand segmentation with depth data is not stable in fast motion case. In order to avoid this problem, a novel Sparse Observation (SO) representation is proposed, which also largely speed up the SLR. To better measure the similarity of two signs represented by two SOs sequences, the stable matched pairs should be found. Such SO alignment is formulated as a variant of SMP and the similarity between two SOs is calculated by the prior hand posture relationship map. The prior map can not only reduce the feature dimension but also remove some posture outliers. When combined with 3D hand motion trajectory, the proposed method can be used to accurately find the best matching of the sign candidates for a large vocabulary efficiently by a variant of classical Gale-Shapley solution to the SMP. The advantages of the proposed method lie on the following points: (1) the sign representation from the dense frame based model is simplified to this kind of SO; (2) the hand tracking and segmentation are more accurate for the SO due to the lower motion speed of hands; (3) the observation alignment problem of two signs can be solved effectively. To evaluate our method, two large vocabulary datasets were collected by us and part of the datasets has already been released for the research usage. The proposed method is also evaluated on the well-known Chalearn Multi-modal Gesture Dataset [18] and achieves good results.

Briefly speaking, the main contributions of this work are:

1. Different from the traditional dense frame based model, we propose SO representations, based on which, a framework is presented to realize the efficient and effective sign language recognition.
2. We formulate the similarity measurement between SO sequences as a variant of SMP. The classical Gale–Shapley solution is extended to get the order preserving matched SO pairs.
3. We construct a posture relationship map to generate the distinguished low-dimensional features of SOs, which are robust for versatile postures and can speed up the comparison.

The remaining part of this paper is organized as follows. Section 2 gives the technical overview. Section 3 describes the generation and alignment of sparse observations. More details about the implementation of our SLR are presented in Section 4. Section 5 reports the experimental results and also the comparison with HMM and DTW based methods. Section 6 concludes the paper.

2. Method overview

Fig. 1 illustrates the flowchart of SLR using the proposed method. The input is RGB-D data captured by Kinect sensor. As widely used, hand posture and motion trajectory are two key cues extracted from RGB-D data to characterize a sign. Hand motion is represented by left and right 3D hand trajectories T. Hand postures P are represented by HOG features in our sparse observations. To reduce the high dimension, the feature is re-mapped into a vector $\boldsymbol{p}$ by the prior of hand posture relationship map. The two kinds of cues P and T are shown in Fig. 1(b). In our framework, the hand posture and motion trajectory are used individually to give the similarity scores between query sample and gallery. Nevertheless, our paper emphasizes on the posture based SLR. Core steps of this procedure are given below and the details of implementation are illustrated in Section 4.

Sparse observations generation: Given the input data, the SOs are determined through the key posture fragments by considering the motion speed constraints. Thus a sign video is represented by a discrete SO sequence.

Recognition with posture relationship map: Once having the SOs, the recognition score can be obtained by solving a Stable Marriage

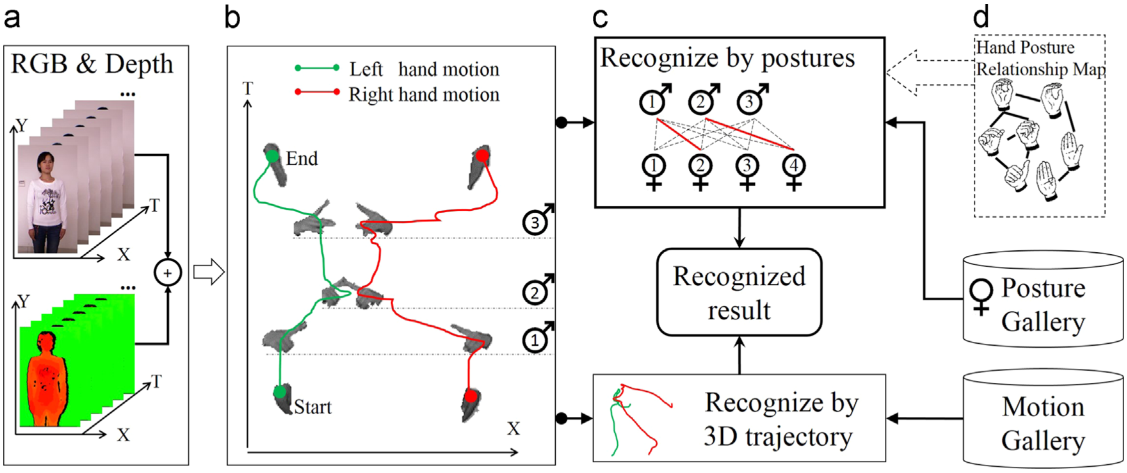


Fig. 1. The flowchart of SLR. (a) Input Stream of a sign "People". (b) Sparse observations of hand postures, left and right hand's motion trajectories in 2D (X-T). (c) Recognition. (d) The posture gallery, motion gallery and hand posture relationship map.

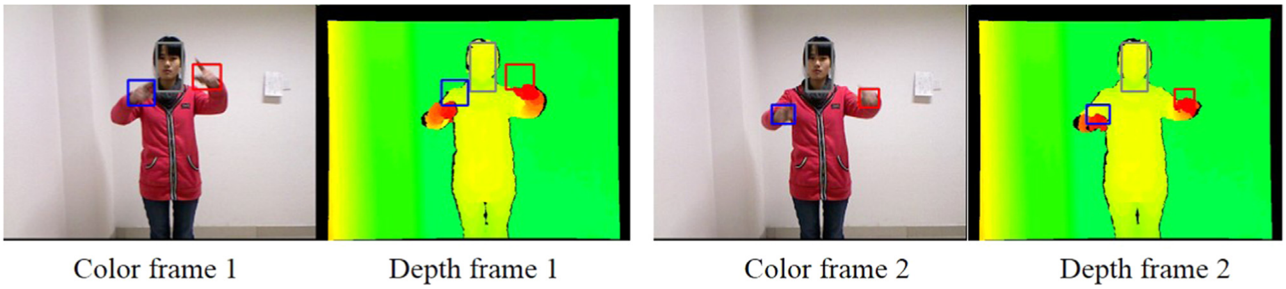


Fig. 2. Examples for inconsistency between color and depth data.

Problem using the feature generating from the posture relationship map.

To achieve a better recognition accuracy, the final result is obtained by fusing the matching scores from hand posture and motion trajectory.

3. Sparse observations

This section gives the details on sparse observation generation and alignment.

3.1. Sparse observation generation

As a kind of visual feature, hand posture can express much meanings for a sign. The most significant hand postures appear in video fragments with slow hands motions. Therefore, in this paper, we try to describe the sign by several significant hand postures corresponding to the key posture fragments. The discrete hand postures are called sparse observations. Experimentally, this representation lead to an excellent SLR correct rate on large size of vocabulary. This representation has two advantages, i.e., robust representation and efficient computing. It is robust because it avoids using the probably inaccurate hand segmentation in fast motion cases, which are shown in Fig. 2 with two examples. In Fig. 2, it can be found that though color and depth data have been carefully registered, the bounding boxes of hands in the RGB images are inconsistent with those in the depth maps. Our representation is efficient because it reduces the computational complexity compared with the dense frame based method since only very sparse observations are selected and used for the subsequent recognition.

Generally, there are two steps in the SO generation. First, in temporal, it is needed to find the key posture fragments with the hand speed constraints. An adaptive method is used to segment key posture fragments from the continuous video. The key posture fragment begins when the speed of the ith frame is lower than the average speed of the previous m frames (see Eq. (1)) and ends when the speed is larger than the average speed of frames from the beginning of this fragment (with the frame ID f) to the current frame i, see Eq. (2).

$$v_i < \alpha \frac{1}{m} \sum_{j=i-m}^{i-1} v_j \tag{1}$$

$$v_i > \beta \frac{1}{i-f} \sum_{j=f}^{i-1} v_j \tag{2}$$

Second, in spatial, hand posture segmentation in a frame takes advantage of both depth map and color image. Since signer tends to face the Kinect and his/her face can be detected robustly, a skin color model is initialized by the skin of human face and updated by skin pattern of hand region. The adaptive skin color model is used to search coarse hand region. Meanwhile, with depth constraint, most of the background can be removed perfectly. Obviously, it is fast and accurate, especially in the frames with low hand motion, where the registration of depth and color image is more accurate than the others. For the case when face, neck and hands are overlapping in RGB image, the depth of head will be

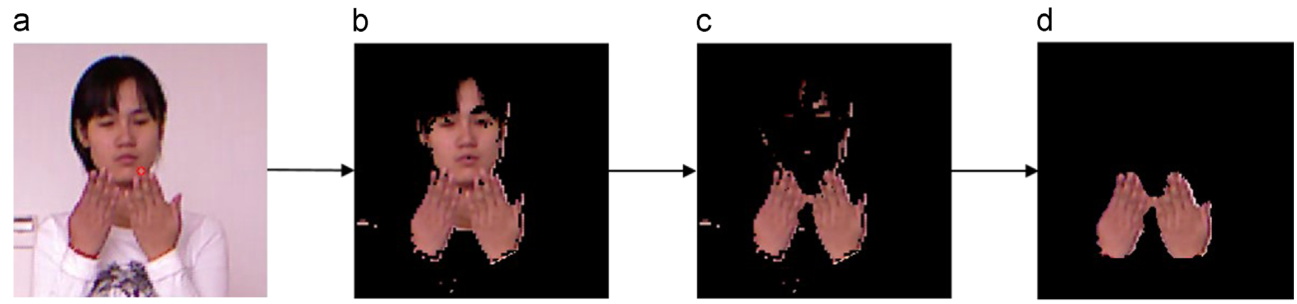
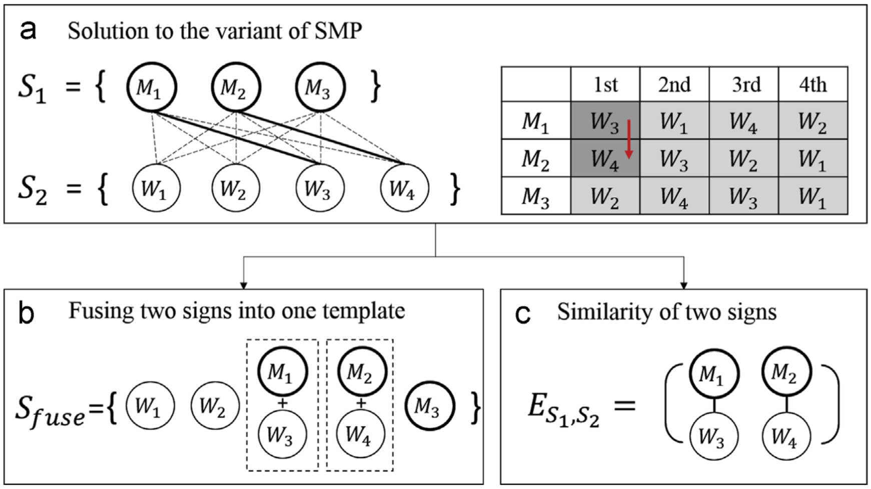


Fig. 3. Hand segmentation in case when face, neck and hands are overlapping. (a) Color image. (b) Skin region including face, neck and hands. (c) Most region of face and neck is removed by the depth. (d) Optimized hand region.


a Solution to the variant of SMP
$S_1 = \{ M_1 \; M_2 \; M_3 \}$
$S_2 = \{ W_1 \; W_2 \; W_3 \; W_4 \}$
b Fusing two signs into one template
$S_{fuse}=\{ W_1 \; W_2 \; M_1 + W_3 \; M_2 + W_4 \; M_3 \}$
c Similarity of two signs
$E_{S_1,S_2} = \begin{pmatrix} M_1 & M_2 \\ W_3 & W_4 \end{pmatrix}$

	1st	2nd	3rd	4th
M_1	W_3	W_1	W_4	W_2
M_2	W_4	W_3	W_2	W_1
M_3	W_2	W_4	W_3	W_1

Fig. 4. Sparse observation alignment. (a) Observations alignment on two sign videos by solving the variant of SMP. The solid lines connect stable matched pairs. (b) Gallery generation by fusing the matched observations. (c) Similarity of two signs.

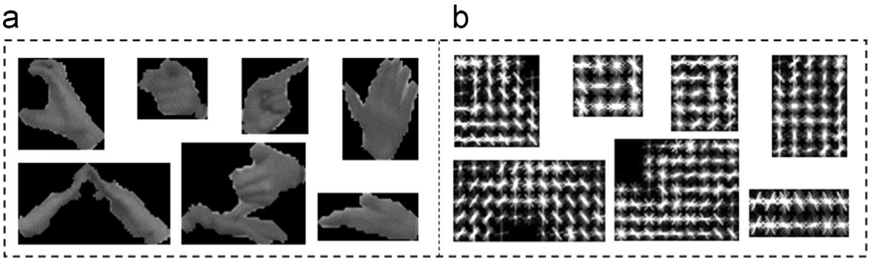


Fig. 5. HOG description. (a) The segmented hand posture. (b) The HOG feature illustration.

considered to distinguish the hand from the confused region. Fig. 3 gives an example of hand segmentation result in such cases.

3.2. Sparse observations alignment

Stable Marriage Problem (SMP) is the problem to find the best matching pairs between two sets (men and women in the prototype) of elements. David Gale and Lloyd Shapley [32] proved that, for any equal number of men and women, it is always possible to solve the SMP and makes all marriages stable. Our SO alignment problem just can be formulated as a variant of SMP and obtain stable matched SO pairs in a sequential order. In details, the SOs (men or women) are regarded as elements in the set, and the numbers may be not equal in two different sets (see S_1 and S_2 in Fig. 4). Most importantly, the matching of two sequential SOs must take the order of the elements into consideration since the SOs of a sign are performed in chronological order. Fig. 4(a) illustrates the variant of classical Gale-Shapley solution for SO alignment of two signs. The SO alignment contributes to the posture gallery generation and also the matching in SLR, as shown in Fig. 4(b) and (c).

Solving SMP is typically a “propose-engage” procedure as in Algorithm 1. The SOs in one sign video are acted as “men” and in the other sign video are acted as “women”. Each man ranks all the women in the order of preference. The classical Gale-Shapley solution is used to get the matching pairs between “men” and “women” from step 5 to step 19 in Algorithm 1. Thought the numbers of men and women are different, the typical one to one mapping can still be obtained on the sub-set, where the numbers of men and women are equal (the unmatched ones are

Algorithm 1. The procedure of SO alignment.

```
1  Given sign $M = \{P_1^1, P_2^1, ..., P_{l_1}^1\}$ and $W = \{P_1^2, P_2^2, ..., P_{l_2}^2\}$ ;
2  Define the elements in $M$ as men “$m$” and elements in $W$ as women “$w$”;
3  Compute the similarities of all the pairs between elements in $M$ and $W$
   using hand posture relationship map.
   $S(m_1, w_1), S(m_1, w_2), ..., S(m_{l_1}, w_{l_2})$. There are $l_1 \times l_2$ pairs totally;
4  Each man in $M$ ranks all the women in $W$ in the order of preference;
5  Set all the $m$ and $w$ free;
6  while Exists free man $m$ has potential women $w$ to propose to do
7      $w$ is $m$’s highest ranked women, who has not be proposed by $m$;
8      if $w$ is free then
9          ($m$, $w$) become engaged;
10     else
11         Some pair ($m'$, $w$) exists.;
12         if $w$ prefer $m$ to $m'$ then
13             ($m$, $w$) become engaged;
14             $m'$ become free;
15         else
16             ($m'$, $w$) keeps engaged;
17         end
18     end
19 end
20 while Exist cross engaged pairs do
21     Find the matched pair ($m$, $w$) with the maximum similarity score
       (larger than a threshold);
22     Detect the matched pairs ($m^*$, $w^*$), which have crossed connections
       with ($m$, $w$), and their total number $N$;
23     if $N == 1$ then
24         Delete the connection of ($m^*$, $w^*$);
25     else
26         if Exist 2 pairs of ($m^*$, $w^*$) that are not crossed then
27             Delete the connection of ($m$, $w$);
28         else
29             Delete all the connections of ($m^*$, $w^*$);
30         end
31     end
32 end
33 Output the stable matched observation pairs in sign $M$ and $W$;
```

eliminated). However, in order to preserve the observations' orders, cross matching pairs are not allowed. For example, the connecting lines of the pairs (M_1,W_3) and (M_2,W_1) in Fig. 4 are crossed and one of them should be deleted. Therefore, a pruning operation is adopted to maintain the sequential constraint by detecting the crossed engaged pairs and deleting the pairs with smaller similarities (step 20–22). It ensures the reasonable matching between two sign videos. Since the procedure of pruning we added has the complexity of $O(n^2)$, the total complexity of our solution for the variant of SMP is the same with classical Gale-Shapley algorithm ($O(n^2)$).

4. Implementation

In our SLR problem, both SO generation and alignment (Section 3) aim to achieve a good recognition performance and fast processing by searching the stable matched observation pairs in two sign videos.

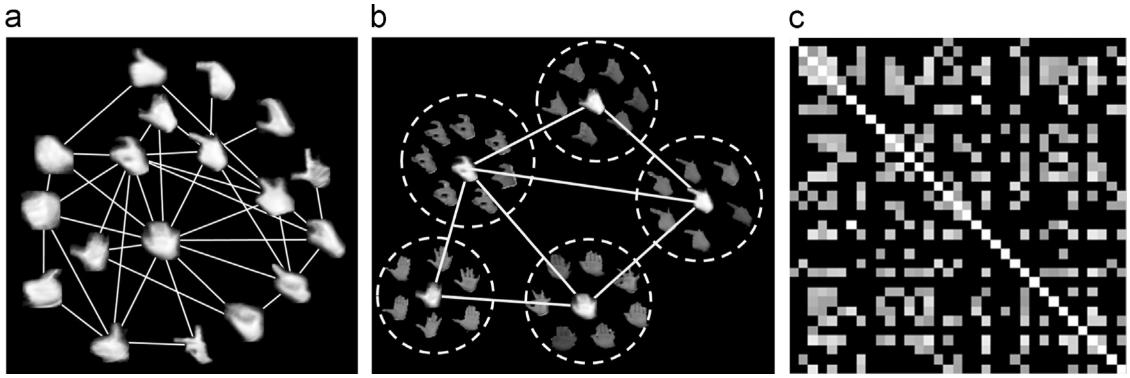


Fig. 6. Hand posture relationship map. (a) Example of a local hand posture relationship map for dominated hand (right hand). (b) The cluster centers and within class samples. (c) The sparse matrix for hand posture relationship.

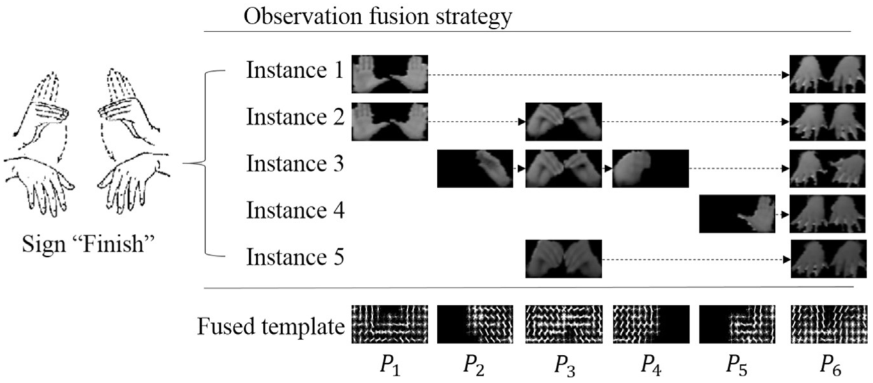


Fig. 7. Template fusion strategy of sign "Finish". We fuse two images from different instances by averaging their HOG features.

Followed that, this section introduces two important steps of SLR implementation: gallery generation and sign recognition. In both steps, the similarity between two SOs is computed with hand posture relationship map, which is introduced in the first sub-section. Finally, the SO represented hand posture P and motion trajectory T is fused to get better and more stable SLR result.

4.1. Hand posture relationship map

For each SO, we should consider the situations when left and right hands are apart or touching each other. Here we denote L as the left hand, R as the right hand and B as the two touching hands. We take the case "touching hand" into consideration since it is unwise to segment two overlapped hands arbitrarily. $\{L, R, B\}$ are denoted as the three channels to individually represent SOs. In sign language performing, the hand postures have tremendous variations for the flexible character. However, they can be clustered into many groups. These prior knowledge on the posture clusters and also their relations are very important to hand posture representation. Therefore, different from the direct matching [33], the hand posture relationship map is constructed to record the similarities between different SOs in each channel (L, R, B).

First step is the hand posture clustering. In order to describe the hand posture properly, Histogram of Oriented Gradient (HOG) is adopted. With a good segmentation, the edge of hand posture is clear and the gradients of regions around fingers are discriminative enough, as shown in Fig. 5. With HOG feature, unsupervised K-means is used to obtain the hand posture clusters automatically from plenty of training data in each channel.

In the second step, the connections among these clusters are built. Suppose K clusters are trained and μ_k are their cluster centers. The matrix consists of similarities of all the μ_k pairs is the original hand posture relationship map. However, most of the values in the map are small. They can be reset to zero if below a threshold, which makes the relationship map sparse. This sparse structure has the advantage of fast computing as well as less storage. With relationship map, feature dimension can be reduced tremendously. For example, a posture image with D dimensions of HOG feature ($\boldsymbol{h}$) can be represented as a sparse vector $\boldsymbol{p}$ with K (K is the number of the cluster and $K \ll D$) dimension. In details, suppose a hand posture P is classified to the class k with a probability of Pr_k. We select the vector $\boldsymbol{m}_k$ from the kth column of the sparse hand posture relationship matrix as in Fig. 6(c). The $\boldsymbol{p}$ is computed by Eq. (3).

$$\boldsymbol{p} = Pr_k \times \boldsymbol{m}_k, \tag{3}$$

where $\boldsymbol{p}$ is a vector with K dimensions and is served as the feature of hand posture in the rest of this paper.

The postures in three different channels (L, R, B) are clustered individually. Fig. 6(a) visualizes an example of a local relationship map of the dominant hand (right hand), as well as within class samples in Fig. 6(b). Thus, the similarity between two SOs can be obtained. The formulation of hand posture similarity is given as follows:

$$w_{\boldsymbol{p}_1,\boldsymbol{p}_2} = \sum_{\theta \in \{L,R,B\}} \lambda_\theta \frac{|\boldsymbol{p}_1^\theta - \boldsymbol{p}_2^\theta|}{K_\theta}, \tag{4}$$

where $\lambda_\theta \in \{0, 1\}$ represents the situation of two hands. K_θ is the class number of hand channel θ. $\boldsymbol{p}_1^\theta$ and $\boldsymbol{p}_2^\theta$ are the proposed feature vector of hand channel θ with posture relationship maps.

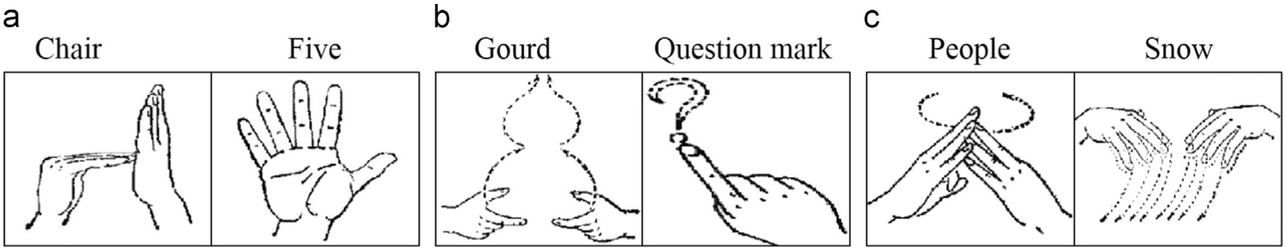


Fig. 8. Examples of three categories of Chinese sign language. (a) Posture dominant signs. (b) Trajectory dominant signs. (c) Together represented signs.

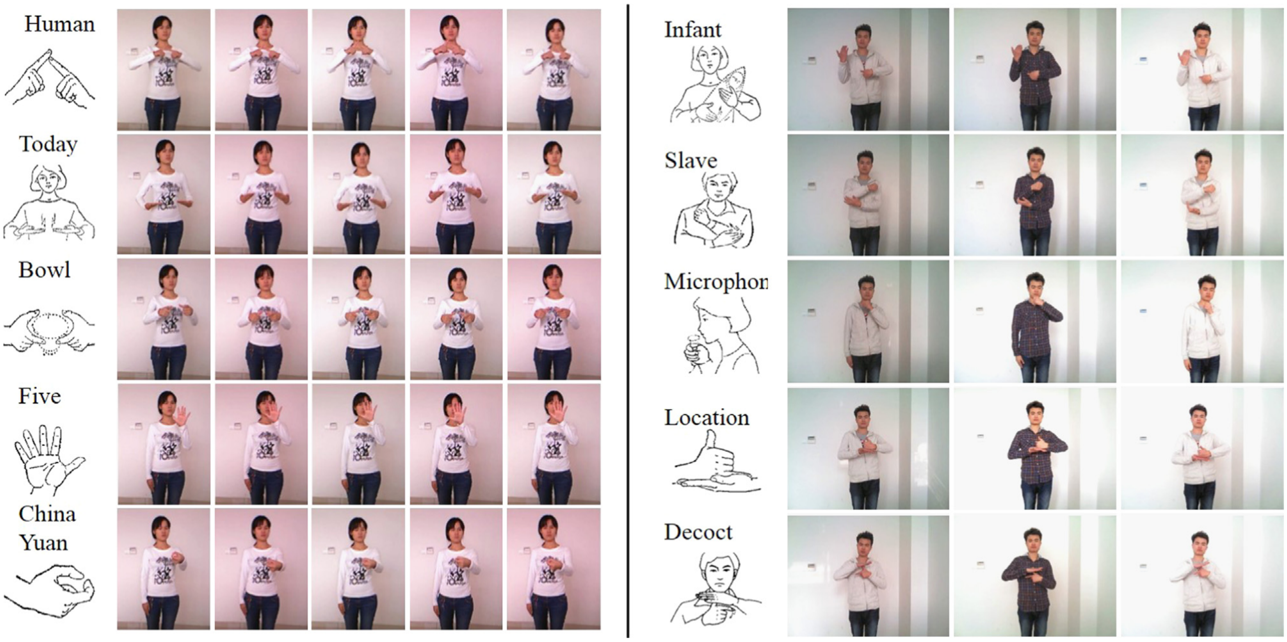


Fig. 9. Left: Daily SL samples performed by a signer for 5 times. Right: Large vocabulary SL samples performed by a signer for 3 times.

4.2. Gallery generation

In most template matching based methods, there are usually multiple templates in the gallery. To speed up the computation, we fuse these multiple sign samples into one general template. The stable matched pairs by SO alignment contribute heavily to the template fusion. The average feature of all the matched observations is set as the fused one in the final template. As shown in the Fig. 4(b), S_{fuse} has 5 elements, among which there are two fused observations. For example, the pair M_1 and W_3 are fused due to the high similarity score. After the postures being fused, the rest postures are arranged according to the similarities to the nearest fused postures.

Fig. 7 illustrates the fused result by the case of sign “Finish”. After observations in all the instances being aligned correctly, its total number decreases from 13 to 6 in the final template. The template has the property of generality and diversity. It is general due to the fused P_1, P_3, P_6 by multiple samples. It is diverse since individual postures P_2, P_4, P_5 are also included in the template.

4.3. Recognition based on sparse observations

Once obtain the gallery templates, any given probe sign can be recognized by template matching. Concretely speaking, it includes the following main steps. First, the SOs corresponding to the given probe sign are extracted. Then, the similarity between the SOs of the probe and any of the gallery templates are computed. The key to measure two different signs is the SO alignment, which can find the best matched pairs in chronological order. We define the stable matched pair set as Γ, and the total number is N. The final similarity score is computed only by calculating the similarities of these stable matched pairs as shown in Eq. (5),

$$E_{S_1,S_2} = \frac{1}{N}\sum_{(i,j)\in\Gamma} w_{\mathbf{p}_i,\mathbf{p}_j} + \lambda_N b(N), \tag{5}$$

where $\mathbf{p}_i \in S_1$ and $\mathbf{p}_j \in S_2$. $b(N)$ is the bonus term for the matched pair number in case two SOs differ too much in length. As stated in 4.1, the hand posture relationship map is used to fast obtain the similarity score $w_{\mathbf{p}_i,\mathbf{p}_j}$. Finally, the recognition result is determined to be the class with the maximum similarity E_{S_1,S_2}.

4.4. SLR by integrating SO and motion

Intrinsically, sign is a kind of 3D activity. Therefore, the depth data is a helpful compensation to 2D appearance feature for generating more discriminative feature. Actually, Chinese sign language words can be classified into three categories, i.e., posture dominant signs, trajectory dominant signs and together represented signs. Fig. 8 gives some examples of the three categories of signs. Therefore, motion trajectory is also used in our algorithm.

The dense 3D motion trajectory is used individually in our implementation. To characterize the motion, both the left and right hand movements are recorded as 3D trajectory vectors.

The trajectories are normalized in two steps. First, each trajectory is translated, rotated and resized according to the position and the size of signer. Second, an even distribution with fixed points number M is generated by deleting or interpolation. Given two 3D motion trajectories, similarity between them is measured as:

$$E_{T_i,T_j} = 1 \Big/ \exp\left(\lambda_T \cdot \sum_{m=1}^{M} dis(T_i^m, T_j^m)\right). \qquad (6)$$

where $dis(T_i^m, T_j^m)$ is the Euclidean distance of two points in normalized 3D trajectories and λ_T is the weight coefficient.

When given query sample and gallery templates of sign classes, both sparse observations and motion trajectories are used to estimate the similarity between query sample and each class template k. We merge the results from the two cues (E_{S,S_k} and E_{T,T_k}) by weight γ and get the recognized result by Eq. (7).

$$\arg\max_k(\gamma \times E_{S,S_k} + (1-\gamma) \times E_{T,T_k}) \qquad (7)$$

5. Experiment

In this section, firstly, we will give a brief introduction to our collected data sets. Then the experimental results of SLR on the two datasets and another published dataset are provided in detail, including the comparison with the HMM, two DTW based methods and the method of Chai [3].

5.1. Data sets

To the best of our knowledge, there is no public dataset of large vocabulary Chinese sign language using Kinect sensor. Therefore, we collected the high quality data with the help of deaf signers. Totally, two datasets are built in our work for different purposes.

1. *Daily SL dataset*: This dataset includes 370 different signs selected from the daily used words. The signs are performed by 1 signer 5 times. We call it *Dataset A*, which has 370 × 5 sign sequences totally. Fig. 9 (left) gives some samples.
2. *Large vocabulary SL dataset*: This dataset includes 1000 isolated signs performed by another signer 3 times. We call it *Dataset B*, which has 1000 × 3 sign sequences totally. Fig. 9 (right) gives some samples.

In our data collection procedure, all the signers were standing in front of the Kinect sensor with a distance around 1.5 m. As captured by Kinect, the data sequence including depth maps and color images are recorded. Simultaneously, the skeleton tracking results provided by Kinect Windows SDK are also recorded.

We also evaluated the proposed method on Chalearn Multimodal Gesture Dataset [18]. The dataset contains several people performing gestures from a vocabulary of 20 Italian sign gesture categories. It is designed for small-vocabulary signer-independent continuous gesture recognition. We select part of the dataset in this paper to show the performance of our method.

Table 1
The experimental results of using different features on dataset A. "P" represents hand posture, "T" represents motion trajectory. "P&T" is the result using two cues. Column "P-BF" is the results from Brute-Force (BF) matching on SO.

Test instance	P	P-BF	T	P&T
Ins. 1	0.446	0.414	0.754	0.840
Ins. 2	0.695	0.624	0.773	0.930
Ins. 3	0.746	0.714	0.759	0.932
Ins. 4	0.743	0.708	0.781	0.940
Ins. 5	0.719	0.670	0.754	0.930
Average	0.670	0.626	0.764	**0.915**

5.2. Experimental results and analysis

In this part, the experiments on isolated sign classification are conducted to evaluate the effectiveness of the SO generation and alignment. Three datasets are used for the experiments and evaluations. First, an elaborate evaluation on using different features is given to show the performance of postures and motion trajectories, also the result of the fusion features is listed. Then the performance is compared to HMM [5] and DTW based methods (fast-DTW [29] and Statistical DTW (SDTW) [30]) to show the validity of the proposed sparse observation representation and also recognition. We denote "Ins. n" as the No.*n* instance, which includes all the sign words sampled in one time by each signer. In the following experiments, the leave-one-out cross-validation strategy is adopted. For example, if "Ins. n" is used for testing, the other instances are used for training.

In our implementation, the hand posture image is resized to 64 × 64 pixels for either the left or right hands and 128 × 64 pixels for the hands touching case. In the HOG transformation, the block size is 16 × 16, cell size is 8 × 8, bin size is 9, and the feature vector dimension of HOG feature $\boldsymbol{h}$ is 1764 in total. Hand posture relationship maps for left, right and both channels are all learned from hand postures in more than 2000 random selected frames. After K-means clustering, left hand has 35 classes, right hand has 100 classes and touching hand has 50 classes, i.e., K_L=35, K_R=100, K_B=50 and the dimension of $\boldsymbol{p}^L$ is 35, $\boldsymbol{p}^R$ is 100, $\boldsymbol{p}^B$ is 50 in Eq. (4).

5.2.1. Comparison on features

The experiments on SLR with separated feature cue (sparse observation (P) and motion trajectory (T)) and also the combined cues are all conducted on dataset A. See from Table 1, a good promotion is achieved when combining features from two cues together. Another matching method (named as brute-force matching) is served as a comparison to our proposed stable matching solution in column "P-BF". Brute-force matching simply using all the m sparse observations of query sample to the matched n elements in templates. Theoretically, it is time consuming for requiring of totally $C_n{}^m$ comparison times. Its performance may also be affected due to the situations like observation superfluous or absence in the query samples. While our proposed stable matching solution excludes those observations effectively and achieves 5% superiority to brute-force matching. Considering for the good performance, in our following experiments, we fix the feature to be "P&T".

5.2.2. Evaluation on daily SL

We conduct the isolated sign classification on dataset A. The recognition of daily used signs is utterly important to the SLR applications. The experimental results with the comparison of HMM based method and DTW based methods are given in Table 2. The result of HMM-based method is evaluated by the code used in the [5] with dense frame based hand posture and hand motion trajectory. Two DTW methods also use the same feature in the dense frames. All the experiments use the same segmentation of hand posture. Our method achieves 91.5% average recognition accuracy for top 1 matching in this small vocabulary set. It should be figured out that the average top 1 recognition accuracy exceeds that of HMM based method by 7.9 percentage points, fast-DTW based method by 13.1 percentage points and SDTW by 8.3 percentage points. We conduct one-way ANOVA (Analysis Of Variance) on the dataset and the results are listed in Table 3 and illustrated in Fig. 10. It can be seen from the table and the

Table 2
The experimental results and the comparison in dataset A (370 vocabulary set).

Test instance		Ins. 1	Ins. 2	Ins. 3	Ins. 4	Ins. 5	Ave.	Time Cost
The Proposed Method	Top 1	0.840	0.930	0.932	0.940	0.930	**0.915**	18.5 ms/sign
	Top 5	0.945	0.984	0.991	0.986	0.986	**0.979**	
	Top 10	0.962	0.986	0.994	0.989	0.989	**0.984**	
HMM [5]	Top 1	0.840	0.851	0.881	0.816	0.791	0.836	176 ms/sign
	Top 5	0.940	0.946	0.962	0.943	0.921	0.943	
	Top 10	0.976	0.978	0.978	0.967	0.967	0.973	
Fast-DTW [29]	Top 1	0.830	0.821	0.881	0.775	0.614	0.784	185 ms/sign
	Top 5	0.949	0.927	0.951	0.892	0.770	0.898	
	Top 10	0.976	0.957	0.973	0.935	0.824	0.933	
SDTW	Top 1	0.772	0.819	0.865	0.857	0.849	0.832	170 ms/sign
	Top 5	0.895	0.927	0.957	0.924	0.911	0.923	
	Top 10	0.927	0.951	0.970	0.949	0.946	0.949	
Chai et al. [3] Only trajectory feature	Top 1	0.768	0.824	0.824	0.849	0.786	0.810	11.2 ms/sign
	Top 5	0.892	0.919	0.930	0.941	0.911	0.919	
	Top 10	0.927	0.951	0.951	0.970	0.930	0.946	

Table 3
Results of One-way ANOVA on dataset A.

Source	SS	df	MS	F	Prob > F
Columns	0.04753	4	0.01188	3.74	0.0199
Error	0.06358	20	0.00318		
Total	0.11111	24			

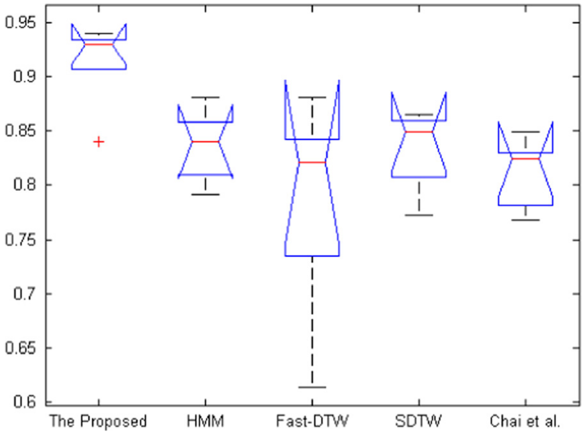


Fig. 10. The boxplot of One-way ANOVA on dataset A.

Table 4
Evaluations of λ_N in Eq. (5). The average recognition accuracies are given.

Values of λ_N	Top 1	Top 5	Top 10
0.00	0.897	0.979	0.986
0.01	0.902	0.980	0.986
0.03	**0.915**	**0.979**	**0.984**
0.05	0.904	0.981	0.986
0.10	0.878	0.971	0.980
1.00	0.658	0.777	0.809

Table 5
Evaluations of γ in Eq. (7). The average recognition accuracies are given.

Values of γ	Top 1	Top 5	Top 10
0.3	0.808	0.928	0.959
0.4	0.821	0.938	0.966
0.5	0.840	0.949	0.972
0.6	0.864	0.964	0.981
0.7	0.891	0.977	0.986
0.8	**0.915**	**0.979**	**0.984**
0.9	0.885	0.967	0.978
1.0	0.671	0.844	0.878

Table 6
The experimental results and the comparison on dataset B (1000 vocabulary set).

Test instance		Ins. 1	Ins. 2	Ins. 3	Ave.
The Proposed Method	Top 1	0.721	0.77	0.741	**0.744**
	Top 5	0.877	0.913	0.882	**0.891**
	Top 10	0.914	0.942	0.922	**0.926**
HMM [5]	Top 1	0.607	0.637	0.608	0.617
	Top 5	0.806	0.818	0.779	0.801
	Top 10	0.872	0.867	0.833	0.857
Fast-DTW [29]	Top 1	0.542	0.822	0.802	0.722
	Top 5	0.714	0.938	0.928	0.86
	Top 10	0.772	0.957	0.954	0.894
SDTW	Top 1	0.738	0.747	0.670	0.718
	Top 5	0.897	0.886	0.851	0.878
	Top 10	0.927	0.923	0.893	0.914
Chai et al. [3] Only trajectory feature	Top 1	0.603	0.618	0.552	0.591
	Top 5	0.778	0.797	0.734	0.770
	Top 10	0.827	0.841	0.788	0.819

illustration of boxplot, the *p*-value (0.0199) are less than the significance level ($p < 0.05$). Therefore, the results of SO are statistically significant and not just likely chance occurrences in the experiments.

Another important advantage of our method is the fast speed of processing. The SO and hand posture relationship map greatly speed up the processing. The time cost for the proposed method is 18.5 ms per sign while the cost for HMM based method is 176 ms per sign, almost 10 times when compared to ours. As for fast-DTW and SDTW, almost 185 and 170 ms cost per sign respectively, which are also unacceptable in real-time system. While the SO generation is real time by hand motion speed constraint (there is approximate 4 SOs in average for a word of Chinese SL) and SO alignment is fast by classical Gale-Shapley solution. In addition, by using the hand posture relationship map, the 1764 dimension HOG

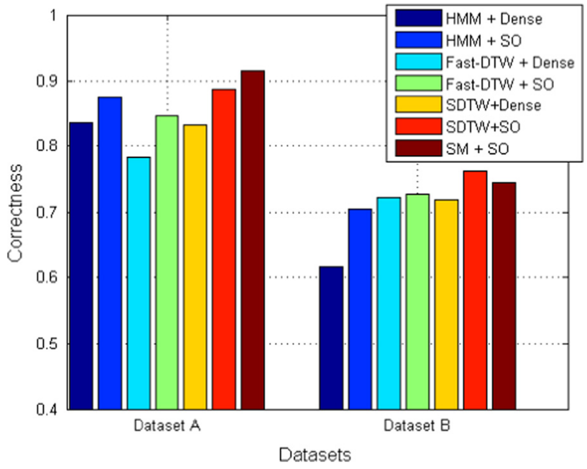


Fig. 11. Comparison the dense and sparse frames in different methods.

feature is transformed into discriminated vector with dimension less than 100.

λ_N in Eq. (5) is fixed to 0.03 in our experiments. The other values for λ_N are also evaluated on dataset A as shown in Table 4. The parameter γ in Eq. (7) is the weight of merging hand posture scores and motion trajectory scores. The performance with different values on dataset A is shown in Table 5. It can be seen from Table 5, the hand postures, i.e., sparse observations, contribute much (0.8 for hand posture and 0.2 for motion trajectory) to the final recognition results. To this end, we fix the parameter γ to 0.8 in all of the experiments.

5.2.3. Evaluation on large vocabulary SL

We also carry out the experiment on the large vocabulary SLR, which is a very challenging task for the wide difference on the variation of hand postures, as well as motion trajectories. The experiment configuration is similar with the above experiment and the experimental results are shown in Table 6. In this experiment, the whole performance is lower than that of the dataset A, but our proposed method also outperforms HMM, fast-DTW and SDTW and achieves an average accuracy of 74.4%, 89.1% and 92.6% in top 1, top 5 and top 10 respectively. Meanwhile, in this large vocabulary SLR, HMM can only get the average accuracy of 61.7% in top 1 with a roughly 12.7 percentage points decreasing compared with the proposed method. As stated before, in HMM, some inaccurate hand segmentation results, especially for the fast motion cases, are inevitably used in the HMM training and testing, which will affect the performance. Fast-DTW performs the best in the round “Ins. 2” and “Ins. 3” while the worst in the round “Ins. 1”. The large variance as well as the heavy computational burden makes Fast-DTW based method inferior to the proposed method. SDTW conquers the problem of large variance in Fast-DTW and has approximate recognition accuracies of the three rounds of evaluations. Comparatively speaking, the performance of our algorithm is relative stable. That is benefited from using well segmented hand posture and the proposed sparse observation representation.

From Table 6, we can also see that our proposed method achieves nearly 15 percentage points promotion in top 5 compared to top 1 case. Therefore, the performance can be improved dramatically within top 5 matching candidates if the context-based language model is integrated.

5.2.4. Evaluation between dense frames and SOs

The experiments for HMM and DTW are conducted on dense frame based signs in the above experiments. To verify the superiority of the proposed sparse observation, we also conduct experiments to evaluate HMM and DTW based methods only using the SO. However, the number of SOs from one sign is so small that even less than the state number of HMM. We just use all the frames in the key fragment, where SOs are extracted. Fig. 11 shows the experimental results. The correctness of HMM+SO and DTW+SO outperforms the HMM+dense and DTW+dense respectively, while in a more fast speed due to the less frames to be involved in processing. Though, the recognition accuracy of SDTW+SO is better than SM+SO, the time cost is 247 ms per sign for SDTW+SO and only 30 ms per sign for SM+SO. In other aspect, we can also see that the novel SO representation is effective for the performance enhancement not only for our used SM strategy, but also for other methods, such as HMM, Fast-DTW and SDTW.

Table 7
The experimental results and the comparisons on Chalearn Multi-modal Gesture Dataset.

Test group	Our method	HMM [5]	Fast-DTW	SDTW
G01	0.80	0.70	0.75	0.30
G02	0.85	0.80	0.85	0.45
G03	0.95	0.85	0.70	0.45
G04	0.95	0.90	0.80	0.55
G05	0.95	0.90	0.85	0.65
G06	1.00	0.95	0.90	0.60
G07	0.85	1.00	0.90	0.60
G08	0.90	0.95	0.85	0.70
G09	0.95	0.85	0.95	0.65
G10	0.85	0.80	0.85	0.50
G11	0.90	0.80	0.80	0.50
G12	1.00	0.95	0.85	0.70
G13	1.00	1.00	0.90	0.85
G14	0.90	0.90	0.80	0.75
G15	0.85	0.90	0.80	0.70
G16	0.90	0.80	0.75	0.60
G17	0.95	1.00	0.90	0.70
G18	0.90	0.95	0.90	0.70
G19	0.90	0.95	0.95	0.75
Average	**0.913**	0.892	0.845	0.616

Table 8
Results of One-way ANOVA on Chalearn Multi-modal Gesture Dataset.

Source	SS	df	MS	F	Prob > F
Columns	1.06668	3	0.35556	43.74	3.31234e – 16
Error	0.58526	72	0.00813		
Total	1.65194	75			

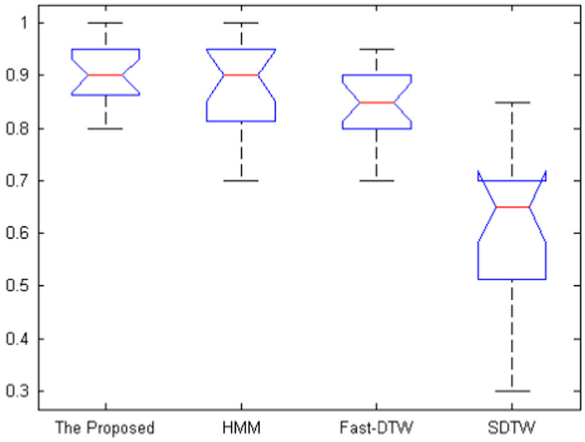


Fig. 12. The boxplot of the One-way ANOVA on Chalearn Multi-modal Gesture Dataset.

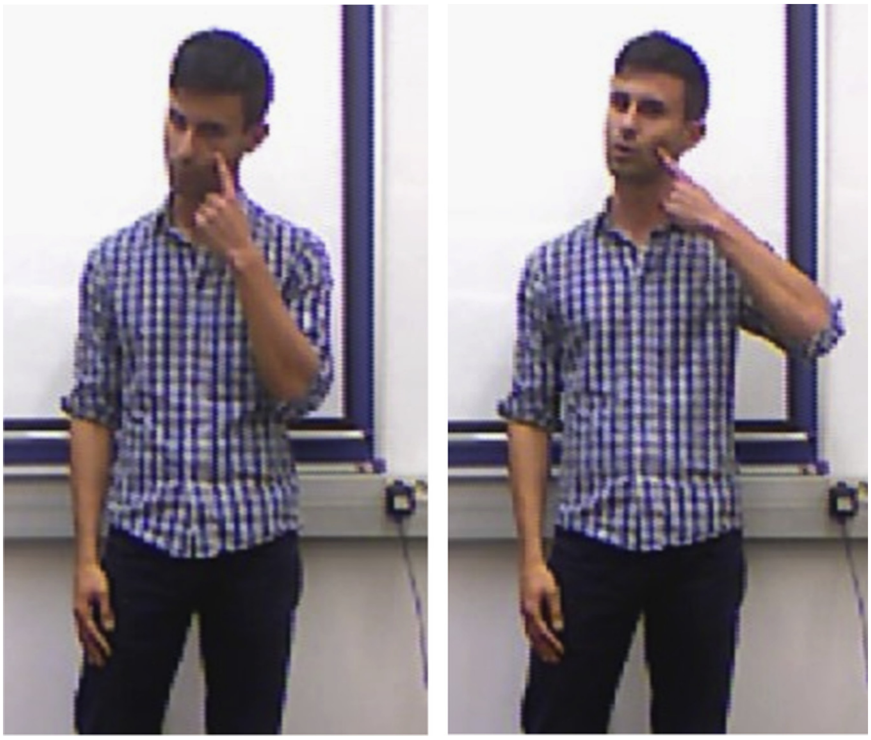


Fig. 13. One pair of confused gestures: furbo (left) and buonissimo (right).

5.2.5. *Evaluation on Chalearn Multi-modal Gesture Dataset*

Chalearn Multi-modal Gesture Dataset is designed for small-vocabulary continuous gesture recognition. Though there are only 20 categories gestures, several pairs of them are confused and hard to be discriminated. Moreover, the way of signing are variant even for the same person in two repetitions. For example, the first gesture “vattene” is performed randomly by right or left hand. Since we focus on isolated SLR, we use all the data from the first signer (including 20 gestures × 19 samples) in the folder “train”, which provides manually segmentations, and apply leave-one-out cross-validation.

In each round of testing, one of the samples is used for test and the rest for training. For each sign, sparse observations are extracted from the 18 training samples, which will be fused into one model by the gallery generation method proposed in Section 4.2. See from Table 7, the recognition results of the proposed method outperforms HMM, Fast-DTW and SDTW. The statistical analyses ANOVA is also given in Table 8 and illustrated in Fig. 12. The results are statistically significant with a very small *p*-value. It can be seen that the performance of HMM is comparable to the proposed method in this dataset since the number of training samples (18 samples for each) is larger than that in Dataset A (4 samples for each). As we known, the more training data, the better HMM performs. However, since the sign data collection is difficult and time consuming, the training samples are always limited. From this point of view, the proposed method is better than HMM especially when the training samples are less. As mentioned above, for the same kind of gesture, the signing pattern maybe different, which leads to inferior statistical models in SDTW. Therefore, SDTW based method has low recognition accuracy. As to the failed cases by our method, some of them are indeed very confused. For instance, Fig. 13 shows one example of confused pairs. Dealing with these challenging cases is beyond this paper and will be our future works.

6. Conclusion and future works

This paper proposes a novel sparse observation representation and a framework for efficient SLR. The matched SO pairs of two signs are found through SO alignment by solving a variant of SMP. The similarity between the stable matched observations is efficiently evaluated by a prior hand posture relationship map, which transfers the high dimension HOG feature into a low dimensional sparse vector. Heavily benefit from the SO alignment and hand posture relationship map, a general and diverse gallery template is built by fusing multiple SO represented samples and is used for SLR. Experimental results on three datasets convincingly show that our method is superior to the traditional HMM and DTW based methods on both accuracy and processing speed. While experiments of this paper are conducted on user dependent datasets, our future work mainly focuses on signer independent SLR on large vocabulary datasets. In addition, with well-trained language model for Chinese sign language, continuous SLR can be realized.

Acknowledgment

This work was partially supported by the Microsoft Research Asia, and Natural Science Foundation of China under contract Nos. 61390511, 61472398, and the FiDiPro program of Tekes.

References

[1] S. Ong, S. Ranganath, Automatic sign language analysis: a survey and the future beyond lexical meaning, IEEE Trans. Pattern Anal. Mach. Intell. 27 (6) (2005) 873–891, http://dx.doi.org/10.1109/TPAMI.2005.112.

[2] Z. Zafrulla, H. Brashear, P. Yin, P. Presti, T. Starner, H. Hamilton, American sign language phrase verification in an educational game for deaf children, in:

International Conference on Pattern Recognition, 2010, pp. 3846–3849. http://dx.doi.org/10.1109/ICPR.2010.937.

[3] X. Chai, G. Li, Y. Lin, Z. Xu, Y. Tang, X. Chen, M. Zhou, Sign language recognition and translation with kinect, in: IEEE Conference on AFGR, 2013.

[4] W. Gao, G. Fang, D. Zhao, Y. Chen, Transition movement models for large vocabulary continuous sign language recognition, in: IEEE Conference on Automatic Face and Gesture Recognition, 2004, pp. 553–558. http://dx.doi.org/10.1109/AFGR.2004.1301591.

[5] C. Wang, W. Gao, S. Shan, An approach based on phonemes to large vocabulary chinese sign language recognition, in: IEEE Conference on Automatic Face and Gesture Recognition, 2002, pp. 411–416. http://dx.doi.org/10.1109/AFGR.2002.1004188.

[6] W.W. Kong, S. Ranganath, Automatic hand trajectory segmentation and phoneme transcription for sign language, in: IEEE Conference on Automatic Face and Gesture Recognition, 2008, pp. 1–6. http://dx.doi.org/10.1109/AFGR.2008.4813462.

[7] Yung-Hui Lee, Cheng-Yueh Tsai, Taiwan sign language (tsl) recognition based on 3d data and neural networks, Expert Syst. Appl. 36 (2009) 1123–1128.

[8] H. Wang, A. Stefan, S. Moradi, V. Athitsos, C. Neidle, F. Kamangar, A system for large vocabulary sign search, in: Trends and Topics in Computer Vision, Springer, Crete, 2012, pp. 342–353.

[9] Y. Li, X. Chen, X. Zhang, K. Wang, Z. Wang, A sign-component-based framework for chinese sign language recognition using accelerometer and sEMG data, IEEE Trans. Biomed. Eng. 59 (10) (2012) 2695–2704, http://dx.doi.org/10.1109/TBME.2012.2190734.

[10] V. Kosmidou, L. Hadjileontiadis, S. Panas, Evaluation of surface EMG features for the recognition of American sign language gestures, in: Engineering in Medicine and Biology Society, 2006, pp. 6197–6200. http://dx.doi.org/10.1109/IEMBS.2006.259428.

[11] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, A. Blake, Real-time human pose recognition in parts from single depth images, in: IEEE Conference on Computer Vision and Pattern Recognition, 2011, pp. 1297–1304. http://dx.doi.org/10.1109/CVPR.2011.5995316.

[12] I. Oikonomidis, N. Kyriazis, A. Argyros, Efficient model-based 3d tracking of hand articulations using kinect, in: Proceedings of the British Machine Vision Conference, Dundee, BMVA Press, 2011, pp. 101.1–101.11.

[13] N. Pugeault, R. Bowden, Spelling it out: Real-time asl fingerspelling recognition, in: ICCV Workshops, 2011, pp. 1114–1119. http://dx.doi.org/10.1109/ICCVW.2011.6130290.

[14] J. Charles, M. Everingham, Learning shape models for monocular human pose estimation from the microsoft xbox kinect, in: ICCV Workshops, 2011, pp. 1202–1208. http://dx.doi.org/10.1109/ICCVW.2011.6130387.

[15] Z. Zafrulla, H. Brashear, T. Starner, H. Hamilton, P. Presti, American sign language recognition with the kinect, in: Proceedings of the 13th International Conference on Multimodal Interfaces, ACM, New York, 2011, pp. 279–286.

[16] J. Wan, Q.Q. Ruan, W. Li, S. Deng, One-shot learning gesture recognition from RGB-D data using bag of features, J. Mach. Learn. Res. 14 (2013) 2549–2582.

[17] H. Cooper, B. Holt, R. Bowden, Sign language recognition, in: Visual Analysis of Humans, Springer, 2011, pp. 539–562. ISBN 978-0-85729-996-3.

[18] S. Escalera, J. Gonzàlez, X. Baró, M. Reyes, O. Lopes, I. Guyon, V. Athitsos, H. Escalante, Multi-modal gesture recognition challenge 2013: dataset and results, in: Proceedings of the 15th ACM on International Conference on Multimodal Interaction, Sydney, ACM, 2013, pp. 445–452.

[19] I. Guyon, V. Athitsos, P. Jangyodsuk, B. Hamner, H.J. Escalante, Chalearn gesture challenge: design and first results, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2012, pp. 1–6.

[20] T. Pfister, J. Charles, A. Zisserman, Domain-adaptive discriminative one-shot learning of gestures, ECCV 2014 (2014) 814–829.

[21] W.L. Hoo, C.S. Chan, Zero-shot object recognition system based on topic model, IEEE Trans. Hum.-Mach. Syst. 41(5), 2015, 518-525, http://dx.doi.org/10.1109/THMS.2014.2358649.

[22] Rung-Huei Liang, Ming Ouhyoung, A sign language recognition system using hidden Markov model and context sensitive search, in: Virtual Reality Software and Technology, 1996, pp. 59–66.

[23] M. Martínez-Camarena, O. Mogrovejo, J. Antonio, T. Tuytelaars, Towards sign language recognition based on body parts relations, in: International Conference on Image Processing, 2015, pp. 1–5.

[24] H.-D. Yang, S. Sclaroff, S.-W. Lee, Sign language spotting with a threshold model based on conditional random fields, IEEE Trans. Pattern Anal. Mach. Intell. 31 (7) (2009) 1264–1277, http://dx.doi.org/10.1109/TPAMI.2008.172.

[25] C.P. Vogler, American sign language recognition: reducing the complexity of the task with phoneme-based modeling and parallel hidden Markov models (Ph.D. thesis), Citeseer, 2003.

[26] E.-J. Ong, H. Cooper, N. Pugeault, R. Bowden, Sign language recognition using sequential pattern trees, in: IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 2200–2207. http://dx.doi.org/10.1109/CVPR.2012.6247928.

[27] E.-J. Ong, O. Koller, N. Pugeault, R. Bowden, Sign spotting using hierarchical sequential patterns with temporal intervals, in: IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1931–1938.

[28] S. Celebi, A. Aydin, T. Temiz, T. Arici, Gesture recognition using skeleton data with weighted dynamic time warping, in: VISAPP, 2013, pp. 620–625.

[29] S. Salvador, P. Chan, Toward accurate dynamic time warping in linear time and space, Intell. Data Anal. 11 (5) (2007) 561–580.

[30] J. Lichtenauer, E. Hendriks, M. Reinders, Sign language recognition by combining statistical DTW and independent classification, IEEE Trans. Pattern Anal. Mach. Intell. 30 (11) (2008) 2040–2046, http://dx.doi.org/10.1109/TPAMI.2008.123.

[31] S.G.M. Almeida, F.G. Guimarães, J.A. Ramírez, Feature extraction in brazilian sign language recognition based on phonological structure and using RGB-D sensors, Expert Syst. Appl. (2014) 7259–7271.

[32] D. Gale, L.S. Shapley, College admissions and the stability of marriage, Am. Math. Mon. (1962) 9–15.

[33] W.T. Freeman, M. Roth, Orientation histograms for hand gesture recognition, in: FG Workshop, vol. 12, 1995, pp. 296–301.

Hanjie Wang received the B.S. degree from University of Science and Technology Beijing (USTB), Beijing, China, in 2010. He is currently a Ph.D. candidate in Institute of Computing Technology, Chinese Academy of Science (CAS). His research interests include computer vision and pattern recognition, especially sign language recognition. He has published technical papers in the area of gesture recognition and sign language recognition.

Xiujuan Chai received the B.S., M.S., and Ph.D. degrees in computer science from the Harbin Institute of Technology, Harbin, China, in 2000, 2002, and 2007, respectively. She was a Post-doctorial researcher in Nokia Research Center (Beijing), from 2007 to 2009. She joined the Institute of Computing Technology, Chinese Academy Sciences, Beijing, in July 2009 and now she is an Associate Professor. Her research interests cover computer vision, pattern recognition, and multimodal human-computer interaction. She especially focuses on sign language recognition related research topics.

Xilin Chen received the B.S., M.S., and Ph.D. degrees in computer science from the Harbin Institute of Technology, Harbin, China, in 1988, 1991, and 1994, respectively, where he was a Professor from 1999 to 2005. He has been a Professor with the Institute of Computing Technology, Chinese Academy of Sciences (CAS), since 2004. He is also a FiDiPro Professor from 2012 to 2015 in Oulu University. He has authored one book and over 200 papers in refereed journals and proceedings in the areas of computer vision, pattern recognition, image processing, and multimodal interfaces. He is a fellow of the China Computer Federation.