Multi-view Common Space Learning for Emotion Recognition in the Wild

Jianlong Wu[1,2], Zhouchen Lin[1,2], Hongbin Zha[1,2]
[1]Key Laboratory of Machine Perception (MOE), School of EECS, Peking University, Beijing 100871, P. R. China
[2]Cooperative Medianet Innovation Center, Shanghai Jiao Tong University, Shanghai 200240, P. R. China
{jlwu1992, zlin}@pku.edu.cn, zha@cis.pku.edu.cn

ABSTRACT

It is a very challenging task to recognize emotion in the wild. Recently, combining information from various views or modalities has attracted more attention. Cross modality features and features extracted by different methods are regarded as multi-view information of the sample. In this paper, we propose a method to analyse multi-view features of emotion samples and automatically recognize the expression as part of the fourth Emotion Recognition in the Wild Challenge (EmotiW 2016). In our method, we first extract multi-view features such as BoF, CNN, LBP-TOP and audio features for each expression sample. Then we learn the corresponding projection matrices to map multi-view features into a common subspace. In the meantime, we impose $\ell_{2,1}$-norm penalties on projection matrices for feature selection. We apply both this method and PLSR to emotion recognition. We conduct experiments on both AFEW and HAPPEI datasets, and achieve superior performance. The best recognition accuracy of our method is 55.31% on the AFEW dataset for video based emotion recognition in the wild. The minimum RMSE for group happiness intensity recognition is 0.9525 on HAPPEI dataset. Both of them are much better than that of the challenge baseline.

CCS Concepts

•Computing methodologies → Activity recognition and understanding; Image representations;

Keywords

Emotion Recognition; Multi-view Learning; Common Space Learning; EmotiW 2016 Challenge

1. INTRODUCTION

Automatic emotion recognition attracts more and more attention in computer vision due to its important role in many applications, such as human computer interaction (HCI) and psychological research. Many methods [25, 38] have been proposed for expression recognition during the past decades. However, former researchers mainly focus on static images based emotion recognition under lab-controlled environment. In recent years, with the organizing of several emotion recognition competitions such as Audio Video Emotion Challenges (AVEC) [26] and Emotion Recognition challenge in the Wild [6, 5], video based emotion recognition in the wild has been greatly promoted. Compared with previous static images based recognition under controlled environment, video based emotion recognition in the wild is more challenging as it has large pose and illumination variations caused by uncontrolled real-world environment.

Several methods have been proposed to recognize video based emotion in the wild and achieved very good performance. For instance, Zhao et al. [39] used LBP-TOP features to encode spatial-temporal patterns in dynamic images sequence. Liu et al. [14] used Riemannian manifold kernels to represent each expression video clip. Sikka et al. [22] applied multiple kernel learning to combine different features. Yao et al.[37] encoded facial feature relations with graph structure. Wu et al. [33] extracted multiple features from video clips and fused them based on the partial least square regression (PLSR) [32].

As facial emotion video clips contain much spatial-temporal and multi-modality information, it's important to represent it from multi different views [30]. Another issue is that how to make full use of these multi-view features to improve emotion recognition. Towards these two issues, in this paper, we first extract different kinds of features such as LBP-TOP [39], BoF [12], CNN [19] and audio features. Those features are regarded as various views of emotion samples for better representation. After multi-view features extraction, we propose a common space learning method to utilize multi-view and multi-modality features to improve the classification. In the common space learning method, we use common space projection to measure the relevance among multi-view features, and $\ell_{2,1}$-norm penalty term is used to select relevant and discriminative features. An iterative algorithm is presented to solve the regularized linear regression problem. On the other hand, we also utilize PLSR [32] to calculate the regression score of each view features as [33] did. Finally, we combine the projected features together. An overview of our proposed multi-view common space learning method is shown in Figure 1.

The rest of this paper is organized as follows. We first introduce the extracted multi-view features in Section 2. Then we present the details of multi-view common space learning method and PLSR in Section 3. In Section 4, we

ICMI'16, November 12–16, 2016, Tokyo, Japan
© 2016 ACM. 978-1-4503-4556-9/16/11...$15.00
http://dx.doi.org/10.1145/2993148.2997631

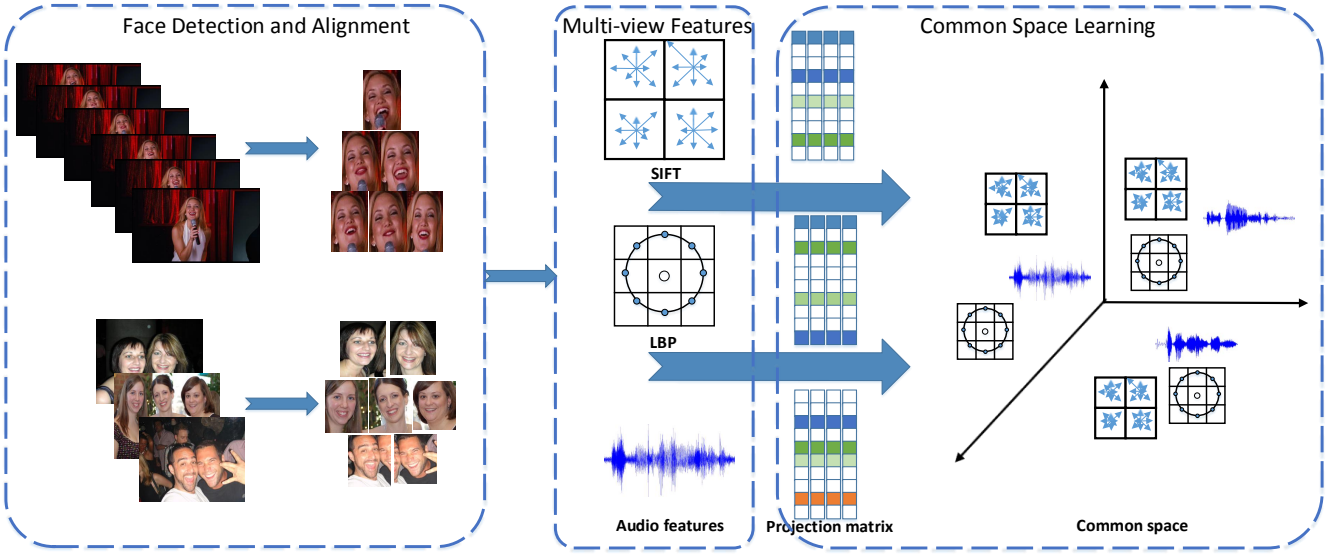


Figure 1: Framework of the proposed multi-view common space learning method. For video or image emotion samples, we first detect and align the face. Then we extract multi-view features, such as SIFT, LBP, audio features and et al. Finally, we learn the projection matrix for each view to map the data into one common space. In the common space, we can combine the mapped multi-view features to improve the recognition.

conduct extensive experiments on both AFEW and HAPPEI datasets. Finally, Section 5 concludes our paper.

2. MULTI-VIEW FEATURES

2.1 Audio Features

For video based emotion recognition, audio information plays an important role [2]. We use the openSMILE toolkit [8], an open-source feature extractor that unites feature extraction algorithms from the speech processing and the Music Information Retrieval communities, to extract audio features based on INTERSPEECH 2010 audio template [21]. We use 21 energy & spectral related functionals and 19 voicing related functionals to extract corresponding low-level descriptors and delta regression coefficients. With another 2 voiced/unvoiced durational features, there are 1582 dimensional features in total. For detailed feature information, please refer to [9].

2.2 LBP-TOP Features

Local Binary Patterns from Three Orthogonal Planes (LBP-TOP) [39], extending from the widely used LBP [18] operator, is proposed to handle the influence of varying rotation and lighting condition on dynamic textures. LBP-TOP considers the co-occurrence statistics of dynamic textures in three directions, concatenating LBP on three orthogonal planes: XY, XT, and YT, where the XY plane provides the spatial texture information, and the XT and YT planes provide information about the spacetime transitions. Features of LBP-TOP [39] are robust to gray-scale and rotations variations. It has been successfully applied to video based emotion recognition while the video can be regarded as a sequence of dynamic facial expression images. We adopt LBP-TOP to extract dynamic features for expression recognition in the wild.

2.3 Features of the BoF Model

Bag of Features(BoF) [3] is one of the most popular and effective image classification frameworks in recent literature. It has achieved the state-of-the-art performance in many classification tasks [12], including the emotion recognition. The commonly used BoF framework generally consists of four basic modules: local features extraction, codebook generation, descriptors encoding and spatial pyramid pooling. In this paper, we use two kinds of coding methods to extract video features.

We first divide each image or each frame of the emotion video into many overlapped grid blocks with a fixed step size, and then extract SIFT [16] features on each block. The extracted SIFT [16] features are invariant to image scale and rotation. Based on the SIFT features, we simply use the classical K-means [15] clustering algorithm to learn the dictionary for encoding. Then we adopt locality-constrained linear coding (LLC) [29] and group saliency coding (GSC) [34], which are two commonly used encoding methods of BoF. The core idea of LLC [29] is to reconstruct features with closest codewords via resolving a least square based optimization problem with locality constraints on the codewords. Different from reconstruction based coding method LLC, GSC [34] is developed from the saliency based coding [11]. For detailed information and comparison of these encoding methods, please refer to [12]. Finally, we use the spatial pyramid matching (SPM) [13] method to partition the image into increasingly finer spatial subregions. Max pooling [36] is adopted to pool all responses on each codeword in a specific subregion into one value. Final representation is obtained by concatenating descriptions of all blocks.

After feature extraction of the BoF [3] model, the dimension of extracted feature vectors is very high, especially when the number of spatial pyramids levels is large. High dimensional features will influence both the efficiency and accuracy of classification. In this case, it is necessary to reduce the

features dimension. We use the principle component analysis (PCA) [17] to reduce dimension. The core idea of PCA [17] is to maximize the total variance of projection.

2.4 CNN Features

Convolutional Neural Networks (CNN) have achieved the state-of-the-art performance in many computer vision tasks, such as face recognition [19]. One of the most famous CNN architectures is the deep convolutional network [23] designed for ImageNet Challenge 2014. For image based emotion recognition, we adopt the 16 weight layers CNN network presented in [23] to extract CNN features. This network contains 13 convolution and 3 fully-connected(FC) layers. For detailed architecture information, please refer to [23]. We directly use the parameters trained on the face dataset of [19], which contains 2.6 million faces of 2622 celebrities. In our experiment, the features after the first FC layer are used to represent the images.

3. MULTI-VIEW ANALYSIS

In this section, we first present the framework to learn the common space for multi-view features. Then, we give an iterative algorithm to optimize the linear regression problem. Finally, we introduce another regression method PLSR for emotion recognition.

3.1 Multi-view Common Space Learning

As we have multi-view features of emotion clips or images, there are mainly two important issues that we need to take into consideration for improving the recognition accuracy. On the one hand, we need to measure the similarity among various views features. On the other hand, we need to select the relevant and discriminative features during learning. In this case, we propose a multi-view common space learning (MCSL) method with $\ell_{2,1}$-norm penalty to achieve the above two requirements.

3.1.1 Problem Formulation of MCSL

Let $X_q = [x_1^q, x_2^q, \cdots, x_n^q] \in \mathbb{R}^{d_q \times n}, q = 1, 2, \cdots, M$ denote the q-th view labeled data matrices, where n is the number of train samples, M is the total number of views and d_q is the dimension of each feature in the q-th view. Each pair$\{x_i^1, x_i^2, \cdots, x_i^M\}, i = 1, 2, \cdots, n$ represents M views features of i-th sample and belongs to the same class. Let $Y = [y_1, y_2, \cdots, y_n]^T \in \mathbb{R}^{n \times c}$ denote the class label matrix, where c is the number of classes. The class indication matrix Y satisfies that $y_{ij} = 1$ if data point x_i belongs to the j-th class, $y_{ij} = 0$ otherwise. The common space learning method aims to map the multi-view features into the common space defined by the class labels by learning a projection matrix for each view features. In the meantime, we impose $l_{2,1}$-norm on the projection matrices for feature selection. Then, we can get the objective function for common space learning:

$$\min_{W_1, \cdots, W_M} \sum_{q=1}^{M} \|X_q^T W_q - Y\|_F^2 + \lambda \sum_{q=1}^{M} \|W_q\|_{2,1}, \qquad (1)$$

where $W_q \in \mathbb{R}^{d_q \times c}$ is the projection matrix for q-th view data. For matrix $U \in \mathbb{R}^{n \times m}$, let $U^{(i)}$ denote its i-th row. The Frobenius norm of the matrix U is defined as $\|U\|_F = \sqrt{\sum_{i=1}^{n} \|U^{(i)}\|_2^2}$. The $\ell_{2,1}$-norm [1] of U is defined as the sum of the ℓ_2-norm of the rows of M: $\|U\|_{2,1} = \sum_{i=1}^{n} \|U^{(i)}\|_2$. In the objective function Eq. 1, the first term is multi-view linear regression, which can help us to map all different view data into one common space and compute their similarity. The second term is used for feature selection. As the $\ell_{2,1}$-norm encourages the sparsity of W's columns, the discriminative features that are relevant to the class label will get large weights.

3.1.2 Optimization Algorithm for MCSL

The optimization of Eq. 1 is equal to optimize the following M subproblems:

$$\min_{W_q} \|X_q^T W_q - Y\|_F^2 + \lambda \|W_q\|_{2,1}, \quad q = 1, 2, \cdots, M. \qquad (2)$$

As the sub-problem in Eq. 2 contains a nonsmooth regularization terms of $\ell_{2,1}$-norm, it's complicated to solve W_q directly. Thus, we use an alternative iterative algorithm to solve this problem. When the ℓ_2-norm of i-th row of q-th view projection matrix equals to zero, that is $\|W_q^{(i)}\|_2 = 0$, then Eq. 1 is not differentiable. Following [28, 31], we can introduce a small perturbation ε to ℓ_2-norm of each row, and $\|W_q\|_{2,1}$ can be replaced with $\sum_{i=1}^{n} \sqrt{\|W_q^{(i)}\|_2^2 + \varepsilon}$. Here, ε is usually set to be a small constant value. It is easy to verify that when $\varepsilon \to 0$, the derived minimization problem is obviously equal to the problem Eq. 2.

Then, we can get the derivative of the objective function in Eq. 2 with respect to W_q, and set it to zero. We can obtain that:

$$X_q X_q^T W_q - X_q Y + \lambda D_q W_q = 0, \quad q = 1, 2, \cdots, M, \qquad (3)$$

where D_q is a diagonal matrix with the i-th diagonal element as $\frac{1}{\sqrt{\|W_q^{(i)}\|_2^2 + \varepsilon}}$. Further we have:

$$W_q = (X_q X_q^T + \lambda D_q)^{-1} X_q Y, \quad q = 1, 2, \cdots, M. \qquad (4)$$

We need to mention that D_q is dependent on W_q which is still unknown variable. Under this circumstance, we use an iterative algorithm to solve the problem Eq. 4. The algorithm is described in Algorithm 1.

Algorithm 1 An Iterative Algorithm for Multi-view Common Space Learning (MCSL)

Input:

Multi-view data $X_q \in \mathbb{R}^{d_q \times n},\ q = 1, 2, \cdots, M$,
class label matrix $Y \in \mathbb{R}^{n \times c}$.

Initialize:

Set $t = 1$ and initialize W_q^1 by solving:
$\min_{W_q} \|X_q^T W_q - Y\|_F^2,\ q = 1, 2, \cdots, M$.

while not converge **do**

1. Calculate the diagonal matrix D_q^t for $q = 1, 2, \cdots, M$, where the i-th diagonal element is $\frac{1}{\sqrt{\|W_q^{(i)}\|_2^2 + \varepsilon}}$.
2. Compute W_q^{t+1} for $q = 1, 2, \cdots, M$ according to Eq. 4.
3. $t = t + 1$.

end while

Output:

Projection matrices $W_q \in \mathbb{R}^{d_q \times c},\ q = 1, 2, \cdots, M$.

In Algorithm 1, we first initialize the projection matrix W_q^1 by solving the simple linear regression problem without

(a) Some example frames of expression videos in the wild

(b) Some example images for group happiness intensity recognition in the wild

Figure 2: Some example images of AFEW 6.0 and HAPPEI datasets.

penalty. This can be achieved by $W_q^1 = (X_q X_q^T)^{-1} X_q Y, \quad q = 1, 2, \cdots, M$. While the algorithm does not converge, we compute the diagonal matrix D_q^t in step 1. In step 2, we compute the optimal projection matrices W_q^{t+1} for each view data.

The computation cost of MCSL is very low. According to Algorithm 1, since it's easy to compute the diagonal matrix D_q^t with W_q, the computation cost of step 1 is trivial. In step 2, instead of directly computing the matrix inverse with cubic complexity, we can update W_q^{t+1} by solving a system of linear equations with quadratic complexity. Let d represents the largest dimension value among d_q, that is $d = \max(d_q),\ q = 1, 2, \cdots, M$. Then the time complexity of MCSL is about $\mathcal{O}(kd^2)$, where k is the total number of iterations until Algorithm 1 converges. In our experiments, it takes less than 10 iterations before the algorithm converges. Therefore, the whole algorithm of MCSL can be solved very efficiently.

3.2 PLSR

Besides the MCSL, we also apply the PLSR to emotion recognition in the wild. We adopt the same PLSR manner as that in [14, 33]. PLSR can be regarded as the combination of PCA [17] and canonical correspondence analysis (CCA) [24] . For each category, we design an one-vs-all PLSR to calculate the regression value. Let X be feature variables and Y be the 0-1 labels. According to [20], PLSR decomposes these variables into:

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

where T and U contain the latent vectors, P and Q are orthogonal loading matrices, and E and F are residuals. PLSR tries to find the optimal weights w_x and w_y to get the maximum covariance such that:

$$[cov(t,u)]^2 = \max_{|w_x|=|w_y|=1} [cov(Xw_x, Yw_y)]^2. \tag{6}$$

Then we can get the regression projection matrix B [20] as:

$$B = X^T U (T^T X X^T U)^{-1} T^T Y. \tag{7}$$

The regression score can be estimated by:

$$S = XB. \tag{8}$$

Following the above process, we can calculate the regression value of test samples.

3.3 Combination Strategy

For each view features, we utilize both the MCSL and the PLSR methods to learn their corresponding projection matrix, which is used to map the data into the common space defined by the label matrix. The projected result of each view data represents the confidence score that this sample belongs to each class. Then we adopt the score level combination method and assign specific weight to the confidence score of each view data:

$$S^{comb} = \sum_{q=1}^{M} \alpha_q S^q, \tag{9}$$

where S^{comb} denotes the combined confidence score, and S^q represents the confidence score computed on q-th view features. The weights which vary from view to view are relevant to the performance of each view. We learn the optimal weights on the validation dataset. The final predicted label of emotion sample is the category with the largest combined confidence score.

4. EXPERIMENTAL RESULTS

4.1 EmotiW 2016 Challenge

The emotion recognition in the wild challenge (EmotiW 2016) [5] contains two sub-challenges. One is video based emotion recognition challenge and the other is image based group level happiness recognition challenge. All the data of EmotiW 2016 are collected in the wild, which is very close to real world conditions. We take part in both two sub-challenges to evaluate the performance of our methods.

The dataset for video based emotion recognition (VER) is AFEW [7] 6.0 dataset, which includes 773 train video clips, 383 validation video clips and 593 test video clips. All train and validation video clips are collected from movies. The test

Table 1: Recognition accuracy comparison between PLSR and MCSL on validation dataset of AFEW with single view features for video based emotion recognition.BoF^{LLC} stands for the LLC based BoF method and BoF^{GSC} denotes the GSC based BoF method.

	Audio features	LBP-TOP features	LLC based BoF (BoF^{LLC})	GSC based BoF (BoF^{GSC})
PLSR	33.96 %	37.74%	47.44%	45.82%
MCSL	33.42%	38.54%	45.82%	49.06%

Table 2: Performance comparison of different methods on both validation and test datasets for video based emotion recognition.

Methods		Accuracy	
		Val	Test
Baseline (LBP-TOP)		38.81%	40.47%
Audio+ Video	Multi-view features with MCSL	49.87%	51.43%
	Multi-view features with PLSR	49.87%	51.94%
	Multi-view features with PLSR (Customized)	–	55.31%

data consists of both movies data and reality TV data, which is the major difference between EmotiW 2016 and earlier years' challenge. Figure 2(a) shows some example images of seven expressions taken from video clips. The major task of VER is to classify each video clip into seven basic expression types, such as angry (AN), disgust (DI), fear (FE), happy (HA), neutral (NE), sad (SA) and surprise (SU).

HAPPEI [4] dataset is collected for group level happiness recognition sub-challenge. This dataset contains 1500 train images, 1138 validation images and 496 test images. Each image contains group of people. The task is to infer the happiness mood intensity of the group as a whole on a scale from 0 to 5. In Figure 2(b), we present some example images of six levels happiness.

4.2 Parameter Setting

For our MCSL method, we fine tune the parameter λ in Eq. 1 by searching the grid of $\{10^{-3}, 10^{-2}, \cdots, 10^{2}, 10^{3}\}$. λ is set to 1 in all experiments.

For video based emotion recognition, organizers first apply pre-trained face models [40] to detect faces in each video clip. Then, they adopt the intraface tracking library [35] to align the detected facial images. And each facial image is aligned to size 128×128.

LBP-TOP [39] features are extracted from non-overlapping spatial 4×4 blocks by the organizers. We directly use the aligned facial images as well as the extracted audio and LBP-TOP features provided by organizers.

Similar to LBP-TOP features, BoF features are also extracted for both video based emotion recognition and group happiness intensity recognition. We divide each facial image into overlapped blocks with step 1 and size 16×16. Then we use the Vlfeat [27] to extract 128-dimensional SIFT features in each block. Based on the SIFT features, we learn the dictionary by the K-means [15] clustering algorithm with 1024 centres. Both the nearest neighbours number for LLC [29] and groups number for GSC [34] are set to 5. During the pooling process, we employ the SPM with levels of $[1 \times 1, 2 \times 2, 4 \times 4, 8 \times 8]$. We adopt max pooling to pool the features in each region of the image for group happiness intensity recognition and across all the frames of each video for video based emotion recognition. Under the above settings, dimension of the final BoF representation for each sample is $1024 \times 85 = 87040$. We further use PCA [17] to reduce dimension with principle components ratio 97%.

We only extracted CNN features for image based group level happiness intensity recognition. We adopt the 16 weight layers CNN network presented in [23]. For detailed architecture information, please refer to [23]. We directly use the parameters trained on the face dataset of [19], which contains 2.6 million faces of 2622 celebrities. In our experiments, we use the features after the first FC layer to represent the images. The dimension of each CNN feature is 4096.

In the combination process, as the multi-view features for two tasks are different, the combination methods for these two tasks are different. For video based emotion recognition:

$$S_V^{comb} = \alpha_1 S^{audio} + \alpha_2 S^{LBP-TOP} + \alpha_3 S^{LLC} + \alpha_4 S^{GSC}. \tag{10}$$

we set the optimal weights as $\alpha_1 = 0.25$, $\alpha_2 = 0.15$, $\alpha_3 = 1.00$ and $\alpha_4 = 0.50$ for PLSR, and $\alpha_1 = 0.95$, $\alpha_2 = 0.16$, $\alpha_3 = 1.00$ and $\alpha_4 = 0.11$ for MCSL, respectively. For group happiness intensity recognition:

$$S_G^{comb} = \beta_1 S^{GSC} + \beta_2 S^{LBP-TOP} + \beta_3 S^{CNN}. \tag{11}$$

We set the optimal weights as $\beta_1 = 1.0$, $\beta_2 = 0.2$ and $\beta_3 = 1.5$ for PLSR, and $\beta_1 = 1.0$, $\beta_2 = 0.44$ and $\beta_3 = 0.08$ for MCSL, respectively. All the weights are learned on the validation dataset.

4.3 Video Based Emotion Recognition

In Table 1, we compare the performance of PLSR and MCSL with different single view features of validation dataset. We can see that the recognition accuracy of MCSL is comparable to that of PLSR. Especially on GSC based BoF features, the proposed MCSL can achieve 49.06%, which is the highest recognition accuracy on single view data. Table 2 shows the multi-view results of MCSL and PLSR on both validation and test datasets. The baseline recognition accuracy is 38.81% and 40.47% on validation and test datasets, respectively. With multi-view features of both audio and video, both the MCSL and PLSR reach 49.87% on validation dataset. On the test dataset, the MCSL achieves 51.43%, which is a little lower than the performance 51.94% of PLSR.

Table 3: Recognition RMSE comparison between PLSR and MCSL on validation dataset of HAPPEI with single view features for group happiness intensity recognition. Strategy 1 uses the label corresponding to the largest combined confidence score as the predicted intensity, while strategy 2 sums the product of each intensity and its corresponding confidence score.

		CNN	LBP-TOP	GSC based BoF
Strategy 1	PLSR	0.5463	0.7379	0.5408
	MCSL	0.5426	0.6196	0.5399
Strategy 2	PLSR	0.4225	0.8543	0.4510
	MCSL	0.4170	0.6205	0.3987

Table 4: RMSE comparison of PLSR and MCSL on validation and test dataset of HAPPEI dataset for group happiness intensity recognition.

Methods	RMSE	
	val	test
Baseline (LBP-TOP)	0.78	1.3
Multi-view features with PLSR	0.3686	0.9536
Multi-view features with MCSL	0.3666	0.9525

The recognition results on both two datasets largely surpass the baseline.

In Table 4, we present the confusion matrices of MCSL and PLSR on the test dataset, which are similar to each other. We can easily find that angry, happy and neutral expressions are easily to be recognized correctly, while other expressions such as disgust, fear, sad and surprise are more likely to be misclassified. We also notice that it is difficult to recognize surprise and disgust expression, and fear expression samples are easily misclassified to surprise for PLSR. This phenomenon might relate to few train samples of surprise and high correlation between surprise and fear. We need to note that total sample numbers of fear and surprise expressions on the test dataset are 66 and 28, respectively. By analysing the statistics in Figure 4(b), we further customize our method slightly. For predicted surprise and disgust expressions of PLSR, we use the category with the second largest confidence score instead of the largest value as the predicted label. The corresponding recognition result is shown in Figure 3, and the overall recognition accuracy become 55.3%.

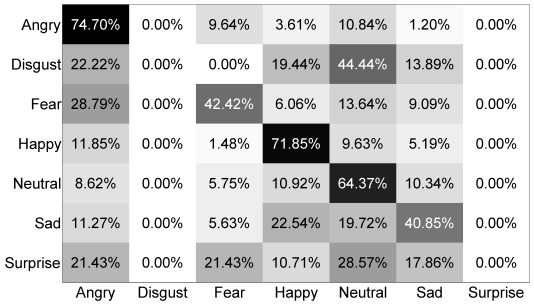


Figure 3: Confusion matrix of customized method on the test dataset of AFEW for video based emotion recognition.

4.4 Group Happiness Intensity Recognition

For group happiness intensity recognition [4, 10], we first detect and alignment the facial images in each group image. Then, we classify the happiness level of each face with the proposed methods. The group level happiness is simply decided by the mean confidence score of all faces in this image. The performance is evaluated by the Root mean square error(RMSE), which is defined by:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(I_{pre} - I_{gnd})^2}{n}}, \tag{12}$$

where n is number of test samples, I_{pre} denotes the predicted intensity and I_{gnd} stands for the ground truth intensity of test sample.

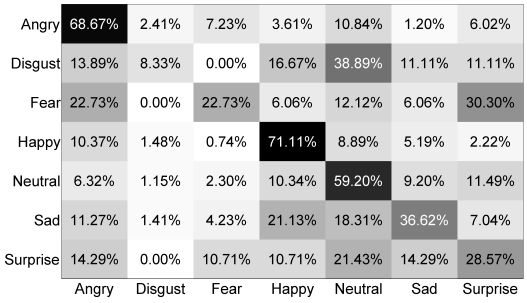
Besides using the label corresponding to the largest combined confidence score as the predicted intensity, we propose another strategy to decide the predicted intensity based on the combined confidence score. As the intensity of group happiness is continuous, we sum the product of each intensity and its corresponding confidence score, that is:

$$I_{pre} = round(\sum_{I=0}^{5} I * P_I), \tag{13}$$

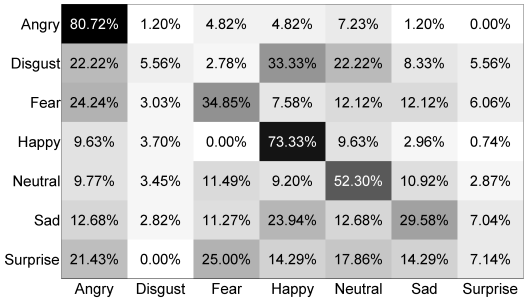
where I denotes the intensity label, P_I denotes the probability that this sample belongs to intensity I, and $round$ means choosing the nearest intensity as the predicted label I_{pre}.

Table 3 shows the recognition results of PLSR and MCSL on validation dataset of HAPPEI with single view features under two different strategy. It can be easily seen from the table that the result with the second strategy is much better than that of the first one. On the other hand, the performance of the proposed MCSL is better than that of the PLSR.

In Table 4, we compare the multi-view performance of the PLSR and MCSL on HAPPEI dataset for group happiness intensity recognition. On both validation and test datasets of HAPPEI, the RMSE of MCSL is slightly better than that of PLSR. On the test dataset, the minimum RMSE of MCSL is 0.9525, which is much better than the baseline 1.3.



(a) Confusion matrix of PLSR on the test dataset



(b) Confusion matrix of MCSL on the test dataset

Figure 4: Confusion matrices of the PLSR and MCSL on the test dataset of AFEW for video based emotion recognition.

5. CONCLUSIONS

In this paper, we propose a multi-view common space learning method for emotion recognition in the wild. We first extract audio features, LBP-TOP, BoF and CNN features as multi-view features of emotion sample. Then we learn the projection matrix for each view to map the features into one common space defined by the class label matrix. In the meantime, $\ell_{2,1}$-norm is imposed for feature selection. In the projected common space, we assign different weights to different view results and combine them together to improve the recognition. We apply both MCSL and PLSR for emotion recognition. We evaluate the performance of our methods on both the AFEW 6.0 dataset and the HAPPEI dataset as part of the fourth Emotion Recognition in the Wild Challenge (EmotiW 2016). Our method achieves very good performances on both two sub-challenges. As multi-view features can well represent the video and image emotion samples for emotion recognition in the wild, in the future, we will further investigate how to make full use of multi-view information to facilitate recognition.

6. ACKNOWLEDGMENTS

Z. Lin is supported by National Basic Research Program of China (973 Program) (grant no. 2015CB352502), National Natural Science Foundation of China (NSFC) (grant nos. 61272341 and 61231002), and Microsoft Research Asia Collaborative Research Program.

7. REFERENCES

[1] A. Argyriou, T. Evgeniou, and M. Pontil. Convex multi-task feature learning. *Machine Learning*, 73(3):243–272, 2008.

[2] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan. Analysis of emotion recognition using facial expressions, speech and multimodal information. In *Proceedings of the ACM International Conference on Multimodal Interfaces*, pages 205–211, 2004.

[3] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *Proceedings of European Conference Computer Vision Workshop on Statistical Learning in Computer Vision*, pages 1–2, 2004.

[4] A. Dhall, R. Goecke, and T. Gedeon. Automatic group happiness intensity analysis. *IEEE Transactions on Affective Computing*, 6(1):13–26, 2015.

[5] A. Dhall, R. Goecke, J. Joshi, J. Hoey, and T. Gedeon. Emotiw 2016: Video and group-level emotion recognition challenges. In *Proceedings of the ACM International Conference on Multimodal Interaction (ICMI)*, 2016.

[6] A. Dhall, R. Goecke, J. Joshi, K. Sikka, and T. Gedeon. Emotion recognition in the wild challenge 2014: Baseline, data and protocol. In *Proceedings of the ACM International Conference on Multimodal Interaction (ICMI)*, pages 461–466, 2014.

[7] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon. Collecting large, richly annotated facial-expression databases from movies. *IEEE Transactions on Multimedia*, 19(3):34–41, 2012.

[8] F. Eyben, F. Weninger, F. Gross, and B. Schuller. Recent developments in opensmile, the munich open-source multimedia feature extractor. In *Proceedings of the ACM International Conference on Multimedia (MM)*, pages 835–838, 2013.

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

[10] X. Huang, A. Dhall, G. Zhao, R. Goecke, and M. Pietikäinen. Riesz-based volume local binary pattern and a novel group expression model for group happiness intensity analysis. In *Proceedings of the British Machine Vision Conference (BMVC)*, pages 34.1–34.12, 2015.

[11] Y. Huang, K. Huang, Y. Yu, and T. Tan. Salient coding for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1753–1760, 2011.

[12] Y. Huang, Z. Wu, L. Wang, and T. Tan. Feature coding in image classification: A comprehensive study. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(3):493–506, 2014.

[13] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2169–2178, 2006.

[14] M. Liu, R. Wang, S. Li, S. Shan, Z. Huang, and X. Chen. Combining multiple kernel methods on Riemannian manifold for emotion recognition in the wild. In *Proceedings of the ACM International Conference on Multimodal Interaction (ICMI)*, pages 494–501, 2014.

[15] S. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982.

[16] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal on Computer Vision*, 60(2):91–110, 2004.

[17] B. Moore. Principal component analysis in linear systems: Controllability, observability, and model reduction. *IEEE Transactions on Automatic Control*, 26(1):17–32, 1981.

[18] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):971–987, 2002.

[19] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *British Machine Vision Conference*, volume 1, page 6, 2015.

[20] R. Rosipal and N. Krämer. Overview and recent advances in partial least squares. In *Proceedings of the International Conference on Subspace, Latent Structure and Feature Selection*, pages 34–51. Springer, 2006.

[21] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. A. Müller, S. S. Narayanan, et al. The interspeech 2010 paralinguistic challenge. In *InterSpeech*, volume 2010, pages 2795–2798, 2010.

[22] K. Sikka, K. Dykstra, S. Sathyanarayana, G. Littlewort, and M. Bartlett. Multiple kernel learning for emotion recognition in the wild. In *Proceedings of the ACM International Conference on Multimodal Interaction (ICMI)*, pages 517–524, 2013.

[23] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. 2015.

[24] C. J. Ter Braak. Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. *Ecology*, 67(5):1167–1179, 1986.

[25] Y.-L. Tian, T. Kanade, and J. F. Cohn. Facial expression analysis. *Handbook of face recognition*, pages 247–275, 2005.

[26] 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 *Proceedings of the ACM International Workshop on Audio/Visual Emotion Challenge (AVEC)*, pages 3–10, 2013.

[27] A. Vedaldi and B. Fulkerson. Vlfeat: An open and portable library of computer vision algorithms. In *Proceedings of the ACM International Conference on Multimedia*, pages 1469–1472, 2010.

[28] H. Wang, F. Nie, and H. Huang. Multi-view clustering and feature learning via structured sparsity. In *International Conference on Machine Learning*, pages 352–360, 2013.

[29] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong. Locality-constrained linear coding for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3360–3367, 2010.

[30] K. Wang, R. He, L. Wang, W. Wang, and T. Tan. Joint feature selection and subspace learning for cross-modal retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP(99):1–1, 2015.

[31] K. Wang, R. He, W. Wang, L. Wang, and T. Tan. Learning coupled feature spaces for cross-modal matching. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2088–2095, 2013.

[32] H. Wold. Partial least squares. *Encyclopedia of statistical sciences*, pages 581–591, 1985.

[33] J. Wu, Z. Lin, and H. Zha. Multiple models fusion for emotion recognition in the wild. In *Proceedings of the ACM International Conference on Multimodal Interaction (ICMI)*, pages 475–481, 2015.

[34] Z. Wu, Y. Huang, L. Wang, and T. Tan. Group encoding of local features in image classification. In *Proceedings of the IEEE International Conference on Pattern Recognition (ICPR)*, pages 1505–1508, 2012.

[35] X. Xiong and F. de la Torre. Supervised descent method and its applications to face alignment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 532–539, 2013.

[36] J. Yang, K. Yu, Y. Gong, and H. Thomas. Linear spatial pyramid matching using sparse coding for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1794–1801, 2009.

[37] A. Yao, J. Shao, N. Ma, and Y. Chen. Capturing au-aware facial features and their latent relations for emotion recognition in the wild. In *Proceedings of the ACM International Conference on Multimodal Interaction (ICMI)*, pages 451–458, 2015.

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

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

[40] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2879–2886, 2012.