Automated Facial Expression Recognition for Product Evaluation

Abstract

Facial expression play an important role in signaling interest and opinion about a product. In this paper we present a negative affect classifier that detects nose wrinkles (AU9) or brow lowerer (AU4). Our approach uses an automatic mechanism for extracting the frames of interest in videos. We employ an automatic facial feature tracker to detect the area of interest in each frame, we then apply Gabor Jets for feature extraction followed by applying a dynamic Bayesian network classifier. We validate our classifier on two corpora: the first is of posed videos, the second is from a corpus of spontaneous videos of people evaluating two different beverages. We achieved an average accuracy of 97% for posed videos and 96% for spontaneous videos. We also show that negative facial expressions captured using our machine vision technique are more predictive than self-reports. This was proven by correlating between self reports and reports provided by human coders which is used as our ground truth for our classifier.

1. Introduction

Quantifying customers’ experience using machine vision is an emerging application nowadays. Many companies often conduct market research to build market share, competitive advantage and to predict how people would like their product. The problem with predicting customer’s experience is that the current evaluation methods such as relying on customers’ self reports is very subjective. People are not always feeling comfortable revealing their true emotions. They may inflate their degree of happiness or satisfaction in self reports. Participants report higher wellbeing in face-to-face interviews than they do in mail surveys or self-administered questionnaires [12]. This is because participants are unwilling to reveal their true emotions to strangers.

In case of interviews, this interviewer effect disappears when the interviewer is severely handicapped [13]. Participants would like to give a positive feelings to others but would rather not exaggerate when faced with another's unfortunate situation. This can show that self reports are very subjective and affected by external factors.

Due to the inaccuracy of self reports, market researchers are trying to find new channels by which they can capture the users’ affective states without asking them for their direct opinion. One of these expressive channels is facial expressions. Facial expressions are considered to be a very powerful means for humans to communicate their emotions [3]. Facial expression constitutes 55 percent of the effect of a communicated message [10] and is hence a major modality in affective state recognition.

In this paper we present a system for detecting negative affective states in a video corpus of naturally evoked videos from a study of people evaluating two different beverages. First, we were able to extract the frames of interest based on an automatic sip detection algorithm. We then localized the area of interest where the facial muscles change as a result of negative affective state. We then applied Gabor Jets for feature extraction and dynamic bayesian networks for classification. Finally, we compared the affective state generated from the system and the affective state generated from manual coding of sips against participants’ self reports to show that self reports are subjective and an automated technique is needed for better prediction.

In this paper, we were faced with the challenge of tagging hours of videos to quantify participants’s reactions to sipping different beverages. The paper makes three principal contributions: (1) We present a system for detecting negative affective state by first detecting frames of interest in videos and then applying machine perception to the automated detection of negative affect; (2) Testing and training on thousands of posed and spontaneous images with an accurate detection of negative affective state in the presence of substantial head motion and occlusions, changes in lighting (most facial analysis systems are not tested on natural videos); and (3) an analysis of participants’ self reports that shows that self reports are not accurate in conveying affective state and are affected by previous experience. Our approach can be generalized and applied to detect positive affective states by detecting other AUs such as cheek raiser (AU6), Lip Corner Puller (AU12), Cheek Puffer (AU13)

The paper is organized as follows: section 2 surveys related works in the area of emotion detection in video. Section 3 discusses the corpus of videos from a customer satisfaction study. Section 4 gives an overview of our methodology for detecting negative affective states in videos. Sec-

replacemeinfinal/09/$25.00 ©2009 IEEE
tion 5 and section 6 describes the feature extraction and classification techniques we used in this research. Experimental results and evaluation are provided in section 7. Section 8 provides some interesting results about self reports and its correlation with reports from human coders. Section 9 concludes the paper and outlines future direction of research in the area of detecting affective states based on different action units.

2. Related works

To capture facial expressions in videos, expressions are classified according to some scheme. The most prevalent approaches are based on the existence of six basic emotions (anger, disgust, fear, joy, sorrow and surprise) as argued by Ekman [4] and the Facial Action Coding System (FACS), developed by Ekman and Friesen [5], which codes expressions as a combination of 44 facial movements called Action Units.

The first step in detecting emotions in videos is to accurately detect action units. There are a number of systems that are implemented for detecting action units with high accuracy. Bartlett et al [1] present one of the most successful systems for detecting AUs using Gabor filters followed by support vector machines (SVMs). Faces are first localized, scaled to 96x96 pixels and then passed through a bank of Gabor filters before classification. Vural et al [14], improved the work done by Bartlett et al. [1] by retraining the system on a larger dataset. They reached an accuracy of 93% on posed images and 75% on spontaneous images. The problem with these two systems is that accuracy of AU detection decreases as the training sample decreases. Another disadvantage is that they are using SVMs and as a result they are not making use of the temporal information and the relation between the different frames in videos.

This paper extends action unit detection by accurately nose wrinkle (AU9) or brow lowerer (AU4) in thousands of spontaneous and posed images with substantial degrees of head motion and changes in the lighting. We also propose the use of dynamic Bayesian Networks with only 92 training images which results in average accuracy of 97.2% in posed videos and 96.5% in spontaneous videos.

3. A video corpus of naturally-evoked videos

The corpus of naturally-evoked videos were collected from a sipping study that was conducted in collaboration with a major beverage company. Thirty–five customers were recruited (equal number of males and females), each customer is seated in front of a laptop (with a built–in web-cam) and given a choice of two beverages that were located on the left and right of the laptop. A customer is then asked to take a sip of one of the beverages and answer several online questions about their experience. The customer has to write down a number between 1 and 9 to indicate whether s/he liked or disliked the drink where 1 indicates totally dislike, 5 indicates neutral and 9 indicates totally likes. This sequence of sipping then answering questions were repeated 30 times, for an average duration of 30 minutes per customer.

We note that while the customers were aware of being recorded, they were not given any instructions to limit their face or body motion. As a result, there is considerable head and body motion in all the videos, especially as the customers turned to pick a beverage and there is substantial individual variances in expressiveness.

We were interested in looking at the time window after sipping one of the two beverages. We hired two human experts to code the facial valence for each trial, into positive (1) /neutral (0) / negative (-1) responses.

Focusing on the most expressive subset of the participants, we analyzed ten videos out of this corpus. According to the two coders’ data, these ten videos contain 300 sips. 113 out of the 300 sips represent a positive affective state, 105 sips represent a neutral affective state and 82 sips represent a negative affective state. An example of a sip that has a negative affective state can be shown in Fig. 2 and another sip that has a positive affective state can be shown in Fig. 3.

We noticed that 11% of the sips that represent negative affective state contains nose wrinkles. Fig. 4 shows the percentage of different action units that appeared on the 82 dif-
different sips representing negative affective states as reported by the two coders. Note that nose wrinkles represent the second most frequent action units in these sips. The first most frequent state is the frown state. However, frown is a general state that indicates a negative affect. It is considered to be a combination of different action units that may include inner brow raiser (AU1), outer brow raiser (AU2), upper lid raiser (AU5) and nose wrinkles (AU9). We are focusing our research in predicting negative states based on detecting nose wrinkle (AU9) or brow lowerer (AU4). Fig. 1 shows examples of negative affective states in posed videos that resulted from the presence of either nose wrinkles (AU9) or brow lowerer (AU4).

3.1. Coding application

To set the ground truth that we will compare the results of our classifier against, we recruited another four coders to code the videos of interest using our coding application shown in Fig. 5.

Since it would be a tedious and time consuming task to ask each coder to code videos that last for thirty minutes each and since these videos have lots of irrelevant information to our research, each video is processed using our automatic sip detection algorithm[] to compute the range of frames where each sip occurred. We then add two seconds before and after each sip to capture the facial expressions that occurred just before and after the sip. Applying the automatic sip detection algorithm[] will make the coders code 2-4 minutes of relevant information instead of coding 30 minutes as shown in Fig. 6.

After reducing the video to only the sipping frames, each sip will be viewed to the user along with the images of 22 action units as described in the Facial Action Coding System (FACS)[5]. The video is viewed in slow motion so that the coders can view the different action units and they are given the ability to repeat the sip. After the sip is over, the coders have to say whether the affective state of the participant was positive or negative or neutral.

4. Approach

Since the purpose of this research is to detect the affective state of the participants after taking sips from the drinks, we found that the video frames that are relevant to our approach are the ones that are just before and after the sip event. Consequently, we processed the video using the implementation of automatic sip detector proposed in[]. This sip detector makes use of the head gestures detected by a network of Hidden Markov Models combined with event information that is automatically logged by the sip application to form a semantic representation of the sipping events.

As shown in Fig. 7, we present a multi-level approach to detect negative affective states in video. After extracting the relevant frames from the videos using the automatic sip detector, features are then extracted from these frames of interest and then fed into the affective classifier.

Since the purpose of this research is to deduce the affective state of the participant through detecting AU9 or AU4, there was no need to extract all the facial features and train the action unit classifier on features that are not relevant.
Consequently, for every incoming frame of the video, we first locate the region of interest, area above the nose and between the two eyes. This region is then passed through a bank of Gabor Jets to generate distinctive features which are then fed into a classifier.

For classification, we used Dynamic Bayesian Network (DBN) to make use of the temporal information found in the videos. The output of the classifier will then be passed to the affective state classifier which will state whether the participant has a positive or negative state. Finally, the output of the affective state classifier is compared against the participant’s self-report collected from the sip application. It is also compared against four coders’ reports which are collected from our coding application.

4.1. Negative affective state position localization

To detect faces in an image, we used Google’s face-tracker [6]. The tracker uses a generic face template to bootstrap the tracking process, initially locating the position of 22 facial land-marks including the eyes, mouth, eyebrows and nose. We use the feature point located at the top of the nose to locate the region where we can test for the presence or absence of negative affective state as shown in Fig. 8. The feature point located at the top of the nose represents the center of the X coordinate of the area of interest. This feature point is also considered to be the minimum Y coordinate of the area of interest. We then draw a rectangle out of the given information and start processing this rectangle.

5. Feature extraction

After generating the images of interest from the video frames, we wanted to extract features from these images to be used in our classification. We decided to use Gabor filters to account for the substantial degree of head motion which gives different scales and orientations for the face in videos. Gabor filters convolve the image with a Gaussian function multiplied by a sinusoidal function. The Gabor filters are considered to be orientation and scale tunable edge detector. The statistics of these features can be used to characterize the underlying texture information [9]. The Gabor function is defined as

\[ g(t) = ke^{i\theta}w(at)s(t) \]

Where \( w(t) \) is a Gaussian function and \( s(t) \) is a sinusoidal function. To make sure that Gabor filters will be a discriminative feature extractor, we applied Gabor filters convolution on two images which can be shown in Fig. 9. From the images and the values obtained, it is noticed that Gabor filters are in fact discriminative feature extractor.

One major disadvantage of Gabor filters is that they are computationally expensive, making it difficult to be applied in real-time applications [7]. To detect video images in real-time, we decided to use Gabor Jets which describe the local image contrast around a given pixel in angular and radial directions [7]. Gabor Jets are characterized by the radius of the ring around which the Gabor computation will be applied. We chose the center of our Gabor Jets to be the center...
of pupil. We experimented with both images of size 3x3 and images of size 4x2 on posed videos. The extracted image is then passed to the Gabor filters with 4 scales and 6 orientations to generate 216 features in case of 3x3 or 192 features in case of 4x2 representing the magnitude of the Gabor filters.

6. Training and classification

6.1. Training

To train our classifiers, we chose 92 images captured from two different videos representing two different participants. One of the videos is posed while the other is spontaneous video. 46 out of the 92 images were representative set of the existence of nose wrinkles and the other 46 were representative set of the absence of nose wrinkles.

To make sure that the Gabor Jet features extracted from each image are representative of the class they belong to, we applied K-Means algorithm, an unsupervised learning algorithm, to group the training images into two clusters. This will help us make sure that all the images representing nose wrinkles will be grouped in one cluster and all the images representing the absence of nose wrinkles will be grouped in another cluster. The output of the K-Means algorithm is shown in Fig. 10. The figure shows that some of the images that were supposed to be a training examples of nose wrinkles were clustered with the examples representing the absence of nose wrinkles. We removed the images that were not accurate and reapplied K-Means algorithm to make sure that all the nose wrinkle images belong to the same cluster and so do the images that represent the absence of nose wrinkles. The results of applying this process several times is shown in Fig 11.

We then computed the mean and average of each Gabor feature, which constituted the training set for our Dynamic Bayesian Network classifiers.

6.2. Dynamic bayesian networks (DBN)

To make use of the temporal relations between AUs, we experimented with Dynamic Bayesian Networks. DBNs are defined by number of hidden states (N) and number of observed states (M) and number of parameters \( \lambda_j = (\pi; A) \):

- \( N \), the number of states in the model \( S = \{S_1, ..., S_N\} \); \( S_1 \) is a discrete hidden node representing whether there is a nose wrinkle or not. This is because the state of the presence or the state of the absence of a nose wrinkle are complementary. \( S_2, S_3, ..., S_N \) are continuous observed states representing the features generated from the Gabor Jets;
- \( A = \{a_{ij}\} \), is an N x N matrix to represent the topology of the network where \( a_{ij} = 1 \) indicates an edge from node i to node j. The structure of our the dynamic Bayesian Network is shown in Fig. 12;
- \( \pi_i \) is the probability of state i in case of a discrete node such as the node representing the presence or the absence of nose wrinkle. \( \pi_i \) can also be the mean and variance of state i as in the case of continuous nodes representing Gabor features.

The probability of the presence of a nose wrinkle is defined by

\[
P(X_1, X_2, X_3, ..., X_n) = \prod_{i=1}^{n} P(X_i|\text{parents}(X_i))
\]

We have to also, define the inter relation between the different time slices. In our case, we found that our hidden node at time t is dependent only on the hidden node at t-1. The model is given by the joint probability distribution:

\[
P(X_i, Y, \Theta) = P(Y|X_i, B_\theta)P(X_i|A, \Pi)
\]

Where \( \Theta \) represents the number of time slice and the observation function \( B_\theta \) is parameterized by the conditional probability distribution that model the dependency between the two nodes. We detected the desired AU based on 5 previous time slices.
Table 1. Results of applying our DBN classifiers on Gabor Jet features generated after convolving on 3x3 and 4x4 images.

<table>
<thead>
<tr>
<th>AU</th>
<th># images</th>
<th>3x3 Gabor Jet</th>
<th>4x2 Gabor Jet</th>
</tr>
</thead>
<tbody>
<tr>
<td>wrinkle = 1</td>
<td>672</td>
<td>True 90.4</td>
<td>True 97.2</td>
</tr>
<tr>
<td>wrinkle = 0</td>
<td>478</td>
<td>True 98.3</td>
<td>True 96.3</td>
</tr>
<tr>
<td></td>
<td>1150</td>
<td>97.2</td>
<td>1050</td>
</tr>
</tbody>
</table>

7. Experimental evaluation

We have created a large database of 1200 images taken from posed videos. The videos used for testing were recorded at 30 frames per second, and last between 5 to 8 seconds, compared to a mean duration of 0.67 seconds per sequence in the Cohn-Kanade database [2]. The resolution is 320x240. The videos were acted by 30 actors of varying age ranges and ethnic origins. All the videos were frontal with a uniform white background. The process of labeling the videos involved a panel of 10 judges who were asked could this be the emotion name? When 8 out of 10 judges agreed, a statistically significant majority, the video was included. We selected the images from the disgusted and deserted databases where nose wrinkles are found in the disgusted emotion videos and absent in the deserted emotion videos.

We used different images for testing than those used for training. The chosen images for testing have head pitches which range from -6.4 to 17.0 degrees, head yaws which range from -19.34 to 27.25 and head rolls which range from -20.0 to 16.2. We tested our methodology on 1150 images drawn from 9 different videos representing three different participants. The classifier was trained on images extracted from the first participant only.

We experimented our approach with Gabor Jets on images of size 4x2 and 3x3. Table 1 shows the accuracy of applying our DBN classifier to the extracted images and Fig. 13 shows the accuracy of the DBN classifier for each participant. Table 1 shows that the accuracy of applying the Gabor Jets on 3x3 are higher than applying Gabor Jets on 4x2.

Table 2. Results of applying our approach on spontaneous videos.

<table>
<thead>
<tr>
<th>Subject</th>
<th>presence of wrinkle</th>
<th>absence of wrinkle</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>sips</td>
<td>images</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>15</td>
<td>100</td>
</tr>
<tr>
<td>3</td>
<td>24</td>
<td>21</td>
</tr>
</tbody>
</table>

After getting an average accuracy of 97.2% on posed videos, we decided to generalize our approach and test on spontaneous videos taken from our corpus mentioned in section 3. We focused our experimentation on Gabor Jets applied to 3x3 images since it gave better accuracy than 4x2. We made use of our coding application to know which videos and which sips have nose wrinkles (AU9). We tested our approach on 15 sips taken from the videos of three different subjects. 5 out of the 15 sips have nose wrinkles as coded by the coders and the rest of sips are used to make sure that the absence of nose wrinkles are detected by our approach on spontaneous videos. Our hypothesis is that the presence of nose wrinkle will indicate that the participant did not like the drink and this will be compared against the participant’s self report and against four coders’ reports. Table 2 shows the accuracy of our classifier in detecting nose wrinkles for 3 different subjects. The four coders agreed that the participants did not like the drink in the five sips where there were nose wrinkles.

8. Discussion

Coding facial expressions by human coders is a subjective and labor-intensive process, but this step is necessary to build a more objective classification system by a computer. The two human experts who were recruited to code the facial valence of each trial, did not perfectly agree with each other. For the 10 participants we studied, the two human coders had moderate agreement in terms of Cohens Kappa for classifying sip facial valence responses: Observed agreement = 66.7% with Cohen’s kappa = 0.4595 (meaning moderate agreement) [8].

Relying on customers’ self reports to measure customers’ experience and product satisfaction is not considered to be a very accurate measure. We found out that there are two reasons why self reports are not an accurate measure based on our study of our corpus. First of all, they are not actually reflecting the affective state of the users that can be deducted from their facial expressions. We compared the self reports against reports given by coders. We found that participants are not very accurate in expressing their affective state. We applied Pearson correlation [11] between coders’ reports and participants’ self reports given by the following equation.

\[ \rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \cdot \sigma_Y} \]
We got a Pearson coefficient of 0.375 and since the value is closer to zero, this indicates that self reports and coders’ reports are independent.

Another problem with self reports is that they are affected by the previous trials. If the customer liked one of the two drinks, s/he keeps on giving positive feedbacks on both drinks and vice versa. This was found in four out of the examined 10 videos. Fig. 14 shows that while the coders’ data are fluctuating based on the participant’s facial expression, the participant keeps on giving positive feedbacks on both drinks although he liked one of the two drinks only.

These two findings are in fact supporting our point that self reports are not a good measure and we need a machine vision technique to automatically detect affective states.

Our approach misclassified some of the images used for testing. The reason why there are some misclassified images is because our methodology depends mainly on the accuracy of the tracker. Since the center of the Gabor Jet is determined by making use of the location of one of the feature points generated by the tracker, any substantial drift in this feature point will result in misclassification of the nose wrinkle (AU9).

9. Conclusion and future directions

This paper describes a methodology for detecting negative affective state by detecting nose wrinkles. Detecting affective state is important for many different applications such as market survey and human-computer interaction. We presented the results of applying DBN classifiers on Gabor Jets features. We reached an average accuracy of 97.2% in posed videos and 96.5% in spontaneous videos. We showed that our approach works better for detecting affective state than using self report. This was proved by comparing coders’ reports and participants’ self reports.

Our next step is to work on detecting other action units as shown in Fig. 15 that can help in inferring affective states. We will manage to detect inner brow raiser (AU1), outer brow raiser (AU2) and upper lid raiser (AU5) to be able to capture the frown state. Our approach can be generalized to detect positive affective state. We will manage to detect action units such as cheek puffer (AU13), cheek raiser (AU6) and lip corner pull (AU12) which can act as predictors for positive affective state.

Figure 14. The figure shows the persistence of the self report over the 30 sips. The red and green charts show the affective state captured by the coders based on the participant’s facial expressions.

Figure 15. negative affective state action units: Inner Brow Raiser (AU1), Outer Brow Raiser (AU2), Upper Lid Raiser (AU5). Positive affective state action units: cheek raiser (AU6), Lip Corner Puller (AU12), Cheek Puffer (AU13)

References