Isolated Sign Language Recognition with Grassmann Covariance Matrices

HANJIE WANG and XIUJUAN CHAI, Institute of Computing Technology, Chinese Academy of Sciences
XIAOPENG HONG and GUOYING ZHAO, University of Oulu
XILIN CHEN, Institute of Computing Technology, Chinese Academy of Sciences

In this article, to utilize long-term dynamics over an isolated sign sequence, we propose a covariance matrix–based representation to naturally fuse information from multimodal sources. To tackle the drawback induced by the commonly used Riemannian metric, the proximity of covariance matrices is measured on the Grassmann manifold. However, the inherent Grassmann metric cannot be directly applied to the covariance matrix. We solve this problem by evaluating and selecting the most significant singular vectors of covariance matrices of sign sequences. The resulting compact representation is called the *Grassmann covariance matrix*. Finally, the Grassmann metric is used to be a kernel for the support vector machine, which enables learning of the signs in a discriminative manner. To validate the proposed method, we collect three challenging sign language datasets, on which comprehensive evaluations show that the proposed method outperforms the state-of-the-art methods both in accuracy and computational cost.

CCS Concepts: • **Computing methodologies** → **Activity recognition and understanding;**

Additional Key Words and Phrases: Hearing loss, sign language, covariance matrix, Grassmann manifold

ACM Reference Format:
Hanjie Wang, Xiujuan Chai, Xiaopeng Hong, Guoying Zhao, and Xilin Chen. 2016. Isolated sign language recognition with Grassmann covariance matrices. ACM Trans. Access. Comput. 8, 4, Article 14 (May 2016), 21 pages.
DOI: http://dx.doi.org/10.1145/2897735

1. INTRODUCTION

Hearing loss affects 360 million people worldwide [Magariños et al. 2015]. In China, there are 21 million people suffering from hearing loss, according to the statistics of China Disabled Persons' Federation in 2010, and daily communication is a big challenge for them. Sign language is the major means of communication in the deaf community. However, the communication between a deaf person and others is extremely difficult. Deaf persons have to use inefficient texting, very limited simple gestures, or expensive professional interpreters in specific scenarios. Therefore,

This work was partially supported by Microsoft Research Asia and the Natural Science Foundation of China under contracts 61472398 and 61572205, and the Academy of Finland, Fidipro Program of Tekes, and Infotech Oulu.
Authors' addresses: H. Wang, X. Chai, and X. Chen, Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing, 100190, China, and also with Cooperative Medianet Innovation Center, China; emails: hanjie.wang@vipl.ict.ac.cn, chaixiujuan@ict.ac.cn, xlchen@ict.ac.cn; X. Hong and G. Zhao, Faculty of Information Technology and Electrical Engineering, University of Oulu, Finland; emails: xiaopeng.hong@ee.oulu.fi, gyzhao@ee.oulu.fi.

DOI: http://dx.doi.org/10.1145/2897735

automatic sign language recognition (SLR) is needed to bridge the language gap of communication and hence attracts the attention of researchers in the field of pattern recognition [Cooper et al. 2011; Gao et al. 2004; Ong et al. 2012; Ong and Ranganath 2005] in past decades. An automatic SLR system would be designed with the function of translating the visual sign language into written or spoken language in real time. Communication would be much easier in many scenarios with such a system. For example, a doctor who is not familiar with sign language can understand a patient who is deaf. On the other side, a speech-to-sign synthesizer could help deaf persons understand others. Lu and Huenerfauth [2014] developed a sign language animation generation system to improve the accessibility of information for deaf persons.

This article focuses on the recognition of Chinese sign language (CSL). CSL has a history of more than 100 years and keeps growing at both the lexical and morphological levels. At the lexical level, the authorized dictionary named *Chinese Sign Language* [China-Deaf-Assoc. 2003] collected 5,586 words. At the morphological level, the linguistic aspects of CSL have also been studied in the literature (e.g., Shen [1998] and Yang and Fischer [2002]) in past decades. A sign should be represented as a feature to be automatically recognized. In the early years of sign language linguistic studies, Stokoe [2005] indicated that a sign can be described with manual features such as hand shape, position, and motion, as well as nonmanual features such as facial expressions. Different from human body actions [Sanin et al. 2013; Wang et al. 2012b], most signs are only related to the upper body (especially focusing on hands and arms) and subtle facial expressions [Cooper et al. 2011]. Although beyond the scope of this article, some authors have examined facial expressions. For example, Von Agris et al. [2008] examined the facial feature significance for SLR and proved the importance of integrating facial expressions. Koller et al. [2015] also experimentally showed the importance of facial information. Meanwhile, the size of the sign vocabulary is relatively large. For example, the American sign language vocabulary contains more than 5,000 words [Sternberg 1998]. Even for the limited area of weather forecast in German sign language, 1,558 classes of signs are covered in the RWTH-PHOENIX-Weather 2014 dataset [Forster et al. 2014]. As a result, learning the subtle differences of hand motions and shapes between signs remains a big challenge for SLR.

With a large vocabulary of signs, discriminative features extracted to describe signs are extremely important. In SLR, numerous features and representations have been applied. According to the different ways of using the temporal information, the features for SLR can be briefly categorized into two classes: frame-based features and spatiotemporal features. As to the first class, a sign clip is represented by sequential observations. To train a model or gain recognition from these sequential observations, hidden state–based methods like the hidden Markov model (HMM) [Liang and Ouhyoung 1996; Starner et al. 1998; Gao et al. 2004; Pitsikalis et al. 2011] and conditional random fields (CRF) [Yang et al. 2009; Kong and Ranganath 2014] were frequently used. Dynamic time warping (DTW) [Celebi et al. 2013] was also widely used to compute the distance between two sign samples. Different from the frame-based methods, the ST features are extracted from the ST cuboids to describe the long-term dynamics of signs. For example, 3D low-level features have been popular in recent years. Most of them were extended from corresponding 2D features by taking the temporal dimension into account, such as the 3D Harris corner detector [Laptev 2005], the 3D scale-invariant feature transform (SIFT) [Scovanner et al. 2007], the 3D speeded-up robust features (SURF) [Willems et al. 2008], the 3D histogram of oriented gradients (HOG) [Klaser et al. 2008], and sequential patterns (SPs) [Ong et al. 2012, 2014]. Recently, Huang et al. [2015] extracted ST features from a raw video stream by the 3D convolutional neural network (CNN) to recognize 25 signs. However, since the vocabulary is very large in SLR and sign data collection is expensive and difficult,

each sign only has few samples, which is not enough for training a robust CNN. In the early years, bare hands visual SLR was realized by only using a common camera. For example, Starner et al. [1998] tested both desk and wearable computer-based videos in their work. Today, there is a general consensus that no single feature leads to optimal performance, as features from different modalities usually supplement each other. Zafrulla et al. [2010] used both colored gloves and embedded accelerometers to track signers' hand movements and built the CopyCat system. In their following work [Zafrulla et al. 2011], Kinect was applied to the CopyCat system. Our work also applies Kinect and proposes a novel sign representation, which focuses on the fusion of multimodal features and the encoding of long-term dynamics over a sign sequence. More specifically, we use the covariance matrix [Tuzel et al. 2006] to describe the correlations between any two features of frames and between any two feature dimensions.

It is well known that covariance matrices are not in Euclidean space and hence are commonly formulated as points on the Riemannian manifolds [Arsigny et al. 2007; Pennec et al. 2006; Wang et al. 2012a]. However, most existing distances based on the Riemannian metric, such as log-Euclidean distance (LED) [Tuzel et al. 2008; Vemulapalli et al. 2013], cannot globally preserve the topological structure of the points on the manifold by simply mapping them to the tangent space at one point. Therefore, in this work, we measure the proximity of covariance matrices on the Grassmann manifold. However, the inherent Grassmann metric measured by principal angles [Golub and Van Loan 2012; Hamm and Lee 2008] cannot be directly applied to the covariance matrix. Therefore, we tackle the problem by evaluating and selecting the most significant singular vectors of covariance matrices of sign sequences. This results in a compact and discriminative representation, which is called the Grassmann covariance matrix (GCM).

The major contributions of this work are summarized as follows. First, a covariance matrix that is able to fuse multiple features and encode the long-term dynamics is suitably used for sign representation. Second, we present a novel manifold-based method for efficient SLR by proposing the GCM representation with the principal angles-based Grassmann metric to tackle the drawback inherited from the commonly used Riemannian metric. Finally, since publicly available datasets are limited in this area, we collected three datasets to evaluate the proposed method of SLR. To facilitate the research in SLR, we released the three datasets, which are publicly available.[1]

The article is organized as follows. Section 2 describes multimodal features and their covariance representations. Section 3 provides an introduction to the proposed GCM. Section 4 presents experiments and evaluations. Finally, conclusions are given in Section 5.

2. MULTIPLE FEATURES AND COVARIANCE REPRESENTATION

In our work, signs are captured with a Kinect. The data include both the RGB image and depth map. From each frame in a sign video, multiple features including hand shape and body skeleton are extracted. To fuse the features effectively, a covariance matrix serves as the descriptor for a sign sequence. We first introduce the features extracted in each frame before the generation of covariance representation.

2.1. Features

We take two typical features, namely hand shapes and body skeletons, as an example to show the effectiveness of the proposed method. Hereinafter, the appearance feature for hand shape in a segmented hand region is denoted as $\boldsymbol{p}$, and the geometric upper body skeleton feature is used to characterize the motion and denoted as $\boldsymbol{s}$.

[1]http://vipl.ict.ac.cn/homepage/KSL/data4evaluation.html.

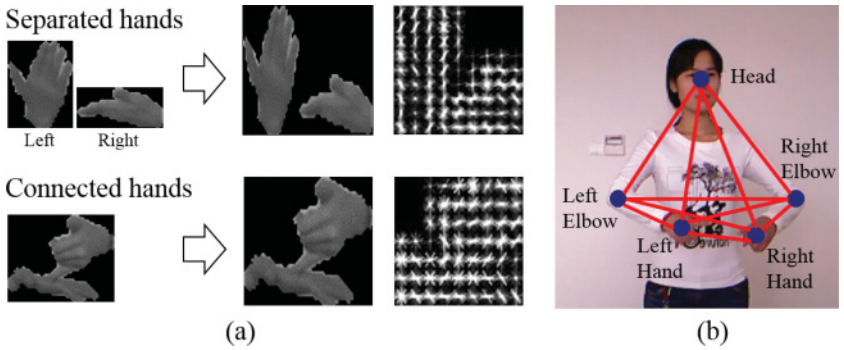

Fig. 1. Two kinds of features. (a) Hand shape is represented by HOG features. (b) Skeleton pairwise features.

In each frame, $\boldsymbol{p}$ and $\boldsymbol{s}$ are concatenated as $\boldsymbol{x}^T = [\boldsymbol{p}^T, \boldsymbol{s}^T]$. For a sign video with M available frames, the feature can be represented as follows:

$$X = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_M] \in \mathbb{R}^{D \times M}, \tag{1}$$

where D is the dimension of the concatenated feature $\boldsymbol{x}$ and M is the total frame number. Representation X is the concatenated feature matrix. Therefore, a set of signs can be represented as $\{(X_i, L_i)_{i=1}^K\}$, where L_i is the sign label and K is the total sign class number. The rest of this section introduces the generation of features $\boldsymbol{p}$ and $\boldsymbol{s}$ in one frame (see illustrations in Figure 1).

Hand shape feature $\boldsymbol{p}$. We extract HOG [Dalal and Triggs 2005] features $\boldsymbol{p}$ from segmented hand shapes in each frame. Hand segmentation, as the preprocessing for extracting feature of hand shape, is utterly important [Erol et al. 2007]. In the early years, unstable RGB-based hand segmentation methods suffered from variant illuminations, cluttered background, and disturbance of human faces. With the depth information from Shotton et al. [2013], a method was proposed to classify each pixel belonging to a specific body part. Consequently, with the hand joints' positions and the depth constraints, hand segmentation in a frame is realized as follows. First, hand regions are roughly determined according to the positions of hand joints. Second, the human face in the same frame is detected with the face detection tool proposed in Yan et al. [2008] to build a user-specific skin model. The self-adaptive skin model performs better than a fixed skin model due to its resistance to variant illuminations and skin colors of different signers. Third, hand candidates are extracted with the skin model in the relatively small regions around hand joints. When performing sign language, hands may cover the face in some cases. Since we have located the human face in the second step, the face will not be selected as hand candidates. Figure 2 presents such an example. Finally, the maximum connected regions are determined as hands. In addition, the left and right hands can be separated or connected in a frame. If two hands are detected to be separated, they will be combined together into one image. Otherwise, if two hands are connected (i.e., only one region containing both hands is detected), this procedure can be omitted. A normalization procedure is executed to the segmented hand images under both situations. Figure 1(a) provides a better understanding of hand shape feature generation.

Body skeleton feature $\boldsymbol{s}$. A pairwise relative position feature [Wang et al. 2012b] is employed for describing the skeleton $\boldsymbol{s}$. Different from body action in Wang et al. [2012b], sign language is only related with joints in the upper body. Therefore, we chose five joints, including the head joint (J_0), left elbow (J_1), left hand (J_2), right elbow

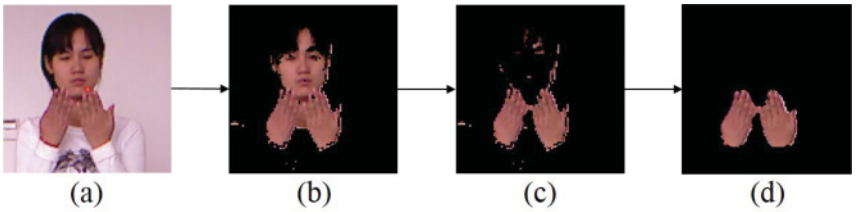

Fig. 2. Hand segmentation in a case when the face and hands are overlapping. (a) Color image. (b) Skin region including face and hands. (c) Most regions of the face are removed by the depth. (d) Optimized hand region.

(J_3), and right hand (J_4). J_i is represented by a 3D coordinate. The distances between 10 ($C_5^2 = 10$) pairs of joints are computed. Considering the variant human sizes, each distance is normalized by dividing the largest one. The distance between two joints is normalized as follows:

$$d_{k,l} = \frac{\|J_k - J_l\|^2}{\max\limits_{i,j\in[0,\ldots,4]} \{\|J_i - J_j\|^2\}}, \tag{2}$$

where J_k and J_l are two different joints ($k, l \in [0, 1, 2, 3, 4], k < l$). The 10 normalized distances are arrayed in predefined orders to serve as 10 dimensions of the skeleton feature $\boldsymbol{s}$. Figure 1(b) provides an illustration of our skeleton pairwise feature.

2.2. Modeling Sign with Covariance Matrix

For the observation sequence X of a sign, its covariance matrix C^* can be computed by Equation (3) based on the following formulation:

$$C^* = \frac{1}{M-1}\sum_{i=1}^{M}(\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^T, \tag{3}$$

where $\boldsymbol{x}_i$ is the feature of the i_{th} frame of the sequence X, M is the total frame number, $\boldsymbol{\mu}$ is the mean of all column vectors $\boldsymbol{x}$, and the size of C^* is $D \times D$. To ensure that C^* is full rank to avoid the singular problem, we add a regularization term to the formulation as

$$C = C^* + \lambda \times I, \tag{4}$$

where I is the identity matrix and λ can be very small (0.001 in our article).

Define $C(i, j)$ as the element in covariance matrix C. To be specific, $C(i, j)$ sums the products of dimensions i and j from all frames in the sign video. Herein, two aspects are analyzed to reveal the advantages of sign modeling with the covariance matrix. In the spatial domain, the dimensions i and j, which may belong to two modalities, are fused simply by a multiplication in each frame to describe their correlations. Hence, covariance provides a natural way of fusing features in the spatial domain (in one frame). In the temporal domain, since $C(i, j)$ is obtained from a statistic over all frames, it is able to resist noises. The noises are usually caused by inaccurate hand tracking or blurred hand shapes. Although noises are inevitable in SLR due to the limited measurement accuracy of Kinect or inaccurate skeleton tracking, covariance is robust in the sense of statistics. In conclusion, first, covariance matrices fuse multiple features that might be correlated. The diagonal elements of the covariance matrix are the variances of feature $\boldsymbol{x}$, and the nondiagonal elements are the correlations between different features. Second, the covariance matrix is able to filter out numerous small noises, which have negative influence on the SLR. In addition, the regularization term ensures that C is full rank even when the duration of a sign is short (i.e., $D > M$).

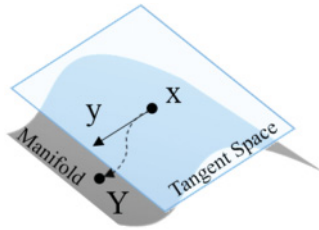

Fig. 3. Illustration of the tangent space. (Reproduced from Harandi et al. [2012]).

Usually, covariance matrices are endowed with Riemannian metrics and thus form a Riemannian manifold. In this manifold, given the specific metric, the geodesic distance between two points is defined as the length of the shortest curve that connects them. The geodesic distance is related to the distance in the tangent space, which is the hyperplane tangent to the surface of the manifold at one point. In practice, the geodesic distance induced by the Riemannian metric is commonly measured by the affine-invariant distance (AID) [Pennec et al. 2006] or LED [Arsigny et al. 2007]. LED is the most popular one. However, the Euclidean space is just locally similar to the manifold, which is a topological space [Tuzel et al. 2008]. Therefore, a logarithm mapping with a universally fixed tangent pole used in the LED does not globally preserve the topological structure of points on the manifold. Figure 3 offers an illustration of the tangent space at point x on the Riemannian manifold. The solid line $x - y$ represents the mapped distance on the tangent space, and the dotted line shows the geodesic starting at x and ending at Y. Tangent spaces preserve only the local structure at the point of an identity matrix (the pole of projection, e.g., x in Figure 3) on the manifold. Hence, only the distances between points and the tangent pole are equal to the true geodesic distances, whereas most distances are measured inaccurately. To alleviate this problem, we turn to the Grassmann manifold for a better proximity measurement of covariance matrices. The following sections briefly review the Grassmann manifold and provide details of the proposed method.

3. GRASSMANN COVARIANCE MATRIX

Orthogonal matrices can be regarded as points on the Grassmann manifold. The covariance matrix is symmetric and positive definite. Hence, there is a corresponding orthogonal matrix, which can be easily computed by singular value decomposition (SVD). Consequently, it is able to measure the proximity of covariance matrices on the Grassmann manifold. This section provides details of the proposed GCM representation and analyzes the advantages for SLR application. First, we offer a brief introduction to the Grassmann manifold.

3.1. Grassmann Manifold Introduction

This section reviews the definition of the Grassmann manifold, the calculation of principal angles, and the corresponding Grassmann distance. Interested readers may refer to Hamm and Lee [2008] for more details.

Grassmann manifold. The Grassmann manifold, $G_{m,D}$, is the set of m-dimension linear subspaces of the R^D. An element of $G_{m,D}$ can be represented by an orthonormal matrix Y of size $D \times m$. $span(Y)$ denotes the subspace spanned by column vectors of Y.

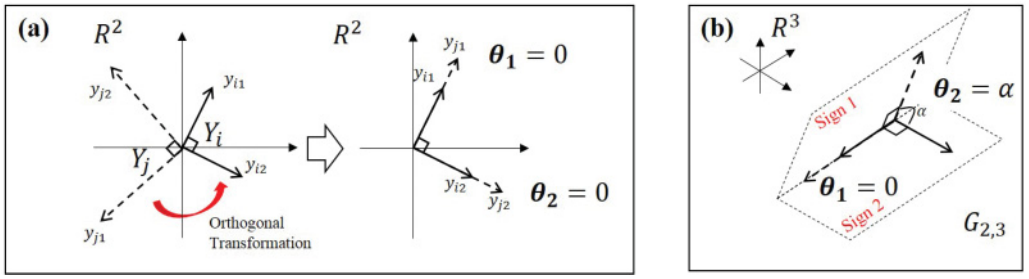

Fig. 4. (a) $G_{2,2}$, an example of case $d = D$ in $G_{d,D}$. Y_i and Y_j are both basis vectors (the vectors of Y_j are slightly stretched to better identify them). The distance between the two subspaces $\mathcal{D}_G^2(Y_i, Y_j)$ is $\theta_1^2 + \theta_2^2 = 0$, as defined in Equation (6). (b) Example when signs are represented as GCMs on $G_{2,3}$. The distance between two signs is measured by the principal angles.

If and only if $span(Y_i) = span(Y_j)$ are the two elements Y_i and Y_j considered to be the same.

Principal angles. Define $\boldsymbol{u}_k$ and $\boldsymbol{v}_k$ as the orthogonal bases of subspaces $span(Y_i)$ and $span(Y_j)$. The principal angles ($0 \leq \theta_1 \leq \cdots \leq \theta_m \leq \pi/2$) between the two subspaces are defined recursively by

$$\cos\theta_k = \max_{\boldsymbol{u}_k \in span(Y_i)} \max_{\boldsymbol{v}_k \in span(Y_j)} \frac{\boldsymbol{u}_k^T \boldsymbol{v}_k}{|\boldsymbol{u}_k|\,|\boldsymbol{v}_k|}, \tag{5}$$

subject to $\boldsymbol{u}_k^T \boldsymbol{u}_k = \boldsymbol{v}_k^T \boldsymbol{v}_k = 1$ and $\boldsymbol{u}_k^T \boldsymbol{u}_p = \boldsymbol{v}_k^T \boldsymbol{v}_p = 0, (p = 1, 2, \ldots, k-1)$.

Grassmann distance. In earlier years, Wong [1967] pointed out that principal angles are related to the geodesic distance by

$$\mathcal{D}_G^2(Y_i, Y_j) = \sum_{k=1}^{m} \theta_k^2. \tag{6}$$

3.2. From Covariance Matrix to GCM

From Section 3.1, one may find that it is still not feasible to apply the intrinsic metric of Grassmann to the covariance matrix straightforwardly. The reasons are as follows. The covariance matrix with a regularization term is a full rank matrix with a size of $D \times D$, where D is the dimension of the concatenated multimodal feature. For m-dimension linear subspaces of the R^D, if $m = D$, the so-called subspace is actually R^D itself. In other words, $span(\bar{Y}_i) = span(\bar{Y}_j) = R^D$. Therefore, $\theta_k = 0$ for all k according to Equation (5), and the distances between any two points are zeros according to Equation (6). This means that the distance between any two covariance matrices is zero. Readers can induce the characteristic easily from the case $G_{2,2}$ in Figure 4(a). When the significant singular vectors of a sign's covariance matrix are extracted, we actually obtain the subspace of the R^D to describe the sign. The signs represented by different subspaces are distinctive. Figure 4(b) gives an example when signs are represented as 2D subspaces in a 3D space. As we know, the representations are two planes, and the distance between the two planes in 3D space can be measured by the angles (θ_1 and θ_2), as shown in Figure 4(b). Therefore, the distance based on the principal angles reflects the two signs' dissimilarity.

As a result, usually when classifying the m-dimension linear subspaces by principal angles, the value of m should be smaller than D in the definition of the Grassmann manifold $G_{m,D}$. In this article, we evaluate and select the most significant singular

vectors of covariance matrices of sign sequences and obtain the resulting GCM. In what follows, a projection metric is used to measure the distances while preserving the topological structure. SLR is fast due to the small dimension of subspace by significant singular vector selection. In addition, the distance defined by principal angles is more intuitive and computationally efficient than the Riemannian distance (Golub and Van Loan [2012]), and thus SLR technology can be more effective when incorporating these principal angles. The generation of GCM is as follows. First, we orthogonalize the symmetric covariance matrices C by SVD. Then, we select the subspace of the orthonormal matrix (i.e., GCM). The formulations are as follows.

For a real symmetric matrix, the C with size $D \times D$ can be decomposed via SVD by

$$C = Y \Sigma Y^T, \tag{7}$$

where $Y = [\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_D]$ is an orthonormal matrix and Σ is a rectangular diagonal matrix with nonnegative real numbers on the diagonal. The first d ($d \ll D$) column vectors of Y contain the most important feature dimensions with large eigenvalues. The parameter d is determined based on the experimental performance. We denote them as $\bar{Y}$, where $\bar{Y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_d]$. The sets of d-dimensional subspaces of the R^D form the Grassmann manifold $G_{d,D}$. The representation $\bar{Y}$ is denoted as GCM. Therefore, given two GCMs $\bar{Y}_i$ and $\bar{Y}_j$, principal angles between them are required for measuring the distance. The principal angles in Equation (5) can be computed from the SVD of $\bar{Y}_i^T \bar{Y}_j$ straightly as follows [Hamm and Lee 2008]:

$$\bar{Y}_i^T \bar{Y}_j = U \Sigma_{ij} V^T, \tag{8}$$

where $U = [\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_d]$, $V = [\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_d]$, and $\Sigma_{ij} = diag(cos\theta_1, \ldots, cos\theta_d)$. With these principal angles $(\theta_1, \ldots, \theta_d)$, the distance can thus be measured by Equation (6).

The improvement by using GCM is twofold. First, the representation is more compact, as we only select the most significant singular vectors of the covariance matrix. Second, compared to the LED metric, the Grassmann metric (by principal angles) is more capable of preserving the topological structure. In SLR, interclass variances are quite smaller and the intraclass variances are larger, especially in user-independent cases with a large vocabulary size. GCM can achieve better performance than the original covariance matrix endowed in the Riemannian manifold, as GCM avoids the inaccurate distance computing of covariance matrices on the approximated tangent space.

3.3. SLR with GCM

For the GCM representations of two signs, the distance can be measured according to the projection metric $\mathcal{M}_p$ [Vemulapalli et al. 2013; Hamm and Lee 2008], which is based on the principal angles, as follows:

$$\mathcal{M}_p(\bar{Y}_i, \bar{Y}_j) = \left(\sum_{k=1}^{d} \sin^2 \theta_k \right)^{1/2} = \left(d - \sum_{k=1}^{d} \cos^2 \theta_k \right)^{1/2}, \tag{9}$$

where $\bar{Y}_i$ and $\bar{Y}_j$ are GCMs and θ_k is one of the d principal angles.Thus, the Grassmann manifold can be connected to Euclidean spaces by using Mercer kernels. Here, the adopted projection kernel is the one most popularly used [Vemulapalli et al. 2013; Hamm and Lee 2008], which is denoted as $\mathcal{K}_{GCM}$ and adjusted to GCM by

$$\mathcal{K}_{GCM}(X_i, X_j) = \mathcal{K}(\bar{Y}_i, \bar{Y}_j) = \|\bar{Y}_i^T \bar{Y}_j\|^2, \tag{10}$$

where X_i, X_j are two sign sequences and $\mathcal{K}$ is the symbol of the kernel.

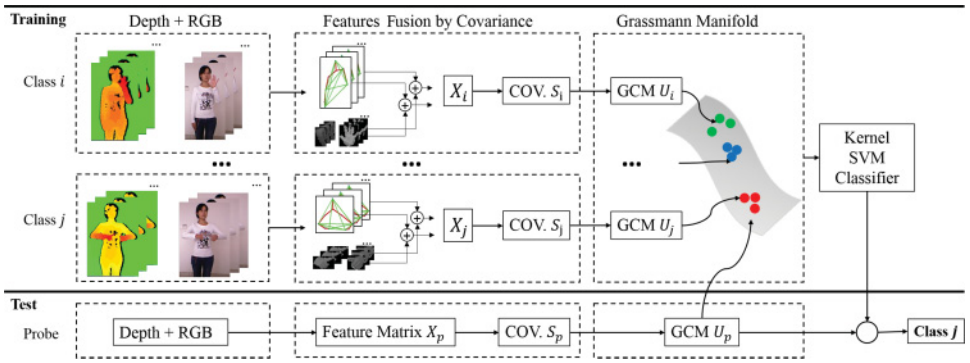

Fig. 5. The method of SLR recognition by GCM representations on the Grassmann manifold.

For the recognition task, an SVM classifier based on the metric $\mathcal{M}_p$ on the Grassmann manifold is learned. The $\mathcal{K}_{GCM}$ can be used as a kernel function in the SVM's context. In implementation, the isolated signs are collected and labeled manually. The corresponding GCMs are also assigned to the same label. We calculate the distances between all GCM pairs on the Grassmann manifold. Suppose that there are N labeled training samples; a large matrix with size $N \times N$ can be obtained and termed as M_{tr}. M_{tr} serves as the input of LibSVM [Chang and Lin 2011] with the kernel_type "PRECOMPUTED"[2] to train the kernel SVM model. Similarly, the distance matrix between the probe and training samples should also be computed before classifying by the trained SVM model. In detail, the similarity of two signs is actually the similarity of two GCMs, whose similarity is calculated via principal angles on the Grassmann manifold. For example, the N training samples are denoted as $[\bar{Y}_1, \bar{Y}_2, \ldots, \bar{Y}_N]$. The distance between samples $\bar{Y}_k$ and $\bar{Y}_l$ can be calculated and denoted as $d_{k,l}$. In the training stage, all N training data calculate the distances between each other to get a feature matrix for SVM training. Each row of the matrix is a feature of a sample. For example, $[d_{k,1}, d_{k,2}, \ldots, d_{k,N}]$ is the feature of $\bar{Y}_k$ and is manually labeled as L_k. In the test stage, given a test sample $\bar{Y}_0$, the distances between training and test data can also be calculated and denoted as $[d_{0,1}, d_{0,2}, \ldots, d_{0,N}]$, which serves as the feature for SVM classification. The trained SVM classifier then gives the label to the $\bar{Y}_0$. Overall, the function shown in Equation (10) is actually a kernel. Figure 5 illustrates the flowchart of employing GCM for SLR.

4. EXPERIMENTS

To evaluate the proposed GCM, we conducted several SLR experiments on three datasets that we collected. Cross validations were applied in experiments on all datasets. Detailed analysis and evaluations including the performance of different features, evaluation of different d, and time comparisons with other methods are provided. Statistical analyses wereperformed to show that the result of the proposed method is statistically significant, especially on the signer-independent dataset. In addition, GCM was evaluated on the published ChaLearn multimodal gesture dataset.

4.1. Data Collection

4.1.1. Capture Setting. We adopted Microsoft Kinect to capture both RGB and depth images. A capture environment was set up and fixed as shown on the left side of

[2]Kernel_type = PRECOMPUTED, svm_type = C_SVC, degree = 3, gamma = 0.0000, coef0 = 0, nu = 0.5, cache_size = 100, C = 1, eps = 0.001, p = 0.1, shrinking = 1, probability = 0, nr_weight = 0, weight_label = NULL, weight = NULL.

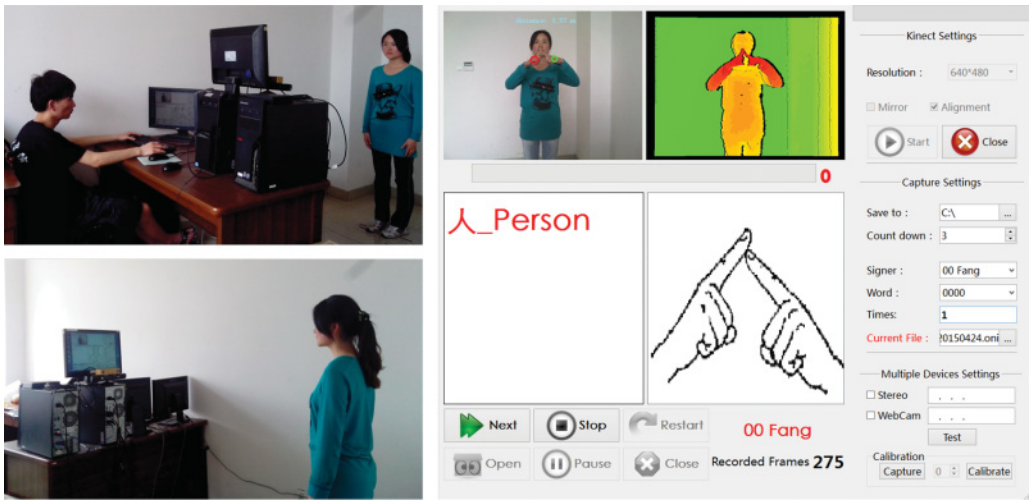

Fig. 6. Capture scenario and the interface of the capture system.

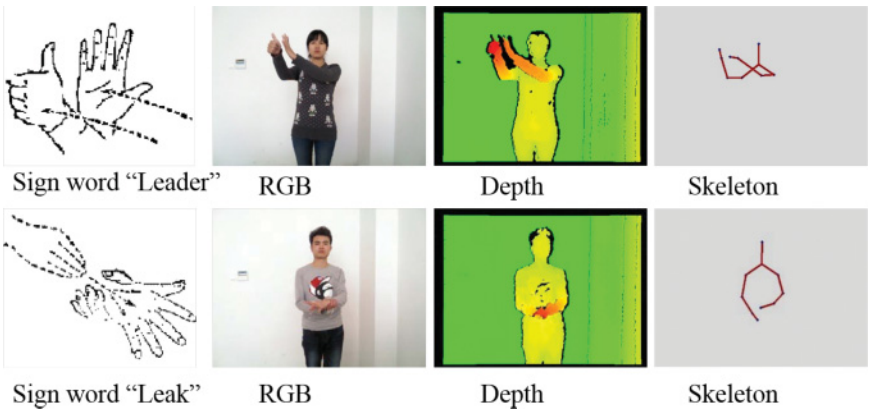

Fig. 7. Samples of the captured data (depth map and skeleton data are also represented as images).

Figure 6. The distance between the signer and Kinect was around 1.5m. Kinect was set approximately 1.5m high from the ground. Multiple deaf persons (both female and male) from Special Education College who use sign language to communicate in daily life were enrolled in the data collection.

The sign data were collected as follows. A signer performed signs in front of the Kinect, and another volunteer recorded these data with a capture tool. The right side of Figure 6 is the interface of the capture tool. In the capture tool, the figure illustration and its corresponding meaning for each sign were shown to instruct for the signing. The RGB and depth images of the capturing region were also displayed on the top of the interface. In addition, the signer's ID and the recording repetition were all listed and embedded in the file name of the saved signing data.

4.1.2. Data Description. The record frame rate was 30 frames per second for RGB images, depth maps, and skeleton data, which are shown in Figure 7. The RGB image had a resolution of 640×480. Depth data records the real depth values, and the depth color image in Figure 7 is just for pretty illustration. Depth maps and RGB images were

Fig. 8. Samples from three datasets. Dataset A: Daily sign language dataset (one signer, five repetitions, sign "PEOPLE"). Dataset B: Large-vocabulary signer-dependent sign language dataset (one signer, three repetitions, sign "INFANT"). Dataset C: Large-vocabulary signer-independent sign language dataset (seven signers, one time, sign "SEAT").

carefully calibrated. The skeleton data was provided by Kinect. Since the movements of sign language focus on the upper body, 10 relative joints were recorded: the head, shoulder center, two shoulders, two elbows, two wrists and two hands, respectively.

Three datasets were collected for different purposes, and the details are given next:

—*Dataset A*: This dataset included 370 daily used signs selected from the standard Chinese sign language [China-Deaf-Assoc. 2003]. These signs were performed by one female signer. Each sign was performed five times. The dataset has 370×5 isolated sign sequences in total. We denoted A_1, A_2, A_3, A_4 and A_5 as group IDs to indicate the 370 signs collected each time.

—*Dataset B*: This dataset was collected for large-vocabulary signer-dependent SLR evaluations. Dataset B included 1,000 isolated signs performed by one male signer. Each sign was performed three times. This set has $1{,}000 \times 3$ isolated sign sequences in total. Similar to the Dataset A, we defined B_1, B_2 and B_3 as group IDs to indicate the 1,000 signs collected each time.

—*Dataset C*: This dataset was collected for large-vocabulary signer-independent SLR evaluations. Dataset C was the most challenging dataset, including 1,000 isolated signs performed by seven different signers. Each signer performed each sign only once. The dataset has $1{,}000 \times 7$ sign sequences in total. We also defined C_1, C_2, C_3, C_4, C_5, C_6 and C_7 as group IDs to indicate the 1,000 signs signed by different signers.

Figure 8 shows some samples from the three datasets.

4.2. Experimental Setup

In our implementation, the HOG feature has 324 dimensions and the pairwise skeleton feature has 10 dimensions. All images of hand regions were resized to 64×64 as shown in Figure 1(a) with the function "cvResize" in OpenCV (*interpolation* = *CV_INTER_LINEAR*). The size of blocks are 32. The size of block strides and cells are both 16, and the number of bins is 9. The value of the skeleton feature $\boldsymbol{s}_i$ is in the range of $[0, 1]$, whereas the value of the HOG feature is roughly in the range of $[-0.0295, 0.0149]$. Therefore, the HOG feature should be normalized. We term the original HOG feature in one frame as $\boldsymbol{h}$ and the normalized HOG feature as $\boldsymbol{p}$. Each

Table I. Experimental Results of Basic Evaluations and the Comparison of Different Methods in Dataset A (370 Vocabulary Set, Signer Dependent)

Probe ID	HMM	DTW	LED_SVM	ARMA_SVM	GCM_SVM
A_1	0.924	0.927	0.938	0.932	**0.949**
A_2	0.945	0.935	0.959	0.968	**0.973**
A_3	0.951	0.973	0.981	0.970	**0.986**
A_4	0.948	0.941	0.946	0.949	**0.951**
A_5	0.913	0.876	0.905	0.935	**0.941**
Ave.	0.936	0.930	0.946	0.941	**0.960**

dimension $\boldsymbol{p}_i$ is computed as follows:

$$\boldsymbol{p}_i = \frac{\frac{\boldsymbol{h}_i}{\sum_{j=1}^{n} \boldsymbol{h}_j} - h_{min}}{h_{max} - h_{min}}, \tag{11}$$

where $h_{max} = 0.0149$ and $h_{min} = -0.0295$ are parameters used for normalization. n is the dimension of the HOG feature (324). Therefore, $\boldsymbol{x}$ in Equation (1) is a feature vector with 334 dimensions, and the value of each dimension is in the range of [0, 1]. In the generation of GCM, there is only one parameter—that is, the number of selected column d from Y to $\bar{Y}$ in Section 3.2. d is set to 10 experimentally. We used LibSVM [Chang and Lin 2011] to build the classifier. The kernel type is "PRECOMPUTED" and the SVM type is "C_SVC." Other parameters keep default.

The proposed method was compared to other four methods: HMM, DTW, LED [Arsigny et al. 2007], and autoregressive moving average (ARMA) [Xu et al. 2014], on the aspects of accuracy and computational cost. LED and ARMA both use SVM as classifiers in our isolated SLR task. In the training and test procedures of HMM and DTW methods, the dense frame-based sequential observation $\boldsymbol{x}$ is taken as the input feature. We exhaustively searched the fine parameters for HMM and DTW for good SLR performance. We use the source code of HMM from Wang et al. [2002]. There are two factors that can directly influence the recognition accuracy of HMM. The first is the number of states Ns, and the second is the number of mixture components Nm. In our experiment, $Ns = 3$, $Nm = 3$, and the Gaussian mixture model serve as the components. In addition, the 334 dimensional features used in HMM are reduced to 61 dimensions through principal component analysis by reserving 99% data energy. We compute the LED metric according to the literature of Arsigny et al. [2007] and classify the covariance matrices' represented signs by SVM. ARMA is used to model a sequence of observations (Equation (2) of Xu et al. [2014]). Each observation is dependent on a hidden state. In Xu et al. [2014], the "ordered patches" are the observations, whereas in our work, features from frames are the observations. A sign is represented as a sequence of observations to build an ARMA model. After sign modeling, we measure the distance between two signs with the Grassmannian distance (refer to Equation (7) of Xu et al. [2014]) and choose kernel SVM as the classifier.

Multiple rounds of leave-one-out cross validation are performed on basic evaluations of each dataset, and the accuracies are averaged over all rounds. For example, in one round to evaluate the group B_1 on dataset B, the model is trained from the data of B_2 and B_3. After all three rounds, an average accuracy score can be computed.

4.3. Experiments on Dataset A

Table I presents the basic evaluations of all methods. In this dataset, GCM outperforms the others with accuracy of 96% on average. In addition, the accuracies in all rounds of cross validation exceed the other methods.

Fig. 9. Some samples of the confused sign pairs.

Table II. Experimental Results of Basic Evaluations and the Comparison of Different Methods in Dataset B (1,000 Vocabulary Set, Signer Dependent)

Probe ID	HMM	DTW	LED_SVM	ARMA_SVM	GCM_SVM
B_1	0.792	0.817	0.833	0.868	**0.907**
B_2	0.864	0.919	0.894	0.915	**0.941**
B_3	0.841	0.874	0.886	0.889	**0.924**
Ave.	0.832	0.870	0.871	0.891	**0.924**

In this daily used signs dataset, only few signs fail with GCM, and most of them are shown in Figure 9. It can be seen that even humans will be confused with most of the pairs. For example, signs 0308 (ZERO) and 0352 (O) are almost the same and hard to be classified except in specific semantic contexts. Signs 0110 (BYE) and 0212 (NO) can only be classified if facial expression or head movement is considered. People always smile when signing BYE and show a disgusted expression when signing NO. The only difference between signs 0168 (QUESTION) and 0169 (QUESTIONNAIRE) is that the ending of QUESTION is a dot while the ending of QUESTIONNAIRE is a circle. The two signs cannot be separated without a more accurate skeleton estimation. The preceding experimental results exactly show that the proposed GCM works in general cases except for these special conditions.

4.4. Experiments on Dataset B

This section presents the experiments conducted on dataset B, whose vocabularly size is 1,000. In addition to the basic performance evaluation using the strategy of cross validation, the experiment to evaluate and determine the parameter d is given. Finally, we evaluate the features and show the computational cost for all methods.

Basic evaluation. Table II shows the results and comparisons in dataset B. It can be seen from the table that GCM_SVM outperforms the other methods. Our method achieves 92.4%, increasing at least by 3.3 percentage points from the results of the other four methods. HMM, as the classical method for SLR, performs worse than the GCM_SVM by 9.2 percentage points on this dataset. Since the interclass variation is tiny and the intraclass variation is large with the 1,000-sign vocabulary, the distances on global tangent space are less accurate using the LED metric. This experimental result confirms that GCM performs better than the corresponding covariance matrices on the Riemannian manifold. Note that the result of HMM is slightly worse than DTW in dataset B, as there are only two samples for training. On the contrary, HMM outperforms DTW in dataset A with more training data, yet it is still worse than GCM_SVM (see Table I). Compared to the result of GCM_SVM on dataset A, although

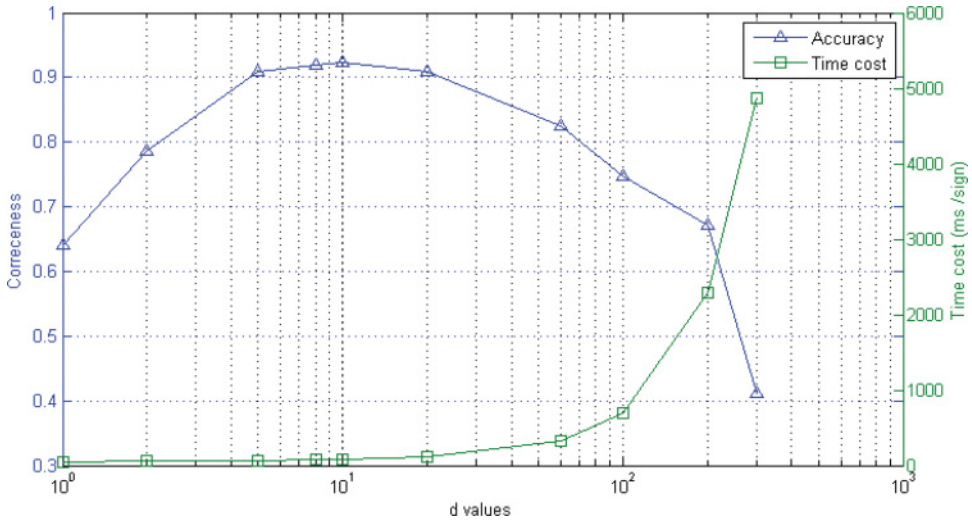

Fig. 10. Accuracy (blue line) and computational cost (green line) with changed $d \in [1, D]$.

the size of vocabulary becomes larger and the model is trained with fewer training data, the accuracy on dataset B is only decreased by 3.6 percentage points.

Evaluation on $\boldsymbol{d}$. How to select an appropriate d is a crucial problem in GCM. In this experiment, a simulated curve of the trend is shown in Figure 10 for different values of d. The highest recognition rate appears near the range $d \in [8, 20]$ when $D = 334$. With larger d, the accuracy decreases while the cost of processing time increases very fast. On the Grassmann manifold, the distance between two subspaces (i.e., two GCMs) is computed based on their principal angles. Two subspaces both with high dimensionality d will share most of the orthogonal bases in the Grassmann manifold $G_{d,D}$. Therefore, given two GCMs, most of the principal angles will be zeros. An extreme example is given in Section 3.2 and Figure 4(a) shows that in $G_{2,2}$, where d is equal to D, the distances between any two subspaces are zeros. Therefore, d is fixed to 10 in our experiments.

Evaluation on features. Although the skeleton pairwise feature $\boldsymbol{s}$ has only 10 dimensions in $\boldsymbol{x}$ (Equation (1)), it is a beneficial complement to the hand shape features. An additional experiment is conducted to test the individual hand shape feature $\boldsymbol{p}$. Without loss of generality, the results of all methods are shown in Figure 11. It is clear that the result of combined feature, with the additive body skeleton information, is significantly better than that of the single hand shape feature for all methods.

Computational cost. For SLR, it is important to consider the cost of processing time, especially for some real-time applications. Computational costs for all methods are listed in Table III. The experiments are conducted on dataset B with 1,000 signs on a regular desktop computer equipped with an Intel Core i7 and 12GB RAM. Since the HMM-based method is realized based on the source code of Wang et al. [2002], a process that they used named *beam search* is also applied in this experiment. In Viterbi searching, a threshold "BeamThreshold" is setin their method. Only the hypothesis whose likelihood is larger than "BeamThreshold" is considered for further growth. Therefore, the searching space is pruned to conserve the computing and memory resources. The computational costs for GCM_SVM is comparable to the ARMA_SVM method. They are almost two times faster than the other methods in the test stage.

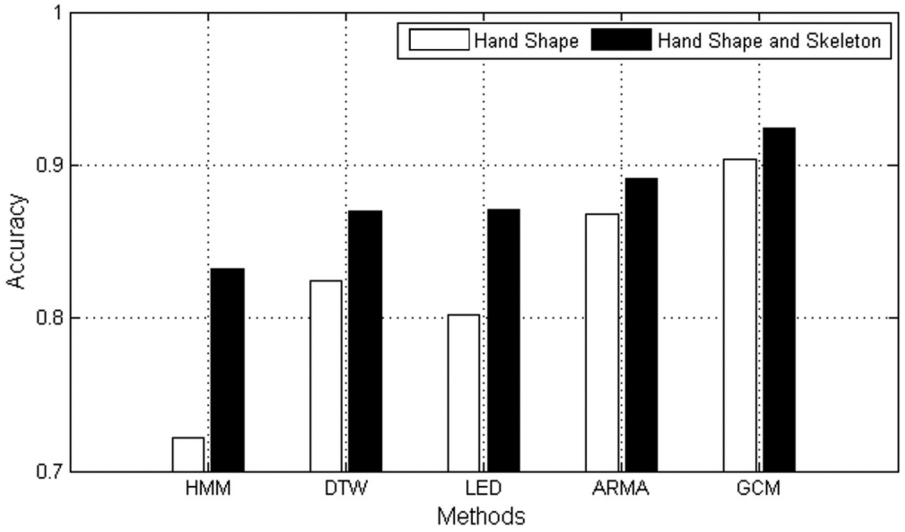

Fig. 11. Evaluation using individual and combined features on dataset B.

Table III. Computational Cost for All Methods on Dataset B

	HMM	DTW	LED_SVM	ARMA_SVM	GCM_SVM (d = 10)
Train (ms/sign)	154	—	985	239	268
Test (ms/sign)	260	1300	316	140	174

Table IV. Experimental Results of Basic Evaluations and the Comparison of Different Methods in Dataset C (1,000 Sign Words, Signer Independent)

Probe ID	HMM	DTW	LED_SVM	ARMA_SVM	GCM_SVM
C_1	0.574	0.617	0.668	0.658	**0.726**
C_2	0.571	0.605	0.633	0.652	**0.735**
C_3	0.587	0.666	0.670	0.666	**0.754**
C_4	0.559	0.606	0.562	0.649	**0.728**
C_5	0.559	0.335	0.645	0.617	**0.703**
C_6	0.610	0.492	0.709	0.714	**0.785**
C_7	0.474	0.162	0.425	0.470	**0.530**
Ave.	0.562	0.498	0.616	0.632	**0.709**

4.5. Experiments on Dataset C

We also conducted experiments on a more challenging dataset containing 1,000 signs performed by seven different signers, and the results are shown in Table IV. Since there are seven different signers, more variants of observations exist. The observed variants are caused by different genders, illuminations, and the variant habits of signing. Some of those variations can be seen in Figure 8(c). In this challenging dataset, the accuracy of our GCM_SVM exceeds the rest of the methods significantly by 7.7 to 14.7 percentage points in all rounds of cross validation without exception.

On dataset C, as well as datasets A and B, statistical analyses are performed to determine whether our results are statistically significant. The p-values are given by the student's t distribution and are listed in Table V. On the signer independent dataset C, it can be seen from the table that the p-values are all less than the significance level ($p < 0.05$). Therefore, the results of GCM_SVM are statistically significant and not just likely chance occurrences in the experiments. In addition, in the case of HMM and

Table V. The *p*-Values Given by the Student's *t* Distribution

Dataset	HMM/GCM	DTW/GCM	LED/GCM	ARMA/GCM
Dataset A	0.033635	0.067322	0.186935	0.224794
Dataset B	0.008631	0.078766	0.034696	0.058853
Dataset C	**0.000598**	**0.008676**	**0.038171**	**0.049690**

DTW, the *p*-values are even smaller than 0.01. Under the signer-dependent settings on dataset B, the superiorities of GCM_SVM are significant when compared to that of LED_SVM and HMM. On dataset A, the *p*-values of HMM/GCM are still less than the significance level.

4.6. Experiments on ChaLearn Dataset

The ChaLearn multimodal gesture dataset covers 20 main gesture categories from an Italian gesture dictionary. It is also collected via Kinect and provides audio, a skeletal model, a user mask, RGB, and depth images. Here, 27 signers appear, and 13,858 gesture samples are recorded in the dataset. The focus of the dataset is on signer-independent multimodal feature-based gesture learning. The signers perform the gestures continuously in three separated sets, namely for development, validation, and test. Please refer to the literature of Escalera et al. [2013] for more details.

To evaluate the performance of the continuous gesture sequence in the ChaLearn dataset, the Levenshtein distance between the recognition result and the ground truth is computed. The distance is also known as the edit distance, which is a criterion of quantifying the dissimilarity of two strings. The evaluation score is the sum of the Levenshtein distances for all sentences, divided by the total number of signs in the ground truth. For simplicity, we use the abbreviation "LD" for *Levenshtein distance score* in this article. The LD can be computed as follows:

$$LD = \frac{Ins + Del + Sub}{N}, \tag{12}$$

where, Ins, Del, and Sub correspond to the insertion, deletion, and substitution errors, respectively, and N is the total number of test signs. Besides the LD, we also provide recall and precision for comprehensive analyses.

The segmentation in the ChaLearn dataset is relatively easy, as it contains audio information, and most signers tend to put down their hands to rest positions before and after making a gesture. We do not use the provided segmentations for development and validation data. Instead, we designed a segmentation based on both audio and hand positions. In each frame, the distance between the head joint and two hands is denoted as D_{hand}, and the distance between the head joint and body center joint is denoted as D_{center}. When the rate $r = D_{hand}/D_{center}$ is larger than a threshold 1.3, the hands are most probably in rest positions and the frame belongs to a nongesture period (see Figure 12(a)). However, several unintentional movements result in failed judgments (e.g., dressing one's hair in Figure 12(b)). Therefore, audio-based video segmentation is carried out. We use the real-time endpoint detection algorithm proposed by Zhou and Ji [2010]. Finally, visual- and audio-based segmentations are merged to obtain the video segmentation.

According to the organizers' test protocol, our GCM model is trained on development and validation sets and evaluated on the test set. In our implementation, we define two channels (i.e., the visual feature channel (skeleton and hand posture) and audio feature channel (voice)) that are processed separately, and their results are fused in the decision stage. The dimensions of visual and audio cues are 374 and 36 (Mel-frequency cepstral coefficients along with their first and second derivatives), respectively. The best fusing weight is 0.3 for the audio cue and 0.7 for the visual cue. The detailed precision-recall curve is given in Figure 13 with different fusing weights. Table VI

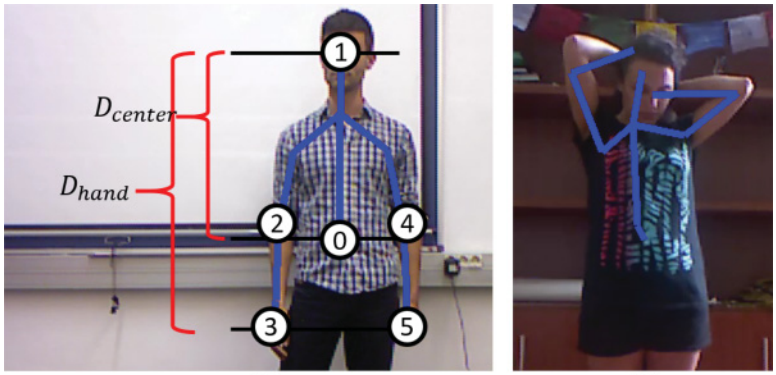

Fig. 12. Visual-based video segmentation. (a) Distance D_{hand} and D_{center}. (b) Failure judgment based on visual-based segmentation.

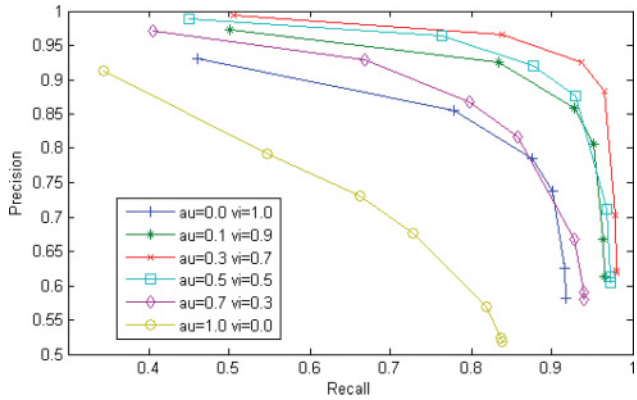

Fig. 13. Precision-recall with different fusion weights of audio and visual channels (au, audio; vi, visual).

Table VI. Evaluations and Comparisons on the ChaLearn Dataset

Methods	LD	Recall (%)	Precision (%)
IVA MM (Rank 1) [Wu et al. 2013]	0.12756	—	—
W WEIGHT (Rank 2) [Escalera et al. 2013]	0.15387	—	—
ET (Rank 3) [Bayer and Silbermann 2013]	0.16813	—	—
LRS (Rank 6) [Neverova et al. 2013]	0.1773	89.39	90.72
GCM	**0.1173**	**94.02**	**93.07**

gives the experimental results of GCM, as well as the top-ranked methods (rank 6 provides recall and precision). The test score is the LD suggested by the organizers. It can be seen from Table VI that the result of GCM is comparable to and even exceeds that of the rank 1 team. The method proposed by rank 6 achieves 89.39% recall and 90.72% precision in Neverova et al. [2013], which are both lower than ours.

Recently, we noticed that Pitsikalis et al. [2015] achieved a quite low LD of 0.0667. The concrete performance comparison is given in Table VII. In their work, activity detection, especially voice activity detection, is helpful. However, in our method, the audio feature is not fully explored, and it is also beyond the scope of this work. Clearly, by pure visual features, our performance exceeds that of Pitsikalis et al. The LD with

Table VII. Evaluations with Individual Features on the ChaLearn Dataset

Methods	Audio	Skeleton	Hand Posture	S+H	All
Pitsikalis et al.	**0.128**	0.509	0.798	—	0.0667
GCM	0.452	0.509	**0.366**	**0.279**	0.1173

Note: We use *LD* instead of *accuracy*, which is given in Pitsikalis et al. [2015].

single hand posture in Pitsikalis et al. [2015] is almost twice as much as that in the GCM model.

5. CONCLUSION

To help smooth communication between those who are deaf and those with whom they are communicating, automatic SLR is needed to translate sign language into written or spoken language. This article proposes an efficient method to recognize Chinese sign language. In addition, the algorithm can also be extended to the recognition of other sign languages.

Technically, a sign language word can be represented as a sequence of body motion, especially focusing on the hands and arms. We propose modeling dynamics with covariance matrices, which can naturally fuse multimodal features of a sign language word. By selecting the most significant singular vectors from a covariance matrix, a sign word can be further represented as a compact GCM. Therefore, the classification of signs is converted to measure the distance between probe and gallery signs. We comprehensively evaluate the proposed GCM_SVM and compare it with others, including HMM, DTW, LED_SVM, and ARMA_SVM. The experimental results show that GCM_SVM outperforms the others with higher accuracy and is less time consuming. The additional statistical tests show that our results are statistically significant.

The covariance matrix takes the long-term correlations into consideration, yet is limited in describing the temporal information of a sign sequence. Multiple GCMs may be extracted from a single sign sequence to compensate for the temporal information to some extent in our future work.

Actually, GCM representation can be extended to realize continuous SLR. As GCM is calculated over time, it naturally reduces the temporal resolution. To overcome the problem, a sliding window will be used to segment the continuous sequence in the temporal domain. We will classify the segmented subsequence in each sliding window, where the middle frame is to be labeled. Consequently, a continuous sign sequence can be labeled frame by frame. We will also optimize the labeling through Viterbi-like decoding and infer good sign spotting.

Based on the proposed method, a real-time SLR system has been developed. In future work, a field test will be conducted to evaluate the system. The scenarios could be hospitals, supermarkets, or classrooms. It is our hope that feedback from the user studywill help us improve the SLR method and system.

REFERENCES

Vincent Arsigny, Pierre Fillard, Xavier Pennec, and Nicholas Ayache. 2007. Geometric means in a novel vector space structure on symmetric positive-definite matrices. *SIAM Journal on Matrix Analysis and Applications* 29, 1, 328–347.

Immanuel Bayer and Thierry Silbermann. 2013. A multi modal approach to gesture recognition from audio and video data. In *Proceedings of the 15th ACM International Conference on Multimodal Interaction (ICMI'13)*. ACM, New York, NY, 461–466.

Sait Celebi, Ali Selman Aydin, Talha Tarik Temiz, and Tarik Arici. 2013. Gesture recognition using skeleton data with weighted dynamic time warping. In *Proceedings of the 8th International Joint Conference on Computer Vision, Imaging, and Computer Graphics Theory and Applications (VISAPP'13)*. 620–625.

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* 2, 3, 27.

China-Deaf-Assoc. 2003. *Chinese Sign Language (in Chinese)*. Huaxia Publishing House. ISBN: 9787508030050

Helen Cooper, Brian Holt, and Richard Bowden. 2011. Sign language recognition. In *Visual Analysis of Humans*. Springer, 539–562.

N. Dalal and B. Triggs. 2005. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition*, Vol. 1. IEEE, Los Alamitos, CA, 886–893.

Ali Erol, George Bebis, Mircea Nicolescu, Richard D. Boyle, and Xander Twombly. 2007. Vision-based hand pose estimation: A review. *Computer Vision and Image Understanding* 108, 1C2, 52–73.

Sergio Escalera, Jordi Gonzàlez, Xavier Baró, Miguel Reyes, Oscar Lopes, Isabelle Guyon, Vassilis Athitsos, and Hugo Escalante. 2013. Multi-modal gesture recognition challenge 2013: Dataset and results. In *Proceedings of the 15th International Conference on Multimodal Interaction (ICMI'13)*. ACM, New York, NY, 445–452.

Jens Forster, Christoph Schmidt, Oscar Koller, Martin Bellgardt, and Hermann Ney. 2014. Extensions of the sign language recognition and translation corpus RWTH-PHOENIX-weather. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC'14)*. 1911–1916.

Wen Gao, Gaolin Fang, Debin Zhao, and Yiqiang Chen. 2004. Transition movement models for large vocabulary continuous sign language recognition. In *Proceedings of the 6th IEEE International Conference on Automatic Face and Gesture Recognition*. IEEE, Los Alamitos, CA, 553–558.

Gene H. Golub and Charles F. Van Loan. 2012. *Matrix Computations*. Vol. 3. JHU Press. ISBN: 9781421407944

Jihun Hamm and Daniel D. Lee. 2008. Grassmann discriminant analysis: A unifying view on subspace-based learning. In *Proceedings of the 25th International Conference on Machine Learning (ICML'08)*. 376–383.

Mehrtash Tafazzoli Harandi, Conrad Sanderson, Arnold Wiliem, and Brian C. Lovell. 2012. Kernel analysis over Riemannian manifolds for visual recognition of actions, pedestrians and textures. In *Proceedings of the Workshop on the Applications of Computer Vision (WACV'12)*. 433–439.

Jie Huang, Wengang Zhou, Houqiang Li, and Weiping Li. 2015. Sign language recognition using 3D convolutional neural networks. In *Proceedings of the 2015 IEEE International Conference on Multimedia and Expo (ICME'15)*. IEEE, Los Alamitos, CA, 1–6.

Alexander Klaser, Marcin Marszalek, and Cordelia Schmid. 2008. A spatio-temporal descriptor based on 3D-gradients. In *Proceedings of the British Machine Vision Conference (BMVC'08)*. 275:1–275:10.

Oscar Koller, Jens Forster, and Hermann Ney. 2015. Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. *Computer Vision and Image Understanding* 141, 108–125.

W. W. Kong and S. Ranganath. 2014. Towards subject independent continuous sign language recognition: A segment and merge approach. *Pattern Recognition* 47, 3, 1294–1308.

Ivan Laptev. 2005. On space-time interest points. *International Journal of Computer Vision* 64, 2–3, 107–123.

Pengfei Lu and Matt Huenerfauth. 2014. Collecting and evaluating the CUNY ASL corpus for research on American sign language animation. *Computer Speech and Language* 28, 3, 812–831.

Marta Magariños, Marta Milo, and Isabel Varela-Nieto. 2015. Editorial: Aging, neurogenesis and neuroinflammation in hearing loss and protection. *Frontiers in Aging Neuroscience* 7, 1–2.

Natalia Neverova, Christian Wolf, Giulio Paci, Giacomo Sommavilla, Graham W. Taylor, and Florian Nebout. 2013. A multi-scale approach to gesture detection and recognition. In *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW'13)*. IEEE, Los Alamitos, CA, 484–491.

Eng-Jon Ong, Helen Cooper, Nicolas Pugeault, and Richard Bowden. 2012. Sign language recognition using sequential pattern trees. In *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'12)*. IEEE, Los Alamitos, CA, 2200–2207.

Eng-Jon Ong, Oscar Koller, Nicolas Pugeault, and Richard Bowden. 2014. Sign spotting using hierarchical sequential patterns with temporal intervals. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1923–1930.

Sylvie C. W. Ong and Suhas Ranganath. 2005. Automatic sign language analysis: A survey and the future beyond lexical meaning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27, 6, 873–891.

Xavier Pennec, Pierre Fillard, and Nicholas Ayache. 2006. A Riemannian framework for tensor computing. *International Journal of Computer Vision* 66, 1, 41–66.

Vassilis Pitsikalis, Athanasios Katsamanis, Stavros Theodorakis, and Petros Maragos. 2015. Multimodal gesture recognition via multiple hypotheses rescoring. *Journal of Machine Learning Research* 16, 1, 255–284.

Vassilis Pitsikalis, Stavros Theodorakis, Christian Vogler, and Petros Maragos. 2011. Advances in phonetics-based sub-unit modeling for transcription alignment and sign language recognition. In *Proceedings of the 2011 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW'11)*. IEEE, Los Alamitos, CA, 1–6.

R.-H. Liang and M. Ouhyoung. 1996. A sign language recognition system using hidden Markov model and context sensitive search. In *Proceedings of the ACM Symposium on Virtual Reality and Technology*. 59–66.

Andres Sanin, Conrad Sanderson, Mehrtash T. Harandi, and Brian C. Lovell. 2013. Spatio-temporal covariance descriptors for action and gesture recognition. In *Proceedings of the 2013 IEEE Workshop on Applications of Computer Vision (WACV'13)*. IEEE, Los Alamitos, CA, 103–110.

Paul Scovanner, Saad Ali, and Mubarak Shah. 2007. A 3-dimensional sift descriptor and its application to action recognition. In *Proceedings of the 15th International Conference on Multimedia*. ACM, New York, NY, 357–360.

Yu Lin Shen. 1998. Shouyu Xingzhi Fenxin [Analyzing qualities of sign language]. *Teshu Jiaoyu Yanjiu [Research on Special Education]* 2, 6–10.

Jamie Shotton, Toby Sharp, Alex Kipman, Andrew Fitzgibbon, Mark Finocchio, Andrew Blake, Mat Cook, and Richard Moore. 2013. Real-time human pose recognition in parts from single depth images. *Communications of the ACM* 56, 1, 116–124.

Thad Starner, Joshua Weaver, and Alex Pentland. 1998. Real-time American sign language recognition using desk and wearable computer based video. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20, 12, 1371–1375.

Martin L. A. Sternberg. 1998. *American Sign Language*. HarperCollins. ISBN: 978-0062716088

William C. Stokoe. 2005. Sign language structure: An outline of the visual communication systems of the American deaf. *Journal of Deaf Studies and Deaf Education* 10, 1, 3–37.

Oncel Tuzel, Fatih Porikli, and Peter Meer. 2006. Region covariance: A fast descriptor for detection and classification. In *Proceedings of the 9th European Conference on Computer Vision (ECCV'06)*. 589–600.

Oncel Tuzel, Fatih Porikli, and Peter Meer. 2008. Pedestrian detection via classification on Riemannian manifolds. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30, 10, 1713–1727.

Raviteja Vemulapalli, Jaishanker K. Pillai, and Rama Chellappa. 2013. Kernel learning for extrinsic classification of manifold features. In *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'13)*. 1782–1789.

Ulrich Von Agris, Moritz Knorr, and Karl-Friedrich Kraiss. 2008. The significance of facial features for automatic sign language recognition. In *Proceedings of the 8th IEEE International Conference on Automatic Face and Gesture Recognition (FG'08)*. IEEE, Los Alamitos, CA, 1–6.

Chunli Wang, Wen Gao, and Shiguang Shan. 2002. An approach based on phonemes to large vocabulary Chinese sign language recognition. In *Proceedings of the 5th IEEE International Conference on Automatic Face and Gesture Recognition*. IEEE, Los Alamitos, CA, 411–416.

Jiang Wang, Zicheng Liu, Ying Wu, and Junsong Yuan. 2012b. Mining actionlet ensemble for action recognition with depth cameras. In *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'12)*. 1290–1297.

Ruiping Wang, Huimin Guo, Larry S. Davis, and Qionghai Dai. 2012a. Covariance discriminative learning: A natural and efficient approach to image set classification. In *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'12)*. 2496–2503.

Geert Willems, Tinne Tuytelaars, and Luc Van Gool. 2008. An efficient dense and scale-invariant spatio-temporal interest point detector. In *Computer Vision—ECCV 2008*. Lecture Notes in Computer Science, Vol. 5303. Springer, 650–663.

Yung-Chow Wong. 1967. Differential geometry of Grassmann manifolds. *Proceedings of the National Academy of Sciences of the United States of America* 57, 3, 589.

Jiaxiang Wu, Jian Cheng, Chaoyang Zhao, and Hanqing Lu. 2013. Fusing multi-modal features for gesture recognition. In *Proceedings of the 15th ACM International Conference on Multimodal Interaction (ICMI'13)*. ACM, New York, NY, 453–460.

Chunyan Xu, Tianjiang Wang, Junbin Gao, Shougang Cao, Wenbing Tao, and Fang Liu. 2014. An ordered-patch-based image classification approach on the image Grassmannian manifold. *IEEE Transactions on Neural Networks and Learning Systems* 25, 4, 728–737.

Shengye Yan, Shiguang Shan, Xilin Chen, and Wen Gao. 2008. Locally assembled binary (LAB) feature with feature-centric cascade for fast and accurate face detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'08)*. IEEE, Los Alamitos, CA, 1–7.

H.-D. Yang, S. Sclaroff, and S.-W. Lee. 2009. Sign language spotting with a threshold model based on conditional random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31, 7, 1264–1277.

Jun Hui Yang and Susan Fischer. 2002. Expressing negation in Chinese sign language. *Sign Language and Linguistics* 5, 2, 167–202.

Zahoor Zafrulla, Helene Brashear, Thad Starner, Harley Hamilton, and Peter Presti. 2011. American sign language recognition with the Kinect. In *Proceedings of the 13th International Conference on Multimodal Interfaces*. ACM, New York, NY, 279–286.

Zahoor Zafrulla, Helene Brashear, Pei Yin, Peter Presti, Thad Starner, and Harley Hamilton. 2010. American sign language phrase verification in an educational game for deaf children. In *Proceedings of the 2010 20th International Conference on Pattern Recognition (ICPR'10)*. IEEE, Los Alamitos, CA, 3846–3849.

Ming-Zhong Zhou and Li-Xin Ji. 2010. Real-time endpoint detection algorithm combining time-frequency domain. In *Proceedings of the 2010 2nd International Workshop on Intelligent Systems and Applications (ISA'10)*. IEEE, Los Alamitos, CA, 1–4.

Received May 2015; revised February 2016; accepted February 2016