Vinereactor: Crowdsourced Spontaneous Facial Expression Data

Edward Kim
Department of Computing Sciences
Villanova University
800 E. Lancaster Ave, Villanova, PA
edward.kim@villanova.edu

Shruthika Vangala
Department of Computing Sciences
Villanova University
800 E. Lancaster Ave, Villanova, PA
svangal1@villanova.edu

ABSTRACT

Although machines are more pervasive in our everyday lives, we are still forced to interact with them through limited communication channels. Our overarching goal is to support new and complex interactions by teaching the computer to interpret the expressions of the user. Towards this goal, we present Vinereactor, a new labeled database for face analysis and affect recognition. Our dataset is one of the first to explore human expression recognition *in response to* a stimulus video, enabling a new facet of affect analysis research. Furthermore, our dataset is the largest of its kind, nearly a magnitude larger than its closest related work.

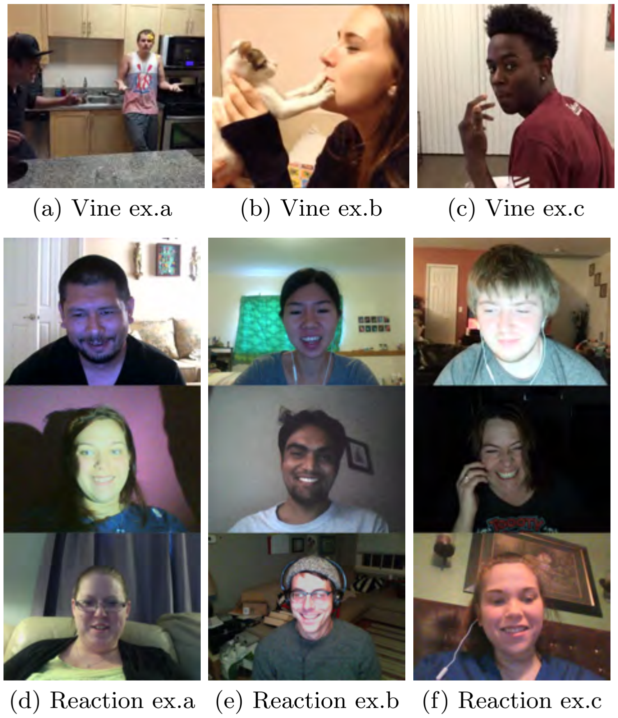
Keywords

affect recognition, emotion classification, face analysis, video database, crowdsourcing

1. INTRODUCTION

Basic face detection and recognition have made notable progress in the past couple years due to the advancements in machine learning, high resolution cameras, and availability of standard datasets. However, in order to further the communication capabilities between humans and machines, a computer must be able to incorporate affect recognition and understand how a person is feeling. Recognition and interpretation of human non-verbal communication cues can help inform a huge variety of computer aided applications, including pain medicine, online education, and marketing.

Towards this goal, we are researching novel ways to use computer vision algorithms to collect and analyze spontaneous facial expressions. In our application, we collected a random database of short internet video clips that have been categorized by genre (comedic, shocking, etc.) and have a high probability of eliciting an emotion response. We then play these "stimulus" videos for human observers and record their reactions, see Figure 1. This recording task was crowdsourced over Amazon Mechanical Turk and hundreds of gigabytes of video data were collected and analyzed using face analysis algorithms. Our spontaneous facial expression dataset has been annotated with facial landmark points and FACS labels (Facial Action Coding System). FACS are one of the most comprehensive methods of categorizing facial activity in a quantitative way [4]. The muscle activations in the face are labeled using action units (AUs) and the combination of certain AUs infer a particular emotion.


(a) Vine ex.a (b) Vine ex.b (c) Vine ex.c

(d) Reaction ex.a (e) Reaction ex.b (f) Reaction ex.c

Figure 1: A screenshot from three example stimulus videos are shown in (a)(b)(c). These videos are played to a set of crowdsourced observers whose reactions are recorded and analyzed (d)(e)(f).

Our data contributes to two major objectives. First, our large annotated dataset can be used to improve face analysis and emotion classification algorithms. We are freely distributing the largest annotated databases of spontaneous facial expression data to date. Second, the recorded spontaneous reaction data is linked with a database of stimulus videos. This opens up a whole new avenue of research where one could use our data to begin looking at the interaction between general videos and human reactions. For example, if we can quantify the emotion displayed by a human

ICMR'16, June 06 - 09, 2016, New York, NY, USA

ISBN 978-1-4503-4359-6/16/06...$15.00
DOI: http://dx.doi.org/10.1145/2911996.2912038

watching these videos, we can attempt to transfer the emotion classified from the observer to the video for indexing and retrieval. Furthermore, we can begin looking at general, non-face video emotion classification (stimulus videos) by knowing how people react. Currently, even given the state-of-the-art computer vision classification models, this idea of general video emotion classification is essentially an impossible problem to solve without the use of additional manual annotation. We hypothesize that our reaction data can provide the annotation data by just analyzing people's facial reactions instead of explicitly asking for their input.

2. BACKGROUND

A very large number of datasets for face analysis already exists. In this section, we will mention some of the most popular and most closely related datasets to our work in both image-based and video-based datasets.

2.1 Image & Video Datasets for Face Analysis

The CK+ dataset [10] - This popular dataset is fully FACS encoded with emotion labels and Active Appearance Models (AAMs) [2] face landmarks. The dataset contains 593 image sequences across 123 subjects.

The HELEN dataset [9] - The HELEN dataset consists of 2,000 training images and 330 test images with highly accurate, detailed and consistent annotations of facial components. This dataset is primarily an effort to build facial feature localization algorithms.

Multi-PIE dataset [5] - The Multi-PIE database is one of the largest databases, containing over 750,000 images with various camera angles, facial expressions, and illuminations. The data consists of 337 subjects with AAM style labels for 6,152 of the images. As a side note, this database is not freely available.

Labeled Faces in the Wild dataset [6] - This database contains 13,000 face photographs collected from the web. The faces contain a label of the name of the person pictured, which would be useful for unconstrained face recognition.

UNBC-McMaster (Pain AU) [11]- The UNBC-McMaster Shoulder Pain Archive (Pain AU) is one of the largest databases of AU coded videos of spontaneous facial expressions. There are 200 sequences across 25 subjects and all frames are FACS encoded with intensity labels.

YouTube Faces Database [13] - This dataset contains 3,425 videos of 1,595 different people. All the videos were downloaded from YouTube. An average of 2.15 videos are available for each subject.

BP4D [14] - The BP4D dataset is a 3D video database of spontaneous facial expressions in a diverse group of young adults. Well-validated emotion inductions were used to elicit expressions of emotion and paralinguistic communication.

UvA-NEMO [3] - This is a large-scale smile database which has 1240 smile videos (597 spontaneous and 643 posed) from 400 subjects. Ages of subjects vary from 8 to 76 years.

AM-FED [12] - This dataset contains 242 facial videos (168,359 frames) recorded in real world conditions. The data contains frame by frame FACS labels and 22 automatically detected landmark points. This work is the closest to ours where the data generated were people's responses to three internet advertisement videos.

This list is by no means comprehensive, but there are generalizations that we can highlight that differentiate our dataset from the existing image and video datasets for facial analysis. The first difference is that our dataset is the first to provide a database of stimulus videos *and* their associated human reactions. This combination is essential if we are to move beyond basic one-way analysis and begin analyzing complex interactions that humans have with machines and their environment. Second, our dataset is a magnitude larger than its closest related work, AM-FED, that has only three internet advertisements and 168,359 video frames. Our work has 200 general stimulus videos and over 1.3 million video response frames.

3. METHODOLOGY

Our process consisted of three major tasks. In the first task, we searched and downloaded random stimulus videos, known as vine videos. In our second task, we crowdsourced the collection of reactions to these videos, and in the final step, we analyzed the reaction videos to provide face landmark annotation and AU classifications.

3.1 Stimulus Videos

For the stimulus videos, we decided to use Vine videos because they were limited in size (maximum duration of 7 seconds) and manually categorized into different genres. We collected 200 random vine videos from the comedy vine.co[1] channel by scraping the featured videos from the website sporadically between May 2015 and September 2015. The videos have a duration range between 2 - 7 seconds and come with further associated meta data including url, number of likes, number of comments, and number of "revines".

3.2 Crowdsourced Reactions

Before the reaction data collection, we obtained ethics approval to conduct our research in the online crowdsourcing platform, Amazon Mechanical Turk. Each stimulus video was shown to a set of 30 unique viewers. The crowdsourcing task, HIT (human intelligence task), also asks for permission to activate the user's webcam and then instructs the user to center their face in the webcam's field of view. The worker can see their activated webcam (and face) next to an embedded video that requires an additional click to play the video. When they click to play the video, simultaneously, the stimulus video begins and their webcam begins to record their expressions. They are only able to watch the vine once, and for each 7-second vine video, we collected approximately 0.5 GB of total reaction video. After the video is complete, they are asked two questions. The first question is, how amusing is the video on a scale from 1-7, where 1 is not funny, 4 is fairly amusing, and 7 is extremely funny. The second question is related to giving consent to have their image, video, and likeness used for research purposes. This is a lengthy disclaimer describing our confidentiality procedures, and their rights as a participant in this study. It is explicitly stated that they do not have to give us permission to display and distribute their videos in order to be paid for the HIT. If permission is not granted, their data will not be released, distributed, or processed in any identifiable way. An optional comments section is available to the participant for providing any additional feedback. The workers are paid $0.10 USD for their participation. The total cost for data collection was $840.00 USD, which includes both the rewards and fees.

[1] https://vine.co/channels/comedy

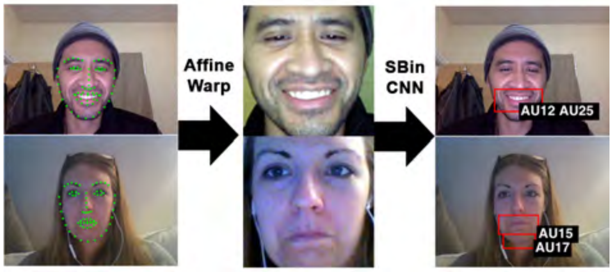


Figure 2: The face is first processed through a CLM [1] keypoint detection algorithm, then the cropped face area is affine warped to a base face shape and processed by the SBinCNN [8] algorithm to produce the AU activations.

3.3 Face Analysis

Keypoints - After collecting the reaction video, each reaction frame is processed to obtain the annotated 68 facial landmarks, see Figure 2. After an evaluation of several methods, we chose the CLM [1] framework which utilizes a constrained local neural field for landmark detection. In our experiments, it was the most robust in video annotation with variable lighting conditions, and had a relatively fast processing time ($<$ 0.5 seconds per frame on an Intel i7).

AU Labels - Given the 68 landmark annotations, we are able to move to the next step of our annotation process, i.e. action unit classification. For this process, we chose the SBinCNN [8] method as it had the highest f1 score in comparison to all the methods we tested. In the first step of this process, each face is aligned to a mean shape through an affine warp of the outer 27 keypoints of the face and 4 keypoints of the nose bridge. A full 68 keypoint warp was not used as it could distort the facial muscle activations. A contrast stretch is then applied to the image, and the resulting face is used for AU classification. This classification system is based upon a deep learning methodology. It uses a convolutional neural network training loss specific to AU intensity that utilizes a binned cross entropy method to fine-tune an existing network, e.g. CaffeNet [7]. The processing time for AU labeling is less than 5ms per image frame on a Tesla K40. The 10 AUs labeled in our dataset were the following, {1,2,4,6,7,12,15,17,25,26}. The most common AUs (and their approximate semantic meaning) that show up in our data are AU4 (brow lowering), AU12 (smile), AU15 (frown), AU17 (chin raised), and AU25 (lips apart).

4. RESULTS AND DATABASE STATISTICS

In this section, we present our analysis of the stimulus and reaction data, statistics of the video dataset, and failure cases which we will address in future work.

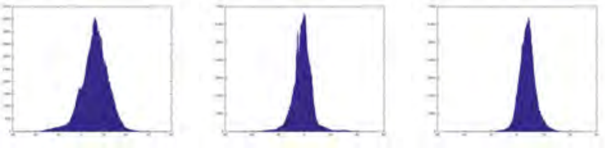
4.1 Stimulus and Reaction Videos

Stimulus Videos - We scraped 200 vine videos and used these videos as our stimulus to elicit reactions. The vine videos are stored in MP4 format and take up 280 MB of disk space. Each vine video has an associated XML file with the meta data from vine.co. Through our crowdsourcing process, we are also able to assign each video a numeric amusement value between 1-7 that corresponds to an average of 30 independent observers. The amusement values have an average standard deviation of 0.306 across all videos.

Reaction Videos-We collected 6,029 video responses from 343 unique mechanical turk workers in response to the 200 stimulus videos, with a minimum of 30 reaction videos per stimulus video. The total number of video frames in the database is 1,380,343 at a resolution of 320x240. All of the reactions are stored in JPG format and the total size of the reaction video data is 143 GB. See Figure 4 for an example of the stimulus video and processed reactions. Of the 6,029 video responses, 3455 (57.3%) of the reactions were approved to be used for research purposes and will be open and available for the community as long as the data is not redistributed or used for commercial applications. This public dataset includes 222 unique subjects with an average of 17.275 reactions per stimulus video.

4.2 Pitch, Yaw, Roll of Reactions

We further characterize the database by computing the distribution of the pitch, yaw, and roll of all the reaction faces and present them in Figure 3. The mean values for pitch, yaw, and roll are 6.23, -1.58, and 3.90 respectively. The variance is 35.26, 53.69, and 9.16, which shows more variability than datasets collected in controlled conditions. To put this in perspective, the UNBC-McMaster dataset [11] has variances of 23.58, 40.82, and 33.28.



(a) Pitch (b) Yaw (c) Roll

Figure 3: The distribution of yaw, pitch, and roll in our dataset, where the measurements are in degrees.

4.3 Failure Cases

There are three typical failure cases with our post processing pipeline. The first failure case is the failure of the webcam to record; this produces a completely black sequence of images. This failure is rare, accounting for less than 0.3% of the total dataset. The second failure case was the missed detection of a face by our face detection cascade (which includes a fast matlab implementation, [15], [1]). In these cases, the faces were not visible or below a face detection threshold due to various lighting effects, occlusion, or perspective transformations. These two failure cases account for 5.2% of the total dataset and will be discarded in the public dataset.

The final failure case is an incorrect or mis-alignment of the facial keypoints. This is a much more difficult failure case to detect and we note that our future work will be research into both automatic and crowdsourced validation of the machine annotated labels in our dataset.

5. CONCLUSION

In summary, we present a new database for face analysis and affect recognition. Our database was collected "in the wild" and is fully labeled with 68 facial landmarks and AU classifications. Our contribution to the community is two fold. Our dataset is one of the first that addresses the complex problem of stimulus and reaction video. Furthermore,

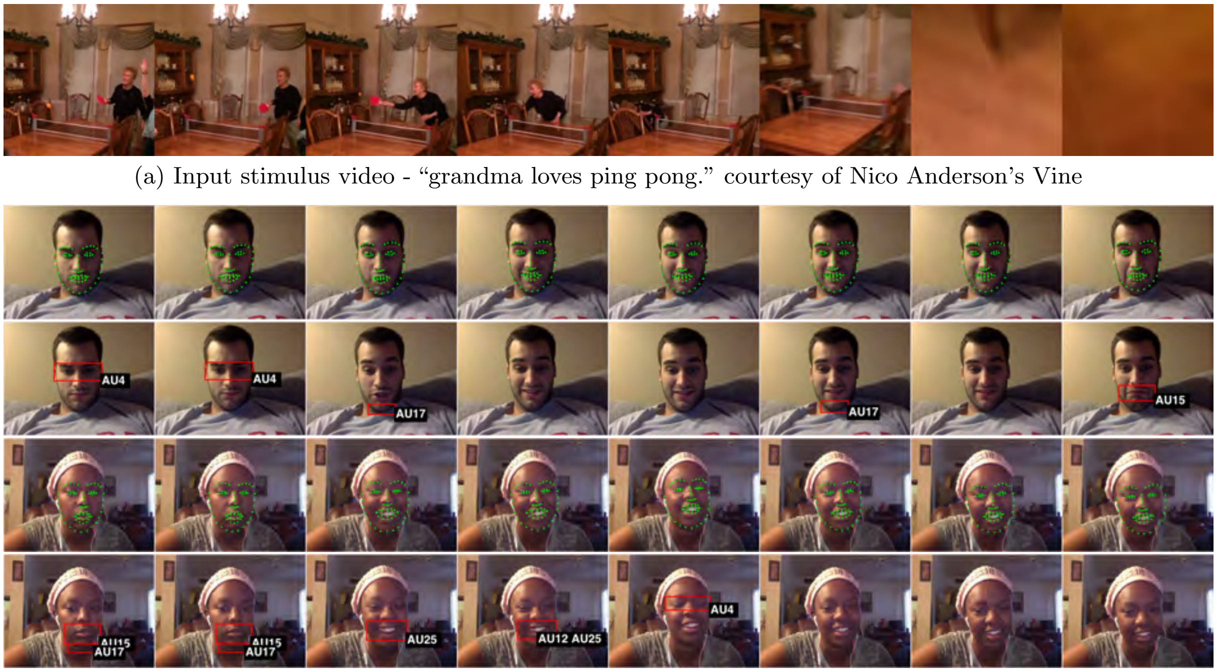

(a) Input stimulus video - "grandma loves ping pong." courtesy of Nico Anderson's Vine

Figure 4: An example of an input stimulus video shown to crowdsourced workers (a) and a subset of responses to the video with the computed 68 facial landmarks annotated in green and AU activations boxed in red.

our dataset is the largest of its kind, nearly one magnitude larger than its closest related work. We plan to share the outcomes created during this project including the data, as well as the algorithms used to classify the reaction faces.

6. ACKNOWLEDGMENTS

We thank all of the mechanical turkers for their participation in this research project. This work was supported by an AWS in Education Grant award.

7. REFERENCES

[1] T. Baltrusaitis, P. Robinson, and L.-P. Morency. Constrained local neural fields for robust facial landmark detection in the wild. In *International Conference on Computer Vision Workshops*, pages 354–361, 2013.

[2] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (6):681–685, 2001.

[3] H. Dibeklioğlu, A. A. Salah, and T. Gevers. Are you really smiling at me? spontaneous versus posed enjoyment smiles. In *ECCV*, pages 525–538. 2012.

[4] 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, 1997.

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

[6] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

[7] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ACM International Conference on Multimedia*, pages 675–678, 2014.

[8] E. Kim and S. Vangala. Deep action unit classification using a binned intensity loss and semantic context model. Technical report, Villanova University, 2016.

[9] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang. Interactive facial feature localization. In *ECCV*, pages 679–692. 2012.

[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*, pages 94–101, 2010.

[11] P. Lucey, J. 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*, pages 57–64, 2011.

[12] D. McDuff, R. Kaliouby, and R. W. Picard. Crowdsourcing facial responses to online videos. *Affective Computing*, 3(4):456–468, 2012.

[13] L. Wolf, T. Hassner, and I. Maoz. Face recognition in unconstrained videos with matched background similarity. In *CVPR*, pages 529–534, 2011.

[14] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, P. Liu, and J. M. Girard. Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing*, 32(10):692–706, 2014.

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