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/

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