Emotion Recognition In The Wild Challenge 2013*

Abhinav Dhall
Res. School of Computer Science
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

Michael Wagner
HCC Lab
University of Canberra/
Australian National University
michael.wagner@canberra.edu.au

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

ABSTRACT

Emotion recognition is a very active field of research. The Emotion Recognition In The Wild Challenge and Workshop (EmotiW) 2013 Grand Challenge consists of an audio-video based emotion classification challenges, which mimics real-world conditions. Traditionally, emotion recognition has been performed on laboratory controlled data. While undoubtedly worthwhile at the time, such lab controlled data poorly represents the environment and conditions faced in real-world situations. The goal of this Grand Challenge is to define a common platform for evaluation of emotion recognition methods in real-world conditions. The database in the 2013 challenge is the Acted Facial Expression In Wild (AFEW), which has been collected from movies showing close-to-real-world conditions.

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, Performance, Algorithms

Keywords

Audio-video data corpus, Facial expression

1. INTRODUCTION

Realistic face data plays a vital role in the research advancement of facial expression analysis. Much progress has been made in the fields of face recognition and human activity recognition in the past years due to the availability of realistic databases as well as robust representation and classification techniques. With the increase in the number of video clips online, it is worthwhile to explore the performance of emotion recognition methods that work 'in the wild'.

Emotion recognition traditionally has been been based on databases where the subjects posed a particular emotion [1] [2]. With recent advancements in emotion recognition various spontaneous databases have been introduced [3] [4]. For providing a common platform for emotion recognition researchers, challenges such as the Facial Expression Recognition & Analysis (FERA) [3] and Audio Video Emotion Challenges 2011 [5], 2012 [6] have been organised. These are based on spontaneous database [3] [4].

Emotion recognition methods can be broadly classified on the bases of the emotion labelling methodology. The early methods and databases [1] [2] used the universal six emotions (*angry*, *disgust*, *fear*, *happy*, *neutral*, *sad* and *surprise*) and *contempt/neutral*. Recent databases [4] use continuous labelling in the *Valence* and *Arousal* scales. Emotion recognition methods can also be categorised on the bases of the number of subjects in a sample. Majority of the research is based on a single subject [3] per sample. However with the popularity of social media, users are uploading images and videos from social events which contain groups of people. The task here then is to infer the emotion/mood of the group of a people [7].

Emotion recognition methods further can be categorised on the type of environment: lab-controlled and 'in the wild'. Traditional databases and methods proposed on them have lab-controlled environment. This generally means uncluttered (generally static) backgrounds, controlled illumination and minimal subject head movement. This is not the correct representative of real-world scenarios. Databases and methods which represent close-to-real-world environments (such as indoor, outdoor, different color backgrounds, occlusion and background clutter) have been recently introduced. Acted Facial Expressions In The Wild (AFEW) [8], GENKI [9], Happy People Images (HAPPEI) [8] and Static Facial Expressions In The Wild (SFEW) [10], are recent emotion databases representing real-world scenarios.

For moving the emotion recognition systems from labs to the real-world, it is important to define platforms where re-

*Initial pre-published version, will be updated in the future.

ICMI'13, December 9–12, 2013, Sydney, Australia

http://–enter the whole DOI string from rightsreview form confirmation.

searchers can verify their methods on data representing the close-to-real-world scenarios. Emotion Recognition In The Wild (EmotiW) challenge aims to provide a platform for researchers to create, extend and verify their methods on real-world data.

The challenge seeks participation from researchers working on emotion recognition intend to create, extend and validate their methods on data in real-world conditions. 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.

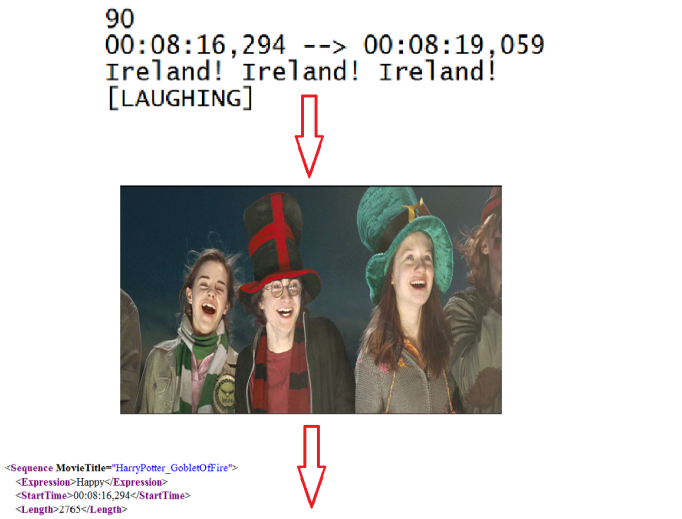

Figure 1: The screenshot described the process of database formation. For example in the screenshot, when the subtitle contains the keyword 'laughing', the corresponding clip is played by the tool. The human labeller then annotates the subjects in the scene using the GUI tool. The resultant annotation is stored in the XML schema shown in the bottom part of the snapshot. Please note that the structure of the information about a sequence containing multiple subjects. The image in the screenshot is from the movie 'Harry Potter and The Goblet Of Fire'.

Ideally, one would like to collect spontaneous data. However, as anyone working in the emotion research community will testify, collecting spontaneous databases in real-world conditions is a tedious task. For this reason, current spontaneous expression databases, for example SEMAINE, have been recorded in laboratory conditions. To overcome this limitation and the lack of available data with real-world or close-to-real-world conditions, the AFEW database has been recorded, which is a temporal database containing video clips collected by searching closed caption keywords and then validated by human annotators. AFEW forms the bases of the EmotiW challenge. While movies are often shot in somewhat controlled environments, they provide close to real world environments that are much more realistic than current datasets that were recorded in lab environments. We are not claiming that the AFEW database is a spontaneous facial expression database. However, clearly, (good) actors attempt mimicking real-world human behaviour in movies. The dataset in particular addresses the issue of emotion recognition in difficult conditions that are approximating real world conditions, which provides for a much more difficult test set than currently available datasets.

It is evident from the experiments in [8] that automated facial expression analysis in the wild is a tough problem due to various limitations such as robust face detection and alignment, and environmental factors such as illumination, head pose and occlusion. Similarly, recognising vocal expression of affect in real-world conditions is equally challenging. Moreover, as the data has been captured from movies, there are many different scenes with very different environmental conditions in both audio and video, which will provide a challenging testbed for state-of-the-art algorithms, unlike the same scene/backgrounds in lab controlled data.

Therefore, it is worthwhile to investigate the applicability of multimodal systems for emotion recognition in the wild. There has been much research on audio only, video only and to some extent audio-video multimodal systems but for translating emotion recognition systems from laboratory environments to the real-world multimodal benchmarking standards are required.

2. DATABASE CONSTRUCTION PROCESS

Databases such as the CK+, MMI and SEMAINE have been collected manually, which makes the process of database construction long and erroneous. The complexity of database collection increases further with the intent to capture different scenarios (which can represent a wide variety of real-world scenes). For constructing AFEW, a semi-automatic approach is followed [8]. The process is divided into two steps. First, subtitles from the movies using both the Subtitles for Deaf and Hearing impaired (SDH) and Closed Captions (CC) subtitles are analysed. They contain information about the audio and non-audio context such as emotions, information about the actors and the scene for example '[SMILES]', '[CRIES]', '[SURPRISED]', etc. The subtitles

Attribute	Description
Length of sequences	300-5400 ms
No. of sequences	1832 (AFEW 3.0) EmotiW: 1088
No. of annotators	2
Expression classes	Angry, Disgust, Fear, Happy, Neutral, Sad and Surprise
Total No. of expressions (some seq. have mult. sub.)	2153 (AFEW 3.0) EmotiW: 1088
Video format	AVI
Audio format	WAV

Table 2: Attributes of AFEW database.

Database	Challenges	Natural	Label	Environment	Subjects Per Sample	Construction Process
AFEW [8]	EmotiW	Spontaneous (Partial)	Discrete	Wild	Single & Multiple	Semi-Automatic
Cohn-Kanade+ [1]	-	Posed	Discrete	Lab	Single	Manual
GEMEP-FERA [3]	FERA	Spontaneous	Discrete	Lab	Single	Manual
MMI [2]	-	Posed	Discrete	Lab	Single	Manual
Semaine [2]	AVEC	Spontaneous	Continous	Lab	Single	Manual

Table 1: Comparison of AFEW database which forms the bases of the EmotiW 2013 challenge.

Expression	StartTime	Length	Person
			Pose NameOfActor AgeOfActor ExpressionOfPerson

```
<Sequence MovieTitle="HarryPotter_GobletOfFire">
  <Expression>Happy</Expression>
  <StartTime>00:08:16,294</StartTime>
  <Length>2765</Length>
  <Person Pose="Frontal" AgeOfCharacter="14" NameOfActor="Emma Watson" AgeOfActor="14" ExpressionOfPerson="Happy" Gender="Female"/>
  <Person Pose="Frontal" AgeOfCharacter="14" NameOfActor="Daniel Radcliffe" AgeOfActor="15" ExpressionOfPerson="Happy" Gender="Male"/>
  <Person Pose="Frontal" AgeOfCharacter="14" NameOfActor="Bonnie Wright" AgeOfActor="13" ExpressionOfPerson="Happy" Gender="Male"/>
</Sequence>
```

Figure 2: The figure contains the annotation attributes in the database metadata and the XML snippet is an example of annotations for a video sequence. Please note the expression tags information was removed in the xml meta-dat distributed with EmotiW data.

are extracted from the movies using a tool called VSRip[1]. For the movies where VSRip could not extract subtitles, SDH subtitles are downloaded from the internet[2]. The extracted subtitle images are parsed using Optical Character Recognition (OCR) and converted into .srt subtitle format[3]. The .srt format contains the start time, end time and text content with milliseconds accuracy.

The system performs a regular expression search with keywords[4] describing expressions and emotions on the subtitle file. This gives a list of subtitles with time stamps, which contain information about some expression. The extracted subtitles containing expression related keywords were then played by the tool subsequently. The duration of each clip is equal to the time period of appearance of the subtitle on the screen. The human observer then annotated the played video clips with information about the subjects[5] and expressions. Figure 1 describes the process. In the case of video clips with multiple actors, the sequence of labelling was based on two criteria. For actors appearing in the same frame, the ordering of annotation is left to right. If the actors appear at different time stamps, then it is in the order of appearance. However, the data in the challenge contains videos with single subject only. The labelling is then stored in the XML metadata schema. Finally, the human observer estimated the age of the character in most of the cases as the age of all characters in a particular movie is not available on the internet. The database version 3.0 contains information from 75 movies [6]

[1]VSRip http://www.videohelp.com/tools/VSRip extracts .sub/.idx from DVD movies.

[2]The SDH subtitles were downloaded from www.subscene.com, www.mysubtitles.org and www.opensubtitles.org.

[3]Subtitle Edit available at www.nikse.dk/se is used.

[4]Keyword examples: [HAPPY], [SAD], [SURPRISED], [SHOUTS], [CRIES], [GROANS], [CHEERS], etc.

[5]The information about the actors was extracted from www.imdb.com.

[6]The seventy-five movies used in the database are: 21, About a boy, American History X, And Soon Came The Darkness, Black Swan, Bridesmaids, Change Up, Chernobyl Diaries, Crying Game, Curious Case Of Benjamin Button, December Boys, Deep Blue Sea, Descendants, Did You Hear About the Morgans?, Dumb and Dumberer: When Harry Met Lloyd, Four Weddings and a Funeral, Friends with Benefits, Frost/Nixon, Ghoshtship, Girl With A Pearl Earring, Hall Pass, Halloween, Halloween Resurrection,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, I Am Sam, It's Complicated, I Think I Love My Wife, Jennifer's Body, Juno, Little Manhattan, Margot At The Wedding, Messengers, Miss March, Nany Diaries, Notting Hill, Oceans Eleven, Oceans Twelve, Oceans Thirteen, One Flew Over the Cuckoo's Nest, Orange and Sunshine, Pretty in Pink, Pretty Woman, Pursuit of Happiness, Remember Me, Revolutionary Road, Runaway Bride, Saw 3D, Serendipity, Solitary Man, Something Borrowed, Terms of Endearment, There Is Something About Mary, The American, The Aviator, The Devil Wears Prada, The Hangover, The Haunting of Molly Hartley, The Informant!, The King's Speech, The Pink Panther 2, The Social Network, The Terminal, The Town, Valentine Day, Unstoppable, Wrong Turn 3 You've Got Mail.

2.1 Database Annotations

The human labelers densely annotated the subjects in the clips. Figure 2 displays the annotations in the database. The details of the schema elements are described as follows:

StartTime - This denotes the start time stamp of the clip in the movie DVD and is in the hh:mm:ss,zzz format.

Length - It is the duration of the clip in milliseconds.

Person - This contains various attributes describing the actor in the scene described as follows: *Pose* - This denotes the pose of the actor, based on the human labeler's observation.

AgeOfCharacter - This describes the age of the character based on human labeler's observation. In few cases the age of the character available www.imdb.com was used. But this was frequent in case of lead actors only.

NameOfActor - This attribute contains the real name of the actor.

AgeOfActor - This describes the real age of the actor. The information was extracted from www.imdb.com by the human labeler. In very few cases the age information was missing for some actors!, therefore the observational values were used.

Gender - This attribute describes the gender of the actor, again entered by the human labeler.

3. EMOTIW DATA PARTITIONS

The challenge data is divided into three sets: '*Train*', '*Val*' and '*Test*'. *Train*, *Val* and *Test* set contain 380, 396 and 312 clips respectively. The AFEW 3.0 dataset contains 1832 clips, for EmotiW challenge 1088 clips are extracted. The data is subject independent and the sets contains clips from different movies. The motivation behind partitioning the data in this manner is to test methods for unseen scenario data, which is common on the web. For the participants in the challenge, the labels of the testing set are unknown. The details about the subjects is described in Table 3.

4. VISUAL ANALYSIS

For face and fiducial points detection the Mixture of Parts (MoPs) framework [11] is applied to the video frames. MoPs represents the parts of an object as a graph with n vertices $V = \{v_1, \ldots, v_n\}$ and a set of edges E. Here, each edge $(v_i, v_j) \in E$ pair encodes the spatial relationship between parts i and j. A face is represented as a tree graph here. Formally speaking, for a given image I, the MoPs framework computes a score for the configuration $L = \{l_i : i \in V\}$ of parts based on two models: an appearance model and a spatial prior model. We follow [12]'s mixture-of-parts formulation.

The **Appearance Model** scores the confidence of a part specific template w_p applied to a location l_i. Here, p is a view-specific mixture corresponding to a particular head pose. $\phi(I, l_i)$ is the histogram of oriented gradient descriptor [13] extracted from a location l_i. Thus, the appearance model calculates a score for configuration L and image I as:

$$App_p(I, L) = \sum_{i \in V_p} {w_i}^p . \phi(I, l_i) \quad (1)$$

The **Shape Model** learns the kinematic constraints between each pair of parts. The shape model (as in [12]) is defined as:

$$Shape_p(L) = \sum_{ij \in E_p} a_{ij}^p dx^2 + b_{ij}^p dx + c_{ij}^p dy^2 + d_{ij}^p dy \quad (2)$$

Here, dx and dy represent the spatial distance between two parts. a, b, c and d are the parameters corresponding to the location and rigidity of a spring, respectively. From Eq. 1 and 2, the scoring function S is:

$$Score(I, L, p) = App_p(I, L) + Shape_p(L) \quad (3)$$

During the inference stage, the task is to maximise Eq. 3 over the configuration L and mixture p (which represents a pose). The fiducial points are used to align the faces. Further, spatio-temporal features are extracted on the aligned faces.

The aligned faces are shared with participants. Along with MoPS, aligned faces computed by the method of Gehrig and Ekenel [14] is also shared.

4.1 Volume Local Binary Patterns

Local Binary Pattern - Three Orthogonal Planes (LBP-TOP) [15] 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:

$$d = \sum_{\mathcal{O}}^{XY,XT,YT} \sum_{p} \sum_{i=1}^{k} 2^{i-1} I(\mathcal{O}_p, N_i) \quad (4)$$

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

5. AUDIO FEATURES

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

Set	Num of Subj.	Max Age	Avg Age	Min Age	Males	Fe--males
Train	99	76y	32.8y	10y	60	39
Val	126	70y	34.3y	10y	71	55
Test	90	70y	36.7y	8y	50	40

Table 3: Subject description of the three sets.

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

	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise	Overall
Val$_{audio}$	42.37	12.00	25.93	20.97	12.73	14.06	9.62	19.95
Test$_{audio}$	44.44	20.41	27.27	16.00	27.08	9.30	5.71	22.44
Val$_{video}$	44.00	2.00	14.81	43.55	34.55	20.31	9.62	27.27
Test$_{video}$	50.00	12.24	0.00	48.00	18.75	6.97	5.71	22.75
Val$_{audio\text{-}video}$	44.07	0.00	5.56	25.81	63.64	7.81	5.77	22.22
Test$_{audio\text{-}video}$	66.67	0.00	6.06	16.00	81.25	0.00	2.86	27.56

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

Affect Recognition (openEAR) [18] toolkit backend openSMILE [19].

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 5 describe the details of LLD features and functionals.

6. BASELINE EXPERIMENTS

For computing the baseline results, openly available libraries are used. Pre-trained face models (Face_p146_small, Face_p99 and MultiPIE_1050) available with the MoPS package [7] were applied for face and fiducial points detection. The models are applied in hierarchy.

The fiducial points generated by MoPS is used for aligning the face and the face size is set to 96×96. 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. Non-linear SVM is learnt for emotion classification. The video only baseline system achieves 27.2% classification accuracy on the *Val* set. The audio baseline is computed by extracting features using the OpenSmile toolkit. A linear SVM classifier is learnt. The audio only based system gives 19.5% classification accuracy on the *Val*. Further, a feature level fusion is performed, where the audio and video features are concatenated and a non-linear SVM is learnt. The performance drops here and the classification accuracy is 22.2%. On the *Test* set which contains 312 video clips, audio only gives 22.4%, video only gives 22.7% and feature fusion gives 27.5%.

Table 6, describes the classification accuracy for the *Val* and *Test* for audio, video and audio-video systems. For the *Test* set the feature fusion increases the performance of the system. However, the same is not true for the *Val* set. The confusion matrices for val and test are described in **Val$_{audio}$**: Table 7, **Val$_{video}$**: Table 8, **Val$_{audio\text{-}video}$**: Table 9, **Test$_{audio}$**: Table 10, **Test$_{video}$**: Table 11, **Test$_{audio\text{-}video}$**: Table 12.

[7] http://www.ics.uci.edu/ xzhu/face/

	An	Di	Fe	Ha	Ne	Sa	Su
An	25	10	7	6	1	4	6
Di	13	6	4	9	7	5	6
Fe	12	8	14	6	4	8	2
Ha	20	3	8	13	8	4	6
Ne	8	10	5	16	7	6	3
Sa	12	15	12	6	2	9	8
Su	14	7	7	7	8	4	5

Table 7: Val$_{audio}$: Confusion matrix describing performance of the audio subsystem on the *Val* set.

	An	Di	Fe	Ha	Ne	Sa	Su
An	26	0	2	6	8	11	6
Di	15	10	4	6	7	7	1
Fe	18	3	8	5	6	5	9
Ha	20	1	5	27	3	5	1
Ne	8	5	7	7	19	2	7
Sa	15	3	4	6	13	13	10
Su	11	5	4	8	11	8	5

Table 8: Val$_{video}$: Confusion matrix describing performance of the video subsystem on the *Val* set.

The automated face localisation on the database is not always accurate, with a significant number of false positives and false negatives. This is attributed to the varied lightning conditions, occlusions, extreme head poses and complex backgrounds.

7. CONCLUSION

Emotion Recognition In The Wild (EmotiW) challenge is platform for researchers to compete with their emotion

	An	Di	Fe	Ha	Ne	Sa	Su
An	26	1	2	7	17	3	3
Di	4	0	0	14	30	1	1
Fe	11	2	3	14	17	4	3
Ha	11	0	2	16	30	2	1
Ne	7	1	0	12	35	0	0
Sa	7	0	2	17	28	5	5
Su	2	0	3	7	33	4	3

Table 9: $\text{Val}_{\text{audio-video}}$: Confusion matrix describing performance of the audio-video fusion system on the *Val* set.

recognition methods on 'in the wild' data. The audio-visual challenge data is based on the AFEW database. The labelled 'Train' and 'Val' sets were shared along with unlabelled 'Test' set. Meta-data containing information about the actor in the clip are shared with the participants. The performance of the different methods will be analysed for insight on performance of the state-of-art emotion recognition methods on 'in the wild' data.

8. REFERENCES

[1] Patrick Lucey, Jeffrey F. Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *CVPR4HB10*, 2010.

[2] Maja Pantic, Michel François Valstar, Ron Rademaker, and Ludo Maat. Web-based database for facial expression analysis. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, ICME'05, 2005.

[3] Michel Valstar, Bihan Jiang, Marc Mehu, Maja Pantic, and Scherer Klaus. The first facial expression recognition and analysis challenge. In *Proceedings of the Ninth IEEE International Conference on Automatic Face Gesture Recognition and Workshops*, FG'11, pages 314–321, 2011.

[4] Gary McKeown, Michel François Valstar, Roderick Cowie, and Maja Pantic. The semaine corpus of emotionally coloured character interactions. In *IEEE ICME*, 2010.

[5] Björn Schuller, Michel François Valstar, Florian Eyben, Gary McKeown, Roddy Cowie, and Maja Pantic. Avec 2011-the first international audio/visual emotion challenge. In *ACII (2)*, pages 415–424, 2011.

[6] Björn Schuller, Michel Valstar, Florian Eyben, Roddy Cowie, and Maja Pantic. Avec 2012: the continuous audio/visual emotion challenge. In *ICMI*, pages 449–456, 2012.

[7] Abhinav Dhall, Jyoti Joshi, Ibrahim Radwan, and Roland Goecke. Finding happiest moments in a social context. In *ACCV*, 2012.

[8] Abhinav Dhall, Roland Goecke, Simon Lucey, and Tom Gedeon. A semi-automatic method for collecting richly labelled large facial expression databases from movies. *IEEE Multimedia*, 2012.

[9] Jacob Whitehill, Gwen Littlewort, Ian R. Fasel, Marian Stewart Bartlett, and Javier R. Movellan. Toward Practical Smile Detection. *IEEE TPAMI*, 2009.

[10] Abhinav Dhall, Roland Goecke, Simon Lucey, and Tom Gedeon. Static Facial Expression Analysis In Tough Conditions: Data, Evaluation Protocol And Benchmark. In *ICCVW*, BEFIT'11, 2011.

[11] P.F. Felzenszwalb and D.P. Huttenlocher. Pictorial Structures for Object Recognition. *IJCV*, 2005.

[12] Xiangxin Zhu and Deva Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *CVPR*, pages 2879–2886, 2012.

[13] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *CVPR*, pages 886–893, 2005.

[14] Tobias Gehrig and Hazım Kemal Ekenel. A common framework for real-time emotion recognition and facial action unit detection. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2011 IEEE Computer Society Conference on*, pages 1–6. IEEE, 2011.

[15] Guoying Zhao and Matti Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 2007.

	An	Di	Fe	Ha	Ne	Sa	Su
An	24	4	6	9	2	3	6
Di	14	10	2	9	7	4	3
Fe	8	4	9	2	4	2	4
Ha	17	4	4	8	5	7	5
Ne	6	8	6	7	13	6	2
Sa	12	6	6	7	3	4	5
Su	6	5	6	9	2	5	2

Table 10: $\text{Test}_{\text{audio}}$: Confusion matrix describing performance of the audio subsystem on the *Test* set.

	An	Di	Fe	Ha	Ne	Sa	Su
An	27	3	3	4	6	4	7
Di	14	6	4	7	6	4	8
Fe	9	4	0	4	9	2	5
Ha	9	5	1	24	1	4	6
Ne	11	13	1	5	9	6	3
Sa	8	3	3	11	10	3	5
Su	7	5	6	5	7	3	2

Table 11: $\text{Test}_{\text{video}}$: Confusion matrix describing performance of the video subsystem on the *Test* set.

	An	Di	Fe	Ha	Ne	Sa	Su
An	36	0	1	2	14	0	1
Di	13	0	1	15	18	1	1
Fe	8	1	2	4	16	0	2
Ha	12	1	2	8	22	1	4
Ne	5	0	0	3	39	1	0
Sa	16	1	1	8	13	0	4
Su	10	1	2	10	9	2	1

Table 12: $\text{Test}_{\text{audio-video}}$: Confusion matrix describing performance of the audio-video fusion system on the *Test* set.

[16] Björn Schuller, Michel Valstar, Florian Eyben, Gary McKeown, Roddy Cowie, and Maja Pantic. Avec 2011–the first international audio/visual emotion challenge. In *Affective Computing and Intelligent Interaction*, pages 415–424. Springer Berlin Heidelberg, 2011.

[17] Björn Schuller, Stefan Steidl, Anton Batliner, Felix Burkhardt, Laurence Devillers, Christian A Müller, and Shrikanth S Narayanan. The interspeech 2010 paralinguistic challenge. In *INTERSPEECH*, pages 2794–2797, 2010.

[18] Florian Eyben, Martin Wollmer, and Bjorn Schuller. OpenearâĂŤintroducing the munich open-source emotion and affect recognition toolkit. In *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*, pages 1–6. IEEE, 2009.

[19] Florian Eyben, Martin Wöllmer, and Björn Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. In *ACM Multimedia*, pages 1459–1462, 2010.