From individual to group-level emotion recognition: EmotiW 5.0

Conference Paper · November 2017

DOI: 10.1145/3136755.3143004

Abhinav Dhall
Indian Institute of Technology Ropar

Roland Goecke
University of Canberra

Shreya Ghosh
Curtin University

Jyoti Joshi
University of Canberra

From Individual to Group-Level Emotion Recognition: EmotiW 5.0

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

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

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

Jyoti Joshi
University of Waterloo
Canada
jyoti.joshi@uwaterloo.ca

Jesse Hoey
University of Waterloo
Canada
jhoey@cs.uwaterloo.ca

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

ABSTRACT

Research in automatic affect recognition has come a long way. This paper describes the fifth Emotion Recognition in the Wild (EmotiW) challenge 2017. EmotiW aims at providing a common benchmarking platform for researchers working on different aspects of affective computing. This year there are two sub-challenges: a) Audio-video emotion recognition and b) group-level emotion recognition. These challenges are based on the acted facial expressions in the wild and group affect databases, respectively. The particular focus of the challenge is to evaluate method in 'in the wild' settings. 'In the wild' here is used to describe the various environments represented in the images and videos, which represent real-world (not lab like) scenarios. The baseline, data, protocol of the two challenges and the challenge participation are discussed in detail in this paper.

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

ACM Reference Format:
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 19th ACM International Conference on Multimodal Interaction (ICMI'17).* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3136755.3143004

ICMI'17, November 13–17, 2017, Glasgow, UK

ACM ISBN 978-1-4503-5543-8/17/11…$15.00
https://doi.org/10.1145/3136755.3143004

1 INTRODUCTION

The Fifth Emotion Recognition in the Wild (EmotiW)[1] challenge is a grand challenge in the ACM International Conference on Multimodal Interaction 2017. The aim of this challenge is to provide a platform for researchers to evaluate their affect recognition methods on 'in the wild' data. 'In the wild' here refers to the real-life, uncontrolled conditions such as diverse backgrounds (indoor/outdoor), illumination conditions, head motion occlusion, multiple people in an image and spontaneous expression etc. [2]. EmotiW 2017 comprises of two sub-challenges: a) audio-video emotion recognition (VReco) and b) group-level emotion recognition (GReco).

EmotiW 2017 is the fifth iteration of the challenge. Over the past four years the challenge tasks have been audio-video emotion recognition, image-level facial expression recognition and group-level happiness intensity estimation. Table 1 summarizes the tasks in the EmotiW challenges. The task during the first EmotiW challenge at ACM ICMI 2013 was the VReco challenge [4]. The Acted Facial Expressions in the Wild (AFEW) 3.0 dataset was used as the dataset for evaluation in this challenge. Some interesting methods have been proposed in VReco sub-challenge [10] [15] [21]. EmotiW 2014 [3] continued with the VReco task albeit with more data and cleaner labels.

In EmotiW 2015 [9], we introduced the static image based facial expression recognition sub-challenge. The task was to predict the facial expression of a subject in a given image. The universal emotion categories are *Angry*, *Disgust*, *Fear*, *Happy*,*Neutral*, *Sad* and *Surprise*. A total of 22 teams participated in the new sub-challenge and the Static Facial Expressions in the Wild (SFEW) 2.0 database [5] was used. SFEW is extracted from AFEW database by computing key-frames based on facial points clustering. The top performing methods were based on deep learning techniques [13] [22]. Researchers also proposed transfer learning based methods for dealing with smaller sized training data [17].

The EmotiW 2016 [2] challenge was organized at the ACM ICMI 2016, Tokyo. There were two major changes as compared to the earlier years. The first change was that in the VReco sub-challenge, the AFEW 6.0 data was collected from movies and reality TV shows. The rationale was to add *more spontaneous* data to AFEW 6.0. Further, automatic group-level emotion recognition, was introduced. This sub-challenge was based on the HaPpy PeoplE Images (HAPPEI)

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

database [1]. The GReco sub-challenge is motivated by the tremendous increase in the number of images and videos, which are shared on social networking websites such as Facebook, Instagram etc. From the perspective of automatic affect recognition, this poses a new challenge due to the presence of multiple people in an image [2].

In our earlier work [1], we conducted a user study to understand the salient attributes, which characterizes the perception of users towards the mood of a group. Based on that understanding, we can broadly categorize the approaches as top-down and bottom-up. Top-down here means the effect of the scene, background and group structure on the perceived mood of a group. Bottom-up approaches focus on group members individually. The task is to understand the contribution of each group member towards the perception of his/her group's mood. This is based on various person level attributes such as attractiveness, age, gender, clothes, spontaneous expression etc. For details about the survey, please refer to the paper [1]. A latent Dirichlet allocation based attribute augmented model was proposed to infer the intensity of happiness at a group-level. The data was extracted from Flickr based on keyword search. This data (HAPPEI [7]) was the basis of the GReco sub-challenge during the EmotiW 2016.

An interesting experiment was conducted at MIT [11], where four cameras were installed at various locations and the group-level emotion was computed as the average score of presence of smiles among the passerbys. In GReco sub-challenge 2016, an ensemble of features from LSTM were used for happiness intensity prediction by Li et al. [14]. This was the top performing method in the EmotiW 2016 challenge. Further, it was also observed that face-level geometrical features though simple are effective for inferring happiness intensity Vonikakis et al. [18]. In an interesting work, Mou et al. [16] proposed fusion of both body and face level features for the task of group-level emotion in images using the Valence and arousal emotion annotations. Huang et al. [12], proposed local binary pattern based features and used conditional random field based technique for inferring the perceived group-level happiness intensity.

Figure 1: The figure shows samples from three classes in the GReco sub-challenge [8].

This year in EmotiW 2017, there are two sub-challenges: VReco and GReco. During EmotiW 2016, GReco was based on intensity estimation. This year, the task is to classify the perceived mood of a group into three categories: *Positive*, *Neutral* & *Negative*. Figure 1 shows sample images in the GReco task with its classes. The task in VReco remains the same as last year, the different is that sitcom TV data has also been used. Table 1 summarizes the five EmotiW challenges on the basis of the problem tasks and the databases involved.

2 DATA

The data used in the two sub-challenges is discussed below:

VReco - This sub-challenge is the continuation of the audio-video based bimodal emotion recognition task, which has been part of the last four EmotiW challenges. A third modality in the form of meta-data was also shared with the challenge participants. The meta-data consisted of the subject age, gender and identity. The data in this challenge is created by a simple sentiment analysis of the closed captions in movies and TV series [6]. A dictionary of words related to affect were searched in the closed captions and were used for generating weak emotion labels. These labels were then modified by the labellers, if required. An obvious advantage of using the closed captions for data collection is time saving as finding affect-wise salient video segments in a long movie is a laborious task. Another advantage is that we get access to rich meta-data from internet repositories such as IMDB[2]. This can be used for adding context information in prediction techniques. The AFEW data is divided into three data partitions: *Train* (773 samples), *Val* (383 samples) and *Test* (653 samples). Similar to EmotiW 2016, the TV show data has been added to the Test set only. Last year reality TV data was added and this year we also added data from sitcom TV series. This adds to the variety in the environments in which the data has been recorded. The Train and Val set data this year is same as that in the EmotiW 2016 VReco sub-challenge. Also, it is to be noted that the three data partitions are subject and movie/TV source 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 2 discusses the details about the video samples in the database. 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 (these details are same as for the VReco sub-challenge in EmotiW 2016 [2]).

[2] www.imdb.com

Table 1: The summary of the sub-challenges in the EmotiW challenge series. Here VReco - Audio-Video emotion recognition, SReco - frame-level Static facial expression recognition, GReco - Group-based emotion recognition.

EmotiW	Challenge 1	Challenge 2	Comments
Sydney 2013 [4]	VReco	-	GReco - AFEW consists of movie data.
Istanbul 2014 [3]	VReco	-	GReco - AFEW consists of movie data.
Seattle 2015 [9]	VReco	SReco	VReco - AFEW consists of movie data. SReco - SFEW consists of frames from AFEW.
Tokyo 2016 [2]	VReco	GReco	VReco - AFEW consists of movie & reality TV data [7] GReco - Group-level Happiness intensity - HAPPEI database [1]
Glasgow 2017	VReco	GReco	VReco - AFEW - movie + TV GReco - Three class categorical problem - GAF database [8]

GReco - The Group Affect Database 2.0 [8] forms the basis of this sub-challenge. The images are sourced from Google images and Flickr on the basis of keyword search. The keywords are based on different events both happy and sad (eg: festival, party, silent protest, violence etc.). Similar to the VReco sub-challenge the database is divided into three data partitions i.e. *Train* (3630 samples), *Val* (2068 samples) and *Test* (773 samples). The classes to be predicted are *Positive*, *Neutral* and *Negative*.

3 BASELINE EXPERIMENTS

3.1 VReco Sub-challenge

The baseline pipeline for VReco is similar to the one in EmotiW 2016. For localizing the face, we used the pre-trained face models shared in the Mixture of PartS (MoPS) library [24]. The output from MoPS is used to initialize of the Intraface library tracker [20]. Affine warping is applied for aligning the faces into a 128×128 grid. The Local Binary Pattern-Three Orthogonal Planes (LBP-TOP) [23] is computed on the aligned frames. Each frame is divided spatially into non-overlapping 4×4 blocks. The LBP-TOP is a standard texture based feature, which has been extensively used for face-based affect classification [4] [3]. The LBP-TOP feature from each block are concatenated to create one feature vector. Non-linear Chi-square kernel based SVM is trained for emotion classification (*Anger*, *Disgust*, *Fear*, *Happiness*, *Neutral*, *Sadness* and *Surprise*). The video only baseline system achieves 38.81% and 41.07% classification accuracy for the *Val* and *Test* sets, respectively. Please note that the evaluation metric unweighted classification accuracy. The list of movies used in both VReco is mentioned in Section 6.

Table 2: Attributes of the subset of the AFEW 7.0 database used in the EmotiW 2017 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	1809
Video format	AVI
Audio format	WAV

3.2 GReco Sub-challenge

For the GReco sub-challenge in EmotiW 2016, we had extracted the CENsus TRansform hISTogram (CENTRIST) descriptor [19] from the images. The CENTRIST feature descriptor is computed by applying a local binary pattern like Census transform it is rich texture descriptor. As the descriptor is computed on an image, it captures both the top-down and bottom-up attributes as discussed in [1]. CENTRIST is based on the Census tranform, which is similar to the local binary pattern. For encoding the structural information an image is divided into 4×4 non-overlapping blocks and the CENTRIST is computed block-wise. Support Vector Regression with a non-linear Chi-square kernel was used to train the classification model. Unweighted classification accuracy is used as the evaluation metric. The simple method achieved 52.97% on the *Val* set and 53.62% on the *Test* set.

4 CHALLENGE RESULTS

Similar to the EmotiW 2016 challenge, this year we received over 100 challenge registrations. In the VReco sub-challenge 25 teams submitted labels generated by their methods for evaluation. A total of 14 teams submitted the labels in the GReco sub-challenge. At the end of the *Test* evaluation phase, 22 teams submitted the papers discussing their methods. Based on the relative performance and manuscript reviews 13 papers were accepted for publications. The team's with Top 3 performing methods have been asked to share code/library with us for evaluation. This phase is still going on and the current leader-board for GReco and VReco sub-challenges is mentioned in Table 3 and Table 4, respectively.

5 CONCLUSION

The Fifth Emotion Recognition in the Wild is a grand challenge in the ACM International Conference on Multimodal Interaction 2017, Glasgow. There are two sub-challenges in EmotiW 2017. The first sub-challenge deals with the task of emotion recognition into the universal emotion categories from samples containing audio-video data. The second sub-challenge contains images collected from the internet. The task is to classify the collective emotion at the group-level in the images. The challenge received over 100 registrations. The top performing methods in both the sub-challenges are based on the deep learning methodologies. A peer review process was followed to select high quality papers describing the high performing

Table 3: The Table shows the Leader-board for the VReco sub-challenge. Please note that this is not the final ranking as the final code evaluation for the top performing team was going on by the time of the camera ready submission of this paper was made. For the final ranking, please refer to the challenge website.

Team Name	Classification Accuracy (%)
ILC	60.34
NTechLab	60.03
BUPT	59.72
OL-UC	58.80
RUC-IBM	58.50
I2R	57.27
INHA	57.12
SJTU	55.28
LC	52.58
Tsinghua	52.53
NLPR	51.14
UD-GPB	51.15
EURECOM-UNIMORE	49.92
BNU	49.77
HAHA	49.46
SIAT	48.69
FACEALL_BUPT	42.26
CNU	41.80
Baseline	41.07
SRI-CVT	40.27
Knowledge-technology	40.12
NEU	39.66
OB	39.35
Smartear_abalia	35.37
Seres	33.23
Shenzhen University	32.77

methods in the two sub-challenges. In the next EmotiW challenge, we plan to extend the data and make the emotion labels finer.

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.

Table 4: The Table shows the Leader-board for the GReco sub-challenge. Please note that this is not the final ranking as the final code evaluation for the top performing team was going on by the time of the camera ready submission of this paper was made. For the final ranking, please refer to the challenge website.

Team Name	Classification Accuracy (%)
SIAT	80.89
UD-GPB	80.61
BNU	79.78
AMD	78.53
AmritaEEE	75.07
Nanjing_university	74.79
Omega_3	70.64
CVI_SZ	69.67
THAPAR UNIVERSITY	66.34
CRNS	64.68
UoN	63.43
jci_garage	57.89
Baseline	53.62

ACKNOWLEDGEMENT

We are grateful to the ACM ICMI 2017 chairs, the EmotiW program committee members, the reviewers and the challenge participants.

REFERENCES

[1] Abhinav Dhall, Roland Goecke, and Tom Gedeon. 2015. Automatic group happiness intensity analysis. *IEEE Transactions on Affective Computing* 6, 1 (2015), 13–26.

[2] 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.

[3] Abhinav Dhall, Roland Goecke, Jyoti Joshi, Karan Sikka, and Tom Gedeon. 2014. Emotion Recognition In The Wild Challenge 2014: Baseline, Data and Protocol.

In *Proceedings of the ACM on International Conference on Multimodal Interaction (ICMI)*.
[4] Abhinav Dhall, Roland Goecke, Jyoti Joshi, Michael Wagner, and Tom Gedeon. 2013. Emotion Recognition In The Wild Challenge 2013. In *Proceedings of the ACM on International Conference on Multimodal Interaction (ICMI)*. 509–516.
[5] Abhinav Dhall, Roland Goecke, Simon Lucey, and Tom Gedeon. 2011. 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*. 2106–2112.
[6] 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.
[7] Abhinav Dhall, Jyoti Joshi, Ibrahim Radwan, and Roland Goecke. 2012. Finding Happiest Moments in a Social Context. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*. 613–626.
[8] 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.
[9] Abhinav Dhall, OV Ramana Murthy, Roland Goecke, Jyoti Joshi, and Tom Gedeon. 2015. Video and image based emotion recognition challenges in the wild: Emotiw 2015. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM, 423–426.
[10] Samira Ebrahimi, Chris Pal, Xavier Bouthillier, Pierre Froumenty, SÃĬbastien Jean, Kishore Reddy Konda, Pascal Vincent, Aaron Courville, and Yoshua Bengio. 2013. Combining Modality Specific Deep Neural Networks for Emotion Recognition in Video. In *Proceedings of the ACM on International Conference on Multimodal Interaction (ICMI)*. 543–550.
[11] Javier Hernandez, Mohammed Ehsan Hoque, Will Drevo, and Rosalind W Picard. 2012. Mood meter: counting smiles in the wild. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*. 301–310.
[12] Xiaohua Huang, Abhinav Dhall, Guoying Zhao, Roland Goecke, and Matti Pietikäinen. 2015. 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)*. 34–1.
[13] Bo-Kyeong Kim, Hwaran Lee, Jihyeon Roh, and Soo-Young Lee. 2015. Hierarchical committee of deep cnns with exponentially-weighted decision fusion for static facial expression recognition. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM, 427–434.
[14] Jianshu Li, Sujoy Roy, Jiashi Feng, and Terence Sim. 2016. Happiness Level Prediction with Sequential Inputs via Multiple Regressions. In *Proceedings of the ACM on International Conference on Multimodal Interaction (ICMI)*.
[15] Mengyi Liu, Ruiping Wang, Shaoxin Li, Shiguang Shan, Zhiwu Huang, and Xilin Chen. 2014. Combining multiple kernel methods on Riemannian manifold for emotion recognition in the wild. In *Proceedings of the 16th International Conference on Multimodal Interaction*. ACM, 494–501.
[16] Wenxuan Mou, Oya Celiktutan, and Hatice Gunes. 2015. 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)*, Vol. 5. IEEE, 1–6.
[17] Hong-Wei Ng, Viet Dung Nguyen, Vassilios Vonikakis, and Stefan Winkler. 2015. Deep learning for emotion recognition on small datasets using transfer learning. In *Proceedings of the 2015 ACM on international conference on multimodal interaction*. ACM, 443–449.
[18] Vonikakis Vonikakis, Yasin Yazici, Viet Dung Nguyen, and Stefan Winkler. 2016. Group happiness assessment using geometric features and dataset balancing. In *Proceedings of the ACM on International Conference on Multimodal Interaction (ICMI)*.
[19] Jianxin Wu and Jim M Rehg. 2011. CENTRIST: A visual descriptor for scene categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33, 8 (2011), 1489–1501.
[20] 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.
[21] Jingwei Yan, Wenming Zheng, Zhen Cui, Chuangao Tang, Tong Zhang, and Yuan Zong. 2016. Multi-Clue Fusion for Emotion Recognition in the Wild. In *Proceedings of the ACM on International Conference on Multimodal Interaction (ICMI)*.
[22] Zhiding Yu and Cha Zhang. 2015. Image based static facial expression recognition with multiple deep network learning. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM, 435–442.
[23] 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.
[24] 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.