SEMANTICS CONSTRAINED DICTIONARY LEARNING FOR SIGNER-INDEPENDENT SIGN LANGUAGE RECOGNITION

Fang Yin[1,2] *Xiujuan Chai*[1] *Yu Zhou*[3] *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] University of Chinese Academy of Sciences, Beijing 100049, China
[3] Institute of Information Engineering, Chinese Academy of Sciences, Beijing, 100093, China

ABSTRACT

In this paper, a sparse coding based framework is proposed for sign language recognition (SLR), especially for the signer-independent case. To deal with the inter-signer variation, a dictionary capturing the common features among different signers is learnt by considering the semantic constraint. Thus for a given sign from an unknown signer, the sparse representation, which maintains more information of this specific sign class while neglecting the identity information as much as possible, can be generated. In our implementation, each sign is partitioned into a fixed number of fragments and the features fusing hand shape and moving trajectory are extracted from the fragments. The dictionary learnt from the training fragments can be taken as the basic subunits of signs and each fragment of sign video can be coded by these basis vectors. Finally, the recognition result is achieved through SVM with the concatenated sparse coding features of the fragments. The experiments and comparisons show that our method is more effective for the signer-independent recognition problem than other baseline methods. At the same time, it also performs well for the signer-dependent case.

Index Terms— Sparse coding, dictionary learning, sign language recognition, signer-independent, semantic constraint

1. INTRODUCTION

Although the research in sign language recognition has been somewhat successful, most of the previous work focused on the signer-dependent situation. In real life, many inter-signer variations appear in the signing procedure, mainly existing in different hand shapes and movements. Unfortunately, when confronted with this kind of variations, the performance of the traditional signer-dependent SLR system will decrease dramatically. So currently, one of the big challenges in SLR is the signer-independent problem.

For signer-independent SLR, the existing methods can be broadly classified into two categories. One is to extract the feature or use the recognition system which is robust to different signers[1, 2]. Zieren and Kraiss[1] used the features normalized for signer-independence and realized the classification in a HMM framework. The other category for signer-independent recognition is using the new signer's data to adapt the generalized model[3–6]. Borrowing from speech recognition, Agris et al.[3] used Maximum Likelihood Linear Regression(MLLR) and Maximum A Posteriori(MAP) for signer adaptation. Farhadi et al.[4] introduced transfer learning to signer-independent sign language recognition. They proposed a comparative feature and obtained recognition rate of 64.2% on a new signer with 90 signs.

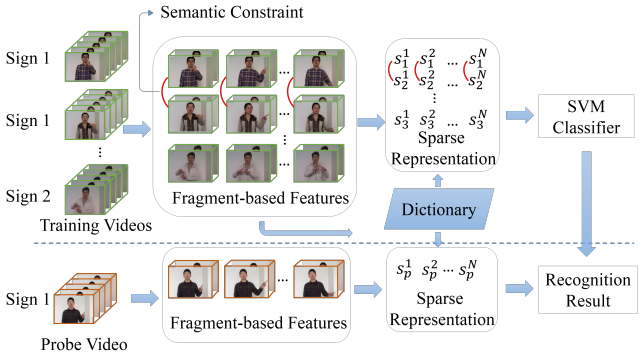

Fig. 1. Framework of our proposed SLR method.

As a sequential problem, currently, the most widely used model in sign language recognition is HMM or its variations[6–9]. There are also some other models for SLR, such as DTW[10], sequential pattern trees[11, 12], trajectory matching[13] and CRF[2]. In this paper, we propose a new framework for signer-independent SLR based on sparse coding and SVM. Sparse coding has been used in many areas of computer vision such as image classification [14], action recognition[15] and signal classification[16]. Many researchers extended conventional sparse coding for different scenarios. Yang et al. [17] modeled sparse coding as a sparsity-constrained robust regression problem. Zheng et al.[18] added Laplacian regularization to sparse coding.

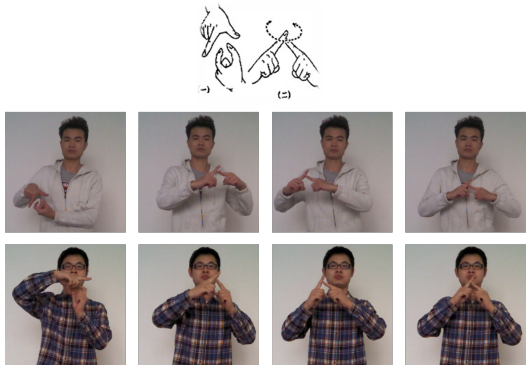

Fig. 2. Example of the sign *Citizen* performed by two different persons, and its standard illustration figure.

To address the signer-independent problem, a sparse coding based framework is proposed as shown in Fig.1. By considering the semantics consistence, the common feature is learnt in a fragment level for the effective sign representation while regardless of the inter-signer variation. Thus the signer-independent codings for the same vocabulary can be generated for robust representation.

The contribution of our work mainly lies in three folds. First of all, a fragment-based feature is designed to characterize the subunits of signs. Secondly, the semantics constrained dictionary learning is proposed to discover the typical common feature. Thirdly, a new sparse coding based framework to tackle signer-independent SLR. To our knowledge, sparse coding is introduced into SLR arera for the first time.

The rest of this paper is organized as follows. Section 2 is the basic idea and formulation of our method. Section 3 is the detailed implementation of the proposed SLR method. Section 4 shows the experimental results to prove the effectiveness of our method. Finally, section 5 is the conclusion.

2. BASIC IDEA AND FORMULATION

In this section, we will first give the overview of the proposed SLR framework, and followed by the mathematical formulation.

2.1. Basic Idea

In real life, the inter-signer variation usually appears in the forms of different motion trajectories and different hand shapes. An example of the same sign performed by different signers is shown in Fig.2. So the features directly extracted from the sign videos of different signers will have big diversity and lead to the wrong classification. The basic idea of our method is to get common features which are robust to different signers. In other words, we want to get the robust feature representation, which only relates to the class of the sign, while being irrelevant to the signer. Therefore, a sparse coding based framework is proposed to tackle SLR, which is capable of handling inter-signer variation by considering on the semantic constraint in dictionary learning stage.

As shown in Fig.1,in our method, first of all, each sign sequence from both training data and the probe is partitioned into multiple fragments equally and the fused feature is extracted accordingly. With the fragment-based feature of training data, a semantics constrained dictionary is learnt.The dictionary can capture the common feature of all the fragments from different signers and it can be seen as the standard basis vectors for sign fragments representation. With this learnt dictionary, any probe sign can be represented by concatenating the sparse coding coefficients of its multiple fragments. Finally, a linear SVM is used for sign classification.

2.2. Formulation

Suppose we have the gallery data $X = [\mathbf{x}_1, ..., \mathbf{x}_n] \in \mathbb{R}^{m \times n}$, which is composed of n data points and each column $\mathbf{x}_i$ is m-dimensional feature vector extracted from a sign. Note that all samples of the same sign might come from different signers. $B = [\mathbf{b}_1, ..., \mathbf{b}_k] \in \mathbb{R}^{m \times k}$ is a dictionary matrix learnt from X and each column of B is a basis vector. Given a probe feature $\mathbf{x}_p$, it can be recovered from the dictionary B as $\mathbf{x}_p = B\mathbf{s}_p + \epsilon$, where $\mathbf{s}_p \in \mathbb{R}^k$ is the reconstructed coefficient vector, ϵ is the residual term. So a sparse representation $\mathbf{s}_p$ of probe feature $\mathbf{x}_p$ is gotten with the dictionary B.

In the traditional sparse coding framework, the dictionary B can be learnt by optimizing:

$$\begin{aligned} &\min_{\mathbf{B},\mathbf{S}} \|X - BS\|_F^2 + \lambda \sum_{i=1}^{n} |\mathbf{s}_i|_1 \\ &s.t. \|\mathbf{b}_d\|^2 \leq c, d = 1, ..., k \ , \end{aligned} \tag{1}$$

where $S = [\mathbf{s}_1, ..., \mathbf{s}_n] \in \mathbb{R}^{k \times n}$ is the coding matrix of gallery data X, regularization parameter λ could control sparsity so that only a sparse number of elements in S is nonzero. $\|\mathbf{b}_i\|^2 \leq c$ is a norm constraint for basis vectors and c is a constant value.

In our method, the target is to learn a dictionary that could capture the common feature of the signs so that the codings represented by this dictionary are insensitive to inter-signer variation. Therefore, we assume that the codings of the features with same semantics performed by different signers should be similar. Based on this assumption, a graph G is constructed and each coding $\mathbf{s}_i$ is a vertices of graph G. The weight W_{ij} of the edges between $\mathbf{s}_i$ and $\mathbf{s}_j$ in graph G is either 0 or 1. If $\mathbf{s}_i$ and $\mathbf{s}_j$ correspond to the same semantics, $W_{ij} = 1$, otherwise, $W_{ij} = 0$. The detailed explanation of semantic constraint is given in Section 3.2.

So our semantic constraint regularization can be defined as $\sum_{i,j} (\mathbf{s}^i - \mathbf{s}^j)^2 W_{ij}$, which characterizes how similar the fragments with same semantics are. So the dictionary that

captures the intrinsic feature of the signs could learnt by incorporating semantic constraint regularization to conventional sparse coding objection function:

$$\min_{B,S} \|X - BS\|_F^2 + \gamma \sum_{i,j} (\mathbf{s}^i - \mathbf{s}^j)^2 W_{ij} + \lambda \sum_{i=1}^{n} |\mathbf{s}_i|_1 \quad (2)$$
$$s.t. \|\mathbf{b}_d\|^2 \leq c, d = 1, ..., k$$

where $\gamma \geq 0$ is a tunable parameter for semantic constraint regularization.

Define a diagonal weight matrix D and $D_{ii} = \sum_{j=1}^{n} W_{ij}$. $L = D - W$ is the Laplacian matrix. Referring to [19], it can be proved that $\frac{1}{2} \sum_{i,j} (\mathbf{s}^i - \mathbf{s}^j)^2 W_{ij} = tr(SLS^T)$. Then (2) can be written as:

$$\min_{B,S} \|X - BS\|_F^2 + \gamma tr(SLS^T) + \lambda \sum_{i=1}^{n} |\mathbf{s}_i|_1 \quad (3)$$
$$s.t. \|\mathbf{b}_d\|^2 \leq c, d = 1, ..., k.$$

Expression (3) is not convex in B and S together but it can be solved by alternatingly optimizing with respect to B and S while fixing the other one. This optimization is similar with that in [18]. Because of the limitation of paper length, detailed optimization algorithm is omitted.

3. IMPLEMENTATION

3.1. Fragment based Feature

Intrinsically, sign language can be regarded as a temporal ensemble of some action subunits. Inspired from this point, a fragment based feature is proposed to characterize the approximate subunits.

First of all, a sign is partitioned into N fragments equally. For each fragment, trajectory feature and hand shape feature are extracted respectively, and then fused into a whole feature vector.

Kinect is taken as the capture device in our work, and the 3D coordinates of 5 skeleton joints (left hand, right hand, left elbow, right elbow and head) provided by Kinect are used for trajectory representation. There are two steps in the trajectory feature extraction: normalization and resampling, as illustrated in Fig.3. For visualization, the 3D coordinates are projected to $x - y$ plane. In the normalization step, all the points are normalized using head point as reference. The points of four joints in the fragment could form four trajectories. Then resampling is performed along each trajectory by equidistant linear interpolation. Finally, we get L equally distributed points in each trajectory. In our implementation, $L = 10$, and the dimension of the trajectory feature for each fragment is 120 (4×10×3).

The hands in each frame can be well segmented with the adaptive skin color model and the depth constraint. HOG[20] feature is extracted from the segmented hand and the dimension of HOG is reduced from 324 to 51 by PCA technique. In each fragment, only one typical frame is selected to describe the appearance of the hand, which is most similar to the average of all the hand regions in the fragment.

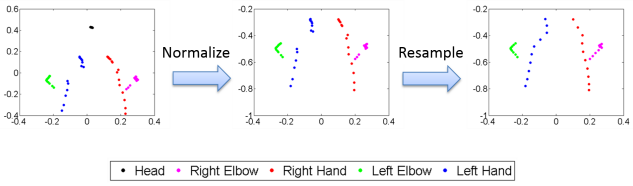

Fig. 3. Fragment-based trajectory feature extraction.

Finallya 171 dimensional vector is generated to describe each fragment by concatenating the trajectory and hand shape feature.

3.2. Semantics Constrained Dictionary Learning

To get a completed dictionary, a large training set is used, which can be denoted as $X^{tr} = \{X_1, X_2, ..., X_n\}$, where n is the total number of training samples. Each sign sample is $X_i = \{\mathbf{x}_i^1, \mathbf{x}_i^2, ..., \mathbf{x}_i^N\}$, where N is the fragment number for each sign. Thus totally $N \times n$ fragments are collected for training the dictionary B with our proposed semantics constrained dictionary learning method. For two fragments $\mathbf{x}_j^i$ and $\mathbf{x}_q^p$, if X_j and X_q correspond to the same sign and $i = p$, then these two fragments have the same semantics. Further, the semantic constraint means that the sparse codings of the fragments with same semantic should be as consistent as possible. The final dictionary is generated by optimizing expression (3).

3.3. Sign Language Recognition

With the learnt dictionary B, each fragment of training data $\mathbf{x}_j^i$ can be coded as the coefficients $\mathbf{s}_j^i$. Then the sparse representation of a sign is generated by concatenating the coefficients of the N fragments, *i.e.*, $S_i = [\mathbf{s}_i^{1T}, \mathbf{s}_i^{1T}, ..., \mathbf{s}_i^{NT}]^T$. In our implementation, the training data for dictionary learning is also used for classifier training with a linear SVM.

For a given probe sign $X_p = \{\mathbf{x}_p^1, \mathbf{x}_p^2, ..., \mathbf{x}_p^N\}$, the sparse representation for each fragment can be gotten by optimizing $\min_{\mathbf{s}_p^i} \{\|\mathbf{x}_p^i - B\mathbf{s}_p^i\|^2 + \lambda|\mathbf{s}_p^i|_1\}$, where $i = 1...N$, and B is the pre-learnt dictionary. Thus the whole feature vector of the probe sign can be represented by concatenating these codings. With the input feature vector, the final recognition result can be obtained by the trained SVM classifier.

4. EXPERIMENTS

In this section, we conduct experiments on two 3D sign language datasets to evaluate the proposed method on both signer-independent and signer-dependent SLR tasks. Below,

Table 1. Parameters evaluation on Dataset II.

Accuracy	$\lambda = 0.01$	$\lambda = 0.05$	$\lambda = 0.1$	$\lambda = 0.5$
$\gamma = 0.01$	0.897	0.930	0.909	0.416
$\gamma = 0.05$	0.754	**0.947**	0.912	0.427
$\gamma = 0.1$	0.370	0.940	0.901	0.394

we first introduce two SL vocabulary datasets and the experiment settings. Then the proposed method is evaluated with comparisons to other baseline methods.

4.1. Datasets and experiment settings

To evaluate the performance of the proposed method, we collect two 3D SL vocabulary datasets with Kinect. Dataset I contains 1000 signs of standard Chinese Sign Language. The data is performed by 7 deaf students and each signer performed once. This dataset is used for the signer-independent test. Dataset II with vocabulary size of 370 is performed by one deaf student with five repetitions, which is designed for the signer-dependent test. In sparse coding related methods, the dictionary size is fixed to be 256. Based on our experience, the number of fragments N is fixed to 5.

4.2. Evaluation on Different parameters

In this section, we evaluate the effectiveness of parameters γ and λ, which are semantic constraint regularizer parameter and sparsity parameter respectively. The experiment is conducted on Dataset II (370 vocabularies) by leave-one-out cross-validation, and the accuracy is the average recognition rate of cross validation. All the experiments in the following are conducted with leave one out cross-validation strategy. The results with different parameter values are listed in Table 1. It shows how the parameters affect the performance. Since the best performance is obtained when $\gamma = 0.05$ and $\lambda = 0.05$, the paraments are fixed as $\gamma = 0.05$ and $\lambda = 0.05$ in the following experiments.

4.3. Performance Evaluation and Comparisons

4.3.1. Baseline methods

We set HMM and Dynamic Time Warping (DTW) as our baseline method because they are two classic models for sign language recognition. The input of HMM is the dense frame-based features composed by concatenating hand shape feature and trajectory feature. HOG is extracted as hand shape feature. In terms of the trajectory, we tried two different features. One is the point-based coordinate feature, referred as HMM(Point). Another is skeleton pairwise feature[21], referred as HMM(Pair). The input data for DTW is same as HMM(Pair). Besides HMM and DTW, the conventional sparse coding algorithm (SC) and Graph sparse coding (GraphSC)[18] are also used as comparisons.

Table 2. Signer-independent SLR on Dataset I.

Group	1	2	3	4	5	6	7	Avg
HMM(Point)	0.55	0.50	0.51	0.55	0.54	0.65	0.49	**0.54**
HMM(Pair)	0.57	0.57	0.59	0.56	0.56	0.61	0.47	**0.56**
DTW	0.62	0.61	0.67	0.61	0.34	0.49	0.16	**0.50**
SC	0.58	0.53	0.49	0.65	0.41	0.64	0.49	**0.53**
GraphSC	0.63	0.62	0.50	0.64	0.50	0.69	0.53	**0.59**
Our Method	0.66	0.68	0.58	0.69	0.56	0.74	0.62	**0.64**

Table 3. Signer-dependent SLR on Dataset II.

Group	1	2	3	4	5	Avg
HMM(Point)	0.910	0.943	0.964	0.935	0.929	**0.936**
HMM(Pair)	0.924	0.945	0.951	0.948	0.913	**0.936**
DTW	0.927	0.935	0.973	0.940	0.875	**0.930**
SC	0.913	0.935	0.951	0.918	0.932	**0.930**
GraphSC	0.912	0.958	0.953	0.937	0.947	**0.941**
Our Method	0.935	0.967	0.959	0.935	0.940	**0.947**

4.3.2. Experiments on Signer-Independent SLR

To validate the effectiveness of the proposed method on signer-independent SLR, we carried out the experiment on Dataset I. Table 2 gives all the recognition accuracy corresponding to each group and also the average accuracy. The results clearly show that our method outperforms others significantly and the proposed framework is effective for signer-independent SLR problem.

4.3.3. Experiments on Signer-Dependent SLR

This subsection gives the experiment on the signer-dependent SLR conducted on Dataset II. The results are given in Table 3. From this table, we can see that even for the signer-dependent SLR test, our method can also have sightly advantage over other methods.

5. CONCLUSION

In this paper we build a novel framework to introduce sparse coding technique into sign language recognition for the first time. A fragment-based feature including both trajectory and hand shape information is designed to characterize the subunits of signs. On these fragment-based features, a dictionary is learnt. To explore the common feature irrelevant to different signers, a semantics constrained dictionary learning algorithm is proposed. The experiments convincingly show that our method outperforms many other methods on signer-independent SLR and it also works well on signer-dependent SLR.

6. ACKNOWLEDGEMENT

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] Jörg Zieren and Karl-Friedrich Kraiss, "Robust person-independent visual sign language recognition," in *Pattern recognition and image analysis*, pp. 520–528. Springer, 2005.

[2] WW Kong and Surendra Ranganath, "Towards subject independent continuous sign language recognition: A segment and merge approach," *Pattern Recognition*, vol. 47, no. 3, pp. 1294–1308, 2014.

[3] Ulrich Von Agris, Daniel Schneider, Jörg Zieren, and K-F Kraiss, "Rapid signer adaptation for isolated sign language recognition," in *Computer Vision and Pattern Recognition Workshop, 2006. CVPRW'06. Conference on*. IEEE, 2006, pp. 159–159.

[4] Ali Farhadi, David Forsyth, and Ryan White, "Transfer learning in sign language," in *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*. IEEE, 2007, pp. 1–8.

[5] Ulrich Von Agris, C Blomer, and K-F Kraiss, "Rapid signer adaptation for continuous sign language recognition using a combined approach of eigenvoices, mllr, and map," in *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*. IEEE, 2008, pp. 1–4.

[6] Yu Zhou, Xiaokang Yang, Weiyao Lin, Yi Xu, and Long Xu, "Hypothesis comparison guided cross validation for unsupervised signer adaptation," in *Multimedia and Expo (ICME), 2011 IEEE International Conference on*. IEEE, 2011, pp. 1–4.

[7] Chunli Wang, Wen Gao, and Shiguang Shan, "An approach based on phonemes to large vocabulary chinese sign language recognition," in *Automatic Face and Gesture Recognition, 2002. Proceedings. Fifth IEEE International Conference on*. IEEE, 2002, pp. 411–416.

[8] Masaru Maebatake, Iori Suzuki, Masafumi Nishida, Yasuo Horiuchi, and Shingo Kuroiwa, "Sign language recognition based on position and movement using multi-stream hmm," in *Universal Communication, 2008. ISUC'08. Second International Symposium on*. IEEE, 2008, pp. 478–481.

[9] M Al-Rousan, Khaled Assaleh, and A Talaa, "Video-based signer-independent arabic sign language recognition using hidden markov models," *Applied Soft Computing*, vol. 9, no. 3, pp. 990–999, 2009.

[10] Haijing Wang, Alexandra Stefan, Sajjad Moradi, Vassilis Athitsos, Carol Neidle, and Farhad Kamangar, "A system for large vocabulary sign search," in *Trends and Topics in Computer Vision*, pp. 342–353. Springer, 2012.

[11] Eng-Jon Ong, Helen Cooper, Nicolas Pugeault, and Richard Bowden, "Sign language recognition using sequential pattern trees," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 2200–2207.

[12] Eng-Jon Ong, Oscar Koller, Nicolas Pugeault, and Richard Bowden, "Sign spotting using hierarchical sequential patterns with temporal intervals," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. IEEE, 2014, pp. 1931–1938.

[13] Xiujuan Chai, Guang Li, Yushun Lin, Zhihao Xu, Yili Tang, Xilin Chen, and Ming Zhou, "Sign language recognition and translation with kinect," in *IEEE Conf. on AFGR*, 2013.

[14] Jianchao Yang, Kai Yu, Yihong Gong, and Thomas Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 1794–1801.

[15] Yan Zhu, Xu Zhao, Yun Fu, and Yuncai Liu, "Sparse coding on local spatial-temporal volumes for human action recognition," in *Computer Vision–ACCV 2010*, pp. 660–671. Springer, 2011.

[16] Ke Huang and Selin Aviyente, "Sparse representation for signal classification," in *NIPS*, 2006, pp. 609–616.

[17] Meng Yang, D Zhang, and Jian Yang, "Robust sparse coding for face recognition," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 625–632.

[18] Miao Zheng, Jiajun Bu, Chun Chen, Can Wang, Lijun Zhang, Guang Qiu, and Deng Cai, "Graph regularized sparse coding for image representation," *Image Processing, IEEE Transactions on*, vol. 20, no. 5, pp. 1327–1336, 2011.

[19] Mikhail Belkin and Partha Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering.," in *NIPS*, 2001, vol. 14, pp. 585–591.

[20] Navneet Dalal and Bill Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*. IEEE, 2005, vol. 1, pp. 886–893.

[21] Jiang Wang, Zicheng Liu, Ying Wu, and Junsong Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 1290–1297.