EmotiW 2016: Video and Group-Level Emotion Recognition Challenges

Abhinav Dhall
David R. Cheriton School of Computer Science
University of Waterloo, Canada
abhinav.dhall@uwaterloo.ca

Roland Goecke
Human-Centred Technology Centre
University of Canberra, Australia
roland.goecke@ieee.org

Jyoti Joshi
David R. Cheriton School of Computer Science
University of Waterloo, Canada
jyoti.joshi@uwaterloo.ca

Jesse Hoey
David R. Cheriton School of Computer Science
University of Waterloo, Canada
jhoey@cs.uwaterloo.ca

Tom Gedeon
Information Human Centred Computing
Australian National University, Australia
tom.gedeon@anu.edu.au

ABSTRACT

This paper discusses the baseline for the Emotion Recognition in the Wild (EmotiW) 2016 challenge. Continuing on the theme of automatic affect recognition 'in the wild', the EmotiW challenge 2016 consists of two sub-challenges: an audio-video based emotion and a new group-based emotion recognition sub-challenges. The audio-video based sub-challenge is based on the Acted Facial Expressions in the Wild (AFEW) database. The group-based emotion recognition sub-challenge is based on the Happy People Images (HAPPEI) database. We describe the data, baseline method, challenge protocols and the challenge results. A total of 22 and 7 teams participated in the audio-video based emotion and group-based emotion sub-challenges, respectively.

CCS Concepts

•Computing methodologies → Computer vision; Machine learning algorithms; •Information systems → *Multimedia databases;*

Keywords

Audio-video data corpus, Emotion recognition, Group-level emotion recognition, Facial expression challenge, Affect analysis in the wild

1. INTRODUCTION

The paper presents the baseline and results of the Emotion Recognition in the Wild (EmotiW) 2016 challenge. 'In the wild' here refers to the real-life, uncontrolled conditions such as diverse background situations (indoor/outdoor), illumination conditions, head motion occlusion, multiple people in an image and spontaneous expression etc. There are a large number of research works in lab-controlled conditions. As a stepping stone to move to automatic affect recognition in diverse conditions, the EmotiW challenge series is based on data representing real-world scenarios. EmotiW challenge series aims at providing a platform for researchers to benchmark the performance of their methods on 'in the wild' data. This year, EmotiW 2016 comprises of two sub-challenges – a) Video based emotion Recognition (VReco); b) Group-level Emotion Recognition (GReco).

The EmotiW series has been run as a grand challenge as part of the ACM International Conference on Multimodal Interaction (ICMI) since 2013. The task during the first EmotiW challenge at ACM ICMI 2013 was the VReco challenge [6]. The Acted Facial Expressions in the Wild (AFEW) 3.0 dataset was used as the basis for the challenge. A total of 27 teams registered for the challenge and 9 teams submitted the test labels. The data was divided into three sets: *Train*, *Validation* and *Test*. The second EmotiW challenge was organised at ACM ICMI and had 9 papers. It is interesting to note that the performance of the methods [18] in the second EmotiW challenge is better than the methods in first EmotiW challenge. However, there is much room for improvement for automatic emotion recognition in the wild.

The problem of emotion recognition in varied condition is multi-dimensional. Recent survey studies [3] [20] present details of the challenges and the state-of-the-art in affect analysis. To address one of the challenges, the third EmotiW challenge [5] added a new sub-challenge: static image based facial expression sub-challenge. The new sub-challenge focussed on facial expression classification in the scenario, where only a single frame is available. A total of 22 teams participated in the new sub-challenge and the Static Facial Expressions in the Wild (SFEW) 2.0 database [8] was used as the benchmarking data. This year, the fourth EmotiW challenge is being organised at the ACM ICMI 2016, Tokyo. There are two major additions this year: the VReco sub-challenge is based on the newer version 6.0 of the AFEW database. AFEW 6.0 has data from movies and reality TV

ICMI'16, November 12–16, 2016, Tokyo, Japan
© 2016 ACM. 978-1-4503-4556-9/16/11...$15.00
http://dx.doi.org/10.1145/2993148.2997638

Figure 1: The figure shows a sample image from the GReco sub-challenge. For representation purposes the group has been enclosed in a rectangle and the score value = 3 is the group-level happiness intensity.

shows. The intent is to add *more spontaneous* data to the EmotiW benchmarking effort. The new sub-challenge added this year is the automatic group-level emotion recognition, which is based on the HaPpy PeoplE Images (HAPPEI) database [4].

The rise of recent trend of photo sharing on the internet via social networking medium has resulted in an exponential increase in the amount of visual data from different social events. From the perspective of automatic affect recognition this poses a new challenge due to the presence of multiple people in an image. In these images, the group of people may be posing for a photograph or may be clicked unaware during an event, as a candid shot. Within the context of this sub-challenge, the task is to infer the perception of the overall group-emotion of the people in a group. Furthermore, the gamut of emotions is limited to the range of happiness intensity. Figure 1 shows a sample image, the bounding box represents the group of people. The value 3 is the group-level happiness intensity. The range of the happiness intensity is in the HAPPEI database is [0-5], where 0 corresponds to 'neutral' and 5 being the highest score corresponding to happiness ('thrilled').

From the user study in [4], it was noticed that the perception of the group-level emotion is effected by attributes, which can be broadly divided into bottom-up and top-down factors. Bottom-up here refers to the attributes of the group members such as age, attractiveness, gender, spontaneous expression and occlusion etc. The top-up attributes refer to the context, the scene background and the effect of the group structure etc.

There is not much prior research in the field of automatic group-level emotion recognition. A bottom-up approach based on analysing the contribution of each member of the group towards the overall group mood was proposed in [9]. A topic model with low-level features based bag-of-words and facial attributes was trained on the HAPPEI data. Weighted soft-assignment was used to compute the histogram for the bag-of-words. The weights here correspond to the contribution of a person in the group towards overall perception of the mood of the group. It is shown that the perception of group mood is based on different attributes and is not an averaging problem. An interesting experiment to access the mood of the passerby was conducted in [13].

Huang et al. [15] proposed a graphical model and local binary patterns based technique to infer the group-level emotion. Body and face level features were explored for inferring the group-level emotion in images using the Valence and arousal emotion annotations by Mou et al. [19]. For face analysis, they used the quantised local Zernike moments descriptor [21] and for the body pose, structural statistics are extracted using the pyramid of histogram of gradients [2]. In another work, automatic group-level emotion detection was described as a three emotion class problem. The new database: group affect database [10] is labelled with broad emotion categories (*positive*, *neutral* and *negative*) within the valence range.

Table 1: Attributes of the subset of the AFEW 6.0 database used in the EmotiW 2016 challenge.

Attribute	Description
Length of sequences	300-5400 ms
No. of annotators	3
Expression classes	Anger, Disgust, Fear, Happiness, Neutral, Sadness and Surprise
Total No. of expressions	1749
Video format	AVI
Audio format	WAV

2. DATA

EmotiW 2016 has two sub-challenges and there are two datasets as described below:

VReco - The first sub-challenge is the video based emotion recognition sub-challenge. The sub-challenge is being run for the fourth time and is based on the AFEW 6.0 database. AFEW is developed using a semi-automatic process. Subtitle for Deaf & Hearing impaired (SDH) closed captions are parsed for presence of keywords related to emotion such as 'angry', 'cry', 'sad' etc. Short sequences which contain the keyword are selected by the labeller if it contains relevant data. The details of database collection are discussed in [7]. The major difference between AFEW 6.0 and AFEW 5.0 (which formed the base of EmotiW 2015 [11]) is the addition of data from reality TV shows. The assumption here is that the expressions of the actors in reality TV shows can be more spontaneous as compared to the ones in movies.

Similar to the last year, EmotiW 2016, VReco sub-challenge data is divided into three sets: *Train* (773 samples), *Val* (383 samples) and *Test* (593 samples). The reality TV show data has been added to the Test set only. The rationale behind this is to test the generalisation performance of the participating methods, which are trained on the movie data and tested on data from both movies and reality TV shows. The data of the sub-challenge can be accessed via the challenge site[1] and contains two labelled sets. It is to be noted that the three data partitions are subject and movie independent i.e. the data in the three sets belongs to mutually exclusive movies and actors.

The sub-challenge's task is to classify a sample audio-video clip into one of the seven categories: *Anger*, *Disgust*, *Fear*, *Happiness*, *Neutral*, *Sadness* and *Surprise*. Table 1 discusses the details about the video samples in the database. The labeled training and validation sets were made available early in April and the new, unlabeled test set was made

[1] https://sites.google.com/site/emotiw2016/

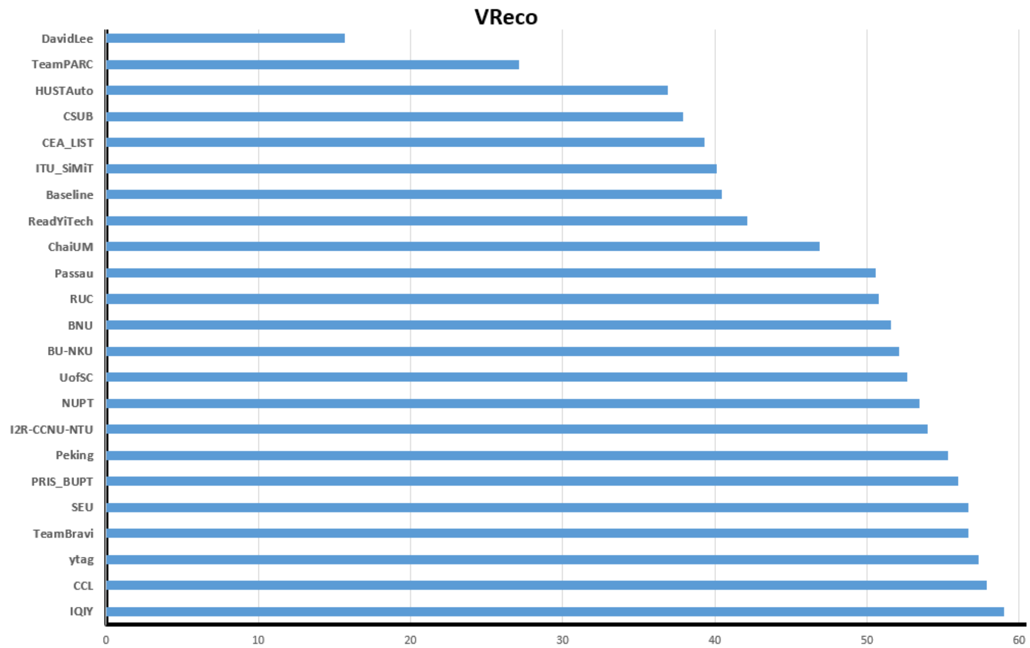

Figure 2: The graph compares the classification accuracy performance of participants in the VReco sub-challenge. Higher accuracy (along the X axis) represents better performance.

available in July 2016. There are no separate video-only, audio-only, or audio-video challenges. Participants are free to use either modality or both. Results for all methods will be combined into one set in the end. Participants are allowed to use their own features and classification methods. The labels of the testing set are unknown. Participants will need to adhere to the definition of training, validation and testing sets. In their papers, they may report on results obtained on the training and validation sets, but only the results on the testing set will be taken into account for the overall Grand Challenge results.

GReco - The HAPPEI database [9] forms the basis of the group-level emotion recognition sub-challenge. The HAPPEI database has been created from Flickr. Keywords related to social events such as 'convocation', 'marriage', 'party' etc were used to fetch images from difference social events. Images with multiple faces are further short-listed by running a face detector on the downloaded images. The images are further labelled with person-level meta-data. This meta-data consists of face locations in an image, occlusion intensity, facial happiness intensity and frontal/non-frontal pose information for each subject in images. All the fields in the meta-data have been labelled by human labellers and is only shared for the *Train* and the *Val* set. The face-level happiness intensity's range is [0-5]. This range of happiness intensity can be loosely mapped to: ['neutral' $\rightarrow$ 'small smile' $\rightarrow$ 'large smile' $\rightarrow$ 'small laugh' $\rightarrow$ 'large laugh' $\rightarrow$ 'thrilled']. Similar to the AFEW 6.0 database, HAPPEI has been divided into three sets: *Train* (1500 samples), *Val* (1138 samples) and *Test* (496 samples).

3. BASELINE EXPERIMENTS

3.1 VReco Sub-challenge

For computing the baseline results for the VReco sub-challenge, publicly available libraries are used. Pre-trained face models [29] are applied for face detection and initialisation of the Intraface tracking libray [25]. The fiducial points generated by Intraface are used for aligning the face. The face size is set to 128×128 pixels. Post aligning Local Binary Pattern-Three Orthogonal Planes (LBP-TOP) [28] features are extracted from non-overlapping spatial 4×4 blocks. LBP-TOP is a standard texture based feature, which has been extensively used for face-based affect classification [6] [5]. The LBP-TOP feature from each block are concatenated to create one feature vector. A non-linear Chi-square kernel based SVM is learnt for emotion classification (*Anger*, *Disgust*, *Fear*, *Happiness*, *Neutral*, *Sadness* and *Surprise*). The video only baseline system achieves 38.81% and 40.47% classification accuracy for the *Val* and *Test* sets, respectively. Please note that these are unweighed accuracies. The list of movies used in both VReco is mentioned in Section 7.

3.2 GReco Sub-challenge

As discussed in [2], the perception of the overall happiness (mood) intensity of a group is effected by top-down and bottom-up components. The top-down attributes are the factors external to a group member for eg: scene, neighbours etc. Bottom-up here refers to a group member's characteristics such as facial expression, facial attributes etc. Following [10], the baseline feature used in this challenge is the CENsus TRansform hISTogram (CENTRIST) descriptor [24]. CENTRIST is based on the Census tranform, which is similar to the local binary pattern. Since it is a scene descriptor, it is computed on the whole image by dividing the image into 4×4 non-overlapping blocks and takes into consideration both the bottom-up and top-down attributes. Support Vector Regression with a non-linear Chi-square kernel was

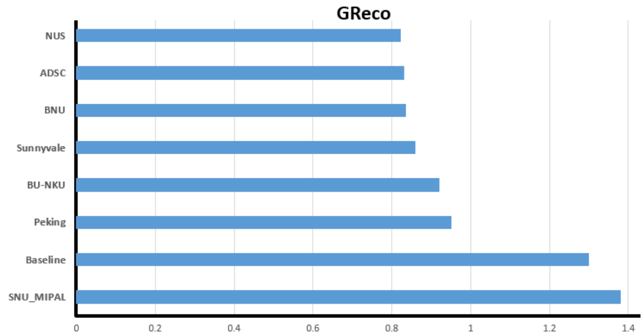

Figure 3: The figure shows a sample from the GReco sub-challenge, the score is the group-level happiness score. Lower RMSE shows better performance.

used to train the regression model. Root Mean Square Error (RMSE) is the proposed comparison metric. The parameters for regression were empirically found and the RMSE for Val and Test sets are 0.78 and 1.30, respectively. The winner method of this sub-challenge is the one with the smallest RMSE.

4. CHALLENGE RESULTS

Combined together, both the sub-challenges received 100 registrations. A total of 18 papers were submitted post the test phase. The VReco sub-challenge received test label sets for evaluation from 22 teams. Figure 2 shows the performance comparison of the participating methods based on the unweighted classification accuracy. The top performing Team Xers (Fan et al. [12]) proposed a pipeline based on 3D Convolutional Neural Networks (CNN) and recurrent neural network. Team CCL (Yao et al. [27]) proposed the HoloNet pipeline based on CNN and the method is the first runner-up. The second runner-up position was tied between two teams: TeamBravi (Bargal et al. [1]) and Team SEU (Yan et al. [26]).

In the GReco sub-challenge, a total of 7 teams submitted Test label sets for evaluation. Figure 3 shows the performance comparison of the proposed methods based on the RMSE metric. The top performing entry is from Team NUS (Li et al. [17]), the technique is based on ensemble of features in Long Short Term Memory (LSTM) [14] and ordinal regression. The first runner-up is the method from Team ADSC (Vonikakis et al. [23]), which is based on geometric features extracted from faces in an image. Partial least square regression is used to infer the group-level happiness intensity. Team BNU (Sun et al. [22]) proposed a LSTM based approach and fined tuned the AlexNet model [16] by training on the Static Facial Expressions in the Wild [8] and the HAPPEI databases.

5. CONCLUSION

The fourth Emotion Recognition in the Wild 2016 challenge provides a platform for researchers to benchmark and compete with their emotion recognition methods. This year the two sub-challenges: Audio-video emotion recognition and the group-level emotion recognition are based on the Acted Facial Expressions in the Wild database 6.0 and the Happy People Images databases, respectively. A total of 22 teams participated from industry and university labs in the audio-video emotion recognition sub-challenge. In the second sub-challenge a total 7 teams participated. The top 3 teams in both the sub-challenges were required submitted code/libraries/executables for validation.

It is to be noted that the performance of the top performing teams in the VReco sub-challenge is close and most of the teams use deep learning techniques. Extra data is also used and fine tuning is performed on convolutional neural network based pre-trained models. Not all teams use the aligned faces, which are provided and tried different deep learning based techniques. In the future, we will increase the amount of data and explore different emotion labelling techniques. For the group-level emotion recognition sub-challenge in the future image samples with more emotion labels will be added. The group affect database will be extended and more finer emotion label categories will be added.

6. ACKNOWLEDGEMENT

We are grateful to the ICMI'16 chairs, EmotiW program committee members and the reviewers. Part of the work was done by Abhinav Dhall, when at the University of Canberra and was supported by the Australian Research Council Discovery project grant ARCDP130101094. We also acknowledge the support by AGE-WELL NCE Inc., a member of the Networks of Centres of Excellence program and Alzheimer's Association grant ETAC-14-321494.

7. APPENDIX

Movie Names: 21, 50 50, About a boy, A Case of You, After the sunset, Air Heads, American, American History X, And Soon Came the Darkness, Aviator, Black Swan, Bridesmaids, Captivity, Carrie, Change Up, Chernobyl Diaries, Children of Men, Contraband, Crying Game, Cursed, December Boys, Deep Blue Sea, Descendants, Django, Did You Hear About the Morgans?, Dumb and Dumberer: When Harry Met Lloyd, Devil's Due, Elizabeth, Empire of the Sun, Enemy at the Gates, Evil Dead, Eyes Wide Shut, Extremely Loud & Incredibly Close, Feast, Four Weddings and a Funeral, Friends with Benefits, Frost/Nixon, Geordie Shore Season 1, Ghoshtship, Girl with a Pearl Earring, Gone In Sixty Seconds, Gourmet Farmer Afloat Season 2, Gourmet Farmer Afloat Season 3, Grudge, Grudge 2, Grudge 3, Half Light, Hall Pass, Halloween, Halloween Resurrection, Hangover, Harry Potter and the Philosopher's Stone, Harry Potter and the Chamber of Secrets, Harry Potter and the Deathly Hallows Part 1, Harry Potter and the Deathly Hallows Part 2, Harry Potter and the Goblet of Fire, Harry Potter and the Half Blood Prince, Harry Potter and the Order Of Phoenix, Harry Potter and the Prisoners Of Azkaban, Harold & Kumar go to the White Castle, House of Wax, I Am Sam, It's Complicated, I Think I Love My Wife, Jaws 2, Jennifer's Body, Life is Beautiful, Little Manhattan, Messengers, Mama, Mission Impossible 2, Miss March, My Left Foot, Nothing but the Truth, Notting Hill, Not Suitable for Children, One Flew Over the Cuckoo's Nest, Orange and Sunshine, Orphan, Pretty in Pink, Pretty Woman, Pulse, Rapture Palooza, Remember Me, Runaway Bride, Quartet, Romeo Juliet, Saw 3D, Serendipity, Silver Lining Playbook, Solitary Man, Something Borrowed, Step Up 4, Taking Lives, Terms of Endearment, The American, The Aviator, The Caller, The Crow, The Devil Wears Prada, The Eye, The Fourth Kind, The Girl with Dragon Tattoo, The Hangover, The Haunting, The Haunting of Molly Hartley, The Hills have Eyes 2, The Informant!, The King's Speech, The Last King of Scotland, The Pink Panther 2, The Ring 2, The Shinning, The Social Network, The Terminal, The Theory of Everything, The Town, Valentine Day, Unstoppable, Uninvited, Valkyrie, Vanilla Sky, Woman In Black, Wrong Turn 3, Wuthering Heights, You're Next, You've Got Mail.

8. REFERENCES

[1] S. Bargal, E. Barsoum, C. Ferrer Canton, and C. Zhang. Emotion Recognition in the Wild from Videos using Images. In *Proceedings of the ACM on International Conference on Multimodal Interaction (ICMI)*, 2016.

[2] A. Bosch, A. Zisserman, and X. Munoz. Representing Shape with a Spatial Pyramid Kernel. In *Proceedings of the ACM international conference on Image and video retrieval (CIVR)*, pages 401–408, 2007.

[3] C. A. Corneanu, M. Oliu, J. F. Cohn, and S. Escalera. Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: history, trends, and affect-related applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016.

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

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

[6] A. Dhall, R. Goecke, J. Joshi, M. Wagner, and T. Gedeon. Emotion recognition in the wild challenge 2013. In *Proceedings of the ACM on International Conference on Multimodal Interaction (ICMI)*, pages 509–516, 2013.

[7] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon. Acted Facial Expressions in the Wild Database. In *Technical Report*, 2011.

[8] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon. Static Facial Expression Analysis In Tough Conditions: Data, Evaluation Protocol And Benchmark. In *Proceedings of the IEEE International Conference on Computer Vision and Workshops BEFIT*, pages 2106–2112, 2011.

[9] A. Dhall, J. Joshi, I. Radwan, and R. Goecke. Finding Happiest Moments in a Social Context. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, pages 613–626, 2012.

[10] A. Dhall, J. Joshi, K. Sikka, R. Goecke, and N. Sebe. The more the merrier: Analysing the affect of a group of people in images. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–8, 2015.

[11] A. Dhall, O. Ramana Murthy, R. Goecke, J. Joshi, and T. Gedeon. Video and image based emotion recognition challenges in the wild: Emotiw 2015. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 423–426. ACM, 2015.

[12] Y. Fan, X. Lu, D. Li, and Y. Liu. Video-based Emotion Recognition Using CNN-RNN and C3D Hybrid Networks. In *Proceedings of the ACM on International Conference on Multimodal Interaction (ICMI)*, 2016.

[13] J. Hernandez, M. E. Hoque, W. Drevo, and R. W. Picard. Mood meter: counting smiles in the wild. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pages 301–310, 2012.

[14] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[15] X. Huang, A. Dhall, G. Zhao, R. Goecke, and M. Pietikäinen. Riesz-based volume local binary pattern and a novel group expression model for group happiness intensity analysis. In *Proccedings of the British Machine Vision Conference (BMVC)*, pages 34–1, 2015.

[16] 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.

[17] J. Li, S. Roy, J. Feng, and T. Sim. Happiness Level Prediction with Sequential Inputs via Multiple Regressions. In *Proceedings of the ACM on International Conference on Multimodal Interaction (ICMI)*, 2016.

[18] 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 16th International Conference on Multimodal Interaction*, pages 494–501. ACM, 2014.

[19] W. Mou, O. Celiktutan, and H. Gunes. Group-level arousal and valence recognition in static images: Face, body and context. In *Proceedings of the IEEE International Automatic Face and Gesture Recognition Conference and Workshops (FG)*, volume 5, pages 1–6. IEEE, 2015.

[20] E. Sariyanidi, H. Gunes, and A. Cavallaro. Automatic analysis of facial affect: A survey of registration, representation, and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1113–1133, 2015.

[21] E. Sariyanidi, H. Gunes, M. Gökmen, and A. Cavallaro. Local zernike moment representation for facial affect recognition. In *Proccedings of the British Machine Vision Conference (BMVC)*, 2013.

[22] B. Sun, Q. Wei, L. Li, Q. Xu, J. He, and L. Yu. LSTM for Dynamic Emotion and Group Emotion Recognition in the Wild. In *Proceedings of the ACM on International Conference on Multimodal Interaction (ICMI)*, 2016.

[23] V. Vonikakis, Y. Yazici, V. Dung Nguyen, and S. Winkler. Group happiness assessment using geometric features and dataset balancing. In *Proceedings of the ACM on International Conference on Multimodal Interaction (ICMI)*, 2016.

[24] J. Wu and J. M. Rehg. CENTRIST: A visual descriptor for scene categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(8):1489–1501, 2011.

[25] 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.

[26] J. Yan, W. Zheng, Z. Cui, C. Tang, T. Zhang, and Y. Zong. Multi-Clue Fusion for Emotion Recognition in the Wild. In *Proceedings of the ACM on International Conference on Multimodal Interaction (ICMI)*, 2016.

[27] A. Yao. HoloNet: Towards Robust Emotion Recognition in the Wild. In *Proceedings of the ACM on International Conference on Multimodal Interaction (ICMI)*, 2016.

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

[29] 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.