Crowdsourced data collection of facial responses

Crowdsourced Data Collection of Facial Responses

Daniel McDuff
MIT Media Lab
Cambridge
02139, USA
djmcduff@media.mit.edu

Rana el Kaliouby *
MIT Media Lab
Cambridge
02139, USA
kaliouby@media.mit.edu

Rosalind Picard *
MIT Media Lab
Cambridge
02139, USA
picard@media.mit.edu

ABSTRACT

In the past collecting data to train facial expression and affect recognition systems has been time consuming and often led to results that do not include spontaneous expressions. We present the first crowdsourced data collection of dynamic, natural and spontaneous facial responses as viewers watch media online. This system allowed a massive corpus of 3,268 videos to be collected in under two months.

We characterize the data in terms of viewer demographics, position, scale, pose and movement of the viewer within the frame, and illumination of the facial region. We compare statistics from this corpus to those from the CK+ and MMI databases and show that distributions of position, scale, pose, movement and luminance of the facial region are significantly different from those represented in these traditionally used datasets.

We demonstrate that it is possible to efficiently collect massive amounts of ecologically valid responses, to known stimuli, from a diverse population using such a system. In addition facial feature points within the videos can be tracked for greater than 90% of the frames. These responses were collected without need for scheduling, payment or recruitment. Finally, we describe a subset of data (over 290 videos) that will be available for the research community.

Categories and Subject Descriptors

I.5.4 [**Image Processing and Computer Vision**]: Applications - Computer vision, Signal processing; I.5.5 [**Image Processing and Computer Vision**]: Interactive systems; J.4 [**Social and Behavioral Sciences**]: Psychology

General Terms

Design, Human Factors, Measurement, Performance

*Dr. Kaliouby and Dr. Picard also hold positions with Affectiva, Inc.

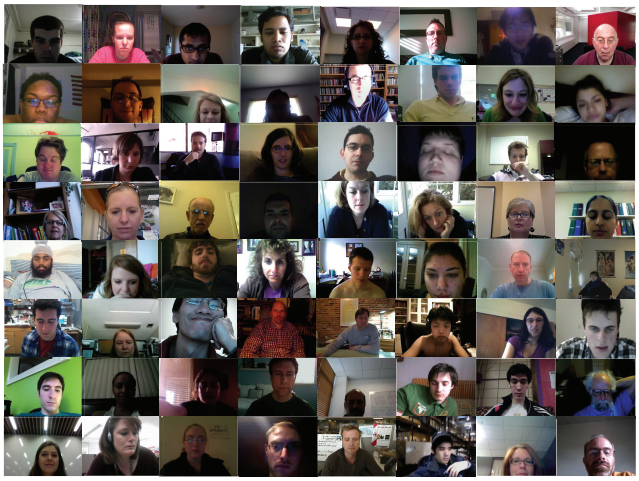

Figure 1: A sample from the 3268 videos collected. There are significant variations in position, scale, pose, lighting and movement in the responses. These represent a subset of the public data.

Keywords

Facial responses, non-verbal communication, crowdsourcing.

1. INTRODUCTION

Computer-based machine learning and pattern analysis depends hugely on the number of training examples. 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. In addition these datasets often feature posed expressions and/or are collected in an artificial setting of a laboratory or studio. It has been shown in numerous studies that spontaneous expressions differ from those that are deliberate [19] and that they can communicate different information [27]. Further to this, it has been shown that in day-to-day life people exhibit complex and subtle affective states [1] such as worry and concentration [18].

There are cultural and gender difference in non-verbal responses [11, 8]. In designing systems that are able to detect non-verbal cues and affective states it is therefore necessary to collect data that reflects these nuances and contains a significant amount of examples across these different categories. Until now datasets have been limited to a relatively small number of examples due to the difficulty and expense involved in collection. Similarly, the populations from which

these data are collected are often limited and do not reflect intra- and inter-person variability across different cultures, genders, ages and personalities.

This work was motivated by the desire to collect a large dataset of natural and spontaneous interactions. We present the first findings from an experiment to crowdsource data of facial responses over the internet. Until now it has not been clear what quality of data could be collected via crowdsourcing, what the natural limits on pose, lighting and image quality might be and how feature trackers would perform.

The main contribution of this paper is presenting the first-in-the-world crowdsourced facial expression corpus. We believe that this method of collecting and analyzing facial video can truly accelerate research in automated understanding of facial expressions and gestures. The method can also provide a mechanism to ask entirely new research questions, and to answer those questions with data that is ecologically valid. We present a massive dataset, collected via the internet over just 54 days, containing 3,268 videos captured in natural environments whilst the viewers were presented with known stimuli, one of three commercials.

We compare the demographics, position, scale, pose, movement and lighting for these data collected with sets of videos from the CK+ [10] and the MMI [16] databases, datasets traditionally used for training and testing facial expression recognition systems. We contrast the dynamic ranges of the respective features across each of the datasets.

2. RELATED WORK

The internet provides the ability to crowd-source lots of useful information [17]. People are willing to engage and share visual images from their webcams [21] and these can be used for training automatic algorithms for learning. Inspired by these approaches, we capture videos of natural engagement with media online and show that this can be elicited without payment, providing motivation for the viewers by combining the experiment with popular and engaging media shown during recent Super Bowl television coverage.

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 is tedious work - requiring a lot of effort which many researchers may not have the resources to do. In the area of facial expression analysis, the Cohn-Kanade database, in its extended form named CK+, played a key role in advancing the state of the art in this area. The CK+ database, contains 593 recordings of posed and non-posed sequences. The sequences are recorded under controlled conditions of light and head motion, and range between 9-60 frames per sequence. Each sequence represents a single facial expression that starts with a neutral frame and ends with a peak facial action. Transitions between expressions are not included. Several systems use the CK, or CK+, databases for training and/or testing. Since it was first published, a number of papers have been published that were trained and/or tested on this data set including: Bartlett et al. [2], Cohen et al. [5], Cohn et al. [6], Littlewort et al. [9] and Michel & El Kaliouby [14]. Since then, a few other databases have emerged, including: MMI [16], SEMAINE [13], RU-FACS [3], SAL [7]. A survey of databases and affect recognition systems can be found in [25]. However, there is a need for mechanisms to quickly and efficiently collect numerous examples of natural and spontaneous responses. Lab-based studies pose numerous challenges including recruitment, scheduling and payment. Efforts have been made to collect significant amounts of spontaneous facial responses, however the logistics of a laboratory based study typically limits the number of participants to under 100, e.g. 42 in [12]. By using the internet we can make data collection efficient, asynchronous and less resource intensive, and get at least an order of magnitude more participants.

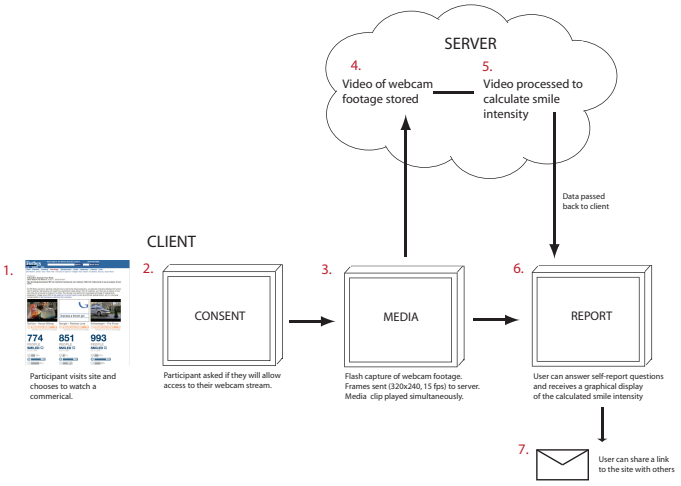

Figure 2: Overview of what the user experience was like and Affectiva's (www.affectiva.com) web-based framework that was used to crowdsource the facial videos. From the viewer's perspective, all that is needed is a browser with Flash support and a webcam. The video from the webcam is streamed in real-time 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.

3. CROWDSOURCING PLATFORM

Figure 2 shows the web-based framework that was used to crowdsource the facial videos and provides an overview of 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 15 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). All three ads were somewhat amusing and were designed to elicit smile or laughter responses.

On selecting a commercial to watch, visitors are asked to

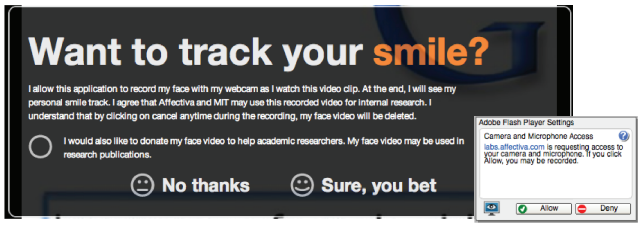

Figure 3: The consent forms that the viewers were presented with before watching the commercial and before the webcam stream began.

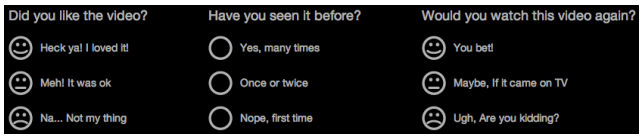

Figure 4: The self-report questions the viewers were presented with after watching the commercial.

1) grant access to their webcam for video recording and 2) to allow Affectiva and MIT 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 publically are shown in this paper. This data collection protocol was approved by the Massachusetts Institute of Technology 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 3. 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 responses (if any) to the self report questions. No other data is stored.

Following each commercial, the webcam is automatically stopped and a message clearly states that the "webcam has now been turned off". Viewers could then 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 4. Finally, viewers were provided with a graphical representation of their smile intensity during the clip compared to other viewers who watched the same video; viewers were also given the option to tweet their result page or email it to a friend. All in all, it took under 5 seconds to turn around the facial analysis results once the video was completed so viewers perceived the results as instantaneous. Viewers were free to watch one, two or three videos and could watch a video as many times as they liked. In this paper we focus on the general characteristics of the collected videos (e.g., pose and lighting) and leave the analysis of the facial and self-report responses to future work as there is not space to discuss them fully here.

4. THE DATASET

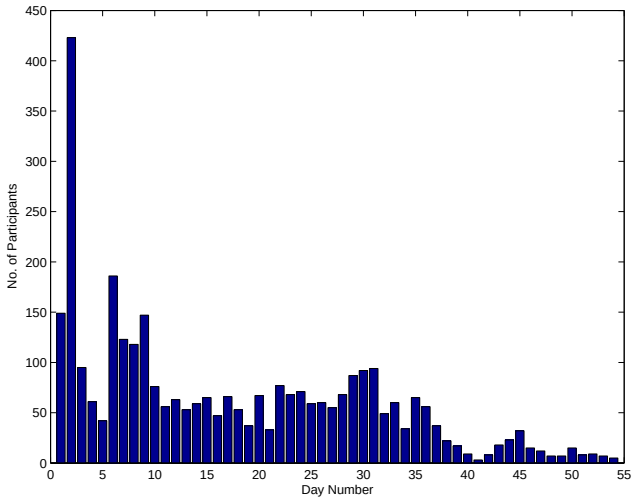

Figure 5: Histogram of the number of viewers that successfully completed the study on each of the 54 consecutive days (from 3/3/2011) that it was live.

Figure 6: Map showing the location of the 3268 viewers, based on their IP address. No viewers IP were located outside of the latitudes shown.

Using the framework described we collected 3,268 videos (2,615,800 frames) over a period of 54 days from 03/03/2011 to 04/25/2011. The application was promoted on the Forbes website - http://www.forbes.com/2011/02/28/detect-smile-webcam-affectiva-mit-media-lab.html. We refer to the data collected as the Forbes dataset. The number of visitors who clicked a video was 16,366. Of these 7,562 (46.2%) opted-in to allow webcam access. 5,268 (32.2%) completed the experiment. For the analysis here we disregard videos for which we were unable to identify a face within at least 90% of frames as they were significantly beyond the performance of the Nevenvision tracker, this left 3,268 videos (20.0%). Figure 5 shows the number of these videos that were completed on each of the 54 days. All videos were recorded with a resolution of 320x240 and a frame rate of 15 fps.

As this is the first experiment to collect this kind of data in the wild we compared these data to examples from other datasets collected in laboratories. We compare the statistics for these data collected with sets of videos from the CK+ and MMI databases, data traditionally used for training and testing facial expression and affect recognition systems.

For position, scale, pose, movement and illumination analysis we took all 722 videos from the MMI database that featured participants filmed with a frontal pose (14,360 frames) and all 593 videos from the CK+ dataset (10,708 frames).

Table 1: Table showing the number of videos for each commercial broken down by continent and gender (no. of females shown in brackets).

	No. of viewers (female)		
Continent	Doritos	Google	VW
Africa	14 (4)	14 (8)	18 (8)
Asia	74 (22)	68 (20)	88 (24)
Europe	226 (75)	228 (65)	222 (61)
North America	681 (245)	730 (273)	714 (260)
South America	42 (13)	43 (15)	43 (12)
Oceania	23 (6)	21 (5)	19 (5)
Total	695 (365)	718 (386)	735 (369)

5. CHARACTERIZING THE DATA

5.1 Demographics

The following section concerns statistics about the demographic profiles of the data. We compare these statistics of the viewers of the Forbes study with the participants from the CK+ and MMI datasets.

We use IP information to provide statistics on the locations of viewers by finding the latitude and longitude corresponding to each address. Statistics for gender were obtained by a labeler who watched the videos. IP addresses have been shown to be a reliable measure of location [20]. The IP address geo-location was performed using IPInfoDB [1]. We could not guarantee that the same viewer would watch all three of the commercials or that some may watch them more than once. As we do not have identifiable information from the viewers and we do not have the number of distinct viewers who took part, only a coarse calculation can be provided by the number of distinct IP addresses 1,495 (45.8%). This suggests that on average each location successfully completed the task for two viewings. Table 1 shows the number of viewers in each continent and in brackets the number of females. A majority of the viewers were located in North America and Europe. The geographic location of each of the viewers is shown on the map in Figure 6. Of the 3,268 videos 1,120 (34.3%) featured females as the main subject. The age of viewers was restricted to those over the age of 13 or with a parent or legal guardian's consent. In 924 (28.3%) of the videos the viewer was wearing glasses.

The CK+ dataset was obtained from a subset of data collected in America featuring 210 adults aged 18 to 50 years of age, 69% females, 81% Euro-American, 13% Afro-American, and 6% others. In none of the videos was the participant wearing glasses.

The MMI dataset initially consisted of data from 19 participants (44% female) aged between 19 to 62 from European, Asian, or South America ethnic background. It has since been extended [22] with data from 25 more participants (48% female) aged between 20 and 32 years from similar ethnic backgrounds. In 249 (34.5%) of the 722 frontal videos the participant was wearing glasses.

5.2 Position, Scale and Pose

A facial feature tracker, the Nevenvision tracker[2], was used to automatically find 22 facial features within each frame of the videos. The locations of the points detected are shown in Figure 7. The following metrics need to be considered in the limitations of the facial feature tracker used. About three axes of pitch, yaw (turning) and roll (tilting), the limits are 32.6 (std=4.84), 33.4 (std=2.34) and 18.6 (std=3.75) degrees from the frontal position respectively (deviations reflect variability in performance in different lighting). We do not consider comparisons outside of these ranges.

The number of frames tracked for each of the datasets were as follows: Forbes 2,554,325 of 2,615,800 (97.7%), MMI 14,360 of 14,360 (100%) and CK+ 10,708 of 10,708 (100%)

Position within the frame was determined by position of feature point three, a fixed point at the nose root. We normalized these points to a relative position within a frame 320x240. Three Euler angles for the pose of the head, pitch, yaw and roll were calculated. The head scale within the frame was also calculated using the distance between the feature points; this can be approximated as an inverse measurement of the face from the camera.

Figure 8 shows a spatial representation of the mean loca-

Figure 7: Figure showing the location of the 22 feature points tracked by the Nevenvision tracker. The dashed line highlights the facial region using for evaluating illumination.

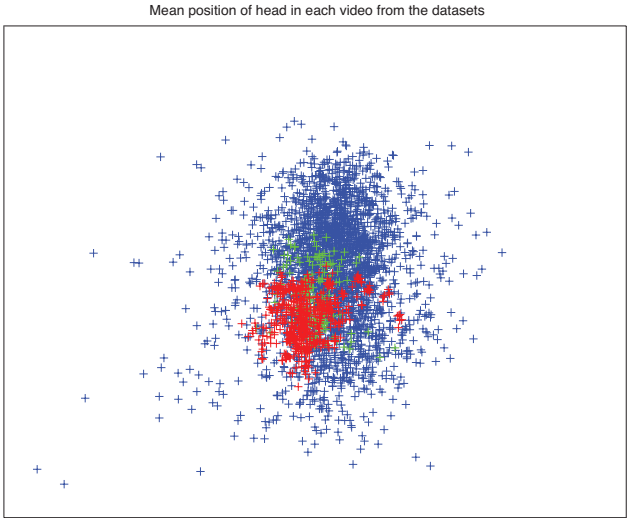

Figure 8: Scatter plot showing the mean location of the viewers' head within the frame for each video. The nose root was used as the location of the head. CK+ (red), MMI (green), Forbes (blue).

[1]http://www.ipinfodb.com/ip_location_api.php
[2]Licensed from Google, Inc.

tion of the faces within the videos. We can see that there is significantly greater variance in the positions of the participants in the Forbes videos than those in both the MMI and CK+ sets. The means of these distributions are similar and are close to the center of the frame. As we discarded any videos in which our tracker did not find a face for more than 90% of the frames it is possible that for some participants the webcam was pointed in the wrong direction and therefore they were not in the frame at all.

Figure 9 shows a histogram of the relative scales of the faces detected. Examples shown for scales of 0.5, 1 and 1.5. There is greater deviation in the scales for the Forbes set (std=0.256) than both the MMI (std=0.122) and CK+ (std=0.116) sets. The mean is significantly lower for the Forbes set (mean=0.987) versus the MMI (mean=1.39) and CK+ sets (mean=1.22), $p<0.05$. This suggests that in normal web interactions people sit further from the camera than represented in the lab and more significantly they fill less of the field of view of the camera. The average head scale was 19% smaller in the Forbes set relative to the scales in CK+.

Figure 10 shows a histogram of the pose angles (pitch, jaw and roll) for each of the detected faces, as calculated using the Nevenvision tracker. As with the position and scale distributions, the variance in the Forbes dataset is greater for all three angles compared to the other two sets. The overriding factor affecting the pose could be the position of the camera relative to the participant and not necessarily the participant's movement within each video. The greatest difference is in the case of head yaw, for which there is a standard deviation of 0.143 for the Forbes set and 0.0668 and 0.0579 for the MMI and CK+ sets.

5.3 Movement

Movement within the videos was evaluated by two methods. Firstly, sparse optical flow between frames was calculated using the Shi and Tomasi and the Pyramidal Lucas-Kanade [4] algorithms, to quantify the movement within the whole frame. Secondly, tracked motion of the head was used to compute the viewer head motion. The movement of the viewer within the frame was calculated as the absolute difference between the subsequent positions locations of the viewers head (nose root) for each second of the video.

Figure 11 shows the distribution of the magnitudes of the optical flow features across the frames for each dataset (bottom). Due to the extra head and body movement and the less intense expressions of the participants in the Forbes dataset the distributions are a lot less defined. Examples of the optical flow in videos for each of the datasets are also shown (top). Motion in specific regions of the face, in particular around the mouth and eyes, are identifiable for the examples from the CK+ and MMI datasets but not for the Forbes set as overall head motion dominates. This was typical for the data.

Figure 12 shows the mean movement trajectories for each of the stimuli. All three have the same form with considerably greater movement at the start of the clips. This is movement relative to the camera and therefore much of this movement could be due to viewers adjusting the direction of their webcam at the start and not their motion, this was observed in a number of the videos. The examples from the CK+ and MMI databases are relatively very short in duration and the participants move only slightly, therefore they do not reflect natural movement that could be expected.

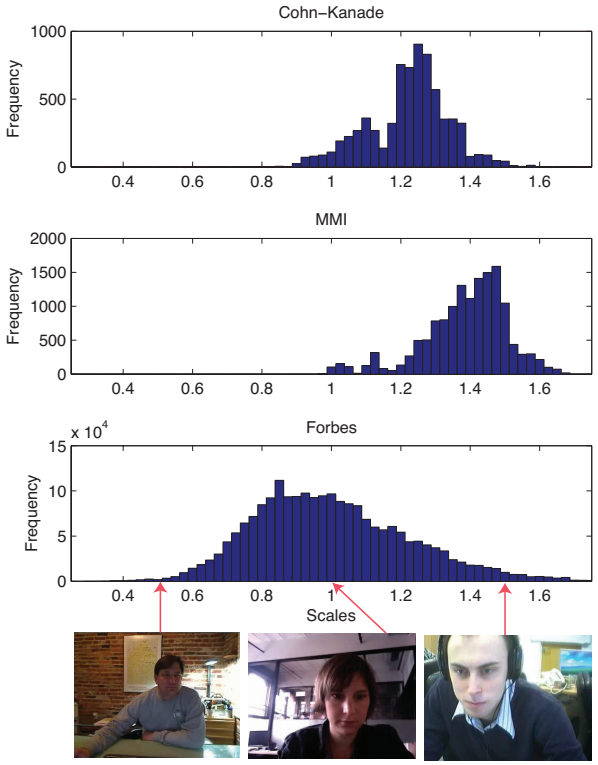

Figure 9: Histogram of head scales for the CK+ (top), MMI (center) and Forbes (bottom) datasets. The head scale was calculated for every frame in which a head was tracked. Examples of face scales of 0.5, 1 and 1.5 are shown below.

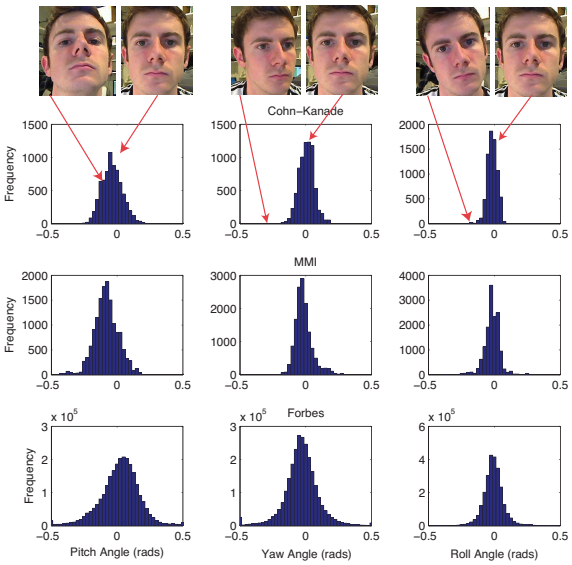

Figure 10: Histograms showing the pose angles of the heads in the CK+ (top), MMI (center) and Forbes (bottom) datasets. Examples of poses with pitch=-0.13 rads, jaw=-0.26 rads and roll=-0.19 rads are shown.

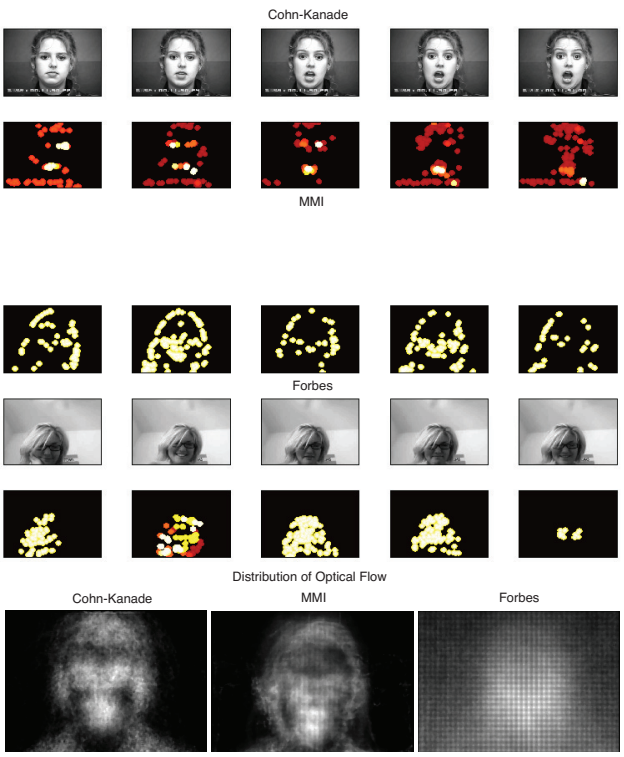

Figure 11: Distribution of optical flow across the frame for the CK+, MMI and Forbes datasets (bottom) and an example of the flow in one of the videos from each dataset (top). CK+ images ©JeffCohn.

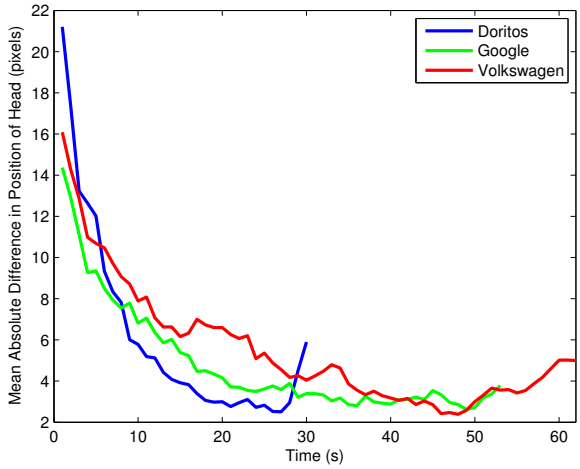

Figure 12: Mean absolute difference of the position of the viewer's heads (pixels) for each second during the videos. The data is divided into responses to each of the stimuli.

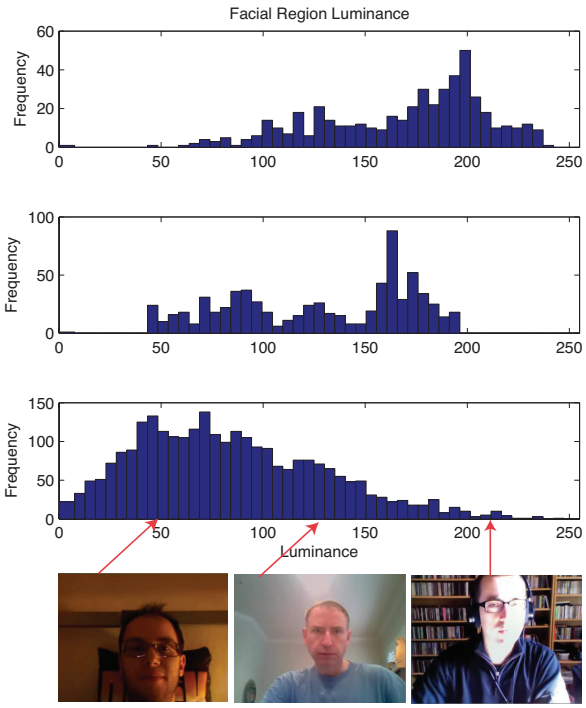

Figure 13: Histograms of the average luminance for the facial region for CK+ (top), MMI (center) and Forbes (bottom) datasets. Examples are shown for luminance values of 50, 125 and 216.

5.4 Illumination

We calculated image quality metrics for the first tracked frame of each video. The metrics were calculated for the facial region within the image, using a box with corners defined by the rigid feature points 4, 10 and 12:

$$\begin{aligned} \text{UL} &= [12.X - 0.2 * (10.X - 12.X), 12.Y + (12.Y - 4.Y] \\ \text{LR} &= [12.X + 0.2 * (10.X - 12.X), 4.Y - (12.Y - 4.Y)] \end{aligned}$$

This region is also shown graphically in figure 7. Two illumination metrics were used. Since the CK+ set were grayscale images we evaluate only grayscale metrics.

Luminance was calculated as the average pixel intensity for the facial region. As this only indicates the brightness of the facial region we also compute the contrast of the region.

Figure 13 shows histograms of the mean luminance calculated for the facial region of the first tracked frame of each of the videos. There is a significant difference between the mean luminance in the videos from the Forbes database. The mean is significantly lower for the Forbes set (mean=84.3) versus the MMI (mean=128) and CK+ sets (mean=168), $p<0.05$. However, the deviation in the average luminance for the Forbes (std=45.2), MMI (std=44.0) and CK+ (std=41.1) sets were all similar.

Michelson contrast [15] was calculated using the maximum and minimum pixel luminance values in the facial region. The formula for Michelson Contrast is shown in 1.

$$\text{Contrast} = \frac{L_{max} - L_{min}}{L_{max} + L_{min}} \quad (1)$$

Where L_{max} and L_{min} are the maximum and minimum val-

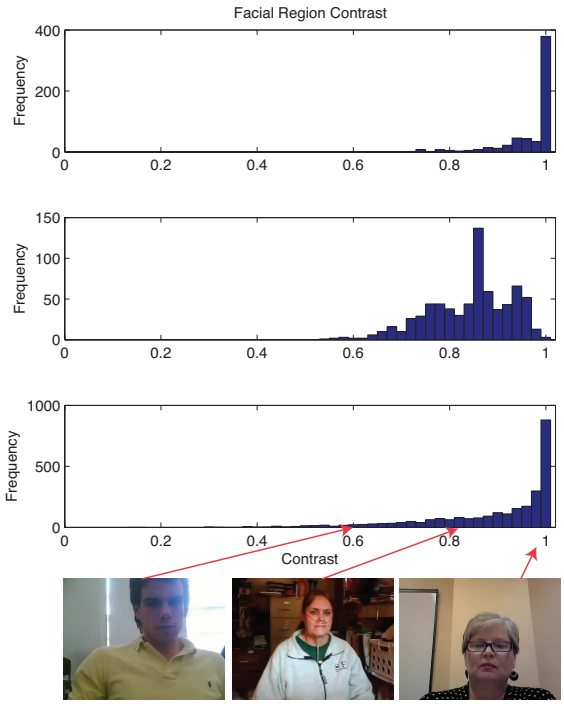

Figure 14: Figure showing histograms of the Michelson contrast for the facial region for CK+ (top), MMI (center) and Forbes (bottom) datasets. Examples are shown for luminance values of 0.60, 0.82 and 1.0.

ues of the luminance within the facial region of the image under consideration.

Figure 14 shows histograms of the Michelson contrast calculated for the facial region of the first tracked frame in each of the videos. There is a marked difference between the contrast in the videos from the MMI database compared to those from the CK+ and Forbes databases. The contrast seems to be stronger on average for the Forbes set compared to the MMI set. Whether this means that the facial features will be more defined is not clear. However, it demonstrates that the current datasets do seem to cover realistic ranges of contrast even though they differ significantly in luminance.

6. CONCLUSIONS

We have presented results from the first crowdsourced collection of natural and spontaneous facial responses over the web. The framework allows very efficient collection of examples of natural and spontaneous responses from a large and varied population. We collected 5,268 videos, of which 3,268 were trackable in over 90% of the frames, over 54 days from locations across the world. These responses are aligned with stimuli that were simultaneously presented to the participants. The method did not require payment or recruitment or the viewers but rather used popular media to motivate opt-in participation.

We have shown that there are marked differences between the position, scale and pose of participants in these natural interactions compared to those in datasets traditionally used for training expression and affect recognition systems,

Figure 15: Examples of some of the challenges involved in working with crowdsourced data: variable lighting, contrast, color, focus, position, pose, occlusions and number of viewers.

the MMI and CK+ datasets. In particular we showed that position along the vertical axis of the frame, scale of the face within the field of view of the camera and jaw of the head had significantly different distributions to those in traditional lab-based datasets in which these degrees-of-freedom are often constrained. The results suggest that we need to include significantly more examples that accurately represent the full extent of these ranges in data used for training and testing systems that might be used in the wild. The average head scale was 19% smaller for the Forbes set compared to the other datasets. In addition we identify that there is much greater head and body movement in the data collected when compared to other data sets. Facial expressions occurred simultaneously with these movements and gestures which were often larger than the movements due to the facial expressions. Movement relative to the camera was greatest at the beginning and end of the clips.

Similarly, we identified a statistically significant difference between the average luminance within the facial region between the Forbes dataset and the CK+ and MMI sets, although the variance of the luminance and the distributions of contrast were not significantly different.

Although, these data demonstrate that the dynamic range of viewer position, pose, movement and illumination are greater than those represented in existing datasets we have shown that we were able to collect thousands of trackable videos via the crowdsourcing platform. This presents a lot of promise for obtaining data for training and testing future algorithms.

One obstacle that remains in the collection of large datasets such as this is how to obtain groundtruth labels. If this is to be done by manual coding it could be very time consuming. However, by utilizing other crowdsourcing methods such as the service provided by Amazon's Mechanical Turk it could be feasible to label vast amounts of examples efficiently [23]. Although the labelers may not be "experts" there are several methods for calculating the reliability of such labels [24]. Methods of unsupervised learning for non-verbal data have also been demonstrated [26].

7. FUTURE WORK

Over 290 viewers opted to share their response videos with the research community and to allow them to be used in publications, we will make these available soon. We plan to add supplementary facial response labels to these videos.

The data present many challenges from a computer vision perspective in terms of dealing with extreme lighting conditions, occlusions and subtle facial expressions. Some

challenging examples are shown in Figure 15.

8. ACKNOWLEDGMENTS

To be included in final version. Richard Sadowsky, Oliver Wilder-Smith and Affectiva provided generous support in providing access to the crowdsourcing platform. Brian Staats provided a great front end design for the site. Jon Bruner and Forbes kindly promoted the work on their site. This work was funded by the Media Lab Thing's Consortium and Proctor and Gamble.

9. REFERENCES

[1] S. Baron-Cohen. *Mind reading: the interactive guide to emotions*. Jessica Kingsley Publishers, 2003.

[2] M. Bartlett, G. Littlewort, I. Fasel, and J. Movellan. Real Time Face Detection and Facial Expression Recognition: Development and Applications to Human Computer Interaction. 2003.

[3] 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.

[4] J. Bouguet. Pyramidal implementation of the affine lucas kanade feature trackerâ description of the algorithm. Technical report, Technical report). Intel Corporation, 2001.

[5] I. Cohen, N. Sebe, A. Garg, L. Chen, and T. Huang. Facial expression recognition from video sequences: temporal and static modeling. *Computer Vision and Image Understanding*, 91(1-2):160–187, 2003.

[6] J. Cohn, L. Reed, Z. Ambadar, J. Xiao, and T. Moriyama. Automatic analysis and recognition of brow actions and head motion in spontaneous facial behavior. In *Systems, Man and Cybernetics, 2004 IEEE International Conference on*, volume 1, pages 610–616. IEEE, 2004.

[7] E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. Mcrorie, J. Martin, L. Devillers, S. Abrilian, A. Batliner, et al. The humaine database: Addressing the collection and annotation of naturalistic and induced emotional data. *Affective computing and intelligent interaction*, pages 488–500, 2007.

[8] J. Hall, J. Carter, and T. Horgan. 5. Gender differences in nonverbal communication of emotion. *Gender and emotion: Social psychological perspectives*, page 97, 2000.

[9] G. Littlewort, M. Bartlett, I. Fasel, J. Chenu, and J. Movellan. Analysis of machine learning methods for real-time recognition of facial expressions from video. In *Computer Vision and Pattern Recognition*. Citeseer, 2004.

[10] P. Lucey, J. 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.

[11] D. Matsumoto. Cultural similarities and differences in display rules. *Motivation and Emotion*, 14(3):195–214, 1990.

[12] 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 (CVPRW), 2010 IEEE Computer Society Conference on*, pages 17–24. IEEE.

[13] G. Mckeown, M. Valstar, R. Cowie, and M. Pantic. The semaine corpus of emotionally coloured character interactions. In *Proceedings of IEEE Int'l Conf. Multimedia, Expo (ICME'10), Singapore*, pages 1079–1084, July 2010.

[14] P. Michel and R. El Kaliouby. Real time facial expression recognition in video using support vector machines. In *Proceedings of the 5th international conference on Multimodal interfaces*, pages 258–264. ACM, 2003.

[15] A. Michelson. *Studies in optics*. Dover Pubns, 1995.

[16] M. Pantic, M. Valstar, R. Rademaker, and L. Maat. Web-based database for facial expression analysis. In *2005 IEEE International Conference on Multimedia and Expo*, page 5. IEEE, 2005.

[17] A. Quinn and B. Bederson. Human Computation: A Survey and Taxonomy of a Growing Field. In *Proceedings of CHI*, 2011.

[18] P. Rozin and A. Cohen. High frequency of facial expressions corresponding to confusion, concentration, and worry in an analysis of naturally occurring facial expressions of Americans. *Emotion*, 3(1):68, 2003.

[19] K. Schmidt, Z. Ambadar, J. Cohn, and L. Reed. Movement differences between deliberate and spontaneous facial expressions: Zygomaticus major action in smiling. *Journal of Nonverbal Behavior*, 30(1):37–52, 2006.

[20] D. Svantesson. Geo-location technologies and other means of placing borders on the 'borderless' internet. *J. Marshall J. Computer & Info. L.*, 23:101–845, 2004.

[21] G. Taylor, I. Spiro, C. Bregler, and R. Fergus. Learning Invariance through Imitation.

[22] M. Valstar and M. Pantic. Induced disgust, happiness and surprise: an addition to the mmi facial expression database. In *The Workshop Programme*, page 65.

[23] C. Vondrick, D. Ramanan, and D. Patterson. Efficiently scaling up video annotation with crowdsourced marketplaces. *Computer Vision–ECCV 2010*, pages 610–623, 2010.

[24] P. Welinder and P. Perona. Online crowdsourcing: Rating annotators and obtaining cost-effective labels. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pages 25–32. IEEE, 2010.

[25] Z. Zeng, M. Pantic, G. Roisman, and T. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(1):39–58, 2009.

[26] F. Zhou, F. De la Torre, and J. Cohn. Unsupervised discovery of facial events. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

[27] M. Zuckerman, J. Hall, R. DeFrank, and R. Rosenthal. Encoding and decoding of spontaneous and posed facial expressions. *Journal of Personality and Social Psychology*, 34(5):966, 1976.