J Multimodal User Interfaces (2016) 10:113–124
DOI 10.1007/s12193-015-0204-5

ORIGINAL PAPER

Video modeling and learning on Riemannian manifold for emotion recognition in the wild

Mengyi Liu[1] · **Ruiping Wang**[1] · **Shaoxin Li**[1] · **Zhiwu Huang**[1] · **Shiguang Shan**[1] · **Xilin Chen**[1]

Received: 15 December 2014 / Accepted: 14 October 2015 / Published online: 11 November 2015

Abstract In this paper, we present the method for our submission to the emotion recognition in the wild challenge (EmotiW). 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 extensive evaluations on the EmotiW 2014 challenge data (including validation set and blind test set), and evaluate the effects of different components in our pipeline. It is observed that our method has achieved the best performance reported so far. To further evaluate the generalization ability, we also perform experiments on the EmotiW 2013 data and two well-known lab-controlled databases: CK+ and MMI. The results show that the proposed framework significantly outperforms the state-of-the-art methods.

✉ Shiguang Shan
sgshan@ict.ac.cn

[1] Key Laboratory of Intelligent Processing, Institute of Computing Technology, CAS, Chinese Academy of Sciences (CAS), No. 6 Kexueyuan South Road, Beijing 100190, People's Republic of China

Keywords Emotion recognition · Video modeling · Riemannian manifold · EmotiW 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 spontaneous or wild databases have been introduced for emotion recognition challenges, such as the facial expression recognition and analysis (FERA) [48], audio video emotion challenges (AVEC) [47], and emotion recognition in the wild (EmotiW) [10]. These challenges have provided common benchmarks for emotion recognition researchers.

Previous works on emotion recognition can be broadly categorized into two groups [55]: static image based methods [28,41,58] and video based methods [30,54,57]. The video based methods tend to utilize dynamic information extracted from image sequences for improving the performance. For instance, Zhao et al. [57] encoded spatial-temporal patterns in facial image sequences using LBP-TOP features. Liu et al. [30] 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 [17,27,49,50], which have been proved to be more robust to image misalignment. In EmotiW 2013 challenge [31], we proposed to model each video (image set) as a linear subspace lying on Grassmannian manifold [17], and conduct a one-vs-all manner partial least squares for classification, which achieved promising results using single type of image feature. In our EmotiW 2014 submission [32], we extend the former work by introducing various modeling methods on manifold to further improve the performance. Specifically, each emotion video clip can be represented 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 Fig. 1. The proposed method is evaluated in EmotiW2014 challenge with its results reported in our conference paper [32]. This paper extends the conference version by providing more in-depth investigation of different components in our framework and conducting more extensive evaluations of their effects on the final performance. To further evaluate the generalization ability of our method, we also add experimental results on the EmotiW 2013 data and two well-known lab-controlled databases: CK+ and MMI. In the following sections, we will detail the whole procedure and the extensive evaluations.

2 The proposed method

2.1 Image feature

2.1.1 HOG

The histogram of oriented gradients (HOG) [6] 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 [8,43].

2.1.2 Dense SIFT

The scale-invariant feature transform (SIFT) [34] combines a feature detector and a feature descriptor. The detector extracts a number of interested points from an image in a way that is

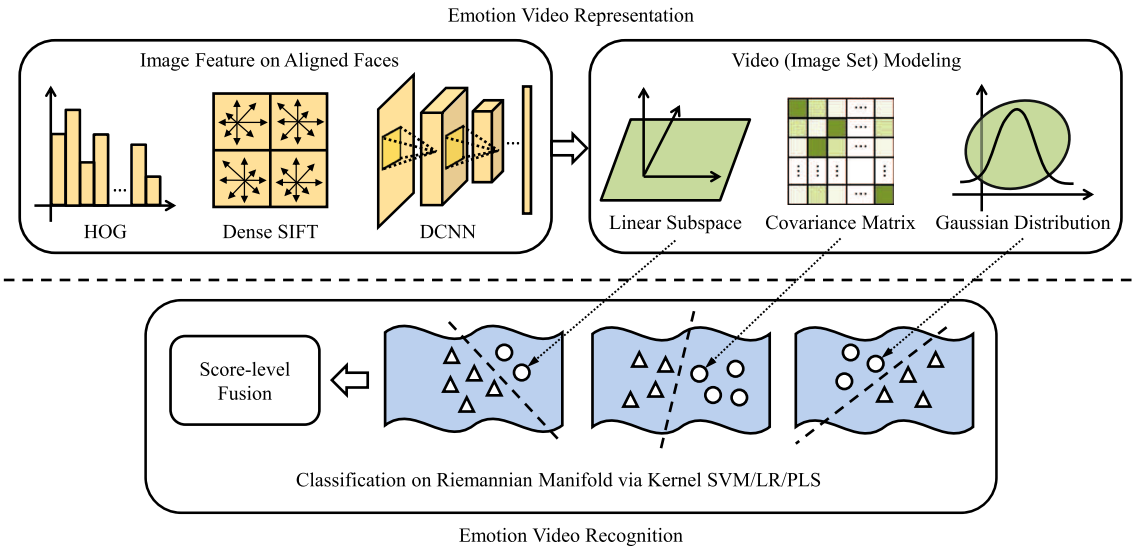

Fig. 1 An overview of the proposed method. The whole procedure includes two stages: emotion video representation and recognition. In representation stage, different image features are first extracted from the coarsely aligned faces, then different image set models are employed on frame features respectively for representing each video clip. In recognition stage, classification on Riemannian manifold spanned 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

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 DCNN feature

Convolutional neural network (CNN) [26] 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 sub-regions of input space, named receptive fields [21]. 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 [24] 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 [14,25,44] have shown, this architecture, even the pre-trained model via ImageNet data without any further specific design changes, can be well generalized to many other problems while 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, and n is the number of frames in the video clip. Based on the feature vector set, we exploit three types of image set models, linear subspace [17], covariance matrix [50], and Gaussian distribution [1, 40], 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:

$$\sum_{i=1}^{n} f_i f_i^T = P \Lambda P^T, \tag{1}$$

where $P = [p_1, p_2, ..., p_r]$, p_j is the j-th leading eigenvector, and r is the dimension of the subspace. All of the video samples can be modeled as a collection of linear subspaces [17,53], which are also the data points on Grassmann manifold $Gr(r, d)$ (Grassmann manifold is a special case of Riemannian manifold [17]).

2.2.2 Covariance matrix

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

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

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 [50]. 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:

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

$$\begin{aligned}\Sigma &= E[(f_i - \mu)(f_i - \mu)^T] \\ &= \frac{1}{n-1}\sum_{i=1}^{n}(f_i - \mu)(f_i - \mu)^T.\end{aligned} \tag{4}$$

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 [27].

2.2.4 Discussion

It is interesting to discuss the relationship among the three different video modeling alternatives (i.e. linear subspace, covariance matrix, and Gaussian distribution) presented above. From Eqs. 1 and 2, our linear subspace ("P") can be viewed as obtained by performing an SVD on the covariance matrix ("C"), where the sample mean information is kept while the eigenvalues that capture the relative importance (magnitude) of different variance directions are discarded.

Moreover, compared to covariance matrix, Gaussian distribution explicitly incorporates "sample mean" into the feature set modeling, which captures both first-order and second-order statistics information.

2.3 Riemannian kernels

The Riemannian kernels enable the classifiers to operate in an extrinsic feature space without computing the coordinates of data in original space. The kernel mapping (e.g. $\Phi_{Proj.}$ and Φ_{LED}) can generates the kernel function $k(\cdot,\cdot)$ by

$$\mathcal{K}_{i,j} = k(x_i, x_j) = \Phi(x_i)^T \Phi(x_j). \tag{5}$$

In the following, we introduce the specific forms of kernel mappings/functions for different set models.

2.3.1 Kernels for linear subspace

Suppose we have N video samples, as presented in Sect. 2.2.1, these video samples can be modeled as a collection of linear subspaces which correspond to points lying on Grassmann manifold $Gr(r,d)$ (also in Riemannian space). We denote the collection of N linear subspaces by $\mathcal{P} = \{P_i\}_{i=1}^N$. 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 [17]. One popularly used kernel [17,18,31] is the Projection kernel originated from the principle angles between two subspaces given by (see Fig. 2):

$$\mathcal{K}_{i,j}^{proj.-poly.} = (\gamma \cdot ||P_i^T P_j||_F^2)^\alpha, \tag{6}$$

where $\mathcal{K}_{i,j}^{proj.-poly.}$ is an element in the kernel matrix $\mathcal{K}$. The corresponding mapping is $\Phi_{proj.} = P_i P_i^T$. Then a form of RBF kernel [49] can be generated using $\Phi_{proj.}$ via:

$$\mathcal{K}_{i,j}^{proj.-rbf} = exp(-\gamma ||\Phi_{proj.}(P_i) - \Phi_{proj.}(P_j)||_F^2). \tag{7}$$

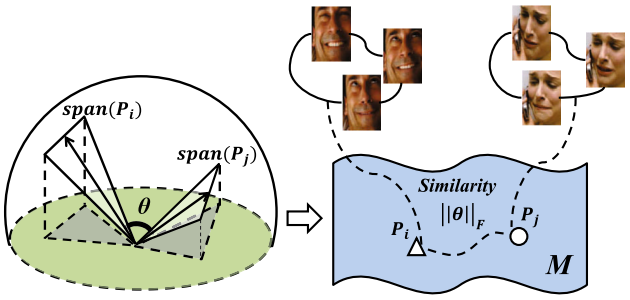

Fig. 2 An illustration of principal angles of linear subspaces and their projection metric distances on Grassmann manifold $Gr(r,d)$

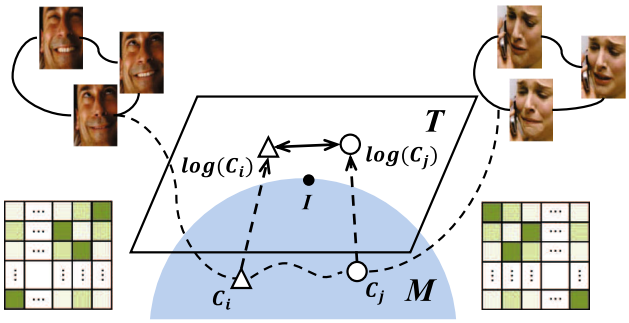

Fig. 3 An illustration of mapping covariance matrices from the SPD Riemannian manifold M to the tangent space T (vector space) at the point of identity matrix I on M

2.3.2 Kernels for covariance matrix

The $d \times d$ SPD matrices, i.e. non-singular covariance matrices $\mathcal{C} = \{C_i\}_{i=1}^N$, can be formulated as data points on SPD Riemannian manifold [37]. A commonly used distance metric for SPD matrices is the Log-Euclidean distance (LED) [2]. Based on LED, [50] 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 Fig. 3).

$$\mathcal{K}_{i,j}^{LED-poly.} = (\gamma \cdot trace[log(C_i) \cdot log(C_j)])^\alpha. \tag{8}$$

The mapping corresponding to $\mathcal{K}_{i,j}^{LED-poly.}$ is given by $\Phi_{LED} = log(C_i)$. Similarly a form of RBF kernel [49] can be generated using Φ_{LED} by:

$$\mathcal{K}_{i,j}^{LED-rbf} = exp(-\gamma ||\Phi_{LED}(C_i) - \Phi_{LED}(C_j)||_F^2). \tag{9}$$

2.3.3 Kernels for Gaussian distribution

The space of d-dimensional multivariate Gaussians is a Riemannian manifold and can be embedded into the space of symmetric positive definite (SPD) matrices [33], 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:

$$\mathcal{N}(\mu, \Sigma) \sim G = |\Sigma|^{-\frac{1}{d+1}} \begin{bmatrix} \Sigma + \mu\mu^T & \mu \\ \mu^T & 1 \end{bmatrix} \tag{10}$$

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

$$\mathcal{K}_{i,j}^{LED-poly.} = (\gamma \cdot trace[log(G_i) \cdot log(G_j)])^\alpha. \tag{11}$$

$$\mathcal{K}_{i,j}^{LED-rbf} = exp(-\gamma ||\Phi_{LED}(G_i) - \Phi_{LED}(G_j)||_F^2). \tag{12}$$

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

An SVM classifier in the kernel space is given by

$$f(x) = \overrightarrow{w}^{*T} \Phi(x) + b^*, \tag{13}$$

where $\Phi(x)$ is the mapping (e.g. $\Phi_{Proj.}$ and Φ_{LED}). The weight vector $\overrightarrow{w}^*$ and bias b^* are given by

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

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:

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

For this work, we employ the Liblinear [13] implementation for optimization.

2.4.3 Partial least squares

Similar to Logistic Regression above, we also apply the partial least square classifier [52] to the kernel matrices. We exploit it in a one-vs-all manner to especially deal with the difficult and confusion categories as in [31].

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

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

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 such that

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

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:

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

thus we can predict $\widehat{Y} = X\beta$ [39]. 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 determined to be the recognition result.

Since PLS optimization problem is quite related to canonical correlation analysis (CCA) [19], it is interesting to compare these two techniques here, as what is done in several previous work [42]. According to their objective function, CCA tends to maximize the correlation between the latent scores, while PLS tends to correlate the latent score of regressor and response as well as captures the variations presenting in the regressor/response space too. Since CCA only correlates the latent score, it may not be able to generalize well to unseen testing points and even may fail to differentiate between training samples in the latent space under some kind of special conditions.

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) [12]. A weighted term λ is introduced at decision level for video–audio fusion:

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

Similarly, the category corresponding to the maximum value of the score vector is chosen as the recognition result.

3 Experiments

3.1 EmotiW 2013/2014 challenge

The emotion recognition in the wild challenge (EmotiW) [9,10] 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. The database in the challenge is the acted faical expression in wild (AFEW) [11], 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 (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 EmotiW organizers. All images are resized to 64×64 pixels. As introduced in Sect. 2.1, three kinds of image features are employed on the aligned faces: HOG, Dense SIFT, and DCNN feature.

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

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

	An	Di	Fe	Ha	Ne	Sa	Su	Total
(a) EmotiW 2013								
Train	58	40	50	65	63	52	52	380
Val	59	50	54	62	55	64	52	396
Test	54	49	33	50	48	43	35	312
(b) EmotiW 2014								
Train	92	66	66	105	102	82	54	567
Val	59	39	44	63	61	59	46	371
Test	58	26	46	81	117	53	26	407

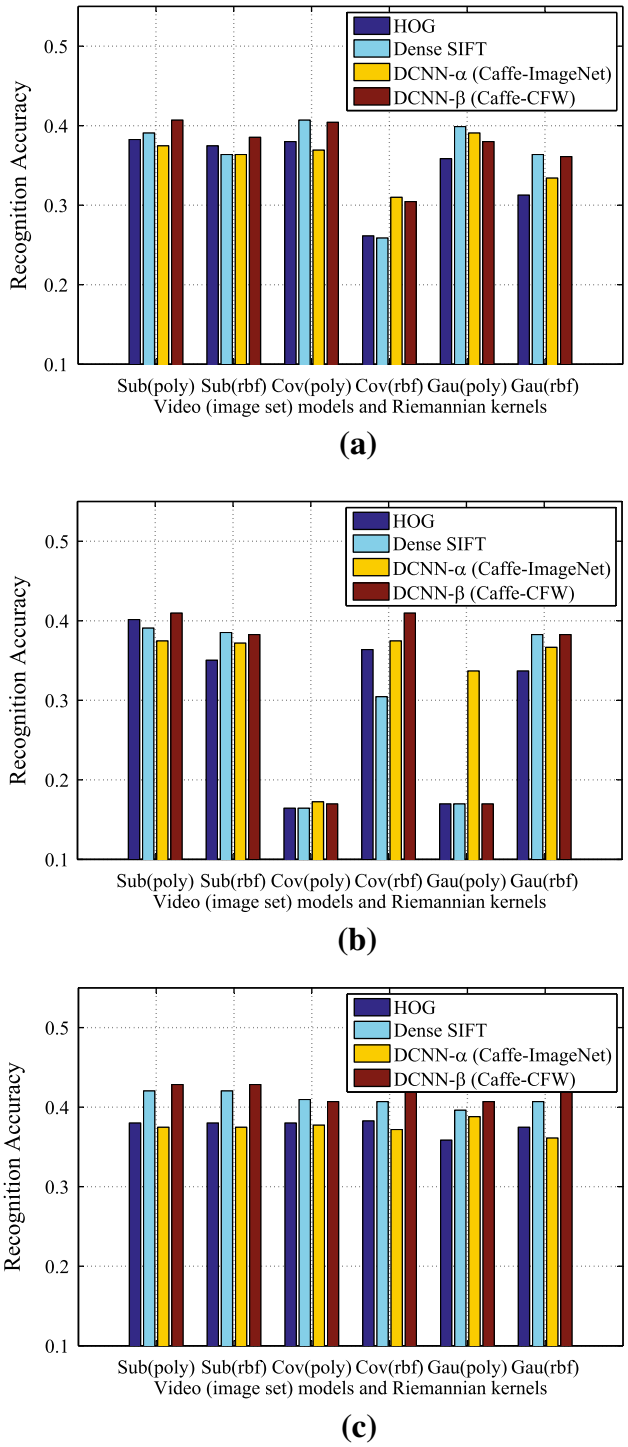

Fig. 4 Emotion recognition accuracy on validation set of the EmotiW 2014 data based on different classifiers. **a** Kernel SVM. **b** Logistic regression. **c** Partial least squares

Table 2 Emotion recognition accuracy on validation set of the EmotiW 2014 data based on different image features

	Linear subspace		Covariance matrix		Gaussian distribution	
	proj-poly	proj-rbf	LED-poly	LED-rbf	g-LED-poly	g-LED-rbf
(a) HOG						
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						
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-α (Caffe-ImageNet)						
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-β (Caffe-CFW)						
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

Bold indicates the best performance

the whole image, we have $128 \times 49 = 6272$ dimensions feature.

For DCNN, we employ the Caffe [22] implementation, which is commonly used in several latest works [14,44]. Two types of DCNN models are trained by feeding different training data: ImageNet ILSVRC-2012 [7] and Celebrity Faces in the Wild (CFW) [56]. 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 [24], 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 1520 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 $256 \times 7 \times 7 = 12{,}544$ nodes' values of the last convolutional layer are used for final features.

PCA dimension reduction was conducted on the high-dimensional image features mainly for consideration of computational cost. We had empirically chosen the ratio=0.90 to achieve a tradeoff between performance and computation complexity.

3.3 Results comparisons

3.3.1 Results on EmotiW 2014 data

We conduct extensive evaluation on EmotiW 2014 challenge data. The emotion recognition results on validation set based on different classifiers are illustrated in Fig. 4. For each single classifier, the DCNN features have shown promising performance on the task, especially the feature extracted by Caffe trained on CFW has achieved better results than the specific hand-crafted feature HOG and Dense SIFT.

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. As shown, PLS achieves the best performance for its one-vs-all manner which deals with each category individually and thus can pay specific attention to those difficult and confusion categories.

The overall recognition results are obtained by one-vs-all PLS classifier using decision-level fusion over different kernels. As presented in Sect. 2.4.4, an equally weighted linear fusion is conducted among the prediction scores based on the six Riemannian kernels with different features, and the weight parameter for video–audio fusion is set as $\lambda = 0.3$ in the final submission. Different fusion strategies and their corresponding results on validation and test sets are listed in Tables 3 and 4.

The confusion matrix of the final submission method are shown in Fig. 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

Table 3 Emotion recognition accuracy on validation set of the EmotiW 2014 data based on multiple kernel methods fusion via PLS

Features	Accuracy (Val)
HOG	38.01
Dense SIFT	43.94
DCNN-α (Caffe-ImageNet)	39.35
DCNN-β (Caffe-CFW)	43.40
HOG + Dense SIFT	44.47
HOG + Dense SIFT + DCNN-α	44.74
HOG + Dense SIFT + DCNN-β	**45.28**

Bold indicates the best performance

Table 4 Emotion recognition accuracy on both validation and test set of the EmotiW 2014 data based on Audio–Video (A + V) fusion via PLS

Methods	Accuracy	
	Val	Test
Audio (OpenSMILE [12])	30.73	–
A + V(HOG + Dense SIFT)	46.36	46.68
A + V(HOG + Dense SIFT + DCNN-α)	46.90	47.91
A + V(HOG + Dense SIFT + DCNN-β)	**48.52**	**50.37**

Bold indicates the best performance

	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise
Angry	0.85	0.03	0.00	0.02	0.05	0.05	0.00
Disgust	0.10	0.18	0.03	0.28	0.33	0.05	0.03
Fear	0.27	0.07	0.27	0.14	0.11	0.09	0.05
Happy	0.05	0.00	0.00	0.83	0.10	0.03	0.00
Neutral	0.13	0.00	0.02	0.08	0.70	0.07	0.00
Sad	0.14	0.03	0.07	0.24	0.29	0.22	0.02
Surprise	0.17	0.04	0.28	0.09	0.33	0.02	0.07

(a)

	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise
Angry	0.81	0.03	0.00	0.05	0.10	0.00	0.00
Disgust	0.12	0.04	0.04	0.35	0.23	0.15	0.08
Fear	0.26	0.00	0.24	0.11	0.20	0.15	0.04
Happy	0.09	0.00	0.01	0.64	0.11	0.15	0.00
Neutral	0.08	0.02	0.05	0.09	0.63	0.12	0.01
Sad	0.11	0.00	0.04	0.25	0.25	0.34	0.02
Surprise	0.12	0.00	0.19	0.08	0.35	0.19	0.08

(b)

Fig. 5 Confusion matrices of the final submission method. **a** Validation set. **b** Test set

Table 5 Performance comparisons of participants in the second EmotiW 2014 challenge [9]

Participants	Accuracy	
	Val	Test
ICT, CAS [32]	48.52	**50.37**
BNU [45]	45.55	47.17
HKPU [4]	40.21	45.21
Bogazici [23]	44.20	44.23
Ulm [36]	–	41.77
Oulu [20]	45.82	41.52
Kielce [15]	–	37.84
Munchen [38]	36.97	35.27
Baseline [9]	33.15	33.66

Bold indicates the best performance

are provided for learning and testing compared with 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 those in lab-controlled data.

Table 5 compares the overall classification accuracy of EmotiW 2014 participants with the video/audio baselines and among each other. Our final submission (denoted as "ICT, CAS" in the figure) achieves 50.37 % on test set, with a significant gain of 16.7 % above the challenge baseline 33.66 %, and wins the first place in the challenge.

3.3.2 Results on EmotiW 2013 data

We also evaluate the proposed method on EmotiW 2013 data and conduct comparisons with our former work [31] submitted to EmotiW 2013 challenge. Using the same settings introduced in Sect. 3.2, we can obtain the recognition results on validation set based on different features as illustrated in Table 6. According to the experience on EmotiW 2014 data, we only employ the DCNN-β features and Partial Least Squares for classification. Due to the emotional labeling inaccuracy of EmotiW 2013 data [9], the recognition performance degrade significantly from ∼40 to ∼30 %. Moreover, the three different image features (i.e. HOG, Dense SIFT, and DCNN-β) obtain similar accuracies (see Tables 6 and 7), which is different from the results on EmotiW 2014 data (please revisit Table 2).

Similar to the EmotiW 2014 data, we demonstrate the fusion results of six different Riemannian kernels on validation set in Table 7 and the final multi-modal (i.e. audio-video) fusion results in Table 8. We can see that the best performance achieved on EmotiW 2013 data is 38.64 %, with a gain of 2.8 % above the results 35.86 % reported in our former work [31]. Accordingly, the confusion matrix is shown in Fig. 6.

Table 6 Emotion recognition accuracy on validation set of the EmotiW 2013 data based on different image features

	Linear subspace		Covariance matrix		Gaussian distribution	
	proj-poly	proj-rbf	LED-poly	LED-rbf	g-LED-poly	g-LED-rbf
(a) HOG						
Partial least squares	29.54	29.54	33.84	**34.85**	32.58	32.07
(b) Dense SIFT						
Partial least squares	32.07	32.07	33.33	33.33	**35.10**	33.84
(c) DCNN-β (Caffe-CFW)						
Partial least squares	32.84	32.84	32.58	**33.08**	32.58	31.31

Bold indicates the best performance

Table 7 Emotion recognition accuracy on validation set of the EmotiW 2013 data based on multiple kernel methods fusion via PLS classifiers

Features	Accuracy (Val)
HOG	34.85
Dense SIFT	35.10
DCNN-β (Caffe-CFW)	33.33
HOG + Dense SIFT + DCNN-β	**36.11**

Bold indicates the best performance

Table 8 Emotion recognition accuracy on validation set of the EmotiW 2013 data based on Audio–Video (A + V) fusion

Methods	Accuracy (Val)
Audio (OpenSMILE [12])	24.24
A + V (HOG)	35.86
A + V (Dense SIFT)	37.12
A + V (DCNN-β)	36.36
A + V (HOG + Dense SIFT + DCNN-β)	**38.64**

Bold indicates the best performance

	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise
Angry	0.56	0.00	0.12	0.07	0.15	0.08	0.02
Disgust	0.26	0.04	0.04	0.26	0.34	0.02	0.04
Fear	0.13	0.06	0.22	0.15	0.20	0.07	0.17
Happy	0.26	0.03	0.08	0.18	0.31	0.02	0.13
Neutral	0.20	0.02	0.00	0.22	0.47	0.05	0.04
Sad	0.16	0.05	0.13	0.23	0.25	0.11	0.08
Surprise	0.12	0.04	0.04	0.29	0.40	0.02	0.10

Fig. 6 Confusion matrix of the final results on EmotiW 2013 validation set

3.3.3 Results on CK+ and MMI databases

To further evaluate the generalization ability, In this section, we also perform experiments on two well-known lab-controlled databases: CK+ and MMI.

Table 9 The number of samples for each expression in CK+ and MMI database

	An	Co	Di	Fe	Ha	Sa	Su	Total
CK+	45	18	59	25	69	28	83	327
MMI	31	–	32	28	42	32	40	205

The CK+ database [35] consists of 593 sequences from 123 subjects, which is an extended version of Cohn-Kanade (CK) database. The image sequence vary in duration from 10 to 60 frames and incorporate the onset (neutral face) to peak formation of the facial expression. The validated expression labels are only assigned to 327 sequences which are found to meet the criteria for 1 of 7 discrete emotions (anger, contempt, disgust, fear, happiness, sadness, and surprise) based on facial action coding system (FACS). We adopt leave-one-subject-out cross-validation (118 folds) following the general setup in [35].

The MMI database [46] includes 30 subjects of both sexes and ages from 19 to 62. In the database, 213 sequences have been labeled with six basic expressions, in which 205 sequences were captured frontal view. Each of the sequence reflects the whole temporal activation patterns (onset → apex → offset) of a single facial expression type. In our experiments, all of these data were used and also a person-independent tenfold cross-validation was conducted as in several previous work [16,30]. Compared to CK+, MMI is thought to be more challenging for the subjects pose expressions non-uniformly and usually wear some accessories (e.g. glasses, moustache). The number of samples for each expression in CK+ and MMI is illustrated in Table 9.

The average emotion recognition accuracy results based on different image features (i.e. HOG, Dense SIFT. Note that, for fair comparison, we don't employ DCNN feature which involve external data.) are illustrated in Tables 10 and 11 for CK+ and MMI respectively. We can observe that the covariance matrix based modeling achieves the best performance on both databases for different features (much better than linear subspace based modeling). However on wild data (please revisit Table 2), linear subspace shows better perfor-

Table 10 Average emotion recognition accuracy on CK+ database based on different image features

	Linear subspace		Covariance matrix		Gaussian distribution	
	proj-poly	proj-rbf	LED-poly	LED-rbf	g-LED-poly	g-LED-rbf
(a) HOG						
Logistic regression	78.01	74.63	14.29	**88.81**	35.42	73.72
Partial least squares	80.09	83.61	**92.69**	92.37	83.28	83.61
(b) Dense SIFT						
Logistic regression	78.64	78.39	39.30	**87.17**	54.98	79.81
Partial least squares	84.39	84.39	**91.63**	91.38	85.63	86.40

Bold indicates the best performance

Table 11 Average emotion recognition accuracy on MMI database based on different image features

	Linear subspace		Covariance matrix		Gaussian distribution	
	proj-poly	proj-rbf	LED-poly	LED-rbf	g-LED-poly	g-LED-rbf
(a) HOG						
Logistic regression	59.73	58.16	13.91	**63.11**	30.50	55.08
Partial least squares	59.27	59.27	**66.42**	63.58	61.09	60.85
(b) Dense SIFT						
logistic regression	58.82	55.36	13.91	**63.15**	47.88	**63.15**
Partial least squares	57.76	57.76	**67.78**	63.94	57.71	57.74

Bold indicates the best performance

Table 12 Fusion results of different Riemannian kernels and different image features on two databases (a) CK+ database, (b) MMI database

Features	Average Acc	Overall Acc
(a) CK+ database		
HOG	93.42	95.72
Dense SIFT	92.20	95.11
HOG + Dense SIFT	**94.82**	**96.64**
(b) MMI database		
HOG	66.64	70.24
Dense SIFT	67.78	70.24
HOG + Dense SIFT	**71.38**	**74.63**

Bold indicates the best performance

mance compared to covariance and gaussian modeling. The reason might be that the expression variations in these two datasets are from lab controlled setting, so there exist a whole procedure for performing the exaggerated expression, which favors the COV model.

The fusion results of different Riemannian kernels and different image features are listed in Table 12. Two criteria, average recognition accuracy (per category) and overall classification accuracy are measured for performance comparison. The corresponding confusion matrices are shown in Fig. 7. We can observe that even on lab-controlled data, it is still hard to recognize some difficult and confusion emotion categories such as "contempt" and "fear" (especially serious for "fear" on MMI database).

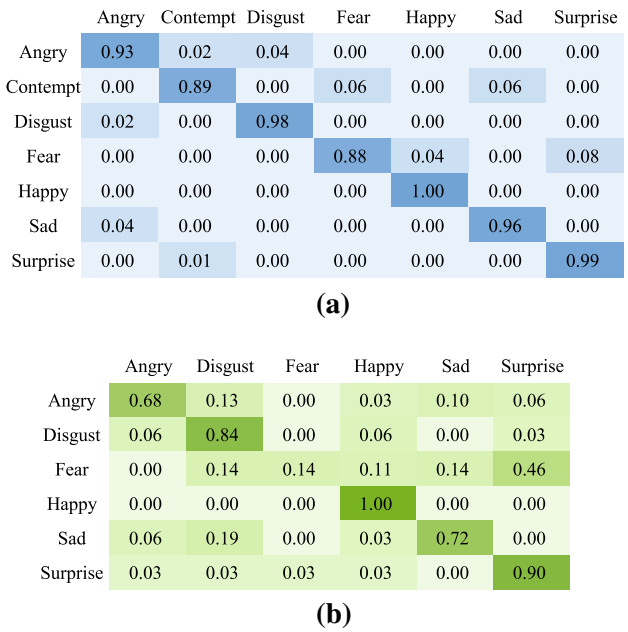

Fig. 7 Confusion matrices of the final results. **a** CK+ database. **b** MMI database

Finally, we also compare the our results with several state-of-the-art methods as in Table 13. The results demonstrate that the proposed framework outperforms the existing methods significantly.

Table 13 Performance comparison with state-of-the-art methods (a) CK+ database, (b) MMI database

Methods	Average Acc	Overall Acc
(a) CK+ database		
AAM [35]	83.3	88.3
HMM [51]	83.5	–
ITBN [51]	86.3	88.8
3DCNN [29]	78.0	85.9
3DCNN-DAP [29]	87.9	92.4
MCF [5]	89.4	–
Ours	**94.82**	**96.64**
(b) MMI database		
HMM [51]	51.5	–
ITBN [51]	59.7	60.5
3DCNN [29]	50.7	53.2
3DCNN-DAP [29]	62.2	63.4
Ours	**71.38**	**74.63**

Bold indicates the best performance

4 Conclusions

In this paper, we propose a method for video-based emotion recognition in real-world condition. Each emotion video clip is simply regarded as an image set and different kinds of image set models are explored 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 measuring distance metrics. At last, a score-level fusion of classifiers learned based on different kernel methods and different modalities is conducted for producing final recognition result. The method is evaluated on EmotiW 2013/2014 data and has achieved very 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.

Acknowledgments This work is partially supported by 973 Program under contract No. 2015CB351802, Natural Science Foundation of China under contracts Nos. 61390511, 61222211, 61379083, and Youth Innovation Promotion Association CAS No. 2015085.

References

1. Arandjelovic O, Shakhnarovich G, Fisher J, Cipolla R, Darrell T (2005) Face recognition with image sets using manifold density divergence. IEEE Comput Vis Pattern Recognit 1:581–588
2. Arsigny V, Fillard P, Pennec X, Ayache N (2007) Geometric means in a novel vector space structure on symmetric positive-definite matrices. SIAM J Matrix Anal Appl 29(1):328–347
3. Chang C-C, Lin C-J (2011) Libsvm: a library for support vector machines. ACM Trans Intell Syst Technol (TIST) 2(3):27
4. Chen J, Chen Z, Chi Z, Fu H (2014) Emotion recognition in the wild with feature fusion and multiple kernel learning. ACM Int Conf Multimodal Interact 1:508–513
5. Chew SW, Lucey S, Lucey P, Sridharan S, Conn JF (2012) Improved facial expression recognition via uni-hyperplane classification. IEEE Comput Vis Pattern Recognit 1:2554–2561
6. Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. IEEE Comput Vis Pattern Recognit 1:886–893
7. Deng J, DongW, Socher R, Li L-J, Li K, Fei-Fei L (2009) Imagenet: a large-scale hierarchical image database. IEEE Comput Vis Pattern Recognit 1:248–255
8. Dhall A, Asthana A, Goecke R, Gedeon T (2011) Emotion recognition using phog and lpq features. IEEE Autom Face Gesture Recognit 1:878–883
9. Dhall A, Goecke R, Joshi J, Sikka K, Gedeon T (2014) Emotion recognition in the wild challenge 2014: baseline, data and protocol. ACM Int Conf Multimodal Interact 1:461–466
10. Dhall A, Goecke R, Joshi J, Wagner M, Gedeon T (2013) Emotion recognition in the wild challenge 2013. ACM Int Conf Multimodal Interact 1:509–516
11. Dhall A, Goecke R, Lucey S, Gedeon T (2012) Collecting large, richly annotated facial-expression databases from movies. IEEE MultiM 19(3):34–41
12. Eyben F, Wöllmer M, Schuller B (2010) Opensmile: the munich versatile and fast open-source audio feature extractor. ACM Int Conf Multimed 1:1459–1462
13. Fan R-E, Chang K-W, Hsieh C-J, Wang X-R, Lin C-J (2008) Liblinear: a library for large linear classification. J Mach Learn Res 9:1871–1874
14. Girshick R, Donahue J, Darrell T, Malik J (2013) Rich feature hierarchies for accurate object detection and semantic segmentation. arXiv preprint; arXiv:1311.2524
15. Grosicki M (2014) Neural networks for emotion recognition in the wild. ACM Int Conf Multimodal Interact 1:467–472
16. Guo Y, Zhao G, Pietikäinen M (2012) Dynamic facial expression recognition using longitudinal facial expression atlases. Eur Conf Comput Vis 1:631–644
17. Hamm J, Lee DD (2008) Grassmann discriminant analysis: a unifying view on subspace-based learning. Int Conf Mach Learn 1:376–383
18. Harandi MT, Sanderson C, Shirazi S, Lovell BC (2011) Graph embedding discriminant analysis on grassmannian manifolds for improved image set matching. IEEE Comput Vis Pattern Recognit 1:2705–2712
19. Hotelling H (1936) Relations between two sets of variates. Biometrika 321–377
20. Huang X, He Q, Hong X, Zhao G, Pietikainen M (2014) Improved spatiotemporal local monogenic binary pattern for emotion recognition in the wild. ACM Int Conf Multimodal Interact 1:514–520
21. Hubel DH, Wiesel TN (1968) Receptive fields and functional architecture of monkey striate cortex. J Physiol 195(1):215–243
22. Jia Y (2013) Caffe: an open source convolutional architecture for fast feature embedding. http://caffe.berkeleyvision.org
23. Kaya H, Salah AA (2014) Combining modality-specific extreme learning machines for emotion recognition in the wild. ACM Int Conf Multimodal Interact 1:487–493
24. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. Adv Neural Inf Process Syst 1:1097–1105
25. Lan Z-Z, Jiang L, Yu S-I, Rawat S, Cai Y, Gao C, Xu S, Shen H, Li X, Wang Y, et al (2013) Cmu-informedia at trecvid 2013 multimedia event detection. TRECVID 2013 Workshop
26. LeCun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. Proc IEEE 86(11):2278–2324

27. Li P, Wang Q, Zhang L (2013) A novel earth mover's distance methodology for image matching with gaussian mixture models. IEEE Int Conf Comput Vis 1:1689–1696
28. Liu M, Li S, Shan S, Chen X (2013) Au-aware deep networks for facial expression recognition. IEEE Autom Face Gesture Recognit 1:1–6
29. Liu M, Li S, Shan S, Wang R, Chen X (2014) Deeply learning deformable facial action parts model for dynamic expression analysis. Asian Conf Comput Vis 1:143–157
30. Liu M, Shan S, Wang R, Chen X (2014) Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition. IEEE Comput Vis Pattern Recognit 1:1749–1756
31. Liu M, Wang R, Huang Z, Shan S, Chen X (2013) Partial least squares regression on grassmannian manifold for emotion recognition. ACM Int Conf Multimodal Interact 1:525–530
32. Liu M, Wang R, Li S, Shan S, Huang Z, Chen X (2014) Combining multiple kernel methods on riemannian manifold for emotion recognition in the wild. ACM Int Conf Multimodal Interact 1:494–501
33. Lovrić M, Min-Oo M, Ruh EA (2000) Multivariate normal distributions parametrized as a riemannian symmetric space. J Multivar Anal 74(1):36–48
34. Lowe DG (2004) Distinctive image features from scale-invariant keypoints. Int J Comput Vision 60(2):91–110
35. Lucey P, Cohn JF, Kanade T, Saragih J, Ambadar Z, Matthews I (2010) The extended cohn-kanade dataset (ck+): a complete dataset for action unit and emotion-specified expression. IEEE Comput Vis Pattern Recognit Workshops 1:94–101
36. Meudt S, Schwenker F (2014) Enhanced autocorrelation in real world emotion recognition. ACM Int Conf Multimodal Interact 1:502–507
37. Pennec X, Fillard P, Ayache N (2006) A riemannian framework for tensor computing. Int J Comput Vision 66(1):41–66
38. Ringeval F, Amiriparian S, Eyben F, Scherer K, Schuller B (2014) Emotion recognition in the wild: incorporating voice and lip activity in multimodal decision-level fusion. ACM Int Conf Multimodal Interact 1:473–480
39. Rosipal R, Krämer N (2006) Overview and recent advances in partial least squares. Subspace Latent Struct Featur Select 34–51
40. Shakhnarovich G, Fisher JW, Darrell T (2002) Face recognition from long-term observations. Eur Conf Comput Vis 1:851–865
41. Shan C, Gong S, McOwan PW (2009) Facial expression recognition based on local binary patterns: a comprehensive study. Image Vision Comput 27(6):803–816
42. Sharma A, Jacobs DW (2011) Bypassing synthesis: pls for face recognition with pose, low-resolution and sketch. IEEE Comput Vis Pattern Recognit 1:593–600
43. Sikka K, Dykstra K, Sathyanarayana S, Littlewort G, Bartlett M (2013) Multiple kernel learning for emotion recognition in the wild. ACM Int Conf Multimodal Interact 1:517–524
44. Simonyan K, Zisserman A (2014) Two-stream convolutional networks for action recognition in videos. arXiv preprint; arXiv:1406.2199
45. Sun B, Li L, Zuo T, Chen Y, Zhou G, Wu X (2014) Combining multimodal features with hierarchical classifier fusion for emotion recognition in the wild. ACM Int Conf Multimodal Interact 1:481–486
46. Valstar M, Pantic M (2010) Induced disgust, happiness and surprise: an addition to the mmi facial expression database. Int Conf Lang Resour Eval 1:65–70
47. Valstar M, Schuller B, Smith K, Eyben F, Jiang B, Bilakhia S, Schnieder S, Cowie R, Pantic M (2013) Avec 2013: the continuous audio/visual emotion and depression recognition challenge. ACM Int Workshop Audio/Vis Emot Chall 1:3–10
48. Valstar MF, Jiang B, Mehu M, Pantic M, Scherer K (2011) The first facial expression recognition and analysis challenge. IEEE Autom Face Gesture Recognit 1:921–926
49. Vemulapalli R, Pillai JK, Chellappa R (2013) Kernel learning for extrinsic classification of manifold features. IEEE Comput Vis Pattern Recognit 1:1782–1789
50. Wang R, Guo H, Davis LS, Dai Q (2012) Covariance discriminative learning: a natural and efficient approach to image set classification. IEEE Comput Vis Pattern Recognit 1:2496–2503
51. Wang Z, Wang S, Ji Q (2013) Capturing complex spatio-temporal relations among facial muscles for facial expression recognition. IEEE Comput Vis Pattern Recognit 1:3422–3429
52. Wold H (1985) Partial least squares. In: Kotz S, Johnson NL (eds) Encyclopedia of statistical sciences, vol 6. Wiley, New York, pp 581–591
53. Yamaguchi O, Fukui K, Maeda K-I (1998) Face recognition using temporal image sequence. IEEE Autom Face Gesture Recognit 1:318–323
54. Yang P, Liu Q, Metaxas DN (2007) Boosting coded dynamic features for facial action units and facial expression recognition. IEEE Comput Vis Pattern Recognit 1:1–6
55. Zeng Z, Pantic M, Roisman GI, Huang TS (2009) A survey of affect recognition methods: audio, visual, and spontaneous expressions. Pattern Anal Mach Intell IEEE Trans 31(1):39–58
56. Zhang X, Zhang L, Wang X-J, Shum H-Y (2012) Finding celebrities in billions of web images. Multimed IEEE Trans 14(4):995–1007
57. Zhao G, Pietikainen M (2007) Dynamic texture recognition using local binary patterns with an application to facial expressions. Pattern Anal Mach Intell IEEE Trans 29(6):915–928
58. Zhong L, Liu Q, Yang P, Liu B, Huang J, Metaxas DN (2012) Learning active facial patches for expression analysis. IEEE Comput Vis Pattern Recognit 1:2562–2569