EmotiW 2019: Automatic Emotion, Engagement and Cohesion Prediction Tasks

Abhinav Dhall
Monash University
Indian Institute of Technology Ropar
abhinav.dhall@monash.edu

Roland Goecke
Univesity of Canberra
roland.goecke@ieee.org

Shreya Ghosh
Indian Institute of Technology Ropar
shreya.ghosh@iitrpr.ac.in

Tom Gedeon
Australian National University
tom@cs.anu.edu.au

ABSTRACT

This paper describes the Seventh Emotion Recognition in the Wild (EmotiW) Challenge. The EmotiW benchmarking platform provides researchers with an opportunity to evaluate their methods on affect labelled data. This year EmotiW 2019 encompasses three sub-challenges: a) Group-level cohesion prediction; b) Audio-Video emotion recognition; and c) Student engagement prediction. We discuss the databases used, the experimental protocols and the baselines.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision**; **Machine learning**.

KEYWORDS

Affect recognition, Multimodal Analysis, Neural networks, Emotion Recognition

ACM Reference Format:
Abhinav Dhall, Roland Goecke, Shreya Ghosh, and Tom Gedeon. 2019. EmotiW 2019: Automatic Emotion, Engagement and Cohesion Prediction Tasks. In *2019 International Conference on Multimodal Interaction (ICMI '19), October 14–18, 2019, Suzhou, China.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3340555.3355710

1 INTRODUCTION

The research in Human-Centred Artificial Intelligence (HCAI) is gaining momentum in various areas of computer science. This is primarily driven by the availability of data and processing power, which is allowing the researchers to investigate the usability of methods in understanding of user's intent and affect. This information is important to a machine for it to assist the user for increasing the productivity and wellbeing of a user. For progress in the development of methods, data labelled with affect is required. The seventh Emotion Recognition in the Wild (EmotiW) challenge is a benchmarking effort for researchers to evaluate their affect prediction methods on common data. EmotiW 2019[1] is being organised as part of ACM International Conference on Multimodal Interaction 2019, Suzhou, China. This year there are three sub-challenges - a) Group-level Cohesion (GC) prediction; b) Audio-Video (AV) emotion recognition and c) Student Engagement Predcition (EP). Details of the earlier EmotiW challenges can be tracked to EmotiW 2017 & EmotiW 2018.

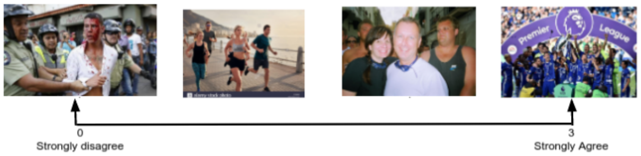

Figure 1: The figure shows the data [10] in the Group Cohesion sub-task of EmotiW 2019. The perceived cohesion is labelled in the range of [*strongly disagree - strong agree*].

2 GROUP-LEVEL COHESION PREDICTION

Group level affect related sub-challenges are being organized for the past three years at EmotiWs. The main motivation of this challenge is affect analysis in challenging conditions mainly in real-time. In fourth EmotiW challenge [5], the HAPpy PEople Images (HAPPEI) database [3] was used with CENsus TRansform hISTogram (CENTRIST) [24] descriptor as a baseline. CENTRIST is rich texture descriptor which

ICMI '19, October 14–18, 2019, Suzhou, China
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-6860-5/19/10…$15.00
https://doi.org/10.1145/3340555.3355710

[1] http://sites.google.com/view/emotiw2019

is computed by applying a local binary pattern like Census transformation. This descriptor can leverage both the top-down and bottom-up group-level attributes. The evaluation matrix for this challenge was the Root Mean Square Error (RMSE) value. Similarly, in fifth EmotiW challenge [4] Group AFfect Database 2.0 (GAF 2.0) is used with CENTRIST descriptor as a baseline. In the sixth EmotiW challenge [8], Inception V3 network is used as a baseline. The database was Group AFfect Database 3.0 (GAF 3.0). The evaluation matrix was overall accuracy. The accepted papers [11–13, 15, 16, 18, 19, 22, 23] mainly used deep learning-based facial, scene feature ensemble network to predict overall group-level emotion.

Group level cohesion is defined as the tendency of the group members to remain in unity in order to accomplish a common goal in the most well-organized way. In group dynamics, one of the most important requirements for the collaborative effort and effective teamwork is cohesion. The cohesiveness of a group is an essential indicator of the emotional state, structure and success of a group of people. The primary motivation behind this is to be able to predict the cohesiveness score of a group of people. The task of the sub-challenge is to classify a group's perceived cohesiveness in the range [0-3] where 0 represent low cohesion and 3 represents strong cohesion. This [0-3] score is measured in Group Cohesion Score (GCS).

To create the database [10], we used and extended the images from the GAF 3.0 database. GAF 3.0 was created via web crawling of various keywords related to social events (for eg: *world cup winners, wedding, family, laughing club, birthday party, siblings, riot, protest and violence* etc.). Cohesion labels were added to GAF 3.0. The total number of images in the sub-challenge are 16,443. We split the data into three parts: 9,300 images for training, 4,244 images for validation and 2,899 images for testing purposes.

The cohesion score in the GAF 3.0 database was labelled by 5 annotators (3 females and 2 males) of age group 21-30 years. In order to annotate data, the survey results assist in human perception regarding group cohesion. The annotators labelled each image for its cohesiveness in the range [0-3]. Treadwell et al. [20] argued that it is better to have these four 'anchor points' (i.e., *strongly agree, agree, disagree and strongly disagree*) instead of having low to high scores. The low to high score scaling may vary perception-wise from person to person. Thus, these soft scaled 'anchor points' are reliable (Figure 1). Along with GCS, GAF 3.0 database is also labelled with three group emotions (positive, negative and neutral) across the valance axis. Before the annotation, the annotators are familiarized with the concepts of GCS labels with corresponding images.

Rank	Team Name	Overall (MSE)
1	SML	0.41
2	SIAT	0.43
3	UD_ECE	0.44
4	CNU_ECE	0.47
5	CaeitInnov	0.48
6	CNU_MIP	0.49
7	Beijing Normal University	0.49
***	Baseline	0.50
8	IDAC	0.54

Table 1: MSE based leaderboard for the GC sub-challenge.

For computing the baseline, we trained the Inception V3 network followed by three fully connected layers (each having 4096 nodes) for the three classification task. We initialize the network with ImageNet pre-trained weights and fine-tune the network with SGD optimizer having a learning rate of 0.001 and momentum 0.9 without any learning rate decay. We use mean square error as the evaluation matrix. The performance of Inception V3 on *Validation* and *Test* sets are 0.84 and 0.50 respectively. Table 1 shows the performance of teams as compared to the baseline in this sub-challenge.

3 AUDIO-VIDEO EMOTION RECOGNITION SUB-CHALLENGE

The AV sub-challenge is the oldest running task in the EmotiW series [6]. This sub-challenge deals with the classic problem of prediction of discrete emotion labels (*anger*, *disgust*, *fear*, *happy*, *neutral* and *surprise*) from videos. The Acted Facial Expressions in the Wild (AFEW) [7] database is used for the tasks of training and evaluation. Each video contains a subject showing an emotion. The AFEW database has been collected with a semi-automatic process. Subtitles for people with hearing impairments contain words related to the emotion shown by the subject in a scene. The subtitles were parsed to generate noisy emotion labels. The labellers could either chose or discard the assigned label. The data in AFEW has been collected from movies and sitcoms, this gives an interesting variation the data in terms of head movement, occlusion, different illumination and (close to) spontaneous expressions. From the perspective of the audio modality, sample may contain background noise, low volume, music in the background. The expectation is that the methods in the AV sub-challenge will use both the audio and the video modality. Interesting methods have been proposed in the earlier EmotiWs for the AV task [2], [21] [9].

The AV baseline is based on computing the Local Binary Patterns in Three Orthogonal Planes (LBP-TOP) [26] facial descriptor. Zhu and Ramanan face detector [17] is used to initialize the face tracking process [25]. LBP-TOP is computed

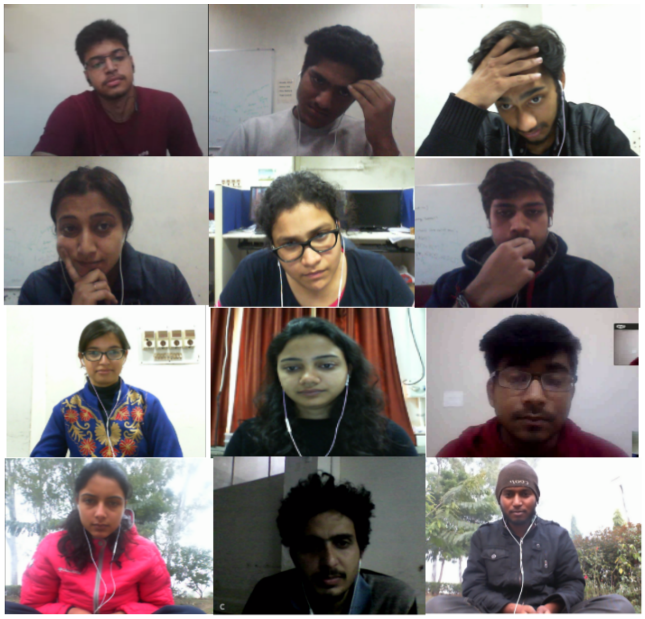

Figure 2: The figure shows the different environments in which the EngageWild database [14] has been recorded.

on blocks by dividing affine warped faces into a grid of 4 × 4. For classification, non-linear support vector machine was trained. For the *Validation* and the *Test* sets the classification accuracy achieved is 38.81% and 41.07%, respectively. Table 2 shows the performance of the 22 teams, which submitted *Test* set labels generated from their method for evaluation.

Rank	Team Name	Overall (%)
1	VAR	63.39
2	AIPL	62.78
3	USTC_NELSLIP & SIAT_MMLab	62.48
4	SeekTruthLab	62.02
5	CNU_MIP	61.56
6	KDDIResearch	58.65
7	ADLER	58.34
8	CNU_ECE	57.88
9	Beijing Normal University	52.98
10	LEI	51.45
11	KI	51.30
12	huochaitiantang	49.61
13	YCT	46.55
14	NeurodataLab	42.41
15	HEU-408	41.96
***	Baseline	41.07
16	CRIM	35.22
17	USTC_NELSLIP	34.15
18	NTUST	28.17
19	CobraLab	27.87
20	KB435	27.41
21	NWNU-FERT	24.042
22	Kaitou	19.44

Table 2: Classification accuracy (%) based comparison of the AV sub-challenge participating methods.

4 STUDENT ENGAGEMENT PREDICTION SUB-CHALLENGE

The EP task is to predict the engagement intensity of a subject in a video. This sub-challenge is based on the EngageWild database [14]. In a sample in the database, a subject watched an educational video (MOOC). The average duration of the video is 5 minutes. The range of intensity of engagement is *not engaged* (distracted) and *highly engaged*. The data has been recorded in diverse conditions and across different environments (Figure 2). For more details about the database, please refer to Kaur et al. [14]. The baseline of the sub-challenge is based on the head pose and the eye gaze features, which are extracted using the OpenFace 0.23 library [1]. The video is divided into segments. Each segment is represented by computing standard deviation across the head movement directions from the frames present in a segment. The eye gaze movement is also represented as average variance of the points returned for gaze of left and right eye in a particular segment compared to the mean eye points of the video. As a result, both the eye and head pose features are concatenated resulting in a 9 dimensional feature vector. Each video is represented using a collection of segments, where each segment is represented as a fused feature containing information of head pose and eye gaze. These features are passed through a Long short-term Memory (LSTM) layer, which returns activation for each segment of the video, passed to the flatten layer and then flattened feature vector is passed to the network of three dense layers followed by average pooling which gives the regressed value of engagement level of a video. MSE is used as the evaluation metric and the MSE for *Validation* and *Test* sets are 0.10 and 0.15, respectively. Table 3 shows the final leaderboard for the EP sub-challenge.

5 CONCLUSION

The paper presents the details of the EmotiW 2019 benchmarking. This year three tasks constituted as sub-challenges. Overall, each sub-challenge received good participation with the audio-video emotion recognition sub-challenge receiving the highest number of entries. It is interesting to note that the methods are focused towards deep learning. In the future, we will continue with this benchmarking and introduce newer problems related to affective computing.

Rank	Team Name	Overall (MSE)
1	SML	0.059
2	Tokyo AI Team	0.061
3	SIAT	0.062
4	UD-ECE	0.064
5	YCT	0.066
6	Tester	0.073
7	IntIntLab	0.077
8	Beijing Normal University	0.080
***	Baseline	0.150

Table 3: MSE based comparison of the EP sub-challenge participating methods.

ACKNOWLEDGMENTS

We are grateful to the ACM ICMI 2019 chairs, the EmotiW program committee members, the reviewers and the challenge participants. We acknowledge the support of Nvidia for their generous GPU grant.

6 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 Big Bang Theory, 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.

REFERENCES

[1] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. 2016. Openface: an open source facial behavior analysis toolkit. In *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*. IEEE, 1–10.

[2] Sarah Adel Bargal, Emad Barsoum, Cristian Canton Ferrer, and Cha Zhang. 2016. Emotion recognition in the wild from videos using images. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, 433–436.

[3] Abhinav Dhall, Roland Goecke, and Tom Gedeon. 2015. Automatic group happiness intensity analysis. *IEEE Transactions on Affective Computing* (2015), 13–26.

[4] Abhinav Dhall, R Goecke, S Ghosh, J Joshi, J Hoey, and T Gedeon. 2017. From Individual to Group-level Emotion Recognition: EmotiW 5.0. *ACM ICMI* (2017).

[5] Abhinav Dhall, Roland Goecke, Jyoti Joshi, Jesse Hoey, and Tom Gedeon. 2016. Emotiw 2016: Video and group-level emotion recognition challenges. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, 427–432.

[6] Abhinav Dhall, Roland Goecke, Jyoti Joshi, Michael Wagner, and Tom Gedeon. 2013. Emotion recognition in the wild challenge 2013. In *Proceedings of the 15th ACM on International conference on multimodal interaction*. ACM, 509–516.

[7] Abhinav Dhall, Roland Goecke, Simon Lucey, Tom Gedeon, et al. 2012. Collecting large, richly annotated facial-expression databases from movies. *IEEE multimedia* 19, 3 (2012), 34–41.

[8] Abhinav Dhall, Amanjot Kaur, Roland Goecke, and Tom Gedeon. 2018. EmotiW 2018: Audio-Video, Student Engagement and Group-Level Affect Prediction. In *ACM ICMI*.

[9] Yingruo Fan, Jacqueline CK Lam, and Victor OK Li. 2018. Video-based emotion recognition using deeply-supervised neural networks. In *Proceedings of the 2018 on International Conference on Multimodal Interaction*. ACM, 584–588.

[10] Shreya Ghosh, Abhinav Dhall, Nicu Sebe, and Tom Gedeon. 2019. Predicting Cohesiveness in Images. In *IJCNN*.

[11] Xin Guo, Luisa F Polanía, and Kenneth E Barner. 2017. Group-level emotion recognition using deep models on image scene, faces, and skeletons. In *ACM ICMI*.

[12] Xin Guo, Bin Zhu, Luisa F Polanía, Charles Boncelet, and Kenneth E Barner. 2018. Group-Level Emotion Recognition using Hybrid Deep Models based on Faces, Scenes, Skeletons and Visual Attentions. In *Proceedings of the 2018 on International Conference on Multimodal Interaction*. ACM, 635–639.

[13] Aarush Gupta, Dakshit Agrawal, Hardik Chauhan, Jose Dolz, and Marco Pedersoli. 2018. An Attention Model for group-level emotion recognition. In *Proceedings of the 2018 on International Conference on Multimodal Interaction*. ACM, 611–615.

[14] Amanjot Kaur, Aamir Mustafa, Love Mehta, and Abhinav Dhall. 2018. Prediction and localization of student engagement in the wild. In *2018 Digital Image Computing: Techniques and Applications (DICTA)*. IEEE, 1–8.

[15] Ahmed Shehab Khan, Zhiyuan Li, Jie Cai, Zibo Meng, James O'Reilly, and Yan Tong. 2018. Group-Level Emotion Recognition using Deep Models with A Four-stream Hybrid Network. In *Proceedings of the 2018 on International Conference on Multimodal Interaction*. ACM, 623–629.

[16] Jianshu Li, Sujoy Roy, Jiashi Feng, and Terence Sim. 2016. Happiness level prediction with sequential inputs via multiple regressions. In *ACM ICMI*.

[17] Deva Ramanan and Xiangxin Zhu. 2012. Face detection, pose estimation, and landmark localization in the wild. In *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Citeseer, 2879–2886.

[18] Bo Sun, Qinglan Wei, Liandong Li, Qihua Xu, Jun He, and Lejun Yu. 2016. LSTM for dynamic emotion and group emotion recognition in the wild. In *ICMI*. ACM, 451–457.

[19] Lianzhi Tan, Kaipeng Zhang, Kai Wang, Xiaoxing Zeng, Xiaojiang Peng, and Yu Qiao. 2017. Group emotion recognition with individual facial emotion CNNs and global image based CNNs. In *ACM ICMI*.

[20] Thomas Treadwell, Nicole Lavertue, VK Kumar, and Venkatesh Veeraraghavan. 2001. The group cohesion scale-revised: reliability and validity. *Journal of Group Psychotherapy, Psychodrama and Sociometry* 54, 1 (2001), 3.

[21] Valentin Vielzeuf, Corentin Kervadec, Stéphane Pateux, Alexis Lechervy, and Frédéric Jurie. 2018. An occam's razor view on learning audiovisual emotion recognition with small training sets. In *Proceedings of the 2018 on International Conference on Multimodal Interaction*. ACM, 589–593.

[22] Vonikakis Vonikakis, Yasin Yazici, Viet Dung Nguyen, and Stefan Winkler. 2016. Group happiness assessment using geometric features and dataset balancing. In *ACM ICMI*.

[23] Qinglan Wei, Yijia Zhao, Qihua Xu, Liandong Li, Jun He, Lejun Yu, and Bo Sun. 2017. A new deep-learning framework for group emotion recognition. In *ACM ICMI*. 587–592.

[24] Jianxin Wu and Jim M Rehg. 2010. Centrist: A visual descriptor for scene categorization. *IEEE transactions on pattern analysis and machine intelligence* 33, 8 (2010), 1489–1501.

[25] Xuehan Xiong and Fernando De la Torre. 2013. Supervised descent method and its applications to face alignment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 532–539.

[26] Guoying Zhao and Matti Pietikainen. 2007. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Transactions on Pattern Analysis & Machine Intelligence* 6 (2007), 915–928.