EmotiW 2018: Audio-Video, Student Engagement and Group-Level Affect Prediction

Abhinav Dhall
Indian Institute of Technology Ropar, India
abhinav@iitrpr.ac.in

Amanjot Kaur
Indian Institute of Technology Ropar, India
amanjot.kaur@iitrpr.ac.in

Roland Goecke
University of Canberra, Australia
roland.goecke@ieee.org

Tom Gedeon
Australian National University, Australia
tom.gedeon@anu.edu.au

ABSTRACT

This paper details the sixth Emotion Recognition in the Wild (EmotiW) challenge. EmotiW 2018 is a grand challenge in the ACM International Conference on Multimodal Interaction 2018, Colarado, USA. The challenge aims at providing a common platform to researchers working in the affective computing community to benchmark their algorithms on 'in the wild' data. This year EmotiW contains three sub-challenges: a) Audio-video based emotion recognition; b) Student engagement prediction; and c) Group-level emotion recognition. The databases, protocols and baselines are discussed in detail.

KEYWORDS

Emotion Recognition; Affective Computing;

ACM Reference Format:
Abhinav Dhall, Amanjot Kaur, Roland Goecke, and Tom Gedeon. 2018. EmotiW 2018: Audio-Video, Student Engagement and Group-Level Affect Prediction. In *2018 International Conference on Multimodal Interaction (ICMI '18), October 16–20, 2018, Boulder, CO, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3242969.3264972

1 INTRODUCTION

The sixth Emotion Recognition in the Wild (EmotiW)[1] challenge is a series of benchmarking effort focussing on different problems in affective computing in real-world environments. This year's EmotiW is part of the ACM International Conference on Multimodal Interaction (ICMI) 2018. EmotiW is a challenge series annually organised as a grand challenge in ICMI conferences. The aim is to provide a competing platform for researchers in affective computing. For details about the earlier EmotiW challenge, please refer to EmotiW 2017's baseline paper [3]. There are other efforts in the affective computing community, which focus on different problems such as depression analysis (Audio/Video Emotion Challenge [14]) and continous emotion recognition (Facial Expression Recognition and Analysis [16]). Our focus is affective computing in 'in the wild' environments. Here 'in the wild' means different real-world conditions, where subjects show head pose change, have varied illumination on the face, show spontaneous facial expression, there is background noise and occlusion etc. An example of the data captured in different environments can be seen in Figure 1.

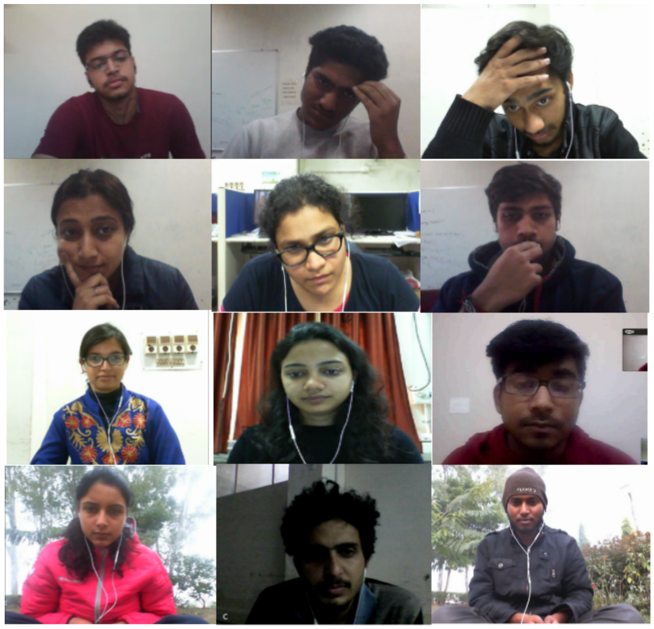

Figure 1: The images of the videos in the student engagement recognition sub-challenge [9]. Please note the varied backgrounds environment and illumination.

EmotiW 2018 contains three sub-challenges: a) Student Engagement Prediction (EngReco); Audio-Video Emotion Recognition (VReco); and c) Group-level Emotion Recognition (GReco). EngReco is a new problem introduced this year. In total there were over 100 registrations in the challenge. Below, we discuss the three sub-challenges, their baseline, data, evaluation protocols and results.

2 STUDENT ENGAGEMENT RECOGNITION

Student engagement in MOOCs is a challenging task. Engagement is one of the affective state which is a link between the subject and

[1] https://sites.google.com/view/emotiw2018

ICMI '18, October 16–20, 2018, Boulder, CO, USA
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-5692-3/18/10…$15.00
https://doi.org/10.1145/3242969.3264972

resource. It has various aspects such as emotional, cognitive and behavioral aspect. Challenge involved in engagement level detection of a user is that it does not remain same always, while watching MOOC material. To help the students to retain their attention level or track those parts of the video where they loss the attention it is mandatory to track student engagement based on various social cues such as looking away from the screen, feeling drowsy, yawning, being restless in the chair and so on. User engagement tracking is vital for other application such as detecting vehicle driver 's attention level while driving, customer engagement while reviewing new product. With the advent of e-learning environment in the education domain automatic detection of engagement level of students based on computer vision and machine learning technologies is the need of the hour. An algorithmic approach for automatic detection of engagement level requires dataset of student engagement. Due to unavailability of datasets for student engagement detection in the wild new dataset for student engagement detection is created in this work. It will address the issue of creating automatic student engagement tracking software. It will be used for setting performance evaluation benchmark for student engagement detection algorithms. In the literature, various experiments are conducted for student engagement detection in constrained environment. Various features used for engagement detection are based on Action Units, facial landmark points, eye movement, mouse clicks and motion of head and body.

2.1 Data Collection and Baseline

Our database collection details are discussed in the Kaur et al. [9]. Student participants were asked to watch five minutes long MOOC video. The data recording was done with different methods: through Skype, using a webcam on a laptop or computer and using a mobile phone camera. We endeavored to capture data in different scenarios. This is inorder to simulate different environments, in which students watch learning materials. The different environments used during the recording are computer lab, playground, canteen, hostel rooms etc. In order to introduce unconstrained environment effect different lighting conditions are also used as dataset is recorded at different times of the day. Figure 1 shows the different environments represented in the engagement database.

The data was divided into three sub-sets: *Train*, *Validation* and *Test*. In the dataset total 149 videos for training and 48 videos for validation are released. Testing data contains 67 videos. Dataset split follows subject independence i.e no subject is repeated among the three splits. The class wise distribution of data is as follows: 9 videos belong to level 0, 45 videos belong to level 1, 100 videos belong to level 2 and remaining 43 videos belong to level 3. The dataset has total 91 subjects (27 females and 64 males) in total. The age range of the subjects is 19-27 years. The annotation of the dataset is done by 6 annotators. The inter reliability of the annotators is measured using weighted Cohen's K with quadratic weights as the performance metric. This is a regression problem so the labels are in the range of [0 - 1].

For the baseline eye gaze and head movement features are computed. The eye gaze points and head movement w.r.t to camera movement are extracted using the OpenFace library [1]. The approach is as follows: firstly, all the videos are down sampled to same number of frames. The video is then divided into segments with 25% overlap. For each segment statistical features are generated such as standard deviation of the 9 features (from OpenFace). As a result each video has 100 segments, where each segment is represented with the help of the 9 features. By learning a long short term memory network the Mean Square Error (MSE) are 0.10 and 0.15 for the *Validation* and the *Test* sets, respectively. The performance of the competing teams on the *Test* set can be viewed in Table 1. A total of 6 teams submitted the labels for evaluation during the testing phase. Please note that this list is preliminary as the evaluation of code of the top three teams is underway. The same applies to the other two sub-challenges.

3 GROUP-LEVEL EMOTION RECOGNITION

This sub-challenge is the continuation of EmotiW 2017's GReco sub-challenge [3]. The primary motivation behind this is to be able to predict the emotion/mood of a group of people. Given the large increase in the number of images and videos, which are posted on social networking platforms, there is an opportunity to analyze affect conveyed by a group of people. The task of the sub-challenge is to classify a group's perceived emotion as *Positive*, *Neutral* or *Negative*. The labeling is representation of the Valence axis. The images in this sub-challenge are from the Group Affect Database 3.0 [5]. The data is distributed into three sets: *Train*, *Validation* and *Test*. The *Train*, *Validation* and *Test* sets contain 9815, 4346 and 3011 images, respectively. As compared to the EmotiW 2017 the amount of data has increased three folds.

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 use stochastic gradient descent optimizer without any learning rate decay to train the model. The classification accuracy for the *Validation* and *Test* sets are 65.00% and 61.00%, respectively. The performance of the competing teams in this sub-challenge are reported in the Table 2. A total of 12 teams submitted labels for evaluation during the testing phase.

4 AUDIO-VIDEO BASED EMOTION RECOGNITION

The VReco sub-challenge is the oldest running task in the EmotiW challenge series. The task is based on the Acted Facial Expressions in the Wild (AFEW) database [4]. AFEW database has been collected from movies and TV serials using a keyword search. Subtitles for hearing impaired contain keywords, which may correspond to the emotion of the scene. The short sequences with subtitles containing emotion related words were used as candidate samples. The database is then curated with these candidate audio-video samples. The database similar to the other two databases in EmotiW has been divided into three subsets: *Train*, *Validation* and *Test*.

The task is to predict the emotion of the subject in the video. Universal categorical emotion representation (*Angry*, *Disgust*, *Fear*, *Happy*, *Neutral*, *Sad* and *Surprise*) is used for representing emotions.

The baseline is computed as follows: face detection [22] is performed for initializing the tracker [19]. The face volume of aligned faces is divided into non-overlapping patches of 4 × 4 and Local Binary Patterns in Three Orthogonal Planes (LBP-TOP)[21] is computed. LBP-TOP captures the spatio-temporal changes in the texture.

Engagement Prediction Challenge		
Rank	**Team**	**MSE**
1	SIAT [20]	0.06
2	VIPL_Engagement [13]	0.07
3	IIIT_Bangalore [15]	0.08
4	Liulishuo [2]	0.08
5	Touchstone	0.09
6	**Baseline**	0.15
7	CVSP_NTUA_Greece	2.97

Table 1: The Table shows the comparison of participants in the student engagement prediction sub-challenge (RMSE) on the *Test* set. Note that this is the initial ranking and may change before the event.

Group-level Emotion Recognition		
Rank	**Team**	**Class. Accuracy (%)**
1	UD-ECE [7]	68.08
2	SIAT [18]	67.49
3	UofSC [10]	66.29
4	LIVIA [8]	64.83
5	ZJU_CADLiu_HanchaoLi	62.94
6	ZJU_IDI	62.90
7	FORFUN	62.11
8	SituTech	61.97
9	midea	61.31
10	**Baseline**	61.00
11	Beijing Normal University	59.28
12	UNIMIB-IVL	57.82
13	AMIKAIST	39.46

Table 2: The Table shows the comparison of participants in the Group-level emotion recognition sub-challenge (Classification Accuracy) on the *Test* set. Note that this is the initial ranking and may change before the event.

Audio-Video Emotion Recognition		
Rank	**Team**	**Class. Accuracy (%)**
1	SituTech [11]	61.87
2	E-HKU [6]	61.10
3	AIPL [12]	60.64
3	OL_UC [17]	60.64
5	UoT	60.49
6	NLPR	60.34
7	INHA	59.72
8	SIAT	58.04
9	TsinghuaUniversity	57.12
10	AIIS-LAB	56.51
11	VU	56.05
12	UofSC	55.74
13	VIPL-ICT-CAS	55.59
14	Irip	55.13
15	ZBC_Lab	54.98
16	Summerlings	54.82
17	EmoLab	54.21
18	CNU	53.75
19	Mind	53.60
20	Midea	53.45
21	KoreaUniversity	53.14
22	Kaitou	51.76
23	Beijing Normal University	50.54
24	USTC_NELSLIP	48.70
25	PopNow	48.09
26	BUCT	45.94
27	BIICLab	42.57
28	CobraLab	41.81
29	**Baseline**	41.07
30	17-AC	35.83
31	SAAMWILD	33.84
32	Juice	25.27

Table 3: The Table shows the comparison of participants in the audio-video emotion recognition sub-challenge (Classification Accuracy) on the *Test* set. Note that this is the initial ranking and may change before the event.

For classification, we trained a non-lines support vector machine. The classification accuracy (%) on the *Validation* and *Test* set are 38.81% and 41.07%, respectively. The data in this sub-challenge is similar to that of EmotiW 2017 [3]. Table 3 shows the comparison of the classification accuracy for 31 teams in this sub-challenge. It is notable that the performance of most the teams outperforms the baseline. Most of the proposed techniques are based on deep learning.

5 CONCLUSION

The sixth Emotion Recognition in the Wild is a challenge in the ACM International Conference on Multimodal Interaction 2018, Boulder. There are three sub-challenges in EmotiW 2018. Engagement detection of students while watching MOOCs is the first sub-challenge which deals with the task of engagement recognition from the recorded videos of subjects while watching the stimuli. The second sub-challenge is related to emotion recognition based on the universal emotion categories from the audio-visual data collected from movies. The third sub-challenge is related to collective emotions at group-level from the images. Different interesting methods were proposed by the challenge participants to solve these sub-challenges. The top performing methods are based on deep learning in all the sub-challenges, specifically based on ensemble of networks.

6 ACKNOWLEDGEMENT

We are grateful to the ACM ICMI 2018 chairs, the EmotiW program committee members, the reviewers and the challenge participants. We aslso acknowledge the support of Nvidia for their generous GPU grant.

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 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] Cheng Chang, Chen Zhang, Lei Chen, and Yang Liu. 2018. An Ensemble Model Using Face and Body Tracking for Engagement Detection. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*. ACM.

[3] Abhinav Dhall, Roland Goecke, Shreya Ghosh, Jyoti Joshi, Jesse Hoey, and Tom Gedeon. 2017. From individual to group-level emotion recognition: EmotiW 5.0. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. ACM, 524–528.

[4] Abhinav Dhall, Roland Goecke, Simon Lucey, and Tom Gedeon. 2012. Collecting large, richly annotated facial-expression databases from movies. *IEEE Multimedia* 19, 3 (2012), 0034.

[5] Abhinav Dhall, Jyoti Joshi, Karan Sikka, Roland Goecke, and Nicu Sebe. 2015. 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)*. 1–8.

[6] Yingruo Fan. 2018. Video-based Emotion Recognition Using Deeply-Supervised Neural Networks. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*. ACM.

[7] Xin Guo. 2018. Group-Level Emotion Recognition using Hybrid Deep Models based on Faces, Scenes, Skeletons and Visual Attentions. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*. ACM.

[8] Aarush Gupta, Dakshit Agrawal, Hardik Chauhan, Jose Dolz, and Marco Pedersoli. 2018. An Attention Model for group-level emotion recognition. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*. ACM.

[9] Amanjot Kaur, Aamir Mustafa, Love Mehta, and Abhinav Dhall. 2018. Prediction and Localization of Student Engagement in the Wild. *arXiv preprint arXiv:1804.00858* (2018).

[10] 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 20th ACM International Conference on Multimodal Interaction*. ACM.

[11] Chuanhe Liu. 2018. Multi-Feature Based Emotion Recognition for Video Clips. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*. ACM.

[12] Cheng Lu and Wenming Zheng. 2018. Multiple Spatio-temporal Feature Learning for Video-based Emotion Recognition in the Wild. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*. ACM.

[13] Xuesong Niu, Hu Han, Jiabei Zeng, Xuran Sun, Shiguang Shan, and Xilin Chen. 2018. Automatic Engagement Prediction with GAP Feature. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*. ACM.

[14] Fabien Ringeval, Björn Schuller, Michel Valstar, Jonathan Gratch, Roddy Cowie, Stefan Scherer, Sharon Mozgai, Nicholas Cummins, Maximilian Schmitt, and Maja Pantic. 2017. Avec 2017: Real-life depression, and affect recognition workshop and challenge. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*. ACM, 3–9.

[15] Chinchu Thomas, Nitin Nair, and Dinesh Babu J. 2018. Predicting Engagement Intensity in the Wild Using Temporal Convolutional Network. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*. ACM.

[16] Michel F Valstar, Enrique Sánchez-Lozano, Jeffrey F Cohn, László A Jeni, Jeffrey M Girard, Zheng Zhang, Lijun Yin, and Maja Pantic. 2017. Fera 2017-addressing head pose in the third facial expression recognition and analysis challenge. In *Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*. IEEE, 839–847.

[17] Valentin Vielzeuf, Corentin Kervadec, Alexis Lechervy, Stephane Pateux, and Frederic Jurie. 2018. An Occam's Razor View on Learning Audiovisual Emotion Recognition with Small Training Sets. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*. ACM.

[18] Kai Wang, Yu Qiao, Jianfei Yang, Xiaojiang Peng, Xiaoxing Zeng, Debin Meng, and Kaipeng Zhang. 2018. Cascade Attention Networks For Group Emotion Recognition with Face, Body and Image Cues. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*. ACM.

[19] 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 (CVPR)*. 532–539.

[20] Jianfei Yang, Kai Wang, Xiaojiang Peng, and Yu Qiao. 2018. Deep Recurrent Multi-instance Learning with Spatio-temporal Features for Engagement Intensity Prediction. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*. ACM.

[21] Guoying Zhao and Matti Pietikainen. 2007. Dynamic Texture Recognition Using Local Binary Patterns with an Application to Facial Expressions. *IEEE Transaction on Pattern Analysis and Machine Intelligence* 29, 6 (2007), 915–928.

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