Fast Sign Language Recognition Benefited From Low Rank Approximation

Hanjie Wang[1], Xiujuan Chai[1], Yu Zhou[2] and Xilin Chen[1]

[1] Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing, 100190, China

[2] Institute of Information Engineering, Chinese Academy of Sciences No. 89, Minzhuang Road, Haidian District, Beijing, 100093, China

Abstract— **This paper proposes a framework based on the Hidden Markov Models (HMMs) benefited from the low rank approximation of the original sign videos for two aspects. First, under the observations that most visual information of a sign sequence typically concentrates on limited key frames, we apply an online low rank approximation of sign videos for the first time to select the key frames. Second, rather than fixing the number of hidden states for large vocabulary of variant signs, we further take the advantage of the low rank approximation to independently determine it for each sign to optimise predictions. With the key frame selection and the variant number of hidden states determination, an advanced framework based on HMMs for Sign Language Recognition (SLR) is proposed, which is denoted as Light-HMMs (because of the fewer frames and proper estimated hidden states). With the Kinect sensor, RGB-D data is fully investigated for the feature representation. In each frame, we adopt Skeleton Pair feature to character the motion and extract the Histograms of Oriented Gradients as the feature of the hand posture appearance. The proposed framework achieves an efficient computing and even better correct rate in classification. The widely experiments are conducted on large vocabulary sign datasets with up to 1000 classes of signs and the encouraging results are obtained.**

I. INTRODUCTION

Sign Language (SL) is very important to exchange information in communication within deaf community as well as between deaf and hearing societies. Hence, Sign Language Recognition (SLR) has great potential applications in the fields such as SL translation, tutor and education [1] [2] [3].

Early researchers achieved great successes when using data gloves [4][5]. One typical work was from Gao et al. [5], which achieved a good performance of 90.8% over more than 5000 signs in signer-dependent test. However, the expensive price and wearable character made the SLR system difficult to popularization. Gradually, pure vision based SLR attracted the researchers' attentions, but it was difficult for precise hand tracking and segmentation. Zafrulla et al. [6] tackled the problem by using colored gloves, where a pair of accelerometers were strapped on, to make segmentation and hands tracking easily. Microsoft Kinect, with its real time provision of RGB and depth data [7], contributes to SLR vastly. Selebi et al. [8] used skeleton data for gesture recognition with weighted Dynamic Time Warping method. Several other researchers integrated skeleton feature and hand posture feature to realize more robust SLR[3][9]. In [9], a discriminative exemplar coding method was proposed and obtained about 85.5% recognition rate with 73 classes of the American Sign Language (ASL) signs.

Although there are published datasets on *body actions* captured by Kinect sensor [10] [11], for *sign language*, there is still lacking of available large vocabulary datasets, especially those captured by Kinect sensor to the best of our knowledge. Some experiments were conducted on small datasets. For example, on a datasets with vocabulary of twelve with depth cue, Kurakin et al. [12] proposed a real-time system for dynamic hand gesture recognition on ASL. The other experiments were conducted on large datasets and yet without depth cue. For example, Eng-Jon Ong et al. [13] created a discriminative, multi-class classifier based on sequential pattern trees recently. Their experiment showed 74.1% correct rate on 982 signs in singer-dependent test, which can be taken as the state-of-the-art in SLR research. In the work of Wang et al. [14], in user-independent experiments, the correct sign in the top 10 was 78% with a system vocabulary of 1,113 signs. Chai et al. [15] recently proposed a method and developed a system to recognize signs using trajectory feature with depth cues. To further promote the performance of SLR by the assistant of the depth cue, we collected three datasets using Kinect sensor for research usage. Two of them contains up to 1000 classes of signs.

Inspired by the improvement of speech recognition, many researchers used generative models such as Hidden Markov Models (HMMs) to model a large size of vocabulary [3][5][16][17]. We also recur to HMMs to tackle the SLR problems by considering the powerful modeling ability. However, confronted with the high dimensional visual features, the traditional dense frame based HMMs consume lots of time. Motivated by the observation that most discriminative information concentrates on the features extracted from few key frames, the sign videos can potentially be represented more compact. If the key frames selection procedure can be embedded into the model constructing online, the traditional HMMs are able to be optimized for SLR. Thus, we propose an online low rank approximation to obtain key frames from a sign video. The low rank approximation is realized by minimizing the Residual Sum of Squares (RSS) of the previously selected features before current time. The corresponding parameters will be updated online as well. What's more, it can also adaptively determine

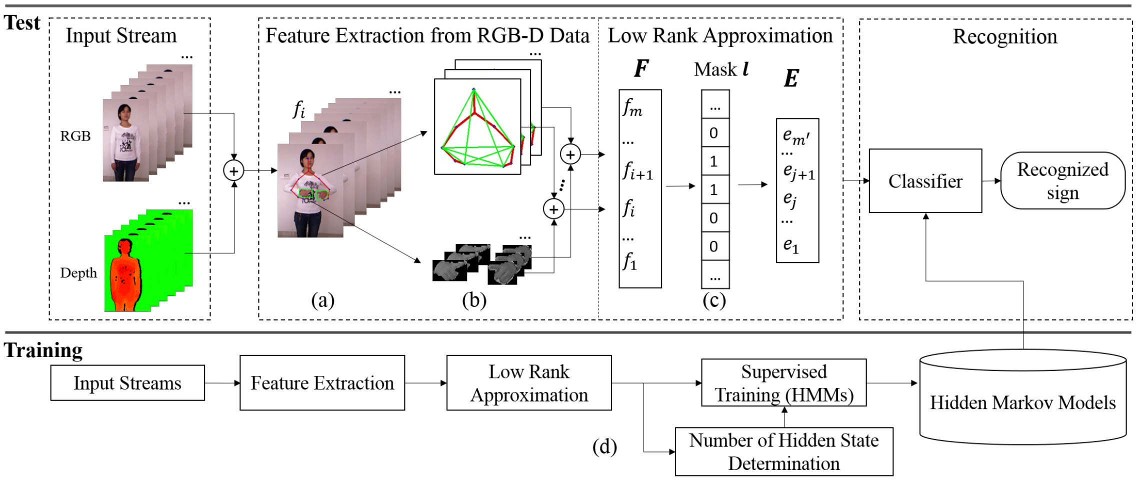


Fig. 1. The framework of Light-HMMs for SLR.(a) Face, hand detection and skeleton detection. (b) Skeleton and hand posture extraction. (c) Selected frames mask using low rank approximation approach. (d) The variant hidden states are determined by low rank approximation in training stage.

the proper number of the HMMs' hidden states for each sign independently. Hence, it is not necessary to fix an arbitrary hidden state number for all the signs. Based on the traditional HMMs and aiming to push the SLR technique to be practice, this paper selects the key frames through low rank approximating and adaptively determine the number of hidden state in HMMs. We denote such HMMs, with fewer frames and proper estimated numbers of hidden states, as the Light-HMMs in this paper. The signs are characterized by combining both the appearance and skeleton pair feature. Figure 1 shows the framework of our algorithm.

Our main contributions are summarized as follows. First, benefited from the low rank approximation, time cost of the HMMs can be efficiently reduced by key frame selection while maintaining or even promoting the recognition accuracy. Second, the number of hidden state for sign models are also automatically determined by the low rank approximation approach and further speed up the recognition. Third, a discriminative and compact feature representation is generated by fusing the posture appearance and skeleton pair feature from RGB-D data captured by Kinect sensor. In addition, datasets used in this paper are subsets of a large SL database, which has been partial released recently ("*http://vipl.ict.ac.cn/homepage/KSL/home.html*"). The remaining part of this paper is organized as follows. Section II shows the basic framework and brief formulations. Section III is the detailed implementations, including the strategies of RGB-D feature, key frame selection and the number of hidden states determination by low rank approximation. Section IV gives the experimental results and analysis. Section V is our conclusion and future work.

II. FRAMEWORK AND FORMULATION

Figure 1 illustrates both the training and test procedures based on the Light-HMMs benefited from low rank approximation from two aspects. Figure 1 (c) shows that low rank approximation removes the abundant frames and selects the key frames for HMMs training and test. Figure 1 (d) shows that instead of using a fixed hidden state number for all the sign class models, low rank approximation adaptively determine the number for each sign model independently.

Since a sign sequence typically concentrates on features from several key frames, our target is to select the most discriminative frames. The key frames selection should be an online procedure that can be embedded in the HMMs framework. Denote the feature matrix from all the frames in the sequence as $F = [\boldsymbol{f}_1, \boldsymbol{f}_2, ..., \boldsymbol{f}_K]$, where $\boldsymbol{f}_i$ is the feature vector of a frame and K is the total frame number. We propose an online approach that applying the criterion of RSS to select a compact feature matrix E, which is the subset of F.

To model the signs, a typical HMMs are represented as $\lambda = (A, B, \pi)$, where A is the matrix of state transition probability, π is the initial states and $B = b_i(O)$ is the likelihood, which is evaluated by Gaussian Mixture Model as shown in Eq. 1

$$b_i\left(O\right) = \sum_{j=1}^{M} c_{ij} G\left[\mu_{ij}, \mathcal{E}_{ij}, O\right], 1 \leq i \leq N, \qquad (1)$$

where N is the state number, M is the number of mixture components and c_{ij} is the mixture weight. Along with the generation of E by low rank approximation, the corresponding mask $\boldsymbol{l}$ is recorded. The mask $\boldsymbol{l}$ indicates whether the frames in F are selected or not by label 1 or 0. Since the temporal region sharing the same labels belongs to one segmentation, $\boldsymbol{l}$ can also be acted as indicator to segment the sign video. Given the probe sign, the most likely class is found by straightforward Viterbi algorithm in the test stage.

III. IMPLEMENTATION

This section gives a detailed description of feature extraction on both hand postures and motion trajectories. More importantly, a key frame selection method by low rank approximation is proposed. The novel Light-HMMs

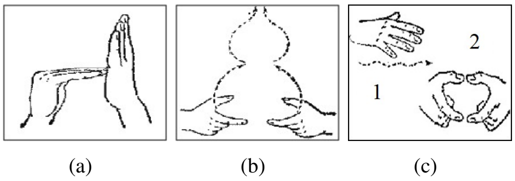


Fig. 2. (a) Posture dominated signs. (b) Trajectory dominated signs (c) Together represented signs.

also features on their adaptive numbers of hidden states, which is automatically determined by the number of signs' segmentations.

A. Feature Extraction from RGB-D Data

Intrinsically, sign is multi-modalities activity containing at least posture appearance and hand movement. The depth information can be used to accurately segment the human hands and the 3D movement is able to explore the skeleton movements along Z coordinate rather than merely on the $X-Y$ plane. The Chinese sign language vocabularies can be classified into three categories, which are posture dominant signs, trajectory dominant signs and together represented signs as shown in Figure 2. Therefore, we combine the hand posture appearance feature and the skeleton feature together to form a powerful feature representation. The column vectors of F are the combined feature as in Eq. 2.

$$\boldsymbol{f}_i = [\boldsymbol{f}_p, \boldsymbol{f}_s]_i, i = 1, 2, \ldots, k, \tag{2}$$

where $\boldsymbol{f}_p$ is the feature of hand posture and $\boldsymbol{f}_s$ is the skeleton motion feature. k is the total number of all the selected key frames. The details of generating $\boldsymbol{f}_p$ and $\boldsymbol{f}_s$ are introduced below.

1) Hand posture feature: HOG feature [18] is extracted from hand region, which is segmented by using self-adaptive skin model and depth constraint. The self-adaptive skin model is initialized by the skin of human face and updated by the skin of the detected human hands in previous frames. The details are omitted due to the limited length of the paper. Since the dimension of the original HOG is too high, Principal Component Analysis is applied for dimensionality reduction and to retain only most salient dimensions. Therefore, the hand posture feature $\boldsymbol{f}_p$ is obtained with the reduced dimension d.

2) Skeleton Pair feature: For the skeleton feature, a direct way is to use the normalized (x, y, z) coordinates. To better extract the intrinsic motion feature, we consider the pairwise relative position features proposed by Wang et al. [11]. Since the signs are essentially upper-body activities, we only select 5 important skeleton points including head, left elbow, right elbow, left hand and right hand. The feature has a dimension of $C_5^2 = 10$ and is invariant to rotation, scaling and translation. Figure 1(b) gives the illustration of the skeleton pairs connected by green lines.

B. Frame Selection Strategy In Light-HMMs

Dense frame based HMMs is practical and reasonable when using data glove due to the accurate measurement and small size of feature dimensions. However, there are many drawbacks when HMMs is applied to visual based SLR. First, the high dimensionality of the visual feature costs much computing time in the traditional frame-based way. Second, since the measurement accuracy of visual based SLR is limited, the features vary even within class. One method to solve the problem is to select the key frames, which have discriminative features. Therefore, not only frames down sampling but also key frames selection are desired for HMMs. The basic idea is that the frames with linearly independent features will be selected for the HMMs' training and test while the others will be discarded. Such online key frame selection procedure is proposed for the first time for SLR and is proved to be efficient.

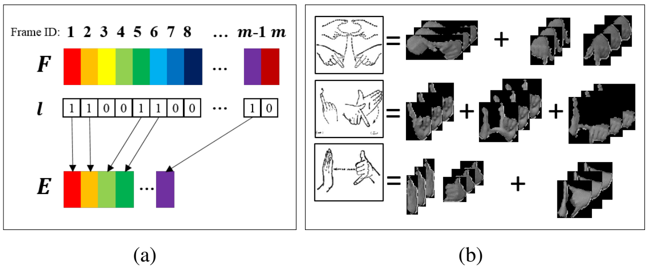


Fig. 3. Low rank approximation strategy for SLR. (a) Concise illustration of relationship among F, $\boldsymbol{l}$ and E. Colorful blocks represent features from frames. (b) Three examples of the frames selection. Segmentations are separated by symbol "+".

In the online implementation, if the feature of current frame can be linearly represented by previous selected frames with less residual, current frame will be removed from the video of the sign. While F is the matrix consists of features from all the frames in a sign video, we define E as the incrementally expanded matrix consisting of all the features of selected frames, and define $\boldsymbol{f_c}$ as the feature of current frame. The goal is to minimize the RSS ϵ of current frame feature by $\epsilon = (\boldsymbol{f_c} - E\boldsymbol{\beta})^T(\boldsymbol{f_c} - E\boldsymbol{\beta})$, where the coefficient $\boldsymbol{\beta}$ is searched to minimize the RSS. A unique solution to $\boldsymbol{\beta}$ can be given by

$$\boldsymbol{\beta} = (E^T E)^{-1} E^T \boldsymbol{f_c} = M E^T \boldsymbol{f_c}, \tag{3}$$

where $M = (E^T E)^{-1}$ is denoted as the core matrix. Therefore, the RSS can be computed as follows

$$\begin{aligned} \epsilon &= (\boldsymbol{f_c} - E\boldsymbol{\beta})^T(\boldsymbol{f_c} - E\boldsymbol{\beta}) \\ &= \left\| \boldsymbol{f_c} - E(M E^T \boldsymbol{f_c}) \right\|^2 . \end{aligned} \tag{4}$$

If ϵ is smaller than a threshold ϵ_0, then current frame $\boldsymbol{f_c}$ will be abandoned. Otherwise, $\boldsymbol{f_c}$ will be selected and used to update the E, as well as the other corresponding parameters.

Algorithm 1 shows more details on key frame selection with our low rank approximation. We can see that it is an online procedure with only two parameters (i.e., the parameter N and threshold ϵ_0 in the initial step). The matrix E and the vector $\boldsymbol{l}$ are incrementally expanded in Algorithm 1. The core matrix is updated by

$$M = \begin{bmatrix} M + \boldsymbol{\beta}^T\boldsymbol{\beta}/\epsilon & -\boldsymbol{\beta}/\epsilon \\ -\boldsymbol{\beta}^T/\epsilon & 1/\epsilon \end{bmatrix} \tag{5}$$

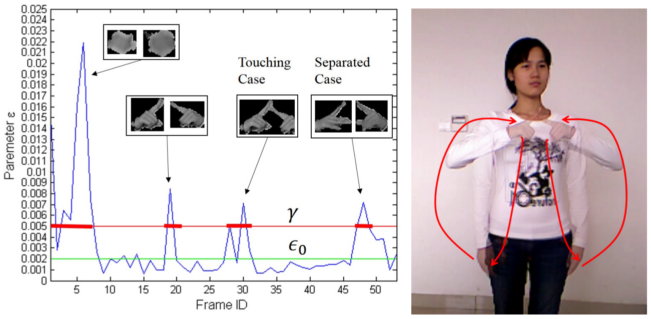
```
Input: Feature matrix $F$ of a sign video.
Output: Low rank matrix $E$ and the corresponding
        mask $\boldsymbol{l}$.
Given the matrix $F$ consists of features vectors;
Down sample the $F$ to $\hat{F}$ every $N$ frames, and the final
frame number is $m$, i.e., $\hat{F} = [\boldsymbol{f}_1, \boldsymbol{f}_2, ..., \boldsymbol{f}_m]$ ;
Define the threshold $\epsilon_0$;
Initial $E = \boldsymbol{f}_1$, $M = (E^T E)^{-1} = 1/(\boldsymbol{f}_1^T \boldsymbol{f}_1)$;
Initial the mask $\boldsymbol{l} = [\,]$;
for $k = 2 : m$ do
    Choose the current frame, i.e., the $k_{th}$ frame and it
    is denoted as $\boldsymbol{f_c} = \boldsymbol{f}_k$;
    Compute the coefficient $\boldsymbol{\beta} = ME^T \boldsymbol{f_c}$;
    Compute the RSS
    $\epsilon = (\boldsymbol{f_c} - E\boldsymbol{\beta})^T (\boldsymbol{f_c} - E\boldsymbol{\beta}) = \left\| \boldsymbol{f_c} - EME^T \boldsymbol{f_c} \right\|^2$;
    if $\epsilon > \epsilon_0$ then
        Add column $\boldsymbol{f_c}$ to $E$ as $E = [E, \boldsymbol{f_c}]$;
        Update the core matrix
        $M = \begin{bmatrix} M + \boldsymbol{\beta}^T \boldsymbol{\beta}/\epsilon & -\boldsymbol{\beta}/\epsilon \\ -\boldsymbol{\beta}^T/\epsilon & 1/\epsilon \end{bmatrix}$;
        Update the mask $\boldsymbol{l} = [\boldsymbol{l}, 1]$;
    else
        Update the mask $\boldsymbol{l} = [\boldsymbol{l}, 0]$;
        Continue;
    end
end
```

Algorithm 1: The procedure of low rank approximation.

After all the frames being processed, a mask $\boldsymbol{l}$ is determined. Figure 3 (a) is the concise illustration of the relationship among F, $\boldsymbol{l}$ and E. Figure 3 (b) shows the results of some example signs. Figure 4 gives a real example of sign "PEOPLE".

C. Number Of Hidden States In Light-HMMs

The number of hidden states should be determined in the training stage. Traditional HMMs commonly have fixed number of hidden states for all the classes despite their variances. However, the actions and durations of signs different from each other vastly. For example, there are approximate 89 frames for the sign "CHAIR" with single key action (Figure 2 (a)) while 150 frames for the sign "FRUIT" with two key actions (Figure 2 (c)). Obviously, it is unwise to share a fixed number of hidden states for the two signs. If the fixed number is too small, the models have weak description ability for those complex signs. On the contrary, with larger fixed number of hidden states, the time cost increases, which is actually unnecessary for those simple signs. To tackle this problem, the hidden states are set to be variant according to the "segmentation" result of low rank approximation. Frames with independent features are labeled with "1" ("0" otherwise) in the mask $\boldsymbol{l}$. Frames sharing the same label are regarded as one segmentation. Hence, we segment the sign video according to the label of "1" in the mask $\boldsymbol{l}$. For example, "00110000111000" indicates that two hidden states HMMs are suitable for the signs since there are two temporal regions of label "1". In the implementation, the mask will be smoothed by a median filter to remove outliers. The number of segmentations is determined by a threshold γ after the frames selecting procedure. See from Figure 4 (a), with the threshold γ, there are fewer selected frames. Thus, the hidden state number can be determined according to the segmentation number (there are four hidden states in Figure 4 (a) and are emphasized as red bold lines).



(a) (b)

Fig. 4. Low rank approximation for the sign PEOPLE and the HMMs hidden state number determination. (a) The trend of ϵ, where $\epsilon_0 = 0.002$ (the green line) and $\gamma = 0.005$ (the red line). (b) The illustration of the corresponding Chinese sign PEOPLE.(Best view in color.)

In the test stage, since both the likelihood and transition probabilities are small than 1, the scores turn to be smaller for classes with larger number of hidden states in the model. Therefore, in our implementation, for longer states number $S > 3$, each translation probability will be multiplied by a weight $S/(3 \times 1.3)$. We set the weight experimentally.

IV. EXPERIMENTS

We conduct several SLR experiments on three datasets. Firstly, the experiment on different features is conducted to test their discriminative abilities. Then the performances in efficiency and accuracy are evaluated with our proposed Light-HMMs, the traditional HMMs, Fast-DTW method [19] and the method proposed by Chai et al. [15].

A. Datasets and Settings

We have collected three datasets by using Microsoft Kinect sensor to evaluate our algorithm. The signers are all deaf students. The distance between Kinect sensor and signer is $1.5 \sim 2$ meters. The Dataset I contains 370 daily signs of Chinese sign language performed by 1 signer with 5 repetitions. To evaluate the recognition performance on large vocabulary dataset, we built another challenging Dataset II, which is composed of 1000 signs from Chinese sign language performed by 1 signer with 3 repetitions. Further to conduct the signer-independent SLR, we collected Dataset III, which has the vocabulary size of 1000 and the data is signed by 7 signers. Table I shows the details of our datasets.

In the next subsections, the effectiveness of low rank approximation and adaptive hidden states will be verified to prove that the Light-HMMs is more suitable for SLR. The leave-one-out cross-validation strategy is adopted to evaluate the performance and the average recognition rate will be

TABLE I

THE DETAILS OF OUR COLLECTED DATASETS.

Datasets	Vocabulary	Signer	Repetition	Total videos
I	370	1 (female)	5	1850
II	1000	1 (male)	3	3000
III	1000	7 (both)	1	7000

TABLE II

COMPARISON OF THE CS, SP, HOG, SP+HOG FEATURES.

Feature	Dimension	Top 1	Top 5	Top 10
CS	15	0.452	0.696	0.771
SP	10	0.625	0.820	0.875
HOG	51	0.710	0.856	0.896
SP+HOG	61	**0.842**	**0.946**	**0.965**

given. The baseline HMMs, Fast-DTW [19] and the method proposed by Chai et al. [15] are evaluated as comparisons. The evenly sample parameter N in Algorithm 1 is fixed to 1. The parameter ϵ of low rank approximation has two values for key frame selection ($\epsilon_0 = 0.001, 0.002$) and the number of hidden state determination ($\gamma = 0.0025, 0.005$) respectively. For a better comparison, the hidden state number of the HMMs is fixed to 3 for the evaluations of key frame selection strategy and will be changed for testing our variant hidden states.

B. Evaluation on Different Features

In SL representation, different features reveal different aspects of signs. For example, hand trajectories describe the dynamic motion and hand postures describe the static appearance. Obviously, they have different discriminative abilities. This section shows the experimental results on the evaluation of different features. The experiments are conducted on the Dataset II. Under the proposed Light-HMMs framework, we compare the results using 4 different features, which are directly skeleton coordinates feature (SC), skeleton pair feature (SP), HOG feature (HOG) and the fused SP+HOG feature.

Table II shows the comparison of the recognition results with the 4 different features. For the two dynamic features (SP and SC), it can be seen that the SP feature is superior to the directly SC feature with about 17 percentage points improvement. The appearance feature (HOG) performs the best when compared with the former two dynamic features in our experiments. A great promotion is achieved when the two features (HOG and SP) are combined since the two features complement each other as illustrated in Figure 2 (c). With depth cue, the SLR can reach 84.2% correct rate on 1000 vocabularies dataset using Light-HMMs. Considering for the good performance, in our following experiment, we fix the feature to be SP+HOG.

C. Experiment on Signer Dependent Test

In this part, we will compare the proposed Light-HMMs framework with the baseline HMMs which uses the dense frame based features. The experiments are conducted on two datasets, which are Dataset I and Dataset II respectively. The performance comparisons are all given in Table III, which shows both recognition rate and the time cost. The item "Time" denotes the processing time for each sign.

TABLE III

PERFORMANCE COMPARISON BETWEEN DIFFERENT METHODS.

Dataset	Framework	Top 1	Top 5	Top 10	Time
Dataset I 370 signs	Fast-DTW	0.885	0.965	0.978	110.4 ms
	Chai et al. [15](Trajectory)	0.810	0.919	0.946	11.2 ms
	HMMs	0.922	0.991	0.997	110 ms
	Light-HMMs($\epsilon_0 = 0.001$)	**0.940**	0.993	0.999	42 ms
	Light-HMMs($\epsilon_0 = 0.002$)	0.931	0.995	0.998	**37** ms
Dataset II 1000 signs	Fast-DTW	0.758	0.887	0.920	1168 ms
	Chai et al. [15](Trajectory)	0.591	0.77	0.819	34 ms
	Baseline	0.832	0.942	0.964	265 ms
	Light-HMMs($\epsilon_0 = 0.001$)	**0.842**	0.946	0.965	109 ms
	Light-HMMs($\epsilon_0 = 0.002$)	0.830	0.942	0.961	**94** ms

TABLE IV

EVALUATION OF ϵ WITH VARIANT VALUES.

Values	Accuracy	Time cost (ms/sign)
0.0001	0.943	87
0.0005	0.940	71
0.001	**0.940**	**42**
0.002	0.931	37
0.003	0.917	36
0.004	0.913	25

From this table, we can see that the proposed Light-HMMs framework reduces the time cost while maintaining or even promoting the high recognition rate. The low rank approximation reduces the frame numbers from 60 to 20 in average. Meanwhile, the processing time of the Light-HMMs is around 1/3 when compared to the baseline on two datasets respectively. The correct recognition rate is also higher than the baseline on both Dataset I ($\epsilon = 0.001, 0.002$) and Dataset II ($\epsilon = 0.001$) since the selected frames are the most salient frames with distinctive discriminative power benefited from low rank approximation. There are only 0.2 percentage point decreasing on Dataset II when the $\epsilon = 0.002$. The parameter ϵ is evaluated with different values as shown in Table IV. In conclusion, the low rank approximation strategy makes HMMs fast, as well as accurate and robust. The performance (94%, 25 frames per second) on Dataset I, which consists of daily used signs, ensure the practicability of online SLR.

D. Experiment on Signer Independent Test

Experiments of user independent SLR are conducted on Dataset III, where 7 groups' signs performed by 7 different signers. The results of Leave-One-Out cross validation are listed in the Table V. The average recognition rates of traditional HMMs algorithm are also listed as reference. Light-HMMs maintains the similar recognition score with approximate 1/3 processing time when compared with the traditional HMMs. In this challenging dataset, the top 1 correct rate decreases by 27 percentage points when compared to the result on user dependent benchmark on dataset II due to the fact that the same sign with different signers, variant illuminations or action speed can result in a large variation on either hand postures or arm skeletons. The correct rate

TABLE V

THE SIGNER-INDEPENDENT TEST RESULT ON DATASET III. $\epsilon = 0.001$

	Light-HMMs	Baseline
Top 1	**0.561**	0.562
Top 5	**0.773**	0.768
Top 10	**0.836**	0.833
Time (/sign)	**125 ms**	322 ms
Frames (/sign)	**28.10**	70.61

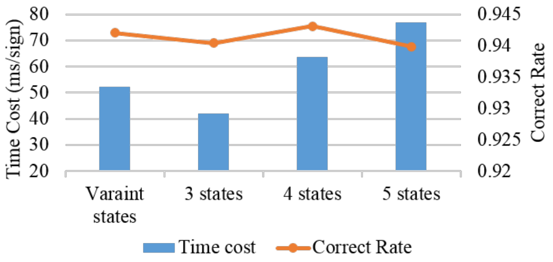


Fig. 5. Part of the results of comparison between variant hidden states and fixed hidden states.

for top 10 can still reach 83.6%.

E. Evaluation On Variant Hidden States

The number of hidden states of the HMMs can be characterised as a trade-off between the time cost and the correct recognition rate. Our variant hidden state strategy can automatically estimate the state number for all the signs and reduce the cost time, which further makes the proposed Light-HMMs fast. We conduct experiments that comparing fixed hidden state and variant hidden states on the Dataset I. Figure 5 shows that variant hidden states strategy costs less time than the cases of 4, 5 hidden states while obtaining higher correct rate than the cases of 3 and 5 (lower than case of 4 by 0.2 percentage points). It can be seen that the variant hidden states strategy is a good balance between time cost and correct rate. Note that the correct rates decrease with hidden state number of 5. That is because the intrinsic hidden state number of signs in Dataset I, which are daily used, is averagely less than 5. This experiment proves that our strategy can automatically determine proper hidden state numbers. The result preferably balances the time cost and the correct recognition rate and makes SLR more practical.

V. CONCLUSIONS AND FUTURE WORKS

This paper proposes a HMMs based framework (Light-HMMs) benefited from low rank approximation for two aspects. First, low rank approximation removes redundant frames and selects key frames for the training and test of HMMs to be faster. Second, the segmentations generated by low rank approximation contribute to determine the number of hidden states for training HMMs. Under the novel Light-HMMs, the encouraging results are obtained in widely tests by using the fused posture appearance features (HOG) and the motion skeleton pair features (SP). Compared with the baseline HMMs method, our Light-HMMs framework costs around 1/3 time while maintaining or even promoting the high recognition rate. From the experiments, it can be seen that the performance is decreased dramatically when facing the signer-independent situation. To make SLR more robust to different signers, further exploration on the deep fusion among hand posture, body skeleton, and the depth map should be the focuses of our future work.

VI. ACKNOWLEDGMENTS

This work was partially supported by the Microsoft Research Asia, and Natural Science Foundation of China under contract Nos. 61390511, 61472398, 61303170, and the FiDiPro program of Tekes.

REFERENCES

[1] T. Pfister, J. Charles, and A. Zisserman, “Large-scale learning of sign language by watching tv (using co-occurrences),” in *BMVC*, 2013.

[2] H. Cooper, B. Holt, and R. Bowden, “Sign language recognition,” in *Visual Analysis of Humans*, pp. 539–562. Springer, 2011.

[3] Z. Zafrulla, H. Brashear, T. Starner, H. Hamilton, and P. Presti, “American sign language recognition with the kinect,” in *ICMI*, 2011, pp. 279–286.

[4] C. Wang, W. Gao, and S. Shan, “An approach based on phonemes to large vocabulary chinese sign language recognition,” in *FG*, 2002, pp. 411–416.

[5] W. Gao, G. Fang, D. Zhao, and Y. Chen, “Transition movement models for large vocabulary continuous sign language recognition,” in *FG*, 2004, pp. 553–558.

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

[7] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, “Real-time human pose recognition in parts from single depth images,” in *CVPR*, 2011, pp. 1297–1304.

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

[9] C. Sun, T. Zhang, B. Bao, C. Xu, and T. Mei, “Discriminative exemplar coding for sign language recognition with kinect,” *Cybernetics, IEEE Transactions on*, vol. 43, no. 5, pp. 1418–1428, 2013.

[10] W. Li, Z. Zhang, and Z. Liu, “Action recognition based on a bag of 3d points,” in *CVPRW*, 2010, pp. 9–14.

[11] J. Wang, Z. Liu, Y. Wu, and J. Yuan, “Mining actionlet ensemble for action recognition with depth cameras,” in *CVPR*, 2012, pp. 1290–1297.

[12] A Kurakin, Z Zhang, and Z Liu, “A real time system for dynamic hand gesture recognition with a depth sensor,” in *EUSIPCO*, 2012, pp. 1975–1979.

[13] Eng-Jon Ong, H. Cooper, N. Pugeault, and R. Bowden, “Sign language recognition using sequential pattern trees,” in *CVPR*, 2012, pp. 2200–2207.

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

[15] X. Chai, G. Li, Y. Lin, Z. Xu, Y. Tang, X. Chen, and M. Zhou, “Sign language recognition and translation with kinect,” in *FG*, 2013.

[16] Rung-Huei Liang and M. Ouhyoung, “A sign language recognition system using hidden Markov model and context sensitive search,” in *Proceedings of the ACM Symposium on Virtual Reality and Technology*, 1996, pp. 59–66.

[17] V. Pitsikalis, S. Theodorakis, C. Vogler, and P. Maragos, “Advances in phonetics-based sub-unit modeling for transcription alignment and sign language recognition,” in *CVPRW*, 2011, pp. 1–6.

[18] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in *CVPR*, 2005, vol. 1, pp. 886–893.

[19] S. Salvador and P. Chan, “Toward accurate dynamic time warping in linear time and space,” *Intelligent Data Analysis*, vol. 11, no. 5, pp. 561–580, 2007.