Smile or Smirk? Automatic Detection of Spontaneous Asymmetric Smiles to Understand Viewer Experience

Thibaud Sénéchal, Jay Turcot and Rana el Kaliouby

***Abstract*— Asymmetric facial expressions, such as a smirk, are strong emotional signals indicating valence as well as discrete emotion states such as contempt, doubt and defiance. Yet, the automated detection of asymmetric facial action units has been largely ignored to date. We present the first automated system for detecting spontaneous asymmetric lip movements as people watched online video commercials. Many of these expressions were subtle, fleeting and co-occurred with head movements. For each frame of the video, the face is located, cropped, scaled and flipped around the vertical axis. Both the normalized and flipped versions of the face feed a right hemiface trained (RHT) classifier. The difference between both outputs indicates the presence of asymmetric facial actions on a frame-basis. The system was tested on over 500 facial videos that were crowdsourced over the Internet, with an overall 2AFC score of 88.2% on spontaneous videos. A dynamic model based on template matching is then used to identify asymmetric events that have a clear onset and offset. The event detector reduced the false alarm rate due to tracking inaccuracies, head movement, eating and non-uniform lighting. For an event that happens once every 20 videos, we are able to detect half of the occurrences with a false alarm rate of 1 event every 85 videos. We demonstrate the application of this work to measuring viewer affective responses to video content.**

***Index Terms*— Emotion, Affective Computing, Facial asymmetry, Asymmetric Facial Expressions, Dynamic event detection**

I. Introduction

The face is one of the richest sources of communicating social and emotional information, and is capable of generating tens of thousands of expressions, including asymmetric ones. Asymmetric facial expressions occur when the intensity, or muscular involvement, on both sides of the face differ. As shown in Fig. 1, asymmetric facial expressions such as smirks play an important role in emotion communication and inform overall valence as well as discrete emotion states such as contempt, skepticism and defiance [7]. Charles Darwin was the first to mention facial asymmetry in his discussion of "Sneering and Defiance" [5], [6]. The Facial Action Coding System (FACS) [8] defines asymmetric action units as those that happen on both sides but with different intensities; if the movement happens strictly on one side of the face, the facial action is referred to as unilateral. For the purposes of this paper, we will use asymmetry to include both unilateral as well as asymmetric expressions.

We first got interested in facial asymmetry while analyzing the thousands of crowdsourced facial videos of Internet viewers watching commercials. Even though some commercials were meant to be humorous, many viewers who did not find the content funny or were skeptical about the message exhibited asymmetric lip expressions, which we will refer to broadly as asymmetric smiles. We realized that, despite being ignored in the automated facial analysis community, asymmetric expressions are frequent and well-documented in the psychology literature (see [1] for a review).

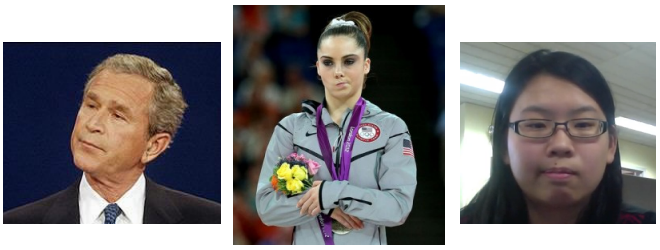

Fig. 1. Asymmetric lip expressions: a) George W. Bush's popular smirk, which people criticized as portraying smugness; b) McKayla Maroney, an American artistic gymnast, looking unimpressed after she won the silver medal in the London 2012 Olympics; c) A viewer looking skeptical in response to claims of an advertisement.

Facial nerves carry neural impulses that, through contracting the various facial muscles, generate a wide gamut of facial expressions [18]. Anatomically, the left facial nerve is independent of the right facial nerve. Often, identical signals are sent to both the left and right facial nerves and the resulting facial expression appears symmetric. Occasionally the facial nerve supplies the muscles of only one side of the face, resulting in asymmetric expressions. Because the muscles of the lower two-thirds of the face are contralateral, i.e. the left hemisphere controls the right hemiface, and vice versa, most humans can easily perform asymmetric movements of the lips, such as a smirk, and these expressions are common. Muscles higher up the face are more bilateral, making it harder to, as an example, raise only one eyebrow. A number of studies have reported that asymmetry is more dominant on the left face in posed expressions of emotion [3], [9], [19] and tends to happen with negative expressions of emotion [2]. For spontaneous expressions, there seems to be equal occurrence of left vs. right hemiface asymmetry [1].

In this paper, we describe an automated system for detecting asymmetric facial actions such as smirks in spontaneous video. The main contributions are as follows: 1) To the best of our knowledge, this is the first system that detects spontaneous asymmetric smiles and reliably distinguishes those expressions from symmetric smiles of happiness; 2) Our

T. Sénéchal, J. Turcot, R. el Kaliouby are with Affectiva Inc., Waltham, MA, USA. {thibaud.senechal,jay.turcot,kaliouby}@affectiva.com

method exploits the natural symmetry of the face to detect asymmetric emotion expressions by applying a classifier to both an original face frame as well as a flipped version of that frame; 3) We use dynamics, namely the detection of the onset and offset of asymmetric facial events, to increase the robustness of our system to real-world conditions; 4) We go beyond just reporting accuracy scores to additionally demonstrate use in a real-world video viewing application to identify moments where viewers exhibit doubt.

II. Related work

Despite the extensive psychology literature on facial asymmetry, the vast majority of research in automated facial analysis assumes that facial expressions are symmetric, possibly due to the lack of databases with facial asymmetry. There are a few exceptions. Liu et al. [14] combine facial asymmetry information with EigenFace and FisherFace to improve identity recognition. They report improved identity recognition on 110 subjects from the FERET database [17], and 55 subjects from the Cohn-Kanade database [12]. Hu [11] explores the use of facial asymmetry to estimate head yaw using Gabor filters and linear discriminant analysis.

Mitra and Liu [16] use facial asymmetry features to improve the recognition of joy, anger and disgust from the Cohn-Kanade database. Fasel and Luttin [10] estimate facial action asymmetry under hugely constrained conditions, such as perfect lighting and no head movement. Difference images (compared to a neutral frame) are projected into a sub-space using PCA or ICA then nearest neighbor classification. The authors artificially create an asymmetric dataset by masking out half of the face and using the neutral face to complete that half, which results in unrealistic images. Moreover, their approach is person-dependent, which is impractical in many real-world applications. Valstar et al., [23] use the fact that upper facial actions are largely symmetric to differentiate between posed and spontaneous eyebrow movements.

Our work differs from previous work in several ways. Early work focused on posed asymmetry in images with good lighting and no head movement. In this paper, we detect asymmetric smiles on challenging spontaneous facial videos that have non-uniform lighting, substantial head movement, viewers who are eating and expressions that are subtle and fleeting. Also, we do not assume the presence of neutral frame at the start of the video, since that is unreliable in real-world data. Finally, our work is unique in that we demonstrate that our system is deployable in the real world.

III. System overview and data

A. Data collection

We use a similar web-based framework as the one described in [15], to crowdsource facial videos as people watch online video commercials. Viewers opt-in to turn their webcam on and to watch short videos while their facial expressions are recorded. On the viewer's machine, all that is needed is a browser with Flash support and a webcam. The video from the webcam is streamed in real-time, at approximately 14 frames a second with a resolution of 320x240, to a server where automated facial expression analysis is performed. The commercials range from amusing to those without high emotion inducing content.

TABLE I

Facial expressions in our spontaneous dataset. Indicates the number of frames containing the expression, and the number of videos or subjects that display at least the facial expression in one frame.

	Asymmetric smile	True smile	AU04	AU02	All
frames	5125	114,000	57,622	16,105	655,000
videos	201	868	254	308	2265

The videos were recorded in real-world conditions. They exhibit non-uniform frame rate and non-uniform lighting. In some cases, the screen of a laptop is the only source of illumination. The camera position is not fixed relative to the viewer and the screen. Viewers may not be paying attention, are fidgeting, and occasionally exhibit challenging behavior such as, eating, speaking to other people in the room or on the phone. Videos collected also contain viewers with glasses and hair bangs that occlude portions of the upper face.

For each video, three FACS trained human labelers marked the onset and offset of several expressions at frame level accuracy. The expressions labeled consisted of: asymmetric smile, true smile (symmetric AU12), AU02 and AU04. Other FACS certified experts then reviewed the labels and videos with low inter-coder reliability were returned for re-labeling. The asymmetric smile label consists of either the asymmetric lip corner pull (AU12) or the asymmetric dimpler (AU14), often referred to as a smirk (Fig. 1a, 1c). We also occasionally observed a lip pucker (AU18) (Fig. 1b) happening in combination with AU12.

To date, we have collected and labeled 2265 videos of spontaneous data with lengths ranging between 30 seconds to a minute. Table I shows the AU occurrences in these videos. 201 videos contain at least one asymmetric smile. As many asymmetric smiles occur on the right hemiface as on the left (97 and 104 respectively). This is consistent with previous literature [1], where equal occurrence of left vs. right asymmetry is reported for spontaneous expressions. About 90% of the asymmetric expressions last less than 5 seconds. One sixth of the frames contain a true smile, which is one of the challenges faced by our system: it must not fire on symmetric smiles (representing happiness) while detecting rare asymmetric smiles (representing skepticism). The videos are grouped into 10 datasets, where each set represents viewers from Brazil, China, India, Germany, Mexico and the United States watching different commercials, thus covering a range of ethnicities: Caucasian, Asian and Hispanic.

In addition to the spontaneous data, we collected 200 posed videos in our locals with the same framework. Participants were asked to pose various mouth expressions, e.g., symmetric and asymmetric AU12 and AU14, AU20 (lip stretcher), grimace etc., while slightly shifting their

head pose (up and down, left and right turn, forward and backward). In general, the posed videos had better lighting, no occlusion and clearer expressions as people were mostly focusing on the task.

B. System overview

The system is comprised of two main components (Fig. 2). The first component is a frame-by-frame detector that yields a confidence index on the presence of an asymmetric smile per frame. For each frame of the video, the face is extracted, cropped and rescaled. Then, Histogram of Oriented Gradients (HOG) [4] features are computed on the normalized face as well as a flipped version of this face. The two descriptors are input, individually, to a Right Hemiface Trained (RHT) classifier that recognizes a right asymmetric smile. The higher the difference between the two outputs, the more likely there is an asymmetric expression on the face.

The second component uses dynamics to detect asymmetric "events" that have a clear onset and offset. The rationale is as follows: we observe, from the FACS labels, that the majority of asymmetric smiles have a short duration with a clear onset and offset. In the case where a viewer's natural facial configuration has asymmetry around the lip region, it is unlikely that the expression is due to a reaction to the content. As a result, the presence of an onset and offset can be used to reduce false alarms. The input to the event detector is the absolute difference between the two outputs of the RHT classifier, computed for each frame of the video. We fit a template—a rectangular function with width W and height H—to the difference signal using a sliding window approach. The following sections explain each component in detail.

IV. Frame-by-frame detection

The frame-by-frame detector aims to recognize the presence of an asymmetric expression in each frame of the video.

A. Features extraction

To compute the features, we first use the Google tracker [13] to detect three facial points: the top of the mouth and the two outer eye corners. We then extract, crop and warp the face to a 96x96 pixel image with fixed inter-ocular distance and vertical scale. Feature extraction was done using the OpenCV implementation of HOG. In addition, Local Binary Patterns (LBP) and Local Gabor Binary Patterns (LGBP) were tested, due to their superior performance for the AU recognition task on the GEMEP-FERA database [22].

The HOG descriptor, first applied in object detection [4], counts the occurrence of gradient orientation in localized portions of an image and improves robustness by using overlapping local contrast normalization. This descriptor represents the face as a distribution of intensity gradients and edge directions and has the advantages of being robust to scale and translation [4]. We explored different parameters for HOG (see section V-A.2); we had the best results using 4x4 cell blocks of 8x8 pixel cells with an overlap of half the block, and histograms of 9 channels evenly spread over 0 to $180°$. The dimension of such a HOG descriptor on a 96x96 image is 3600 (25 blocks x 16 cells x 9 bins).

As shown in Fig. 2, we compute HOG for the flipped image. The image is flipped around the vertical line in the middle of the face. The symmetry plane is determined from the tracker points and therefore relies on accurate tracking. Further work on refining the symmetry plane using visual appearance may increase the robustness to tracker inaccuracies. Flipping the gray level image is equivalent, in the HOG feature space, to flipping the gradient orientations. Rather than computing the flipped HOG feature directly, the histograms of the original HOG feature can be permuted, and the cells reordered to produce the same feature. We use these properties to avoid unnecessary computations and to ensure that the system runs in real-time.

B. Classification

The recognition of an asymmetric smile in the right hemiface is treated as a binary classification problem. Images containing right asymmetric expressions were used as positive samples (target class) and all other images as negative samples (non-target class). The classification performance of several classifiers was tested: Support vector machines (SVM) and random forests.

In case of SVM classifiers, training samples composed of HOG histograms x_i associated with labels y_i (target or non-target), the classification function of the SVM computes the distance d to the SVM hyperplane of the new sample x:

$$d = \sum_{i=1}^{m} \alpha_i k(x_i, x) + b \tag{1}$$

where α_i is the dual representation of the hyperplane's normal vector [20] and k is the kernel function resulting from the dot product in a transformed infinite-dimensional feature space. In our experiments, we evaluate the Gaussian Radial Basis Function (RBF) kernel, the Histogram Intersection Kernel (HIK) [22] [21] and the linear kernel.

C. Detecting asymmetry

To detect the presence of an asymmetric expression on both sides of the face, we use the right-hemiface trained (RHT) classifier on two HOG features: one computed from the original image x, the other one from the flipped image $\bar{x}$. The final score s predicting the likelihood of an asymmetric smile is computed as the absolute difference between the two outputs of the classifier:

$$s = \mid \sum_{i=1}^{m} \alpha_i (k(x_i, x) - k(x_i, \bar{x})) \mid \tag{2}$$

With this approach, a perfectly symmetric expression, e.g., a symmetric smile, would yield a score of 0. Also, because the classifier is trained to recognize only asymmetric smiles, it would not react to other asymmetric facial actions, such as a wink, as classifier outputs for the original image and the flipped one should be both low. Still, this approach can be easily applied to other asymmetric expressions such as a wink if a proper RHT classifier was trained.

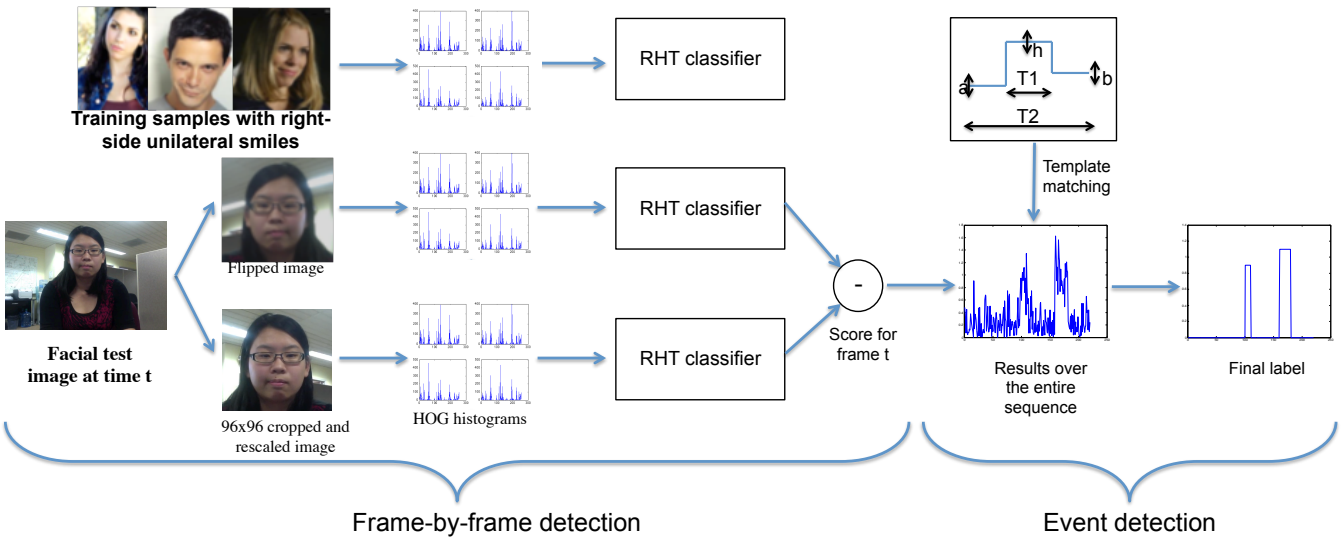

Fig. 2. Flowchart of asymmetric smile detection. Frame-by-frame detection: for each frame, the face is located, cropped, scaled and flipped around the vertical axis. Both the normalized and flipped images are input to a right hemiface trained (RHT) classifier. The difference between both outputs indicates the presence of asymmetric facial actions. Event detection: template matching is used to identify asymmetric events that have a clear onset and offset.

V. Experimental results for frame-by-frame detection

A. Experimental setup

Our first set of experiments was aimed at training the best classifier to recognize an asymmetric smile on the right hemiface. The second set of experiments validates the use of this approach to detect asymmetric smiles on both hemifaces.

1) Training dataset: The classifier was trained on posed data, as well as 1000 spontaneous expression videos of our 2265 video dataset. Images where at least 2 out of 3 FACS labelers identified an asymmetric smile were used as positive samples. Images where none of the labelers identified an asymmetric smile were used as negative samples. Images where only 1 out of 3 labelers identified an asymmetric smile were considered ambiguous and disregarded. This process yielded 10,700 positive samples of asymmetric smiles from 157 persons and 381,000 negative samples from 1200 persons. To reduce the number of training samples, one frame every 200 ms was selected as a positive sample. Negative frames were uniformly sub-sampled to balance the number of positive and negative samples. As a result, 4000 positive samples and 4000 negative samples were used for training.

2) Tuning procedure: One sixth of the training data was used as a validation set to tune the parameters of the system:

- HOG descriptor parameters: cell size, block size, # of channels per cell and # of overlapping blocks;
- Training samples: sampling rate per video to select positive samples, ratio of positive to negative samples;
- SVM classifier: trade-off and RBF kernel spread.

3) Test dataset: A set of 500 spontaneous expression videos was used to evaluate classification performance. The test data contained separate datasets from the training data (different subjects, different ethnic backgrounds , viewing different commercials), ensuring that our system generalized well. The facial responses are all spontaneous and often subtle responses to watching commercials. As in training, images where only 1 out of 3 labelers labeled an asymmetric smile, were disregarded from our evaluation. The resulting set had 720 positive samples from 40 persons and 460,000 negative samples from 500 persons.

4) Performance measure: Because the test data had significantly more negative samples than positive samples, the area under the ROC curve was chosen as performance measure. This measure is desirable as it does not depend on ratio of positive to negative samples, unlike other measures such as a simple accuracy rate. By using the distance to the hyperplane of each sample and varying a decision threshold, we plot hit rate (true positives) against false alarm rate (false positives). The area under this curve is equivalent to the percent of correct detections in a 2-alternative forced choice task (2AFC), in which the system must choose which of two images contains the target.

B. Features and classifier comparison

Table II reports the results of experimenting with different features and classifiers for the RHT classifier. The random forest classifier parameters were tuned on the validation dataset in the same way that the SVM parameters were tuned. Table II shows that the HOG features outperforms other type of histograms like LBP or LGBP. For the classifiers, the SVM with a RBF kernel performs better than the linear SVM and SVM with a HIK. This justifies our choice of using HOG and a RBF kernel SVM classifier.

C. Posed versus spontaneous data

In this experiment, we evaluated the performance of our system when trained and then tested on combinations of posed and/or spontaneous data. The two subsets of the spontaneous data used to train and to test the classifier

TABLE II

COMPARISON BETWEEN DIFFERENT FEATURES AND CLASSIFIERS ON THE RHT CLASSIFIER PERFORMANCE. SEE SECTION V-B FOR ABBREVIATIONS

Features	Classifier	2AFC (%) on test dataset
HOG	RBF kernel SVM	**84.1**
	Linear SVM	83.4
	HIK SVM	83.5
	Random forest	79.1
HOG	RBF kernel SVM	**84.1**
LBP		70.4
LGBP		75.5
LGBP	HIK	77.7

TABLE III

2AFC SCORES WITH POSED AND SPONTANEOUS DATA.

Trained on	Tested on	
	posed	spontaneous
posed	98.5	81.0
spontaneous	87.8	81.1
both	96.2	84.1

were the same as in section V-A. For the case where we exclusively trained and tested on posed data, we used a 6-fold person-independent cross-validation: for each fold, we remove one sixth of the posed data from the training, to use as test data. Results are reported in table III.

As expected, classification scores on the posed data are consistently higher than that of spontaneous data. This highlights the difficulty of recognizing subtle spontaneous expressions in difficult conditions (out-of-plane head rotations, occlusions, bad lighting, etc.) compared to exaggerated posed expressions that only contain head motions (cf. section III-A). Cross-validated training and testing on posed data, led to a 2AFC score of 98.5%, highlighting that training and testing on the same dataset is an easier task than training and on disparate datasets. In contrast, the spontaneous data test was cross database: in each of the 10 datasets, viewers were watching a different commercial and were recorded in different countries / cultures. The resulting 2AFC scores reflected this challenging data. Finally, the best results for the spontaneous data (84.1%) were achieved by training with a mix of spontaneous and posed data. This important finding underscores the importance of having varied data for training, when the test data is heterogeneous.

D. Exploiting facial symmetry

We ran a number of experiments to test the concept introduced in this paper, namely comparing two classifier outputs, that obtained from a facial image and a flipped version of that image, summarized as follows:

- The RHT classifier: we flipped all the left hemiface asymmetric smiles in the test data, so that all the asymmetric expressions effectively occurred on the right hemiface, greatly simplifying the classification task.
- A classifier trained on both right hemiface and left hemiface asymmetric smiles (RLHT classifier). The flipped version of each asymmetric smile was also used to double the number of training samples.
- Two classifiers, one for each hemiface (RHT+LHT), were trained independently but from the same data. The final classification output was the max of both.
- The proposed system: an image and its flip are used separately as input to the RHT classifier and the difference between the two outputs is a measure of asymmetry.

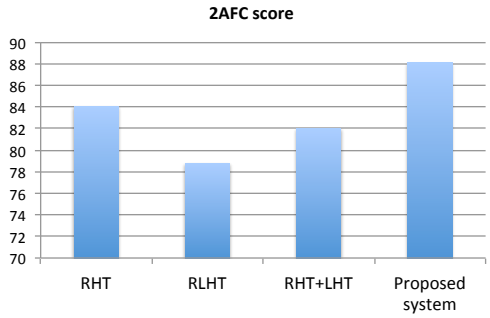

Fig. 3. Experimental results to recognize both right hemiface and left hemiface asymmetric smiles. See text for explanation of each system.

As shown in Fig. 3, using one classifier to learn the right and left asymmetric smiles (RLHT) led to the worst performance (78.8%). This may be due to the fact that the learned expressions had too much variation to be captured by a single classifier. Using two classifiers, one for the right side, and one for the left side, significantly increased the performance to 82.0%. Using the RHT alone gave better results (84.1%), due to the left-side asymmetric smiles being flipped in the test dataset, making the recognition task easier. In fact, the RHT classification score of 84.1% can be seen as an upper-bound of the RHT+LHT classifier performance.

In contrast, our proposed system achieved the highest overall results, with an 2AFC score of 88.2%. This score reflects the performance on the original asymmetric smile detection task, and outperformed the RHT classifier's performance on the simplified task. By applying a single RHT classifier on a facial image and its flipped version, we were able to exploit the relative symmetry of a face to increase the system's performance. Symmetric expressions were penalized, yielding a low asymmetric expression score, which resulted in a large improvement to the 2AFC score. When compared to the classical approach of one classifier trained on both sides (RHT+LHT), the proposed system scored almost 10% higher.

E. Is frame-by-frame detection good enough?

To decide if this detector could be used in a real-world application of understanding the viewer reaction to video content, we plotted the ROC curve and precision/recall curves (Fig. 4). While the 2AFC score was pretty high (88.2%), due to the very low ratio of positive samples to negative samples, the precision score is unacceptably low: 10% for a recall score of 40%. In other words, each time we predict a frame with an asymmetric smile, 9 times out of 10

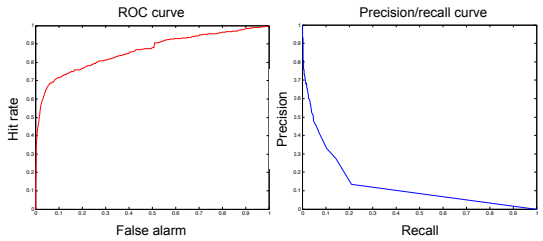

Fig. 4. ROC and Recall/precision curves on all frames of test data.

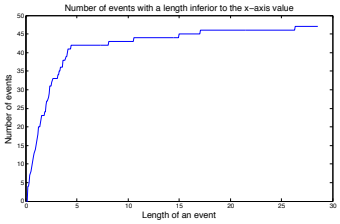

Fig. 5. Length distribution of asymmetric smile events.

we will have a false detection. To make things worse, 60% of true asymmetric smiles will be missed altogether.

With this performance, it is hard to see how the detector could be useful for analyzing viewer experience. We needed to reduce the number of false alarms. We observed that the majority of false alarms come from the same video, for which the detector reported a asymmetric smile for the whole video. This may occur for a variety of reasons: poor face tracker localization performance, large changes in head yaw ($>30\%$) or cases where only half the face is illuminated. Since these types of false alarms effect the entire video uniformly, the dynamics of the asymmetric smile can be leveraged to improve accuracy. As noted earlier, most of the asymmetric smiles only last a few seconds; they appear (onset phase) and disappear (offset phase) quickly. We concluded that detecting an asymmetric smile event (from the onset to offset) may improve the results.

VI. Event detection

A. Definition of an event

An event is defined by the portion of the video that spans the onset of an asymmetrical smile, followed by an apex through to an offset, (i.e. the resting state of the face would be symmetric). Using the FACS human labels, we define a positive and negative event as follows:

- A positive event starts and ends when 3 labelers agree that there is no asymmetric smile and requires that at least 2 labelers agrees on the presence of an asymmetric smile between these two moments.
- A negative event is a section of a video when all labelers agree that there is no asymmetric smile at all.

Using the test dataset presented in section V-A.3, we found 47 positive events in the labeled data, with event lengths shown in Fig. 5. It should be noted that 42 events of 47 had a duration less than 5 seconds. Making use of this prior knowledge, events occurring within a time window of five seconds were targeted. 42 positive events were obtained by centering a 5 seconds window around the 42 sections of videos representing an asymmetric smile. 180,000 negative events were obtained by taking overlapping sections of video in which no asymmetric smilee were labeled.

B. Template matching

To recognize if an event is positive or negative, we take the output of our frame-by-frame system over the full event, and fit the template shown in Fig. 2. This template of length $T_2 = 5$ seconds represents a centered box function of variable length T_1 with $0 < T_1 < T_2$. The amplitudes of the box are defined by 3 values: a, the value of the output before the peak, h with $h \geq a, b$, the value during the peak and b the value after the peak. The best fit, with optimized values of T_1, a, b, h , is found by minimizing the error between the template and the output of the detector. In practice, we start with one value of T_1, and compute the minimum error by setting a, b, h to the mean of the output over their respective section. After computing the error for all values of T_1, we select the value T_1 that leads to the overall minimum error.

Once fit, the parameter of interest is the value of the peak (h) in the output of our frame-by-frame system, alongside the presence of an onset and an offset. If the frame-by-frame output begins low (a low), rises (h high), and returns to a low level again (b low), this is likely that we have detected a real asymmetric smile. We can express this behavior as a single event detection score, s_d, by computing the harmonic mean between the height of the onset $h - a$ and the height of the offset $h - b$:

$$s_d = \sqrt{(h-a) * (h-b)} \tag{3}$$

TABLE IV

False alarm for a hit rate of at least 40%. We repot results for: all data and data for which the tracker was judged accurate by hum a labelers.

	Events with accurate tracking	All events
On a frame basis	0.81%	1.5%
Max within the event	0.23%	0.79%
Template matching	0.07%	0.23%

C. Experimental results

1) Event recognition scores: The distribution of event scores s_d for positive samples shows that 40% of the positive events have a score significantly higher than that of negative event. The remaining positive events yield scores which cannot be distinguished from negative events without the introduction of false alarms. Thus, we chose a threshold for s_d that leads to a detection rate of 40%. We compare our template matching approach (s_d) with a simpler approach, using the maximum value within a window. We report in Table IV results in two cases: for all data, and for data for which the tracker was judged accurate by human labelers

TABLE V

NUMBER OF VIDEOS, POSITIVE EVENTS (POS. EVENTS), HITS, MISS AND FALSE ALARMS (FA) IN 3 DATASETS.

	Videos	Pos. events	Hits	Miss	FA
Dataset 1	137	17	9	8	1
Dataset 2	167	8	5	3	3
Dataset 3	461	14	5	9	5
Total	765	39	19	20	9

Recall	49%
Precision	69%
False alarm rate	1 every 85 videos

(about 70% of the data). This shows that tracker inaccuracies are responsible for an important part of the false alarms. However, applying a max operator per 5 second window reduces the false alarm rate to half that of a frame-based approach. With our approach of template matching, the percent of false alarm is 1/7 of frame-based decision.

2) Event detection on independent datasets: To validate the system, the event detector is applied to a final set of spontaneous videos, grouped into three datasets. As was the case for our training and testing datasets, these three datasets are obtained by asking different viewers to watch different commercials, and none of them were not included in the training. The frame-by-frame detector is applied on the whole video to obtain the scores s. We then use a sliding window to compute the event score s_d. To convert s_d into a probability value between 0 and 1, we apply a soft-threshold in the form of a sigmoid function. The sigmoid was centered on the previously determined threshold that resulted in 40% correct detections in the test dataset.

To count the hits, miss and false alarms reported in Table V, we use the following procedure:

- We count a hit when a positive event shorter than 5 seconds happens and the event detector fires at any moment during the event.
- We count a miss when a positive event shorter than 5 seconds happens and the event detector does not fire during the event.
- We count a false alarm when the detector fires and none of the labelers detected a asymmetric smile.
- For events that are ambiguous (only one labeler detected the asymmetric smile or the event is really long), we ignore the output of the detector.

Based on three datasets, we observe that even if the number of events is rare, 1 in every 20 videos, we still achieve a 69% precision with a recall rate of 49% (for a F1 score of 0.57). This translates to 1 false detection every 85 videos. With these performance characteristics, we can use the system to detect asymmetric smiles on a large set of viewers and discover insights on which portions of a commercial induces skepticism. Fig. 6 shows the results on a single video. Note that the two smirk events are detected, while symmetric smiles are ignored.

We plot the aggregate curve for one of the three datasets, where 137 viewers in Germany watched a commercial for

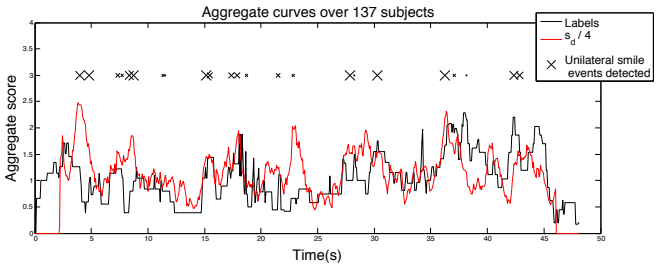

Fig. 7. Aggregate curves of dataset 1. The black curve is computed by adding all percents of labelers detecting a asymmetric smile for the 137 videos. The red curve is computed by adding the scores s_d. The crosses represent at what time we detect a asymmetric smile, and the size of the cross represents the probability of each event.

a consumer beverage that they had not seen before. We compare our system's output to ground truth (Fig. 7). For the ground truth, we sum the percent of labelers detecting an asymmetric smile across the 137 videos. We compare this curve with the aggregation of the output s_d of our classifier. The correlation score between these two curves is 0.4, and both curves show similar moments where the commercial induces more skepticism. Examples of this occur at seconds 4, 8 17, 28, 37 and 43. However, our classifier also shows a smirk event around 23 seconds too, which is not confirmed by human labelers. This false positive does not appear when alternate visualizations are explored: we plot a cross for each asymmetric event detected (after applying a soft-threshold on s_d). The size of each cross reflects the certainty that an asymmetric event is detected. By aggregating the response of 100 viewers, we reduce the effect of outliers and highlight the moments were people were skeptical about the commercial.

VII. CONCLUSION AND FUTURE WORK

This paper presents the first automated system for detecting asymmetric lip expressions, such as smirks, in spontaneous facial videos. While the majority of symmetric smiles portray happiness and enjoyment, asymmetric smiles indicate negative valence as well as discrete emotion states such as contempt, doubt and defiance. Accurate detection of these expressions yields valuable insight into how people engage with video content such as commercials or political debates.

Our approach reliably detects asymmetry in emotion expressions by exploiting the natural symmetry of the face with a classifier applied to both a face frame as well as the flip of that frame. We use dynamics, namely the detection of onset and offset of asymmetric facial events, to increase the robustness of our system to real-world conditions where the facial videos exhibit non-uniform lighting, or viewers are fidgeting and/or eating. We test the system on posed as well as spontaneous expressions, from a variety of countries and ethnicities including Caucasian, Asian and Hispanic. To the best of our knowledge, this is the largest cross-cultural dataset ever used to evaluate an automated facial analysis system.

Our C++ implementation of the frame-by-frame system runs in real-time at 80 frames per second on a Windows

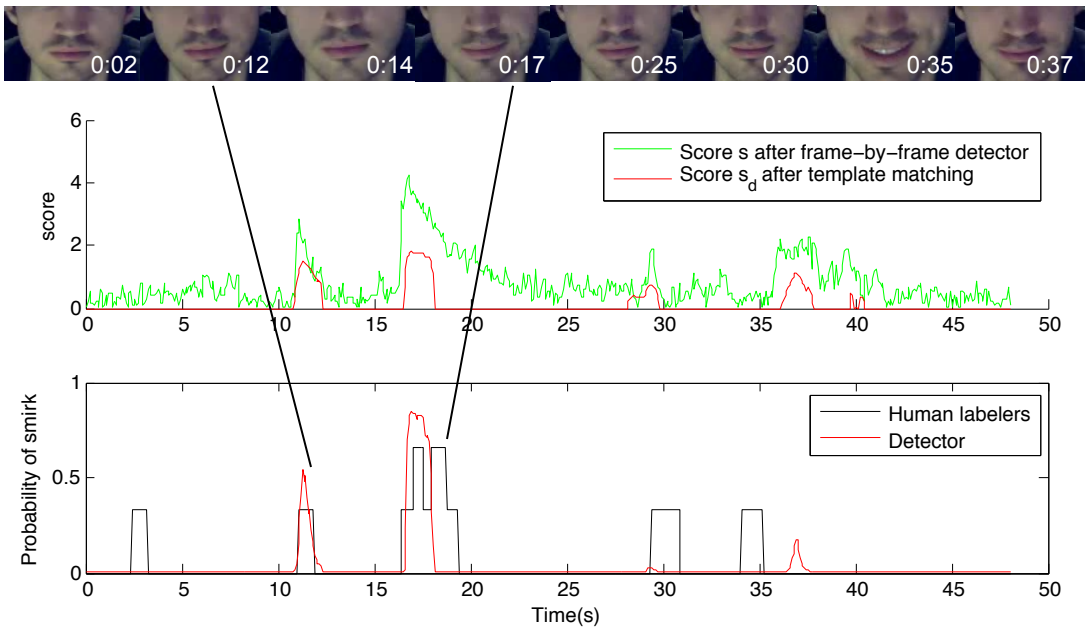

Fig. 6. Asymmetric smile detection on a whole video. Top row: Timestamped mouth regions extracted from the video. Middle row: the score s after the frame-by-frame detector and the score s_d after template matching. Bottom row: the percent of the 3 human labelers detecting an asymmetric smile and s_d after applying a soft-threshold (sigmoid function). Note that the two smirk events are detected, while symmetric smiles are ignored.

machine with an Intel Xeon CPU 3.20Ghz processor. The whole system, including the event detection, processes 70 frames per second with a delay of 2.5 seconds induced by the 5 second sliding window. Finally, we go beyond just reporting accuracy scores to additionally demonstrate how we use the results in a real-world video viewing application. We aggregate the reactions of over 100 viewers to identify moments of skepticism.

There are a number of exciting directions to take the work proposed in this paper. Future work includes testing this model on other asymmetric facial action units such as a wink or asymmetric eyebrow raise. We would also like to apply the idea of event detection to symmetric facial actions and emotion state classifiers. A systematic evaluation of the meaning of these expressions in the context of watching content is needed, e.g., do people report feeling skeptical or contempt when they express these expressions. Finally, we believe that this work can be applied to political polling where feelings of doubt and skepticism are common.

REFERENCES

[1] J.C. Borod, C.S. Haywood, and E. Koff. Neuropsychological aspects of facial asymmetry during emotional expression: a review of the normal adult literature. *Neuropsychology Review*, 7(1), 1997.

[2] J.C. Borod, J. Kent, E. Koff, C. Martin, and M. Alpert. Facial asymmetry while posing positive and negative emotions: Support for the right hemisphere hypothesis. *Neuropsychologia*, 26(5), 1988.

[3] J.C. Borod, E. Koff, and B. White. Facial asymmetry in posed and spontaneous expressions of emotion. *Brain and cognition*, 2(2), 1983.

[4] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition (CVPR)*, 2005.

[5] C. Darwin. *Emotions in man and animals*. 1872.

[6] P. Ekman. Darwin, deception, and facial expression. *Annals of the New York Academy of Sciences*, 1000(1), 2003.

[7] P. Ekman et al. *Asymmetry in facial expression.* American Assn for the Advancement of Science, 1980.

[8] P. Ekman and W.V. Friesen. *Facial action coding system: A technique for the measurement of facial movement*. 1978.

[9] P. Ekman, J.C. Hager, and W.V. Friesen. The symmetry of emotional and deliberate facial actions. *Psychophysiology*, 18(2), 1981.

[10] B. Fasel and J. Luettin. Recognition of asymmetric facial action unit activities and intensities. In *Int'l Conf. on Pattern Recognition (ICPR)*, 2000.

[11] Y. Hu, L. Chen, Y. Zhou, and H. Zhang. Estimating face pose by facial asymmetry and geometry. In *Int'l Conf. on Automatic Face and Gesture Recognition (FG'04)*, pages 651–656. IEEE, 2004.

[12] T. Kanade, J.F. Cohn, and Y. Tian. Comprehensive database for facial expression analysis. In *Int'l Conf. on Automatic Face and Gesture Recognition (FG'00)*, pages 46–53. IEEE, 2000.

[13] M. Kim, S. Kumar, V. Pavlovic, and H.A. Rowley. Face tracking and recognition with visual constraints in real-world videos. In *Computer Vision and Pattern Recognition (CVPR)*, 2008.

[14] Y. Liu, K.L. Schmidt, J.F. Cohn, and S. Mitra. Facial asymmetry quantification for expression invariant human identification. *Computer Vision and Image Understanding*, 91(1), 2003.

[15] D. McDuff, R. El Kaliouby, and R. Picard. Crowdsourcing facial responses to online videos. *IEEE Trans. on Affective Computing*, 2012.

[16] S. Mitra and Y. Liu. Local facial asymmetry for expression classification. In *Computer Vision and Pattern Recognition (CVPR)*.

[17] P.J. Phillips, H. Wechsler, J. Huang, and P.J. Rauss. The feret database and evaluation procedure for face-recognition algorithms. *Image and Vision Computing*, 16(5), 1998.

[18] W.E. Rinn. The neuropsychology of facial expression: a review of the neurological and psychological mechanisms for producing facial expressions. *Psychological bulletin*, 95(1), 1984.

[19] H.A. Sackeim, R.C. Gur, and M.C. Saucy. Emotions are expressed more intensely on the left side of the face. *Science*, 202(4366), 1978.

[20] B. Scholkopf and A.J. Smola. *Learning with Kernels*. 2002.

[21] T. Senechal, K. Bailly, and L. Prevost. Automatic facial action detection using histogram variation between emotional states. In *Intl Conf. on Pattern Recognition (ICPR)*, 2010.

[22] T. Senechal, V. Rapp, H. Salam, R. Seguier, K. Bailly, and L. Prevost. Facial action recognition combining heterogeneous features via multi-kernel learning. *IEEE Trans. on Systems, Man, and Cybernetics, Part B: Cybernetics*, 42(4), 2012.

[23] M.F. Valstar, M. Pantic, Z. Ambadar, and J.F. Cohn. Spontaneous vs. posed facial behavior: automatic analysis of brow actions. In *Int'l Conf. on Multimodal Interfaces*, 2006.