15th IAPR International Conference on Machine Vision Applications (MVA)
Nagoya University, Nagoya, Japan, May 8-12, 2017.

Enhancing Discriminability of Randomized Time Warping for Motion Recognition

Lincon Sales de Souza[1], Bernardo Bentes Gatto[2] and Kazuhiro Fukui[3]
University of Tsukuba, Tsukuba, Japan[1,3]
Federal University of Amazonas, Manaus, Brazil[2]
{lincons[1],kfukui[3]}@cvlab.cs.tsukuba.ac.jp, bernardo[2]@icomp.ufam.edu.br

Abstract

In this paper, we propose a framework of action sequence recognition by combining the representation of randomized time warping (RTW) with the enhanced Grassmann discriminant Analysis (eGDA). RTW is an extension of Dynamic time warping (DTW), and it has been shown to be effective for motion recognition, as it can effectively retain an actions temporal information by generating a low-dimensional subspace from a set of time elastic (TE) features of a video. On the other hand, the eGDA can use the concepts of generalized difference subspace and Grassmann manifold symbiotically to learn a discriminative manifold where video subspaces can be regarded as points. The main advantages of the proposed method are: removing common features between the actions which are not useful for discrimination, thus increasing the distance between subspaces of different classes, and reducing the distance between subspaces of the same class; and estimating a discriminative manifold even if there are few training data. We demonstrate the validity of the proposed method through experiments on motion recognition using two public datasets, namely, the Cambridge gesture database and the KTH action dataset.

1 Introduction

In this paper, we discuss a framework for characterizing and classifying motion image sequences, focusing on hand gestures and human actions. Among the methods for motion analysis, dynamic time warping (DTW) has been one of the most widely used [1]. The core idea of DTW is to search for the best alignment of two sequential patterns by optimizing a warping function, which specifies the sequential correspondence between them. The search is done by dynamic programming which can optimize the alignment score and produce the alignment path of the most similar warped patterns.

DTW has been recently generalized to a faster and more effective method named randomized time warping (RTW) [2], which does not need dynamic programming. The core idea of RTW is to generate a set of time warped patterns, called time elastic (TE) features, through repeated random subsampling, while preserving the original temporal order. This mechanism can be regarded as a simultaneous search for the most similar warped patterns from a number of randomly obtained candidates. Comparing two sets of TE features can be costly as the number of features increase, therefore the comparison is conducted using a subspace based method, in which each set of TE features is represented as a low-dimensional subspace,

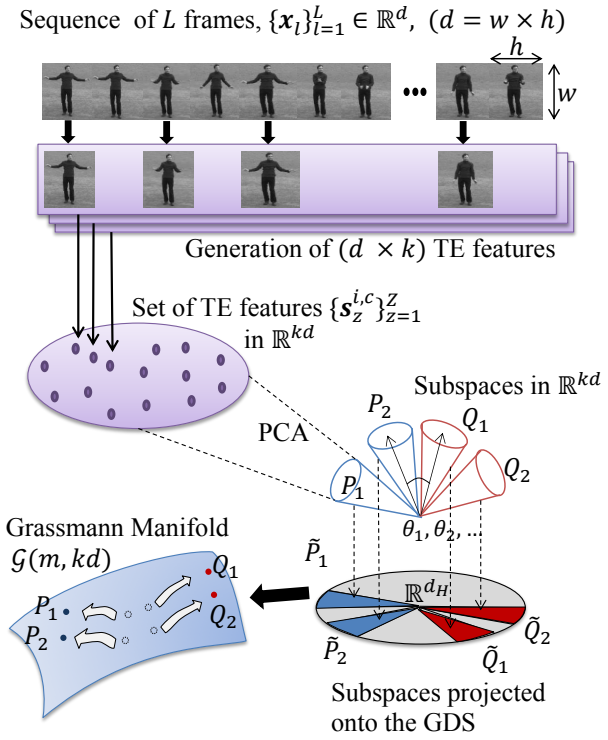


Figure 1. Conceptual diagram of the proposed method. A set of TE features is extracted by randomly sampling images from an image sequence. Next, a hypo subspace is generated by applying PCA to the set. For each image sequence, a hypo subspace is generated in this way. Finally, the hypo subspaces are orthogonalized by projecting them onto the GDS, and then are projected onto the Grassmann Manifold.

called a sequence hypothesis (Hypo) subspace.

The RTW converts the problem of comparing two sequences to comparing two hypo subspaces, which can then in turn be solved by measuring the canonical angles between them. The mutual subspace method (MSM) [3] is well known as a fundamental classification method using canonical angles, which has been used along with RTW.

Comparison of hypo subspaces has also been performed by introducing the Grassmann manifold formulation, which simplifies the complicated procedure of the subspace based method using canonical angles. The Grassmann manifold, symbolized as $\mathcal{G}(m, D)$, is defined as a set of m-dimensional linear subspaces of $\mathbb{R}^D$ [4]. In this framework, a subspace-based method is regarded as a simple classification method on a Grassmann manifold, where each single subspace is

treated as a point, and thereby, each motion video is represented by a point in the manifold. Various types of classification methods have been constructed on a Grassmann manifold [5, 6]. But in particular, RTW formulation has been concretely used along with the discriminant analysis on a Grassmann manifold (GDA), which has been known as one of the useful tools for image set classification. GDA can be easily conducted as a kernel discriminant analysis through the kernel trick with a Grassmann kernel.

Although it has been useful to combine RTW with GDA, some issues arise from this representation:

- TE feature space is usually very high dimensional, and same-class actions may have vary large variation in this space;
- some parts in time of some actions' movement may look similar to that of others, causing overlap of the different actions distributions in the TE feature space;
- GDA is capable of finding the most discriminant directions in a manifold only from the given points on the manifold, and it cannot operate the TE features points directly. Hence, if classes were not separable in the TE feature space, the corresponding data points on the manifold are also not separable.

From this viewpoint, we propose to project class subspaces onto a generalized difference subspace (GDS) [7], before mapping each class subspace on a Grassmann manifold, as can be seen in Fig 1. This idea has been recently useful for subspaces which represent 3D object recognition [8]. As GDS has been shown to magnify the angles between different class subspaces to provide more discriminative sample for GDA, it is expected this mechanism can improve the representation of the RTW hypo subspaces on the Grassmann manifold. The validity of our proposed method is demonstrated through experiments with the Cambridge gesture database [9] and the KTH action dataset [10].

The rest of the paper is organized as follows. In Sec.2, we explain the proposed framework for classifying motion in detail. In Sec.3, we conduct experiments on motion recognition using two public datasets, namely, the Cambridge hand gesture database and the KTH action dataset. Sec.4 concludes the paper.

2 Algorithm of the Proposed Method

In this section, we explain the algorithm of the proposed method.

In our framework, an image with the size $w \times h$ is represented by a $d(= w \times h)$-dimensional vector, so that any given feature vector $\boldsymbol{x} \in \mathbb{R}^d$. Consider N_c training ordered sequences $\{\boldsymbol{x}_l^{i,c}\}_{l=1}^{L_i^c}$ for each c-th class $(c = 1, \ldots, C)$ and an ordered sequence of L_{in} input images $\{\boldsymbol{x}_l^{in}\}_{l=1}^{L_{in}}$. Each of these sequences represent a body motion or hand gesture captured by video, for example.

An $d \times k$ dimensional TE feature vector $\boldsymbol{s} = [\boldsymbol{y}_1^T \boldsymbol{y}_2^T \ldots \boldsymbol{y}_k^T]$ is created by randomly selecting k images from a sequence $\{\boldsymbol{x}_l^{i,c}\}_{l=1}^{L_i^c}$, such that $\boldsymbol{y}_1^T \boldsymbol{y}_2^T \ldots \boldsymbol{y}_k^T \in \{\boldsymbol{x}_l^{i,c}\}_{l=1}^{L_i^c}, t(\boldsymbol{y}_1) < \ldots < t(\boldsymbol{y}_k)$, where $t(\cdot)$ denotes the original order of the image.

Let this procedure of random selection be repeated Z times, such that we obtain $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_Z$. Subsequently, a correlation-like matrix $\boldsymbol{R}_i^c$, which corresponds to the set of the TE feature vectors, can be computed as:

$$\boldsymbol{R}_i^c = \frac{1}{Z} \sum_{z=1}^{Z} \boldsymbol{s}_z^{i,c} \boldsymbol{s}_z^{i,c\top}. \tag{1}$$

We apply principal component analysis (PCA) by computing the eigenvectors of each matrix $\boldsymbol{R}_i^c$ to construct m-dimensional subspaces $\mathcal{Y}_i^c$. The orthogonal basis of each subspace are obtained as the eigenvectors corresponding to the m largest eigenvalues. In the following, each subspace m-dimensional $\mathcal{Y}_i^c$ is represented by the matrix $\boldsymbol{Y}_i^c \in \mathbb{R}^{kd \times m}$, which has the corresponding orthogonal basis as its column vectors. A set of TE features generated from a sequence contains various possible warped patterns, each of which corresponds to one hypothesis. In this sense, the subspace generated from a set of TE features is called a sequence hypothesis (Hypo) subspace.

In order to utilize effectively the feature extraction function of GDS, we introduce the global class subspaces $\mathcal{M}_c$, which is denoted by a matrix $\boldsymbol{M}_c \in \mathbb{R}^{kd \times d_m}$, which represents compactly all the subspaces belonging to the same class c. The orthogonal basis of $\mathcal{M}_c$ can be obtained as the eigenvectors corresponding to the d_m largest eigenvalues of the auto-correlation matrix:

$$\boldsymbol{R}^c = \frac{1}{N_c} \sum_{i=1}^{N_c} \boldsymbol{R}_i^c = \frac{1}{ZN_c} \sum_{i=1}^{N_c} \sum_{z=1}^{Z} \boldsymbol{s}_z^{i,c} \boldsymbol{s}_z^{i,c\top}. \tag{2}$$

Next, to generate a GDS, we calculate the total sum matrix, $\boldsymbol{S}$, which is defined as:

$$\boldsymbol{S} = \sum_{c=1}^{C} \sum_{j=1}^{d_m} \boldsymbol{\Phi}_j^c \boldsymbol{\Phi}_j^{c\top}, \tag{3}$$

where $\boldsymbol{\Phi}_j^c$ is a basis of the d_m-dimensional $\mathcal{M}_c$. The orthogonal basis of the GDS can be obtained as d_h eigenvectors, $\{\boldsymbol{d}_i\}_{i=1}^{d_h}$ corresponding to the d_h smallest eigenvalues of the sum matrix $\boldsymbol{S}$. The subspaces $\boldsymbol{Y}_i^c$ are projected onto the GDS and their projections are denoted by $\{\widetilde{\boldsymbol{Y}}_i^c\}_{i=1}^{N_c} \in \mathbb{R}^{d_h \times m}$. The input subspace of $\boldsymbol{X}$ is also projected onto the GDS and its projection is denoted by $\widetilde{\boldsymbol{X}}$.

We apply the GDA algorithm to these projected subspaces. For example, the kernel matrix, $\boldsymbol{K}$, is calculated as the similarity matrix between class subspaces $\widetilde{\boldsymbol{Y}}_q$ and $\widetilde{\boldsymbol{Y}}_w$. The step-by-step training and testing algorithms of the proposed method are shown in Algorithms 1 and 2, respectively.

3 Experiments

In this section, we discuss the validity of the proposed method through hand gesture and human action recognition tasks.

Algorithm 1: Learning algorithm of the proposed method

input: training ordered sequences $\{\boldsymbol{x}_l^{i,c}\}_{l=1}^{L_i^c}$, with class label c

for $c = 1, \ldots, C$ **do**
  for $i = 1, \ldots, N_c$ **do**
    $\{\boldsymbol{s}_z^{i,c}\}_{z=1}^{Z} \leftarrow \mathrm{TE}(\{\boldsymbol{x}_l^{i,c}\}_{l=1}^{L_i^c})$ // obtain TE features
    $\boldsymbol{R}_i^c \leftarrow \frac{1}{Z}\sum_{z=1}^{Z} \boldsymbol{s}_z^{i,c}\boldsymbol{s}_z^{i,c\top}$ // calculate set covariance matrix
    $\boldsymbol{Y}_i^c \leftarrow \mathrm{EVD}(\boldsymbol{R}_i^c)$ // apply eigendecomposition
  end
  $\boldsymbol{R}^c \leftarrow \frac{1}{N_c}\sum_{i=1}^{N_c} \boldsymbol{R}_i^c$ // calculate class covariance matrix
  $\boldsymbol{M}_c \leftarrow \mathrm{EVD}(\boldsymbol{R}^c)$ // apply eigendecomposition
end
$\boldsymbol{P}, \boldsymbol{H} \leftarrow \mathrm{EVD}(\sum_{c=1}^{C} \boldsymbol{M}_c\boldsymbol{M}_c^\top)$ // obtain GDS and principal subspace
foreach $\boldsymbol{Y}_i^c$ **do** $\widetilde{\boldsymbol{Y}}_i^c \leftarrow H^\top \boldsymbol{Y}_i^c$ // project all subspaces onto the GDS
for $q = 1, \ldots, N$ **do**
  for $w = 1, \ldots, N$ **do**
    $[\boldsymbol{S}_{train}]_q^w \leftarrow k_p(\widetilde{\boldsymbol{Y}}_q, \widetilde{\boldsymbol{Y}}_w)$ // generate similarity matrix
  end
end
$\boldsymbol{\alpha}^* \leftarrow \max_{\boldsymbol{\alpha}} Ra(\boldsymbol{\alpha})$ // solve LDA problem
$\boldsymbol{F}_{train} \leftarrow \boldsymbol{\alpha}^{*\top} \boldsymbol{S}_{train}$ // compute training coefficients
return $\boldsymbol{F}_{train}, \boldsymbol{H}, \boldsymbol{\alpha}^*$ // return dictionary, GDS and GDA projection operators

Algorithm 2: Input evaluation algorithm of the proposed method

input: pattern set with L' input images $\{\boldsymbol{x}^{in}\}$

$\{\boldsymbol{s}_z^{in}\}_{z=1}^{Z} \leftarrow \mathrm{TE}(\{\boldsymbol{x}^{in}\})$ // obtain TE features
$\boldsymbol{R}_{in} \leftarrow \frac{1}{Z}\sum_{z=1}^{Z} \boldsymbol{s}_z^{in}\boldsymbol{s}_z^{in\top}$ // calculate set covariance matrix
$\boldsymbol{X} \leftarrow \mathrm{EVD}(\boldsymbol{R}_{in})$ // apply eigendecomposition
$\widetilde{\boldsymbol{X}} \leftarrow \boldsymbol{H}^\top \boldsymbol{X}$ // project subspace onto the GDS
for $q = 1, \ldots, N$ **do**
  $[\boldsymbol{S}_{test}]_q \leftarrow k_p(\widetilde{\boldsymbol{Y}}_q, \widetilde{\boldsymbol{X}})$ // generate similarity matrix
end
$\boldsymbol{F}_{test} \leftarrow \boldsymbol{\alpha}^{*\top} \boldsymbol{S}_{test}$ // compute test coefficients
$\mathrm{pred}(\boldsymbol{x}^{in}) \leftarrow \mathrm{NN}(\boldsymbol{F}_{train}, \boldsymbol{F}_{test})$ // perform 1-NN classification
return $\mathrm{pred}(\boldsymbol{x}^{in})$ // return a class prediction

Table 1. Results of the Cambridge Hand Dataset Experiment. m refers to the dimension of the hypo subspaces, d_m is the dimension of the global class subspaces, and d_p is the dimension of the principal subspace, which is complementary to the GDS.

	Accuracy (%)	m	d_m	d_p
RTW+GDA	91.56	6	-	-
RTW+eGDA	94.89	5	50	15

3.1 Experiment with Cambridge Hand Dataset

We conducted two types of experiments with the Cambridge hand gesture dataset. This database contains 9 classes of hand gesture videos, each in 5 illumination scenarios, and 20 sample videos for each of the scenarios and classes. The number of frames of each video ranges from 37 to 119. In addition, in the experiments, all the images were resized to 12×16 pixels.

In the first experiment, we performed a qualitative experiment to aid in the visualization of the proposed method mechanism. We utilize three classes of hand gestures from the Cambridge dataset, each containing 50 videos. Figure 2 shows scatter plots of the generated points corresponding to the hypo subspaces, by using the conventional method and the proposed method. The figure suggests that by using the proposed framework, reduction of the distance between subspaces of the same class can be achieved.

In the second experiment, we compared the combination of RTW and conventional GDA with RTW and the enhanced GDA. The number of selected frames to build one TE feature is fixed at 15, the number of TE features for each set is fixed to be 100. The other parameters, namely, dimension of hypo subspaces m, dimension of class subspaces d_m, and dimension of principal subspace d_p were varied and optimized. The results can be seen in Table 1.

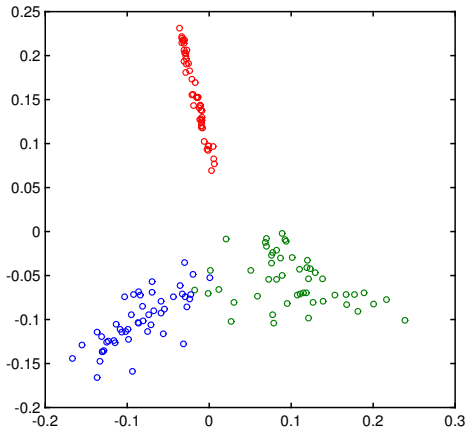
Table 2. Results of the KTH Action Dataset Experiment.

	Accuracy (%)	m	d_m	d_p
RTW+GDA	82.03	10	-	-
RTW+eGDA	83.96	10	120	5

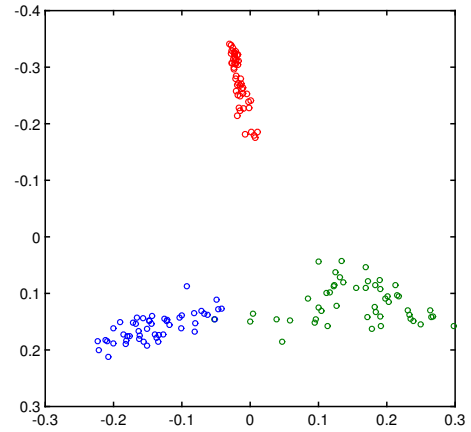
3.2 Experiment with KTH Action Dataset

We also conducted an experiment using the KTH action dataset. This database contains 6 classes of actions performed by humans in videos,namely: boxing, hand clapping, hand waving, running, jogging, and walking. The videos were filmed under 4 different shooting conditions: outdoors, outdoors with variation of zooming, outdoors with different clothes, and indoors.There are 4 sample videos for each of the conditions and classes. The number of frames of each video ranges from 37 to 119. In addition, in the experiments, all the images were resized to 16×16 pixels. In total there are 2391 sequences of actions.

We compare the combination of RTW and conventional GDA with RTW and the enhanced GDA. 2 repetitions were used for testing and 2 for training. The number of selected frames to build on TE feature is fixed at 15, and the number of TE features for each set



(a) Conventional GDA



(b) eGDA

Figure 2. Scatter points of three hand gesture classes by using RTW combined with (a) Conventional GDA; (b) eGDA.

is fixed to be 100. The other parameters, namely dimension of subspaces, class subspaces, and GDS were varied and optimized. The results can be seen in Table 2.

4 Conclusions

In this paper we have proposed a combination of randomized time warping and eGDA, to address more effectively the classification of motion sequences, focusing on the applications of hand gestures and human action classification. The key idea of our enhanced Grassmann manifold is to project class subspaces onto a generalized difference subspace before mapping them on a Grassmann manifold. The GDS projection can extract the differences between classes and generate data points with optimized between-class separability on the manifold, which are more desirable for GDA. The validity of our enhanced Grassmann discriminant analysis was evaluated through classification experiments with Cambridge hand gesture dataset and KTH action dataset, where it outperformed the state-of-the-art method by using RTW and GDA. As a future work, we seek to comprehend the relationship between the two types of mapping in GDS projection and Grassmann manifold more clearly.

Acknowledgement

This work is supported by JSPS KAKENHI Grant Number 16H02842.

References

[1] T. Darrell and A. Pentland, "Space-time gestures," in *Computer Vision and Pattern Recognition, 1993. Proceedings CVPR'93., 1993 IEEE Computer Society Conference on*, pp. 335–340, IEEE, 1993.

[2] C. H. Suryanto, J.-H. Xue, and K. Fukui, "Randomized time warping for motion recognition," *Image and Vision Computing*, vol. 54, pp. 1–11, 2016.

[3] O. Yamaguchi, K. Fukui, and K. Maeda, "Face recognition using temporal image sequence," *Proc. International Conference on Automatic Face and Gesture Recognition*, pp. 318–323, 1998.

[4] Y. Chikuse, "Statistics on special manifolds," *Springer, Lecture. Notes in Statistics*, vol. 174, 2013.

[5] J. Hamm and D. D. Lee, "Grassmann discriminant analysis: a unifying view on subspace-based learning," in *Proceedings of the 25th international conference on Machine learning*, pp. 376–383, ACM, 2008.

[6] P. Turaga, A. Veeraraghavan, A. Srivastava, and R. Chellappa, "Statistical computations on grassmann and stiefel manifolds for image and video-based recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 33, no. 11, pp. 2273–2286, 2011.

[7] K. Fukui and A. Maki, "Difference subspace and its generalization for subspace-based methods," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 37, no. 11, pp. 2164–2177, 2015.

[8] L. Souza, H. Hino, and K. Fukui, "3d object recognition with enhanced grassmann discriminant analysis," in *ACCV 2016 Workshop (HIS 2016)*, 2016.

[9] T.-K. Kim and R. Cipolla, "Canonical correlation analysis of video volume tensors for action categorization and detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 8, pp. 1415–1428, 2009.

[10] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local svm approach," in *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, vol. 3, pp. 32–36, IEEE, 2004.