Combining Multiple Kernel Methods on Riemannian Manifold for Emotion Recognition in the Wild

Mengyi Liu, Ruiping Wang, Shaoxin Li, Shiguang Shan, Zhiwu Huang, Xilin Chen
Key Lab of Intelligence Information Processing of Chinese Academy of Sciences (CAS),
Institute of Computing Technology, CAS, Beijing, 100190, China
{mengyi.liu, shaoxin.li, zhiwu.huang}@vipl.ict.ac.cn,
{wangruiping, sgshan, xlchen}@ict.ac.cn

ABSTRACT

In this paper, we present the method for our submission to the Emotion Recognition in the Wild Challenge (EmotiW 2014). The challenge is to automatically classify the emotions acted by human subjects in video clips under real-world environment. In our method, each video clip can be represented by three types of image set models (i.e. linear subspace, covariance matrix, and Gaussian distribution) respectively, which can all be viewed as points residing on some Riemannian manifolds. Then different Riemannian kernels are employed on these set models correspondingly for similarity/distance measurement. For classification, three types of classifiers, i.e. kernel SVM, logistic regression, and partial least squares, are investigated for comparisons. Finally, an optimal fusion of classifiers learned from different kernels and different modalities (video and audio) is conducted at the decision level for further boosting the performance. We perform an extensive evaluation on the challenge data (including validation set and blind test set), and evaluate the effects of different strategies in our pipeline. The final recognition accuracy achieved 50.4% on test set, with a significant gain of 16.7% above the challenge baseline 33.7%.

Categories and Subject Descriptors

I.5.4 [**Pattern Recognition**]: Applications—*computer vision, signal processing*; I.4.m [**Image Processing and Computer Vision**]: Miscellaneous

Keywords

Emotion Recognition; Riemannian Manifold; Multiple Kernels; EmotiW 2014 Challenge

1. INTRODUCTION

Automatic emotion recognition is a popular and challenging problem in the research fields of cognitive psychology, human-computer interaction, pattern recognition, and so on. Early stage research mostly focuses on the emotion databases collected in "lab-controlled" environment where human subjects posed particular emotions (e.g. angry, happy, and surprise). With recent advances in emotion recognition community, various spontanous or wild databases have been introduced for emotion recognition challenges, such as the Facial Expression Recognition & Analysis (FERA 2011) [33], Audio Video Emotion Challenges (AVEC 2011/2012/2013) [32], and Emotion Recognition in the Wild (EmotiW 2013) [8]. These challenges have provided common benchmarks for emotion recognition researchers.

Previous works on emotion recognition can be broadly categorized into two groups [39]: static image based methods [29, 42, 21] and video based methods [41, 38, 22]. The video based methods tend to utilize dynamic information extracted from image sequences for improving the performance. For instance, Zhao *et al.* [41] encoded spatial-temporal patterns in facial image sequences using LBP-TOP features. Liu *et al.* [22] modeled each emotion clip as a manifold of mid-level features for representing the local spatial-temporal variations on faces. As demonstrated in their experiments, various types of dynamic features are crucial for modeling emotion variations in the recognition task.

Generally, extracting dynamics from successive frames requires accurate image alignment to eliminate the rigid motion effect brought by camera or head pose. However, it is quite difficult especially when dealing with "wild data" due to the large variations caused by uncontrolled real-world environment. As a video clip can be simply regarded as an image set, it is natural to introduce the image-set-based classification methods [13, 35, 34, 20], which have been proved to be more robust to image misalignment. So in this paper, we propose to represent the emotion video clip using three kinds of image set models (i.e. linear subspace, covariance matrix, and Gaussian distribution) respectively, which can all be viewed as points residing on some Riemannian manifolds. Then different Riemanian kernels are employed on these set models correspondingly for similarity/distance measurement. For classification, three types of classifiers, kernel SVM, logistic regression, partial least squares, are investigated for comparisons. Finally, a score-level fusion of classifiers learned based on different kernel methods and different modalities (i.e. video and audio) is conducted to further improve the performance. An overview of the proposed method is illustrated in Figure 1. We will detail the whole procedure in the next section.

WOODSTOCK '97 El Paso, Texas USA

http://dx.doi.org/10.1145/2663204.2666274.

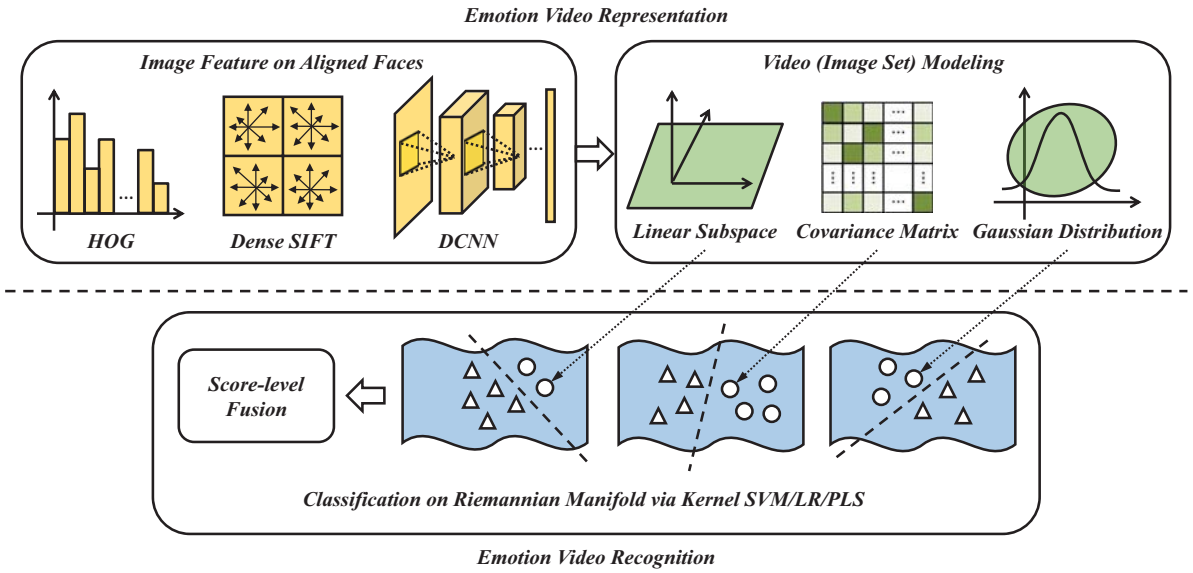


Figure 1: An overview of the proposed method. The whole procedure includes two stages: emotion video representation and recognition. In representation stage, different image features are first extracted from the coarsely aligned faces, then different image set models are employed on frame features respectively for representing each video clip. In recognition stage, classification on Riemannian manifold sapnned by the points (i.e. image sets) is performed using different types of classifiers by exploiting a group of Riemannian kernels. Finally a score-level fusion is conducted to combine the prediction results from different kernels.

2. THE PROPOSED METHOD

2.1 Image Feature

2.1.1 HOG

The Histogram of Oriented Gradients (HOG) [4] feature describes the local shape and appearance of objects by capturing the distribution information of intensity gradients or edge directions. The descriptor decomposes a local region into small squared cells, computes the histogram of different bins of oriented gradients in each cell, and normalizes the results using block-wise pattern (each block contains several cells). HOG is commonly used in computer vision problems, such as object detection and recognition. It has also been successfully used for facial expression analysis in [6, 30].

2.1.2 Dense SIFT

The Scale-Invariant Feature Transform (SIFT) [25] combines a feature detector and a feature descriptor. The detector extracts a number of interested points from an image in a way that is consistent with some variations of the illumination or viewpoint. The descriptor associates to the region around each interest point a signature which identifies its appearance compactly and robustly. For dense SIFT, it is equivalent to performing SIFT descriptor on a dense grid of locations on an image at a fixed scale and orientation. The obtained feature vectors characterizing appearance information are often used for categorization task.

2.1.3 Deep CNN Feature

Convolutional Neural Network (CNN) [19] is a type of feed-forward artificial neural network which is inspired from biology. The individual neurons are designed to simulate cells within visual cortex, which are sensitive to small subregions of input space, named receptive fields [15]. Thus the connections among neurons are tied in such a way that each output neuron only responds to a local region of input neurons. This mechanism is better suited to exploit the strong spatially local correlations presented in natural images. Currently, one of the most popular CNN architectures is the 9-layers deep model [17] designed for ImageNet ILSVRC-2012. There are four convolutional layers with their corresponding pooling layers, and finally followed by an output layer which is constructed according to category labels. As the experiments in some latest works [18, 12, 31] have shown, this architecture, even the pre-trained model via ImageNet data, can be well generalized to many other problems, without any further specific design but maintaining impressive performance.

2.2 Video (Image Set) Modeling

After extracting image features for each video frame, one video clip can be regarded as a set of feature vectors $F = [f_1, f_2, ..., f_n]$, where $f_i \in R^d$ denotes the i-th image with d-dimensional feature description. Based on the feature vector set, we exploit three types of image set models, linear subspace [13], covariance matrix [35], and Gaussian distribution [28, 1], for their desirable capability of capturing data variations to model emotion video.

2.2.1 Linear Subspace

The feature set $F = [f_1, f_2, ..., f_n]$ can be represented by a linear subspace $P \in R^{d \times r}$ via SVD as follows:

$$\sum_{i=1}^{n} f_i f_i^T = P \Lambda P^T, \quad (1)$$

where $P = [p_1, p_2, ..., p_r]$, p_j is the j-th leading eigenvector, r is the dimension of the subspace, and n is the number of frames in the video clip. All of the video samples can be modeled as a collection of linear subspaces [37, 13], which are also the data points on Grassmann manifold $Gr(r, d)$ (Grassmann manifold is a special case of Riemannian manifold [13]).

2.2.2 Covariance Matrix

We can also represent the image feature set with the $d \times d$ sample covariance matrix:

$$C = \frac{1}{n-1}\sum_{i=1}^{n}(f_i - \overline{f})(f_i - \overline{f})^T, \quad (2)$$

where $\overline{f}$ is the mean vector of the image features. As the raw second-order statistic of a set of samples, the covariance matrix makes no assumption about the data distribution, thus providing a natural representation by encoding the feature correlation information specific to each class [35]. It is also well known that the $d \times d$ nonsingular covariance matrices are Symmetric Positive Definite (SPD) matrices Sym_d^+ lying on a Riemannian manifold.

2.2.3 Gaussian Distribution

Suppose the feature vectors $f_1, f_2, ..., f_n$ follow a k-dimensional Gaussian distribution $\mathcal{N}(\mu, \Sigma)$, where μ and Σ are the data mean and covariance respectively:

$$\mu = E(f_i) = \frac{1}{n}\sum_{i=1}^{n} f_i, \quad (3)$$

$$\Sigma = E[(f_i - \mu)(f_i - \mu)^T] = \frac{1}{n-1}\sum_{i=1}^{n}(f_i - \mu)(f_i - \mu)^T. \quad (4)$$

The Gaussian jointly considers the first-order statistic mean and second-order statistic covariance in a single model. By embedding the space of Gaussians into a Lie group or regarding it as a product of Lie groups, we can measure the intrinsic distance between Gaussians on the underlying Riemannian manifold [20].

2.3 Riemannian Kernels

2.3.1 Kernels for Linear Subspace

As presented in Section 2.2.1, the video samples are modeled as a collection of linear subspaces which correspond to points lying on Grassmann manifold M (also in Riemannian space), denoted by $\mathcal{P} = \{P_i\}_{i=1}^N$, where N is the number of video samples. The similarity between two data points P_i and P_j can be measured via mapping the Grassmann manifold to Euclidean space using Mercer kernels [13]. One popularly used kernel [13, 14, 23] is the Projection kernel originated from the principle angles between two subspaces given by (see Figure 2):

$$\mathcal{K}_{i,j}^{Proj.-Poly.} = (\gamma \cdot ||P_i^T P_j||_F^2)^\alpha, \quad (5)$$

where $\mathcal{K}_{i,j}^{Proj.-Poly.}$ is an element in the kernel matrix. The corresponding mapping is $\Phi_{Proj.} = P_i P_i^T$. Then a form of

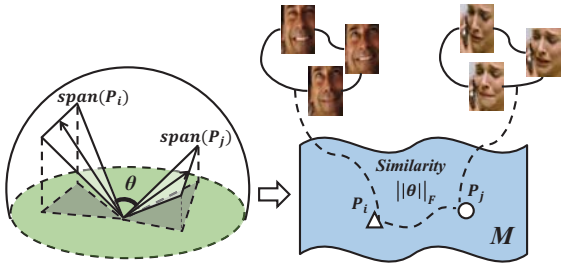


Figure 2: An illustration of principal angles of linear subspaces and their projection metric distances on Grassmann manifold M.

RBF kernel [34] can be generated using $\Phi_{Proj.}$ by:

$$\mathcal{K}_{i,j}^{Proj.-RBF} = exp(-\gamma||\Phi_{Proj.}(P_i) - \Phi_{Proj.}(P_j)||_F^2). \quad (6)$$

2.3.2 Kernels for Covariance Matrix

The $d \times d$ SPD matrices, i.e. non-singlular covariance matrices $\mathcal{C} = \{C_i\}_{i=1}^N$, can be formulated as data points on SPD Riemannian manifold [26]. A commonly used distance metric for SPD matrices is the Log-Euclidean Distance (LED) [2]. Based on LED, [35] proposed a Riemannian kernel that computes the inner-product in a vector space T obtained by mapping data points from the SPD manifold to the tangent space at the identity matrix I via ordinary matrix logarithm operator (see Figure 3).

$$\mathcal{K}_{i,j}^{LED-Poly.} = (\gamma \cdot trace[log(C_i) \cdot log(C_j)])^\alpha. \quad (7)$$

The mapping corresponding to $\mathcal{K}_{i,j}^{LED-Poly.}$ is given by $\Phi_{LED} = log(C_i)$. Similarly a form of RBF kernel [34] can be generated using Φ_{LED} by:

$$\mathcal{K}_{i,j}^{LED-RBF} = exp(-\gamma||\Phi_{LED}(C_i) - \Phi_{LED}(C_j)||_F^2). \quad (8)$$

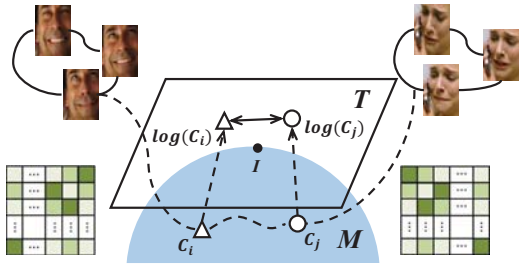


Figure 3: An illustration of mapping covariance matrices from the SPD Riemannian manifold M to the tangent space T (which is a vector space) at the point of identity matrix I on M.

2.3.3 Kernels for Gaussian Distribution

The space of d-dimensional multivariate Gaussians is a Riemannian manifold and can be embeded into the space of Symmetric Positive Definite (SPD) matrices [24], denoted as Sym_{d+1}^+. Thus A d-dimensional Gaussian $\mathcal{N}(\mu, \Sigma)$ is uniquely represented by a $(d+1) \times (d+1)$ SPD matrix G

as follows:

$$\mathcal{N}(\mu,\Sigma) \sim G = |\Sigma|^{-\frac{1}{d+1}} \begin{bmatrix} \Sigma + \mu\mu^T & \mu \\ \mu^T & 1 \end{bmatrix} \quad (9)$$

When obtaining the SPD matrices $\mathcal{G} = \{G_i\}_{i=1}^N$, we can calculate the corresponding Riemannian kernels similarly as in Section 2.3.2:

$$\mathcal{K}_{i,j}^{LED-Poly.} = (\gamma \cdot trace[log(G_i) \cdot log(G_j)])^{\alpha}. \quad (10)$$

$$\mathcal{K}_{i,j}^{LED-RBF} = exp(-\gamma||\Phi_{LED}(G_i) - \Phi_{LED}(G_j)||_F^2). \quad (11)$$

2.4 Classifiers

Based on the above six Riemannian kernels, traditional learning methods operating in vector space can be exploited to classify data points (i.e. image set models) on the Riemannian manifolds for emotion video recognition. In our framework, three types of classifiers are investigated as described below.

2.4.1 Kernel SVM

The Riemannian kernels enable the classifiers to operate in an extrinsic feature space without computing the coordinates of data in original space. An SVM classifier in the kernel space is given by

$$f(x) = \overrightarrow{w}^{*T}\Phi(x) + b^*, \quad (12)$$

where $\Phi(x)$ is the mapping (e.g. $\Phi_{Proj.}$ and Φ_{IP}) which generates the kernel function $k(\cdot,\cdot)$ by

$$\mathcal{K}_{i,j} = k(x_i, x_j) = \Phi(x_i)^T\Phi(x_j). \quad (13)$$

The weight vector $\overrightarrow{w}^*$ and bias b^* are given by

$$\begin{aligned} \overrightarrow{w}^*, b^* &= \underset{\overrightarrow{w},b,\eta}{argmin}\ \frac{1}{2}||\overrightarrow{w}||_2^2 + C\sum_i \eta_i. \\ s.t.\quad & y_i(\overrightarrow{w}^T\Phi(x_i) + b) \geq 1 - \eta_i, \eta_i \geq 0. \end{aligned} \quad (14)$$

For this work, we employ the LibSVM [3] implementation on our pre-calculated Riemannian kernel matrices for classification.

2.4.2 Logistic Regression

According to the Riemanian kernel matrices, the i-th row contains similarities between the i-th video (image set) and all videos in training set, which can be directly treated as a feature vector of this sample. For each sample in the training or test set, we calculate its similarities to all training samples thus obtain the training kernel matrix and test kernel matrix for feature representation. We employ an L2-regularized logistic regression on these features for classification by solving the objective function:

$$\underset{\overrightarrow{w}}{min}(C\sum_i log(1 + exp(-y_i - \overrightarrow{w}^T x_i)) + \frac{1}{2}||\overrightarrow{w}||_2^2). \quad (15)$$

For this work, we employ the Liblinear [11] implementation for optimization.

2.4.3 Partial Least Squares

Similar to Section 2.4.2, we also apply the partial least square classifier [36] to the kernel matrices. We adapt it to a one-vs-all manner to especially deal with the difficult and confusion categories as in [23].

Suppose there are c categories of emotions, we design c one-vs-all PLS to predict each class simultaneously. For a single classifier, given feature variables X and 0-1 labels Y, the PLS decomposes them into

$$\begin{aligned} X &= U_x V_x^T + r_x \\ Y &= U_y V_y^T + r_y \end{aligned} \quad (16)$$

where U_x and U_y contain the extracted latent vectors, V_x and V_y represent the loadings, and r_x and r_y are residuals. PLS is to find weight vectors w_x and w_y that

$$[cov(u_x, u_y)]^2 = \underset{|w|=|v|=1}{max}[cov(Xw_x, Yw_y)]^2, \quad (17)$$

where u_x and u_y are the column vectors of U_x and U_y respectively. $cov(u_x, u_y)$ is the covariance of samples. With the obtained latent vectors, the regression coefficients from X to Y are given by:

$$\begin{aligned} \beta &= W_x(V_x^T W_x)^{-1}U_x Y \\ &= X^T U_y(U_x^T X X^T U_x)^{-1}U_x^T Y, \end{aligned} \quad (18)$$

thus we can predict $\widehat{Y} = X\beta$ [27]. Applying the c one-vs-all PLS to each test sample, we can obtain c regression values respectively. The category corresponding to the maximum value is decided to be the recognition result.

2.4.4 Fusion Scheme

We learn each classifier on the six Riemannian kernels with different image features respectively. An equal-weighted linear fusion is conducted among the prediction scores obtained by the same type of classifiers. Besides the video modality, we also obtain prediction scores on audio features (extracted by OpenSMILE toolkit) [10]). A weighted term λ is introduced at decision level for video-audio fusion:

$$Score^{fusion} = (1-\lambda)Score^{video} + \lambda Score^{audio} \quad (19)$$

Similarly, The category corresponding to the maximum value of the score vector is decided to be the recognition result.

3. EXPERIMENTS

3.1 EmotiW 2014 Challenge

The Emotion Recognition in the Wild Challenge (EmotiW 2014) [7] consists of an audio-video based emotion classification task which mimics real-world conditions. The goal of this challenge is to extend and carry forward the new common platform for evaluation of emotion recognition methods in the wild defined in EmotiW 2013 [8]. The database in the 2014 challenge is the Acted Faical Expression in Wild (AFEW) [9] 4.0, which has been collected from movies showing close-to-real-world conditions. Three sets for training, validation, and testing are available for participants (The numbers of samples for each emotion category in the three sets are illustrated in Table 1). The task is to classify an audio-video clip into one of the seven emotion categories

Table 1: The numbers of samples for each emotion category in the training, validation and testing sets.

	An	Di	Fe	Ha	Ne	Sa	Su
Train	92	66	66	105	102	82	54
Val	59	39	44	63	61	59	46
Test	58	26	46	81	117	53	26

(i.e. angry, disgust, fear, happy, neutral, sad, and surprise). The labels of the testing set are unknown. Participants can learn their models on training set and optimize the parameters on validation set, then report the prediction results on testing set for evaluation.

3.2 Parameter Setting

We simply use the aligned face images provided by Emoti-W 2014 organizers. All images are resized to 64×64 pixels. Three kinds of image features are employed on the aligned faces: HOG, Dense SIFT, and DCNN.

For HOG, we divide each image into $7 \times 7 = 49$ overlapping blocks with the size of 16×16 pixels (i.e. the strides are 8 pixels in both horizontal and vertical directions). The descriptor is applied by computing histograms of oriented gradient on 2×2 cells in each block, and the orientations are quantized into 9 bins, which results in $2 \times 2 \times 9 = 36$ dimensions for each block and $36 \times 49 = 1764$ dimensions for the whole image.

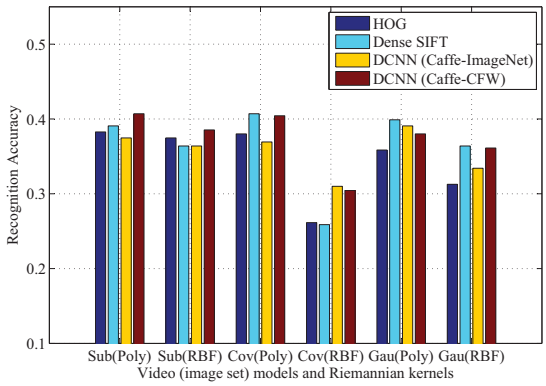
For Dense SIFT, we divide each image into 49 overlapping local regions as done for HOG. In each 16×16 pixels block, we apply the SIFT descriptor to the center point, and obtain a typical $4 \times 4 \times 8 = 128$ dimensions feature vector. For the whole image, we have $128 \times 49 = 6272$ dimensions feature.

For DCNN, we employ the Caffe [16] implementation, which is commonly used in several latest works [12, 31]. Two types of DCNN models are trained by feeding different training data: ImageNet ILSVRC-2012 [5] and Celebrity Faces in the Wild (CFW) [40]. The first one is for evaluating the generalization ability of the deep model and natural image data, so we exactly take use of the same parameters as that in [17], the 9216 nodes' values of the last convolutional layer are used for final image features. The second one is to explore the shared feature representations for both face identities and expressions. Over 150,000 face images from 1,520 people are used for training and the labels are their identities. The architecture is $3@237 \times 237 \rightarrow 96@57 \times 57 \rightarrow 96@28 \times 28 \rightarrow 256@28 \times 28 \rightarrow 384@14 \times 14 \rightarrow 256@14 \times 14 \rightarrow 256@7 \times 7 \rightarrow 4096 \rightarrow 1520$. Similar to the first model, the $256 \times 7 \times 7 = 12544$ nodes' values of the last convolutional layer are used for final image features.

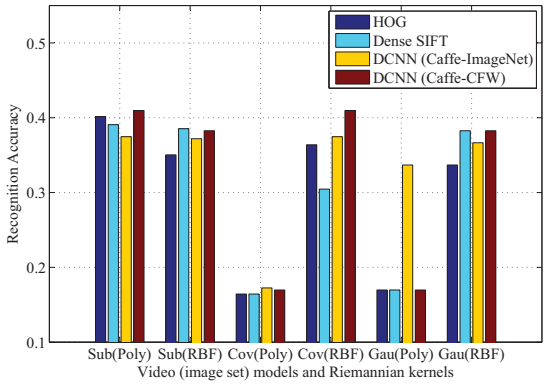
3.3 Results Comparisons

The emotion recognition results on validation set based on different classifiers are illustrated in Figure 4. For each single classifier, the DCNN features have shown promising performance on the task, especially the features extracted by Caffe trained on CFW have achieved better results than the specific hand-crafted features.

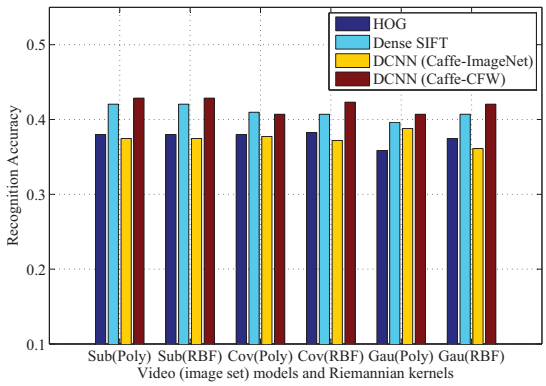
We also demonstrate the results on validation set based on different features in Table 2. For each single feature, the results based on six Riemannian kernels and three classifiers are all listed respectively. As shown, PLS achieves the best performance for its one-vs-all manner which can especially deal with the difficult and confusion categories.



(a) Kernel SVM.



(b) Logistic Regression.



(c) Partial Least Squares.

Figure 4: Emotion recognition accuracy on validation set based on different classifiers.

The overall recognition results are obtained by one-vs-all PLS classifier using decision-level fusion over different kernels. As presented in Section 2.4.4, an equal-weighted linear fusion is conducted among the prediction scores based on the six Riemannian kernels with different features, and the weighted term for video-audio fusion is set as $\lambda = 0.3$ in the final submission. Different strategies of fusion and their corresponding results on validation and test sets are all listed in Table 3.

Table 2: Emotion recognition accuracy on validation set based on different image features.

(a) HOG

	Linear Subspace		Covariance Matrix		Gaussian Distribution	
	Proj.-Poly. Kernel	Proj.-RBF Kernel	LED-Poly. Kenrel	LED-RBF Kernel	LED-Poly. Kenrel	LED-RBF Kernel
Kernel SVM	38.27%	37.47%	38.01%	26.15%	35.85%	31.27%
Logistic Regression	40.16%	35.04%	16.44%	36.39%	16.98%	33.69%
Partial Least Squares	38.01%	38.01%	38.01%	38.27%	35.85%	37.47%

(b) Dense SIFT

	Linear Subspace		Covariance Matrix		Gaussian Distribution	
	Proj.-Poly. Kernel	Proj.-RBF Kernel	LED-Poly. Kenrel	LED-RBF Kernel	LED-Poly. Kenrel	LED-RBF Kernel
Kernel SVM	39.08%	36.39%	40.70%	25.88%	39.89%	36.39%
Logistic Regression	39.08%	38.54%	16.44%	30.46%	16.98%	38.27%
Partial Least Squares	42.05%	42.05%	40.97%	40.70%	39.62%	40.70%

(c) DCNN (Caffe-ImageNet)

	Linear Subspace		Covariance Matrix		Gaussian Distribution	
	Proj.-Poly. Kernel	Proj.-RBF Kernel	LED-Poly. Kenrel	LED-RBF Kernel	LED-Poly. Kenrel	LED-RBF Kernel
Kernel SVM	37.74%	36.39%	36.93%	31.00%	39.08%	33.42%
Logistic Regression	37.47%	37.20%	17.25%	37.47%	33.69%	36.66%
Partial Least Squares	37.47%	37.47%	37.74%	37.20%	38.81%	36.12%

(d) DCNN (Caffe-CFW)

	Linear Subspace		Covariance Matrix		Gaussian Distribution	
	Proj.-Poly. Kernel	Proj.-RBF Kernel	LED-Poly. Kenrel	LED-RBF Kernel	LED-Poly. Kenrel	LED-RBF Kernel
Kernel SVM	40.70%	38.54%	40.43%	30.46%	38.01%	36.12%
Logistic Regression	40.97%	38.27%	16.98%	40.97%	16.98%	38.27%
Partial Least Squares	42.86%	42.86%	40.70%	42.32%	40.70%	42.05%

The confusion matrix of the final submission method are shown in Figure 5. We can see that "angry", "happy" and "neutral" are much easier to be distinguished from other emotions, but it is still hard to do well on some difficult and confusion emotion categories such as "disgust", "fear", and "sad". Moreover, in contrast to the experience in emotion classification on lab-controlled data, in our experiments, "surprise" is hard to be recognized and easy to be confused with some other categories like "neutral" and "fear". The reason may lie in the following two aspects: first, few "surprise" data are provided for learning and testing compared to other categories (as shown in Table 1); second, the "surprise" emotion may not be acted exaggeratedly sometimes in the real-world condition, thus no typical appearance variations (e.g. mouth stretching, upper lip raising) are shown as that in lab-controlled data.

4. CONCLUSIONS

In this paper, we propose a method for video-based emotion recognition in real-world condition. Each emotion video clips is simply regarded as an image set and different kinds of image set models are used to represent the video clips as a collection of data points on Riemannian manifold. Then multiple Riemanian kernels are employed on these set models correspondingly for distance metrics. At last, a score-level fusion of classifiers learned based on different kernel methods and different modalities is conducted for final recognition results. The method is evaluated on EmotiW 2014 data and achieves promising results on both validation and unseen test data. In the future, we will try to deal with the few difficult categories and explore more effective fusion strategy to further improve the performance.

Table 3: Performance comparisons of different strategies on both validation and test set based on PLS.

Methods		Accuracy	
		Val	Test
Baseline (*provided by EmotiW organizers*)		34.4%	33.7%
Audio (*OpenSMILE Toolkit*)		30.73%	--
Video	HOG	38.01%	--
	Dense SIFT	43.94%	--
	DCNN (*Caffe-ImageNet*)	39.35%	--
	DCNN (*Caffe-CFW*)	43.40%	--
	HOG + Dense SIFT	44.47%	--
	HOG + Dense SIFT + DCNN (*Caffe-ImageNet*)	44.74%	--
	HOG + Dense SIFT + DCNN (*Caffe-CFW*)	45.28%	--
Audio + Video (*HOG+Dense SIFT*)		46.36%	46.68%
Audio+Video (*HOG + Dense SIFT + DCNN (Caffe-ImageNet)*)		46.90%	47.91%
Audio+Video (*HOG + Dense SIFT + DCNN (Caffe-CFW)*)		***48.52%***	***50.37%***

	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise
Angry	84.75%	3.39%	0.00%	1.69%	5.08%	5.08%	0.00%
Disgust	10.26%	17.95%	2.56%	28.21%	33.33%	5.13%	2.56%
Fear	27.27%	6.82%	27.27%	13.64%	11.36%	9.09%	4.55%
Happy	4.76%	0.00%	0.00%	82.54%	9.52%	3.17%	0.00%
Neutral	13.11%	0.00%	1.64%	8.20%	70.49%	6.56%	0.00%
Sad	13.56%	3.39%	6.78%	23.73%	28.81%	22.03%	1.69%
Surprise	17.39%	4.35%	28.26%	8.70%	32.61%	2.17%	6.52%

(a) Validation set

	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise
Angry	81.03%	3.45%	0.00%	5.17%	10.34%	0.00%	0.00%
Disgust	11.54%	3.85%	3.85%	34.62%	23.08%	15.38%	7.69%
Fear	26.09%	0.00%	23.91%	10.87%	19.57%	15.22%	4.35%
Happy	8.64%	0.00%	1.23%	64.20%	11.11%	14.81%	0.00%
Neutral	7.69%	1.71%	5.13%	9.40%	63.25%	11.97%	0.85%
Sad	11.32%	0.00%	3.77%	24.53%	24.53%	33.96%	1.89%
Surprise	11.54%	0.00%	19.23%	7.69%	34.62%	19.23%	7.69%

(b) Test set

Figure 5: Confusion matrix of the final submission method.

5. ACKNOWLEDGMENTS

The work is partially supported by Natural Science Foundation of China under contracts nos.61025010, 61222211, 61390511, 61379083.

6. REFERENCES

[1] O. Arandjelovic, G. Shakhnarovich, J. Fisher, R. Cipolla, and T. Darrell. Face recognition with image sets using manifold density divergence. In *CVPR*. IEEE, 2005.

[2] V. Arsigny, P. Fillard, X. Pennec, and N. Ayache. Geometric means in a novel vector space structure on symmetric positive-definite matrices. *SIAM journal on matrix analysis and applications*, 29(1):328–347, 2007.

[3] C.-C. Chang and C.-J. Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.

[4] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*. IEEE, 2005.

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*. IEEE, 2009.

[6] A. Dhall, A. Asthana, R. Goecke, and T. Gedeon. Emotion recognition using phog and lpq features. In *FG*. IEEE, 2011.

[7] A. Dhall, R. Goecke, J. Joshi, K. Sikka, and T. Gedeon. Emotion recognition in the wild challenge 2014: Baseline, data and protocol. In *ACM ICMI*. ACM, 2014.

[8] A. Dhall, R. Goecke, J. Joshi, M. Wagner, and T. Gedeon. Emotion recognition in the wild challenge 2013. In *ACM ICMI*. ACM, 2013.

[9] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon. Collecting large, richly annotated facial-expression databases from movies. *IEEE MultiMedia*, 2012.

[10] F. Eyben, M. Wöllmer, and B. Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. In *ACM MM*. ACM, 2010.

[11] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. Liblinear: A library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874, 2008.

[12] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *arXiv preprint arXiv:1311.2524*, 2013.

[13] J. Hamm and D. D. Lee. Grassmann discriminant analysis: a unifying view on subspace-based learning. In *ICML*. ACM, 2008.

[14] M. T. Harandi, C. Sanderson, S. Shirazi, and B. C. Lovell. Graph embedding discriminant analysis on grassmannian manifolds for improved image set matching. In *CVPR*. IEEE, 2011.

[15] D. H. Hubel and T. N. Wiesel. Receptive fields and functional architecture of monkey striate cortex. *The Journal of physiology*, 195(1):215–243, 1968.

[16] Y. Jia. Caffe: An open source convolutional architecture for fast feature embedding. *http://caffe. berkeleyvision. org*, 2013.

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[18] Z.-Z. Lan, L. Jiang, S.-I. Yu, S. Rawat, Y. Cai, C. Gao, S. Xu, H. Shen, X. Li, Y. Wang, et al. Cmu-informedia at trecvid 2013 multimedia event detection. In *TRECVID 2013 Workshop*, 2013.

[19] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proc. of IEEE*, 86(11):2278–2324, 1998.

[20] P. Li, Q. Wang, and L. Zhang. A novel earth mover's distance methodology for image matching with gaussian mixture models. In *ICCV*. IEEE, 2013.

[21] M. Liu, S. Li, S. Shan, and X. Chen. Au-aware deep networks for facial expression recognition. In *FG*. IEEE, 2013.

[22] M. Liu, S. Shan, R. Wang, and X. Chen. Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition. In *CVPR*, 2014.

[23] M. Liu, R. Wang, Z. Huang, S. Shan, and X. Chen. Partial least squares regression on grassmannian manifold for emotion recognition. In *ACM ICMI*. ACM, 2013.

[24] M. Lovrić, M. Min-Oo, and E. A. Ruh. Multivariate normal distributions parametrized as a riemannian symmetric space. *Journal of Multivariate Analysis*, 74(1):36–48, 2000.

[25] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[26] X. Pennec, P. Fillard, and N. Ayache. A riemannian framework for tensor computing. *International Journal of Computer Vision*, 66(1):41–66, 2006.

[27] R. Rosipal and N. Krämer. Overview and recent advances in partial least squares. In *Subspace, Latent Structure and Feature Selection*, pages 34–51. Springer, 2006.

[28] G. Shakhnarovich, J. W. Fisher, and T. Darrell. Face recognition from long-term observations. In *ECCV*. Springer, 2002.

[29] C. Shan, S. Gong, and P. W. McOwan. Facial expression recognition based on local binary patterns: A comprehensive study. *Image and Vision Computing*, 27(6):803–816, 2009.

[30] K. Sikka, K. Dykstra, S. Sathyanarayana, G. Littlewort, and M. Bartlett. Multiple kernel learning for emotion recognition in the wild. In *ACM ICMI*. ACM, 2013.

[31] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. *arXiv preprint arXiv:1406.2199*, 2014.

[32] M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, and M. Pantic. Avec 2013: the continuous audio/visual emotion and depression recognition challenge. In *ACM Workshop on AVEC*. ACM, 2013.

[33] M. F. Valstar, B. Jiang, M. Mehu, M. Pantic, and K. Scherer. The first facial expression recognition and analysis challenge. In *FG*. IEEE, 2011.

[34] R. Vemulapalli, J. K. Pillai, and R. Chellappa. Kernel learning for extrinsic classification of manifold features. In *CVPR*. IEEE, 2013.

[35] R. Wang, H. Guo, L. S. Davis, and Q. Dai. Covariance discriminative learning: A natural and efficient approach to image set classification. In *CVPR*. IEEE, 2012.

[36] H. Wold. Partial least squares. *Encyclopedia of statistical sciences*, 1985.

[37] O. Yamaguchi, K. Fukui, and K.-i. Maeda. Face recognition using temporal image sequence. In *FG*. IEEE, 1998.

[38] P. Yang, Q. Liu, and D. N. Metaxas. Boosting coded dynamic features for facial action units and facial expression recognition. In *CVPR*. IEEE, 2007.

[39] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(1):39–58, 2009.

[40] X. Zhang, L. Zhang, X.-J. Wang, and H.-Y. Shum. Finding celebrities in billions of web images. *Multimedia, IEEE Transactions on*, 14(4):995–1007, 2012.

[41] G. Zhao and M. Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(6):915–928, 2007.

[42] L. Zhong, Q. Liu, P. Yang, B. Liu, J. Huang, and D. N. Metaxas. Learning active facial patches for expression analysis. In *CVPR*. IEEE, 2012.