Multi-clue Fusion for Emotion Recognition in the Wild

Jingwei Yan[1], Wenming Zheng[1]*, Zhen Cui[1]*, Chuangao Tang[1],
Tong Zhang[1,2], Yuan Zong[1], Ning Sun[3]
[1]Research Center for Learning Science, Southeast University, Nanjing, China
[2]School of Information Science and Engineering, Southeast University, Nanjing, China
[3]Engineering Research Center of Wideband Wireless Communication Technology,
Nanjing University of Posts and Telecommunications, Nanjing, China
{yanjingwei, wenming_zheng, zhen.cui} @seu.edu.cn

ABSTRACT

In the past three years, Emotion Recognition in the Wild (EmotiW) Grand Challenge has drawn more and more attention due to its huge potential applications. In the fourth challenge, aimed at the task of video based emotion recognition, we propose a multi-clue emotion fusion (MCEF) framework by modeling human emotion from three mutually complementary sources, facial appearance texture, facial action, and audio. To extract high-level emotion features from sequential face images, we employ a CNN-RNN architecture, where face image from each frame is first fed into the fine-tuned VGG-Face network to extract face feature, and then the features of all frames are sequentially traversed in a bidirectional RNN so as to capture dynamic changes of facial textures. To attain more accurate facial actions, a facial landmark trajectory model is proposed to explicitly learn emotion variations of facial components. Further, audio signals are also modeled in a CNN framework by extracting low-level energy features from segmented audio clips and then stacking them as an image-like map. Finally, we fuse the results generated from three clues to boost the performance of emotion recognition. Our proposed MCEF achieves an overall accuracy of 56.66% with a large improvement of 16.19% with respect to the baseline.

CCS Concepts

•Security and privacy → Biometrics; •Computing methodologies → Neural networks;

Keywords

AFEW; multi-clue; emotion recognition in the wild; convolutional neural network (CNN); recurrent neural network (RNN)

* The corresponding authors

ICMI'16, November 12–16, 2016, Tokyo, Japan
© 2016 ACM. 978-1-4503-4556-9/16/11...$15.00
http://dx.doi.org/10.1145/2993148.2997630

1. INTRODUCTION

Emotion recognition has become a very hot topic in pattern recognition and affective computing. Its principal target is to automatically label a given image or video with a certain emotion type such as the six basic emotions (angry, disgust, fear, happy, sad, surprise) or neutral. It plays an essential role in a wide range of scenarios such as human behavior analysis, public security, children education and so on.

In the past three decades emotion recognition, including human facial expression recognition (FER) and speech emotion recognition (SER), has been widely studied by many researchers [2, 9, 23, 28]. However, among the various proposed methods, most of them work on the conventional databases collected under the laboratory controlled environment, such as multipie [8], CK+ [18], eNTERFACE [19] and so on. Recognizing emotion states in the wild is much more complicated due to unconstrained conditions involving changes of illuminations, occlusions, head pose variations and background noises. To solve this task, many methods have been proposed during the course of EmotiW challenges.

Yao et al. [26] explored the relationship between expression-specific facial features from muscle motions. Kaya et al. [12] combined a bunch of traditional features and employed least square based learners. Kahou et al. [5] used IRRR [15] to model the video sequence. Wu et al. [25] employed bag of features model to encode video clips. Furthermore, based on the development of deep learning [3, 14, 16], a number of CNN based methods are proposed to attempt to extract high-level emotion features of videos or images, and have achieved the state-of-the-art performance. Although some progresses have been made in emotion recognition in recent years as stated above, the technique of robust emotion feature extraction and fusion still need to be explored because the state-of-the-art accuracies are far from our expectation.

In this paper, to deal with video based emotion recognition problem, we propose a multi-clue emotion fusion (MCEF) framework by modeling human emotion from three mutually complementary sources, facial appearance texture, facial action, and audio, as shown in Fig. 1. To extract high-level emotion features from sequential face images, we employ a CNN-RNN architecture. In the architecture, face image from each frame is first fed into the fine-tuned VGG-Face network [20] to extract high-level face feature, and then the features of all frames are sequentially traversed in a bidirectional RNN [21] so as to capture dynamic changes of facial textures. To attain more accurate facial actions, facial land-

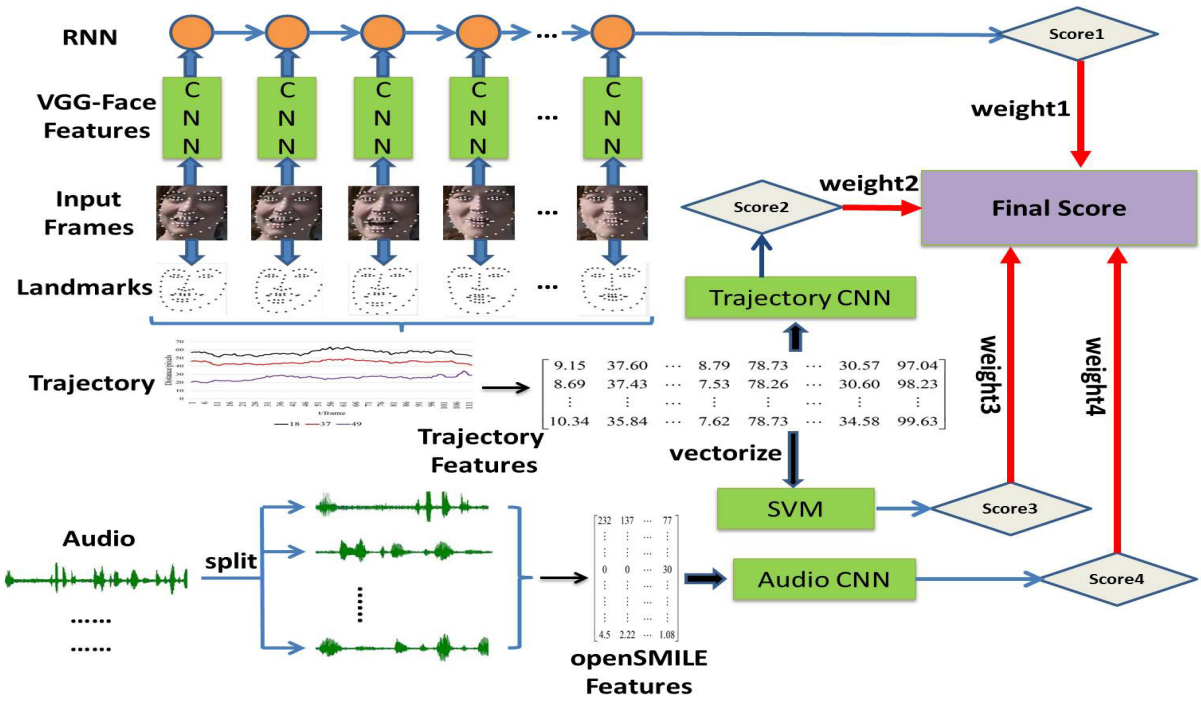

Figure 1: The proposed MCEF framework. From top to bottom, the appearance textures, facial actions and audio signals are respectively modeled as the CNN-RNN model, trajectory based SVM & CNN model, and a temporal-pyramid CNN model. All Scores from these models are finally fused to predict the emotion.

mark trajectories are modeled for explicitly learning emotion variations of facial components. Furthermore, we also model audio signals by using a CNN network. To capture audio information of different temporal regions, we segment an audio sequence into several clips by employing a temporal-pyramid structure and then extract low-level energy features from each clip. These energy features of all clips within an audio sequence are stacked as an image-like map for the CNN network. Finally, we fuse the classification scores generated from all these models to boost the performance of emotion recognition. The entire MCEF framework is shown in Fig. 1. The final results show the superiority of our method, which surpasses the baseline [1] by 16.19%.

2. LEARN FACIAL APPEARANCE TEXTURE

In order to obtain high-level emotion information, we use a fine-tuned VGG-Face model to extract video frame features and then feed the feature sequences to a RNN to model dynamic changes of the sequence. The detail of the CNN-RNN model is shown in the upper left of Fig. 1.

2.1 Detect Faces

Due to the database is collected from classic movies and TV reality shows, majority of the video frames contain a lot of background objects and other irrelevant information. Therefore the original frame can hardly be used directly.

We first detect face region in each frame of a video with a single shot detector (SSD) [17]. SSD is trained to detect multiple objects in the first place. Here we fine-tune the model on the Face Detection Data Set and Benchmark (FDDB) [10] and utilize it as a face detector. Most of the faces are successfully detected while there are still a small amount of frames where the detector failed. For that case we crop them into a smaller region manually and run the detector again until we get the correct ones. After all faces are cropped we resize the face images to a standard size of $224 * 224$ as the input of VGG-Face network.

2.2 Fine-tune VGG-Face Network

Different from other works like [5, 13], where new CNN architectures were designed and trained with additional data, we take advantage of the previously trained VGG-Face [20] model directly. VGG-Face model is a 16-layer or 19-layer CNN architecture which is developed by Visual Geometry Group (VGG) of University of Oxford and has shown great performance in face recognition tasks. As facial expression is closely related to human face, we believe that VGG-Face model can be trained to extract high-level facial texture features which are more discriminative in emotion recognition than other hand-crafted features. In this paper we employ the 16-layer VGG-Face model.

All the cropped face images from the same video clip are regarded as samples whose labels are the same with the video. In order to balance training sample quantity of each emotion category, we adopt some data augment methods such as flipping or rotating the original images to generate the new ones. Note that validation data are used to evaluate the network.

After extensive experiments we find VGG-Face model which is fine-tuned from the fifth convolutional layer can achieve a relatively high accuracy of 36.21% on the validation set while the model is not overfitting or collapsing. As for the

accuracy here seems not high enough, we think there maybe two reasons. First, the accuracy is calculated over the whole image samples rather than video-wise. Therefore, the actual accuracy should be higher than that. Second, as we know, facial expressions are formed gradually and not each of the frames contains a strong enough expression that easy to recognition. So it is acceptable that the VGG-Face network makes mistakes about quite a few frames. However our fine-tuned model is proved to performance well in the experiments we conduct.

2.3 Train Bidirectional RNN

Recurrent neural network (RNN) [7] is able to recognize patterns in sequences of data. It is often used to model temporal relationship because it can keep the previous state and feed that to the next neural node. Here we employ RNN to capture the dynamic differences within a video frame sequence. Therefore facial texture changes can be modeled to recognize facial expressions. For an ordinary RNN, the hidden state $\mathbf{h}_t$ at time step t can be formulated as:

$$\mathbf{h}_t = \Phi(\mathbf{U}\mathbf{h}_{t-1} + \mathbf{W}\mathbf{x}_t) \tag{1}$$

where $\mathbf{x}_t$ is the input at time step t, W is the weight matrix of the input data, U is the weight matrix of the last time step hidden state $\mathbf{h}_{t-1}$, and Φ is the activation function. Equation (1) shows that RNN is able to "memorize" the previous states to some extent. Here we take the last fully connected layer output of VGG-Face, i.e. fc7, which is a 4096-dimension vector, as the extracted frame feature. And the sequence of the frame features is the input of RNN.

Generally we treat video clip as a one-direction frame sequence, e.g. the frames are listed in chronological order. If we reverse the order of frames, there is still a sequence of video frames except that the facial expression process is in a reverse way. Therefore, we can utilize the reversed sequence as additional input data together with the original one and employ a model called Bidirectional RNN (BRNN) [21] to learn the temporal relations between both directions. According to [21], We can modify equation (1) as follows:

$$\begin{aligned} \mathbf{h}_t^f &= \Phi(\mathbf{U}^f \mathbf{h}_{t-1}^f + \mathbf{W}^f \mathbf{x}_t^f) \\ \mathbf{h}_t^b &= \Phi(\mathbf{U}^b \mathbf{h}_{t-1}^b + \mathbf{W}^b \mathbf{x}_t^b) \end{aligned} \tag{2}$$

where notation with superscript f indicates the forward sequence order, while notation with superscript b indicates the opposite. Then we sum the two direction hidden state together at the corresponding location:

$$\mathbf{h}_t = \mathbf{h}_t^f + \mathbf{h}_{T-t}^b \tag{3}$$

where T is the length of the video clip. The following process is the same as the ordinary one-direction RNN. We employ the classic back propagation through time (BPTT) [24] to train the BRNN. The flowchart of CNN-BRNN is shown in the upper left part of Fig. 1. Note that we only draw one-direction RNN for the convenience.

3. LEARN FACIAL LANDMARK TRAJECTORY

In addition to the CNN-BRNN framework adopted in the previous section, we consider the emotion recognition problem in a more straightforward way. Inspired by action unit (AU) [22] which indicates the fact that facial expression is a combination of several AUs' movements, i.e. facial expressions are generated from facial muscle movements. So we can describe facial actions as the trajectories of vital facial landmarks in an explicit manner and use those as some kind of underlying emotion feature. Once we set a fiducial point in the face, usually the apex of nose or the center point of two eyes, the motion of other landmarks can be calculated by a subtraction between the two coordinates. According to Jung et al. [11], it has been proved to be effective for facial expression recognition. Here we propose two models to explicitly leverage the trajectory information. In the first model trajectory feature is stretched into a vector for each video. Then a linear SVM is trained based on the features. In the second model trajectory feature is stretched into an image-like map for each video. Then we design a 4-layer CNN to learn a more discriminative model. The details of these two models are described in 3.2 and 3.3.

3.1 Landmark Trajectory

We use our own facial landmark detector [4] to localize 68 facial landmarks [1]. After landmarks are detected in each frame, the faces are aligned and scaled with the center points of two eyes. We use the six landmark points around each eye to estimate its center. Meanwhile affine transformation is applied to the 68 landmark points.

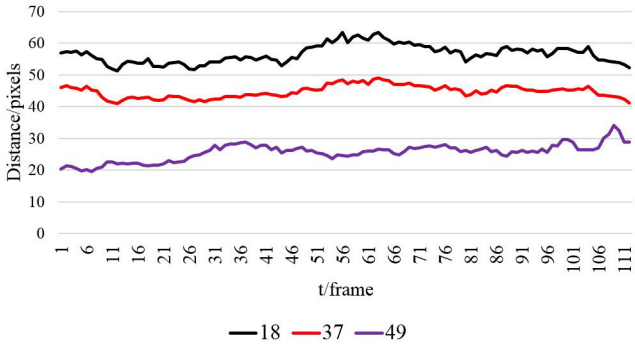

Figure 2: Illustration of some landmarks' trajectories. The sample used here is 004004880.avi from training set whose emotion state is happy. Black, red and purple curves represent ending point of right eyebrow, right eye and right corner of the mouth respectively.

An example of landmarks trajectories is shown in Fig. 2. Black, red and purple curves represent ending points of right eyebrow, right eye and right corner of the mouth respectively. Its sample is from training set and in the video there is an actress laughing happily. So as illustrated in Fig. 2, three related points are in the trajectory of lifting and falling back.

3.2 SVM Model

Following the setting in [11], we exclude the 17 points on the contour of the face, and adopt the 51 points located on the different facial parts, such as nose, eyes, eyebrows and mouth. Landmark points at frame t is presented as follows:

$$\mathbf{p}^{(t)} = [x_1^{(t)}, y_1^{(t)}, x_2^{(t)}, y_2^{(t)}, ..., x_{51}^{(t)}, y_{51}^{(t)}] \tag{4}$$

[1] http://ibug.doc.ic.ac.uk/resources/300-W/

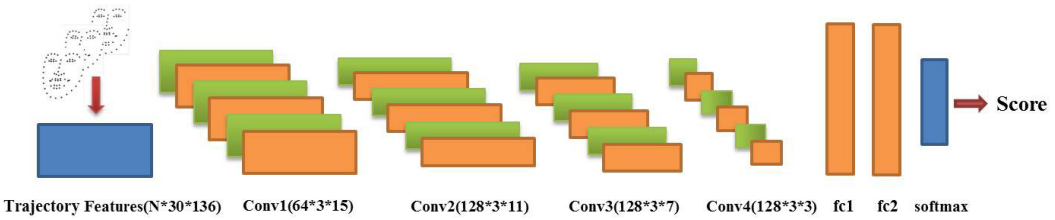

Figure 3: Framework of CNN based facial landmark trajectory model.

where $(x_i^{(t)}, y_i^{(t)})$ is the ith landmark coordinate. We then employ a normalization method:

$$\tilde{x}_i^{(t)} = \frac{x_i^{(t)} - x_c^{(t)}}{\delta_x^{(t)}}, \qquad \tilde{y}_i^{(t)} = \frac{y_i^{(t)} - y_c^{(t)}}{\delta_y^{(t)}} \tag{5}$$

where $x_c^{(t)}, y_c^{(t)}$ is the center landmark coordinate and $\delta_x^{(t)}, \delta_y^{(t)}$ is the standard deviation of (x, y) coordinates at the frame t. After we get the normalized landmark motion feature, features of the same video are concatenated together. Then a linear SVM is adopted to learn the trajectory based emotion recognition model.

3.3 CNN Model

As facial muscle action indicates expression in a visible and direct way, we believe there are more information encoded in the trajectory data that are worth exploring. To this end we propose a trajectory model based on CNN to extract the high-level features. Similar to the previous section, for each video clip 30 frames are randomly selected. For each frame 68 landmarks' trajectories are stretched into a 136-dimension vector. Then each video clip is transferred into a $30 * 136$ image-like map by combining all the feature vectors together. A 4-layer convolutional neural network is designed to train the emotion classification model as shown in Fig. 3.

4. LEARN AUDIO SIGNALS

Audio data plays an important role in emotion recognition. Due to that the AFEW database is collected from movies and TV shows, the audio signal components are quite complex. For example, some audio signal only contain background music, environmental noises or voices which are from other irrelevant characters. Despite the noisy audio data, the audio emotion recognition result can still be used to boost the overall performance.

Inspired by the work [27], a similar framework which uses CNN to deal with audio features is proposed. Different from converting the audio signal to an image by fourier transform [27], we extract low-level energy features of the audio by openSMILE [6] toolbox. For each audio signal, we segment it into many sub-clips by employing a temporal-pyramid structure so that different temporal information can be acquired. After extracting features of each sub-clip, we stack the features of the same audio source together into an image-like map as shown in the bottom of Fig. 1. A 4-layer CNN is then employed to learn a more discriminative feature for emotion classification. Both channels of the audio signal are used. The scheme of the designed CNN is similar to Fig. 3.

Table 1: Performance of 3-fold cross validation.

Model	Accuracy
Textures + CNN-BRNN	44.46%
Trajectory Features + SVM	37.37%
Trajectory Features + CNN	35.73%
Audio + CNN	30.88%
Fusion	49.22%

5. THE FUSION

In the previous three sections we present details of the MCEF framework that we propose. As facial textures, facial action and audio are complementary to some extent, we employ a decision fusion method to boost the MCEF performance. During the predicting phase each model generates a score matrix which indicates the probability of every predicted sample belonging to the related emotions. According to the performance of each model, we assigned each of them a proper weight. Then the final score matrix can be formulated as follows:

$$S = \omega_1 S^{Video} + \omega_2 S^{Tra\_SVM} + \omega_3 S^{Tra\_CNN} + \omega_4 S^{Audio} \tag{6}$$

where ω_i is the weight of each model. Note that in equation (6) there is only one score matrix for each model, however, multiple score matrices from one model can be employed because parameters in a model usually affect the performance and sometimes they can be complementary. We adopt the grid search strategy to optimize the weights on the validation set.

6. EXPERIMENT

6.1 Database and Parameter Setting

We conduct experiments on the competition dataset, i.e. Acted Facial Expressions In The Wild database (AFEW 6.0). In this dataset, the training and validation data consist of short video clips collected from classic movies, while the testing data also includes some reality TV show chips besides movie clips. There are totally 773 training samples, 383 validation samples and 593 testing samples. Each training and validation sample are assigned to one label from six basic emotions and the neutral emotion, while the test samples need to be classified. As the reality TV show is more spontaneous than movie characters' performance, it increases the difficulty of emotion recognition. The sophisticated technique with LBPTOP descriptor and SVR classifier achieves 38.81% and 40.47% respectively on validation and testing data, as provided by the challenge organizer as the baseline.

We strictly follow the competition protocol, and do not use other data. In CNN-BRNN model, we fix the length of input video frames to 40 frames. Videos which are longer than 40 frames are down-sampled by randomly picking or cutting a continuous 40 frames. Meanwhile, for videos whose lengths are shorter than 40 frames, we repeatedly pad the first frame into the beginning part of the sequence so as to form a new sequence with 40 frames. In order to mitigate over-fitting and increase the generalization ability of learnt models, we augment the original samples by flipping images or rotating images in a small angles. These procedures can expand our training set to a relatively large scale. Due to data limitation, we only adopt one layer BRNN. The BRNN learning rate is set to 0.003.

In the CNN based trajectory model, we sample each video into a sequence of 30 frames and then extract the trajectory curves of 68 landmarks. Thus the input of CNN is an $N * 30 * 136$ matrix, where N indicates the batchsize. The convolutional parameters are shown in Fig. 3. In the CNN based audio model, we segment one audio to 28 overlapping clips by using a temporal-pyramid structure, the input dimension is $N*28*1582$, where 1582 is the feature dimension extracted by openSMILE toolbox. The 4-layer convolutional parameters are respectively $32 * 3 * 15$, $32 * 3 * 11$, $32 * 3 * 7$ and $32 * 3 * 3$. The learning rate of both the two networks is set to 0.001.

6.2 Result

Figure 4: Confusion matrix of 3-fold cross validation.

As the scale of AFEW 6.0 database is not large, we combine the training and validation set together and employ 3-fold cross validation to tune our models, including how to configure the used networks and how to fuse the scores. The validation results are reported in Table. 1. Compared to other models, CNN+RBNN performs better and achieves an accuracy of 44.46%. By weighting all the scores, we can obtain an accuracy of 49.22%, which is obviously superior to the performance of each single model. It implies that the information from facial textures, facial action and audio are complementary to some extent. The corresponding confusion matrix is reported in Fig. 4.

During the test phase, our first three submissions are based on CNN-BRNN model alone. The third submission reaches an overall accuracy of 49.92%, which is nearly an 10% improvement with respect to the baseline. It indicates

Figure 5: Confusion matrix of the 3rd submission.

that our proposed CNN-BRNN model can work well compared with the sophisticated LBPTOP method. The corresponding confusion matrix is shown in Fig. 5.

Figure 6: Confusion matrix of the 8th submission.

Our last submission, i.e. the eighth submission, achieves an encouraging result in the challenge. The final overall accuracy is 56.66%, which surpasses significantly 16.19% on the baseline. From the confusion matrix in Fig. 6, our MCEF framework performs better on recognizing neutral, happy and angry samples while none of the test samples is classified as disgust or surprise labels. A possible reason is that, in the fusion stage, despite that there is one or two models label the sample as disgust or fear, however their scores are not large enough to balance other model's scores on the other wrong emotions. On the other hand, disgust and surprise themselves are indeed hard to recognize even for human, and they are often confused with other emotions. For our method, most of the disgust samples are mistaken as neutral and happy while most of the surprise samples are classified to similar emotions like neutral, sad and angry.

Another observation is that we achieve a better performance on the testing than the validation process. We think the main reason should be attributed to the distribution of testing samples. In the validation process, the number of samples for each class is nearly equal according to the provided data. But in the testing process, the sample number of each class becomes extremely unbalanced according to the confusion matrix in Fig 6. Even so, the performance

on each class is almost consistent according to Fig. 4 and Fig. 6, where the disgust and surprise samples are severely misclassified into other classes.

7. CONCLUSIONS

In this paper we presented a multi-clue emotion fusion framework to deal with the challenging emotion recognition in the wild problem. We characterize human emotions from three aspects including facial textures, facial action and audio. The proposed CNN-BRNN model can be successfully used to model the dynamic changes of facial textures. The trajectories of landmarks are employed to capture facial action. For audio signals in the video, the low-level features of temporal-pyramid manners are extracted and then fed into a specified CNN network. Finally we combine the previous scores together by considering their complementarity in recognizing emotions. The results reported in the challenge indicate that our proposed MCEF framework is more promising on the task emotion recognition.

8. ACKNOWLEDGMENTS

This work was partly supported by the National Basic Research Program of China under Grant 2015CB351704, the National Natural Science Foundation of China (NSFC) under Grants 61231002, 61572009 and 61471206, the Natural Science Foundation of Jiangsu Province under Grant BK20130020 and BK20141428.

9. REFERENCES

[1] J. J. Abhinav Dhall, Roland Goecke, J. Hoey, and T. Gedeon. Emotiw 2016: Video and group-level emotion recognition challenges. In *ICMI*. ACM, 2016.

[2] C.-N. Anagnostopoulos, T. Iliou, and I. Giannoukos. Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011. *Artificial Intelligence Review*, 43(2):155–177, 2015.

[3] D. Ciregan, U. Meier, and J. Schmidhuber. Multi-column deep neural networks for image classification. In *CVPR*, 2012.

[4] Z. Cui, S. Xiao, J. Feng, S. Yan, and W. Zheng. Recurrent shape regression. *IEEE TPAMI (under review)*, 2016.

[5] S. Ebrahimi Kahou, V. Michalski, K. Konda, R. Memisevic, and C. Pal. Recurrent neural networks for emotion recognition in video. In *ICMI*, pages 467–474. ACM, 2015.

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

[7] A. Graves, A.-r. Mohamed, and G. Hinton. Speech recognition with deep recurrent neural networks. In *ICASSP*, 2013.

[8] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-pie. *Image and Vision Computing*, 28(5):807–813, 2010.

[9] S. Happy and A. Routray. Automatic facial expression recognition using features of salient facial patches. *IEEE TAC*, 6(1):1–12, 2015.

[10] V. Jain and E. Learned-Miller. Fddb: A benchmark for face detection in unconstrained settings. Technical Report UM-CS-2010-009, University of Massachusetts, Amherst, 2010.

[11] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim. Joint fine-tuning in deep neural networks for facial expression recognition. In *ICCV*, 2015.

[12] H. Kaya, F. Gürpinar, S. Afshar, and A. A. Salah. Contrasting and combining least squares based learners for emotion recognition in the wild. In *ICMI*, pages 459–466. ACM, 2015.

[13] B.-K. Kim, H. Lee, J. Roh, and S.-Y. Lee. Hierarchical committee of deep cnns with exponentially-weighted decision fusion for static facial expression recognition. In *ICMI*, pages 427–434. ACM, 2015.

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[15] Q. V. Le, N. Jaitly, and G. E. Hinton. A simple way to initialize recurrent networks of rectified linear units. *arXiv preprint arXiv:1504.00941*, 2015.

[16] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[17] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. SSD: Single shot multibox detector. *arXiv preprint arXiv:1512.02325*, 2015.

[18] 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*, 2010.

[19] O. Martin, I. Kotsia, B. Macq, and I. Pitas. The enterface'05 audio-visual emotion database. In *ICDE Workshops*, 2006.

[20] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *British Machine Vision Conference*, 2015.

[21] M. Schuster and K. K. Paliwal. Bidirectional recurrent neural networks. *IEEE TSP*, 45(11):2673–2681, 1997.

[22] Y.-I. Tian, T. Kanade, and J. F. Cohn. Recognizing action units for facial expression analysis. *IEEE TPAMI*, 23(2):97–115, 2001.

[23] K. Wang, N. An, B. N. Li, Y. Zhang, and L. Li. Speech emotion recognition using fourier parameters. *IEEE TAC*, 6(1):69–75, 2015.

[24] P. J. Werbos. Backpropagation through time: what it does and how to do it. *Proceedings of the IEEE*, 78(10):1550–1560, 1990.

[25] J. Wu, Z. Lin, and H. Zha. Multiple models fusion for emotion recognition in the wild. In *ICMI*, pages 475–481. ACM, 2015.

[26] 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 *ICMI*, pages 451–458. ACM, 2015.

[27] S. Zhang, C. Liu, H. Jiang, S. Wei, L. Dai, and Y. Hu. Feedforward sequential memory networks: A new structure to learn long-term dependency. *arXiv preprint arXiv:1512.08301*, 2015.

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