Event Detection: Ultra Large-scale Clustering of Facial Expressions

Thomas Vandal, Daniel McDuff and Rana El Kaliouby
Affectiva, Waltham, USA

***Abstract*— Facial behavior contains rich non-verbal information. However, to date studies have typically been limited to the analysis of a few hundred or thousand video sequences. We present the first-ever ultra large-scale clustering of facial events extracted from over 1.5 million facial videos collected while individuals from over 94 countries respond to one of more that 8000 online videos. We believe this is the first example of what might be described "big data" analysis in facial expression research. Automated facial coding was used to quantify eyebrow raise (AU2), eyebrow lowerer (AU4) and smile behaviors in the 700,000,000+ frames. Facial "events" were extracted and defined by a set of temporal features and then clustered using the k-means clustering algorithm. Verifying the observations in each cluster against human-coded data we were able to identify reliable clusters of facial events with different dynamics (e.g. fleeting vs. sustained and rapid offset vs. slow offset smiles). These events provide a way of summarizing behaviors that occur without prescribing the properties. We examined the how these nuanced facial events were tied to consumer behavior. We found that smile events - particularly those with high peaks - were much more likely to occur during viral ads. This data is cross-cultural, we also examine the prevalence of different events across regions of the globe.**

I. INTRODUCTION

The face is a rich channel for communicating non-verbal information and automated facial expression analysis has huge potential. Recently, it has been demonstrated that spontaneous facial responses can be collected efficiently and quickly on a large-scale using webcams and the cloud [1], [2]. Despite the variability in lighting, pose and image quality that exists in these videos, subtle behaviors can be detected [3]. Furthermore, we can make inferences from the observed behaviors, such as viewer preferences for online videos [4] and purchase intent towards advertised brands [5].

The most common taxonomy for coding facial behavior is the facial action coding system (FACS) [6]. Manual labeling of action units (AUs) is time consuming and requires specific training. It is often infeasible to hand-label all or even a subset of AUs. In this work we present analysis of over 700 million video frames of spontaneous data, namely facial events extracted from over 1.5 million facial videos collected while individuals from over 94 countries respond to one of more than 8000 online videos. It would have been impossible to hand code such a large dataset. Computer vision and machine learning techniques have been developed to alleviate these challenges via automatic recognition of facial behavior [7].

Although we can collect facial responses on a large-scale, understanding the types of behavior that exist in that dataset is a challenge. Unsupervised clustering allows us to identify similar groups of observations within data and characterize them. For instance, we can identify if there are some behaviors with faster onsets or offsets, or shorter durations. These may be linked to different emotional responses and resulting behavior.

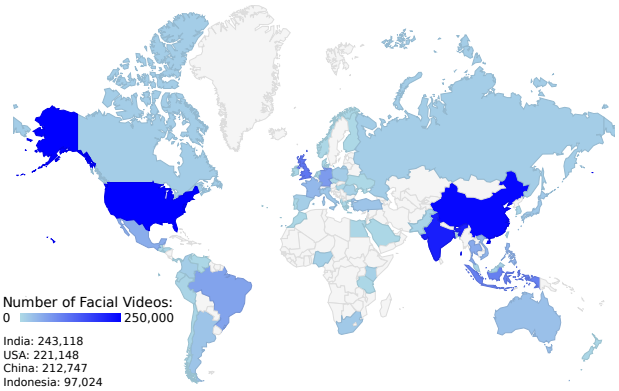

Fig. 1. Map showing the distribution across the world of facial videos in our dataset. India has the greatest number of videos (243,118) followed by the United States, China, Indonesia, and the United Kingdom. We have data from all continents. A facial video is typically 20 to 75 seconds long and recorded at 14 frames per second.

Furthermore, unsupervised clustering could make data labeling much more efficient [8], [9]. It is often only necessary to label a small number of representative samples of different types of observations in order improve performance considerably and clustering can help identify the most appropriate samples to be labeled. For instance, in sparse data, finding positive examples of an action is much more challenging than finding negative examples - our approach makes finding positive examples of different types of behavior much more efficient.

The aim of this work so two-fold: 1) To discover different types of facial events (e.g. smiles with different temporal patterns) from a vast dataset of naturalistic and spontaneous facial responses, 2) To understand the meaning of these events in the context of media effectiveness (specifically virality and video sharing). The large amount of data forced us to use technologies and methods which have not been necessary within the affective computing field to date. Hadoop[1] was used to retrieve and process our data in a reasonable time while making the solution scalable over any size dataset. This architecture is vital as our dataset of video responses is increasing all the time (from thousands of videos in 2011 [1] to millions in 2015). Figure 2 shows our approach.

[1]Hadoop is a tool for scalable data processing which allows one to divide a dataset and process it over multiple machines. http://hadoop.apache.org/

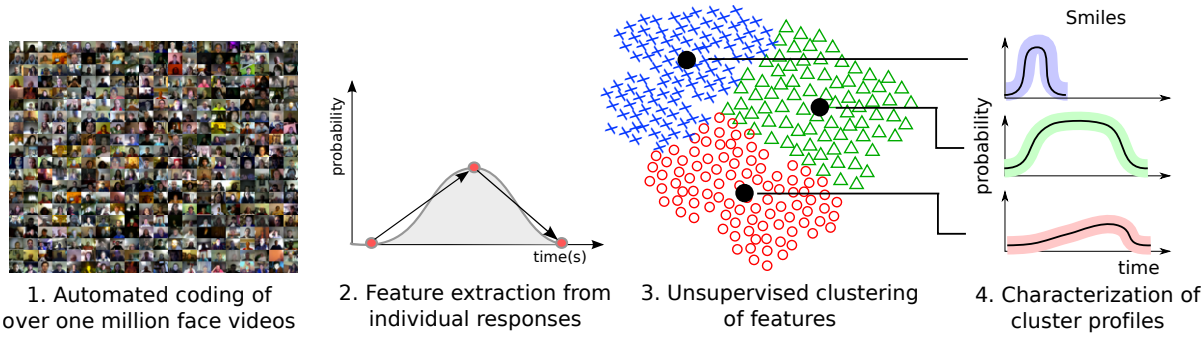

Fig. 2. In this paper we present extremely large-scale clustering of facial events by analyzing dynamic behavior in over 700,000,000 frames of video. The clustering techniques have the potential for improving the efficiency of labeling and discovery of different types of behavior.

The contributions of this work are to: 1) present the first-ever ultra large-scale analysis of over 1.5 million facial videos collected via viewers' webcams (∼700,000,000 frames), 2) to recover clusters of facial dynamics that exist within the data using an unsupervised approach and evaluate the properties of each cluster, 3) to link the occurrence of facial events to the virality of online advertisements, 4) examine the cross-cultural differences in frequency of events.

II. PREVIOUS WORK

A. Clustering

A comprehensive review of techniques for automated facial expression and gesture recognition can be found in Zeng *et al.* [10]. A large majority of the approaches are direct applications of supervised learning techniques. However, some work has considered the application of unsupervised and semi-supervised approaches.

Zhou *et al.* [9] present a novel unsupervised approach (Aligned Cluster Analysis) for clustering similar facial events. De la Torre *et al.* [8] presented a two stage method for temporal segmentation of facial gestures, firstly using spectral graphs to cluster shape and appearance features and secondly grouping these into facial gestures. Another approach for temporal segmentation that clustered frames with similar shapes and cut segments based on shape changes was presented by Zelnik-Manor and Irani [11].

Bettinger *et al.* [12] used active appearance model features to segment long video sequences into shorter segments containing similar distinct actions. The dynamics of the resulting actions were then learnt using a Hidden Markov Model (HMM).

In some cases the techniques described were applied to posed facial data or videos recorded in controlled settings. Without exception, the previous work applied methods to relatively small datasets featuring at most a few hundred participants and video sequences. In contrast we present results showing that meaningful facial behaviors can be recovered from vast amounts of spontaneous and naturalistic data in an unsupervised way.

B. Facial Expression Dynamics

Temporal dynamics help distinguish between different types of facial behavior (i.e. posed and spontaneous eyebrow raises [13] or smiles [14]). Smiles alone can be very complex with many meanings and implications [15]. The dynamics and timing of facial behavior can also influence the interpreted message [16] (i.e. whether the person smiling is perceived as being polite, amused or embarrassed). Automated approaches have used dynamic features to classify smiles responses based on their context (e.g. smiles in the situations of frustration vs. amusement) [17]. However, supervised techniques such as these either require data labeling or assumptions about the properties of different types of responses. We extract temporal features from the facial responses of individuals and learn the properties of different types of responses.

III. DATA

A. Data Collection

The data used for our analysis was collected using a web-based framework much like the method presented by McDuff *et al.* [1]. We leverage the Internet to solicit facial responses from large groups of people, a form of "affective crowdsourcing" [18]. Our framework was deployed to capture responses of participants watching media content via their webcam. The media content principally consisted of video advertisements between 15 and 60 seconds in length. Movie trailers, election debate clips, and TV shows where also used. Participants were asked to opt-in to each study and allow their webcam feed to be recorded. The resulting webcam videos vary in quality due to lighting conditions and Internet bandwidth. As the conditions are unconstrained these videos contain naturalistic responses, including head movements and facial actions caused by external stimuli within the participants environment. These facial videos were streamed to the cloud and processed using the automated facial coding classifiers described below. The time series data output from the facial coding classifiers were stored as comma separated files, each column containing the output from one of the classifiers.

These facial videos include participants from over 94 countries around with world, including the USA, China, Indonesia, United Kingdom, France, Germany. Figure 1 shows the number of sessions collected in each country. The greatest number of sessions were recorded in India (243,118), the United States (221,218), China (212,747) and Indonesia (97,024). The media that participants watched was grouped into 34 ad categories including: groceries, retail, banking and finance, and automotive. The different stimuli and demographics gives us a broad range of data and emotional significance.

B. Automated Facial Coding

Classifiers: The automated facial coding classifiers used in this work all had a similar design. A facial landmark detector (Nevenvision Tracker[2]) was used to identify the location of the face within each frame of video. Histograms of oriented gradients (HOG) features were extracted from the region of interest (ROI) within the frame. A support vector machine (SVM) classifier, with radial basis function (RBF) kernel, for each detector was applied to the HOG features to compute the facial metrics. This is similar to the approach used in [19].

We used three different facial coding classifiers. These were trained on example images coded by at least three human labelers for the presence and absence of an action. The classifiers were:
Eyebrow Raiser - Presence of AU02, outer eyebrow raiser.
Eyebrow Lowerer - Presence of AU04, eyebrow lowerer.
Smile - Presence of a smile (the presence of AU12, lip corner puller, alone was not a sufficient justification.)

Classifier Training: The classifiers were trained on hand-coded data. An inter-coder agreement of 50% was required for a positive example to be used for training and 100% agreement on the absence was required for a negative example to be used for training.

Classifier Evaluation: The facial coding classifiers were tested on webcam data. The testing data were independent to the portion of hand labeled data used for validating the clustering performance in Section V. Each classifier was trained on at least 5,000 FACS labeled example images and tested on a similar number. The area under the receiver operating characteristic (ROC) curves for the eyebrow raiser, eyebrow lowerer and smile classifiers are shown in Table I.

TABLE I
AREA UNDER THE RECEIVER OPERATING CHARACTERISTIC CURVES FOR THE EYEBROW RAISER, EYEBROW LOWERER AND SMILE CLASSIFIERS.

	Classifier		
	Eyebrow R.	Eyebrow L.	Smile.
AUC	77.0	91.6	96.9

[2]Licensed from Google, Inc.

IV. APPROACH

A. Feature Extraction

The facial coding classifiers output a probability of each action being present in every frame that a face could be detected. This results in one 1-d time-series per classifier for each video. The frame rate of the videos was 14 fps.

Each metric was smoothed using a Gaussian filter (σ = 3) to remove high frequency noise that would lead to spurious peaks being detected. Secondly, a simple peak detection algorithm was used to identify peaks and valleys within the metrics. The algorithm finds all local minima of the smoothed curve, this is anytime the derivative of the metric goes from negative to positive. An event was defined using the following criteria:

Algorithm 1 Event Definition

1: *loop (through minima)*:
2: **if** $(0.8 * \text{max} + 0.2 * \text{curr. min.} > \text{next min.} \ \& \ \text{max} > 10)$ **then**
3: segment curr. min to next min. is event.
4: **goto** *loop*.

Where: "max" is the largest maxima between the current minima ("curr. min") and next consecutive minima ("next min."). When the "if" statement was true the segment between the current minima and the next minima is considered an event.

Finally, a set of six features were extracted from each of the resulting events. The features were used in our unsupervised learning model. Figure 3 shows an overview of the feature extraction. The features extracted from each event were:

Event Height (A): Maximum value.
Event Length (B): Duration between onset and offset.
Event Rise (C): Increase from onset to peak.
Event Decay (D): Decrease from peak to next offset.
Rise Speed (E): Gradient of event rise.
Decay Speed (F): Gradient of event decay.

After finding all events, we then filtered the events allowing only those with an **Event Rise** greater than 10% of the maximum possible event rise. After feature extraction and filtering we were left with 834,704 outer eyebrow raiser events, 945,511 eyebrow lowerer events and 459,468 smile events. Although this is a large number of events, the result still indicates that a large majority of the videos do not contain a smile, eyebrow raise, or eyebrow lowerer event. Facial expressions are sparse and thus we need accurate classifiers. This finding supports previous work [20].

Running this algorithm on a single local machine would have taken approximately 200 hours - clearly prohibitive and non-repeatable. Using Hadoop, a distributed computing platform, where we were able to distribute the data on nine machines running in parallel. In total, processing using Hadoop in the cloud took 2.5 hours compared to the 200

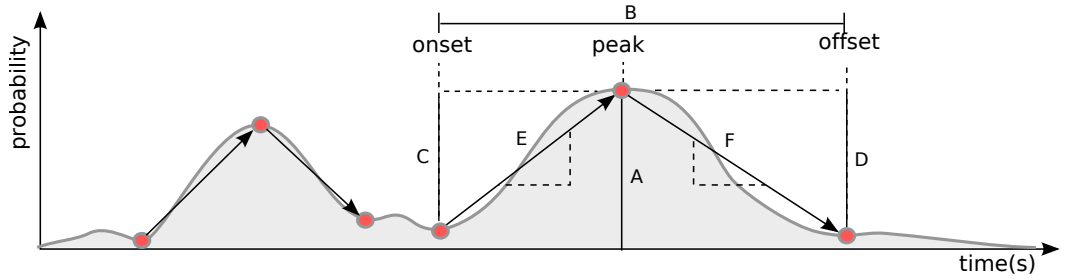

Fig. 3. Definitions of the features extracted for each facial event. 1) A peak detector was used to find peaks and valleys within each signal. 2) A threshold was used to filter out events with height < 0.1. 3) Event height (A), length (B) rise (C), decay (D), rise speed (E) and decay speed (F) were extracted from the resulting events to form a feature vector. The features were normalized using z-score and power transforms before clustering.

hours locally. We believe that as facial analysis and affective computing enters the realm of what is known loosely as "big data", it is necessary for researchers to adopts distributed machine learning and data processing platforms such as Hadoop.

B. Feature Normalization (Power Transformation)

Before performing the unsupervised learning we normalized all the features to the same scale. Before scaling, the event height features ranged from 10 to 100, while event rise speed features ranged from 0 to 10. To normalize the features we used a Box-Cox transformation over each feature:

$$x\prime = \frac{x^{\gamma} - 1}{\gamma} \tag{1}$$

Where γ was computed from the maximum likelihood estimator for each feature and action unit.

Following this we subtracted the mean of each feature and divided by the standard deviation (to result in a z-score with zero mean and unit variance):

$$x\prime\prime = \frac{x\prime - \mu_{x\prime}}{\sigma_{x\prime}} \tag{2}$$

This transformation allows for equal significance of each feature in our k-means model.

C. Clustering

We used an unsupervised approach, K-Means, to cluster the resulting data. Although K-Means is a relatively simple approach it does allow for fast computation (necessary with such large data) and was deemed sufficient.

1) Choosing K in K-means: To choose the number of clusters we computed the Bayesian Information Criterion (BIC_k) for K in 1,2,..,10. The smallest K was chosen where (1-BIC_{k+1}/BIC_k)<0.025. For smile and eyebrow raiser the smallest K corresponded to five clusters and for eyebrow lowerer it corresponded to four clusters.

2) K-means: Given a value for K and our feature set, the K-Means clustering algorithm was used to group each event into its respective category. Once clustering was completed the cluster memberships could then be mapped back to their pre-transformed features for analysis.

V. Results

A. Unsupervised Clustering

K-Means clustering was run on the eyebrow raiser, eyebrow lowerer and smile separately. Eyebrow lowerer events were grouped into four clusters while there were five clusters for the eyebrow raiser events and smile events. Table II shows statistics and attributes of each cluster. For example, the smile clusters are distributed evenly, being made up of 18%, 22%, 21%, 15%,and 24% respectively. Cluster 1 has a higher event peak and rise than the others with a long length while Cluster 2 also has a high event peak and rise with a shorter length. Figure 4 shows how the smile event shapes differ per cluster.

B. Human-Labeled Event Comparison

In order to estimate what proportion of each cluster represented true examples of the action compared to false positives a subset of the extracted events were labeled by human coders. The coders were asked to confirm if the detected event was a true example of the action (e.g. whether a smile event corresponded to a smile or a false positive). This exercise showed different results for each of the classifiers.

1) Smile: As smile is proven to be the most accurate classifier (ROC AUC = 0.96 in Table I) we expect results to show multiple clusters with a large portion of true positives. Clusters 1 and 2 are dramatically more accurate than the other three with 86% and 75% true positive rates. These clusters are characterized by large rise and event values. Cluster 2 represents a fast onset and offset smile with short duration. Cluster 1 represents longer duration smile. Cluster 5, which showed the highest ratio of false positives at 65%, has the lowest average peak value and peak length, which as we can see in Table III. This result suggests that many events in peak 5 may be caused by rigid head motions or other noise artifacts - a very useful observation.

2) Eyebrow Lowerer: Results show the clusters 1 and 2 have similar and relatively high true positive rates while clusters 3 and 4 have lower accuracy. As expected, clusters with lower average peak rise have higher false positive rates.

3) Eyebrow Raiser: Cluster 2 and 3 have true positive rates of 61% and 62% respectively, the only clusters with a true positive rates greater than 50%. Clusters 1, 4, and 5 have nearly a 50/50 chance of giving a true eyebrow raise.

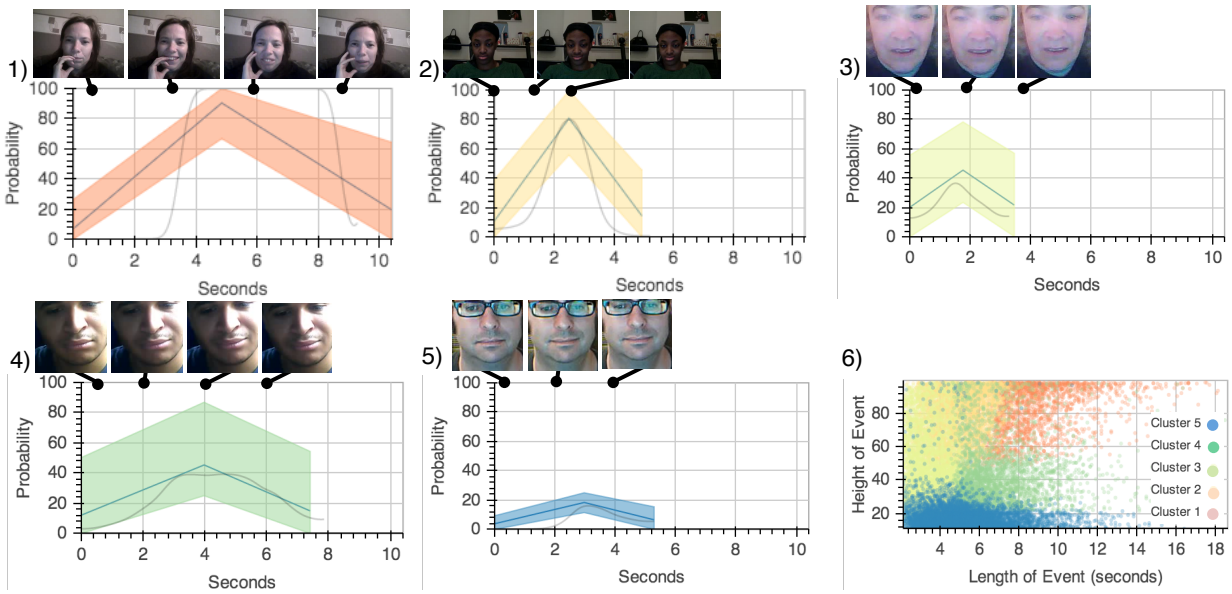

Fig. 4. a-e) Show the distributions of the event heights for the events within each smile cluster. The colors of the events reflects the cluster in the scatter plot (f). Example smile tracker from each cluster are also shown superimposed. For the examples we have shown cropped frames from the face videos corresponding to the example track. Examples 1-4 were smiles, example 5 was not a smile. f) Shows the distribution of event length and event height per cluster for smile. The colors used in Figure f match those in Figures 1-5. Similar profiles were observed for eyebrow raise and eyebrow lowerer. Each of the people shown agreed to allow their video to be shared publicly.

TABLE II

FEATURES OF THE CLUSTER CENTROIDS THAT RESULT FROM THE UNSUPERVISED CLUSTERING OF SMILE, EYEBROW RAISER AND EYEBROW LOWERER EVENTS. THE FEATURES DEFINITIONS CAN BE FOUND IN FIGURE 3. THE CLUSTERING IDENTIFIED FIVE CLUSTERS (FOUR FOR EYEBROW LOWERER) IN EACH CASE THAT HAD DIFFERENT DYNAMICS AND LENGTHS.

Cluster	Events in Cluster	% of Total Events (%)	Event Height (A) (prob.)	Event Length (B) (s)	Event Rise (C) (prob.)	Event Decay (D) (prob.)	Event Rise Speed (E) (prob./sec)	Event Decay Speed (F) (prob./sec)
			Smile					
1	76,277	18	90.21	10.39	83.19	70.97	21.17	15.58
2	94,304	22	80.55	4.94	69.6	66.1	29.98	28.8
3	92,190	21	45.51	3.46	25.31	23.81	14.83	14.18
4	66,676	15	45.28	7.4	33.21	30.42	9.84	10.45
5	106,307	24	18.39	5.29	14.51	11.81	5.81	5.72
			Eyebrow Lowerer					
1	238,630	27	52.06	5.17	45.98	43.95	19.75	19.69
2	156,510	18	43.2	7.78	33.96	26.16	9.68	9
3	204,621	23	33.44	3.11	22.76	20.1	14.74	13.89
4	273,651	31	17.04	4.66	13.95	11.38	6.25	6.31
			Eyebrow Raiser					
1	91,679	20	63.81	4.74	58.03	56.64	26.18	25.78
2	86,254	18	57.7	7.74	50.42	43.86	13.02	14.82
3	108,763	23	39.82	3.59	27.46	25.58	15.77	15.35
4	92,234	20	20.53	6.3	17.4	15.65	5.47	6.7
5	88,945	19	19.18	3.44	14.12	11.1	7.98	7.54

C. *Region Differences*

Previous research has shown differences in facial expressions between cultures and markets [21]. Our clustering work enabled us to find variations from region to region. Our dataset features video responses recorded in markets all over the world (see Figure 1).

Figure 5 shows the proportion of events from the different clusters that were detected in North America, Oceania, Latin America, Europe, the Middle East, Africa, Southern, Southeastern and Eastern Asia. From the previous experiment on human labeled event comparison, we know that clusters 1 and 2 are made up of high peak values with high accuracy. These clusters are more prominent in North America than Asian regions. As a result, Asian regions contain a higher ratio of clusters 4 and 5, which have lower peak values with lower accuracy. The results are particularly interesting and correspond well with previous research. They suggest that individuals in Asia are generally less expressive than those

TABLE III

THE TRUE AND FALSE POSITIVE EVENTS WITHIN EACH CLUSTER FOR EACH OF THE EXPRESSION CLASSIFIERS. THE NUMBER OF HAND LABELED EVENTS USED FOR VALIDATION WERE 3162, 4094, 1929 FOR SMILE, EYEBROW RAISER AND EYEBROW LOWERER RESPECTIVELY. AS FACIAL EVENTS ARE VERY SPARSE A CLUSTER WITH TRUE POSITIVE RATE AT 30% CAN STILL BE USEFUL.

Smile					
Cluster	1	2	3	4	5
True Positive	0.86	0.75	0.48	0.47	0.35
False Positive	0.14	0.25	0.52	0.53	0.65
Eyebrow Lowerer					
Cluster	1	2	3	4	
True Positive	0.65	0.61	0.39	0.32	
False Positive	0.35	0.39	0.61	0.68	
Eyebrow Raiser					
Cluster	1	2	3	4	5
True Positive	0.48	0.61	0.62	0.46	0.44
False Positive	0.52	0.39	0.38	0.54	0.56

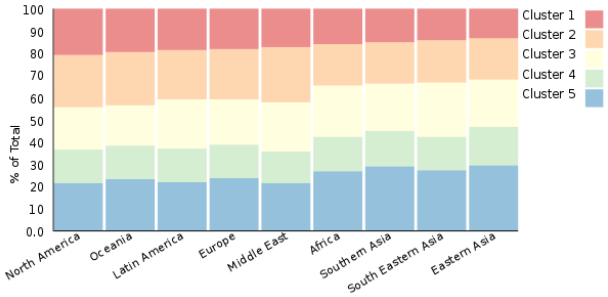

Fig. 5. Stacked bar chart showing the percentage of smile events which fall into each clusters per region. The chart is ordered by proportion of events in cluster one.

in North America. Thus we need classifiers that can pick up more subtle examples of these behaviors. As expressiveness is greater in North America and Oceania the performance of our algorithms will be slightly better in those markets.

D. Evaluating Media Effectiveness

We evaluated the significance of the facial response clusters found with respect to the effectiveness of the media content being viewed. One of the key measures of effectiveness of online video content is virality [22]. Of the over 8,000 ads tested we had measures of self-report sharing likelihood for 170 ads and YouTube statistics for a further 40 ads. We use these in our evaluation.

Intention to Share:

For 74 humorous ads and 96 non-humorous ads (170 ads total) we collected an average of 72 facial responses per ad. The ads were for chocolate, pet care and food products. Following each ad we asked the question:

Q. If you watched this ad on a website such as YouTube how likely would you be to **SHARE** it with someone else?

Very unlikely		Neutral		Very likely
1.	2.	3.	4.	5.

We extracted the events from the automatically coded facial responses and classified them using the clustering algorithm described above.

Figure 7 shows the number of times events from each smile cluster was observed per response for viewers that reported a likelihood to share greater than neutral vs. a likelihood to share less than neutral. We can see that there is a much greater number of smile events for those likely to share. The ratio of smile to eyebrow lowerer events during the non-humorous ads was 0.96 for the group likely to share and 0.52 for the group unlikely to share. The different was even greater (0.98 and 0.49) for the humorous ads. Futhermore, those that reported the highest likelihood of sharing (i.e. 5 on the scale) had a smile to eyebrow lowerer ratio of 1.1 compared to 0.41 for those that reported the lowest likelihood of sharing (i.e. 1 on the scale).

YouTube Statistics:

YouTube is another source from which we can gather virality data for ads. Both the view and like counts give an aggregate measure of the popularity of an online video. We collected facial coding data over 40 ads with 80 facial responses each along with the corresponding YouTube statistics two weeks after publishing. Ads tested where both humorous and non-humorous. Many of the ads were commercials during the 2014 Superbowl and thus had a higher prior probability of going viral than an average ad.

To show the effectiveness of these ads, we chose 2 humorous and 2 non-humorous ads, each pair containing a viral and not viral ad, and extracted events from each session. The corresponding YouTube statistics are shown below.

Ad Type	Virality	Views	Likes
Humorous	Viral	20,617,524	94,449
	Not Viral	1,385,458	450
Not Humorous	Viral	47,026,862	196,767
	Not Viral	11,113,680	27,664

Figure 8 shows the ads considered viral exhibited more smile events than the not viral ads. The humorous viral ad contained many more events in clusters 1 and 2 (stronger and more relible smile events) and less events in cluster 5 (weaker and less reliable smile events, 35% true positive rate) than the not viral ad. The non humorous ads shows similar results. Also, we can see the greater number of smile events exhibited by the humorous ads than the non-humorous ads.

Although the main aim of this paper was not to predict media effectiveness from facial expressions, it is an example for which large-scale clustering results could be very useful. However, there are also many other potential applications for the analysis presented here.

VI. APPLICATIONS

Clustering results have the potential to vastly improve the efficiency of labeling. In sparse data many of the frames have

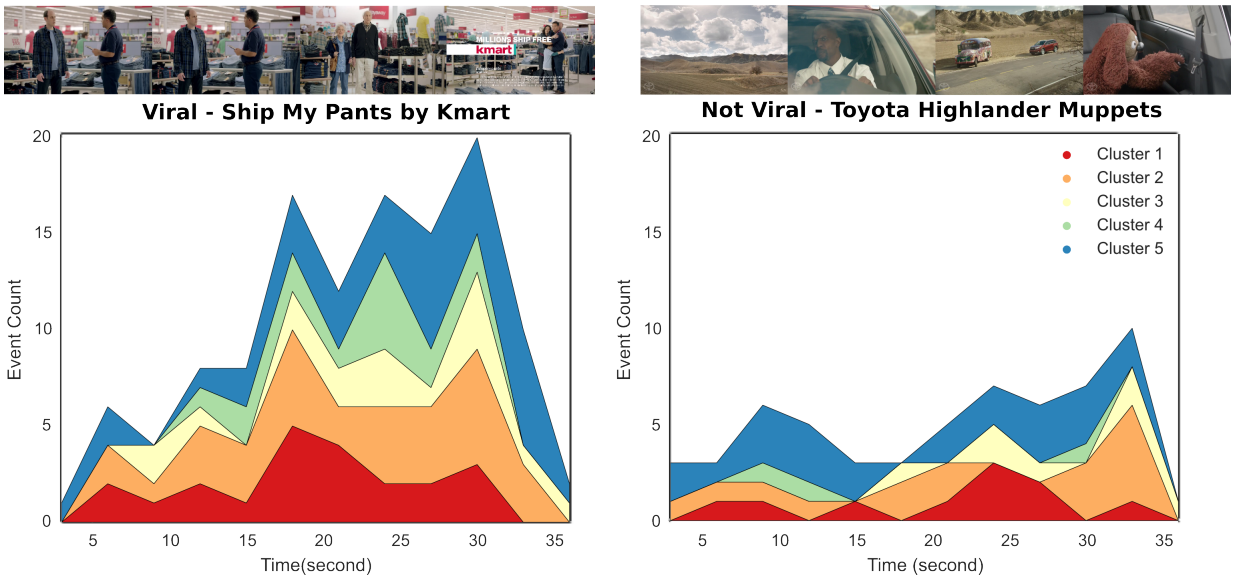

Fig. 6. Stacked area chart showing the number of occurrences of events from each smile cluster detected during two ads, a viral ad (N=72) and non-viral ad (N=84). Both ads were intentionally humorous, however, the non-viral ad elicited far fewer smile events.

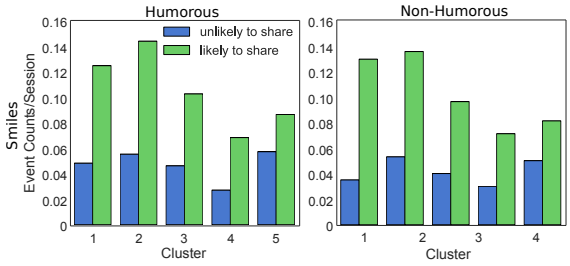

Fig. 7. The number of smile events per session for individuals who report high and low likelihood to share the content they are watching. Left) Results for humorous ads. Right) Results for non-humorous ads. This figure shows that smiles (especially from clusters 1 and 2) were more prevalent when people were more likely to share.

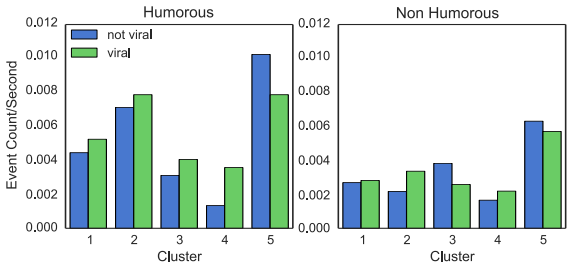

Fig. 8. The number of smile events per second over each participant while watching a given ad. Left) Shows these frequencies for two humorous ads, one which is considered viral and another not viral. Right) Shows frequencies for two non-humorous ads, one which is considered viral and another not viral.

no action units present and so we end up with many more negative labels than positive labels. However, by utilizing the results of our clustering we can prioritize labeling of data with a high prior of being positive examples and making the process much more efficient.

Events can also give an efficient summary of large amounts of facial data. For instance, someone might be interested to know how many instances of a particular facial event occurred within a populations response to a piece of media. However, exactly how to describe these events may be challenging if definitions are not known. In which case event definitions can be generated using an unsupervised approach such as ours.

VII. CONCLUSIONS

In this paper we present ultra large-scale analysis of naturalistic and spontaneous facial responses collected "in-the-wild". We analyzed over 1.5 million videos of facial responses to online media content (∼700,000,000 frames), a task that would have been impossible using manual coding. We use unsupervised clustering to discover groups of dynamic facial behaviors.

Over 2.5 million events were identified and clustered. Verifying the observations in each cluster against human-coded data we were able to identify reliable clusters of responses with different dynamics (e.g. fleeting vs. sustained smiles).

By looking at the occurrence of these events during a set of 170 ads we found that smiles, with high peaks, were much more likely during content which people reported a strong likelihood share and confusion events less likely. The same trends were observed in responses to Super Bowl ads that were shared virally following their launch versus those that

were not. This has the potenial for allowing us to predict what pieces of content might be viraly shared by measuring the responses of only 100 viewers.

Our approach has the potential for vastly decreasing the time and resources required for manual facial action coding. Our method makes the identification of positive examples much more efficient. Future work will involve integrating our approach into an automated system for prioritizing human coding of facial expression data. In addition, the use of event descriptions work as nice features for describing responses to online content. We would like to use clusters such as these to discriminate between different kinds of media (such as very funny and mildy funny). We aim to extend this approach to a greater number of facial behaviors.

REFERENCES

[1] D. McDuff, R. Kaliouby, and R. W. Picard, "Crowdsourcing facial responses to online videos," *IEEE Transactions on Affective Computing*, vol. 3, no. 4, pp. 456–468, 2012.

[2] D. McDuff, R. El Kaliouby, E. Kodra, and R. Picard, "Measuring voter's candidate preference based on affective responses to election debates," in *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*. IEEE, 2013, pp. 369–374.

[3] T. Sénéchal, J. Turcot, and R. el Kaliouby, "Smile or smirk? automatic detection of spontaneous asymmetric smiles to understand viewer experience," in *10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (AFGR), 2013*. IEEE, 2013, pp. 1–8.

[4] D. McDuff, R. E. Kaliouby, T. Senechal, D. Demirdjian, and R. Picard, "Automatic measurement of ad preferences from facial responses gathered over the internet," *Image and Vision Computing*, vol. In press, 2014.

[5] T. Teixeira, R. El Kaliouby, and R. W. Picard, "Why, when and how much to entertain consumers in advertisements? A web-based facial tracking field study," *Marketing Science*, 2014.

[6] P. Ekman and W. Friesen, *Facial action coding system*. Consulting Psychologists Press, Stanford University, Palo Alto, 1977.

[7] 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*, vol. 31, no. 1, pp. 39–58, 2009.

[8] F. De la Torre, J. Campoy, Z. Ambadar, and J. F. Cohn, "Temporal segmentation of facial behavior," in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*. IEEE, 2007, pp. 1–8.

[9] 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.

[10] E. Sariyanidi, H. Gunes, and A. Cavallaro, "Automatic analysis of facial affect: A survey of registration, representation and recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, In press, 2014.

[11] L. Zelnik-Manor and M. Irani, "Event-based analysis of video," in *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, vol. 2. IEEE, 2001, pp. II–123.

[12] F. Bettinger, T. F. Cootes, and C. J. Taylor, "Modelling facial behaviours." in *BMVC*, 2002, pp. 1–10.

[13] M. Valstar and M. Pantic, "Fully automatic facial action unit detection and temporal analysis," in *Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition Workshop (CVPRW), 2006.* IEEE, 2006, pp. 149–149.

[14] M. F. Valstar, H. Gunes, and M. Pantic, "How to distinguish posed from spontaneous smiles using geometric features," in *Proceedings of the 9th international conference on Multimodal interfaces*. ACM, 2007, pp. 38–45.

[15] M. LaFrance, *Lip service: Smiles in life, death, trust, lies, work, memory, sex, and politics*. WW Norton & Company, 2011.

[16] Z. Ambadar, J. Cohn, and L. Reed, "All smiles are not created equal: Morphology and timing of smiles perceived as amused, polite, and embarrassed/nervous," *Journal of Nonverbal Behavior*, vol. 33, no. 1, pp. 17–34, 2009.

[17] M. E. Hoque, D. McDuff, and R. W. Picard, "Exploring temporal patterns in classifying frustrated and delighted smiles," *IEEE Transactions on Affective Computing*, pp. 323–334, 2012.

[18] R. Morris and D. McDuff, "Crowdsourcing techniques for affective computing," *In R.A. Calvo, S.K. DMello, J. Gratch and A. Kappas (Eds). Handbook of Affective Computing.*, 2014.

[19] E. Kodra, T. Senechal, D. McDuff, and R. el Kaliouby, "From dials to facial coding: Automated detection of spontaneous facial expressions for media research," in *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (AFGR)*. IEEE, 2013, pp. 1–6.

[20] D. McDuff, R. E. Kaliouby, J. F. Cohn, and R. Picard, "Predicting ad liking and purchase intent: Large-scale analysis of facial responses to ads," *IEEE Transactions on Affective Computing*, In press, 2014.

[21] N. Mediratta, R. el Kaliouby, E. Kodra, and P. Jha, "Does facial coding generalize across cultures?" in *European Society for Opinion and Marketing Research (ESOMAR) - Asia Pacific, 2013*.

[22] J. Berger and K. Milkman, "What makes online content viral?" *Unpublished manuscript, University of Pennsylvania*, 2011.