Partial Least Squares Regression on Grassmannian Manifold for Emotion Recognition

Mengyi Liu, Ruiping Wang, Zhiwu Huang, Shiguang Shan, Xilin Chen
Key Lab of Intelligence Information Processing
Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China
{mengyi.liu, ruiping.wang, zhiwu.huang, shiguang.shan, xilin.chen}@vipl.ict.ac.cn

ABSTRACT

In this paper, we propose a method for video-based human emotion recognition. For each video clip, all frames are represented as an image set, which can be modeled as a linear subspace to be embedded in Grassmannian manifold. After feature extraction, Class-specific One-to-Rest Partial Least Squares (PLS) is learned on video and audio data respectively to distinguish each class from the other confusing ones. Finally, an optimal fusion of classifiers learned from both modalities (video and audio) is conducted at decision level. Our method is evaluated on the Emotion Recognition In The Wild Challenge (EmotiW 2013). The experimental results on both validation set and blind test set are presented for comparison. The final accuracy achieved on test set outperforms the baseline by 26%.

Categories and Subject Descriptors

I.5.4 [**Pattern Recognition**]: Applications—*computer vision, signal processing*; I.4.m [**IMAGE PROCESSING AND COMPUTER VISION**]: Miscellaneous

General Terms

Experimentation; Performance; Algorithms

Keywords

Emotion Recognition; Grassmannian Manifolds; Partial Least Squares Regression; EmotiW 2013 Challenge

1. INTRODUCTION

In the recent years, automatic emotion recognition has become a popular and challenging problem due to its broad applications, such as Human-Computer Interaction (HCI), multimedia analysis, surveillance, and so on. Early stage research mostly focused on emotion analysis from single static facial images [11]. The investigated approaches can be categorized into two classes: feature-based and template-based. However, As human emotions appear to be temporally dynamic processes, some recent research tend to take use of the image sequences or video clips for improving the emotion recognition performance [19, 20, 18, 9]. For instance, Yang et al [18] extracted dynamic features from image sequences and adopted boosting to learn the classifier. Zhao et al [20] encoded spatial-temporal information in each facial image sequences using LBP-TOP features. As demonstrated in their experiments, various types of dynamic information (e.g. dynamical appearance-based and dynamical geometric-based information) are crucial for modeling emotion variations in the recognition tasks.

However, facial dynamics extraction from successive video frames requires accurate alignment for eliminating the rigid motion of pose and retaining the non-rigid motion of facial muscles. It is quite challenge especially when dealing with real-world data due to the large variations caused by uncontrolled environment. Inspired by some recent research on image-set-based classification [8, 15, 16, 1], we attempt to model all the video frames as an emotional image set (suppose that each video clip represents a single emotion from a single person). Thus the set of images can be modeled by a linear subspace which characterizes the specific person's emotion. For distance metric, the collection of subspaces are treated as points on Grassmannian manifold, and point-to-point distances are induced from the Grassmannian kernels [6, 7].

In this paper, we design a video-based emotion recognition method especially for real-world scenario. First we perform automatic preprocessing to purify the aligned data by filtering out non-face or misaligned-face images. Then video-based feature extraction is conducted using subspace-learning and Grassmannian kernels. After that, One-to-Rest Partial Least Squares (PLS) is learned on video and audio features to distinguish each class from the other confusing ones. In the end, we conduct a multi-modality information fusion at decision level to further improve the performance. An overview of the proposed method is demonstrated in Figure 1. In Section 2, we will detail the key steps of our method.

2. THE PROPOSED METHOD

2.1 Data Preprocessing

Due to large variations caused by pose, illumination, expression, face detection and alignment algorithms can hardly work well on real-world data. Image normalization according to these unreliable results may lead to non-face or

ICMI'13, December 9–13, 2013, Sydney, Australia

http://dx.doi.org/10.1145/2522848.2531738 .

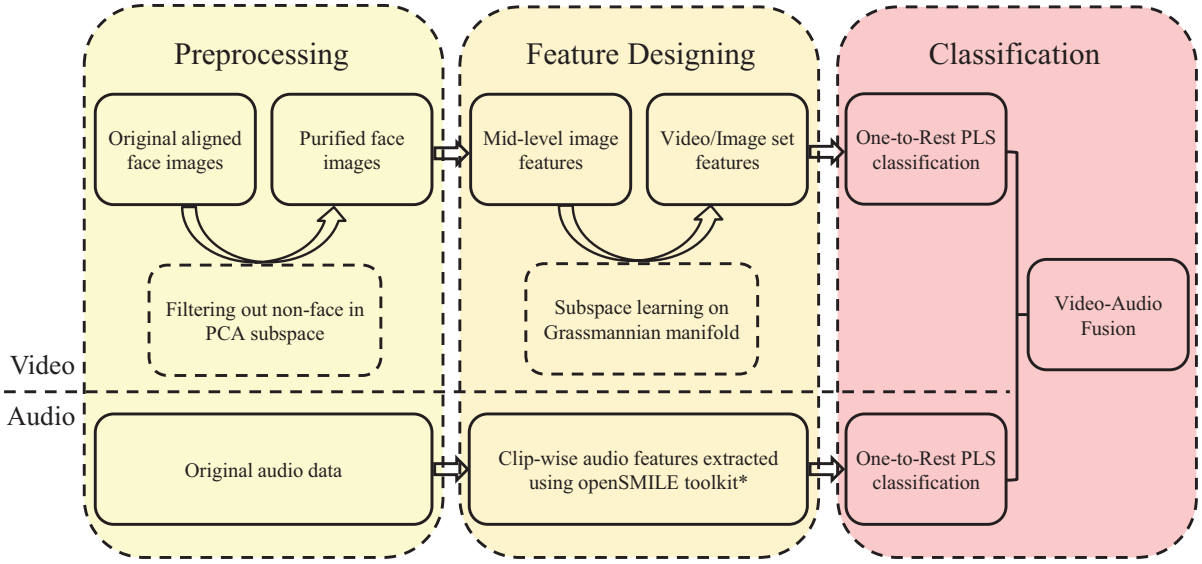

Figure 1: The procedure of the proposed method. (*openSMILE toolkit [4])

misaligned-face samples, which become noises in subspace learning. To purify the "dirty" aligned data, we attempt to construct face subspace defined by "Eigenfaces" [13], in which the non-face image cannot reconstruct itself as well as face image.

Suppose $X = \{x_1, x_2, ..., x_n\}$ (x_i belongs to R^D) is the training set consisting of n face images, we calculate the PCA projection W by preserving relatively low energy of the origin images. The basic idea is that face images do not change radically when projected into the face space, while the projection of non-face images appear quite different [13]. Thus we consider the mean reconstruction error:

$$MeanErr_t = \frac{1}{M} \cdot ||x_t - W^T W x_t||^2 \quad (1)$$

As non-face images are tend to have larger reconstruction error (see Figure 2 and Figure 3), we can set threshold to filter out these non-face samples.

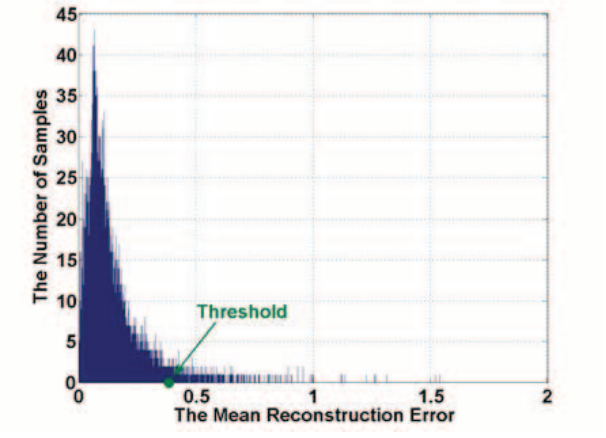

Figure 2: The distribution of mean reconstruction error on training set in EmotiW 2013 Challenge.

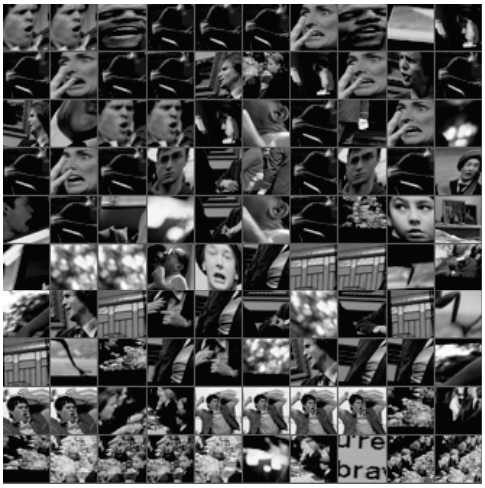

Figure 3: An example of 100 samples with largest mean reconstruction error. Most are non-face images or misalignment results.

2.2 Image/Video Feature Design

2.2.1 Image Feature

As real-world data exhibit large variations, here we employ a kind of mid-level feature introduced in [10] for image representation instead of the simple gray intensity feature used in most image set classification methods [8, 15, 7, 14]. As shown in Figure 4, the feature extraction process includes a convolution layer and a max-pooling layer. First we learn a dictionary (bank of filters) using the dense sampled local patches from training images, and calculate the response value of each filters by convolution. Second, max-pooling is performed over adjacent spacial blocks on each filter map, to generate a robust representation invariant to image translation. For more details please refer to [10].

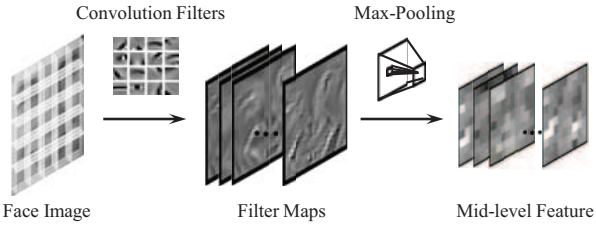

Figure 4: The image feature extraction process.

2.2.2 *Video (Image Set) Feature*

Based on the single image features obtained above, we can construct a linear subspace $P_i \in R^{D\times r}$ via SVD over the ith image set, where r is the dimensionality of the subspace. Thus the whole data can be modeled as a collection of subspaces, which are also the points on Grassmannian manifold M (see Figure 5), denoted by $P = \{P_i\}_{i=1}^{N}$, where N is the number of sets (points). The similarity between two points on Grassmannian manifold can be measured by projection kernels [6]:

$$k_{i,j}^{[proj]} = ||P_i^T P_j||_F^2 \tag{2}$$

Combining the projection kernel with a more complex canonical correlation kernel is proved to be effective in [7]. For P_i and P_j, the canonical correlation kernel is defined as:

$$k_{i,j}^{[CC]} = max_{a_p \in span(P_i)} max_{b_q \in span(P_j)} a_p^T b_q \tag{3}$$

We express the linear combination of two Grassmannian kernels $k_{i,j}^{[proj]}$ and $k_{i,j}^{[CC]}$ using a tunable parameter α:

$$k_{i,j}^{[com]} = k_{i,j}^{[proj]} + \alpha k_{i,j}^{[CC]} \tag{4}$$

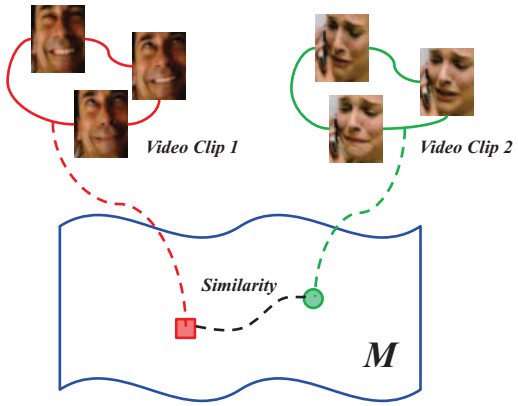

Figure 5: Image sets can be described by linear subspaces and represented as points on Grassmannian manifold.

According to the combined kernel $K^{[com]} = \{k_{i,j}^{[com]}\} \in R^{N\times N}$, the ith row contains similarities between the ith set (point) and all sets (points) in data, which can be treat as a feature vector of the ith set. For each set (point) in training data or test data, we calculate its similarities to all the training sets (points). Thus the training kernels $K_{train} \in R^{N_{train}\times N_{train}}$ and test kernels $K_{test} \in R^{N_{test}\times N_{train}}$ are constructed as the video/image set feature matrices.

2.3 One-to-Rest PLS Classifier

Partial Least Squares (PLS) algorithm can model the relation between sets of observed variables by means of latent variables [12]. Recent work [5, 14] have adapted it to recognition tasks and obtained promising results. In our method, PLS is applied in a One-to-Rest manner to especially deal with several difficult and confusion classes (e.g. disgust, fear, and sad).

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

$$\begin{aligned} X &= TP^T + E \\ Y &= UQ^T + F \end{aligned} \tag{5}$$

where T and U are matrices containing the extracted latent vectors, P and Q represent loadings, E and F are residuals. Based on the nonlinear iterative partial least squares (NIPALS) algorithm [17], PLS finds weight vectors w and v such that

$$[cov(t,u)]^2 = max_{|w|=|v|=1}[cov(Xw, Yv)]^2 \tag{6}$$

where t and u are the column vectors of T and U respectively, $cov(t,u)$ is the sample covariance. With the obtained latent vectors, the regression coefficients between X and Y is estimated by:

$$B = W(P^TW)^{-1}T^TY = X^TU(T^TXX^TU)^{-1}T^TY \tag{7}$$

which results in $\widehat{Y} = XB$. For each test sample, we can obtain c regression values from all the One-to-Rest classifiers. The category corresponding to the maximum value is decided to be the recognition result.

2.4 Video-Audio Fusion

We performed the One-to-Rest PLS on both video and audio data, the regression result in two modalities are $Fit^{video} \in R^{N_{test}\times c}$ and $Fit^{audio} \in R^{N_{test}\times c}$. We conduct a linear fusion at decision level by introducing a weighted term λ:

$$Fit^{fusion} = (1-\lambda)Fit^{video} + \lambda Fit^{audio} \tag{8}$$

3. EXPERIMENTS

3.1 EmotiW 2013 Challenge

The Emotion Recognition In The Wild Challenge (EmotiW 2013) [2] is to define a common platform for evaluation of emotion recognition methods in real-world conditions. The database in challenge is the AFEW [3], which has been collected from movies showing close-to-real-world conditions and divided into three sets for the challenge: training, validation and testing. The task is to classify a sample audio-video clip into one of the seven emotion categories: Anger, Disgust, Fear, Happiness, Neutral, Sadness and Surprise. Participants are free to use either video or audio modality or both, to report an overall results in the end. The labels of the testing set are unknown. Participants can learn their

models only on training set, and optimize the parameter on validation set.

3.2 Experimental Parameters

3.2.1 Basic Image Feature

We simply use the aligned face images provided by EmotiW 2013 organizers. After the data preprocessing described in Section 2, the purified images are resized to 32x32 pixels. According to the parameters in [10], for the first layer, we sample 6-by-6 pixel patches with a stride of 1 pixel on the resized images. Thus each image contains 27-by-27 patches. We learn $K = 100$ filters using K-means among these patches, then the image can be represented by a 27-by-27-by-K dimension vector after convolution. For the second layer, we apply max-pooling over 3-by-3 adjacent patches on each filter map, in the end this yields 9-by-9-by-100 features. To reduce the high dimension, we perform PCA to retain only 1500-dimension as the basic image feature.

3.2.2 The Fusion Weights of Grassmannian Kernels

To evaluate the combination of two Grassmannian kernels $k_{i,j}^{[proj]}$ and $k_{i,j}^{[CC]}$, we tune the parameter α and obtain different results as shown in Figure 6. The trends on training and validation sets are inconsistent. The $k_{i,j}^{[CC]}$ seems useless on predicting training set, but useful on validation set. For test set, we select the best parameter on validation set, i.e. $\alpha = 2^{-5}$ in (4).

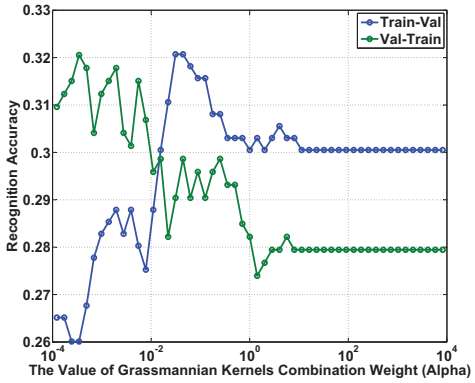

Figure 6: The recognition accuracy on training and validation sets with different combination weights of Grassmannian kernels.("Train-Val" means using Train set to learn model and Validation set to report performance. "Val-Train" means the opposite.)

3.2.3 The Dimensions of One-to-Rest PLS

An important parameter in our method is the dimensions of One-to-Rest PLS classifier. To optimize this parameter, we perform cross-validation on the provided training and validation sets. The experimental results are shown in Figure 7 and Figure 8.

As demonstrated in the figures, we can see the One-to-Rest PLS classifier is very easy to overfit on training set. We can find that when dimension is 10 on video and 5 on audio, it can achieve consistently good performance on both training and validation sets. The settings are also applied on test sets.

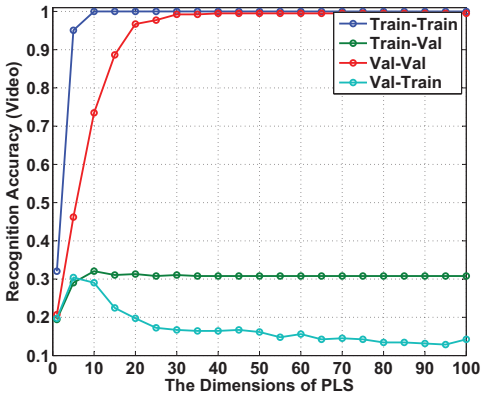

Figure 7: The recognition accuracy on video modality with different dimensions of PLS.

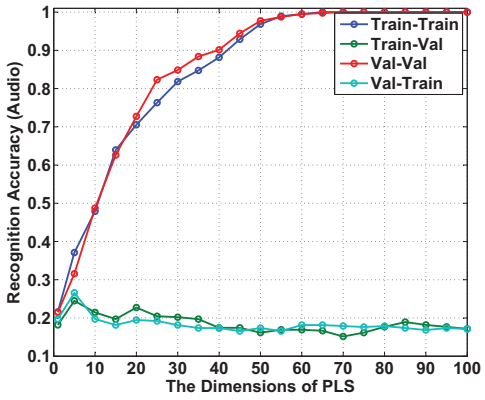

Figure 8: The recognition accuracy on audio modality with different dimensions of PLS.

3.2.4 Video-Audio Fusion Weights

At decision level, we conduct video-audio fusion to further improve the recognition performance. To decide the fusion weights, we tune the λ in (8) and obtain the following results on training and validation sets respectively.

As the inconsistent trend shown in Figure 9, we choose the best parameter on validation set, i.e. $\lambda = 0.25$.

3.3 Results Comparisons

We demonstrate all of our results on validation and test sets (note that the test results are evaluation feedbacks from EmotiW 2013 organizers) in Table 1. For video only part, we compare our method with graph embedding based Grassmannian Discriminant Analysis proposed in [7]. For audio-video fusion, we conduct both feature-level fusion and decision-level fusion. In the last column, We also show the improvement by using purified data. The final accuracy achieved by our method outperform the baseline [2] (which adopted LBP-TOP features for video and kernel SVM for classification) significantly on both validation and test sets.

The confusion matrix of final test results are shown in Table 2. We can see that "Happy" and "Angry" are easy to be distinguished from other expressions, but it is still hard to do well on some difficult and confusion emotion classes such as "Disgust" and "Sad".

Table 1: The comparison results on both validation and test sets.

Performance Comparison		Audio only	Video only		Audio + Video			
					Original data			Purified data
		One-to-Rest PLS	Grassmannian Discriminant Analysis	Grassmannian Kernels + One-to-Rest PLS	Feature-level fusion		Decision-level fusion	Decision-level fusion
					Multi-class LR	One-to-Rest PLS	One-to-Rest PLS	One-to-Rest PLS
Ours	***Val***	24.49 %	30.81%	32.07%	22.48%	24.24%	34.34%	**35.85%**
	Test*	**--**	**24.04%**	**--**	**--**	**26.28%**	**33.01%**	**34.61%**
Baseline	***Val***	19.95%	27.27%		22.22%			
	Test	22.44%	22.75%		27.56%			

Table 2: The confusion matrix of final test results.

	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise
Angry	64.81	0	1.85	12.96	7.4	0	12.96
Disgust	38.77	6.12	0	18.36	6.12	10.2	20.4
Fear	21.21	0	18.18	9.09	21.21	9.09	21.21
Happy	12	0	2	64	12	0	10
Neutral	25	4.16	0	25	39.58	2.08	4.16
Sad	13.95	0	4.65	41.86	18.6	9.3	11.62
Surprise	25.71	0	5.71	11.42	20	11.42	25.71

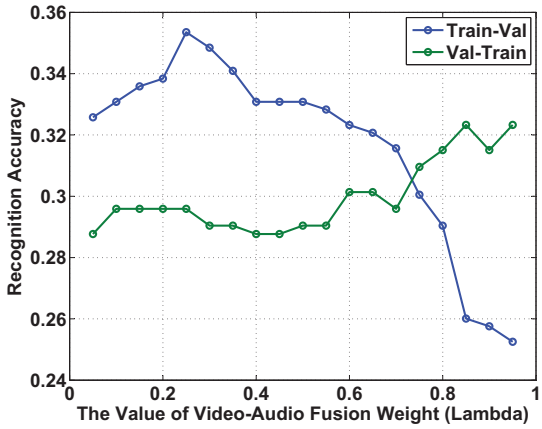

Figure 9: The overall recognition accuracy with different video-audio fusion weights.

4. CONCLUSIONS

In this paper, we propose a method for video-based human emotion recognition in real-world. For each video clip, all frames are treated as an integrated image set, which can be modeled as a linear subspace to be embedded in Grassmannian manifold. In classification, One-to-Rest PLS is applied for both video and audio data, and a fusion of the two modality is conducted at decision level. The method is evaluated on EmotiW 2013 and shows promising result on unseen test data. In the future, we will try to deal with the few difficult categories and design more effective modality fusion strategy to further improve the recognition performance.

5. ACKNOWLEDGMENTS

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

6. REFERENCES

[1] H. Cevikalp and B. Triggs. Face recognition based on image sets. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2567–2573, 2010.

[2] A. Dhall, R. Goecke, J. Joshi, M. Wagner, and T. Gedeon. Emotion recognition in the wild challenge 2013. In *ACM International Conference on Multimodal Interaction (ICMI)*, pages 2496–2503. ACM, 2012.

[3] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon. Collecting large, richly annotated facial-expression databases from movies. *IEEE MultiMedia*, 19(3):0034, 2012.

[4] F. Eyben, M. Wöllmer, and B. Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. In *ACM International Conference on Multimedia*, pages 1459–1462. ACM, 2010.

[5] G. Guo and G. Mu. Simultaneous dimensionality reduction and human age estimation via kernel partial least squares regression. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 657–664, 2011.

[6] 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*, pages 376–383, 2008.

[7] M. T. Harandi, C. Sanderson, S. Shirazi, and B. C. Lovell. Graph embedding discriminant analysis on grassmannian manifolds for improved image set matching. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 2705–2712, 2011.

[8] T.-K. Kim, J. Kittler, and R. Cipolla. Discriminative learning and recognition of image set classes using canonical correlations. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(6):1005–1018, 2007.

[9] S. Koelstra, M. Pantic, and I. Patras. A dynamic texture-based approach to recognition of facial actions and their temporal models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(11):1940–1954, 2010.

[10] M. Liu, S. Li, S. Shan, and X. Chen. Au-aware deep networks for facial expression recognition. In *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, 2013.

[11] M. Pantic and L. J. M. Rothkrantz. Automatic analysis of facial expressions: The state of the art. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(12):1424–1445, 2000.

[12] 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.

[13] M. A. Turk and A. P. Pentland. Face recognition using eigenfaces. In *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on*, pages 586–591, 1991.

[14] R. Wang, H. Guo, L. S. Davis, and Q. Dai. Covariance discriminative learning: A natural and efficient approach to image set classification. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2496–2503, 2012.

[15] R. Wang, S. Shan, X. Chen, and W. Gao. Manifold-manifold distance with application to face recognition based on image set. In *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on*, pages 1–8, 2008.

[16] T. Wang and P. Shi. Kernel grassmannian distances and discriminant analysis for face recognition from image sets. *Pattern Recognition Letters*, 30(13):1161–1165, 2009.

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

[18] P. Yang, Q. Liu, and D. N. Metaxas. Boosting coded dynamic features for facial action units and facial expression recognition. In *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on*, pages 1–6, 2007.

[19] Y. Zhang and J. Qiang. Active and dynamic information fusion for facial expression understanding from image sequences. *Pattern Analysis and Machine Intelligence, IEEE Transaction on*, 27(5):699–714, 2005.

[20] 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.