CHINESE SIGN LANGUAGE RECOGNITION WITH ADAPTIVE HMM

Jihai Zhang, Wengang Zhou, Chao Xie, Junfu Pu, and Houqiang Li

University of Science and Technology of China, Hefei, China
{jihzhang, pjh}@mail.ustc.edu.cn, {zhwg, chaoxie, lihq}@ustc.edu.cn

ABSTRACT

Sign Language Recognition (SLR) aims at translating the sign language into text or speech, so as to realize the communication between deaf-mute people and ordinary people. This paper proposes a framework based on the Hidden Markov Models (HMMs) benefited from the utilization of the trajectories and hand-shape features of the original sign videos, respectively. First, we propose a new trajectory feature (*enhanced shape context*), which can capture the spatio-temporal information well. Second, we fetch the hand regions by Kinect mapping functions and describe each frame by HOG (preprocessed by PCA). Moreover, in order to optimize predictions, rather than fixing the number of hidden states for each sign model, we independently determine it through the variation of the hand shapes. As for recognition, we propose a combination method to fuse the probabilities of trajectory and hand shape. At last, we evaluate our approach with our self-building Kinect-based dataset and the experiments demonstrate the effectiveness of our approach.

Index Terms— Sign language recognition, enhanced shape context, Hidden Markov Models, adaptive hidden states

1. INTRODUCTION

Sign language is the main communication method for the deaf-mute. Sign language is composed by trajectory of the hands, the shape of the hands, the posture of the skeletons, even the face expressions, and so on. Normal people cannot understand most of the signs without learning professional knowledge, which poses an obstacle to the communication between the deaf-mute and normal people. Therefore it is necessary to build a framework that can translate the sign language to text or speech language automatically.

Sign language recognition (SLR) is not as popularly studied as speech language recognition which is partially due to the fact that it pays much more efforts to collect related video data and it is hard to describe the handshapes, postures, and gestures, while tracking the hands and depicting the trajectories are also nontrivial. Early researchers achieve high accuracy rate by using data gloves [1]. The main advantage of the data gloves is that they can capture the finger joints information and hand trajectory accurately. With the accurate features, [1] obtains the accuracy of about 90% in a large vocabulary sign language dataset by using Hidden Markov Model method. But the expense and complexity of the equipment limit its popularity. Consequently, more and more researchers [2] [3] [4] focus on vision based SLR.

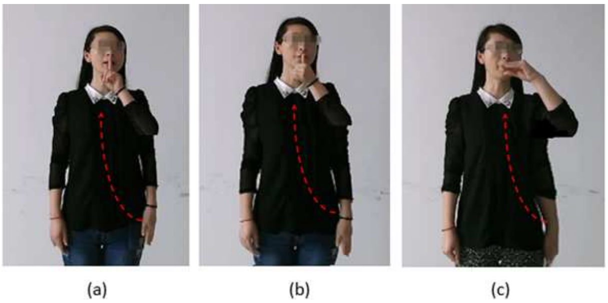

Fig. 1. Three sign words with similar trajectories following the red dash. (a) 'mother' using the forefinger. (b) 'father' using the thumb. (c) 'thin' closing the fingers like 'C'.

One cheap and helpful method is to use the color gloves. [5] uses color gloves to make segmentation and hand tracking easily in American Sign Language recognition. But both data gloves and color gloves are wearable equipment and it may make signers feel uncomfortable. Thanks to Microsoft Kinect, signers can feel free from the wearable equipment. It contributes to SLR vastly with its real time provision of RGB and depth data [6]. Several other researchers focus on skeleton feature and hand posture feature to realize more robust SLR [7] or gesture recognition [8]. In [7], a latent support vector machine method is proposed and it can obtain an accuracy rate of about 85.5% with 73 classes of the American Sign Language isolated signs. In [8], a new feature *Histogram of Oriented Displacement* is proposed and it performs well in gesture recognition. Although the skeleton-based features are robust in some cases, they cannot classify the signs when they share the same trajectory as shown in Fig. 1. To tackle the problem, researchers try more features like HOG [9] to represent the hand shape and it helps to improve the performance in SLR [2] [10].

We can easily find the action recognition dataset based on Kinect, but for sign language recognition, as far as we know, there is still lacking of a public, standard and large Kinect-

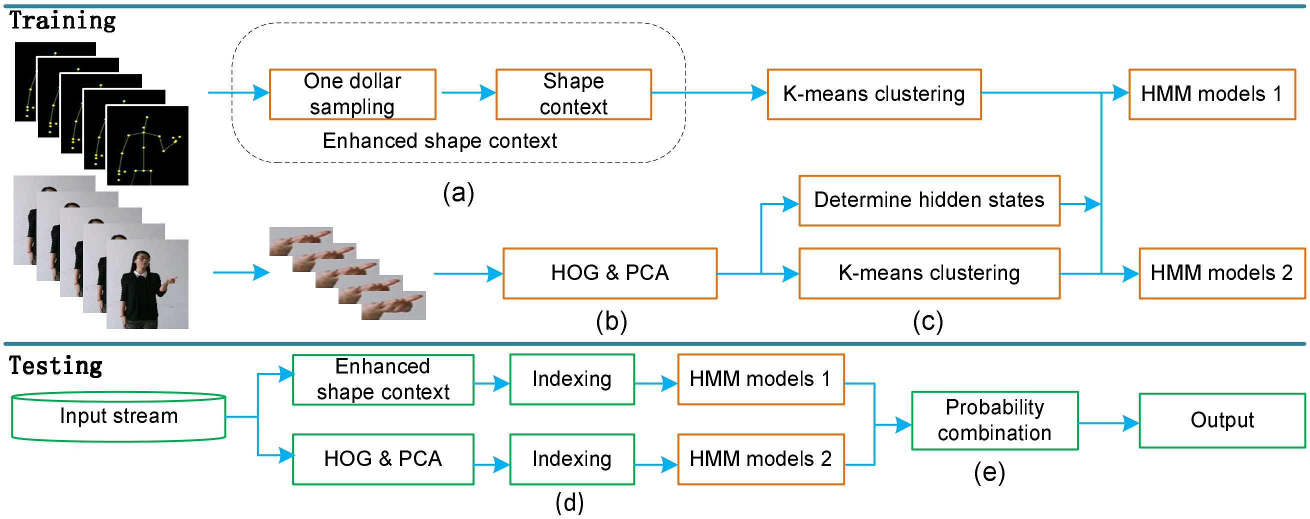

Fig. 2. Our SLR framework. (a) The proposed enhanced shape context feature (introduced in Section 3 in detail). (b) Fetching the hand surroundings by Kinect mapping function and reducing the dimension by PCA. (c) Clustering the HOG features by k-means and using the adaptive hidden states method in Section 3 to train the HMM models. (d) Indexing the features by using the centers obtained in (c). (e) Combining the probabilities using the method in Section 3 and finally, realizing the recognition.

based dataset. Some researchers only perform experiments on small vocabulary datasets. For example, [7] conducts experiments on 73 signs dataset. On 40 German sign dataset, Cooper *et al.* [11] use Hidden Markov Models (HMM) with sub-units to achieve an accuracy of 85.1%. Some researchers also study on large vocabulary datasets. For example, Chai *et al.* use trajectory matching method to realize an accuracy rate of 83.5% on 235 Chinese sign dataset, and they also conducts experiments on 370 Chinese sign dataset with the accuracy of about 90% by using Light-HMM [2]. Eng-Jon Ong *et al.* [12] reach an accuracy rate of 74.1% by using sequential pattern trees on 982 signs in the signer dependent test.

Witnessing the great improvement Hidden Markov Models has made in speech recognition, a lot of researchers use the models to model sign language words [13] [14] [15]. Considering the powerful modeling ability, we also focus on HMM method. Unlike other researchers, we propose a method to adaptively determine the hidden states of the HMM instead of fixing them as a specific value. As for features, we propose a new feature called enhanced shape context (eSC) to represent the spatial and temporal information of the trajectories. In addition, we use HOG feature to describe the hands in video and PCA to reduce the dimension. In recognition stage, we combine the output probabilities with trajectory and hand shape features as the final recognition probability. The framework is shown in Fig. 2.

Our main contributions of this work are summarized as follows.

- We propose a new feature eSC, which consists of one dollar gesture recognizer and shape context. The feature describes the shape of the trajectories well.
- Considering the characteristic of the sign language and benefited from the variation of the hand shape, we propose an HMM with adaptive hidden states to model the sign words instead of fixing the states.
- We propose a combination method to combine the probabilities of the trajectories and hand shapes, and our method performs better than baselines in our large Kinec-based dataset.

The remainder of this paper is organized as follows. Section 2 describes our framework. Section 3 introduces our features, the determination of the hidden states and the combination of the probabilities. Section 4 is the detailed experiments. And then we present our conclusion and future work.

2. FRAMEWORK

Fig. 2 shows both the training and testing procedures. We record the original RGB video data and skeletons information with Kinect 2.0. Fig. 2 (a) denotes the enhanced shape context (eSC) feature extraction, which is described in Section 3. Then we cluster the features with K-means to obtain adaptive hidden states in Section 3 for training the HMM models. For hand shape modeling, we extract the HOG feature and reduce the dimension by PCA, and then train the HMM models with adaptive hidden states. In the testing stage, we extract the eSC feature and HOG feature separately, and index them with the centers obtained in Fig. 2 (c). Then we utilize the combination method in Section 3 to obtain the final recognition result.

3. METHODS

This section contains feature description and our adaptive models.The Kinect sensor provides us 3-D coordinates of 25 skeletons including hands, elbows, head, shoulders, and so on. Besides, it also furnishes 720p RGB videos. With the original data, we propose an enhanced shape context (eSC) feature for trajectory and utilize HOG feature for hand shape representation.

3.1. Enhanced shape context

After obtaining the 3-D skeleton coordinates, we normalize them by the head points and shoulder width of the signer. To further explore the trajectory information, we divide the (x, y, z) coordination into (x, y), (x, z), and (y, z), and then concatenate the three coordination features. As shown in Fig. 3, our eSC feature is extracted in three steps. First, in order to avoid the influence caused by different speeds in SLR, we sample the points, which is benefited from the one dollar recognizer [16]. Then, we extract shape context on each coordination.

3.1.1. Step 1: one dollar sampling

In SLR, even the same person may sign the same word differently with variant speeds. To address the problem, we utilize the one dollar sampling method [16]. The method can sample the path in terms of the point density. To re-sample the trajectory points, we first calculate the total length of the path, which has *M* points. Dividing this length by $N-1$ gives the length of each increment, *I*, between *N* new points. Then the path is stepped through such that when the covered distance exceeds *I*, a new point is added through linear interpolation. Here, *N* is the length when conduct one dollar sampling (we set *N* no larger than 250 in our experiments).

3.1.2. Step 2: shape context

Shape context [17] can describe the distribution of other points when given a reference point [18]. For a point *p*, shape context is defined as a histogram by voting the remaining $N-1$ points to the surrounding bins of *p*, where *N* is the number of the trajectory points. Using the points obtained in step 1, we extract shape context feature in 3 coordinations respectively. In each coordination, we separate the space into 36 bins as shown in Fig. 3 (b), where the plane is divided by 3 circles and 6 lines. Hence, we get a 216-D vector (12 directions x 3 circles x 3 coordinations x 2 hands) for each point.

3.2. HOG of hand shape

HOG feature is one of the most popular image description features. We adopt it to describe the hand shape, which is obtained from the original RGB video with the Kinect mapping function. We fetch the hand center surroundings with the picture within a patch of 70x70 pixels experimentally. We calculate HOG in each 10x10 pixel cell and each block with 2x2 cells. Finally, we obtain a 1296-D HOG feature (6 blocks x 6 blocks x 4 cells x 9 orientations) for each frame. Considering the difficulty of the high dimension processing, we use PCA to reduce the dimension to 50 experimentally. Consequently, each frame is described by a 100-D feature (two hands).

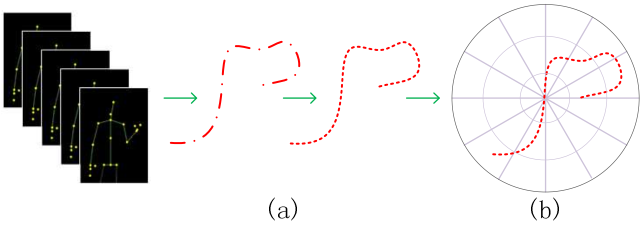

Fig. 3. Steps for eSC. (a) One dollar sampling to tackle the difference caused by different speeds. (b) Shape context extraction with 36 bins in the plane.

3.3. HMMs with adaptive hidden states

HMMs focus on three basis questions, *i.e.*, evaluation, estimation, and decoding problem. The evaluation problem is formulated as: given the observation sequence $\mathbf{O} = O_1, O_2, ..., O_T$, and the model $\lambda = (\pi, A, B)$ to calculate $P(\mathbf{O}|\lambda)$, where A is the states transfer probabilities matrix and B is the observation occurred probabilities matrix when given the states. The second problem refers to the estimation of the model given one or more observations. The decoding problem is: given observations to find the most likely sequence of hidden states.

Rather than using the fixed hidden states, we propose to adaptively determine the number of the hidden states. As shown in Fig. 4, the hand shapes vary from beginning to the end of the video. We obtain the HOG feature per frame as described in Section 3.2. And from the second frame, we calculate the Euclidean distance between the current frame and the former to get a new vector. We observe from the vector that some values are larger than most of others. And inspired by the fact that sign words are composed of some hand shapes, we set the threshold to segment the videos by the variation of the hand shape. This implies that a hidden state may correspond to a series of frames with similar hand shape. In Fig. 4, we set the hand segment threshold as 0.88 experimentally. If we define the corresponding image which is smaller than the threshold as label '1', and the exceeding as label '0'. The sequence is depicted to be the form like "11100011110000...". In the implementation, the segment number of successive '1' after media filter pre-processing is the hidden states of the sign. And the final hidden states for a specific sign word is the mode of the segments on training samples.

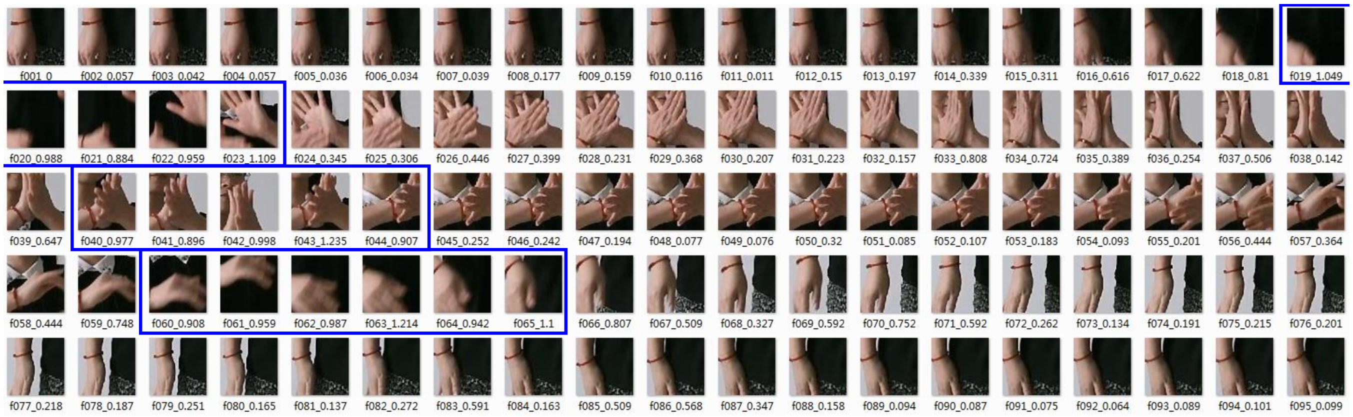

Fig. 4. An example for adapting hidden states in sign 'situation'. Each number under the hand image denotes the difference with the former frame. For example, 'f047_0.194' implies the difference between $47th$ frame and $46th$ frame is 0.194, which is smaller than threshold 0.88 we set experimentally. All the difference values in the blue boxes are larger than 0.88, while others are smaller. The 3 blue solid boxes divide the sequence into 4 segments, which implies that we can determine the hidden state as 4 when modeling the sign 'situation'.

3.4. Combination method

Our eSC feature contains one dollar sampling, which makes it impossible to fuse the trajectory and hand shape feature frame by frame, and the length of trajectory does not equal to that of the original videos. To tackle this problem, we fuse the probabilities of the two kinds of models as in Fig. 2 (e). Since the probability of the HMM decreases when the observation becomes longer, it is unreasonable to add the probabilities directly. To address this problem, we utilize the average probability per point by pre-processing the probability as:

$$V_i = \sqrt[l]{P}_i, i = 1, 2, ..., n \tag{1}$$

where l is the length of the trajectory, n is the number of sign words, P_i is the recognition probability under the ith model, and V_i is the average probability for trajectory under the ith model. In the same way, we can obtain the average probability per frame for hand shape with HOG features. At last, we add the two average probabilities as the final recognition probability and find the maximum. The corresponding model will be recognized as the result.

4. EXPERIMENTS

We build two Kinect-based Chinese sign language datasets, as shown in Table 1, by employing deaf-mute school teachers as the signers and conduct experiments on them by leave-one-out validation. First, we determine the optimal observation number in HMMs. Second, we evaluate the effectiveness of different features. Third, we compare the results between the adaptive hidden states and fixed states. At last, we compare our method with the classical methods including Dynamic Time Warping (DTW), traditional HMMs, and other work on the datasets.

4.1. Datasets

The Kinect-based dataset is collected by professional sign language signers. We record the data at the rate of $30fps$ (30 frames per second). The distance between the signers and the Kinect is about 1.5 meters. We build two datasets. The Dataset I contains 100 sign words by one signer with 5 repetitions and 500 videos in total. To show the recognition performance on large vocabulary dataset, we build dataset II, which contains 500 sign words by one signer with 5 repetitions and 2500 videos in total.

4.2. Evaluation on observation numbers

Unlike hidden states which can be variant in different models, observation number should be consistent. In this part, we evaluate the different observation in eSC feature and HOG feature with fixed hidden states HMMs, respectively. As shown in Fig. 5, the blue polyline describes the accuracy rate of eSC feature, and the orange polyline for the HOG feature. We can find that when the observation number is set to 1900 and 1800 for eSC and HOG, respectively, the performances are 0.750 and 0.694, which are better than other observation numbers.

4.3. Evaluation on different features

Sign language can be represented by different features with different discriminative abilities. For example, trajectory feature can describe the dynamic gesture words and the hand shape can describe the static pose. Furthermore, some words

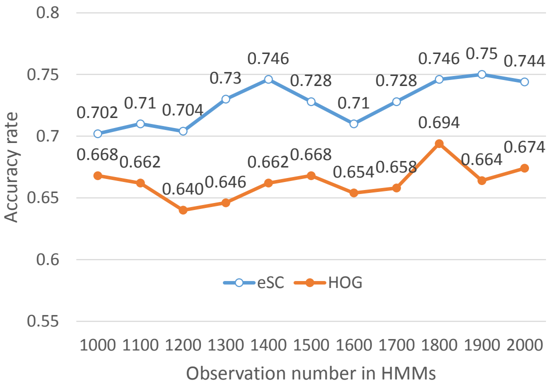

Fig. 5. Evaluation on observation numbers. The orange polyline above describes the accuracy rate of the eSC feature and the blue polyline below describes the accuracy rate of HOG feature.

Table 1. Results of features on Dataset II

Methods	Features	Top1	Top5	Top10
HMMs with fixed states	SC	0.628	0.848	0.902
	eSC	0.702	0.880	0.940
	HOG	0.694	0.886	0.930
	SC+HOG	0.706	0.844	0.970
	eSC+HOG	**0.838**	**0.960**	**0.976**

can only be determined by taking into account both trajectory and hand shape. This section shows the experimental results by evaluating different features for feature selection. We compare the results by using five kinds of features, including normal skeletons coordinates (SC), enhanced shape context (eSC), HOG, SC+HOG and eSC+HOG. The evaluation methods here all use conventional HMM with fixed hidden states 6.

Table 1 shows the recognition results with the five different features. The SC feature contains four skeletons coordinates. The eSC feature is our enhanced shape context feature introduced in Section 3. Top$n$ means the result is accurate in top$n$. We can see from the table that our eSC feature is superior to the normal SC feature with about 8% improvement. Both SC and eSC are dynamic features. As for appearance feature, we calculate HOG feature of the hands and then use PCA to reduce the dimension to 100 per frame (50 each hand). The feature combined by SC and HOG gets higher performance than the former three features, and the feature eSC+HOG performs better than any other features. The comparison demonstrates that our feature has more competitiveness.

Table 2. Results of states with HOG on Dataset II

States	Top1	Top5	Top10	Time(s)
3	0.664	0.872	**0.930**	**0.167**
4	0.678	0.882	0.926	0.211
5	0.676	0.876	0.928	0.261
6	0.694	0.886	**0.930**	0.301
Adaptive states	**0.708**	**0.888**	**0.930**	0.256

Table 3. Results of methods on Dataset I & II

Dataset	method	Top1	Top5	Top10
Dataset I	DTW	0.710	0.800	0.850
100 signs	Lin *et al.* [19]	0.840	0.950	0.970
	HMMs (SC+HOG)	0.760	0.940	0.960
	HMMs (**eSC**+HOG)	0.880	0.970	0.990
	Ours (eSC+HOG)	**0.920**	**0.990**	**1**
Dataset II	DTW	0.666	0.784	0.832
500 signs	Lin *et al.* [19]	0.698	0.904	0.948
	HMMs (SC+HOG)	0.706	0.844	0.970
	HMMs (**eSC**+HOG)	0.838	0.960	0.976
	Ours (eSC+HOG)	**0.860**	**0.968**	**0.988**

4.4. Experiments on variant hidden states & fixed states

Hidden state is important for HMMs, and in SLR, we regard the states as the sub-units of the sign words.Unfortunately, there is no exact sub-unit number and we cannot define or pick out them one by one. For different words, the number of sub-units are likely to be different, and it is unreasonable to fix the same hidden state for all words. All the above inspire us to find a method to automatically determine the hidden states. As show in Section 3, we adapt the states through the change of the hand shape.Table 2 shows the results with fixed and adaptive states. We find that the result of adaptive states is superior to any other fixed states in top1. And the time cost is a trade off among them.

4.5. Experiments on methods

In this part, we compare our proposed method with classical method DTW, HMM, and method in [19] on dataset I and dataset II, respectively. The comparison recognition rates are shown in Table 3.

From the comparison on dataset I in Table 3, we can see that our method can get the accuracy rate of 0.920 in top 1, 0.990 in top 5, and perfect 1 in top 10. In dataset II our results are 0.860 in top 1, 0.968 in top 5, and 0.988 in top 10. In both datasets our results are superior to the others. The performance rate decreases compared with dataset I because while the vocabulary becomes larger, words behave similarly and it increase the difficulty to recognize them accurately.

5. CONCLUSIONS AND FUTURE WORKS

In this paper we propose a sign language recognition framework based on Kinect. The framework contains feature extraction, modeling, and recognition. In feature extraction stage, we propose an enhanced shape context feature, which captures the spatial and temporal information well. As for appearance feature, HOG feature with PCA is used. In modeling stage, rather than using fixed hidden states in HMMs, we proposed an method to obtain the adaptive states inspired by the variation of the hand shapes. We conduct a series of experiments to prove that our eSC feature is better than SC, and our adaptive-hidden-states method is better than the baseline methods. In our future work, we will conduct experiments on larger datasets and explore the deep fusion among skeleton feature, RGB video, data and depth map data to improve the performance.

6. ACKNOWLEDGEMENT

This work is supported in part to Prof. Houqiang Li by the 973 Program under Contract 2015CB351803 and the National Natural Science Foundation of China (NSFC) under Contract 61390514 and Contract 61325009, and in part to Dr. Wengang Zhou by NSFC under Contract 61472378, the Natural Science Foundation of Anhui Province under Contract 1508085MF109, and the Fundamental Research Funds for the Central Universities.

7. REFERENCES

[1] Gaolin Fang, Wen Gao, and Debin Zhao, "Large-vocabulary continuous sign language recognition based on transition-movement models," *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, vol. 37, no. 1, pp. 1–9, 2007.

[2] Hanjie Wang, Xiujuan Chai, Yu Zhou, and Xilin Chen, "Fast sign language recognition benefited from low rank approximation," in *FG*, 2015.

[3] Jie Huang, Wengang Zhou, Houqiang Li, and Weiping Li, "Sign language recognition using real-sense," in *ChinaSIP*, 2015.

[4] Ao Tang, Ke Lu, Yufei Wang, Jie Huang, and Houqiang Li, "A real-time hand posture recognition system using deep neural networks," *ACM Transactions on Intelligent Systems and Technology*, vol. 6, no. 2, pp. 21, 2015.

[5] Z Zafrulla, H Brashear, H Hamilton, and T Starner, "Towards an american sign langauge verifier for educational game for deaf children," in *ICPR*, 2010.

[6] Jamie Shotton, Toby Sharp, Alex Kipman, Andrew Fitzgibbon, Mark Finocchio, Andrew Blake, Mat Cook, and Richard Moore, "Real-time human pose recognition in parts from single depth images," *Communications of the ACM*, vol. 56, no. 1, pp. 116–124, 2013.

[7] Chao Sun, Tianzhu Zhang, Bing-Kun Bao, and Changsheng Xu, "Latent support vector machine for sign language recognition with Kinect.," in *ICIP*, 2013.

[8] Mohammad A Gowayyed, Marwan Torki, Mohamed E Hussein, and Motaz El-Saban, "Histogram of oriented displacements (HOD): describing trajectories of human joints for action recognition," in *IJCAI*, 2013.

[9] Navneet Dalal and Bill Triggs, "Histograms of oriented gradients for human detection," in *CVPR*, 2005.

[10] Jie Huang, Wengang Zhou, Houqiang Li, and Weiping Li, "Sign language recognition using 3d convolutional neural networks," in *ICME*, 2015.

[11] Helen Cooper, Eng-Jon Ong, Nicolas Pugeault, and Richard Bowden, "Sign language recognition using subunits," *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 2205–2231, 2012.

[12] Eng-Jon Ong, Helen Cooper, Nicolas Pugeault, and Richard Bowden, "Sign language recognition using sequential pattern trees," in *CVPR*, 2012.

[13] Jihai Zhang, Wengang Zhou, and Houqiang Li, "A new system for chinese sign language recognition," in *ChinaSIP*, 2015.

[14] Jihai Zhang, Wengang Zhou, and Houqiang Li, "A threshold-based hmm-dtw approach for continuous sign language recognition," in *Proceedings of International Conference on Internet Multimedia Computing and Service*, 2014.

[15] Junfu Pu, Wengang Zhou, Jihai Zhang, and Houqiang Li, "Sign language recognition based on trajectory modeling with hmms," in *MultiMedia Modeling*, 2016.

[16] Jacob O Wobbrock, Andrew D Wilson, and Yang Li, "Gestures without libraries, toolkits or training: A $1 recognizer for user interface prototypes," in *UIST*, 2007.

[17] Jitendra Malik and J Puzicha, "Shape matching and object recognition using shape contexts," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 4, pp. 509–522, 2002.

[18] Wengang Zhou, Yijuan Lu, Houqiang Li, Yibing Song, and Qi Tian, "Spatial coding for large scale partial-duplicate web image search," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010.

[19] Yushun Lin, Xiujuan Chai, Yu Zhou, and Xilin Chen, "Curve matching from the view of manifold for sign language recognition," in *ACCV*, 2014.