Affectiva-MIT Facial Expression Dataset (AM-FED): Naturalistic and Spontaneous Facial Expressions Collected In-the-Wild

Daniel McDuff[†‡], Rana El Kaliouby[†‡], Thibaud Senechal[‡], May Amr[‡], Jeffrey F. Cohn[§], Rosalind Picard[†‡]
‡ Affectiva Inc., Waltham, MA, USA
† MIT Media Lab, Cambridge, MA, USA
§ Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, USA
djmcduff@media.mit.edu, {kaliouby,thibaud.senechal,may.amr}@affectiva.com,
jeffcohn@cs.cmu.edu, picard@media.mit.edu

Abstract

Computer classification of facial expressions requires large amounts of data and this data needs to reflect the diversity of conditions seen in real applications. Public datasets help accelerate the progress of research by providing researchers with a benchmark resource. We present a comprehensively labeled dataset of ecologically valid spontaneous facial responses recorded in natural settings over the Internet. To collect the data, online viewers watched one of three intentionally amusing Super Bowl commercials and were simultaneously filmed using their webcam. They answered three self-report questions about their experience. A subset of viewers additionally gave consent for their data to be shared publicly with other researchers. This subset consists of 242 facial videos (168,359 frames) recorded in real world conditions. The dataset is comprehensively labeled for the following: 1) frame-by-frame labels for the presence of 10 symmetrical FACS action units, 4 asymmetric (unilateral) FACS action units, 2 head movements, smile, general expressiveness, feature tracker fails and gender; 2) the location of 22 automatically detected landmark points; 3) self-report responses of familiarity with, liking of, and desire to watch again for the stimuli videos and 4) baseline performance of detection algorithms on this dataset. This data is available for distribution to researchers online, the EULA can be found at: http://www.affectiva.com/facial-expression-dataset-am-fed/.

1. Introduction

The automatic detection of naturalistic and spontaneous facial expressions has many applications, ranging from medical applications such as pain detection [1], or monitoring of depression [4] and helping individuals on the autism spectrum [10] to commercial uses cases such as advertising research and media testing [14] to understanding non-verbal communication [19]. With the ubiquity of cameras on computers and mobile devices, there is growing interest in bringing these applications to the real-world. To do so, spontaneous data collected from real-world environments is needed. Public datasets truly help accelerate research in an area, not just because they provide a benchmark, or a common language, through which researchers can communicate and compare their different algorithms in an objective manner, but also because compiling such a corpus and getting it reliably labeled, is tedious work - requiring a lot of effort which many researchers may not have the resources to do.

There are a number of publicly available labeled databases for automated facial analysis, which have helped accelerate research in automated facial analysis tremendously. Databases commonly used for facial action unit and expression recognition include; Cohn-Kanade (in its extended edition know as CK+) [11], MMI [23], RU-FACS [2], Genki-4K [24] and UNBC-McMaster Pain archive [12]. These datasets are reviewed in Section 2. However, all (except the Genki-4K and UNBC-McMaster Pain archives) were captured in controlled environments which do not reflect the the type of conditions seen in real-life applications. Computer-based machine learning and pattern analysis depends hugely on the number of training examples [22]. To date much of the work automating the analysis of facial expressions and gestures has had to make do with limited datasets for training and testing. As a result this often leads to over-fitting.

Inspired by other researchers who made an effort to share their data publicly with researchers in the field, we present a database of spontaneous facial expressions that was collected in naturalistic settings as viewers watched video content online. Many viewers watched from the comfort of their homes, which meant that the facial videos contained a range of challenging situations, from nonuniform lighting and head movements, to subtle and nuanced expressions. To collect this large dataset, we leverage Internet crowdsourcing, which allows for distributed collection of data very efficiently. The data presented are natural spontaneous responses to ecologically valid online media (video advertising) and labels of self-reported liking, desire to watch again and familiarity. The inclusion of self-reported labels is especially important as it enables systematic research around the convergence or divergence of self-report and facial expressions, and allows us to build models that predict behavior (e.g, watching again).

While data collection is a major undertaking in and of itself, labeling that data is by far a much grander challenge. The Facial Action Coding System (FACS) [7] is the most comprehensive catalogue of unique facial action units (AUs) that correspond to each independent motion of the face. FACS enables the measurement and scoring of facial activity in an objective, reliable and quantitative way, and is often used to discriminate between subtle differences in facial motion. One strength of FACS is the high level of detail contained within the coding scheme, this has been useful in identifying new behaviors [8] that might have been missed if a coarser coding scheme were used.

Typically, two or more FACS-certified labelers code for the presence of AUs, and inter-observer agreement is computed. There are a number of methods of evaluating the reliability of inter-observer agreement in a labeling task. As the AUs differ in how easy they are identified, it is important to report agreement for each individual label [3]. To give a more complete perspective on the reliability of each AU label, we report two measures of inter-observer agreement for the dataset described in this paper.

The main contribution of this paper is to present a first in the world data set of labeled data recorded over the internet of people naturally viewing online media, the AM-FED dataset contains:

1. **Facial Videos:** 242 webcam videos recorded in real-world conditions.

2. **Labeled Frames:** 168,359 frames labeled for the presence of 10 symmetrical FACS action units, 4 asymmetric (unilateral) FACS action units, 2 head movements, smile, expressiveness, feature tracker fails and gender.

3. **Tracked Points:** Automatically detected landmark points for 168,359 frames.

4. **Self-report responses:** Familiarity with, liking of and desire to watch again for the stimuli videos

5. **Baseline Classification:** Baseline performance of smile, AU2 and AU4 detection algorithms on this dataset and baseline classifier outputs.

To the authors knowledge this dataset is the largest set labeled for asymmetric facial action units AU12 and AU14.

In the remainder of this paper we describe the data collection, labeling and label reliability calculation, and the training, testing and performance of smile, AU2 and AU4 detection on this dataset.

2. Existing Databases

The Cohn-Kanade (in its extended edition known as CK+) [11] has been one of the mostly widely used resource in the development of facial action unit and expression recognition systems. The CK+ database, contains 593 recordings (10,708 frames) of posed and non-posed sequences, which are FACS coded as well as coded for the six basic emotions. The sequences are recorded in a lab setting under controlled conditions of light and head motion.

The MMI database contains a large collection of FACS coded facial videos [23]. The database consists of 1395 manually AU coded video sequences, 300 also have onset-appex-offset annotions. A majority of these are posed and all are recorded in laboratory conditions.

The RU-FACS database [2] contains data from 100 participants each engaging in a 2.5 minute task. In the task, the participants had to act to hide their true position, and therefore one could argue that the RU-FACS dataset is not fully spontaneous. The RU-FACS dataset is not publicly available at this time.

The Genki-4K [24] dataset contains 4000 images labeled as either “smiling” or “non-smiling”. These images were collected from images available on the Internet and do mostly reflect naturalistic smiles. However, these are just static images and not video sequences making it impossible to use the data to train systems that use temporal information. In addition, the labels are limited to presence or absence of smiles and therefore limiting their usefulness.

The UNBC-McMaster Pain archive [12] is one of the largest databases of AU coded videos of naturalistic and spontaneous facial expressions. This is labeled for 10 action units and the action units are coded with levels of intensity making it very rich. However, although of naturalistic and spontaneous expressions the videos were recorded with control over the lighting, camera position, frame rate and resolution.

Multi-PIE [9] is a dataset of static facial expression images using 15 cameras in different locations and 18 flashes to create various lighting conditions. The dataset includes 6 expressions plus neutral. The JAFFE [13] and Semaine [18] datasets contain videos with labeled facial expressions. However, Multi-PIE, JAFFE and Semaine were collected in controlled laboratory settings and are not FACS labeled, but rather have “message judgement” labels, and so are not readily available for training AU detectors.

O'Toole et al. [20] present a database including videos of facial expressions shot under controlled conditions.

3. Data Collection

Figure 1 shows the web-based framework that was used to crowdsource the facial videos and the user experience. Visitors to the website opt-in to watch short videos while their facial expressions are being recorded and analyzed. Immediately following each video, visitors get to see where they smiled and with what intensity. They can compare their "smile track" to the aggregate smile track. On the client-side, all that is needed is a browser with Flash support and a webcam. The video from the webcam is streamed in real-time at 14 frames a second at a resolution of 320x240 to a server where automated facial expression analysis is performed, and the results are rendered back to the browser for display. There is no need to download or install anything on the client side, making it very simple for people to participate. Furthermore, it is straightforward to easily set up and customize "experiments" to enable new research questions to be posed. For this experiment, we chose three successful Super Bowl commercials: 1. Doritos ("House sitting", 30 s), 2. Google ("Parisian Love", 53 s) and 3. Volkswagen ("The Force", 62 s). Viewers chose to view one or more of the videos.

On selecting a commercial to watch, visitors are asked to 1) grant access to their webcam for video recording and 2) to allow MIT and Affectiva to use the facial video for internal research. Further consent for the data to be shared with the research community at large is also sought, and only videos with consent to be shared publicly are shown in this paper. This data collection protocol was approved by the MIT Committee On the Use of Humans as Experimental Subjects (COUHES) prior to launching the site. A screenshot of the consent form is shown in Figure 2. If consent is granted, the commercial is played in the browser whilst simultaneously streaming the facial video to a server. In accordance with MIT COUHES, viewers could opt-out if they chose to at any point while watching the videos, in which case their facial video is immediately deleted from the server. If a viewer watches a video to the end, then his/her facial video data is stored along with the time at which the session was started, their IP address, the ID of the video they watched and self-reported responses (if any) to the self report questions. No other data is stored. A similar web-based framework is described in [16]. Participants were aware that their webcam was being used for recording, however, at no point within the interaction were they shown images from their webcam. This may have had an impact on their behavior but the majority of videos contain naturalistic and spontaneous responses.

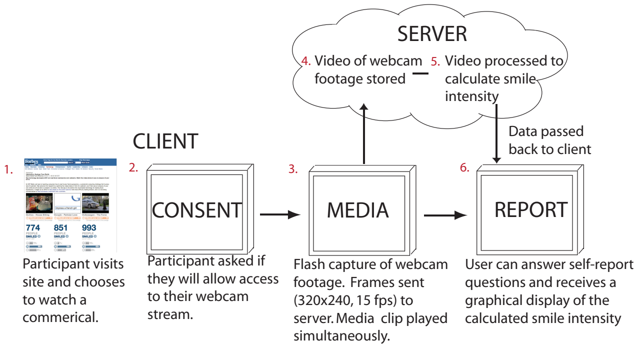
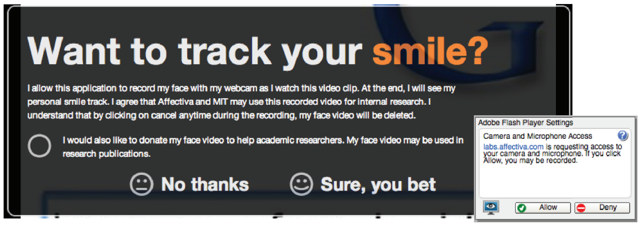

SERVER
4. Video of webcam footage stored
5. Video processed to calculate smile intensity
Data passed back to client
CLIENT
1.
774
851
993
Participant visits site and chooses to watch a commerical.
2.
CONSENT
Participant asked if they will allow access to their webcam stream.
3.
MEDIA
Flash capture of webcam footage. Frames sent (320x240, 15 fps) to server. Media clip played simultaneously.
6.
REPORT
User can answer self-report questions and receives a graphical display of the calculated smile intensity

Figure 1. Overview of what the user experience was like and the web-based framework that was used to crowdsource the facial videos. The video from the webcam is streamed in real-time to a server where automated facial expression analysis is performed, and the results are rendered back to the browser for display. All the video processing was done on the server side.



Want to track your smile?

I allow this application to record my face with my webcam as I watch this video clip. At the end, I will see my personal smile track. I agree that Affectiva and MIT may use this recorded video for internal research. I understand that by clicking on cancel anytime during the recording, my face video will be deleted.

○ I would also like to donate my face video to help academic researchers. My face video may be used in research publications.

Adobe Flash Player Settings

Camera and Microphone Access

labs.affectiva.com is requesting access to your camera and microphone. If you click Allow, you may be recorded.

Figure 2. The consent forms that the viewers were presented with before watching the video and before the webcam stream began.

We collected a total of 6,729 facial videos from 5,268 people who completed the experiment. We disregard videos for which the face tracker was unable to identify a face in at least 90% of frames; this left 3,268 videos (20.0%). As mentioned earlier, the participants were given the option to make their face video available for research purposes. For 489 (7.3%) of the videos this was checked. Due to the considerable effort required in coding 242 of these videos have been hand labeled and are available for public release. We refer to the public portion of the data collected as the AM-FED dataset. All videos were recorded with a resolution of 320x240 and a frame rate of 14 fps. The data contain many challenges from an automated facial action and gesture recognition perspective. Firstly, the lighting is very varied both in terms of illumination and contrast making appearance vary markedly. Secondly, there is considerable range in the pose and scale of the viewers' faces as there were no restrictions applied to the position of the camera and the viewer's were not shown any images from their webcam. Figure 1 (top) shows a selection of frames from the dataset as examples of the diversity. The properties of the larger dataset from which the public data is taken can be found in [15]. This demonstrates that the data contains significantly more varied data, in terms of lighting and pose and position of the participants, than in the CK+ and MMI databases. The gender of subjects and whether they are wearing glasses in the video are labeled in the dataset. The details are provided in Table 1.

Table 1. Demographic, glasses wearing and facial hair information about videos in the dataset.

Gender		Glasses	Facial hair
Male	Female	Present	Present
140	102	86	37

Table 2. Definitions of the labels for the dataset and the number of frames and videos in which each label was present (agreed by majority of labelers). Positive examples of each of the labels are shown in Figure 5

Label	**Definition**	**Frames Present**	**Videos Present**
Gender	Gender of the viewer	-	242
AU2	Outer eyebrow raise	2,587	50
AU4	Brow lowerer	2,274	22
AU5	Upper lid raiser	991	11
AU9	Nose wrinkler	3	1
AU10	Upper lip raiser	26	1
AU14	Symmetrical dimpler	1,161	27
AU15	Lip corner depressor	1	1
AU17	Chin raiser	1,500	30
AU18	Lip pucker	89	7
AU26	Jaw drop	476	6
AU57	Head is forward	253	22
AU58	Head is backward	336	37
Expressiveness	Non-neutral face (may contain AUs that are not labeled)	68,028	208
Smile	Smile (distinct from AU12)	37,623	180
Trackerfail	Frames in which the track failed to accurately find the correct points on the face	18,060	76
Unilateral left AU12	Left asymmetric AU12	467	6
Unilateral right AU12	Right asymmetric AU12	2,330	14
Unilateral left AU14	Left asymmetric dimpler	226	8
Unilateral right AU14	Right asymmetric dimpler	105	4
Negative AU12	AU12 and AU4 together - distinct from AU12 in smile	62	2

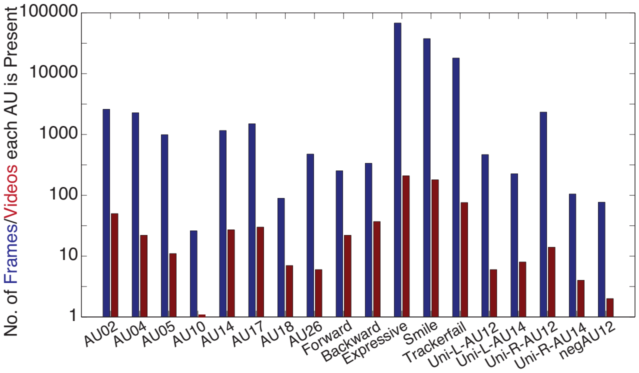
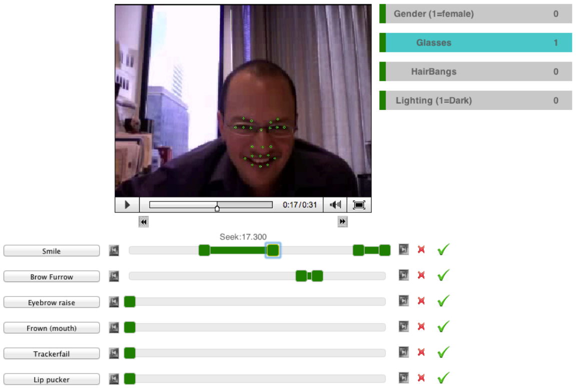


Figure 3. Number of frames in which each label is present (with agreement for >= 50% of labelers).



Gender (1=female) 0
Glasses 1
HairBangs 0
Lighting (1=Dark) 0

Figure 4. Screenshot of the video labeling tool ViDL used to label the videos in the dataset.

4. FACS Coding

Each of the videos were independently labeled, frame-by-frame, by at least three FACS trained coders chosen from a pool 16 coders (labeling stage). All 16 coders had undergone FACS training and three were FACS certified. The labels were subsequently labeled by another independent FACS trained individual (QA stage) and discrepancies within the coding reviewed (relabeling stage). For labeling we used a web-based, distributed video labeling system (ViDL) which is specifically designed for labeling affective data [6]. A version of ViDL developed by Affectiva was used for the labeling task. Figure 4 shows a screenshot of the ViDL interface. The labelers were working independently for the labeling. The coders labeled for presence (binary labels) of AU2, AU4, AU5, AU9, AU12 (unilateral and bilateral), AU14 (unilateral and bilateral), AU15, AU17, AU18 and AU26. Smiles are labeled and are distinct from the labels for AU12 as AU12 may occur in an expression that would not necessary be given the label of smile (e.g. a grimace). The expressiveness label describes the presence of any non-neutral facial expression. The trackerfail label indicates a frame in which the automated Nevenvision facial feature tracker (licensed from Google, Inc.), for which the detected points are provided with the dataset, were not accurately tracking the correct locations on the face. This gives a total of 168,359 FACS coded frames. Definitions of the labels and the number of frames in which they were labeled present by a majority of the labelers are shown in Table 2. Although AU9 and AU15 were coded for, there were only 1 or 2 examples identified by a majority of the coders. Therefore we do not evaluate the reliability of AU9 and AU15. In the smile and action unit classification section of this paper, we assume a label is present if over 50% of the labelers agree it is present and assume that a label is not present if 100% of the labelers agree it is not present. We do not use the frames that do not satisfy these criteria for the classification task.

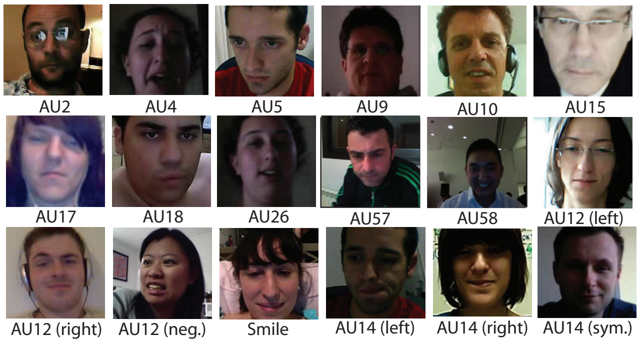


Figure 5. Cropped examples of frames with positive labels for each of the action units coded. Smile and negative AU12 are labeled separately instead of labeling symmetrical AU12.

4.1. Reliability of Labels

A minimum of three coders labeled each frame of the data and agreement between the coders was not necessarily 100%. The labels provided in the archive give the breakdown of all the labelers judgements. We present the reliability of the FACS coding. The reliability for each set of AU labels in a particular video, *p*, is the mean correlation between all pair-wise combinations of the coders labels for that video sequence. Then the "effective reliability" is evaluated using the Spearman-Brown measure of effective reliability [21]. The Spearman-Brown measure of reliability is calculated as:

$$R_{S-B} = \frac{Np}{1+(N-1)p} \tag{1}$$

Where *N* is the number of "tests", in this case the number of coders. The effective reliability accounts for the fact that theoretically employing more that one coder will mean that random errors within the coding begin to cancel out and therefore the effective reliability is greater than the mean reliability for a particular video.

The weighted-mean Spearman-Brown reliability, across all 242 video sequences, for each of the labels is shown in Figure 6. The weighted-mean reliability was calculated by giving the reliability for each video-AU combination a weighting relative to the number of agreed positive examples in that video. As such, a video with very few positive labels that has poor reliability score is down-weighted relative to one that has many positive examples.

As the reliability measure calculated above does not reward agreement by labelers in videos that do not contain any examples of an action unit (i.e. they all label absence of an action for the whole video) we also calculated the percentage agreement across all pairs of labelers and all frames for each of the labels. The mean percentage agreement across all AUs was 0.98, the minimum was for AU26 = 0.87.

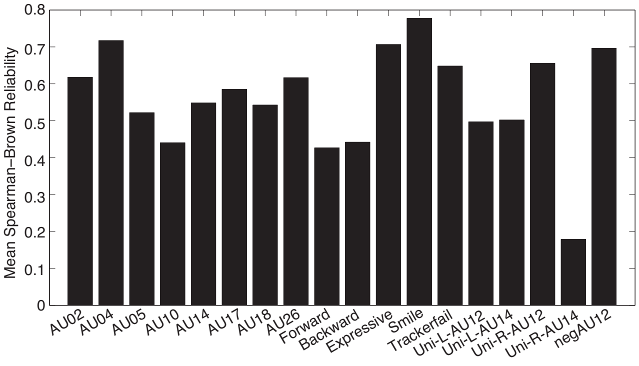


Figure 6. Bar graph showing the mean in the Spearman-Brown reliability for each of the labels

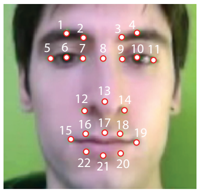


Figure 7. Locations of the 22 landmark points automatically labeled using the Nevenvision tracker that are provided with the dataset.

5. Fiducial Points

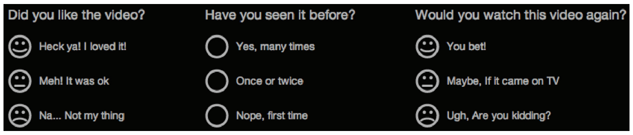
The data is also provided with the frame-by-frame locations of 22 automatically detected landmark points on the face. The points were detected using the Nevenvision tracker. The locations of the points are shown in Figure 7. In some cases the tracker could not identify a face. For these frames the automatic labels (landmark points, smile and action unit classifier outputs) are assigned -1 to indicate that no face was identified.

6. Self-report Responses

Following viewing a commercial viewers could optionally answer three multiple choice questions: "Did you like the video?", "Have you seen it before?" and "Would you watch this video again?". A screenshot of the questions is shown in Figure 8. Viewers were not required to answer the questions and the page would time-out after a time. The responses to the questions (if any) are provided with the dataset. For the publicly available data 234 people answered the likability question, 219 people answered the familiarity question and 194 people answered the desire question. Some preliminary analysis of the relationship between the facial responses and self-report labels collected can be found in [16, 17].

7. Experiments

Whilst this is not a paper focused on AU detection we provide baseline performance for automated AU detection.



Did you like the video?

☺ Heck ya! I loved it!

😐 Meh! It was ok

☹ Na... Not my thing

Have you seen it before?

○ Yes, many times

○ Once or twice

○ Nope, first time

Would you watch this video again?

☺ You bet!

😐 Maybe, If it came on TV

☹ Ugh, Are you kidding?

Figure 8. The self-report questions the viewers were presented with after watching the commercial.

The action units for which we present results are AU2 (outer eyebrow raise), AU4 (brow lowerer), and smile (labelers labeled for presence of a "smile" rather than AU12). The output of each of the classifiers is a probability estimate of the presence of each action unit. The baseline results are set using the following method. The tracker was used to automatically detect the face and track 22 facial landmark points within each frame of the videos. The locations of the facial landmarks are shown in Figure 7. The landmarks were used to locate the face ROI and the segmented face images were rescaled to 120x120 pixels. An affine warp was performed on the bounded face region to account for in-planar head movement. For the smile detection the landmarks were used to locate a region around the mouth and histogram of orientated gradients (HOG) [5] features are computed for the region. The classifiers each use a Support Vector Machine (SVM) with RBF kernel, these showed significantly better performance than random forest classifiers. SVMs have been shown to perform well on smile detection in the past [24]. For the AU2 and AU4 classifiers the landmarks were used to locate a region around the eyes and HOG features were computed for that region.

The AU classifiers were trained and validated on examples from other datasets collected over the web and similar in nature to the data available in the AM-FED dataset. The training and validation sets were independent from one another and were also person-independent. For testing the complete set of public frames in this dataset (168,359 frames) were taken and those for which there was greater than 50% agreement of the present of each action unit or 100% agreement of the absence of each action unit used. For training, validation and testing in the design of the classifiers 16,000 frames were used for the AU2 classifier, 58,000 frames were used for the AU4 classifier and 114,000 frames for the smile classifier. In the validation stage the classifier parameters were selected by maximizing the area under the receiver operating characteristic (ROC) curve. During validation the HOG parameters and the size of facial ROI were optimized. For the SVM classifier the spread of the RBF kernel (γ) and the penalty parameter (C) were optimized.

ROC curves were calculated for each of the AU algorithms, these are shown in Figure 10 respectively. The decision-boundary was varied to calculate the ROC curves shown. The area under the ROC curve for the smile, AU2 and AU4 classifiers was 0.90, 0.72 and 0.70.

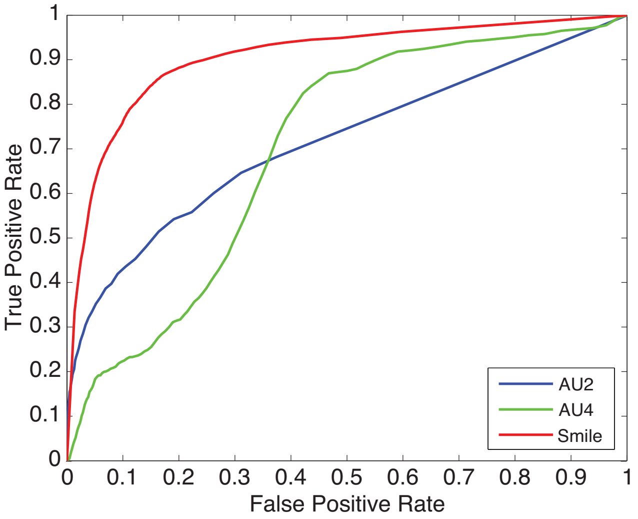


Figure 10. Receiver Operating Characteristic (ROC) curves for the performance of the smile, AU2 and AU4 classifiers on videos from the dataset. Smile AUC = 0.90, AU2 AUC=0.72, AU4 AUC=0.70.

The baseline performance shows that accurate AU detection is possible on this challenging, naturalistic and spontaneous data. However, this paper does not focus on the task of AU detection and there remains room for improvement. In particular the detection of action units is difficult in low illumination conditions. Greater details of the variations in conditions within the larger dataset from which the labeled public data is taken can be found in [15].

8. Distribution Details

Participants provided informed consent for use of their video images for scientific research purposes. Distribution of the dataset is governed by the terms of their informed consent. The data may be used for research purposes and images from the dataset used in academic publications. All of the images in the dataset may be used for research publications. Approval to use the data does not allow recipients to redistribute it and they must adhere to the terms of confidentiality restrictions. The license agreement details the permissible use of the data and the appropriate citation, it can be found at: http://www.affectiva.com/facial-expression-dataset-am-fed/. Use of the dataset for commercial purposes is strictly prohibited.

9. Conclusions and Future Work

The main contribution of this paper is to present a first in the world publicly available dataset of labeled data recorded over the Internet of people naturally viewing online media. The AM-FED contains, 1) 242 webcam videos recorded in real-world conditions, 2) 168,359 frames labeled for the presence of 10 symmetrical FACS action units, 4 asymmetric (unilateral) FACS action units, 2 head movements, smile,

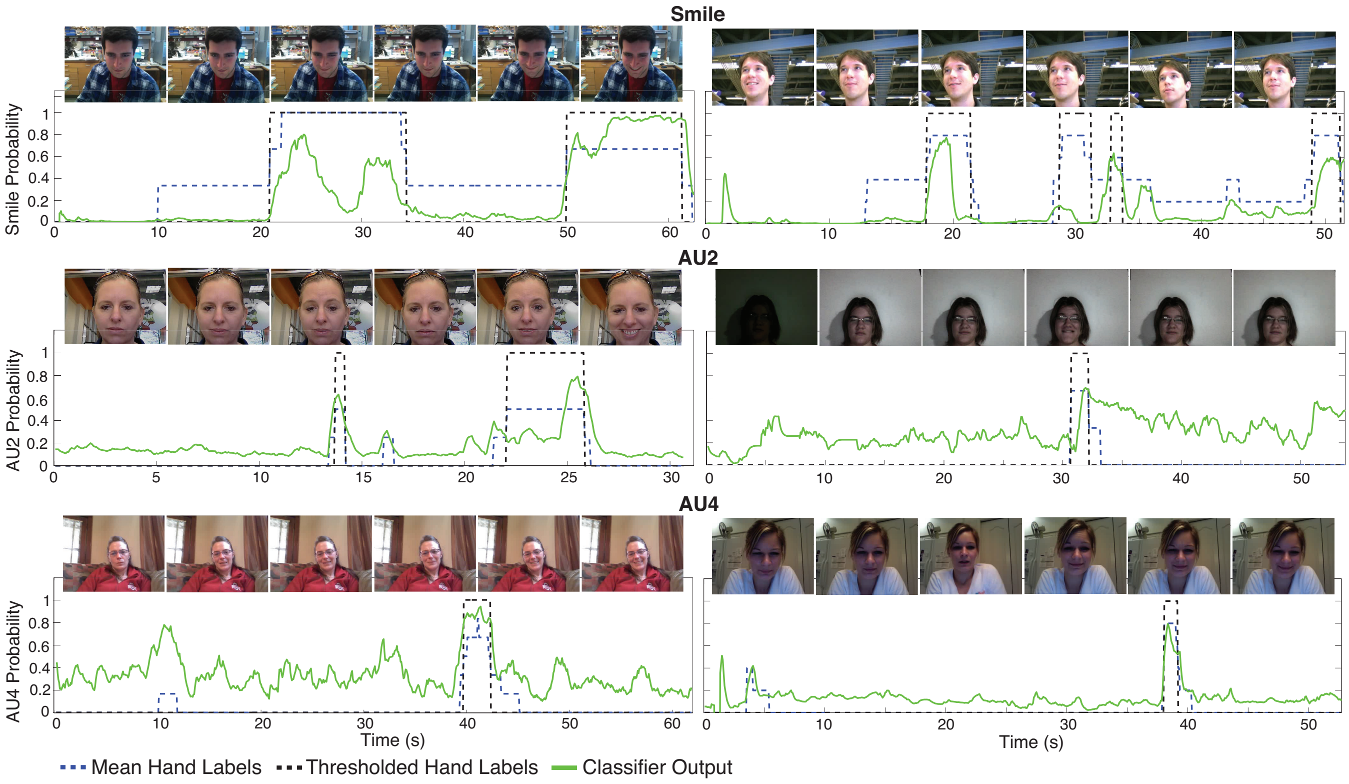


Figure 9. Example comparisons between classifier predictions (green) and manually coded labels (blue and black dashed) for six videos within the dataset. Threshold of hand labels based on >= 0.5 agreement between coders. Frames from the sequences are shown above. Top) Smile classification example, middle) AU2 classification example, bottom) AU4 classification example. The viewer's distance from the camera, their pose and the lighting varies considerably between videos.

general expressiveness, feature tracker fails and gender, 3) locations of 22 automatically detect landmark points, 4) baseline performance of detection algorithms on this dataset and baseline classifier outputs for smile. 5) self-report responses of familiarity with, liking of and desire to watch again for the stimuli videos. This represents a rich and extensively coded resource for researchers working in the domains of facial expression recognition, affective computing, psychology and marketing.

The videos in this dataset were recorded in real-world conditions. In particular, they exhibit non-uniform framerate and non-uniform lighting. The camera position relative the viewer varies from video to video and in some cases the screen of the laptop is the only source of illumination. The videos contain viewers from a range of ages and ethnicities some with glasses and facial hair.

The dataset contains a large number of frames with agreed presence of facial action units and other labels. The most common are smiles, AU2, AU4 and AU17 with over 1,000 examples of these. The videos were coded for the presence of 10 symmetrical FACS action units, 4 asymmetric (unilateral) FACS action units, 2 head movements, smile, general expressiveness, feature tracker fails and gender. The rater reliability (calculated using the Spearman-Brown reliability metric) was good for a majority of the actions. However, in cases where there were only a few examples of a particular action the rate reliability metrics suffered. The labels with the greatest reliability were smile = 0.78, AU4 = 0.72 and expressiveness = 0.71. The labels with the lowest reliability was unilateral AU14 (unilateral) and AU10. This is understandable as the unilateral labels are challenging especially in frames where the lighting is non-uniform in which case the appearance of an asymmetric expression can be amplified. AU10 is also relatively rare, only 26 frames with majority agreed presence, and these come from only one video sequences. Therefore small differences in coders agreement might cause the reliability to be low.

We calculate baseline performance for smile, AU2 and AU4 detection on the dataset, the area under the ROC curves were 0.90, 0.72 and 0.70 respectively. This demonstrates that accurate facial action detection is possible but that there is room for improvement as there are a number of challenging examples. In addition, the labels provide the possibility of testing many other AU classifiers on real-world data.

We hope that the release of this dataset will encourage researchers to test new action unit detection, expression detection and affective computing algorithms on challenging data collected "in-the-wild". We hope that it will also serve as a benchmark, enabling researchers to compare the performance of their systems against a common dataset and that this will lead to greater performance for state-of-the-art systems in challenging conditions.

Acknowledgments

Richard Sadowsky, Oliver Wilder-Smith, Zhihong Zeng, Jay Turcot and Khoulood Ayman provided support with the crowdsourcing system and labeling.

References

[1] A. Ashraf, S. Lucey, J. F. Cohn, T. Chen, Z. Ambadar, K. Prkachin, P. Solomon, and B. Theobald. The painful face: pain expression recognition using active appearance models. In *Proceedings of the 9th international conference on Multimodal interfaces*, pages 9–14. ACM, 2007. 1

[2] M. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan. Automatic recognition of facial actions in spontaneous expressions. *Journal of Multimedia*, 1(6):22–35, 2006. 1, 2

[3] J. F. Cohn, Z. Ambadar, and P. Ekman. *Observer-based measurement of facial expression with the Facial Action Coding System*. Oxford: NY, 2005. 2

[4] J. F. Cohn, T. Kruez, I. Matthews, Y. Yang, M. Nguyen, M. Padilla, F. Zhou, and F. De la Torre. Detecting depression from facial actions and vocal prosody. In *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*, pages 1–7. IEEE, 2009. 1

[5] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. Ieee, 2005. 6

[6] M. Eckhardt and R. Picard. A more effective way to label affective expressions. In *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*, pages 1–2. IEEE, 2009. 4

[7] P. Ekman and W. Friesen. Facial action coding system. 1977. 2

[8] P. Ekman and E. Rosenberg. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA, 1997. 2

[9] R. Gross, I. Matthews, J. F. Cohn, T. Kanade, and S. Baker. Multi-pie. *Image and Vision Computing*, 28(5):807–813, 2010. 2

[10] R. Kaliouby and P. Robinson. Real-time inference of complex mental states from facial expressions and head gestures. *Real-time vision for human-computer interaction*, pages 181–200, 2005. 1

[11] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews. The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pages 94–101. IEEE, 2010. 1, 2

[12] P. Lucey, J. F. Cohn, K. Prkachin, P. Solomon, and I. Matthews. Painful data: The unbc-mcmaster shoulder pain expression archive database. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 57–64. IEEE, 2011. 1, 2

[13] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba. Coding facial expressions with gabor wavelets. In *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*, pages 200–205. IEEE, 1998. 2

[14] D. McDuff, R. El Kaliouby, K. Kassam, and R. Picard. Affect valence inference from facial action unit spectrograms. In *Computer Vision and Pattern Recognition Workshops, 2010 IEEE Computer Society Conference on*. IEEE. 1

[15] D. McDuff, R. El Kaliouby, and R. Picard. Crowdsourced data collection of facial responses. In *Proceedings of the 13th international conference on Multimodal Interaction*. ACM, 2011. 3, 6

[16] D. McDuff, R. El Kaliouby, and R. Picard. Crowdsourcing facial responses to online videos. *IEEE Transactions on Affective Computing*, 3(4):456–468, 2012. 3, 5

[17] D. McDuff, R. El Kaliouby, and R. W. Picard. Predicting online media effectiveness based on smile responses gathered over the internet. In *Automatic Face & Gesture Recognition, 2013 IEEE International Conference on*. IEEE, 2013. 5

[18] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder. The semaine database: annotated multimodal records of emotionally colored conversations between a person and a limited agent. *Affective Computing, IEEE Transactions on*, 3(1):5–17, 2012. 2

[19] D. Messinger, M. Mahoor, S. Chow, and J. F. Cohn. Automated measurement of facial expression in infant–mother interaction: A pilot study. *Infancy*, 14(3):285–305, 2009. 1

[20] A. J. O'Toole, J. Harms, S. L. Snow, D. R. Hurst, M. R. Pappas, J. H. Ayyad, and H. Abdi. A video database of moving faces and people. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(5):812–816, 2005. 3

[21] R. Rosenthal. Conducting judgment studies: Some methodological issues. *The handbook of methods in nonverbal behavior research*, pages 199–234, 2005. 5

[22] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from single depth images. In *CVPR*, volume 2, page 3, 2011. 1

[23] M. Valstar and M. Pantic. Induced disgust, happiness and surprise: an addition to the mmi facial expression database. In *Proceedings of the 3rd International Workshop on EMOTION: Corpora for Research in Emotion and Affect*, page 65, 2010. 1, 2

[24] J. Whitehill, G. Littlewort, I. Fasel, M. Bartlett, and J. Movellan. Toward practical smile detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(11):2106–2111, 2009. 1, 2, 6