Video and Image based Emotion Recognition Challenges in the Wild: EmotiW 2015

Conference Paper · November 2015

DOI: 10.1145/2818346.2829994

5 authors, including:

Abhinav Dhall
Indian Institute of Technology Ropar
88 PUBLICATIONS **2,209** CITATIONS
SEE PROFILE

O.V. Ramana Murthy
Amrita Vishwa Vidyapeetham
39 PUBLICATIONS **608** CITATIONS
SEE PROFILE

Roland Goecke
University of Canberra
167 PUBLICATIONS **3,767** CITATIONS
SEE PROFILE

Jyoti Joshi
University of Canberra
20 PUBLICATIONS **1,014** CITATIONS
SEE PROFILE

Video and Image based Emotion Recognition Challenges in the Wild: EmotiW 2015

Abhinav Dhall
Vision & Sensing Group
University of Canberra
abhinav.dhall@canberra.edu.au

O. V. Ramana Murthy
Vision & Sensing Group
University of Canberra
o.v.ramanamurthy@ieee.org

Roland Goecke
Vision & Sensing Group
University of Canberra/
Australian National University
roland.goecke@ieee.org

Jyoti Joshi
Vision & Sensing Group
University of Canberra
jyoti.joshi@canberra.edu.au

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

ABSTRACT

The third Emotion Recognition in the Wild (EmotiW) challenge 2015 consists of an audio-video based emotion and static image based facial expression recognition sub-challenges, which mimics real-world conditions. The two sub-challenges are based on the Acted Facial Expression in the Wild (AFEW) 5.0 and the Static Facial Expression in the Wild (SFEW) 2.0 databases, respectively. The paper describes the data, baseline method, challenge protocol and the challenge results. A total of 12 and 17 teams participated in the video based emotion and image based expression sub-challenges, respectively.

Categories and Subject Descriptors

I.6.3 [**Pattern Recognition**]: Applications; H.2.8 [**Database Applications**]: Image Databases; I.4.m [**IMAGE PROCESSING AND COMPUTER VISION**]: Miscellaneous

Keywords

Audio-video data corpus; Facial expression challenge; Affect analysis in the wild

1. INTRODUCTION

With the advancement in the computational technology and the exponential jump in the amount of data available over the internet, it is useful to develop methods capable of analysing affect in the wild. 'In the wild' here means uncontrolled conditions. Research in automatic affect analysis has been mostly done in controlled environment (such as lab like conditions). In order to improve the current methods for them to perform well in uncontrolled conditions, there is a need of labelled data, which can represent the 'in the wild' settings. The Emotion Recognition in the Wild (EmotiW) challenge aims at providing a platform for researchers to benchmark the performance of their methods on 'in the wild' data. EmotiW 2015 comprises of two sub-challenges – a) Video based emotion Recognition (VReco); b) Static Image based facial expression Recognition (SReco).

There are several challenges in the affect recognition community, which target different challenges in affect recognition. The Facial Expression Recognition challenge and Analysis (FERA) [14] targets facial expression recognition on universal emotion classes on data created by actors in lab conditions. Audio Video Emotion Challenge (AVEC) [11] [12] focusses on depression intensity and continuous affect intensity estimation. On the other hand, EmotiW targets the efforts required towards affect analysis in the wild.

During the first EmotiW challenge [3], which was organised at the ACM International Conference on Multimodal Interaction (ICMI) 2013, Sydney, a total of 27 teams registered for the challenge and 9 teams submitted the test labels. The database for the challenge was AFEW 3.0 [4]. The data was divided into three sets: *Train*, *Val* and *Test*. Ebrahimi et al. [6], proposed a deep learning based emotion recognition method and their system was the challenge winner with classification accuracy of 41.02%. Sikka et al. [13] proposed a multiple kernel learning based technique, which was the runner-up method during EmotiW 2013. From the classification accuracy of the proposed methods in EmotiW 2013, it can be observed that automatic emotion recognition in the wild is not mature enough to work well in uncontrolled settings and there is a lot of scope for future research in this direction. The second EmotiW challenge [2] was organised as part of ACM ICMI 2014, Istanbul, had participation from 9 teams. Mengyi et al. [9]'s method based on manifold learning and convolutional neural networks performed best during EmotiW 2014. Figure 1 compares the classification accuracy of methods proposed in the EmotiW 2014 challenge. The performance of the proposed techniques improved w.r.t. the entries in 2013. However, it is still far away from practical use in real-world conditions. Therefore, we continue with the challenge series. This year for EmotiW 2015 is being organised as part of ACM ICMI 2015, Seattle. A new sub-challenge has been introduced in EmotiW 2015. The static image based facial expression recognition sub-challenge aims to provide a platform for researchers, who

ICMI '15, November 09-13, 2015, Seattle, WA, USA

DOI: http://dx.doi.org/10.1145/2818346.2829994.

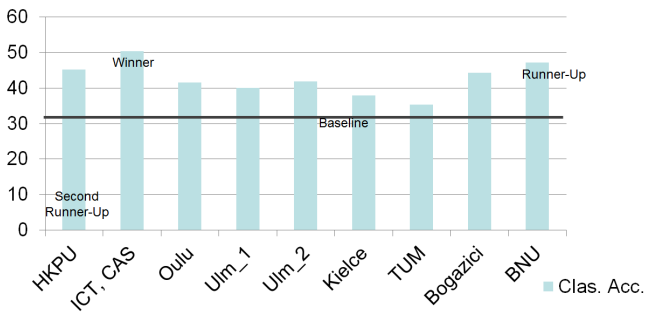

Figure 1: The graph compares the classification accuracy performance of participants in EmotiW 2014 challenge.

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	1645
Video format	AVI
Audio format	WAV

Table 1: Attributes of AFEW 5.0 database.

work on a single image as a sample for facial expression recognition.

2. DATA

EmotiW 2015 has two sub-challenges and there are two datasets as described below:

VReco - For the video based emotion recognition sub-challenge, AFEW 5.0 database is used. 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 [4]. For EmotiW 2015, the database is divided into three sub-sets: *Train* (723 samples), *Val* (383 samples) and *Test* (539 samples). These sets are the extended versions of the EmotiW 2014 [2] sets. The current version of the database, AFEW 5.0 is available at http://cs.anu.edu.au/few contains two labelled sets. sub-challenge 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 available in July 2015. 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

Low Level Descriptors (LLD)	
Energy/Spectral LLD	PCM Loudness MFCC [0-14] log Mel Frequency Band [0-7] Line Spectral Pairs (LSP) frequency [0-7] F0 F0 Envelope
Voicing related LLD	Voicing Prob. Jitter Local Jitter consecutive frame pairs Shimmer Local

Table 2: Audio feature set - 38 (34 + 4) low-level descriptors.

will be taken into account for the overall Grand Challenge results.

SReco - SFEW 2.0 forms the basis of the Image based Static Facial Expression Recognition sub-challenge. SFEW is created from AFEW 5.0 using key-frame extraction method. The video clips as discussed above are created using a semi-automatic technique, we added another step to it. Facial points extracted a video sample in VReco are clustered. The frames closest to the cluster centres are chosen as the key-frames. K-means clustering with number of clusters = 6 was used. Similar to AFEW 5.0, SFEW 2.0 has been divided into three sets: Train (880 samples), Val (383 samples) and Test (372 samples). The list of movies used in both VReco and SReco is mentioned in Section 6.

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 [17] are applied for face detection and initialisation of the Intraface tracking libray [15]. The fiducial points generated by Intraface are used for aligning the face. The face size is set to 128×128. Post aligning Local Binary Pattern-Three Orthogonal Planes (LBP-TOP) [16] 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 [3] [2]. The LBP-TOP feature from each block are concatenated to create one feature vector. 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 36.08% and 39.33% classification accuracy for the *Val* and *Test* sets, respectively. Please note that these are unweighed accuracies.

Functionals
Arithmetic Mean
standard deviation
skewness, kurtosis
quartiles, quartile ranges
percentile 1%, 99%
percentile range
Position max./min
up-level time 75/90
linear regression coeff.
linear regression error(quadratic/absolute)

Table 3: Set of functionals applied to LLD.

In this challenge, a set of audio features similar to the features employed in AVEC 2011 [11] motivated from the INTERSPEECH 2010 Paralinguistic challenge (1582 features) [10] are also shared with the participants. The features are extracted using the open-source Emotion and Affect Recognition (openEAR) [7] toolkit backend openSMILE [8]. The feature set consists of 34 energy & spectral related low-level descriptors (LLD) $\times$ 21 functionals, 4 voicing related LLD $\times$ 19 functionals, 34 delta coefficients of energy & spectral LLD $\times$ 21 functionals, 4 delta coefficients of the voicing related LLD $\times$ 19 functionals and 2 voiced/unvoiced durational features. Table 2 and 3 describe the details of LLD features and functionals, respectively.

3.2 SReco sub-challenge

The MoPS [17] based face and facial parts detector is applied to SFEW images to compute the baseline for SReco. The fiducial points are aligned using affined warp and the face size is set to 128 $\times$ 128. Following our earlier work [1], two feature descriptors are computed. Pyramid of Histogram of Gradients (PHOG) is applied to the aligned face. The parameters are: range = [0-360], pyramid levels = 3 and number of bins = 16. The aligned faces are divided into 4 $\times$ 4 non-overlapping blocks and PHOG is computed from each block. The PHOG histograms from each block are further concatenated. Furthermore, Local Phase Quantisation (LPQ) is computed. LPQ is robust to blur and illumination changes and is computed on the aligned face as a whole. PHOG and LPQ are concatenated at feature level and a non-linear SVM is used. The classification accuracy for the *Val* and *Test* set are 35.93% and 39.13%, respectively.

In total 75 teams, registered for the challenge and 22 papers were submitted. 13 papers were accepted for publication with 6 papers accepted as oral presentations and others as posters. Prof. Marian S. Bartlett (University of California San Diego and Emotient Inc) and Prof. Ira Kemelmacher-Shlizerman (University of Washington) will be presenting the keynote talks at the EmotiW event at ICMI, Seattle.

Figure 2 shows a graph, which compares the classification accuracy of the 17 participating teams w.r.t each other and the baseline in SReco sub-challenge. The winning entries are from the teams of KAIST_CNL, NERV and ADSC. Figure 3 shows a graph, which compares the classification accuracy of the 12 participating teams w.r.t each other and the baseline. The winning entries are from the teams of ILC, Bogazici and Poly-Goethe. It is interesting to note that a large number of the proposed methods used deep learning based techniques. The performance of the methods is limited by – a) the amount of labelled data; b) performance limitations of the current state-of-art face and facial parts detectors. Some teams used extra data for training their deep learning pipelines. For the VReco sub-challenge, the facial features should be able to model the temporal dynamics of emotion. Some frames contain multiple subjects, this in itself is one of the problems, which need to be further looked into. Group-level emotion recognition [5] is a new research area and will be worthwhile to explore in future, in order to improve the automatic emotion recognition methods for them to work on multiple people in a scene.

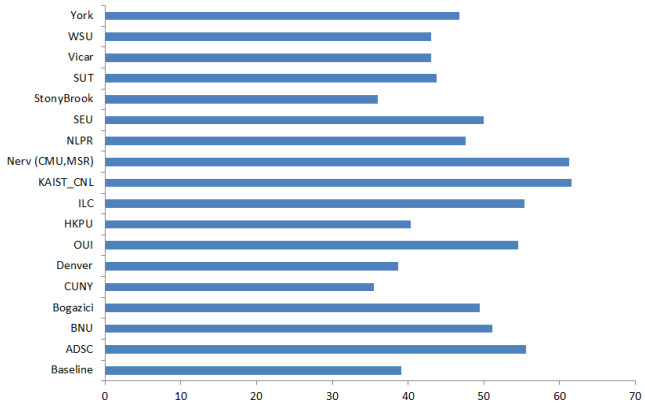

Figure 2: The graph compares the classification accuracy performance of participants in the SReco sub-challenge.

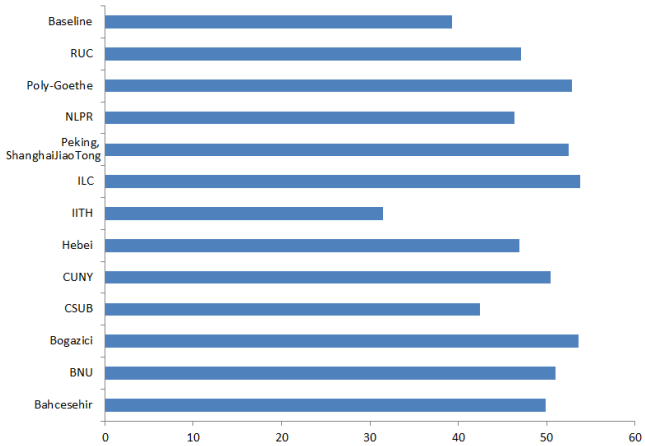

Figure 3: The graph compares the classification accuracy performance of participants in the VReco sub-challenge.

4. CONCLUSION

The third Emotion Recognition in the Wild 2015 challenge provides a platform for researchers to benchmark and compete with their emotion recognition method on the Acted Facial Expressions in the Wild database 5.0 and single image based facial expression recognition methods on the Static Facial Expression in the Wild database 2.0. Emotion recognition in the wild is a challenging problem due to diversity in scenes in the form of head pose, illumination, occlusion and background noise. This year's challenge carry forwards, the platform started by the second EmotiW challenge and adds a new sub-challenge. This paper describes the baselines, data and protocols for the two sub-challenges. The results show that affect analysis in the wild is still an open-problem and there are various sub-problems, which need to be researched upon.

5. ACKNOWLEDGEMENT

We are grateful to the ICMI'15 chairs and EmotiW program committee members. Abhinav Dhall and Roland Goecke

are supported by the Australian Research Council Discovery project grant ARCDP130101094.

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, 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, Ghoshtship, Girl with a Pearl Earring, Gone In Sixty Seconds, 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 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, You're Next, You've Got Mail

7. REFERENCES

[1] A. Dhall, A. Asthana, R. Goecke, and T. Gedeon. Emotion recognition using PHOG and LPQ features. In *Proceedings of the IEEE Conference Automatic Faces & Gesture Recognition workshop FERA*, pages 878–883, 2011.

[2] 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.

[3] 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.

[4] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon. Collecting large, richly annotated facial-expression databases from movies. *IEEE Multimedia*, 19(3):0034, 2012.

[5] 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.

[6] S. Ebrahimi, C. Pal, X. Bouthillier, P. Froumenty, S. Jean, K. R. Konda, P. Vincent, A. Courville, and Y. Bengio. Combining modality specific deep neural networks for emotion recognition in video. In *Proceedings of the ACM on International Conference on Multimodal Interaction (ICMI)*, pages 543–550, 2013.

[7] F. Eyben, M. Wollmer, and B. Schuller. OpenearâĂŤintroducing the munich open-source emotion and affect recognition toolkit. In *Proceedings of the International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–6, 2009.

[8] F. Eyben, M. Wöllmer, and B. Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the ACM International Conference on Multimedia (MM)*, pages 1459–1462, 2010.

[9] 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.

[10] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. A. Müller, and S. S. Narayanan. The interspeech 2010 paralinguistic challenge. In *INTERSPEECH*, pages 2794–2797, 2010.

[11] B. Schuller, M. Valstar, F. Eyben, G. McKeown, R. Cowie, and M. Pantic. Avec 2011–the first international audio/visual emotion challenge. In *Proceedings of the International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 415–424, 2011.

[12] B. Schuller, M. Valster, F. Eyben, R. Cowie, and M. Pantic. Avec 2012: the continuous audio/visual emotion challenge. In *Proceedings of the ACM on International Conference on Multimodal Interaction (ICMI)*, pages 449–456, 2012.

[13] K. Sikka, K. Dykstra, S. Sathyanarayana, G. Littlewort, and M. Bartlett. Multiple kernel learning for emotion recognition in the wild. In *Proceedings of the ACM on International Conference on Multimodal Interaction (ICMI)*, pages 517–524, 2013.

[14] M. F. Valstar, M. Mehu, B. Jiang, M. Pantic, and K. Scherer. Meta-analysis of the first facial expression recognition challenge. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 42(4):966–979, 2012.

[15] 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.

[16] 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.

[17] 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.