Capturing AU-Aware Facial Features and Their Latent Relations for Emotion Recognition in the Wild

Anbang Yao
Intel Labs China,
Beijing, 100190, China
anbang.yao@intel.com

Junchao Shao
Intel Labs China,
Beijing, 100190, China
junchao.shao@intel.com

Ningning Ma
Department of Computer Science and Technology, Tsinghua University,
Beijing, 100084, China
mnn15@mails.tsinghua.edu.cn

Yurong Chen
Intel Labs China
Beijing, 100190, China
yurong.chen@intel.com

ABSTRACT
The Emotion Recognition in the Wild (EmotiW) Challenge has been held for three years. Previous winner teams primarily focus on designing specific deep neural networks or fusing diverse hand-crafted and deep convolutional features. They all neglect to explore the significance of the latent relations among changing features resulted from facial muscle motions. In this paper, we study this recognition challenge from the perspective of analyzing the relations among expression-specific facial features in an explicit manner. Our method has three key components. First, we propose a pair-wise learning strategy to automatically seek a set of facial image patches which are important for discriminating two particular emotion categories. We found these learnt local patches are in part consistent with the locations of expression-specific Action Units (AUs), thus the features extracted from such kind of facial patches are named AU-aware facial features. Second, in each pair-wise task, we use an undirected graph structure, which takes learnt facial patches as individual vertices, to encode feature relations between any two learnt facial patches. Finally, a robust emotion representation is constructed by concatenating all task-specific graph-structured facial feature relations sequentially. Extensive experiments on the EmotiW 2015 Challenge testify the efficacy of the proposed approach. Without using additional data, our final submissions achieved competitive results on both sub-challenges including the image based static facial expression recognition (we got 55.38% recognition accuracy outperforming the baseline 39.13% with a margin of 16.25%) and the audio-video based emotion recognition (we got 53.80% recognition accuracy outperforming the baseline 39.33% and the 2014 winner team's final result 50.37% with the margins of 14.47% and 3.43%, respectively).

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

ICMI '15, November 09-13, 2015, Seattle, WA, USA
 ISBN 978-1-4503-3912-4/15/11…$15.00
DOI: http://dx.doi.org/10.1145/2818346.2830585

General Terms
Algorithms, Experimentation

Keywords
Emotion Recognition; Facial Expression Recognition; Action Unit; Facial Feature Relation; EmotiW 2015 Challenge

1. INTRODUCTION
Understanding of human emotions has been of keen interest in the community of psychology, neuroscience, computer vision, pattern recognition, machine learning, and so forth [11]. This is partially due to its significant role in our daily communication, its potential applications as well as its inherent challenges. Having a large volume of well labeled data is the basic prerequisite for advancing the research on automatic emotion recognition. To this end, numerous databases have been collected for comparatively evaluating algorithms [30]. However, most of the widely used databases such as the Cohn-Kanade facial expression database [16, 22], the Man-Machine Interaction facial expression database [23, 25] and the Banse-Scherer vocal affect database [1] are collected in the constrained environments. When collecting such kind of posed data, subjects were asked to perform a series of deliberately expressed emotions in front of the mounted cameras and/or microphones. A main limitation of this state is that the methods having good performance on these posed databases usually fail to generalize well to spontaneously expressed emotions. Starting from 2011, several spontaneous or wild emotion databases [10, 24, 9, 26] have being collected, annotated, and gradually enlarged in size for using in respective competitions held yearly. In this paper, we present our method concerning the submissions for both sub-challenges of the EmotiW 2015 Challenge.[1]

Different from existing dominant databases merely consisting of posed emotion data, all annotated static images and short lived video clips in the EmotiW 2015 Challenge are directly collected from full length movies. A direct consequence is that the data are filled with a wide variety of changes in illumination, subject pose, context scene, subtle affective display and so on, which makes it difficult to achieve high recognition accuracy in both competition tasks. In the previous two years, the winner teams [15, 19] mainly resort to use deep neural networks [17, 13] or design sophisticated

[1] This work was done when Junchao Shao and Ningning Ma were interns at Intel Labs China.

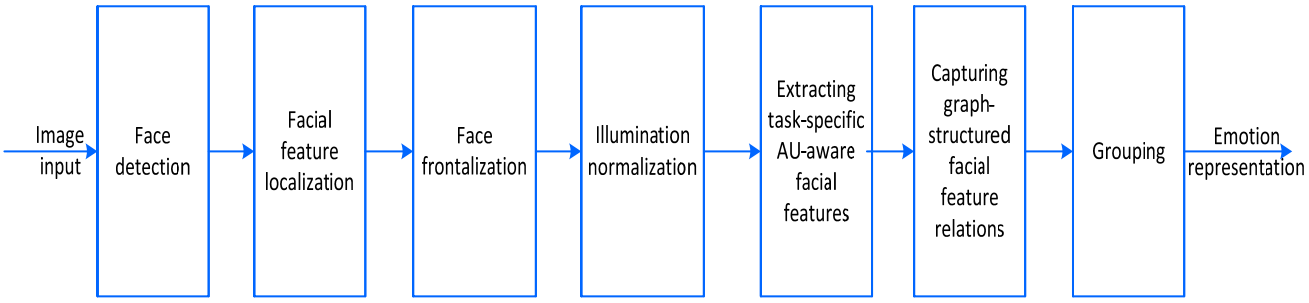


Figure 1. The flowchart of building emotion representation with AU-aware facial feature relations.

method for fusing diverse hand-crafted and deep convolutional features [7, 21, 17]. We also noticed that the accuracy gain from combining audio input into visual input is only ~3% and the performance of audio input alone is relatively poor which implies that visual input plays the key role in the challenge. This motivates us to make a deep study of visual facial expression changes. Existing psychological studies [2, 11] have revealed that facial expressions are the outcomes of human muscle motions, and even a minor change in facial expression is the interaction of the motions of a set of particular muscles. Therefore, we aim to study the problem of classifying emotions spontaneously appeared in the static images and in the dynamic video clips from the perspective of seeking relevant features characterizing expression-specific facial muscle motions and analyzing their latent relations. The contributions of this paper are in four folds:

- A pair-wise learning strategy is introduced to automatically seek a set of discriminative facial patches for classifying two particular emotion categories.
- An undirected graph structure is used to encode inherent task-specific feature relations among learnt facial patches of a respective set.
- A new emotion representation is presented by sequentially concatenating all these task-specific graph-structured facial feature relations.
- Unlike previous teams which usually used outside data to boost the performance of their final models, our models were trained only using given data. Impressively, on both image based static facial expression recognition and audio-video based emotion recognition tasks included in the EmotiW 2015 Challenge, we achieved leading recognition accuracy with large gains compared to the baselines and the 2014 winner team's final result.

We believe that our method provides new insight for handling emotion recognition in the wild. In the following sections, we will describe our detailed method and related experimental results.

2. THE PROPOSED METHOD

Given an image or video frame, Figure 1 shows the flowchart of building respective emotion representation with our proposed AU-aware facial feature relations. Here, we begin with data pre-processing steps (Figure 2 shows illustrative results).

2.1 Data Pre-processing

2.1.1 Face Detection

In our system, face detection is the initial step. We used a self-developed Adaboost based multi-view face detector [27] to locate target faces appeared in the static images or in the first frames of video clips. Although most target faces were successfully detected by our detector, few face instances were missed. For these failed images, we gradually decreased the number of the stages of our face detector until the target faces were finally detected.

2.1.2 Facial Feature Localization

When the faces have been detected and converted to gray-scaled images, the next step is facial feature (i.e., a set of semantic facial landmark points) detection (for static images) or tracking (for video clips). Among many existing methods [6, 3], we used recently proposed Supervised Descent Method (SDM) [29] to detect or track facial features. In tracking scenarios, sometimes it may fail temporally or completely. For temporally failed frames, we directly discarded them. For completely failed video clips, we replaced facial feature tracking by a separate running of face detection and facial feature detection. Based on the detected or tracked facial features, we further cropped out related face regions and rescaled them to a standard size of 156×156 pixels.

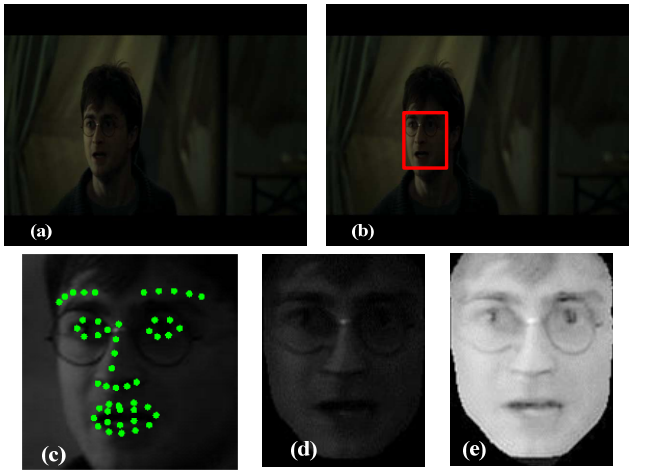


Figure 2. Data pre-processing illustration. (a) Source image; (b) face detection (shown with a red-colored rectangle); (c) facial feature localization (shown with 49 green-colored points) and face cropping; (d) face frontalization; (e) illumination normalization.

2.1.3 Face Frontalization

Once the facial features have been determined, the subsequent step is face registration. Its main goal is to remove the influence of 2D/3D head pose variation, so face images can be geometrically normalized [30]. In our system, we used a straightforward but effective face frontalization method [12] to

perform registration. This method first estimates a coarse 3D face surface by directly projecting detected or tracked 2D facial features to a unified 3D facial feature model, and then builds a frontalized face image by back-projecting source image to the reference coordinate system. The frontalized face images were further rescaled into a standard size of 118×156 pixels.

2.1.4 Illumination Normalization

Illumination changes frequently appeared in the EmotiW 2015 datasets. Therefore, the last data pre-processing step in our system is illumination normalization. We implemented a popular Discrete Cosine Transform (DCT) based method [5] to compensate for illumination variations in the logarithm domain. Figure 2 shows typical example results obtained at our pre-processing steps.

2.2 Emotion Representation

Our method is inspired from two factors. First, from previous works [15, 8, 19], we noticed that the performance of the audio modality alone is relatively poor and the gain from combining audio input into visual input is only ~3%. This motivates us to have a deep study of visual facial expression changes. Second, a well-accepted psychological fact is that facial expressions are the outcomes of the motions of different facial muscles. Even a pretty simple facial expression change is observed as a complex blend of the motions of at least several facial muscles. Consequently, we believe that how to effectively detect task-specific active facial features (characterizing facial muscle motions of two particular expression categories) and how to further describe latent relations (interpreting the interactions of facial muscle motions) among active facial features are the most critical two issues for the recognition of spontaneously displayed expressions. In the current research, the detection of Action Units (AUs) [2, 30] is the main stream to describe related facial muscle motions. However, AU detection still remains a very challenging problem. Recently, some efforts have been reported to use AU-aware facial features from multi-task learning [32, 20] and deep Convolutional Neural Networks (CNNs) [18] to address expression recognition task. In [28], the authors used an improved Bayesian Network to model temporal relations among geometric facial features. Our method aims to present a unified framework to jointly learn task-specific AU-aware facial features and encode their latent relations for robust expression recognition.

2.2.1 Learning AU-Aware Facial Patches

Different from [32, 20] in which AU-aware facial features were collectively learnt by multi-task learning, we present a pair-wise learning strategy to discover relevant AU-aware facial features for classifying two particular emotion categories.

Let S be a training dataset consisting of N normalized images (with a size of $w \times h$ pixels) collected from M facial expression categories. Each expression category has n_m training images, and each face image is segmented into $w_p \times h_p$ local patches with an overlapping stride s. We aim to seek an optimal combination of P facial patches for classifying two particular facial expression categories. We call this class-to-class expression classification as pair-wise learning task. With M expression categories, there are totally $M \times (M-1)/2$ above defined tasks. For the task T, let x_k be the feature set to represent the appearances of Q candidate

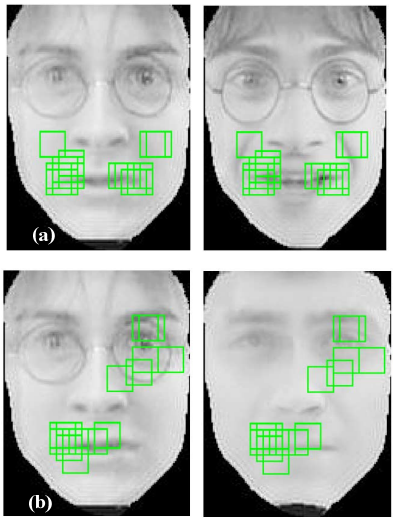


Figure 3. Illustrations of AU-aware facial patches (shown with 16 green-colored rectangles). (a) Samples of Angry-to-Happiness; (b) Samples of Angry-to-Disgust.

local patches of the k^{th} face image from two target expression categories, $y_k \in \{-1,1\}$ be the facial expression label, the objective function in our approach to be minimized can be formulated as

$$\begin{cases} \arg\min_{w,p} \left(\|w\| + \gamma \sum_{k=1}^{N_T} L\left(w, DoT\left(p, x_k\right), y_k\right) \right) \\ Subject \quad to \quad \sum_{i=1}^{Q} p_i = P, and \quad p_i \in \{0,1\} \end{cases}, \tag{1}$$

where w enforces the weights for feature selection, γ is a positive penalty parameter for regularizing classification error, N_t is the number of the images annotated as two target facial expression categories, and L is defined as a logistic loss function

$$L\left(w, DoT\left(p, x_k\right), y_k\right) = \log\left(1 + \exp\left(-y_k w^T DoT\left(p, x_k\right)\right)\right), \tag{2}$$

where operation $DoT\left(p, x_k\right)$ indicates P facial patches are finally selected. Note that this optimization problem can be analytically solved by a greedy searching. For simplicity and efficiency, we present a two-stage method to approximate its analytical solution. In each pair-wise learning task, we first compute the recognition rates for all Q individual candidate image patches, and directly choose αP image patches with the highest recognition rates as the refined candidates. Here, $P \le \alpha P << Q$ thus the searching space is largely narrowed down. Then, an alternative solution can be easily obtained by an iterative searching. In the experiments, we found the combination of local image patches obtained from above two-stage method can well approximate real solution in classification accuracy. In this paper, we just used the histograms of basic Local Binary Pattern (LBP) [31] as the testing features to describe the appearance of each facial patch. Figure 3 illustrates discovered AU-aware facial

patches regarding two different pair-wise tasks. It can be seen that the local patches discovered by our method are

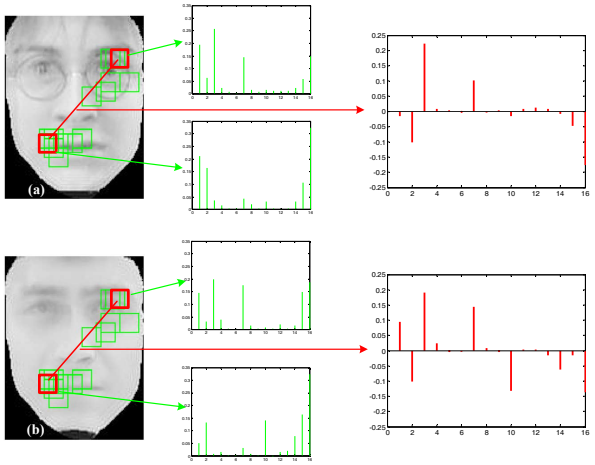


Figure 4. Illustration on computing AU-aware facial feature relations between two local patches determined in Angry-to-Disgust task. In (a) and (b), the left column shows two sample images, the middle column shows LBP histograms extracted from red-colored local patches, and the right column shows respective AU-aware facial feature relations.

mainly located around primary facial components. They are in part consistent with the locations of related expression-specific AUs [2, 30]. Besides, they are visually discriminative.

2.2.2 Encoding Facial Feature Relations with Graph Structure

Once respective P facial patches have been discovered for each pair-wise learning task, we then consider the problem of analyzing the latent relations among the features extracted from these facial patches. To this problem, two main issues should be addressed. Regarding each pair-wise task, the first issue is how to measure the feature relation between any two facial patches. Assuming a metric is available, it may be used to estimate the relation between any two facial patches, which may introduces dimensional curse and redundancy. Therefore, the second main issue is how to select and encode feature relations having better discriminations.

Given a face image I, let x_i and x_j be the respective histograms of LBP features extracted from two facial patches determined in the task T, the related feature relation ρ is directly computed as

$$\rho = x_i - x_j . \quad (3)$$

Here, $x_i, x_j \in \Re^{16}$, and they are the first order statistics of point-wise LBP features. This formed relation metric measures the difference between LBP histograms x_i and x_j. In other words, it is the second order statistics of point-wise LBP features. We found such kind of feature relations is really useful to describe the interactions of facial muscle motions. Figure 4 illustrates the process of computing specific AU-aware facial feature relations.

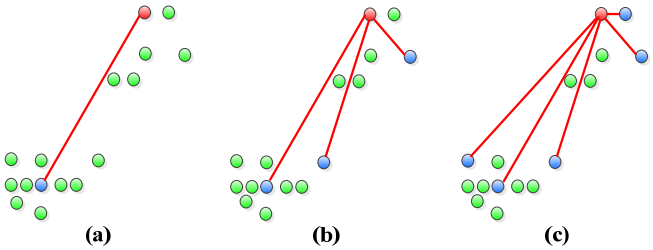


Figure 5. Illustrative graph structures for encoding AU-aware facial feature relations regarding the task of Angry-to-Disgust. A reference vertex (red-colored circle) with (a) one single, (b) three and (c) five connections (red-colored lines). Here, we just draw the connections to one single reference vertex for a better visual effect. The vertices which have connections to the reference vertex are shown as blue-colored circles.

Regarding each pair-wise task, we define an undirected graph structure $G = \{V, E\}$ to encode feature relations among respective facial patches. When constructing a task-specific graph G, the locations of respective P local patches are naturally used as the individual vertices, and the edge connecting any two vertices encodes related feature relation between the respective two local patches of a face image. According to above definition, such kind of task-specific graphs can be either fully connected or highly sparse. Theoretically, there are at most $P-1$ edges to an arbitrary reference vertex. Among $P-1$ candidate edges, $\hat{P}$ edges are finally selected from maximizing mutual information. Figure 5 provides three illustrative examples with $\hat{P} = \{1,3,5\}$.

2.2.3 Constructing Emotion Representation

When AU-aware facial patches and a related sparse graph structure are determined in each pair-wise learning task, two novel emotion representations can be built in three-stages. Given a face image, we first extract AU-aware facial features from local patches automatically discovered by each pair-wise learning task. Then we compute related graph-structured facial feature relations regarding each pair-wise learning task. Based on the resulted AU-aware facial features and their relations, two different emotion representations are finally constructed in a similar manner. The more discriminative emotion representation is constructed by concatenating all task-specific graph-structured facial feature relations sequentially, and the other complementary emotion representation is built by directly concatenating all task-specific facial features. Our results on the EmotiW 2015 Challenge fully testify the power of these two emotion representation methods.

3. PERFORMANCE EVALUATION

3.1 The EmotiW 2015 Challenge

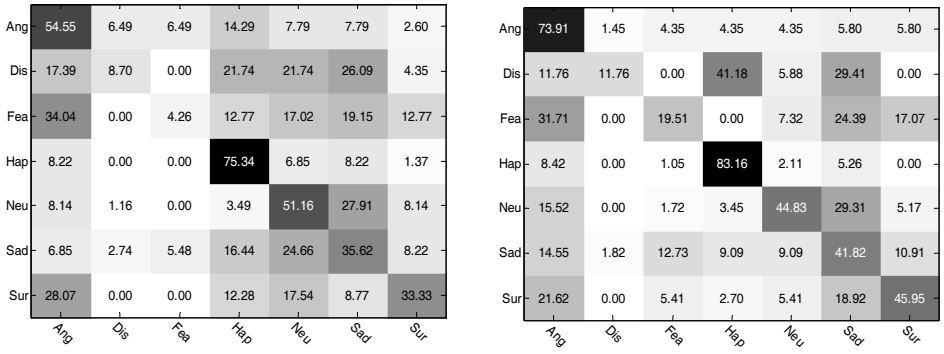
The third Emotion Recognition in the Wild Challenge (EmotiW 2015) has two sub-challenges: (1) audio-video based emotion recognition sub-challenge based on the AFEW 5.0 database [10]; (2) image based static facial expression recognition sub-challenge based on the SFEW 2.0 database [10]. The first sub-challenge is a continuation from the challenge of EmotiW 2013&2014 in which the participators are asked to assign one single emotion label from seven possible emotion categories (including Angry, Disgust, Fear, Happiness, Sadness, Surprise and Neutral) to each testing video clip.

Submission No.	Validation (%)	Testing (%)	Method
1	42.43	52.96	Average of 2 linear SVMs trained with AU-aware facial features/relations (face scale: 118×156 pixels)
5	41.97	54.30	Average of 4 linear SVMs trained with AU-aware facial features/relations (two face scales)
6	41.28	53.23	Maximum of 4 linear SVMs trained with AU-aware facial features/relations (two face scales)
9	43.58	**55.38**	Average of 2 SVMs with the RBF kernel trained with AU-aware facial features/relations (face scale: 118×156 pixels)
10	44.04	52.96	Average of 2 linear SVMs & 2 SVMs with the RBF kernel trained with AU-aware facial features/relations (face scale: 118×156 pixels)

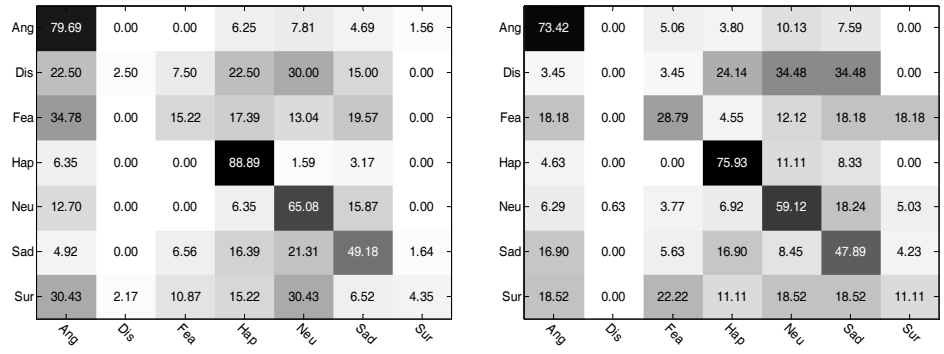
Table 1. Our best 5 submissions on the SFEW 2.0 validation and testing sets.

Submission No.	Validation (%)	Testing (%)	Method
1	44.39	44.53	Average of 5 CNN models (two face scales) + 1 Audio model
7	45.39	49.00	Average of 4 linear SVMs trained with AU-aware facial features/relations (two face scales)
8	46.21	51.58	Average of 3 linear SVMs trained with AU-aware facial features/relations (two face scales) + 1 Audio model
9	49.09	53.25	Average of 3 linear SVMs trained with AU-aware facial feature relations (two face scales) + 1 Audio model + 1 CNN model
10	49.09	**53.80**	Average of another 3 linear SVMs trained with AU-aware facial feature relations (two face scales) + 1 Audio model + 1 CNN model

Table 2. Our best 5 submissions on the AFEW 5.0 validation and testing sets.



(a) SFEW 2.0 validation set **(b) SFEW 2.0 testing set**



(c) AFEW 5.0 validation set **(d) AFEW 5.0 testing set**

Figure 6. Confusion matrices of our final submissions. For each validation set, the models were trained with given training data only. For each testing set, the models were trained on the union of given training and validation sets.

In this sub-challenge, there is no constraint on data modality (e.g., video-only, audio-only or audio-video). Compared with the EmotiW 2014 Challenge, the EmotiW 2015 Challenge has more video clips. Specifically, the AFEW 5.0 databased has 1645 video clips which are divided into three parts: 723 for training, 383 for validation and 539 for testing. The second sub-challenge is a new task in which only static images are available and the participators are asked to label each testing image with one single facial expression label from seven possible expression categories (which are same to those in the first sub-challenge). In the SFEW 2.0 database, there are totally 1766 static images which are also divided into three parts: 958 for training, 436 for validation and 372 for testing. Recall that our approach aims to make a deep understanding of visual emotion changes, thus in what follows we begin with our results on the image based static facial expression sub-challenge.

3.2 Facial Expression Recognition on Wild Static Images

3.2.1 Experimental Settings

All face images in the SFEW 2.0 database were normalized into two standard scales: 118×156 pixels and 128×128 pixels. The normalized face images were segmented into $16\times16-$ *pixels* local patches with an overlapping stride of 8 pixels regardless of their scales. At each face scale, 16 facial patches regarding each related pair-wise learning task were finally determined with our proposed two-stage method. When modeling task-specific graph-structured AU-aware feature relations, the number of the edges connected to each vertex (i.e., facial patch) of a respective graph is 3. For each normalized face image, an AU-aware facial feature relation based representation was computed. To enhance its discrimination, for each face image, we also extracted another representation consisting of sequentially concatenated AU-aware facial features. At the face scale of 118×156 pixels, a linear Support Vector Machine (SVM) and a SVM with Radius Basis Function (RBF) kernel [4] were trained using above two emotion representations independently. At the face scale of 128×128 pixels, 2 linear SVMs were trained. Finally, we trained 6 SVMs using the training set and 6 SVMs using the union of the training and validation sets (with no outside data).

3.2.2 Experimental Results

We take the average or the maximum of the expression scores of any particular combination of trained SVM classifiers to make a final predication on each static face image. Table 1 summarizes the results of our best 5 submissions. It can be seen that a simple combination of our AU-aware facial features and their latent relations demonstrates promising performance on this task, even using linear SVMs as the classifiers. On the validation set, our best recognition accuracy is 44.04% with an improvement of 8.08% compared with the baseline 35.96%. On the testing set, our final recognition accuracy is 55.38% outperforming the baseline 39.13% with a margin of 16.25%. Comparatively, it can be found that a more large accuracy improvement is obtained on the testing set. This is mainly due to the augmentation of the training data by introducing the validation set. Generally, these results fully testify the significance and the effectiveness of our proposed AU-aware facial features and their latent relations. The confusion matrices of our final submissions on the SFEW 2.0 validation and testing sets are shown in Figure 6 (a) and (b), respectively.

3.3 Emotion Recognition on Wild Video Clips

3.3.1 Experimental Settings

On the AFEW 5.0 database, we trained 4 linear SVMs on the training set and 4 linear SVMs using the union of the training and validation sets (with no outside data) under the same parameter settings to those used on the SFEW 2.0 database. To train these SVMs, we only considered the apex frames of the video clips in the training data. Here, we also analyzed the capability of fusing our AU-aware facial features and their latent relations with other popular used features. From the previous works [15, 19], it can be learnt that audio modality can bring ~3% gain in recognition accuracy. So we also trained a SVM with the RBF kernel using audio features extracted with the OpenSmile toolkit. Before extracting audio features, we totally discarded ~20% audio signal at the head and the tail of each clip. On the other side, deep CNNs have been used as the key features in the previous winner teams' solutions [15, 19]. To have a comprehensive comparison, we also trained 5 5-layer CNN models using the Caffe toolkit [14]. These CNN models were trained using the face images with the scale of 118×156 / 128×128 pixels. To these CNN models, the kernels { 7×7 , 5×5 , 3×3 } used in the 3 convolutional layers are completely same, but the number of output feature channels is {{32,32,32},{24,24,24},{20,20,20},{16,16,16},{20,16,16}}. The number of the channels from fully connected layer is 64 or 128. The last layer is a 7-category Softmax classifier. Unlike [15] and [19] in which the authors used outside data to pre-train their CNN models, we just used given data to train our CNN models.

3.3.2 Experimental Results

On the AFEW 5.0 database, we take the average of the emotion scores of any particular combination of trained models to make a final predication on each video clip. On each testing video clip, the emotion scores of a visual feature based model are obtained in two steps. First, it sequentially operates on every frame of each testing video clip. And then the summation of the predicted emotion scores over all frames is used the final emotion scores of this model. The results of our best 5 submissions are summarized in Table 2. Our first six submissions mainly use the combinations of 5 CNN models and one SVM with the RBF kernel trained with audio features. On the validation and the testing sets, the best recognition rates of the CNN dominant combinations are 44.39% and 44.53%, respectively. These results are consistent with those reported in [15] and [19]. Comparatively, a simple combination of our AU-aware facial features and their latent relations shows better performance (45.39% and 49%) just using linear SVMs as the classifiers. By combining them with audio features, we got 51.58% recognition accuracy on the testing set. The performance is further improved by introducing one single CNN model. On the testing set, our final recognition accuracy is 53.8% outperforming the baseline 39.33% and the 2014 winner team's final accuracy 50.37% with the margins of 14.47% and 3.43%, respectively. Comparatively, it can be also noticed that more large accuracy improvements are obtained on the testing set. This is mainly due to the augmentation of the training data by introducing the validation set. The confusion matrices of our final submissions on the AFEW 5.0 validation and testing sets are shown in Figure 6 (c) and (d), respectively.

4. CONCLUSIONS

The EmotiW 2015 Challenge provides a comprehensive dataset collected from full length movies for comparatively evaluating algorithms of recognizing emotions in the wild. In this paper, a new method is proposed to address both tasks included in this challenge. Different from previous works, we aim to analyze spontaneously expressed emotions from the perspective of making a deep exploration of expression-specific AU-aware facial features and their latent relations. In our approach, the AU-aware features characterizing contrasting facial muscle motions regarding two particular emotion categories are first extracted from a set of automatically discovered image patches through a pair-wise learning strategy. To each pair-wise task, we further present an undirected graph structure to capture latent facial feature relations of any pair of image patches. The effectiveness of our emotion representations building from sequentially concatenating all task-specific graph-structured facial features or their latent relations are fully tested in the experiments. We achieved impressively good recognition accuracy in both tasks of the EmotiW 2015 Challenge. The experiments also verify the significance of AU-aware facial features and their latent relations.

5. REFERENCES

[1] R. Banse and K.R. Scherer. Acoustic profiles in vocal emotion expression. *Journal of personality and social psychology*, 70(3):614-636, 1996.

[2] M.S. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan. Recognizing facial expression: machine learning and application to spontaneous behavior. In *CVPR*. IEEE, 2005.

[3] X. Cao, Y. Wei, F. Wen, and J. Sun. Face alignment by explicit shape regression. In *CVPR*. IEEE, 2012.

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

[5] W. Chen, M.J. Er, and S. Wu. Illumination compensation and normalization for robust face recognition using discrete cosine transform in logarithm domain. *Systems, Man, and Cybernetics*, *Part B: Cybernetics*, *IEEE Transactions on*, 36(2):458-466, 2006.

[6] T.F. Cootes, G.J. Edwards, and C.J. Taylor. Active appearance models. *Pattern Analysis and Machine Intelligence*, *IEEE Transactions on*, 23(6):681-685, 2001.

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

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

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

[10] A. Dhall, O.V.R. Murthy, R. Goecke, J. Joshi, and T. Gedeon. Video and image based emotion recognition challenges in the wild: EmotiW 2015. In *ACM ICMI.* ACM, 2015.

[11] P. Ekman and W. Friesen. Facial action coding system: a technique for the measurement of facial movement. *Consulting Psychologists Press Inc*., San Francisco, CA, 1978.

[12] T. Hassner, S. Harel, E. Paz, and R. Enbar. Effective face frontalization in unconstrained images. In *CVPR*. IEEE, 2015.

[13] G. Hinton, L. Deng, et al. Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. *Signal Processing Magazine*, IEEE, 29(6):82-97, 2012.

[14] Y. Jia, E. Shelhamer, et al. Caffe: convolutional architecture for fast feature embedding. In *ACM MM*. ACM, 2014.

[15] S.E. Kahou, C. Pal, et al. Combining modality specific deep neural network models for emotion recognition in video. In *ACM ICMI.* ACM, 2013.

[16] T. Kanade, J. Cohn, and Y. Tian. Comprehensive database for facial expression analysis. In *FG*. IEEE, 2000.

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

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

[19] 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 *ACM ICMI.* ACM, 2014.

[20] P. Liu, J.T. Zhou, I.W.H. Tsang, Z. Meng, S. Han, and Y. Tong. Feature disentangling machine—a novel approach of feature selection and disentangling in facial expression analysis. In *ECCV*. Springer, 2014.

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

[22] P. Lucey, J.F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews. The Extended Cohn-Kanade Dataset (CK+): a complete dataset for action unit and emotion-specified expression. In *CVPR Workshops*. IEEE, 2010.

[23] M. Pantic, M.F. Valstar, R. Rademaker, and L. Maat. Web-based database for facial expression analysis. In *ACM MM*. ACM, 2005.

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

[25] M. Valstar and M. Pantic. Induced disgust, happiness and surprise: an addition to the MMI facial expression database. In *LREC Workshops*, 2010.

[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 *ACM Workshop on AVEC*. ACM, 2013.

[27] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *CVPR*. IEEE, 2001.

[28] Z. Wang, S. Wang, and Q. Ji. Capturing complex spatio-temporal relations among facial muscles for facial expression recognition. In *CVPR*. IEEE, 2013.

[29] X. Xiong and F. de la Torre. Supervised descent method and its applications to face alignment. In *CVPR*. IEEE, 2013.

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

[31] 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.

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