Multiple Models Fusion for Emotion Recognition in the Wild

Jianlong Wu[1], Zhouchen Lin[1,2], Hongbin Zha[1]
[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

Emotion recognition in the wild is a very challenging task. In this paper, we propose a multiple models fusion method to automatically recognize the expression in the video clip as part of the third Emotion Recognition in the Wild Challenge (EmotiW 2015). In our method, we first extract dense SIFT, LBP-TOP and audio features from each video clip. For dense SIFT features, we use the bag of features (BoF) model with two different encoding methods (locality-constrained linear coding and group saliency based coding) to further represent it. During the classification process, we use partial least square regression to calculate the regression value of each model. By learning the optimal weight of each model based on the regression value, we fuse these models together. We conduct experiments on the given validation and test datasets, and achieve superior performance. The best recognition accuracy of our fusion method is 52.50% on the test dataset, which is 13.17% higher than the challenge baseline accuracy of 39.33%.

Categories and Subject Descriptors

I.5.4 [**Pattern Recognition**]: Applications; I.4.m [**Image Processing and Computer Vision**]: [Miscellaneous]

Keywords

Emotion Recognition; Multiple Models Fusion; Bag of Features; EmotiW 2015 Challenge

1. INTRODUCTION

Automatic facial expression recognition has become a hot research topic in computer vision because of its significant role in many applications, such as psychological research and human computer interaction (HCI). The primary task of emotion recognition is to classify the given facial images or videos into seven basic expression types, such as angry (AN), disgust (DI), fear (FE), happy (HA), neutral (NE), sad (SA) and surprise (SU). A variety of methods have been proposed towards this problem during the past decade. One may refer to [24] and [33] for a comprehensive survey. However, most existing works focus on expression recognition from static facial images [35, 21, 18]. When it comes to video based emotion recognition or expression recognition in the wild, the recognition results of previous methods are not very satisfactory. Compared with static facial images based emotion recognition, dynamic images based emotion recognition is more complicated.

In recent years, several emotion recognition competitions such as Audio Video Emotion Challenges (AVEC) [25] and Emotion Recognition in the Wild (EmotiW) [4, 6] greatly promoted the development of video based expression recognition. A few methods have been proposed to automatically recognize expression from video clips. For instance, Zhao et al. [34] used LBP-TOP to extract patterns from dynamic facial image sequences. Kahou et al. [11] proposed a deep neural networks based method for emotion recognition. Sikka et al. [22] fused multiple features using multiple kernel learning. Liu et al. [14] represented image set models of video clip with Riemannian kernels. By fusing multiple features through different approach, these methods achieved the state-of-the-art performance.

As these video clips contain rich spatio-temporal information, it is of great importance for video based emotion recognition to combine multiple visual and acoustical features. Feature extraction and classification are two critical processes for recognition. It is vital to select appropriate feature extraction methods and classification combination method. For feature extraction, as LBP-TOP [34] features are robust to variations of grayscale and dense SIFT [16] features are invariant to scale and rotation transformation, we extract both LBP-TOP [34] and dense SIFT [16] features from each video clip. For audio information, we use openSMILE [7] to extract audio features. While the extracted dense SIFT [16] features are redundant and high-dimensional, we adopt bag of features (BoF) [2] model to further represent the SIFT [16] features. According to [32] and [9], both locality and saliency are very essential for image feature encoding. Therefore, we use two corresponding types of encoding methods in BoF to encode features, respectively. During the classification process, compared with feature level and prediction level fusion, score level fusion can better capture the contribution of each model and is more effective. So we first use partial least square regression to calculate the regression value as the score, and then learn the optimal coefficients to fuse these models. Based on above characteristics, in this paper, we construct a multiple

ICMI 2015, November 9–13, 2015, Seattle, WA, USA.
© 2015 ACM. ISBN 978-1-4503-3912-4/15/11 ...$15.00.
DOI: http://dx.doi.org/10.1145/2818346.2830582.

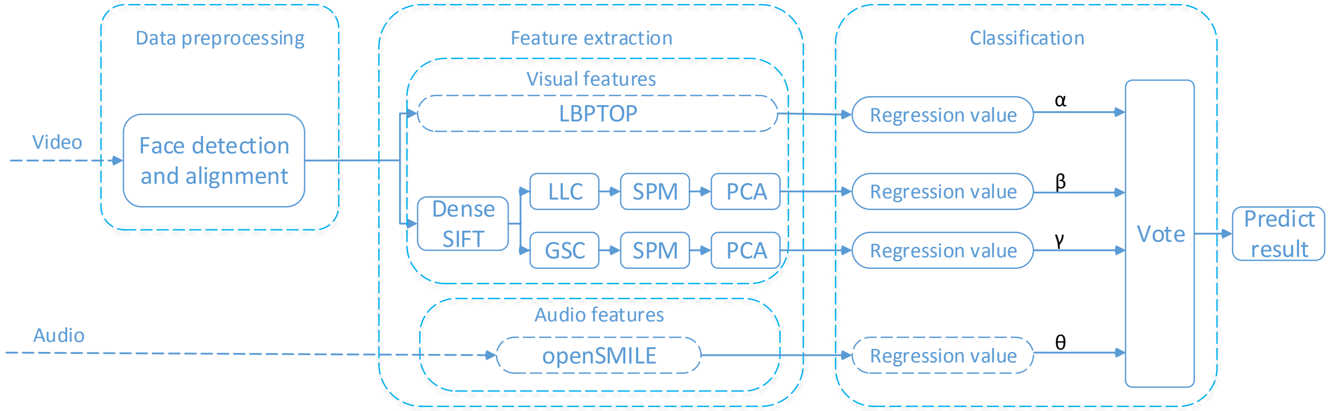


Figure 1: Framework of the proposed multiple models fusion method.

models fusion method for video based emotion recognition in the wild. The proposed method can extract discriminative features for video representation and commendably fuse them for classification.

The rest of this paper is organized as follows. We first introduce the details of BoF model and extracted features in Section 2. Then we present the regression method and our multiple models fusion method in Section 3. In Section 4, we conduct emotion recognition experiments on the given dataset. Finally, Section 5 concludes our paper.

2. VIDEO REPRESENTATION

2.1 Visual Features Extracted with BoF

In this subsection, we first introduce the pipeline of commonly used BoF [2]. Then we present the detailed processes of dense SIFT [16] features extraction and two kinds of encoding methods used in this paper.

2.1.1 The BoF Framework

BoF [2] is one of the most popular and effective image classification frameworks in the recent literature. It is developed from the bag-of-words model in document analysis [10] and has achieved the state-of-the-art performance in many classification tasks, including emotion recognition [23]. As shown in Figure 2, the commonly used BoF framework generally consists of the following three basic modules:

1. **Local features extraction:** In this module, we first divide each input image into many landmark points or dense overlapped patches, and then extract local features such as SIFT [16], HoG [3], and LBP [19] from each block or key point to represent the image. In this paper, we extract SIFT features from dense blocks to represent the facial expression image first.

2. **Descriptors encoding:** Based on the local features of the previous step, we learn a dictionary with the classical K-means [15] clustering algorithm. Each descriptor is encoded into a code vector with codewords in the codebook. By utilizing different encoding methods such as sparse coding [31], saliency coding [8], and LLC [27], we can acquire code vectors with different properties. For a literature survey on encoding methods, one can refer to [12] and [9]. In this paper, we use LLC [27] and GSC [29] to code the descriptors, respectively.

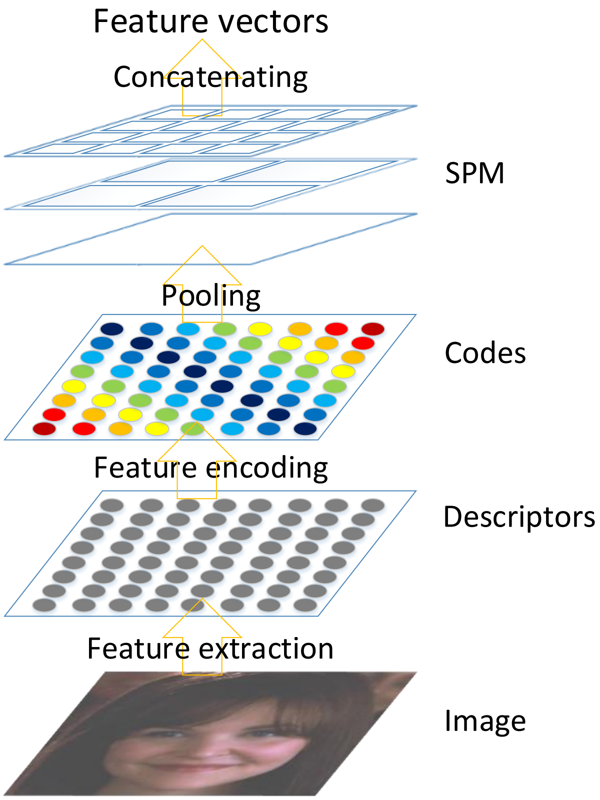


Figure 2: Basic pipeline of the BoF framework.

3. **Spatial pyramid pooling:** In this step, the spatial pyramid matching (SPM) [13] method partitions the image into increasingly finer spatial subregions. Then, pooling process integrates all responses on each codeword in a specific subregion into one value. Max pooling [31] and average pooling [2] are two main pooling methods. We adopt max pooling [31] in this paper. By pooling code vectors in each spatial subregion across different spatial scales, we obtain the local description of every block. The final representation of the image is obtained by concatenating descriptions of all blocks.

2.1.2 Dense SIFT

Scale Invariant Feature Transform (SIFT) [16] is wildly used for feature extraction and image representation. It first detects and selects appropriate keypoints over all scales and image locations, and then computes features in the region around each keypoint. The extracted features are invariant to image scale and rotation. For dense SIFT, we first divide the image into many overlapped grid blocks with a fixed step size, and then compute features on each block. Compared with the original SIFT, dense SIFT does not need to detect keypoints any longer, but the dimension of the extracted features are relatively high. In this case, we adopt BoF [2] for deeper representation.

2.1.3 LLC

Let $X = [x_1, x_2, \cdots, x_N] \in \mathbb{R}^{D \times N}$ be a set of D-dimensional local features extracted from an image. $B = [b_1, b_2, \cdots, b_M] \in \mathbb{R}^{D \times M}$ denotes the codebook with M codewords. Encoded by an encoding algorithm, local features X is converted to N coding vectors $C = [c_1, c_2, \cdots, c_N] \in \mathbb{R}^{M \times N}$.

The core idea of locality-constrained linear coding (LLC) [27] is to reconstruct features with codewords via resolving a least square based optimization problem with locality constraints on the codewords. The objective function of LLC is listed as follows:

$$\begin{aligned} &\arg\min_{C} \sum_{i=1}^{N} (\|x_i - Bc_i\|^2 + \lambda \|d_i \odot c_i\|^2), \\ &s.t. \ \ \mathbf{1}^T c_i = 1, \forall i, \end{aligned} \tag{1}$$

where $\mathbf{1} \in \mathbb{R}^{M \times 1}$ is a column vector of all ones, $\odot$ denotes the element-wise multiplication and $d_i = \exp(\frac{\text{dist}(x_i, B)}{\sigma}) \in \mathbb{R}^M$ is the locality adaptor. Specifically, $\text{dist}(x_i, B) = [\|x_i - b_1\|_2, \cdots, \|x_i - b_M\|_2]^T$. σ is used for adjusting the weight decay speed. The problem defined in Eq. (1) has a closed-form solution:

$$\begin{aligned} &\hat{c_i} = ((B^T - \mathbf{1}x_i^T)(B^T - \mathbf{1}x_i^T)^T + \lambda \text{diag}^2(d_i))^{-1}\mathbf{1}, \\ &c_i = \hat{c_i}/(\mathbf{1}^T \hat{c_i}). \end{aligned} \tag{2}$$

As the solution of LLC only has a few significant values, we can simply use the $K(K < D < M)$ nearest neighbours of x_i in the codebook as the local base $\tilde{B}$ to reconstruct the descriptor x_i, which can speed up the coding process.

2.1.4 GSC

Different from reconstruction based coding method LLC, group saliency coding (GSC) [29] is developed from the saliency based coding [8]. As we use max pooling method to process those encoded features, we believe that saliency is very important for feature coding. We first select K codewords groups for each descriptor x. For each group, GSC [29] calculates the saliency response $\psi^k(x)$, which is then fed back to all the codewords in the group. The final coding result of a feature on each codeword is the maximum response across different group sizes. The computing process of GSC can be formulated as follows:

$$\begin{aligned} c_i &= \max_k \{s_i^k\}, k = 1, 2, \cdots, K, \\ s_i^k &= \begin{cases} \psi^k(x), & \text{if } b_i \in g(x, n_k), \\ 0, & \text{otherwise}, \end{cases} \\ \psi^k(x) &= \sum_{j=1}^{K+1-k} (\|x - \tilde{b}_{k+j}\|_2 - \|x - \tilde{b}_k\|_2), \end{aligned} \tag{3}$$

where s_i^k is the coding result for the kth group and $g(x, n_k)$ denotes the n_k closest codewords set of x for the kth group.

2.1.5 Dimension Reduction

After feature extraction of the BoF [2] 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 dimension of features before classification. We use the classical principle component analysis (PCA) [17] for dimension reduction. The core idea of PCA [17] is to maximize the total variance of projection.

2.2 Visual Features Extracted with LBP-TOP

Local Binary Patterns from Three Orthogonal Planes (LBP-TOP) [34], an extension of the widely used LBP [19] operator, is proposed to handle the influence of varying rotation and lighting condition on dynamic textures. This method considers only 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 [34] are robust to gray-scale and rotations variations. It has been successfully applied to video facial expression recognition while the video can be regarded as a sequence of dynamic facial expression images. We adopt LBP-TOP to extract video features for video based expression recognition.

2.3 Audio Features

Audio information plays an important role in video based emotion recognition [1]. We use the openSMILE toolkit [7], an open-source feature extractor that unites feature extraction algorithms from the speech processing and the Music Information Retrieval communities, to extract audio features. We use 21 energy & spectral related functionals and 19 voicing related functionals to extract corresponding low-level descriptors and delta regression coefficients. By adding 2 voiced/unvoiced durational features, there are 1582 dimensional features in total. For detailed information, please refer to [4].

3. FUSION OF MULTIPLE MODELS

The framework of our proposed method is shown in Figure 1. We first construct four different models to extract features from each facial expression video. For classification, we use the partial least square regression (PLSR) [28] to calculate the regression value of each test sample. Based on the regression value of these four models, we learn the optimal coefficients to fuse them. The final predicted label is the category with the maximum fusion regression value.

Table 1: Performance comparisons of different methods on both validation and test datasets. BoF^{LLC} stands for the LLC based BoF method and BoF^{GSC} stands for the GSC based BoF method.

Methods		Accuracy	
		Val	Test
Baseline (LBP-TOP)		36.08%	39.33%
Audio	Audio	33.96%	–
Video	BoF^{LLC}	47.44%	–
	BoF^{GSC}	45.82%	–
	LBP-TOP+BoF^{LLC}+BoF^{GSC}	48.52%	–
Audio+ Video	Audio+LBP-TOP+BoF^{LLC}+BoF^{GSC}	49.87%	49.35%
	Audio+LBP-TOP+BoF^{LLC}+BoF^{GSC} (Customized)	–	52.50%

3.1 Partial Least Squares Regression

We adopt the same PLSR manner as that in [14]. 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:

$$\begin{aligned} X &= TP^T + E, \\ Y &= UQ^T + F, \end{aligned} \tag{4}$$

where T and U contain the extracted score 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:

$$[cov(t,u)]^2 = \max_{|w_x|=|w_y|=1} [cov(Xw_x, Yw_y)]^2. \tag{5}$$

Then we can get the regression coefficients B as:

$$B = X^T U (T^T X X^T U)^{-1} T^T Y. \tag{6}$$

The regression value can be estimated by:

$$V = XB. \tag{7}$$

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

3.2 Fusion Strategy

For each of these four kinds of feature extraction models (audio, LBP-TOP, LLC based BoF and GSC based BoF), we utilize the PLSR to calculate its corresponding regression value, respectively. Then we adopt the score level fusion method and assign specific weight to each of four previous models:

$$V^{fusion} = \alpha V^{audio} + \beta V^{LBP-TOP} + \gamma V^{LLC} + \theta V^{GSC}, \tag{8}$$

where V represents the regression value. The weight which varies from model to model is relevant to the performance of each model. We learn the optimal weights on the validation dataset.

4. EXPERIMENTS

4.1 Dataset and Parameter Setting

We evaluate the performance of the proposed method on the given AFEW 5.0 dataset [5, 6], which includes 723 train video clips, 383 validation video clips and 539 test video clips. All these video clips are collected from movies that show close-to-real-world conditions. Figure 3 shows example images of seven expressions taken from video clips.

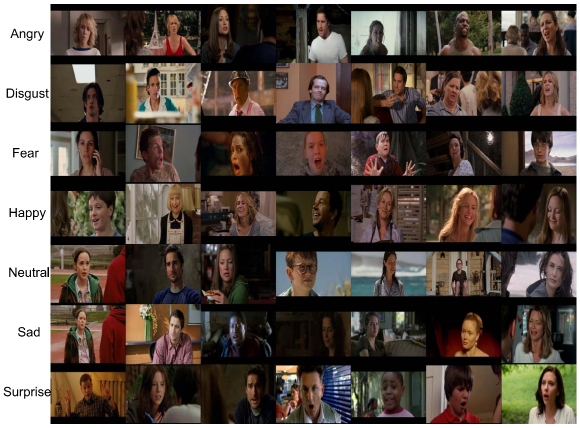
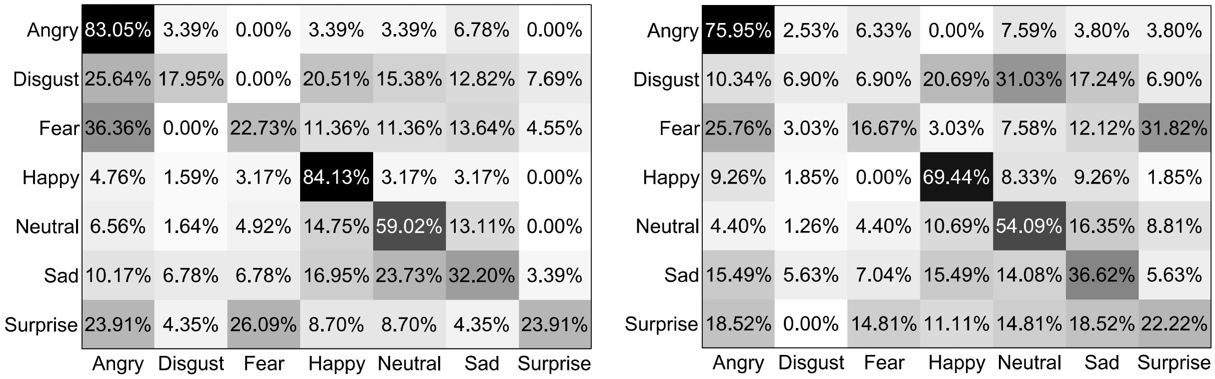


Figure 3: Some example frames of expression videos in the wild.

For each video clip, organizers apply pre-trained face models [36] for face detection and initialization. Then, the intraface tracking library [30] is adopted to align the detected facial images. Each facial image is aligned to size 128 × 128. LBP-TOP [34] features are extracted from non-overlapping spatial 4×4 blocks [6]. We directly use the aligned facial images as well as the extracted audio and LBP-TOP features provided by organizers. During the process of BoF based feature extraction, we set the parameters as follows. We divide each facial image into overlapped blocks with step 1 and size 16×16. For each block, we use Vlfeat [26] to extract 128-dimensional SIFT features. The dictionary is learned by the K-means [15] clustering algorithm with 1024 centres. Both the nearest neighbours number for LLC [27] and groups number for GSC [29] are set to 5. During the pooling process, we employ the SPM with levels of $[1\times1, 2\times2, 4\times4, 8\times8]$. Under the above settings, dimension of the final BoF representation for each frame is $1024 \times 85 = 87040$. We adopt max pooling to process all the frames of each video to get the video representation. We further use PCA [17] to reduce dimension with principle components ratio 97%. For the fusion process, we set the optimal weights as $\alpha = 0.25$, $\beta = 0.15$, $\gamma = 1.00$ and $\theta = 0.50$, which is learned on the validation dataset.



	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise
Angry	83.05%	3.39%	0.00%	3.39%	3.39%	6.78%	0.00%
Disgust	25.64%	17.95%	0.00%	20.51%	15.38%	12.82%	7.69%
Fear	36.36%	0.00%	22.73%	11.36%	11.36%	13.64%	4.55%
Happy	4.76%	1.59%	3.17%	84.13%	3.17%	3.17%	0.00%
Neutral	6.56%	1.64%	4.92%	14.75%	59.02%	13.11%	0.00%
Sad	10.17%	6.78%	6.78%	16.95%	23.73%	32.20%	3.39%
Surprise	23.91%	4.35%	26.09%	8.70%	8.70%	4.35%	23.91%

(a) Confusion matrix on the validation dataset

	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise
Angry	75.95%	2.53%	6.33%	0.00%	7.59%	3.80%	3.80%
Disgust	10.34%	6.90%	6.90%	20.69%	31.03%	17.24%	6.90%
Fear	25.76%	3.03%	16.67%	3.03%	7.58%	12.12%	31.82%
Happy	9.26%	1.85%	0.00%	69.44%	8.33%	9.26%	1.85%
Neutral	4.40%	1.26%	4.40%	10.69%	54.09%	16.35%	8.81%
Sad	15.49%	5.63%	7.04%	15.49%	14.08%	36.62%	5.63%
Surprise	18.52%	0.00%	14.81%	11.11%	14.81%	18.52%	22.22%

(b) Confusion matrix on the test dataset

Figure 4: Confusion matrices of multiple models fusion method for facial expressions on both validation and test datasets.

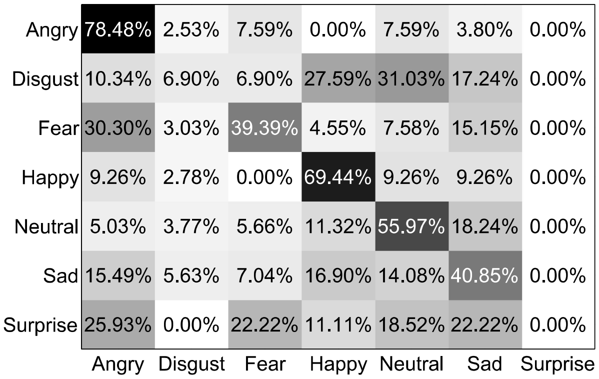
4.2 Results

In Table 1, we present the recognition results of our proposed method. Our proposed method achieves competitive results. The baseline recognition rates given by the EmotiW 2015 organizers on validation dataset and test dataset are 36.08% and 39.33%, respectively. We first test the performance of all the single model on the validation dataset. Among these single models, LLC based BoF achieves the best result, with accuracy 47.44%, while the recognition rate of GSC based BoF is 45.82% on the validation dataset. Both BoF models get satisfactory results. For video only emotion recognition, we fuse the regression values of these three video image set based methods (LBP-TOP, LLC based BoF and GSC based BoF), and achieve 48.52% accuracy. Finally, we fuse all the single models with optimal weights. The recognition rate is further improved to 49.87%, which is 13.79% higher than that of the baseline. On the test dataset, our proposed multiple models fusion method achieves 49.35%, which largely surpasses the baseline. For the customized method shown in Table 1, we will explain it in Section 4.3.

Figure 4 shows the confusion matrices of our multiple models fusion method on both validation and test datasets. From these two matrices, 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.

4.3 Discussion

According to the confusion matrix (computed by the challenge organizers) on the test dataset shown in Figure 4(b), it is difficult to recognize surprise expression, and fear expression samples are easily misclassified to surprise. Few train samples of surprise and high correlation between surprise and fear may account for this phenomenon. We need to note that total sample numbers of fear and surprise expressions on the test dataset are 66 and 26, respectively. By analyzing the statistics in Figure 4(b), we further customize our method slightly. For predicted surprise expression, we use the category with the second largest fusion regression value instead of the largest value as the predicted label. The corresponding recognition result is shown in Figure 5, and the overall recognition accuracy become 52.5%. Comparing the result in Figure 5 with that in Figure 4(b), we can easily see that the improvement of customized method mainly lies in the fact that more than 22% of fear samples, which are misclassified before, are classified correctly this time.



	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise
Angry	78.48%	2.53%	7.59%	0.00%	7.59%	3.80%	0.00%
Disgust	10.34%	6.90%	6.90%	27.59%	31.03%	17.24%	0.00%
Fear	30.30%	3.03%	39.39%	4.55%	7.58%	15.15%	0.00%
Happy	9.26%	2.78%	0.00%	69.44%	9.26%	9.26%	0.00%
Neutral	5.03%	3.77%	5.66%	11.32%	55.97%	18.24%	0.00%
Sad	15.49%	5.63%	7.04%	16.90%	14.08%	40.85%	0.00%
Surprise	25.93%	0.00%	22.22%	11.11%	18.52%	22.22%	0.00%

Figure 5: Confusion matrix of customized method on the test dataset.

5. CONCLUSIONS

In this paper, we propose a multiple models fusion method for video based emotion recognition in the wild. We first extract audio features from the video clips. Then dense SIFT and LBP-TOP visual features are extracted from aligned facial image set of each clip. For dense SIFT features, we further use LLC based BoF and GSC based BoF models to represent them. In the classification process, we first use partial least square regression to calculate the regression value of each single model, and then fuse these models together with the optimal coefficients based on the regression values. We validate the performance of our proposed method on the AFEW 5.0 dataset as part of the third Emotion Recognition in the Wild Challenge (EmotiW 2015). Our method achieves excellent performances on both validation and test datasets. As feature extraction and classification are two key processes for video based emotion recognition, in the future, on the one hand, we will further investigate and mine the connection between frames of video clip. On the other hand, we will try to find effective multi-model fusion method for classification.

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. H. Zha is supported by 973 Program (grant no. 2011CB302202).

7. REFERENCES

[1] 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.

[2] 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.

[3] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 886–893, 2005.

[4] 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.

[5] 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.

[6] A. Dhall, O. R. Murthy, R. Goecke, J. Joshi, and T. Gedeon. Video and image based emotion recognition challenges in the wild: Emotiw 2015. In *Proceedings of the ACM International Conference on Multimodal Interaction (ICMI)*, 2015.

[7] 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.

[8] 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.

[9] 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.

[10] T. Joachims. Text categorization with suport vector machines: Learning with many relevant features. In *Proceedings of the European Conference on Machine Learning (ECML)*. Springer, 1998.

[11] S. E. Kahou, C. Pal, X. Bouthillier, P. Froumenty, c. Gülçehre, R. Memisevic, P. Vincent, A. Courville, and Y. Bengio. Combining modality specific deep neural networks for emotion recognition in video. In *Proceedings of the ACM International Conference on Multimodal Interaction (ICMI)*, pages 543–550, 2013.

[12] A. V. Ken Chatfield, Victor Lempitsky and A. Zisserman. The devil is in the details: an evaluation of recent feature encoding methods. In *Proceedings of the British Machine Vision Conference (BMVC)*, pages 76.1–76.12, 2011.

[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] S. Moore and R. Bowden. Local binary patterns for multi-view facial expression recognition. *Computer Vision and Image Understanding*, 115(4):541–558, 2011.

[19] 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.

[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] O. Rudovic, M. Pantic, and I. Patras. Coupled Gaussian processes for pose-invariant facial expression recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(6):1357–1369, 2013.

[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. Sikka, T. Wu, J. Susskind, and M. Bartlett. Exploring bag of words architectures in the facial expression domain. In *Proceedings of European Conference Computer Vision Workshop and Demonstrations*, pages 250–259, 2012.

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

[25] 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.

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

[27] 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.

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

[29] 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.

[30] 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.

[31] 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.

[32] K. Yu, T. Zhang, and Y. Gong. Nonlinear learning using local coordinate coding. In *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*, pages 2223–2231, 2009.

[33] 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.

[34] 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.

[35] W. Zheng. Multi-view facial expression recognition based on group sparse reduced-rank regression. *IEEE Transactions on Affective Computing*, 5(1):71–85, 2014.

[36] 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.