Emotion Recognition In The Wild Challenge 2014: Baseline, Data and Protocol

Conference Paper · November 2014

DOI: 10.1145/2663204.2666275

5 authors, including:

Abhinav Dhall
Indian Institute of Technology Ropar
112 PUBLICATIONS **3,291** CITATIONS
SEE PROFILE

Roland Goecke
University of Canberra
182 PUBLICATIONS **5,276** CITATIONS
SEE PROFILE

Jyoti Joshi
University of Canberra
22 PUBLICATIONS **1,386** CITATIONS
SEE PROFILE

Karan Sikka
SRI International
46 PUBLICATIONS **1,252** CITATIONS
SEE PROFILE

Emotion Recognition In The Wild Challenge 2014: Baseline, Data and Protocol

Abhinav Dhall
Vision & Sensing Group
University of Canberra/
Australian National University
abhinav.dhall@anu.edu.au

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

Karan Sikka
Machine Perception Lab
Univesity of California
San Diego
karan.sikka@ucsd.edu

Tom Gedeon
Res. School of Computer
Science
Australian National University
tom.gedeon@anu.edu.au

ABSTRACT

The Second Emotion Recognition In The Wild Challenge (EmotiW) 2014 consists of an audio-video based emotion classification challenge, which mimics the real-world conditions. Traditionally, emotion recognition has been performed on data captured in constrained lab-controlled like environment. While this data was a good starting point, such lab controlled data poorly represents the environment and conditions faced in real-world situations. With the exponential increase in the number of video clips being uploaded online, it is worthwhile to explore the performance of emotion recognition methods that work 'in the wild'. The goal of this Grand Challenge is to carry forward the common platform defined during EmotiW 2013, for evaluation of emotion recognition methods in real-world conditions. The database in the 2014 challenge is the Acted Facial Expression In Wild (AFEW) 4.0, which has been collected from movies showing close-to-real-world conditions. The paper describes the data partitions, the baseline method and the experimental protocol.

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

General Terms

Experimentation, emotion recognition

ICMI '14, November 12 - 16 2014, Istanbul, Turkey

http://dx.doi.org/10.1145/2663204.2666275.

Keywords

Audio-video data corpus; Emotion recognition in the wild; EmotiW challenge.

1. INTRODUCTION

This paper describes the baseline, data and experimental protocol for the Second Emotion Recognition In The Wild Challenge (EmotiW) 2014. The recent advancement of social media has given users a platform to socially engage and interact with a larger population. Millions of images and videos are being uploaded every day by users on the web from different events and social gatherings. There is an increasing interest in designing systems capable of understanding human manifestations of emotional attributes and affective displays. Inferring the affective state of the users in images and videos, that are captured under real-world conditions, robust methods capable of performing facial expression analysis 'in the wild' are required. Here, the term **'in the wild'** signifies variability in environments/scenes and backgrounds, illumination conditions, head pose, occlusion *etc.* Automatic emotion analysis has made a significant progress over the last two decades. However, most of these frameworks have been restricted and evaluated on data collected in controlled laboratory settings with frontal faces, perfect illumination and posed expressions. To overcome this limitation, EmotiW 2014 provides a platform for researchers to create, extend and test their methods on a common 'in the wild' benchmarked data.

EmotiW 2014 follows the guidelines of the first EmotiW 2013 challenge and is based on the Acted Facial Expressions in the Wild (AFEW) database [5]. Generally, face related databases have been collected in controlled environments (ambient lighting, simple and consistent background). However, for adapting emotion recognition approaches, that work well on controlled data, to real-world data, databases representing varied scenarios are required. Recently, few databases (e.g. AFEW [5], GENKI [21], Static Facial Expressions in the Wild [4] and Happy People Images [6] etc.), which represent real-world settings have been released.

There are several challenges to be tackled for emotion recognition 'in the wild'. Consider an illustrative example of categorization i.e. assigning an emotion label to a video clip

of a subject(s) protesting at the Tahrir square in Egypt during the 2011 protests. In order to learn an automatic system for inferring the emotion label, labeled data containing video clips representing different emotions in diverse settings are required. A number of standard databases such as the Cohn-Kanade (CK) [13], Multi-PIE [11], FEEDTUM [20] and RU-FACS [1], exist and include both static and dynamic data of subjects displaying a fixed set of expressions. However, all of these databases contain samples with posed/spontaneous expressions under lab-controlled conditions. The field of facial expression recognition should ideally collect spontaneous data in uncontrolled settings to tackle challenges of real-world data. However, collecting spontaneous data in real-world conditions is a tedious task. Therefore, new methods which can speed up the process of creating databases representing real-world conditions are required.

Once the data is available, the next challenge is the detection of face and its constituent parts, followed by Head Pose Normalisation (HPN). It is obvious that a subject in an uncontrolled setting, such as the Tahrir square video clip example, may freely move his/her head leading of out-of-plane head movements. Such a setting gives rise to another challenge and adversely effects the performance of face and facial parts detections, that are required by HPN. It is also known that alignment in case of non-frontal faces is a non-trivial task and can further introduce noise/error during the feature extraction step. During spontaneous expressions, subjects also move their arms and hands as part of the non-verbal communication, leading to the problem of occlusion. Occlusion needs to be then detected and localised for finding accurate fiducial points. The problem becomes more severe with multiple subjects in the scene. Though EmotiW 2013 & 2014 deal with a single subject based emotion recognition, a future challenge for emotion recognition methods is to handle multiple subjects in a scene (e.g. [6]).

From the audio modality perspective, one of the primary challenge is modeling background noise. In case of databases such as AFEW, background noise can be either music or sound, that may describe the context of the scene. Thus it needs to be investigated whether such background is correlated with the underlying emotion. Other related issues for speech modeling are (1) exploring approaches for dealing with variable length samples, (2) selecting robust features to describe the speech signal, among others. Emotion recognition in the wild is a problem involving multiple modalities and thus information cues can also be extracted from non-verbal body gesture and scenes. Thus exploring techniques for combining these multiple modalities for facial expression in the wild is a challenge in itself.

From choosing a classifier perspective, it needs to be investigated as on how beneficial are classifiers which explicitly handle the temporal nature of emotion as compared to regular classifiers? Papers in EmotiW 2013 used classification methods such as the Multiple Kernel Learning (MKL) [19], deep learning [12] etc. In a recent study [15], authors compared the performance of graphical methods such as Conditional Random Field (CRF) with Support Vector Machine (SVM) on the AFEW data. They found that CRF out performed SVM based emotion recognition method.

2. EMOTIW 2013

In the first EmotiW challenge, a total of 27 teams registered for the challenge and 9 teams submitted test labels.

Attribute	Description
Length of sequences	300-5400 ms
No. of annotators	3
Emotion classes	Anger, Disgust, Fear, Happiness, Neutral, Sadness and Surprise
Total No. of clips	2577
Video format	AVI
Audio format	WAV

Table 1: Attributes of AFEW 4.0 database. EmotiW 2014 data is a subset of the AFEW 4.0 data.

Challenge	# of Samples	# of Subjects	Age Range	# of Movies
EmotiW 2013	1088	315	34.60	75
EmotiW 2014	1368	428	34.40	111

Table 2: Comparison of data and subject specifications in EmotiW 2013 and EmotiW 2014 challenges.

The database for the challenge was AFEW 3.0 [5]. The data was divided into three sets: *Train*, *Val* and *Test*. [12], proposed a deep learning based emotion recognition method and their system performed the best (41.02%) out of all participants. They used external data downloaded from Google along with the AFEW data. The second best performance is from the team of [19]. They proposed a MKL based approach, where different modalities are fused as different kernels. They used AFEW data only for training the MKL classifier. The second runner up is from the team of [14]. A method based on partial least square regression on Grassmannian manifolds was proposed. An early noisy aligned face removal is performed using PCA as a preprocessing step. Figure 1 compares the classification accuracy of EmotiW 2013 participants with the baseline and among each other. To validate the methods, a second on-site test was also conducted. A small test set was provided to the participants at the event.

Various important points regarding the problem of emotion recognition in challenging situations were discussed during the EmotiW 2013 event at Sydney. Labelling emotion on data from movie poses various challenges. We have updated the labelling process in EmotiW 2014 (w.r.t. to EmotiW 2013). We had noticed that a biase can be introduced in

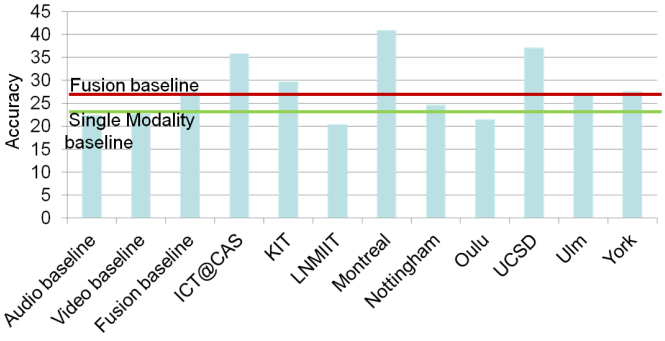

Figure 1: The graph compares the performance of participants of the first EmotiW 2013 challenge [2].

Set	# of Subjects	Max Age	Avg. Age	Min. Age	# of Males	# of Females
Train	177	76	34.09	5	102	75
Val	136	70	35.12	10	78	58
Test	115	88	34.01	5	64	51

Table 3: Subjects attributes of subjects of the EmotiW 2014 data. Age is represented in years.

few of the video clip's labels in some cases, when the labeller knows about the movie (context or generally seen the movie before) and clips are played in sequence from the same movie. Therefore, this year, a sanity check has been added, wherein 3 labellers labelled all the samples in a random order. The new labels were compared with the semi-automatic process and among each other. The samples which had agreement were kept. Samples with no consensus were ignored. In emotion labelling, there are several other challenges such as ethnicity of labellers and type of emotion labelling (discrete/continous). In the current, AFEW 4.0, discrete emotion labelling is used as it is non-trivial to label using continuous labelling or action units in datasets such as the AFEW. Furthermore, it was discussed that some samples in the AFEW database may contain multiple expressions, for e.g. a short duration Neutral expression followed by a longer duration Happy expression (displayed by the same subject). In these cases the label is assigned to the dominating emotion. Even though there may be a secondary expression, which is exhibited by an actor, other cues such as audio, body language and context can help in disambiguating the sample's final emotion label. For AFEW 4.0, the labellers have removed samples which were ambiguous i.e. which had multiple expressions on actor's face and difficult to assign a single emotion label. Though, this issue also sparks an interesting question. Is this the right time to move to assigning multiple labels for to single sample? From the example above, this means assigning both Happy and Neutral to a sample, ranked by their dominance. From a machine learning perspective, weakly labelled learning approaches (which have recently found success in affect analysis task such as pain classification [18]) can also be used for emotion recognition in such scenarios.

3. DATA

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

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 4: Set of functionals applied to LLD.

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

'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 [3]. For EmotiW 2014, the database is divided into three subsets: *Train* (578 samples), *Val* (383 samples) and *Test* (407 samples). EmotiW 2014 data is a subset of the AFEW 4.0 database. The current version, EmotiW 2014, available at http://cs.anu.edu.au/few contains two labelled sets. These are extended versions of the EmotiW 2013 [2] sets are used for training and validation; for testing, new unseen data is used. Table 2 and 3 compare and describe the data statistics of EmotiW 2013 and 2014 data. The task in the challenge is to classify a sample (audio-video clip) into one of the seven emotion categories: *Anger*, *Disgust*, *Fear*, *Happiness*, *Neutral*, *Sadness* and *Surprise*. The labeled *Train* and *Val* sets were made available early in April and the new, unlabeled test set was made available in July 2014. There are no separate audio-only, video-only or audio-video challenges. Participants are free to use either modality or combinations. Participants are allowed to use their own features and classification methods. The labels of the testing set are not shared with the participants. Participants will need to adhere to the definition of *Train*, *Val* and *Test* sets. In their papers, they may report on results obtained on the *Train* and *Val* sets, but only the results on the *Test* set will be taken into account for the overall Grand Challenge results.

4. BASELINE

4.1 Visual Analysis

For face and fiducial points detection the Mixture of Parts (MoPs) framework [10] is applied to the video frames. The Intraface tracker [22] is used to track the fiducial points initialised using MoPs framework. The fiducial points are used to align the faces. Further, spatio-temporal features are extracted on the aligned faces.

4.1.1 Volume Local Binary Patterns

Local Binary Pattern - Three Orthogonal Planes (LBP-TOP) [23] is a popular descriptor in computer vision. It considers patterns in three orthogonal planes: XY, XT and YT, and concatenates the pattern co-occurrences in these three directions. The local binary pattern (LBP-TOP) descriptor assigns binary labels to pixels by thresholding the neighborhood pixels with the central value. Therefore for a center pixel $\mathcal{O}_p$ of an orthogonal plane $\mathcal{O}$ and it's neighboring pixels N_i, a decimal value is assigned to it:

$$d = \sum_{\mathcal{O}}^{XY,XT,YT} \sum_{p} \sum_{i=1}^{k} 2^{i-1} I(\mathcal{O}_p, N_i) \quad (1)$$

LBP-TOP is computed block wise on the aligned faces of a video.

4.2 Audio Features

In this challenge, a set of audio features similar to the features employed in Audio Video Emotion Recognition Challenge 2011 [17] motivated from the INTERSPEECH 2010 Paralinguistic challenge (1582 features) [16] are used. The features are extracted using the open-source Emotion and Affect Recognition (openEAR) [8] toolkit backend openSMILE [9].

The feature set consists of 34 energy & spectral related low-level descriptors (LLD) × 21 functionals, 4 voicing related LLD × 19 functionals, 34 delta coefficients of energy & spectral LLD × 21 functionals, 4 delta coefficients of the voicing related LLD × 19 functionals and 2 voiced/unvoiced durational features. Table 4 and 5 describe the details of LLD features and functionals.

5. BASELINE EXPERIMENTS

For computing the baseline results, openly available libraries are used. Pre-trained face models available with the MoPS [24] is applied for face detection and initialisation of the Intraface tracking libray [22]. The fiducial points generated by Intraface are used for aligning the face and the face size is set to 128 × 128.

Post aligning LBP-TOP features are extracted from non-overlapping spatial 4×4 blocks. The LBP-TOP feature from each block are concatenated to create one feature vector. The concatenation is done left-to-right and top-to-bottom. Non-linear RBF kernel based SVM is learnt for emotion classification. The video only baseline system ($\mathbf{Val_{video}}$) achieves 33.15% classification accuracy on the *Val* set. The audio only baseline system is computed by extracting features using the OpenSmile toolkit. A linear SVM classifier is learnt. The audio only based system ($\mathbf{Val_{audio}}$)gives 26.10% classification accuracy on the *Val* set. A feature level fusion ($\mathbf{Val_{audio\text{-}video}}$) is performed, where the audio and video features are concatenated and a non-linear RBF kernel based SVM is learnt. The performance drops here and the classification accuracy is 28.19%. For 12 samples in the *Val* set, the face detection step failed. Therefore, for $\mathbf{Val_{audio\text{-}video}}$, wherever, face detection failed, audio only model was used.

On the *Test* set, the video only baseline system ($\mathbf{Test_{video}}$) accuracy is 33.66%; audio only based system ($\mathbf{Test_{audio}}$) classification accuracy is 26.78% and audio-video feature fusion ($\mathbf{Test_{audio\text{-}video}}$) is 24.57%. Table 8, Table 7 and Table 9 describe the classification accuracy of audio only, only and audio-video on the *Val* set. Table 11, Table 10 and Table 12 describe the classification accuracy of audio only, video only and audio-video on the *Test* set. The class-wise and overall accuracy are summarised in Table 6. The baseline performance are self-explanatory that emotion recognition in challenging conditions is a non-trivial problem. On investigating the reason for low performance of the baseline methods, several limitations were discovered. For video, there are several samples in the database for which the faces are poorly localised. The localisation error is further propagates in the fiducial point detection step. Furthermore, erroneous facial parts location lead to poor face alignment. Generally, if the face is non-frontal the fiducial parts detection quality can be poor. This leads to error in alignment and feature analysis. Recent, methods such as fiducial points free HPN [7] can be one option for handling non-frontal faces in the wild. Also, features like LBP-TOP may miss the salient frames in some samples from databases such as the AFEW, as it is unknown, when an apex of the expression will occur. Therefore, methods such as the one based on multiple instance learning [19] can be useful. For audio features, it is important for the method to know, if the speech in a scene belongs to the person of interest and not part of the background? Modelling background music score is challenging. The feature fusion performance on the *Val* and *Test* sets results in decrease in performance. This can be attributed to noise in the features and high dimensionality. Feature selection methods will be experimented with in future for increasing the performance of the audio-video analysis method.

	An	Di	Fe	Ha	Ne	Sa	Su
An	32	6	3	8	9	1	5
Di	7	10	3	4	9	2	5
Fe	15	4	7	8	5	4	3
Ha	4	3	9	36	5	6	0
Ne	4	5	4	11	12	13	4
Sa	5	7	6	10	19	10	4
Su	3	5	7	5	14	2	10

Table 7: Classification accuracy performance of $\mathbf{Val_{video}}$: the video system on the *Val* set. For 12 samples in the *Val* test the face detection step failed, these were regarded as failure cases and this was attributed to the final accuracy for the *Val* set.

	An	Di	Fe	Ha	Ne	Sa	Su
An	35	6	4	5	5	5	4
Di	11	1	1	8	13	3	3
Fe	12	5	9	9	5	2	4
Ha	12	6	6	23	9	4	3
Ne	12	7	4	17	22	1	0
Sa	2	14	12	12	10	7	4
Su	11	3	8	10	7	4	3

Table 8: Classification accuracy performance of $\mathbf{Val_{audio}}$: the audio system on the *Val* set.

6. CONCLUSION

The Second Emotion Recognition In The Wild Challenge 2014 provides a platform for researchers to benchmark and compete with their emotion recognition method on the Acted Facial Expressions In The Wild database. Emotion recognition in the wild is a challenging problem due to diversity in the scenes in the form of head pose, illumination, occlusion and background noise. This year's challenge carry forwards the platform started by the First Emotion Recognition In The Wild Challenge. The paper discusses the data partitions, baseline and experimental protocol. The performance of the baseline method is low, which speaks for the scope of work required in developing emotion recognition systems, which work well in the real-world conditions.

	Anger	Disgust	Fear	Happiness	Neutral	Sadness	Surprise	Overall
$\mathbf{Val_{audio}}$	54.68	00.025	19.56	36.50	34.92	11.47	00.065	26.10
$\mathbf{Test_{audio}}$	34.48	15.38	26.08	25.92	30.76	26.41	00.08	26.78
$\mathbf{Val_{video}}$	50.00	25.00	15.21	57.14	34.92	16.39	21.73	33.15
$\mathbf{Test_{video}}$	36.21	34.61	26.08	41.97	40.17	22.64	00.76	33.66
$\mathbf{Val_{audio\text{-}video}}$	68.75	00.00	13.04	25.39	34.92	24.59	10.86	28.19
$\mathbf{Test_{audio\text{-}video}}$	41.37	15.38	21.73	23.45	23.07	24.52	11.53	24.57

Table 6: Classification accuracy for *Val* and *Test* sets for *audio*, *video* and *audio-video* modalities.

	An	Di	Fe	Ha	Ne	Sa	Su
An	44	1	2	6	3	6	2
Di	11	0	2	8	9	9	1
Fe	15	3	6	6	6	7	3
Ha	16	4	4	16	12	7	4
Ne	7	10	2	9	22	7	6
Sa	8	8	7	13	5	15	5
Su	8	3	4	10	9	7	5

Table 9: Classification accuracy performance of $\mathbf{Val_{audio\text{-}video}}$: the audio-video fusion system on the *Val* set.

	An	Di	Fe	Ha	Ne	Sa	Su
An	21	7	6	2	14	5	3
Di	2	9	1	3	5	5	1
Fe	8	5	12	5	9	5	2
Ha	5	5	4	34	8	22	3
Ne	9	14	11	10	47	22	4
Sa	9	7	4	7	10	12	4
Su	4	0	5	4	5	6	2

Table 10: Classification accuracy performance of $\mathbf{Test_{video}}$: the video system on the *Test* set.

	An	Di	Fe	Ha	Ne	Sa	Su
An	20	6	7	7	7	4	7
Di	2	4	2	8	4	6	0
Fe	5	11	12	6	1	4	7
Ha	10	13	6	21	20	6	5
Ne	12	18	16	16	36	12	7
Sa	6	8	12	9	2	14	2
Su	6	1	6	6	2	3	2

Table 11: Classification accuracy performance of $\mathbf{Test_{audio}}$: the audio system on the *Test* set.

	An	Di	Fe	Ha	Ne	Sa	Su
An	24	2	3	7	6	23	4
Di	5	4	2	7	4	3	1
Fe	9	0	10	7	2	7	11
Ha	16	5	7	19	15	13	6
Ne	13	14	8	29	27	15	11
Sa	13	2	6	12	3	13	4
Su	7	1	4	4	3	4	3

Table 12: Classification accuracy performance of $\mathbf{Test_{audio\text{-}video}}$: the audio-video fusion system on the *Test* set.

APPENDIX

Movie Names: 21, About a boy, After the sunset, American, American History X, And Soon Came the Darkness, Aviator, Black Swan, Bridesmaids, Captivity, Carrie, Change Up, Chernobyl Diaries, Children of Men, Crying Game, December Boys, Deep Blue Sea, Descendants, Did You Hear About the Morgans?, Dumb and Dumberer: When Harry Met Lloyd, Elizabeth, Empire of the Sun, 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, 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, Little Manhattan, Messengers, Mama, Mission Impossible 2, Miss March, My Left Foot, Nothing but the Truth, Notting Hill, One Flew Over the Cuckoo's Nest, Orange and Sunshine, Pretty in Pink, Pretty Woman, Remember Me, Runaway Bride, Quartet, Romeo Juliet, Saw 3D, Serendipity, Silver Lining Playbook, Solitary Man, Something Borrowed, Terms of Endearment, The American, The Aviator, The Caller, The Devil Wears Prada, The Girl with Dragon Tattoo, The Hangover, The Haunting of Molly Hartley, The Informant!, The King's Speech, The Pink Panther 2, The Ring 2, The Social Network, The Terminal, The Town, Valentine Day, Unstoppable, Uninvited, Valkyrie, Vanilla Sky, Woman In Black, Wrong Turn 3, You've Got Mail.

A. REFERENCES

[1] M. S. Bartlett, G. Littlewort, M. G. Frank, C. Lainscsek, I. R. Fasel, and J. R. Movellan. Automatic recognition of facial actions in spontaneous expressions. *Journal of Multimedia*, 1(6):22–35, 2006.

[2] 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.

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

[4] 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.

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

[6] 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.

[7] A. Dhall, K. Sikka, G. Littlewort, R. Goecke, and M. Bartlett. A Discriminative Parts Based Model Approach for Fiducial Points Free and Shape Constrained Head Pose Normalisation In The Wild. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, pages 1–8, 2014.

[8] 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.

[9] 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.

[10] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial Structures for Object Recognition. *International Journal on Computer Vision*, 61(1):55–79, 2005.

[11] R. Gross, I. Matthews, J. F. Cohn, T. Kanade, and S. Baker. Multi-PIE. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–8, 2008.

[12] S. E. Kahou, 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.

[13] T. Kanade, J. F. Cohn, and Y. Tian. Comprehensive database for facial expression analysis. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, pages 46–53, 2000.

[14] M. Liu, R. Wang, Z. Huang, S. Shan, and X. Chen. Partial least squares regression on grassmannian manifold for emotion recognition. In *Proceedings of the ACM on International Conference on Multimodal Interaction (ICMI)*, pages 525–530, 2013.

[15] S. Nowee. *Facial Expression Recognition in the Wild: The Influence of Temporal Information*. PhD thesis, University of Amsterdam, 2014.

[16] 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.

[17] 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.

[18] K. Sikka, A. Dhall, and M. Bartlett. Weakly supervised pain localization using multiple instance learning. *Proceedings of the IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–8, 2013.

[19] 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.

[20] F. Wallhoff. Facial expressions and emotion database, 2006. http://www.mmk.ei.tum.de/~waf/fgnet/feedtum.html.

[21] J. Whitehill, G. Littlewort, I. R. Fasel, M. S. Bartlett, and J. R. Movellan. Toward Practical Smile Detection. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 31(11):2106–2111, 2009.

[22] 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.

[23] 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.

[24] 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.