Iterative Reference Driven Metric Learning for Signer Independent Isolated Sign Language Recognition

Fang Yin[1,2,3](✉), Xiujuan Chai[1,2,3], and Xilin Chen[1,2,3]

[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] Cooperative Medianet Innovation Center, Beijing, China
{fang.yin,xiujuan.chai,xilin.chen}@vipl.ict.ac.cn

Abstract. Sign language recognition (SLR) is an interesting but difficult problem. One of the biggest challenges comes from the complex inter-signer variations. To address this problem, the basic idea in this paper is to learn a generic model which is robust to different signers. This generic model contains a group of sign references and a corresponding distance metric. The references are constructed by signer invariant representations of each sign class. Motivated by the fact that the probe samples should have high similarities with their own class references, we aim to learn a distance metric which pulls the samples and their true sign classes (references) closer and push away the samples from the false sign classes (references). Therefore, given a group of references, a distance metric can be exploited with our proposed Reference Driven Metric Learning (RDML). In a further step, to obtain more appropriate references, an iterative manner is conducted to update the references and distance metric alternately with iterative RDML (iRDML). The effectiveness and efficiency of the proposed method is evaluated extensively on several public databases for both SLR and human motion recognition tasks.

Keywords: Sign language recognition · Signer independent · Inter-signer variations · Metric learning · Human motion recognition

1 Introduction

As a key technology to help breaking the communication barrier between the deaf and the hearing, SLR has become an important research area in computer vision. Over the last twenty years, SLR has made some progresses [1–4]. However, most of the researches focused on signer dependent situation, in which the signer of the probe has been seen in the training set. In real applications, the performance will decrease dramatically when the user is new to the system. Since collecting enough training data from each new signer to retrain the SLR model is not

B. Leibe et al. (Eds.): ECCV 2016, Part VII, LNCS 9911, pp. 434–450, 2016.
DOI: 10.1007/978-3-319-46478-7_27

realistic, the signer independent SLR is an urgent problem for the practice of SLR technique.

In this paper, the signer independent isolated SLR is tackled by learning a generic model which is composed of the references for all the sign classes and a corresponding distance metric. Given a group of references, the corresponding distance metric could be learnt by constraining that each sample should be closer to the reference of its own class than any other references. This procedure is realized by our proposed Reference Driven Metric Learning (RDML) algorithm. To obtain the appropriate references, iterative RDML (iRDML) is adopted. In each iteration, the constraint between samples and references could also be used to optimize the references one-by-one with current distance metric. When the iterative optimization is convergent, the final model (references and the distance metric) is robust to different signers because it is derived from the multiple signers' data and captures the generic characteristics. Since each class is represented by a single generic reference, the probe only needs to compare with all the references instead of the training samples. The time cost in recognition stage of our method is rather smaller than the conventional sample-based methods.

The contribution of our work mainly lies in the following three aspects. Firstly, a new framework for signer independent isolated SLR is proposed. In this framework, inter-signer variations are handled by learning a generic model which is robust to different signers. Secondly, RDML is proposed to learn a distance metric based on given references by constraining the distances between all training samples and the references. Thirdly, we propose an iterative manner to optimize references and distance metric alternatively so that a group of more appropriate generic references and the corresponding distance metric could be gotten.

The remainder of this paper is organized as follows: Sect. 2 briefly reviews the related work. Section 3 introduces our proposed method. Section 4 gives the details of the implementation on SLR. The experimental results are presented in Sects. 5 and 6 concludes the paper.

2 Related Work

In this section, we will briefly review the related work in two areas: signer independent SLR and metric learning.

2.1 Signer Independent SLR

There are broadly two kinds of solutions for signer independent SLR problem. One is signer adaptation, i.e. using the data of the new signer to adapt the previous model. The other is generic model, which means only one robust model is used for different signers. Of course, the signer invariant feature extraction also belongs to this category. In the first category, borrowing from speech recognition, Agris et al. [5] used Maximum Likelihood Linear Regression (MLLR)

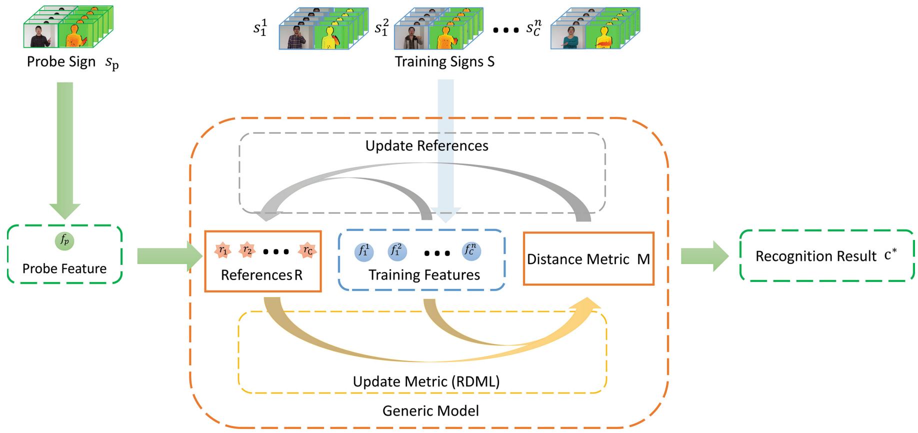

Fig. 1. Framework of signer independent SLR with the proposed iRDML. The sign s_i^n stands for the nth sample for sign class i and f_i^n is the feature extracted from sign s_i^n. r_i is the reference learnt for sign class i.

[6] and Maximum A Posteriori (MAP) [7] for signer adaptation. Later, they extended their work to continuous SLR [8] through combining the eigenvoices [9], MLLR and MAP. Farhadi et al. [10] introduced transfer learning to sign language recognition and addressed signer independence. They designed a comparative feature which is both discriminative and semantically similar based on the assumption that segments of different signs look similar to one another. They obtained recognition rate of 64.2 % on a new signer with 90 dictionary words. In the second category, most of them tried to address signer independence implicitly and concentrated on the inter-signer variations caused by the standing positions of the signers, the signers' heights and different movement epentheses. Zieren and Kraiss [11] used the features normalized for signer independence and adopted HMM to classify. With six different signers, they reached accuracy of 95.0 % and 87.8 % with vocabulary size of 6 and 18 respectively. Shanableh and Assaleh [12] filtered out signer dependent information by encapsulating the movements of the segmented hands in a bounding box. Kong and Ranganath [13] realized signer independent recognition on continuous sign language. They removed movement epenthesis (ME) by using a segment and merge approach to decrease the inter-signer variations in ME and used a two-layer CRF classifier for recognition. The proposed method in this paper also belongs to the second category and our target is to learn a generic model which is robust to different signers. Our method focuses on the inter-signer variations caused by variant movements or hand shapes when performing signs of different signers.

2.2 Metric Learning

Metric learning is widely used in many areas in machine learning, such as image classification [14], ranking [15,16] and kinship verification [17]. It can be categorized into supervised metric learning, unsupervised metric learning and semi-supervised metric learning. In supervised metric learning, most of the methods constrain the relationships between the training samples. Xing et al. [18] formulated the problem as a constrained convex programming algorithm. In [19], a Mahalanobis distance was learnt from the information-theoretic perspective by minimizing the differential relative entropy. RCA introduced in [20] aimed to find a transformation that amplifies relevant variability and suppresses irrelevant variability. Weinberger et al. [21] proposed large margin nearest neighbor classification (LMNN) and the object is to pull the data with same labels closer while pushing data with different labels far apart. Chai et al. [22] extended LMNN by introducing local mean vectors. In unsupervised metric learning, the geometric feature of the data is explored and preserved. The typical methods of unsupervised metric learning include principal component analysis (PCA), locally linear embedding (LLE), locality preserving projections (LPP) and Laplacian eigenmap (LE). There are also some semi-supervised metric learning algorithms fusing both supervised and unsupervised metric learning. Wang et al. [23] used PCA as the unsupervised constraint term and integrated it into RCA. Baghshah and Shouraki [24] considered topological structure of data using the idea of LLE. Niu et al. [25] maximized the entropy on labelled data and minimized the entropy on unlabelled data following entropy regularization.

In this paper, we propose a new supervised metric learning method. Different from the algorithms mentioned above, the metric is learnt by constraining the distances between the training samples and the generic references of classes.

3 Proposed Method

3.1 Basic Idea

Figure 1 is the framework of our signer independent SLR method. Let $S = \{s_1^1, s_1^2, ..., s_C^n\}$ be the training set and each $s_i^n \in S$ is the nth training sample for sign class i. Firstly, the features are extracted $F = \{f_1^1, f_1^2, ..., f_C^n\}$. In the training stage, a generic model which includes a group of references and a distance metric is learnt with the training features. The references and the distance metric are optimized alternately until convergence. Concretely speaking, with a given metric, a group of references can be updated and with the given references, the distance metric can be updated. Specifically, the algorithm learning the distance with given references is the proposed RDML, and the iterative optimization of distance metric and references is iRDML. In the recognition stage, firstly the fragment-based feature f_p is extracted from the probe sign s_p. Then the label of it can be predicted with the previous learnt generic model as follows:

$$c^* = \underset{c \in \{1,...,C\}}{\operatorname{argmin}} \ d(f_p, r_c), \tag{1}$$

where $d(f_p, r_c)$ is the distance between f_p and a specific reference r_c with distance metric M.

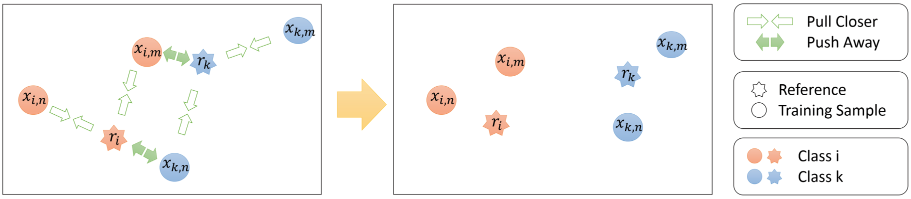

Fig. 2. Illustration of RDML.

The basic idea of our method is to learn a generic model which is robust to different signers. In the initialization of the training stage, firstly, the reference, i.e. a signer invariant description of each sign class is represented approximately by a simple mean of all samples within the specific class. Once the references are given, the generic distance metric can be optimized accordingly. To derive the optimized metric, a constraint that all the samples from different signers should be close to their corresponding references is considered. Concretely speaking, we minimize the distance between the samples and the corresponding references and maximize the distance between the samples and the other references. Actually, the references for various signs are not predefined or accessible in sign language. Thus we use an iterative manner to learn the appropriate sign references and update the corresponding metric from the plenty of training samples performed by multiple signers.

3.2 RDML

Let $X = \{X_1, X_2, ..., X_C\}$ be the labeled training data and $X_i = \{x_{i,1}, x_{i,2}, ..., x_{i,n_i}\}$ are the training samples of ith class. For each class $i = 1, 2, ..., C$, it is assumed that there exists a generic reference r_i. Then the goal of our reference driven metric learning (RDML) is to seek a good metric $d(x_{i,j}, r_k)$ so that the distance from a data point $x_{i,j}$ to its reference $r_i(d(x_{i,j}, r_i))$ is smaller than the distances to other references $r_k(d(x_{i,j}, r_k), k \neq i)$.

Figure 2 is the illustration of the objective function. In Fig. 2, the stars stand for the references and the circles are the training samples. The red and blue colors represent different classes. $x_{i,j}$ is the jth sample of ith class and r_i is the reference of ith class. In the original space, for some training samples, such as $x_{i,m}$ and $x_{k,n}$, the distance to its corresponding reference is larger than the distances to some other references. The goal of our proposed method is to pull the samples closer to their true references and push them away further from the false references.

The distance metric d could be formulated with a square matrix M:

$$d(x_{i,j}, r_k) = (x_{i,j} - r_k)^T M (x_{i,j} - r_k), \tag{2}$$

where M is an $m \times m$ square matrix and m is the dimension of $x_{i,j}$. So the RDML becomes an optimizing problem with the following objective function:

$$\min_M f(M) = f_1(M) + f_2(M), \tag{3}$$

where term $f_1(M)$ is to pull the samples to their corresponding references as close as possible and the term $f_2(M)$ tries to push the samples far away from the false references.

The first term is to minimize the within-class distance, which is defined as the distance between the samples and their true references. It can be directly formulated as the sum of all these kinds of distances:

$$f_1(M) = \sum_{i=1}^{C} \sum_{j=1}^{n_i} d(x_{i,j}, r_i). \tag{4}$$

The second term constrains that each sample should be closer to the true reference than the references of other classes. So only those samples which violate this rule are penalized. Here we use hinge loss $[z]_+ = max(z, 0)$ in our loss function:

$$f_2(M) = \sum_{i=1}^{C} \sum_{j=1}^{n_i} \sum_{k=1}^{C} [1 + d(x_{i,j}, r_i) - d(x_{i,j}, r_k)]_+. \tag{5}$$

So the final objective function is

$$\begin{aligned} \min_M f(M) =& f_1(M) + f_2(M) \\ =& \sum_{i=1}^{C} \sum_{j=1}^{n_i} d(x_{i,j}, r_i) + \sum_{i=1}^{C} \sum_{j=1}^{n_i} \sum_{k=1}^{C} [1 + d(x_{i,j}, r_i) - d(x_{i,j}, r_k)]_+. \\ =& \sum_{i=1}^{C} \sum_{j=1}^{n_i} (x_{i,j} - r_i)^T M (x_{i,j} - r_i) \\ &+ \sum_{i=1}^{C} \sum_{j=1}^{n_i} \sum_{k=1}^{C} \Big[1 + (x_{i,j} - r_i)^T M (x_{i,j} - r_i) \ -(x_{i,j} - r_k)^T M (x_{i,j} - r_k)\Big]_+. \end{aligned} \tag{6}$$

To optimize Eq. (6), gradient descent algorithm is adopted. Let $C_{i,j}^k = (x_{i,j} - r_k)(x_{i,j} - r_k)^T$, Eq. (6) can be rewritten as:

$$f(M) = \sum_{i=1}^{C} \sum_{j=1}^{n_i} tr(MC_{i,j}^i) + \sum_{i=1}^{C} \sum_{j=1}^{n_i} \sum_{k=1}^{C} [1 + tr(MC_{i,j}^i) - tr(MC_{i,j}^k)]_+. \tag{7}$$

Here we define an active triplet set S_a, so that $(i, j, k) \in S_a$ could trigger the hinge loss of $f_2(M)$ in Eq. (5), i.e.,

$$1 + d(x_{i,j}, r_i) > d(x_{i,j}, r_k). \tag{8}$$

So the gradient G of $f(M)$ is

$$G = \frac{\partial f(M)}{\partial M} = \sum_{i=1}^{C} \sum_{j=1}^{n_i} C_{i,j}^{i} + \sum_{(i,j,k) \in S_a} (C_{i,j}^{i} - C_{i,j}^{k}). \tag{9}$$

The optimization of RDML is shown in Algorithm 1, and the initial input distance metric M can be the identity matrix I or any predefined square matrix.

Algorithm 1. Optimization of RDML.

```
Require: Initial distance metric $M$, Training set $X$, References $R$
Ensure: Distance metric $M$
while not converged do
    update $S_a$ based on Equation (8)
    compute $G$ based on Equation (9)
    $M \longleftarrow M - stepsize \times G$
end while
return $M$
```

3.3 iRDML

In above mentioned RDML, the references are assumed to be known already. However, in many situations, the references are not well defined or accessible. Although the center of each class can be used as approximate reference, it does have the difference with the real generic reference for involving many noisy elements. So in this paper, we try to learn more appropriate references in an iterative manner for subsequent modelling and classification.

With a given distance metric M, Eq. (6) can be seen as a function of references R:

$$\begin{aligned} \min_R f(R) &= f_1(R) + f_2(R) \\ &= \sum_{i=1}^{C} \sum_{j=1}^{n_i} d(x_{i,j}, r_i) + \sum_{i=1}^{C} \sum_{j=1}^{n_i} \sum_{k=1}^{C} [1 + d(x_{i,j}, r_i) - d(x_{i,j}, r_k)]_+. \end{aligned} \tag{10}$$

This problem could be solved iteratively by optimizing each reference r_i while fixing other references $r_k (k \neq i)$ with gradient descent algorithm. For each reference r_i, we define two active sets S_1^i and S_2^i:

$$S_1^i = \{(j,k) | 1 + d(x_{i,j}, r_i) > d(x_{i,j}, r_k)\}. \tag{11}$$

$$S_2^i = \{(j,k) | 1 + d(x_{k,j}, r_k) > d(x_{k,j}, r_i)\}. \tag{12}$$

Then the gradient can be represented as:

$$g = \frac{\partial f(r_i)}{\partial r_i} = \sum_{j=1}^{n} 2M(x_{i,j} - r_i) + \sum_{(j,k) \in S_1^i} 2M(x_{i,j} - r_i) - \sum_{(j,k) \in S_2^i} 2M(x_{k,j} - r_i). \tag{13}$$

Algorithm 2. Optimization of References.

```
Require: Initial references $R$, Distance metric $M$, Training set $X$
Ensure: Updated references $R$
while not converged do
    for $i = 1$ to $C$ do
        while not converged do
            update active sets $S_1^i$ and $S_2^i$
            $g = \frac{\partial f(r_i)}{\partial r_i}$
            $r_i \longleftarrow r_i - stepsize \times g$
        end while
    end for
end while
return R
```

The optimization of references is shown in Algorithm 2. The initial references R can be the class centers or any given references.

With a given M, a new group of references R can be learnt one-by-one with fixing the others as described in Algorithm 2. While with a given R, a new distance metric M can be optimized with Algorithm 1. This is a chicken-and-egg problem, so we try to solve it by alternately optimizing M and R iteratively. Algorithm 3 summarizes the procedure of our iterative Reference Driven Metric Learning (iRDML) algorithm.

Algorithm 3. Optimization of iRDML.

```
Require: Training set $X$
Ensure: Updated references $R$, Updated metric $M$
for $i = 1$ to $C$ do
    $r_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{i,j}$
end for
initialize $M$ using Algorithm 1 with $R$
while not converged do
    update $R$ with Algorithm 2
    update $M$ with Algorithm 1
end while
return $R$ and $M$
```

4 Implementation

4.1 Training

In the training stage, the features are extracted from all training samples. The optimal references R and the corresponding distance metric M are learnt with the proposed iRDML method (Algorithm 3). Since M is a positive semi-definite matrix, it could be decomposed as $M = L^T L$. So the Mahalanobis distance between a data point $x_{i,j}$ and a reference r_k with matrix M is

$$\begin{aligned} D(x_{i,j}, r_k) &= \sqrt{(x_{i,j} - r_k)^T M (x_{i,j} - r_k)} \\ &= \sqrt{(Lx_{i,j} - Lr_k)^T (Lx_{i,j} - Lr_k)}. \end{aligned} \tag{14}$$

Equation (14) shows that the Mahalanobis distance with matrix M is equivalent to Euclidean distance of the data in the projected space transformed by matrix L. Therefore, all the references are projected with L for the subsequent recognition.

4.2 Recognition

In the recognition stage, the probe sample x_p is firstly projected with L. Then the class label of probe data x_p can be predicted as:

$$c^* = \underset{c \in \{1,\dots,C\}}{\operatorname{argmin}} \; d_{Eucl}(Lx_p, Lr_c), \tag{15}$$

where d_{Eucl} means Euclidean distance metric.

5 Experiments

To evaluate the performance of the proposed iRDML method, we conduct the experiments on SLR both on public DEVISIGN database and our own collected datasets. Further, we also validate the algorithm on action recognition task on HDM05 dataset.

5.1 Evaluation on DEVISIGN Database

In this subsection, we evaluate our method on DEVISIGN database and the fair comparison on this public database will validate the effectiveness and efficiency of the proposed iRDML.

Datasets. Although there existed some public sign language datasets, there are not many choices to conduct a fair comparison. Many works reported their performance in their own selected subset from the public dataset. Here we choose DEVISIGN [26] dataset for our experiments.

All the experiments conducted on DEVISIGN-L follow the evaluation protocol in [26]. 8 groups of data from 4 signers form the training set and the data from other 4 signers are the test data.

Evaluation on Different Features. In our implementation, we adopt the recent fragment-based feature [27]. Fragment-based feature is designed to describe the sequential data such as sign language data. In the implementation of fragment-based feature, each sign is divided into 5 fragments. The feature of each fragment contains both trajectory and hand shapes. The motions of these 5 joints (left hand, right hand, left elbow, right elbow and head) form the trajectory features. The dimension of trajectory feature in each fragment is 120. For the hand shape feature, HOG descriptor of the typical frame in each fragment is selected to describe the appearance of the hand shape. In order to reduce the computing time, the feature dimension for the final hand shape representation is reduced to 165 from the original 648 by PCA technique. By concatenating the trajectory and hand shape features, the vector to characterize each fragment is generated. With the fragment partition, each sign can be represented by concatenating features of all sequential fragments.

We evaluate our method with two different features. One is the fragment-based feature. The other is frame-based feature, which is widely use in conventional SLR algorithm, such as HMM, DTW etc. In our implementation of frame-based feature, different from the traditional dense frame-based feature, the sparse frame-based feature is generated by linear interpolation to normalize the dimension of the final feature vector so that it can be fed into iRDML. Specifically, skeleton pairwise feature [28] and HOG features of hand shapes are extracted from each frame, and then interpolated into 15 frames. The final dimension of frame-based feature is 2625 and the dimension of fragment-based feature is 1425.

We conducted the experiments on the proposed iRDML and a classical metric learning method (LMNN [21]) to evaluate the representative ability of the two kinds of features mentioned above. In iRDML, the means of the sign classes are used as initial references directly.

From Fig. 3, it is obvious that fragment-based feature has a better performance than frame-based feature, no matter with LMNN or iRDML. Figure 3 also shows that the improvement from frame-based feature to fragment-based feature with LMNN is much more than that with iRDML. The reason should be that comparing with frame-based feature, fragment-based feature is more robust to different signers, which brings the obvious improvement of LMNN. While for iRDML, our generic model is already derived from different signers, so the enhancement of using fragment-based feature is less obvious. In the following experiments on SLR, the inputs for LMNN and our (i)RDML methods are all fragment-based feature.

Comparison Between RDML and iRDML. In our iRDML, since the references and the metric are optimized iteratively, we would like to show how the performance changes with the iterations in this subsection. In our test, the convergence condition of the iterations is set to $|R_t - R_{t-1}| < 0.001$. Figure 4 gives the accuracies of RDML and iRDML with different iterations. Comparing with RDML, the improvement of iRDML is significant. One point should be noticed

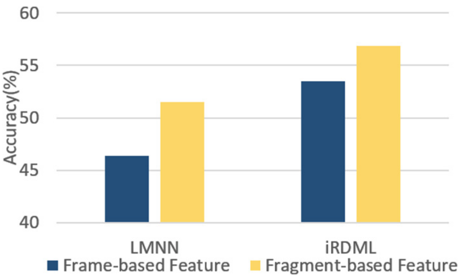

Fig. 3. Performance comparison with different features.

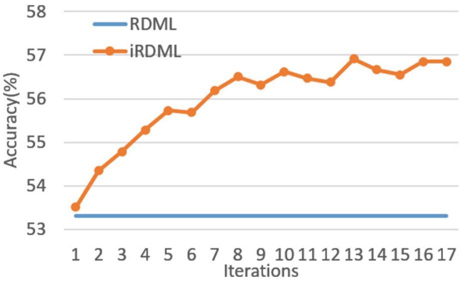

Fig. 4. Accuracies of RDML and iRDML.

that the accuracy of the first iteration in iRDML is higher than RDML. The reason is that in RDML, the class centers are used as references and the metric is learnt accordingly while in the first iteration of iRDML, the references are updated based on the metric learnt in RDML. Besides the comparison between RDML and iRDML, Fig. 4 also clearly shows how the performance of iRDML increases with iterations.

Comparison with Existing Methods. In this part, we will evaluate the performance of our algorithm by comparing it to some other methods reported in [26], which are HMM, DTW, ARMA and GCM respectively. LMNN [21] is one of the most classic metric learning algorithm, so it is also compared in our experiments. In RDML, the class centers are used as references directly, and the class centers are also used as initialization for iRDML.

Comparison on Accuracy. Firstly, the accuracies of different methods are compared and listed in Table 1. From the table we can see that RDML and iRDML consistently outperform the other methods. Comparing with HMM, DTW and ARMA, the accuracy is improved by about 20 percentage points. While for GCM and LMNN, our iRDML still achieves 5 percentage points enhancement.

Comparison on Time Cost. One of the important advantages of our proposed method is that it costs less time in recognition stage. The reason mainly lies in

Table 1. Comparisons with other methods on DEVISIGN database.

Method	Chai et al. [26]				LMNN [21]	Ours	
	HMM	DTW	ARMA	GCM		RDML	iRDML
Acc. (%)	34.44	38.35	39.03	51.81	51.49	53.30	56.85
Time (ms)	507.4	15778	1842	534.7	1.174	0.213	0.213

two folds. On one hand, the computation cost of the linear distance metric is low. On the other hand, in our (i)RDML, the probe only needs to compare with the references instead of all the training samples. We record the recognition time of all the test signs and calculate the average recognition time. Here the recognition time doesn't count the time for feature extraction since all the methods share approximately same amount of time for feature extraction. The experiments are run on a regular PC equipped with Intel Core i7 and 10 GB RAM, which is similar to the experiment conditions in [26]. The time costs of different methods are given in Table 1.

From Table 1, it can be seen that HMM, DTW, ARMA and GCM cost much more time than the metric learning based methods (LMNN, RDML and iRDML). The difference between them is three or four orders of magnitude. Comparing with LMNN, (i)RDML is faster in recognition stage because only the references are used for recognition. This advantage of (i)RDML over LMNN on time cost will be more and more significant when the size of the training data is increasing.

5.2 Evaluation on Our Own Dataset

In order to fully evaluate the proposed method, we have collected two datasets with Kinect by ourselves. These two datasets are used for signer independent and singer-dependent evaluations respectively.

Dataset. Dataset 1 is used for signer independent test. The vocabulary size is 1000. There are 7 signers and each signer performed 1000 signs only once. The 7 groups of data performed by different signers in Dataset 1 are referred as $I1$, $I2$,... and $I7$.

Different from the multiple signers in Dataset 1, Dataset 2 only has one signer who performed the same 1000 vocabulary with three repetitions. This dataset is used for the purpose of signer dependent evaluation.

Signer Independent Evaluation. In this part, we will evaluate our algorithm on Dataset 1 and compare with other methods. Besides HMM [29], DTW [30], ARMA [31] and LMNN [21], we also compare with two other classical metric learning methods: ITML [19] and CSML [32]. For the earlier three methods, the fused frame-based feature is adopted, which is generated by contanating the trajectory and hand shape. Specifically, the trajectory feature is skeleton pairwise feature [28] and the hand shape is described by HOG feature. While for metric learning based methods, we still use the fragment-based feature.

The leave-one-out cross validation is conducted on the seven groups of data from Dataset 1 and all the results are shown in Table 2. Each row with the group name 'In' means the accuracy in this row is evaluated by taking group 'In' as probe data. From the table we can see that iRDML still has the highest performance on this dataset. The last row in Table 2 gives the standard deviations of different methods. iRDML shows relatively stable performance with the

Table 2. Accuracy (%) comparisons on Dataset 1.

Method	$I1$	$I2$	$I3$	$I4$	$I5$	$I6$	$I7$	**Ave.**	**sd.**
HMM	57.4	57.1	58.7	55.9	55.9	61.0	47.4	**56.2**	**4.3**
DTW	61.7	60.5	66.6	60.6	33.5	49.2	16.2	**49.8**	**18.5**
ARMA	65.8	65.2	66.6	64.9	61.7	71.4	47.0	**63.2**	**7.7**
ITML	68	70.2	67.8	74.6	66.9	76.3	64.8	**69.8**	**4.2**
CSML	70.5	73.4	70.8	73.5	68.3	76.9	67.2	**71.5**	**3.3**
LMNN	70.0	72.1	69.0	75.6	69.1	77.7	66.2	**71.3**	**4.0**
iRDML	75.0	78.5	74.4	78.9	74.1	81.4	72.1	**76.3**	**3.3**

Table 3. The p-values given by the Student's t distribution comparing with iRDML.

Baseline/iRDML	HMM/iRDML	DTW/iRDML	ARMA/iRDML	LMNN/iRDML
p-value	0.000003	0.006209	0.000982	0.000046

least standard deviation. We also conduct statistical tests to validate whether the advantage of iRDML is statistically significant comparing with other baseline methods. The p-values are given by the Student's t distribution in Table 3. The statistical tests convincingly show that comparing with other methods, the performance improvement of our proposed iRDML is statistically significant ($p < 0.01$).

Signer Dependent Evaluation. Above experiments are all evaluated in the signer independent case. Although our method is proposed to tackle signer independent problem, we hope it can also work well in signer dependent situation. We conduct this experiment on Dataset 2 in three-fold cross validation. The average accuracies are presented in Fig. 5. It can be seen that iRDML still performs well in this signer dependent dataset although the enhancement is modest. Dataset 2 has only three groups of data and the references are learnt from limited two groups. Therefore, it is reasonable that the improvement of iRDML is indistinctive.

5.3 Experiment on Human Motion Recognition

Although the method is proposed to tackle the SLR problem, it is indeed a generic algorithm for recognition tasks, especially for the person or subject independent case. In this section we evaluate the performance of iRDML on human motion recognition.

The experiment is conducted on the public dataset HDM05 [33]. The motion capture data in HDM05 have been recorded at the Hochschule der Medien (HDM) in the year 2005. The dataset consists of 2337 motion sequences from 65 actions.

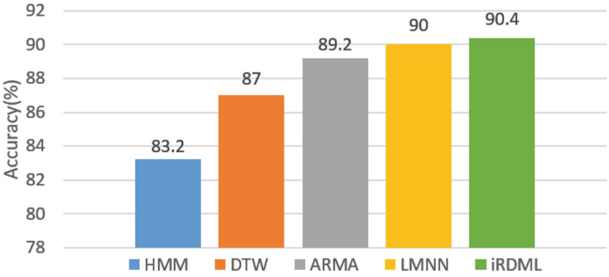

Fig. 5. Accuracy on signer dependent dataset.

We follow the experimental settings in [34]. The data are randomly split into 10 balanced partitions of sequences and 10-fold cross validation is adopted. The results reported in this paper is the average accuracy over 10 folds. The original feature of HDM05 is formed by the 3D coordinates of 31 joints. To align the actions, DTW is adopted. A standard sample is selected randomly in each class and all the other samples are aligned to it. After DTW, each action sequence is sampled to 20 frames. In each frame, we use the same feature as [34] described. Not only the 3D coordinates of joints (PO) but also the temporal differences (TD) between pairs of PO feature are adopted. Finally PCA is used for dimensionality reduction of motion features and the final feature dimension for each action sequence is 1460.

Table 4. Comparisons with other methods on HDM05.

Method	Cho et al. [34]				Ours
	ELM	SVM	MLP	Hybrid MLP	iRDML
Acc. (%)	91.57	94.95	95.20	95.59	95.76

Here we compare our method with the experimental results reported in [34]. The accuracies are shown in Table 4. The performance of iRDML is slightly better than the state-of-the-art MLP and Hybrid MLP with the same feature. Therefore, iRDML can be regarded as a general method to tackle such subject independent problem.

6 Conclusion

This paper proposed a novel iterative Reference Driven Metric Learning (iRDML) method to tackle signer independent SLR problem. We try to seek a generic model which could capture the common character for each sign of different signers. In the generic model, each sign is represented by a signer invariant reference and RDML is proposed to learn the distance between specific references and the training samples. Then an iterative optimizing algorithm is designed to further explore more appropriate references and the corresponding distance metric. Extensive experiments have shown the effectiveness of our proposed iRDML on SLR task. Compared with the state-of-the-art methods, iRDML shows the obvious advantages in both the accuracy and the speed. The extended experiment on human motion recognition suggests that our method can be generalized to other recognition task.

Acknowledgements. This work was partially supported by 973 Program under contract No. 2015CB351802, Natural Science Foundation of China under contracts Nos. 61390511, 61472398, Microsoft Research Asia and the Youth Innovation Promotion Association CAS.

References

1. Lichtenauer, J.F., Hendriks, E.A., Reinders, M.J.: Sign language recognition by combining statistical DTW and independent classification. IEEE Trans. Pattern Anal. Mach. Intell. **30**(11), 2040–2046 (2008)
2. Ong, E.J., Cooper, H., Pugeault, N., Bowden, R.: Sign language recognition using sequential pattern trees. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2200–2207. IEEE (2012)
3. Wang, H., Stefan, A., Moradi, S., Athitsos, V., Neidle, C., Kamangar, F.: A system for large vocabulary sign search. In: Kutulakos, K.N. (ed.) ECCV 2010. LNCS, vol. 6553, pp. 342–353. Springer, Heidelberg (2012). doi:10.1007/978-3-642-35749-7_27
4. Chai, X., Li, G., Lin, Y., Xu, Z., Tang, Y., Chen, X., Zhou, M.: Sign language recognition and translation with kinect. In: IEEE Conference on AFGR (2013)
5. Von Agris, U., Schneider, D., Zieren, J., Kraiss, K.F.: Rapid signer adaptation for isolated sign language recognition. In: IEEE Conference on Computer Vision and Pattern Recognition Workshop, CVPRW 2006, pp. 159–159 (2006)
6. Leggetter, C.J., Woodland, P.C.: Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. Comput. Speech Lang. **9**(2), 171–185 (1995)
7. Gauvain, J.L., Lee, C.H.: Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. IEEE Trans. Speech Audio Process. **2**(2), 291–298 (1994)
8. Von Agris, U., Blomer, C., Kraiss, K.F.: Rapid signer adaptation for continuous sign language recognition using a combined approach of eigenvoices, MLLR, and MAP. In: 19th International Conference on Pattern Recognition, ICPR 2008. IEEE, pp. 1–4 (2008)
9. Kuhn, R., Junqua, J.C., Nguyen, P., Niedzielski, N.: Rapid speaker adaptation in eigenvoice space. IEEE Trans. Speech Audio Process. **8**(6), 695–707 (2000)
10. Farhadi, A., Forsyth, D., White, R.: Transfer learning in sign language. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2007, pp. 1–8, June 2007
11. Zieren, J., Kraiss, K.-F.: Robust person-independent visual sign language recognition. In: Marques, J.S., Pérez de la Blanca, N., Pina, P. (eds.) IbPRIA 2005. LNCS, vol. 3522, pp. 520–528. Springer, Heidelberg (2005). doi:10.1007/11492429_63

12. Shanableh, T., Assaleh, K.: User-independent recognition of Arabic sign language for facilitating communication with the deaf community. Digit. Sig. Process. **21**(4), 535–542 (2011)
13. Kong, W., Ranganath, S.: Towards subject independent continuous sign language recognition: a segment and merge approach. Pattern Recogn. **47**(3), 1294–1308 (2014)
14. Mensink, T., Verbeek, J., Perronnin, F., Csurka, G.: Distance-based image classification: generalizing to new classes at near-zero cost. IEEE Trans. Pattern Anal. Mach. Intell. **35**(11), 2624–2637 (2013)
15. McFee, B., Lanckriet, G.R.: Metric learning to rank. In: Proceedings of the 27th International Conference on Machine Learning (ICML 2010), pp. 775–782 (2010)
16. Lim, D., Lanckriet, G., McFee, B.: Robust structural metric learning. In: Proceedings of the 30th International Conference on Machine Learning, pp. 615–623 (2013)
17. Lu, J., Zhou, X., Tan, Y.P., Shang, Y., Zhou, J.: Neighborhood repulsed metric learning for kinship verification. IEEE Trans. Pattern Anal. Mach. Intell. **36**(2), 331–345 (2014)
18. Xing, E.P., Jordan, M.I., Russell, S., Ng, A.Y.: Distance metric learning with application to clustering with side-information. In: Advances in Neural Information Processing Systems, pp. 505–512 (2002)
19. Davis, J.V., Kulis, B., Jain, P., Sra, S., Dhillon, I.S.: Information-theoretic metric learning. In: Proceedings of the 24th International Conference on Machine Learning, ICML 2007, pp. 209–216. ACM, New York (2007)
20. Shental, N., Hertz, T., Weinshall, D., Pavel, M.: Adjustment learning and relevant component analysis. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002. LNCS, vol. 2353, pp. 776–790. Springer, Heidelberg (2002). doi:10.1007/3-540-47979-1_52
21. Weinberger, K.Q., Blitzer, J., Saul, L.K.: Distance metric learning for large margin nearest neighbor classification. In: Advances in Neural Information Processing Systems, pp. 1473–1480 (2005)
22. Chai, J., Liu, H., Chen, B., Bao, Z.: Large margin nearest local mean classifier. Signal Process. **90**(1), 236–248 (2010)
23. Wang, F.: Semisupervised metric learning by maximizing constraint margin. IEEE Trans. Syst. Man Cybern. Part B Cybern. **41**(4), 931–939 (2011)
24. Baghshah, M.S., Shouraki, S.B.: Semi-supervised metric learning using pairwise constraints. In: IJCAI, vol. 9, pp. 1217–1222. Citeseer (2009)
25. Niu, G., Dai, B., Yamada, M., Sugiyama, M.: Information-theoretic semi-supervised metric learning via entropy regularization. Neural Comput. **26**(8), 1717–1762 (2014)
26. Chai, X., Wang, H., Chen, X.: The devisign large vocabulary of chinese sign language database and baseline evaluations. Technical report VIPL-TR-14-SLR-001. Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS (2014)
27. Yin, F., Chai, X., Zhou, Y., Chen, X.: Weakly supervised metric learning towards signer adaptation for sign language recognition. In: British Machine Vision Conference (2015)
28. Wang, J., Liu, Z., Wu, Y., Yuan, J.: Mining actionlet ensemble for action recognition with depth cameras. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1290–1297. IEEE (2012)

29. Wang, C., Gao, W., Shan, S.: An approach based on phonemes to large vocabulary Chinese sign language recognition. In: Proceedings of Fifth IEEE International Conference on Automatic Face and Gesture Recognition, 2002, pp. 411–416. IEEE (2002)
30. Salvador, S., Chan, P.: Toward accurate dynamic time warping in linear time and space. Intell. Data Anal. **11**(5), 561–580 (2007)
31. Xu, C., Wang, T., Gao, J., Cao, S., Tao, W., Liu, F.: An ordered-patch-based image classification approach on the image Grassmannian manifold. IEEE Trans. Neural Netw. Learn. Syst. **25**(4), 728–737 (2014)
32. Nguyen, H.V., Bai, L.: Cosine similarity metric learning for face verification. In: Kimmel, R., Klette, R., Sugimoto, A. (eds.) ACCV 2010. LNCS, vol. 6493, pp. 709–720. Springer, Heidelberg (2011). doi:10.1007/978-3-642-19309-5_55
33. Muller, M., Roder, T., Clausen, M.: Documentation mocap database HDM05. Technical report CG-2007-2, University of Bonn (2007)
34. Cho, K., Chen, X.: Classifying and visualizing motion capture sequences using deep neural networks. In: 2014 International Conference on Computer Vision Theory and Applications (VISAPP), vol. 2, pp. 122–130. IEEE (2014)