# Online Early-Late Fusion Based on Adaptive HMM for Sign Language Recognition

DAN GUO, Hefei University of Technology
WENGANG ZHOU and HOUQIANG LI, University of Science and Technology of China
MENG WANG, Hefei University of Technology

In sign language recognition (SLR) with multimodal data, a sign word can be represented by multiply features, for which there exist an intrinsic property and a mutually complementary relationship among them. To fully explore those relationships, we propose an online early-late fusion method based on the adaptive Hidden Markov Model (HMM). In terms of the intrinsic property, we discover that inherent latent change states of each sign are related not only to the number of key gestures and body poses but also to their translation relationships. We propose an adaptive HMM method to obtain the hidden state number of each sign by affinity propagation clustering. For the complementary relationship, we propose an online early-late fusion scheme. The early fusion (feature fusion) is dedicated to preserving useful information to achieve a better complementary score, while the late fusion (score fusion) uncovers the significance of those features and aggregates them in a weighting manner. Different from classical fusion methods, the fusion is query adaptive. For different queries, after feature selection (including the combined feature), the fusion weight is inversely proportional to the area under the curve of the normalized query score list for each selected feature. The whole fusion process is effective and efficient. Experiments verify the effectiveness on the signer-independent SLR with large vocabulary. Compared either on different dataset sizes or to different SLR models, our method demonstrates consistent and promising performance.

CCS Concepts: • **Computing methodologies** → **Activity recognition and understanding**; • **Theory of computation** → Online learning theory;

Additional Key Words and Phrases: Sign language recognition, multi-modal feature fusion, query-adaptive, HMM, online algorithm

**ACM Reference format:**
Dan Guo, Wengang Zhou, Houqiang Li, and Meng Wang. 2017. Online Early-Late Fusion Based on Adaptive HMM for Sign Language Recognition. *ACM Trans. Multimedia Comput. Commun. Appl.* 14, 1, Article 8 (December 2017), 18 pages.
https://doi.org/10.1145/3152121

This work was supported in part to Prof. Meng Wang by NSFC under contract no. 61432019 and in part to Dr. Wengang Zhou by NSFC under contract no. 61632019.
Authors' addresses: D. Guo and M. Wang, School of Computer Science and Information Engineering, Hefei University of Technology, Hefei, 230601, P. R. China; emails: guodan@hfut.edu.cn, eric.mengwang@gmail.com; W. Zhou and H. Li, EEIS Department, University of Science and Technology of China, Hefei, 230027, P. R. China; emails: {zhwg, lihq}@ustc.edu.cn. Corresponding authors: Wengang Zhou, Houqiang Li, and Meng Wang.

https://doi.org/10.1145/3152121

## 1 INTRODUCTION

Vision-based sign language recognition (SLR), which facilitates communication between deaf-mute and normal people with less sign language background, is an emerging problem in areas such as human-computer interaction, computer vision, and pattern recognition (Kurakin et al. 2012; Cheng et al. 2016). It has attracted increasing interest on the topic of activity understanding and enhancement in video (Escalera et al. 2014; Alon et al. 2009; Pfister et al. 2013; Lin et al. 2017).

The SLR problem is still challenging due to several complex factors: subtle tempos and styles of articulation in gesture variation, low resolution of two hands in videos, cultural and individual habits of signers, noise of camera channels, and out-of-vocabulary motion disturbance (Neverova et al. 2016). Its application requires real-time efficiency and robustness on variable observation conditions, scenarios, and sign-independent tests. To effectively mine the rule of gesture variation without any prior knowledge of sign and signer, both massive training samples and an effective learning model are needed. However, in real applications, it is difficult to acquire massive samples of sign language. Current works are usually evaluated on limited data collections with few samples for each sign word. In this article, we focus on the SLR problem on a large vocabulary with few training samples. Our dataset contains 370 Chinese language signs, and each sign only has 20 training samples.

Essentially, SLR is a sequence learning problem. In terms of sequence learning, many methods and models have been proposed, such as Dynamic Time Warping (DTW) (Salvador and Chan 2007; Celebi et al. 2013), Support Vector Machine (Sun et al. 2015), Curve Matching (Lin et al. 2014), Hidden Markov Model (HMM) (Wang et al. 2015; Guo et al. 2016; Zhang et al. 2016), and various popular neural network models (NN) (Huang et al. 2015; Wu et al. 2016b; Neverova et al. 2016; Liu et al. 2016). With those efforts, great success has been made. Among the above models, NN has demonstrated promising performance in many computer vision areas, such as image classification (Krizhevsky et al. 2012), video event understanding (Yang et al. 2016), action recognition (Feichtenhofer et al. 2016), outdoor navigation (Ran et al. 2017), and so forth. But the precondition of the NN model is a large number of training samples. In the case of a large vocabulary with very few training samples per sign word, NN is not a good choice. Considering the limited training data in our problem, we choose another excellent model, the HMM model, as the baseline framework for our SLR problem, witnessing its great success in speech recognition.

Each sign has various potential cues among multimodal features. To further improve the recognition precision, we explore the cues from two different perspectives: intrinsic property and mutually complementary relationship. (1) For intrinsic property, the HMM model is adopted to recover hidden variation of tempos and styles of gesture and action. (2) For complementary relationship, we take the fusion idea into the SLR problem. Current research on fusion for SLR is still at a preliminary stage. State-of-the-art fusion technologies still suffer some limitations for the SLR problem (Belongie et al. 1998; Jain et al. 2005; Khan et al. 2012a). For instance, ineffective feature learning is irreversible, such as graph-based methods (Liu et al. 2011; Wang et al. 2012a); the fixed learned weight of each feature or classification model for different queries is not fair (Zhang et al. 2012; Zheng et al. 2015); the weight learning process based on multiple classifier integration is somewhat time-consuming (Terrades et al. 2009); and so forth. Among them, the difficulties of multimodal feature fusion in SLR are mainly as follows:

- At first, the simple combination of feature fusion may not necessarily lead to a better result than either single feature. For example, in our article, the SP feature relates to a distance vector of a 3D coordinate system, while HOG is a visual feature to describe hand sharp. As shown in Figure 1(d), the combined feature (SP-HOG) containing a "bad" result of the HOG feature may drag down the total performance. To address this problem, we proposed

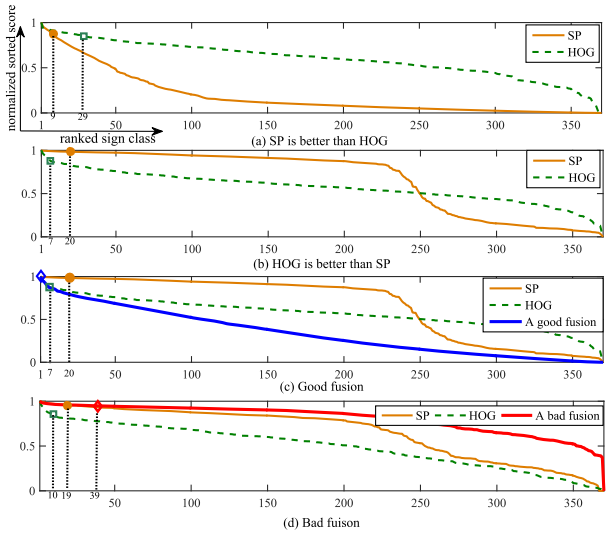



Fig. 1. Multiple feature fusion in SLR. Each symbol on the curves (normalized sorted score lists) marks the correct sign. From (a), (b), and (c), we can observe that a "good" score list is inversely proportional to its curve area. But either feature or score fusion may introduce errors, as shown in (d), which is a "bad" fusion with the combined feature "SP-HOG." In the article, we try to eliminate negative effects among features and combine the merits of two fusion phases (both early and late) to find a "good" fused list.

a feature selection strategy that evaluates characteristics of features (including combined feature) and selects appropriate to-be-fused features.

- Second, for a query, it is important to determine its reasonable scores (relevance probabilities) under different signs. It is nontrivial to adaptively set the model parameters for different signs. The number of hidden variation states in each sign's model indicates key changes of gesture and action. We propose an HMM-state adaptation to determine the state number for each sign model. Then we can adaptively build a learning model of each sign and obtain a score list of the query under these sign models.

- Third, due to individual habit mode and the complexity of motion variation, it is difficult to decide which is always the best feature or score learning model at any query time. As shown in Figure 1, SP is not always the "good" feature on every sign word. Sometimes the score-level performance of SP is better than HOG, but sometimes it is worse. Thus, for a query, we try to realize a query-adaptive score fusion, which query-adaptively preserves useful information at the feature level, assigns appropriate weight at score level, and finally utilizes their merits to integrate the final score.

Therefore, to address the above fusion difficulties, motivated by the score-level late fusion (Zheng et al. 2015), we target SLR and propose an online early (feature)-late (score) fusion framework to effectively utilize both low-level features and high-level decision scores. Actually, our fusion is based on the idea that if the "bad" one greatly drops down the concatenated feature's precision, we drop it and select its complementary (concatenated feature in the early stage) to replace it, and further explore the high-level complementary again in the score fusion learning. The proposed score fusion is query adaptive, unsupervised, and efficient.

As shown in Figure 2, there are four modules in our framework: (1) Early feature fusion: we consider the combined feature, such as SP-HOG in Section 4.1. (2) HMM-state adaptation and

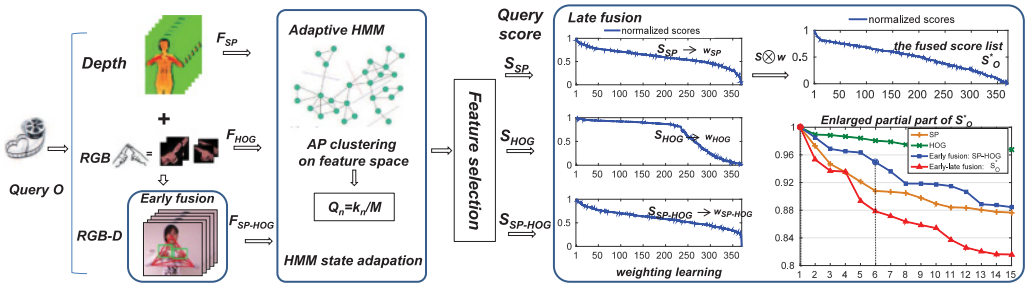



Fig. 2. The framework of our early-late fusion method. At first, we implement the early fusion with RGB-D feature concatenation. Then HMM state adaptation is proposed to adaptively build the learning model of each sign word. After feature selection, we get the score lists of to-be-fused features. In the late fusion scheme, a "good" score list accompanies less area under the score curve than a "bad" one, especially at the front of the score curve with a much steeper shape. We assign the good score list with a large weight and vice versa.

feature selection: we construct our adaptive HMM classifiers under each feature. After feature selection, the probability of query $O$ related to sign words is learned and taken as a score list, namely, a relevance vector. Thus, we obtain multiple to-be-fused score lists. (3) Query-adaptive weight learning: we measure the contribution degrees (weights) of score lists with query $O$ itself. As shown in Figure 1, for a "good" feature, its fusion weight is inversely proportional to the curve area under the $O$'s normalized query score list. The "good" sorted score list always has a relative steep curve, while the "bad" always slowly decreases. This fusion process is query adaptive and unsupervised. (4) Late score fusion: we fuse multiple score lists with assigned weights and rerank signs by the fused score list.

In a nutshell, this article is dedicated to two issues: (1) finding out appropriate hidden states and (2) utilizing multimodal features to improve the recognition precision. The rest of this article is organized as follows: We briefly review related works in Section 2 and introduce our method in Section 3. The experimental results are discussed in Section 4. Finally, we conclude the article in the last section.

## 2 RELATED WORK

In this section, we review related work on three general aspects of the SLR problem. First, traditional approaches for SLR and action recognition are introduced. Then we list currently prevalent SLR datasets and discuss their limitations. After that, some related multimodal fusion work is presented and analyzed.

### 2.1 SLR Model

To learn an effective SLR recognizer, Celebi et al. introduced a weighting scheme with computational complexity comparable to DTW (Salvador and Chan 2007) for gesture recognition (Celebi et al. 2013). Lin et al. (2014) proposed a curve-matching method based on manifold analysis with trajectories of gestures. To improve efficiency, Wang et al. (2015) proposed Light-HMM to select the key frames through low rank approximating and determine the number of hidden states by a Residual Sum of Squares (RSS) threshold. There are also some methods based on neural networks (NNs) for SLR or some similar problems (Kong et al. 2016), such as convolutional neural network (CNN) (Huang et al. 2015; Feichtenhofer et al. 2016; Camgoz et al. 2016), long-short-term memory (LSTM) (Liu et al. 2016), recurrent neural network (RNN) (Neverova et al. 2013), deep dynamic neural networks and HMM (DDNN) (Wu et al. 2016b), recurrent 3D convolutional neural network

(R3DCNN) (Molchanov et al. 2016), temporal convolutions and bidirectional RNN (Pigou et al. 2016), and so forth. All of these prevalent models are effective in continuous sequence learning, such as action and speech recognition domains. However, among these techniques, deep learning methods usually require a large corpus of training data. In our task, the sign vocabulary is large, while the samples per sign word are very few. It is infeasible to apply deep learning to our problem.

### 2.2 SLR Dataset

Cheng et al. pointed out that due to the lack of sign language expertise and the high cost of data collection, training data are usually insufficient in real applications, especially on a large SLR vocabulary (Cheng et al. 2016). As a result, current numerous public datasets in the vision field are usually small in vocabulary size with relatively limited samples, such as the 10-Gesture dataset (Ren et al. 2011), the MSRC-12 Kinect gesture dataset (Fothergill et al. 2012), 12 American Sign Language (ASL) gestures (Kurakin et al. 2012), 24 static ASL sign words (Dong et al. 2015), and 73 ASL signs but with signer-dependent tests (Sun et al. 2013). Even the most popular and well-known gesture dataset, the ChaLearn 2014 dataset (Escalera et al. 2014), only contains 20 gestures. Insufficient data had become a bottleneck for SLR development. ChaLearn 2016 has released its new datasets, isolated and continuous gesture recognition databases, which contain 249 gestures performed by 21 different individuals (Wan et al. 2016). Besides, Wang et al. experimented on a 1,000 Chinese Sign Language (CSL) dataset (Wang et al. 2015). There are also some sentence datasets. For instance, Sun et al. (2015) experimented on an 63-sentence dataset, in which each sentence consisted of two to four sign words. It totally contains 28 sign words. In this article, the involved dataset contains 370 Chinese language signs and each sign only has 20 training samples.

### 2.3 Multimodal Fusion

To explore the complementary SLR learning models, some fusion methods based on multimodal features have pushed the state of the art forward (Ye et al. 2017). Classical multiple feature fusion can be divided into early fusion (Belongie et al. 1998; Khan et al. 2012a, 2012b; Liu et al. 2015; Sun et al. 2016; Wang et al. 2017) and late fusion (Jain et al. 2005; Terrades et al. 2009; Kittler et al. 1998). The early fusion is conducted on the feature level, while the late fusion is conducted on the decision or score level. For late fusion, many efforts have been made to model the distribution of matching similarity (Nandakumar et al. 2008), classifier weighting (Terrades et al. 2009; Kittler et al. 1998), graph-based learning (Liu et al. 2011; Wang et al. 2009, 2012a), and so forth. But there still exist some defects: (1) The fixed learned parameters for different queries are not flexible; (2) some above works consume much more time on fusion optimization with complex computation; and (3) more importantly, in some models, ineffective features may dominate the fusion and drop down the accuracy. Once all features without feature preselection have been taken into account, the fusion process is irreversible. The negative effect of bad features sometimes cannot be eliminated, such as graph-based methods. Besides the above mentioned work, neural network models are also used to perform feature fusion (Wu et al. 2014, 2016a). Wu et al. designed an end-to-end deep learning method for fusing various features (Wu et al. 2014) and further proposed a hybrid deep learning framework integrating both CNN and LSTM to learn multistream multiclass score fusion, which not only weighted the multistream networks for each class but also explored the interclass relationship (Wu et al. 2016a).

In SLR, increasing efforts are focused on the fusion of multimodal features too. Automatic sign language translation was traditionally based on visual recognition techniques (Zhang and Hua 2015; Zhang et al. 2014), until the popularity of Kinect-style depth sensing cameras (Zhao et al. 2014; Cai et al. 2016). In this view, Wang et al. (2015) combined features by skeleton pair feature (SP) and hand HOG feature (HOG) based on the HMM model. As for neural network models, Wu et al.

(2016b) embedded two feature extractor models (i.e., a Deep-Belief Network (DBN) designed for skeletal dynamic data and a 3D CNN for depth and RGB images) into an HMM model to effectively fuse the multimodal gesture datastream. Neverova et al. employed a multiscale and multimodal neural network, termed ModDrop, which targets learning cross-modality correlations between representations of multiple modality channels (Neverova et al. 2016). But constrained to our data limitation, we don't employ neural network fusion in the article. Overall, the work by Wang et al. (2015) is an early fusion, the work by Wu et al. (2016b) is a late fusion, while the work by Neverova et al. (2016) contains both early and late fusion steps based on neural networks. In this article, we combine the merits of two fusion phases (both early and late) and propose an online fusion method for the SLR problem.

## 3 OUR METHOD

Due to the sparsity of training data with very few samples per sign word, we choose the excellent GMM (Gaussian mixture model)-HMM (Hidden Markov Model) model, not the prevalent neural network model, as the basic framework to solve the SLR problem. Given $N$ signs' training data, each sign $n$ has its own HMM model $\lambda_n (1 \leq n \leq N)$, and thus we have $N$ signs' HMMs: $\{\lambda_1, \lambda_2, \ldots, \lambda_N\}$. We use the public toolkit[1] to learn $\{\lambda_n\}$. The recognition process is implemented by the famous Viterbi algorithm, and the most likely sign class $\lambda^*$ of observation sequence $O$ is obtained by Equation (1):

$$\lambda^* = \underset{\{\lambda_1, \lambda_2, \ldots, \lambda_n, \ldots, \lambda_N\}}{argmax} P(\lambda_n | O), \tag{1}$$

where $P(\lambda_n|O)$ learned by the model $\lambda_n (n = 1, \ldots, N)$ indicates the relevance probability of query $O$ related to the $n$th sign.

To further optimize the HMM model, we describe the HMM states' adaptation in Section 3.1 and introduce the early-late fusion in Section 3.2.

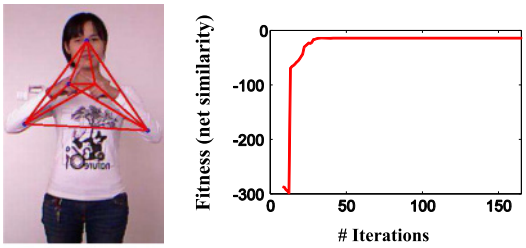
### 3.1 Adaptive HMMs

An excellent HMM model is always tightly related to its inherent latent states. In our task, Chinese signs are characterized by complex and distinct action transitions. It is preferred to adaptively set the parameters of the recognizer models for different sign words. To get a more powerful sign recognizer, we actively learn appropriate latent states for each sign word. We propose an HMM-state adaptation to determine the respective state number $Q_n (1 \leq n \leq N)$ for each sign model.

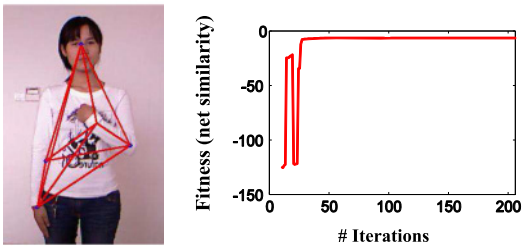
Before learning the HMM model $\lambda_n$, we divide data samples of sign $n$ into reasonable clusters and determine transition types of gesture variation. We adopt affinity propagation (AP) clustering (Frey and Dueck 2007) to adaptively obtain clusters on training data. For sign $n$, we evaluate the distance function in the AP method on any two frame pairs of data samples to construct a frame-similarity net. The net is viewed to obtain the mutual responsibility and availability log-probability ratios between frames $f_j$ and $f_h$. We maximize the similarity optimization function in AP to iteratively find the best frame as exemplar $f_k$, which has larger responsibility weight than all other frames, until no more new exemplars appear. Thus, these best exemplars $\{f_k\}$ are taken as the cluster centers and we can automatically obtain the number of clusters $k_n$.

In our adaptive HMM model, $M$ denotes the cluster number of data distribution in the GMM phase and $Q$ denotes the number of latent states in the HMM phase. Due to our rare samples and chaos characteristic of Gaussian simulation, the effect of $M$ is not very obvious in the SLR problem. Classical SLR approaches make $M$ as a fixed value, and usually set $M = 3$. We follow the rule. The

[1]A public HMM Matlab package: http://www.cs.ubc.ca/murphyk/Software/HMM/hmm.html. Parameters $Q$ and $M$ in the GMM-HMM model are discussed in the article. Other general parameters $A$, $B$, and $\pi$ can be handled by this code package.




(a) Sign "people"




(b) Sign "I"

Fig. 3. Cluster convergence on SP feature with similarity computing. The fitness value is a criterion as in the literature (Frey and Dueck 2007). While it is closer to 0, the clustering convergence is much better. (a) Sign "people." (b) Sign "I."

**ALGORITHM 1:** Early-Late Fusion Based on Adaptive GMM–HMM

**Require:** $N$ signs' training sample sets; Query $O$
**Ensure:** the sign class of query $O$
**Training:**
1: **for** sign $n(1 \le n \le N)$ under feature $i(1 \le i \le m)$ **do**
2: Extract feature set $F_n^{(i)}$ from sign $n$'s training set $Set\_O_n$;
3: Compute the number of clusters $k_n^{(i)}$ on $F_n^{(i)}$ by AP clustering[2];
4: $Q_n^{(i)} = k_n^{(i)}/M$;
5: Learn the GMM-HMM model $\lambda_n^{(i)} = (A, B, \pi)$ with $Set\_O_n$ and $Q_n^{(i)}$;
6: **end for**
**Testing:**
7: Feature selection: e.g., remove "bad" HOG feature in the article;
8: Obtain $O$'s remaining $m'$ score lists $\{s_O^{(i)}\}$ by SLR models $\{\lambda_n^{(i)}\}$;
9: Calculate the fused score list $s_O^*$ by Equation (2)~ Equation (5);
10: $n^* = \arg\max_{n^* \in N} s_O^*$;

number $Q_n(1 \le n \le N)$ is a dominant factor to reflect the number of key gestures and the kinds of their translation relationships. Thus, with fixed $M$-component in the GMM phase of the model, the number of variant states $Q_n$ is proportional to the number of clusters $k_n$, where $k_n$ indicates the total types of key changes of gesture and action. Our adaptive HMM is listed in steps 1 to 5 of Algorithm 1.

[2]As in Frey and Dueck (2007), here we set similarity preference to media similarity in AP.

The state adaptation has good performance on stability and robustness. On the one hand, the clustering method converges quickly on the SLR dataset. For example, action variations of sign "people" and "I" are very different. But as shown in Figure 3 on the fitness of net similarity, which indicates the performance of convergence as in the literature (Frey and Dueck 2007), although sign "I" has much more fluctuation than "people," both of them obviously achieve convergence after 50 iteration times. On the other hand, with the same dataset, the number of converged clusters barely changes by random tests. By LOO cross-validation, we have five groups of 370-sign CSL's training dataset. With 10 times on these five group datasets, in total 18,500 tests, only 20 tests, nearly 0.11% of the generated cluster numbers, change and their difference is very slight. Thus, training data samples can be preprocessed offline to determine $\{Q_1, Q_2, \ldots, Q_N\}$.

### 3.2 Early-Late Fusion

After determining the intrinsic hidden HMM state of each sign, we turn to explore the interrelationship among multimodal features. The early fusion is directly implemented by concatenating different features into a combined feature as detailed in Section 4.1. Here we construct the score list under each feature at the decision level for late fusion and then filter, weight, and fuse score lists to complete the whole method.

*3.2.1 Score List.* At first, we construct the score list of query $O$ under each feature (including the combined feature). Under feature $F^{(i)}(i = 1, \ldots, m)$, we construct the **score list** (a score vector) of query $O$ by $N$ signs' adaptive HMM models $\{\lambda_1^{(i)}, \lambda_2^{(i)}, \ldots, \lambda_N^{(i)}\}$ in Equation (2):

$$s_O^{(i)} = [P(\lambda_1^{(i)}|O), P(\lambda_2^{(i)}|O), \ldots, P(\lambda_N^{(i)}|O)], \tag{2}$$

where $P(\lambda_n^{(i)}|O)$ is obtained by the famous Viterbi algorithm on model $\lambda_n^{(i)}(n = 1, \ldots, N)$. $P(\lambda_n^{(i)}|O)$ indicates the relevance probability of query $O$ related to the $n$th sign under feature $F^{(i)}$. Thus, we obtain $m$ score lists for query $O$: $\{s_O^{(1)}, s_O^{(2)}, \ldots, s_O^{(m)}\}$.

*3.2.2 Feature Selection.* Second, previous early or late fusion works have revealed that a "bad" feature could drop down the overall fusion performance. To avoid this situation, we propose a feature selection strategy: if the performance of a combined feature is better than its single component, we discard the "bad" component feature whose performance is worse than the combined feature. Thus, we can retain the complementary relationship in the combined feature but filter the redundant "bad" information.

In the article, we take the average variance of score lists on training data as our filter criterion. Given a score list (score vector), its variance reflects the deviation degree from its own mean. A smaller variance means that different signs have such similar scores in the list that cannot be distinguished. Instead, a larger variance indicates a good discrimination power. As shown in Figure 4, we implement an LOO cross-validation experiment on a partial small-size dataset: the 50-sign CSL dataset. Under different features, variances and average variances of score lists of a total of 1,000 training samples are respectively illustrated in Figures 4(a) and 4(b). The performance of variance on feature HOG is not as good as the combined feature SP-HOG. Thus, we neglect HOG and select SP and SP-HOG as to-be-fused features.

*3.2.3 Query-Adaptive Weighting.* Third, after feature selection, we assign weights to the remaining $m'$ score lists $\{s_O^{(i)}\}$ . As shown in Figure 1 and Figure 2, the weight is inversely proportional to the area of the normalized sorted score curve. The reason is that a good $s_O^{(i)}$ is assigned a larger weight, while it has a higher score on the correct class and meanwhile a much lower score on other irrelevant classes. In other words, if the sorted score list has a much sharper curve, the

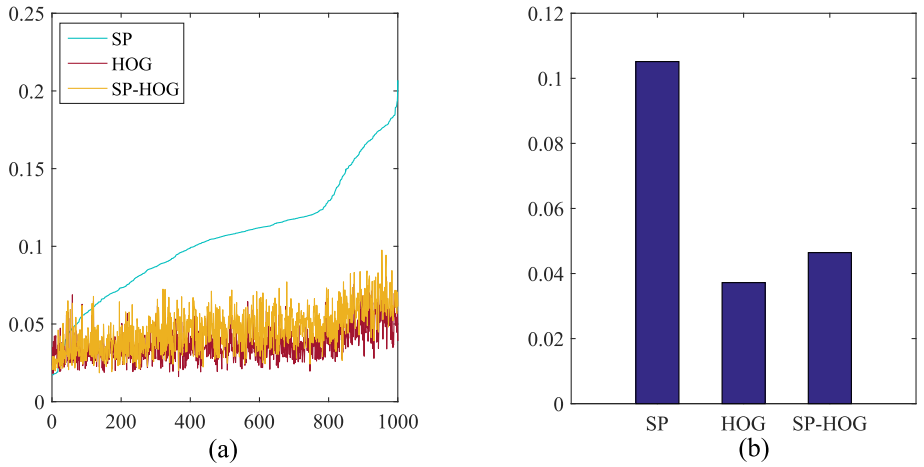



Fig. 4. Variance curves and average variances of score lists of training samples under different features. Feature SP is the best, followed by SP-HOG and HOG. (a) Variance comparison with an arranged sample order that sorted on SP feature. (b) Average variance.

score list with its feature is much more discriminative and helpful. To be more specific, we sort $s_O^{(i)}$ in decreasing order and conduct min-max normalization on it. We denote it as $s_O^{'(i)}$ and weight on $s_O^{'(i)}(1 \le i \le m')$ as follows:

$$\begin{cases} s_O^{'(i)} = \dfrac{s_O^{'(i)} - mins_O^{'(i)}}{maxs_O^{'(i)} - mins_O^{'(i)}} \\ w_O^{(i)} = \dfrac{1/A_{s_O^{'(i)}}}{\sum\limits_{1 \le i \le n} 1/A_{s_O^{'(i)}}}, \end{cases} \tag{3}$$

where $A_{s_O^{'(i)}}$ represents the curve area of the $i$th score list $s_O^{'(i)}$ under feature $F^{(i)}(1 \le i \le m')$. It means that the weighting process is tightly related to $s_O^{'(i)}$, i.e., the query $Q$ itself. The weighting phase is query adaptive and unsupervised.

*3.2.4 Score Fusion.* Finally, we fuse $m'$ score lists. As the product rule usually results in better performance than other rules in biometric multimodality fusion (Kittler et al. 1998; Zheng et al. 2015), the fusion formula is given in Equation (4) or a deformation Equation (5). By the public Matlab code package in footnote 1, we directly implement Equation (5) in a sum format:

$$s_O^* = \left[\prod_{i=1}^{n} (s_O^{'(i)})^{w_O^{(i)}}\right], s.t. \sum_{i=1}^{n} w_O^{(i)} = 1 \tag{4}$$

$$s_O^* = \left[\sum_{i=1}^{n} w_O^{(i)} \cdot \log(s_O^{'(i)})\right], s.t. \sum_{i=1}^{n} w_O^{(i)} = 1. \tag{5}$$

The most likely sign class of query $O$ corresponds to the maximum value in $s_O^*$:

$$n^* = \arg\max_{n^* \in N} s_O^* = \arg\max_{n^* \in N} [s_{O,n}^*], \tag{6}$$

Table 1. Details of the 370-Sign CSL Dataset

| Signs | Dataset | Signer Number | Repetition Time | Sample Number |
|---|---|---|---|---|
| 370 | Training | 4 | 5 | $20 \times 370$ |
| | Testing | 1 | 5 | $5 \times 370$ |

where $s_O^*$ is an $N$-dim vector. Its $n^*$-th component $s_{O,n}^*$ indicates the probability/relevance of query $O$ related to the $n^*$-th sign under feature $F^{(i)}$. The SLR model can be learned offline. Once each sign's SLR classifier is trained, our fusion method is performed online.

## 4 EXPERIMENTS

### 4.1 Experiment Setup

- **Dataset**

Our task focuses on a large vocabulary with limited training samples per sign word. We experiment on the CSL dataset, which is a Kinect RGB-D dataset (Wang et al. 2015). As shown in Table 1, it contains 370 signs played by five signers with five repetitions. The five signers contain both female and male. Their heights and gesture habits are very different. In order to ensure the signer-independent test, we adopt leave-one-out (LOO) cross-validation to test different SLR models in our experiments.

- **Feature representation from RGB-D data**

Current RGB-D SLR datasets have color and depth feature modalities separately. We focus on the consistency and complementary information between the two modalities and their relative importance for SLR tasks. In this article, we take skeleton pair feature (D: 10-dim SP feature), hand feature (RGB: 51-dim HOG feature by PCA dimensionality reduction), and SP-HOG (RGB-D: 61-dim combined feature) as basic features to represent each sign word.

**Hand visual feature (RGB Data):** HOG feature $F_{HOG}$ is extracted from image regions of two hands in videos by using the self-adaptive skin model and depth constraint as in Wang et al. (2015).[3]

**Skeleton pair feature (Depth Data):** For depth data, we extract mutual distances of five skeleton points (head, left elbow, right elbow, left hand, and right hand) and transform them to a 10-dimension SP distance feature $F_{SP}$ (Wang et al. 2012b). Each signer has different body size. To unify gesture posture scales by different signers, we normalize each SP vector by its maximum value. Then we can get each sign sample's SP observation sequence $O_{SP}$ for SLR model training. This feature is invariant to transformations of rotation, scaling, and translation.

**Combined feature (RGB-D Data):** We concatenate $F_{HOG}$ and $F_{SP}$ and denote the new vector as the SP-HOG feature (61-dim combined feature). The SP-HOG feature is deemed as an early fusion. The late fusion is detailed in Section 3.2.

- **Data preprocessing**

**Dimensionality reduction on HOG feature:** Since the dimension of the original HOG feature is too high, PCA (Principal Component Analysis) is applied. We retain about 80% information energy of the dataset by PCA and get the 51-dim HOG feature. To keep sign-independent tests, we obtain the transform matrix on training data. When recognizing a testing sample, we transform the query sample by the obtained matrix.

[3]In this article, we focus on fusion of multimodal features and make no intent to optimize feature extraction. The HOG features in the article were originally extracted with OpenCV with basic parameters, while both the HOG feature and the SP feature used in Wang et al. (2015) are further optimized versions, e.g., some invalid frames are removed.

**Data augmentation on feature SP:** In our experiments, we enrich our limited training data with data augmentation (Chatfield et al. 2014). In order to add appropriate noises and prevent overfitting, we explore a random Gaussian disturbance strategy on skeleton coordinates to augment additional gesture position features.

As a 3D depth skeleton point $(x, y, z)$ collected by Kinect, we take the $x$ coordinate as an example to explain Gaussian disturbance, and $y$ and $z$ coordinates are similar. First, we check the range of $x$ in all training samples under each sign $n$: $[x^n_{max}, x^n_{min}]$. Here is $\Delta x^n = x^n_{max} - x^n_{min}$. Then we set a Gaussian random variable $X \sim N(0, (\eta \Delta x^n)^2)$, where $\eta$ is a disturbance parameter. In our tests, we find that $\eta = 0.01$ is the best. Under sign $n$, an additional $(x', y', z')$ coordinate of the skeleton point is generated as follows:

$$\begin{cases} x' = x + N(0, (\eta \Delta x^n)^2) \\ y' = y + N(0, (\eta \Delta y^n)^2) \\ z' = z + N(0, (\eta \Delta z^n)^2). \end{cases} \tag{7}$$

If this augmentation is conducted $N$ times, we can expand the original dataset $N$ times. However, we downsample every frame and augment the data by $N = 1$. It still has the same number of frames but brings reasonable Gaussian disturbance into the data.

- **Contrasted SLR approaches**

We evaluate our approach on different data sizes and compare with other SLR approaches, such as DTW (Salvador and Chan 2007; Celebi et al. 2013), GMM-HMM, and Light-HMM (Wang et al. 2015). And we also compare our fusion scheme with some fusion works, such as an early fusion for SLR (Wang et al. 2015) and a late fusion (Zheng et al. 2015).

—GMM-HMM (Wang et al. 2015): A good parameter setting for traditional GMM-HMM is $Q = M = 3$, where $Q$ is the number of states in HMM and $M$ is the number of mixture models in GMM.
—Light-HMM (Wang et al. 2015): To trade off precision and runtime, Light-HMM selects key frames and determines adaptive $Q$. Here $M = 3$ and $Q$ is adaptive. To obtain the best performance, we set LightHMM's threshold $\varepsilon_0 = 0.001$ and threshold $\lambda$ to the average value of the RSS score curve of parameter $\varepsilon$.
—DTW (Salvador and Chan 2007; Celebi et al. 2013): The DTW model is quite different form HMM-based models. It does not calculate the probability of query $O$ under each sign class. DTW searches its nearest neighbor in the training dataset and defines the sign class of the nearest neighbor as its class. Therefore, the score list by DTW is set as reciprocal of its distances to all training samples.

### 4.2 Experiment with HMM-State Adaptation

We first test our adaptive HMM. The experimental result on the SP feature is shown in Table 2. In our HMM, parameters $Q$ and $M$ are discussed. Our adaptation is HMM($Q$) with adaptive $Q$ and $M = 3$. And we also test another adaptation by the HMM($M$) with adaptive $M$ and $Q = 3$.

Table 2 shows that adaptive HMM($Q$) is the best. Light-HMM with a few key frames brings down its precision. DTW is much more time-consuming than other HMMs. That is due to the fact that DTW retrieves all training samples to decide its sign class, while HMMs just learn the score of each sign. In our adaptive HMM, $Q$ has much more influence than $M$. $Q$ indicates the number of latent states in the HMM phase and $M$ indicates the number of clusters of data samples in the GMM phase. The former hints at state changes and the later simulates data distribution. Due to

Table 2. Performance of Our Adaptive HMM Compared to Other Methods under a Single SP Feature on the 370-Sign Dataset

| Methods | Top 1 | Top 5 | Top 10 | Testing Time (ms/sign) |
|---|---|---|---|---|
| DTW | 31.59% | 52.84% | 62.45% | 1,730 |
| GMM-HMM | 27.51% | 53.84% | 65.83% | 159 |
| LightHMM | 21.96% | 45.29% | 63.22% | 128 |
| Our adaptive HMM($M$) | 28.10% | 54.04% | 66.62% | 82 |
| Our adaptive HMM($Q$) | **34.82%** | **61.29%** | **70.80%** | 88 |

Table 3. Fusion Types in Our Fusion Framework

| **Fusion I** | **Fusion II** | **Fusion III** | **Our Fusion** |
|---|---|---|---|
| Early fusion | Late fusion | Early-late fusion | Feature selection + early-late fusion |
| SP-HOG feature fusion | SP ⊗ HOG score fusion | SP ⊗ HOG ⊗ SP-HOG (feature + score) fusion | SP ⊗ SP-HOG (feature + score) fusion |

Table 4. Performance of Various Fusion Types on the 370-Sign Dataset

| Feature | Recall @ $R$ | | |
|---|---|---|---|
| | $R = 1$ | $R = 3$ | $R = 5$ |
| SP | 34.82% | 53.26% | 61.29% |
| HOG | 21.52% | 34.03% | 39.89% |
| Fusion I (Wang et al. 2015) | 32.02% | 45.71% | 51.92% |
| Fusion II (Zheng et al. 2015) | 41.21% | 56.27% | 62.75% |
| Fusion III | 41.40% | 56.74% | 62.99% |
| **Our fusion** | **45.32%** | **60.75%** | **67.34%** |

rare samples and chaos characteristic of Gaussian simulation, the effect of $M$ is not very obvious in our problem. But $Q$ is closely related to sign action change. Thus, we still choose adaptive HMM($Q$) as our adaptation strategy.

### 4.3 Comparison on Different Fusion Steps

Here we list different fusion strategies in Table 3. From the results on our signer-independent dataset shown in Table 4, at Recall@$R$=1, the precision on the SP feature is 34.82% and the HOG feature only achieves 21.52%. HOG has a negative fusion effect that leads to the precision of Fusion I (early fusion) being less than that on a single SP feature. Fusion II (late fusion), Fusion III (simple early-late fusion), and our fusion obtain notable improvements. Fusion II has already learned the positive effect of complementarity of features.

Meanwhile, as shown in Figure 5, Fusion II and Fusion III achieve similar performances, but our fusion consistently achieves the best performance. It makes a 13.30% improvement compared to Fusion I, 4.11% compared to Fusion II, and 3.92% compared to Fusion III at Recall@$R$=1. Because of that, our fusion compared to Fusion II and Fusion III further drops the negative effect of "bad" single feature HOG.

### 4.4 Comparison on Different Dataset Sizes

We also evaluate on different sizes. We pick subsets of 370 sign words as small datasets, such as the top 50, 100, and 200 words. As shown in Figure 6(a), with the increase of Recall@$R$, the

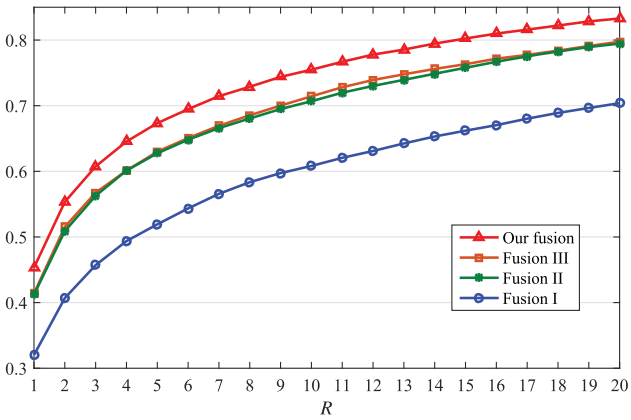



Fig. 5. Recall@$R$ on the 370-sign dataset.

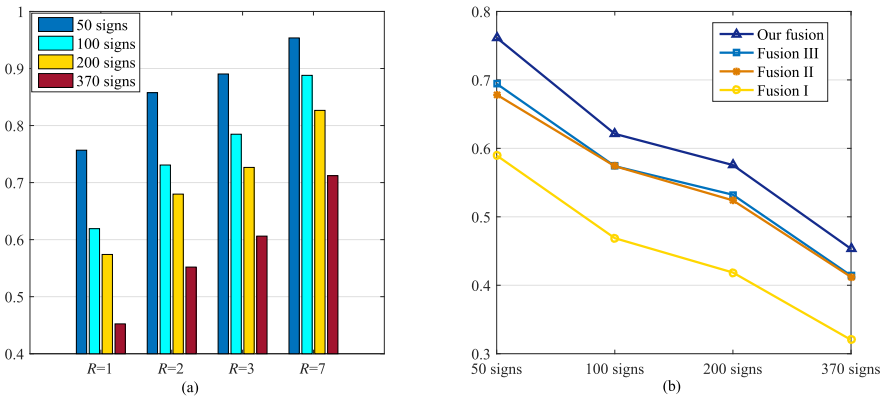



Fig. 6. (a) Precision comparison of our fusion on different datasets. (b) Precision differences of our fusion to Fusions I, II, and III at Recall@$R$=1.

precision steadily grows. On the other hand, when the sign word number increases, the performance declines. The differences of our fusion to other fusion methods are shown in the Figure 6(b). Our fusion still has good precision and stability. It basically improves by 3.92% to 17.20% on different dataset sizes.

### 4.5 Comparison on Different SLR Models

As for comparison on different SLR models, we also evaluate them on the datasets with 50 signs and 370 signs, where the 50-sign dataset is the subset of the 370-sign dataset. As shown in Tables 5 and 6, the HOG feature still achieves poor performance in the SLR problem, especially on the DTW model with the worst precision. Due to the poor performance of HOG, Fusion I (early fusion) performs badly on the 50-sign dataset. On the 370-sign dataset, Fusion I sometimes improves but is still unstable in most cases. Interestingly, Fusion II effectively makes use of HOG, the precision of Fusion III is similar to Fusion II, and our fusion further raises the precision. In short, our method achieves the best performance by exploring the merit of early-late fusion and utilizing complementary.

Table 5. Fusion Comparison on Different SLR Models with 50 Signs

| | DTW | | GMM-HMM | | Light-HMM (Wang et al. 2015) | | Our Adaptive HMM | |
|---|---|---|---|---|---|---|---|---|
| | Reall@1 | Reall@4 | Reall@1 | Reall@4 | Reall@1 | Reall@4 | Recall@1 | Reall@4 |
| SP | 61.84% | 82.08% | 61.60% | 88.96% | 55.44% | 82.00% | 71.68% | 93.60% |
| HOG | 20.24% | 39.68% | 44.64% | 68.64% | 29.12% | 55.92% | 46.88% | 68.24% |
| Fusion I (Wang et al. 2015) | 22.88% | 43.44% | 58.40% | 81.36% | 50.40% | 74.64% | 58.96% | 79.20% |
| Fusion II (Zheng et al. 2015) | 63.12% | 82.96% | 66.08% | 84.96% | 49.92% | 75.76% | **67.84%** | **87.12%** |
| Fusion III | 63.96% | 83.76% | 66.16% | 86.88% | 57.68% | 80.08% | **69.44%** | **87.36%** |
| **Our fusion** | 63.04% | 83.04% | 72.64% | 90.40% | 61.92% | 83.52% | **76.16%** | **91.84%** |

Table 6. Fusion Comparison on Different SLR Models with 370 Signs

| | GMM-HMM | | | Light-HMM | | | Our Adaptive HMM | | |
|---|---|---|---|---|---|---|---|---|---|
| | Reall@1 | Reall@3 | Reall@5 | Reall@1 | Reall@3 | Reall@5 | Recall@1 | Recall@3 | Reall@5 |
| SP | 27.51% | 45.56% | 53.84% | 21.96% | 37.41% | 45.29% | **34.82%** | **53.26%** | **61.29%** |
| HOG | 21.38% | 33.51% | 39.61% | 10.50% | 19.29% | 24.93% | 21.52% | 34.03% | 39.89% |
| Fusion I (Wang et al. 2015) | 32.00% | 46.13% | 52.63% | 24.50% | 38.49% | 45.28% | 32.02% | 45.71% | 51.92% |
| Fusion II (Zheng et al. 2015) | 37.46% | 52.95% | 59.39% | 22.22% | 36.23% | 43.57% | **41.21%** | **56.27%** | **62.75%** |
| Fusion III | 38.93% | 54.42% | 60.79% | 29.04% | 44.41% | 51.72% | **41.40%** | **56.75%** | **62.99%** |
| **Our fusion** | **41.50%** | **57.36%** | **63.91%** | **33.88%** | **49.64%** | **56.67%** | **45.32%** | **60.75%** | **67.34%** |

Table 7. Time Comparison on 50 Signs

| Avg. Testing Time (s) | DTW | GMM-HMM | Light-HMM | Our Adaptive HMM |
|---|---|---|---|---|
| SP | 1.730 | 0.088 | 0.128 | 0.159 |
| HOG | 8.495 | 0.123 | 0.399 | 0.179 |
| Fusion I | 9.025 | 0.124 | 0.217 | 0.156 |
| Fusion II | **0.014** | **0.011** | **0.011** | **0.011** |
| Fusion III | **0.015** | **0.011** | **0.011** | **0.011** |
| Our fusion | **0.014** | **0.011** | **0.011** | **0.011** |

Fusion time in this table merely indicates time of fusion computation.

Meanwhile, Table 7 shows the time cost of different SLR models and fusions. "Our fusion" and "Fusion III" in the table denote the time of fusion computation in Section 3.2.4. DTW is still much more time-consuming than other HMMs. The time cost of various HMMs is close. GMM-HMM has a stable time cost with a fixed value $Q = 3$. Under our dataset, although under a few key frames, the time cost of LightHMM is higher than GMM-HMM, as its average adaptive $Q$ is nearly 4 to 5 times of GMM-HMM's $Q$. It has more complexity of adaptive state transition calculations compared to GMM-HMM, and so does our adaptive HMM, which also has a variable $Q$. Anyway, score fusion time is trivial compared to query time under different SLR models. The fusion computation is efficient for online fusion.

Table 8. Comparison on Different Settings of Late Fusion in Our Fusion Framework

| Datasize | Precision/Time | REF | Our Fusion |
|---|---|---|---|
| 50 signs | Fusion precision | 75.76% | 76.16% |
| | Time (ms) | 16.7 | **11.2** |
| 370 signs | Fusion precision | 45.23% | 45.32% |
| | Time (ms) | 34.6 | **21.6** |

Table 9. Recall@$R$ Comparison Between Fusion III and Our Fusion

| Datasize | Method | Recall@1 | Recall@2 | Recall@3 | Recall@4 | Recall@5 |
|---|---|---|---|---|---|---|
| 50 signs | Fusion III | 69.44% | 80.08% | 85.12% | 87.36% | 89.12% |
| | Our fusion | **76.16%** | **85.92%** | **89.36%** | **91.84%** | **93.36%** |
| 370 signs | Fusion III | 41.41% | 51.54% | 56.75% | 60.11% | 62.99% |
| | Our fusion | **45.32%** | **55.34%** | **60.75%** | **64.54%** | **67.34%** |

### 4.6 Extension on Different Early-Late Fusion Comparisons

At last, we compare two extensions of the early-late fusion framework. One is integrating the reference strategy into our fusion, which is very effective in a late fusion method for image search and person reidentification (Zheng et al. 2015). The other is more comparison between Fusion III and our fusion.

**Reference construction comparison:** We extend the reference construction in Zheng et al. (2015) into our fusion and abbreviate the extension as "REF." Before score fusion, the sorted score list $s_O^{'(i)}$ is subtracted by an irrelative **reference** list $r_O^{(i)}$ (Zheng et al. 2015). Once picking up irrelevant samples, we construct their sorted score lists $\{r\}$. Then $s_O^{'(i)}$ of query $O$ under the feature $F^{(i)}$ is calculated by Equation (8):

$$\begin{aligned} & s_O^{'(i)} = s_O^{'(i)} - r_O^{(i)} \\ & s.t.\ \ r_O^{(i)} = \underset{r_O^{(i)} \in \{r\}}{\arg min} ||s^{'(i)}(u:v) - r(u:v)||_2, \end{aligned} \tag{8}$$

where $u$ and $v$, respectively, denote the beginning and the end position of reference construction on the score curve. We use parameter $k$ in our SLR problem to denote the number of irrelevant sign classes. Thus, $v$ is the total number of sign words minus $k$. Our SLR dataset has very limited samples and signs are not completely interrelated, and thus we merely pick up training samples of the last ranked sign class as the irrelative samples. For 370 words, we set $k = 1$ and $v = 369$. In our test experiments, $u = 15$ is a better parameter setting.

As shown in Table 8, REF and our fusion method have very close performance. Our fusion is slightly better. This may be because due to the complex correlation among different signs, human actions are not completely irrelevant in the real world. Reference construction does not work well in our SLR model. Besides, REF takes an additional time cost on training sample retrieval and reference computation, while our method is more efficient.

**Comparison on different early-late frameworks (Fusion III vs. our fusion):** Here we give more details of our fusion compared to Fusion III. Without feature selection, Fusion III totally fuses the score lists of features SP, HOG, and combined feature SP-HOG. With feature selection, ours fuses score lists of SP and SP-HOG. As shown in Table 9 and compared in Tables 5 and 6, Fusion III is already better than early fusion (Fusion I) and late fusion (Fusion II). But our fusion

is still obviously better than all others, including Fusion III. It's because only the early-fusion framework cannot completely eliminate the overloading negative effects of "bad" features. Our feature selection works well to avoid this and further promote the precision.

## 5 CONCLUSION

We propose an online fusion framework based on adaptive HMM for sign language recognition to integrate early and late fusions. The HMM-state adaptation addresses temporal structure variation of individual habit mode and sign complexity under heterogeneous modalities. Early feature fusion can extract complementary representations in terms of joint performance on a subset or complete set of modalities. We propose an adaptive selection strategy to identify those features to be fused. At last, on the score-level fusion, we further adaptively explore contexts of multimodal features underlying the sorted scores in an unsupervised and query-adaptive manner. Experiments demonstrate the effectiveness and efficiency of our approach.

## REFERENCES

Jonathan Alon, Vassilis Athitsos, Quan Yuan, and Stan Sclaroff. 2009. A unified framework for gesture recognition and spatiotemporal gesture segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31, 9 (2009), 1685–1699.

Serge Belongie, Chad Carson, Hayit Greenspan, and Jitendra Malik. 1998. Color- and texture-based image segmentation using em and its application to content-based image retrieval. In *IEEE Conference on Computer Vision*. 675–682.

Xingyang Cai, Wengang Zhou, Lei Wu, Jiebo Luo, and Houqiang Li. 2016. Effective active skeleton representation for low latency human action recognition. *IEEE Transactions on Multimedia* 18, 2 (2016), 141–154.

Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, and Richard Bowden. 2016. Using convolutional 3d neural networks for user-independent continuous gesture recognition. In *IEEE Conference on Pattern Recognition*. 49–54.

Sait Celebi, Ali Selman Aydin, Talha Tarik Temiz, and Tarik Arici. 2013. Gesture recognition using skeleton data with weighted dynamic time warping. In *International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*. 620–625.

Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Return of the devil in the details: Delving deep into convolutional nets. *Arxiv Preprint Arxiv:1405.3531* (2014).

Hong Cheng, Lu Yang, and Zicheng Liu. 2016. A survey on 3d hand gesture recognition. *IEEE Transactions on Circuits and Systems for Video Technology* 29, 9 (2016), 1659–1673.

Cao Dong, Ming Leu, and Zhaozheng Yin. 2015. American sign language alphabet recognition using microsoft kinect. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 44–52.

Sergio Escalera, Xavier Baró, Jordi Gonzalez, Miguel A Bautista, Meysam Madadi, Miguel Reyes, Víctor Ponce-López, Hugo J. Escalante, Jamie Shotton, and Isabelle Guyon. 2014. Chalearn looking at people challenge 2014: Dataset and results. In *European Conference on Computer Vision Workshop*. 459–473.

Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. 2016. Convolutional two-stream network fusion for video action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*. 1933–1941.

Simon Fothergill, Helena Mentis, Pushmeet Kohli, and Sebastian Nowozin. 2012. Instructing people for training gestural interactive systems. In *The SIGCHI Conference on Human Factors in Computing Systems*. 1737–1746.

Brendan J. Frey and Delbert Dueck. 2007. Clustering by passing messages between data points. *Science* 315, 5814 (2007), 972–976.

Dan Guo, Wengang Zhou, Meng Wang, and Houqiang Li. 2016. Sign language recognition based on adaptive hmms with data augmentation. In *IEEE Conference on Image Processing*. 2876–2880.

Jie Huang, Wengang Zhou, Houqiang Li, and Weiping Li. 2015. Sign language recognition using 3d convolutional neural networks. In *IEEE Conference on Multimedia and Expo*. 1–6.

Anil Jain, Karthik Nandakumar, and Arun Ross. 2005. Score normalization in multimodal biometric systems. *Pattern Recognition* 38, 12 (2005), 2270–2285.

Fahad Shahbaz Khan, Rao Muhammad Anwer, Joost Van de Weijer, Andrew D. Bagdanov, Maria Vanrell, and Antonio M. Lopez. 2012a. Color attributes for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*. 3306–3313.

Fahad Shahbaz Khan, Joost Van de Weijer, and Maria Vanrell. 2012b. Modulating shape features by color attention for object recognition. *International Journal of Computer Vision* 98, 1 (2012), 49–64.

Josef Kittler, Mohamad Hatef, Robert P. W. Duin, and Jiri Matas. 1998. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20, 3 (1998), 226–239.

Tao Kong, Anbang Yao, Yurong Chen, and Fuchun Sun. 2016. HyperNet: Towards accurate region proposal generation and joint object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*. 845–853.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet classification with deep convolutional neural networks. In *International Conference on Neural Information Processing Systems*. 1097–1105.

Alexey Kurakin, Zhengyou Zhang, and Zicheng Liu. 2012. A real time system for dynamic hand gesture recognition with a depth sensor. In *European Signal Processing Conference*. 1975–1979.

Shih Yao Lin, Yen Yu Lin, Chu Song Chen, and Yi Ping Hung. 2017. Recognizing human actions with outlier frames by observation filtering and completion. *ACM Transactions on Multimedia Computing, Communications, and Applications* 13, 3 (2017), 28.

Yushun Lin, Xiujuan Chai, Yu Zhou, and Xilin Chen. 2014. Curve matching from the view of manifold for sign language recognition. In *Asian Conference on Computer Vision*. 233–246.

Tao Liu, Wengang Zhou, and Houqiang Li. 2016. Sign language recognition with long short-term memory. In *IEEE Conference on Image Processing*. 2871–2875.

Wei Liu, Yu Gang Jiang, Jiebo Luo, and Shih Fu Chang. 2011. Noise resistant graph ranking for improved web image search. In *IEEE Conference on Computer Vision and Pattern Recognition*. 849–856.

Zhen Liu, Houqiang Li, Wengang Zhou, Richang Hong, and Qi Tian. 2015. Uniting keypoints: Local visual information fusion for large-scale image search. *IEEE Transactions on Multimedia* 17, 4 (2015), 538–548.

Pavlo Molchanov, Xiaodong Yang, Shalini Gupta, Kihwan Kim, Stephen Tyree, and Jan Kautz. 2016. Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural network. In *IEEE Conference on Computer Vision and Pattern Recognition*. 4207–4215.

Karthik Nandakumar, Yi Chen, Sarat C. Dass, and Anil K. Jain. 2008. Likelihood ratio-based biometric score fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30, 2 (2008), 342–347.

Natalia Neverova, Christian Wolf, Giulio Paci, Giacomo Sommavilla, Graham Taylor, and Florian Nebout. 2013. A multi-scale approach to gesture detection and recognition. In *IEEE Conference on Computer Vision Workshops*. 484–491.

Natalia Neverova, Christian Wolf, Graham Taylor, and Florian Nebout. 2016. Moddrop: Adaptive multi-modal gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38, 8 (2016), 1692–1706.

Tomas Pfister, James Charles, and Andrew Zisserman. 2013. Large-scale learning of sign language by watching tv (using co-occurrences). In *British Machine Vision Conference*.

Lionel Pigou, Aäron Van Den Oord, Sander Dieleman, Mieke Van Herreweghe, and Joni Dambre. 2016. Beyond temporal pooling: Recurrence and temporal convolutions for gesture recognition in video. *International Journal of Computer Vision* (2016), 1–10. https://doi.org/10.1007/s11263-016-0957-7.

Lingyan Ran, Yanning Zhang, Qilin Zhang, and Tao Yang. 2017. Convolutional neural network-based robot navigation using uncalibrated spherical images. *Sensors* 17, 6 (2017), 1341.

Zhou Ren, Junsong Yuan, and Zhengyou Zhang. 2011. Robust hand gesture recognition based on finger-earth mover's distance with a commodity depth camera. In *ACM Conference on Multimedia*. 1093–1096.

Stan Salvador and Philip Chan. 2007. Toward accurate dynamic time warping in linear time and space. *Intelligent Data Analysis* 11, 5 (2007), 561–580.

Chao Sun, Tianzhu Zhang, BingKun Bao, Changsheng Xu, and Tao Mei. 2013. Discriminative exemplar coding for sign language recognition with kinect. *IEEE Transactions on Cybernetics* 43, 5 (2013), 1418–1428.

Chao Sun, Tianzhu Zhang, and Changsheng Xu. 2015. Latent support vector machine modeling for sign language recognition with Kinect. *ACM Transactions on Intelligent Systems and Technology* 6, 2 (2015), 20.

Shaoyan Sun, Wengang Zhou, Qi Tian, and Houqiang Li. 2016. Scalable object retrieval with compact image representation from generic object regions. *ACM Transactions on Multimedia Computing, Communications, and Applications* 12, 2 (2016), 29.

Oriol Ramos Terrades, Ernest Valveny, and Salvatore Tabbone. 2009. Optimal classifier fusion in a non-Bayesian probabilistic framework. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31, 9 (2009), 1630–1644.

Jun Wan, Yibing Zhao, Shuai Zhou, Isabelle Guyon, Sergio Escalera, and Stan Z. Li. 2016. Chalearn looking at people rgb-d isolated and continuous datasets for gesture recognition. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 56–64.

Hanjie Wang, Xiujuan Chai, Yu Zhou, and Xilin Chen. 2015. Fast sign language recognition benefited from low rank approximation. In *IEEE Conference and Workshops on Automatic Face and Gesture Recognition*. 1–6.

Jiang Wang, Zicheng Liu, Ying Wu, and Junsong Yuan. 2012b. Mining actionlet ensemble for action recognition with depth cameras. In *IEEE Conference on Computer Vision and Pattern Recognition*. 1290–1297.

Meng Wang, Xian Sheng Hua, Richang Hong, Jinhui Tang, Guo Jun Qi, and Yan Song. 2009. Unified video annotation via multigraph learning. *IEEE Transactions on Circuits and Systems for Video Technology* 19, 5 (2009), 733–746.

Meng Wang, Hao Li, Dacheng Tao, Ke Lu, and Xindong Wu. 2012a. Multimodal graph-based reranking for web image search. *IEEE Transactions on Image Processing* 21, 11 (2012), 4649–4661.

Meng Wang, Changzhi Luo, Bingbing Ni, Jun Yuan, Jianfeng Wang, and Shuicheng Yan. 2017. First-person daily activity recognition with manipulated object proposals and non-linear feature fusion. *IEEE Transactions on Circuits and Systems for Video Technology* 99 (2017), 1–1. https://doi.org/10.1109/TCSVT.2017.2716819.

Di Wu, Lionel Pigou, Pieter-Jan Kindermans, L. E. Nam, Ling Shao, Joni Dambre, and Jean-Marc Odobez. 2016b. Deep dynamic neural networks for multimodal gesture segmentation and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38, 8 (2016), 1583–1597.

Zuxuan Wu, Yu Gang Jiang, Jun Wang, Jian Pu, and Xiangyang Xue. 2014. Exploring inter-feature and inter-class relationships with deep neural networks for video classification. In *ACM Conference on Multimedia*. 167–176.

Zuxuan Wu, Yu Gang Jiang, Xi Wang, Hao Ye, and Xiangyang Xue. 2016a. Multi-stream multi-class fusion of deep networks for video classification. In *ACM Conference on Multimedia*. 791–800.

Xiaoshan Yang, Tianzhu Zhang, and Changsheng Xu. 2016. Semantic feature mining for video event understanding. *ACM Transactions on Multimedia Computing, Communications, and Applications* 12, 4 (2016), 55.

Jun Ye, Hao Hu, Guo Jun Qi, and Kien A. Hua. 2017. A temporal order modeling approach to human action recognition from multimodal sensor data. *ACM Transactions on Multimedia Computing, Communications, and Applications* 13, 2 (2017), 14.

Jihai Zhang, Wengang Zhou, Chao Xie, Junfu Pu, and Houqiang Li. 2016. Chinese sign language recognition with adaptive HMM. In *IEEE Conference on Multimedia and Expo*. 1–6.

Qilin Zhang and Gang Hua. 2015. Multi-view visual recognition of imperfect testing data. In *ACM Conference on Multimedia*. 561–570.

Qilin Zhang, Gang Hua, Wei Liu, Zicheng Liu, and Zhengyou Zhang. 2014. Can visual recognition benefit from auxiliary information in training? In *Asian Conference on Computer Vision*. 65–80.

Shaoting Zhang, Ming Yang, Timothee Cour, Kai Yu, and Dimitris N. Metaxas. 2012. Query specific fusion for image retrieval. In *European Conference on Computer Vision*. 660–673.

Xin Zhao, Xue Li, Chaoyi Pang, Quan Z. Sheng, Sen Wang, and Mao Ye. 2014. Structured streaming skeleton – a new feature for online human gesture recognition. *ACM Transactions on Multimedia Computing, Communications, and Applications* 11, 1s (2014), 22.

Liang Zheng, Shengjin Wang, Lu Tian, Fei He, Ziqiong Liu, and Qi Tian. 2015. Query-adaptive late fusion for image search and person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*. 1741–1750.

Received January 2017; revised October 2017; accepted October 2017