MEC 2017: Multimodal Emotion Recognition Challenge

Ya Li
National Laboratory of Pattern Recognition, Institute of Automation Chinese Academy of Sciences, Beijing, China
yli@nlpr.ia.ac.cn

Jianhua Tao
National Laboratory of Pattern Recognition, CAS Center for Excellence in Brain Science and Intelligence Technology, Institute of Automation Chinese Academy of Sciences, School of Artificial Intelligence, University of Chinese Academy of Science, Beijing, China
jhtao@nlpr.ia.ac.cn

Björn Schuller
ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany,
Department of Computing, Imperial College London, UK,
Harbin Institute of Technology, China
bjoern.schuller@imperial.ac.uk

Shiguang Shan
Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
sgshan@ict.ac.cn

Dongmei Jiang
Northwestern Polytechnical University, Xi'an, China
jiangdm@nwpu.edu.cn

Jia Jia
Tsinghua University, Beijing, China
jjia@mail.tsinghua.edu.cn

***Abstract*—This paper introduces baselines for the Multimodal Emotion Recognition Challenge (MEC) 2017, which is a part of the first Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia) 2018. The aim of MEC 2017 is to improve the performance of emotion recognition in real-world conditions. The Chinese Natural Audio-Visual Emotion Database (CHEAVD) 2.0 is utilized as the challenge database, which is an extension of CHEAVD as released in MEC 2016. MEC 2017 has three sub-challenges and 31 teams participate in either all or part of them. 27 teams, 16 teams and 17 teams participate in audio (only), video (only) and multimodal emotion recognition sub-challenges, respectively. Baseline scores of the audio (only) and the video (only) sub-challenges are generated from Support Vector Machines (SVM) where audio features and video features are considered separately. In the multimodal sub-challenge, feature-level fusion and decision-level fusion are both utilized. The baselines of the audio (only), the video (only) and the multimodal sub-challenges are 39.2%, 21.7% and 35.7% in macro average precision.**

***Index Terms*—emotion recognition challenges, audio-visual corpus, multimodal features, fusion methods**

I. INTRODUCTION

Automatic emotion recognition is the technology to identify human's emotional states by analyzing human speech, facial expression and body gesture, etc. With the development of artificial intelligence, there is an explosion of interest in realizing more natural human-computer dialogue systems. As an essential aspect in the human-machine interaction, emotion recognition has received a large amount of attention [1-3].

Existing emotion challenges, such as the Audio/Visual Emotion Challenges (AVEC) [2, 4], the INTERSPEECH Emotion Challenge [5] and its predecessors at Interspeech, the Facial Expression Recognition & Analysis (FERA) [6], Emotion Challenge in the Wild Challenge (EmotiW) [1] or further related ones such as tasks in the MediaEval [7, 8] series have been organized. These are mostly based on spontaneous databases , which are an important effort to promote emotion recognition. However, the challenge databases utilized in those efforts do not cover the Chinese language. Since emotion expression varies across different languages and cultures, the Multimodal Emotion Recognition Challenge (MEC) provides a common platform and a common benchmark dataset to promote the research on multimodal emotion recognition for the Chinese language. In MEC 2016, 43 teams registered and 26 teams submitted their results. Most teams utilized traditional methods and deep neural networks to extract multimodal features, combined with fusion methods to boost recognition performance [9-13]. In the multimodal sub-challenge, most teams considered the audio modality and the video modality. Interestingly, [10, 11] utilized the textual modality through automatic speech recognition as well. Besides the work introduced in MEC 2016, [14] discusses various visual descriptors, such as Scale Invariant Feature Transform [15], Histogram of Oriented Gradients [16] and Local Phase Quantization [17] for emotion recognition. [18] learns task-specific AU-aware [19] facial features and encodes their latent relations for the robust expression recognition. [20] utilizes Convolutional Neural Networks, followed with Long-Short Term Memory [21], to extract sequence-level features. [22] introduces 3D convolutional networks [23] into emotion recognition, which models appearance and motion of video simultaneously. All these studies have made a significant progress in multimodal emotion recognition.

MEC 2017 has three sub-challenges following the first MEC 2016, and 31 teams participate in either all or part of these. 27, 16 and 17 teams participate in audio (only), video

This work is supported by the National Natural Science Foundation of China (NSFC) (No.61425017, No. 61773379), the National Key Research & Development Plan of China (No. 2017YFB1002804) and the Major Program for the National Social Science Fund of China (13&ZD189).

We thank the data providers for their kind permission to make their data for non-commercial, scientific use.

978-1-5386-5311-1/18/$31.00 ©2018 IEEE

(only) and multimodal emotion recognition sub-challenges, respectively. The Chinese Natural Audio-Visual Emotion Database (CHEAVD) 2.0 is utilized as the challenge dataset, which is an extension of the CHEAVD [3] that was released in MEC 2016. The extension was made by adding more samples. CHEAVD 2.0 is – just as its predecessor – selected from Chinese movies, soap operas and TV shows, which mimics real-world conditions.

In this paper, we present the baselines for MEC 2017, providing dataset, baseline methods and challenge protocols. Baseline scores of the audio (only) and the video (only) sub-challenges are generated from Support Vector Machines (SVM) where audio features and video features are considered, separately. To generate the baselines for the multimodal sub-challenge, different fusion methods, such as feature level fusion and decision level fusion, are considered as well.

The baseline audio and video feature sets provided by the organizers are free to use – either all or part of them. However, it is highly encouraged to follow the original protocols as outlined here if making comparisons to the participants. Participants will be allowed at most five trials to upload their results for evaluation on the test set for each sub-challenge. The best result among five submissions will be considered as their final scores in the competitions. Each registered team should submit a paper, introducing results and methods the team utilized, which will be peer-reviewed.

This paper is organized as follows. We describe CHEAVD 2.0 in detail in Section II. Baseline features and experimental results are illustrated in Section III and Section IV, separately. Section V concludes the whole paper.

II. Multimodal Emotional Database

A dataset, as a vital aspect in the data-driven approach followed in the challenge, promotes research in particular tasks. In order to provide a basic Chinese resource for the research on emotional multimodal interaction for real-world applications, we collected CHEAVD 2.0 and utilize it as the database for MEC 2017.

Fig. 1. Selected screenshots of videos in the CHEAVD 2.0 database.

CHEAVD 2.0 is an extension of CHEAVD as released in MEC 2016, adding 4178 samples. CHEAVD 2.0 is also selected from Chinese movies, soap operas and TV shows, which contains noise in the background to mimic real-world conditions. Selected screenshots of samples can be found in Fig. 1. CHEAVD 2.0 has 474 minutes of spontaneous emotional segments. 527 speakers, aging from child to elderly, are included in this database. The partition of the recordings with respect to gender distribution is as follows: 58.4% are male subjects, 41.6% are female subjects. The duration of these samples is ranging from 1 second to 19 seconds and the average duration is 3.3 seconds.

The discrete emotion annotation strategy is adopted in MEC 2017. To keep consistent in the emotion labeling, we asked four experienced taggers to label each sample in CHEAVD 2.0. Pairwise kappa coefficients are calculated to evaluate the annotation consistency, which are shown in Table I. Finally, the average of four annotations is adopted as the unique label of each segment by majority vote, and we only selected the top eight major emotion classes, namely, happiness, sadness, worry, anger, anxiety, surprise, disgust and neutral, containing 7030 samples. To assess emotion recognition performance, those samples are divided into three sets: the training set, the validation set and the testing set, which contain 4917, 707 and 1406 samples, respectively. The emotion class distribution of the dataset can be found in Table II. Participants can train their models on the training set, and choose hyper-parameters based on the validation set to find the best emotion recognition model with the highest performance. In the submission stage, the participants should upload their emotion predictions on the testing set.

TABLE I. THE PAIRWISE KAPPA COEFFICIENTS OF THE FOUR ANNOTATORS.

Annotators	***A1***	***A2***	***A3***	***A4***
A1		0.58	0.55	0.43
A2	0.58		0.52	0.41
A3	0.55	0.52		0.42
A4	0.43	0.41	0.42	

TABLE II. FINAL NUMBER OF INSTANCES FOR THE EIGHT EMOTION-CLASSES.

	Train	***Val***	***Test***	***Total***
Neutral	1400	200	400	2000
Angry	884	128	252	1264
Happy	828	119	236	1183
Sad	462	67	132	661
Worried	567	81	162	810
Anxious	457	66	131	654
Surprise	175	25	51	251
Disgust	144	21	42	207
Sum	4917	707	1406	7030

III. Features

A. Acoustic Features

For transparency and easy reproduction, we utilize the *eGeMAPSv01a.conf* specification of features as given in the open-source openSMILE toolkit [24] to extract the extended Geneva Minimalistic Acoustic Parameter Set, which is also the baseline feature set in the AVEC 2016 competition [25] and in MEC 2016 [3]. These features show high robustness for emotion recognition from speech [10].

In the baseline audio feature set, acoustic low-level descriptors cover spectral, cepstral, prosodic and voice quality information, which is shown in Table III in detail. As the audio data contains long continuous recordings, it uses fixed length

segments to extract functionals, which are shifted forward at a rate of 40 milliseconds. Overall, these acoustic baseline features contain 88 audio features.

TABLE III. ACOUSTIC FEATURES OF THE MULTIMODAL EMOTION RECOGNITION CHALLENGE OF MEC 2017.

Energy & spectral low-level descriptors (26)
Sum of auditory spectrum (loudness), α ratio (50-1000 Hz / 1-5 kHz) [1], Energy slope (0-500 Hz, 0.5-1.5 kHz) [1], Hammarberg index[1], MFCC 1-4[2], Spectral flux[2]
Voicing related low-level descriptors (16)
F0 (linear & semi-tone), Formants 1, 2, 3 (freq., bandwith, ampl.), Harmonic difference H1-H2, H1-H3, Log. HNR, jitter (local), shimmer (local)

[1]computed on voiced and unvoiced frames respectively;
[2]computed on voiced, unvoiced and all frames respectively

B. Visual Features

Local Binary Patterns on Three Orthogonal Planes (LBPTOP) [26] is chosen as the baseline visual feature set, which showed its emotion recognition performance in previous works [14, 27, 28].

LBPTOP is a dynamic texture, which extends texture to the temporal domain. Basic LBP has 59 features while utilizing the uniform code. LBPTOP extends basic LBP from two dimensions into three dimensions, applying relevant descriptors on the XY, XT and YT planes independently and concatenating according histograms together (cf. Fig. 2). To gain local information precisely, the block-based method is utilized, where original frames are divided into 2×2 blocks. In the end, $2\times2\times59\times3 = 708$ LBPTOP features are extracted.

To alleviate the background influence, facial pre-processing methods are essential, including grey processing, face detection, face transformation and face normalization. Facial pre-processing methods are following the methods used as in MEC 2016 [3], applying the tracking algorithm and toolkit [29] based on Viola and Jones [30]. As for LBPTOP, we utilize the open-source Matlab code created by Huang based on [26].

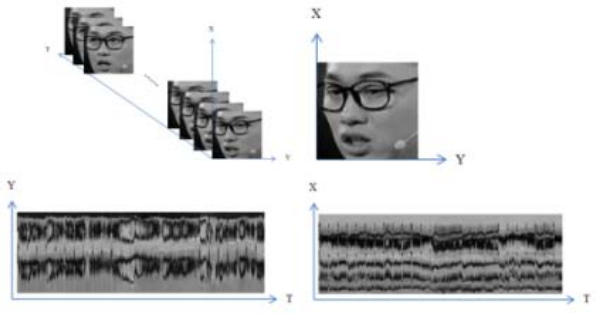

Fig. 2. Three dimensions of LBPTOP.

IV. BASELINE EXPERIMENTS

To ensure full reproducibility of the results，we compute the baseline results entirely relying on the public library scikit-learn. Compared with Random Forests, Adaboost and SVM, we find SVM to be more suitable for small dataset classification tasks. The model is optimized on the training dataset, and we choose the hyper-parameters based on the validation dataset to find the best emotion recognition model with the highest performance.

As emotion states are not evenly distributed in the real world, we choose macro average precision (MAP) as the primary measure in this challenge, and secondly the accuracy (ACC). The calculation methods for MAP and ACC employed here are given in Eqs. (1) – (3), respectively.

$$\mathrm{MAP} = \frac{1}{s}\times\sum_{i=1}^{s} P_i \ , \tag{1}$$

$$P_i = \frac{TP_i}{TP_i + FP_i} \quad \text{and} \tag{2}$$

$$ACC = \frac{\sum_{i=1}^{s} TP_i}{\sum_{i=1}^{s}(TP_i + FP_i)} \ , \tag{3}$$

where s represents the number of the emotion classes. TP_i and FP_i represent the number of true positive prediction and the false positive prediction of the i^{th} emotion class, respectively. P_i is the precision of the i^{th} emotion class.

Table IV shows hyper-parameters and baseline results for the audio (only), the video (only) and the multimodal sub-challenges. Table V compares two fusion methods in the multimodal sub-challenge: feature level fusion and decision level fusion. The confusion matrices of the baseline results for the three sub-challenges are shown in Figs. 3~5, separately.

TABLE IV. ACC (IN %) AND MAP (IN %) ON THE VALIDATION AND TESTING SETS FOR THE AUDIO (ONLY) AND THE VIDEO (ONLY) SUB-CHALLENGES.

	Parameters		***Val***		***Test***	
	Gamma	C	ACC	MAP	ACC	MAP
Audio (only)	3×10^{-3}	5.0	39.9	27.2	40.5	39.2
Video (only)	1×10^{-4}	12.0	36.5	34.1	35.3	21.7

TABLE V. ACC (IN %) AND MAP (IN %) OF FEATURE LEVEL FUSION AND DECISION LEVEL FUSION FOR THE MULTIMODAL SUB-CHALLENGE ON THE TESTING SET.

	ACC (%)	***MAP (%)***
Feature level fusion	43.0	29.1
Decision level fusion	40.3	35.7

From Table IV, one can see that the optimized classifiers have close ACC on the validation set and the testing set, while MAP shows the largest gap. This is because ACC tends to behave in favor of the data distribution, and if the emotion class labels are uniformly distributed, ACC has a higher value. However, the emotion classes are not evenly distributed in the real world, therefore, MAP is a more strict measure which is used to evaluate how the system performs overall across all emotion classes, regardless of a potentially low percentage. Since some of the emotion classes have only a few samples, the MAP is not very stable in some cases. The results show that more efforts need to made to improve the minority emotion classes.

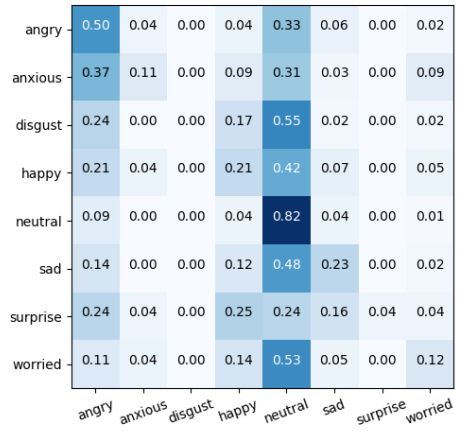

Fig. 3. Confusion matrix: audio baseline system on the testing set.

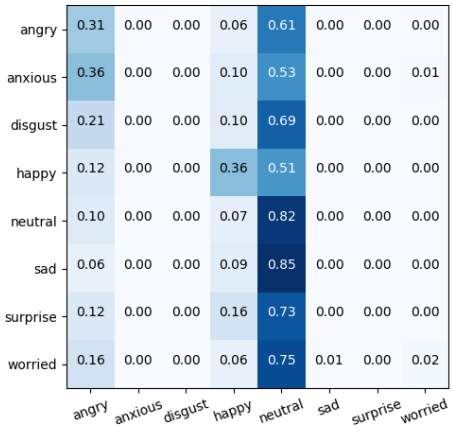

Fig. 4. Confusion matrix: video baseline system on the testing set.

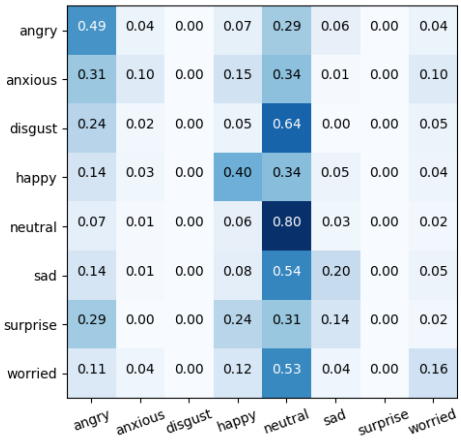

Fig. 5. Confusion matrix: multimodal baseline system on the testing set.

Through Table IV and Table V, the baselines of audio (only), visual (only) and multimodal sub-challenges are 39.2%, 21.7% and 35.7% in MAP. Therefore, the audio modality has the highest MAP among the three sub-challenges on the testing set, while the visual modality (only) is the worst case. However, it is also observed that decision level fusion has a large improvement on MAP.

Through Figs. 3~5, we find that anxiety, disgust, sadness, surprise and worry are hard to classify due to the lack of training samples. Anger and sadness are easily distinct from other emotions through the audio modality. And it appears that one can discriminate happiness well from other emotions through the visual modality. All non-neutral samples have a high chance to be misclassified as neutral due to the unbalanced class distribution which is, however, a natural phenomenon one has to face in a real-world task.

V. Conclusion

This paper introduces the baselines for Multimodal Emotion Recognition Challenge (MEC) 2017, which focuses on the introduction of the data, baseline methods and protocols for challenges. Existing emotion challenges, such as AVEC and EmotiW, are important efforts to promote emotion recognition. However, the challenge dataset used in those efforts does not cover the Chinese language. However, it is interesting to find out the state-of-the-art of emotion recognition for the Chinese language given the cultural differences and a different language that differs significantly from an acoustic point of view given its tonal nature. CHEAVD 2.0 was utilized as challenge dataset, containing 7030 samples and thus being larger than previous attempts at the topic. MEC 2017 has three sub-challenges: the audio (only), the video (only) and the multimodal sub-challenges, respectively. Acoustic features and visual features are extracted by open-source toolkits. Baseline scores for the single-modality sub-challenges are generated from an open-source SVM classifier. To get baseline scores for the multimodal sub-challenge, various fusion methods are considered. The baseline scores of the audio (only), the video (only) and the multimodal sub-challenges are 39.2%, 21.7% and 35.7% in MAP, respectively.

In future ambitions, we aim to continue this exciting challenge series aiming at broadening further up in terms of richness of the database and method inventory.

References

[1] A. Dhall, R. Goecke, J. Joshi, J. Hoey, and T. Gedeon, "Emotiw 2016: Video and group-level emotion recognition challenges," in Proceedings of the 18th ACM International Conference on Multimodal Interaction, 2016, pp. 427-432: ACM.

[2] B. Schuller, M. Valstar, F. Eyben, G. McKeown, R. Cowie, and M. Pantic, "Avec 2011–the first international audio/visual emotion challenge," in Affective Computing and Intelligent Interaction: Springer, 2011, pp. 415-424.

[3] Y. Li, J. Tao, B. Schuller, S. Shan, D. Jiang, and J. Jia, "MEC 2016: the multimodal emotion recognition challenge of CCPR 2016," in Chinese Conference on Pattern Recognition, 2016, pp. 667-678: Springer.

[4] F. Ringeval et al., "Av+ ec 2015: The first affect recognition challenge bridging across audio, video, and physiological data," in Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge, 2015, pp. 3-8: ACM.

[5] B. Schuller, S. Steidl, and A. Batliner, "The interspeech 2009 emotion challenge," in Tenth Annual Conference of the International Speech Communication Association, 2009.

[6] M. F. Valstar, B. Jiang, M. Mehu, M. Pantic, and K. Scherer, "The first facial expression recognition and analysis challenge," in 2011 IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011, pp. 921-926: IEEE.

[7] E. Dellandréa, L. Chen, Y. Baveye, M. V. Sjöberg, and C. Chamaret, "The mediaeval 2016 emotional impact of movies task," in MediaEval 2016 Multimedia Benchmark Workshop Working Notes Proceedings of the MediaEval 2016 Workshop, 2016.

[8] M. Zaharieva, B. Ionescu, A. L. Gînscă, R. L. Santos, and H. Müller, "Retrieving Diverse Social Images at MediaEval 2017: Challenges, Dataset and Evaluation," in Working Notes Proceedings of the MediaEval 2017 Workshop, Dublin, Ireland, 2017.

[9] B. Sun, Q. Xu, J. He, L. Yu, L. Li, and Q. Wei, Audio-Video Based Multimodal Emotion Recognition Using SVMs and Deep Learning. Springer Singapore, 2016.

[10] S. Chen et al., "Video Emotion Recognition via Fusing Multimodal Features," in Chinese Conference on Pattern Recognition, 2016.

[11] X. Sun et al., "MultiModal Emotion Analysis for Video Based on Hybrid Features," in Chinese Conference on Pattern Recognition, 2016.

[12] J. Ye, W. Zheng, Y. Li, Y. Zong, and Z. Cui, "Multimodal Emotion Recognition via Recurrent Neural Network," in Chinese Conference on Pattern Recognition, 2016.

[13] X. Xia, L. Guo, D. Jiang, E. Pei, L. Yang, and H. Sahli, "Audio Visual Recognition of Spontaneous Emotions In-the-Wild," in Chinese Conference on Pattern Recognition, 2016.

[14] H. Kaya, F. Gürpinar, S. Afshar, and A. A. Salah, "Contrasting and combining least squares based learners for emotion recognition in the wild," in Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, 2015, pp. 459-466: ACM.

[15] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," International journal of computer vision, vol. 60, no. 2, pp. 91-110, 2004.

[16] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005, vol. 1, pp. 886-893: IEEE.

[17] J. Heikkila, V. Ojansivu, and E. Rahtu, "Improved blur insensitivity for decorrelated local phase quantization," in 20th International Conference on Pattern Recognition (ICPR), 2010, pp. 818-821: IEEE.

[18] A. Yao, J. Shao, N. Ma, and Y. Chen, "Capturing au-aware facial features and their latent relations for emotion recognition in the wild," in Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, 2015, pp. 451-458: ACM.

[19] 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 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005, vol. 2, pp. 568-573: IEEE.

[20] S. Ebrahimi Kahou, V. Michalski, K. Konda, R. Memisevic, and C. Pal, "Recurrent neural networks for emotion recognition in video," in Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, 2015, pp. 467-474: ACM.

[21] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural computation, vol. 9, no. 8, pp. 1735-1780, 1997.

[22] Y. Fan, X. Lu, D. Li, and Y. Liu, "Video-based emotion recognition using CNN-RNN and C3D hybrid networks," in Proceedings of the 18th ACM International Conference on Multimodal Interaction, 2016, pp. 445-450: ACM.

[23] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4489-4497: IEEE.

[24] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in Proceedings of the 21st ACM international conference on Multimedia, 2013, pp. 835-838: ACM.

[25] F. Povolny et al., "Multimodal emotion recognition for AVEC 2016 challenge," in Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge, 2016, pp. 75-82: ACM.

[26] 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, vol. 29, no. 6, pp. 915-928, 2007.

[27] J. Wu, Z. Lin, and H. Zha, "Multiple models fusion for emotion recognition in the wild," in Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, 2015, pp. 475-481: ACM.

[28] B. Sun et al., "Combining multimodal features within a fusion network for emotion recognition in the wild," in Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, 2015, pp. 497-502: ACM.

[29] X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," in 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 532-539: IEEE.

[30] P. Viola and M. J. Jones, "Robust Real-Time Object Detection," in International Workshop on Statistical and Computational Theories of Vision -- Modeling, Learning, Computing, and Sampling, 2001, p. 87.