Detection of asymmetric eye action units in spontaneous videos

The MIT Faculty has made this article openly available. Please share how this access benefits you. Your story matters.

Citation

As Published
http://dx.doi.org/10.1109/ICIP.2009.5414341

Publisher
Institute of Electrical and Electronics Engineers

Version
Final published version

Accessed
Sun Apr 24 21:20:36 EDT 2016

Citable Link
http://hdl.handle.net/1721.1/60000

Terms of Use
Article is made available in accordance with the publisher’s policy and may be subject to US copyright law. Please refer to the publisher’s site for terms of use.

Detailed Terms

Please share how this access benefits you. Your story matters.
DETECTION OF ASYMMETRIC EYE ACTION UNITS IN SPONTANEOUS VIDEOS

Mina Mikhail
American University in Cairo
Computer Science Department
113 Kasr Al Aini Street
Cairo, Egypt
minamohebn@gmail.com

Rana el Kaliouby
Massachusetts Institute of Technology
Media Laboratory
20 Ames Street
Cambridge MA 02139 USA
kaliouby@media.mit.edu

ABSTRACT
With recent advances in machine vision, automatic detection of human expressions in video is becoming important especially because human labeling of videos is both tedious and error prone. In this paper, we present an approach for detecting facial expressions based on the Facial Action Coding System (FACS) in spontaneous videos. We present an automated system for detecting asymmetric eye open (AU41) and eye closed (AU43) actions. We use Gabor Jets to select distinctive features from the image and compare between three different classifiers—Bayesian networks, Dynamic Bayesian networks and Support Vector Machines—for classification. Experimental evaluation on a large corpus of spontaneous videos yielded an average accuracy of 98% for eye closed (AU43), and 92.75% for eye open (AU41).

Index Terms—Gabor Jets, Dynamic Bayesian Networks (DBN), Support Vector Machines (SVM), Action Units (AU), Spontaneous video

1. INTRODUCTION
Over the past decade there has been an increasing surge of interest in automated facial expression analysis. The majority of these efforts describe facial movements using the Facial Action Coding System (FACS) [1], a catalogue of 44 unique action units (AUs) that correspond to each independent motion of the face. It also includes several categories of head and eye movements. FACS enables the measurement and scoring of facial activity in an objective, reliable and quantitative way. It can also be used to discriminate between subtle differences in facial motion. For these reasons, it has become the leading method in measuring facial behavior. Human trained FACS coders are very adept at picking subtle or fleeting facial actions, which communicate a wide range of information including when a person is lying, depressed, or about to have an epileptic seizure. However, FACS-coding requires extensive training and is a labor intensive task. It takes almost 100 hours of training to become a certified coder, and between one to three hours of coding for every minute of video.

2. RELATED WORK
Bartlett et al [2] 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. The accuracy of AU detection decreases as the training sample decreases. Vural et al [3], improved the work done by Bartlett et al. [2] by retraining the system on a larger dataset. They reached an accuracy of 93% on posed images and 75% on spontaneous images. Tian et al. [4] detect eye state AUs for frontal images by applying Gabor filters on three points of each eye and then feed the results of the Gabor filters into a neural network.

This paper extends eye state detection by accurately detecting eye states in thousands of spontaneous images with substantial degrees of head motion and changes in the light-
As shown in Fig. 1, we present a multi-level approach to detect asymmetric eye open or eye closed AUs in video. Since the purpose of this research is to differentiate between eye open and eye close, there was no need to extract all the facial features and train the eye open/close classifier on features that are not related to the eye. For every incoming frame of the video, we first locate the left and right eye regions. The regions are then passed through a bank of Gabor Jets, which are then fed into a classifier. We used three different classifiers for comparison: static Bayesian network, Dynamic Bayesian Network (DBN) and support vector machines (SVM). Congruency between classification results of the left eye and right eye determines the presence of asymmetry or not.

4. EYE REGION LOCALIZATION

In order to detect faces in an image, we used Google’s face tracker. The tracker uses a generic face template to bootstrap the tracking process, initially locating the position of 22 facial landmarks including the eyes, mouth, eyebrows and nose. We use the eye brow, inner and outer eye corner feature points to locate the eye region as shown in Fig. 2. From each frame, we extract two images representing the left and right eyes. The eye brow feature point represents the maximum Y coordinate for each eye image. The minimum Y coordinate is the reflection of the eye brow feature point on right pupil feature point. The inner and outer eye corner feature points represent the maximum and minimum X for the eye rectangle.

5. FEATURE EXTRACTION

After generating the eye images from the video frames, we wanted to extract features from these images to be used in our classification. Since the videos have substantial degrees of head motions which give different scales and orientations for the face, we decided to use Gabor filters. Gabor filters convolves 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 [5].

One major disadvantage of Gabor filters is that it is computationally expensive, making it difficult to be applied in real-time applications [6]. 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 [6]. 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. So an image of 3x3, as shown in Fig. 3, is passed to the Gabor filters with 4 scales and 6 orientations to generate 216 features representing the magnitude of the Gabor filters.

6. CLASSIFICATION

In order to train our classifiers, we chose 80 eye left images captured from spontaneous videos. 40 out of the 80 images were a representative set of eye open and the other 40 were representative set of the eye closed. We experimented our approach with three different classifiers.

6.1. Static Bayesian Network

We created a Bayesian Network for detecting the eye open or eye closed AU. The Bayesian Network is 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 the eye is open or closed. $S_2, S_3, ..., S_N$ are continuous observed states representing the Gabor Jets generated features;
- $A = \{a_{ij}\}$, is an $N \times N$ 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 Bayesian Network is shown in Fig. 4;
- $\pi_i$ is the probability of state $i$ in case of a discrete node or the mean and variance of state $i$ in case of a continuous node.
**Table 1.** Results of applying the three classifiers to the eye images.

<table>
<thead>
<tr>
<th>AU</th>
<th># images</th>
<th>BN True %</th>
<th>DBN True %</th>
<th>SVM True %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Open</td>
<td>1919</td>
<td>1780</td>
<td>92.7</td>
<td>1809</td>
</tr>
<tr>
<td>Closed</td>
<td>150</td>
<td>147</td>
<td>98</td>
<td>140</td>
</tr>
<tr>
<td></td>
<td>2069</td>
<td>1927</td>
<td>93.1</td>
<td>1949</td>
</tr>
</tbody>
</table>

**6.2. Dynamic Bayesian Network (DBN)**

In order to make use of the temporal relations between AUs, we experimented with Dynamic Bayesian Networks instead of static Bayesian networks. DBNs are defined in the same way like static Bayesian network with some extra parameters. First, we have to 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:

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

**6.3. Support Vector Machines (SVMs)**

Another classifier that was experimented with is an SVM classifier. SVMs view the input data, 216 Gabor Jet features, as two sets of vectors in a 216-dimensional space. SVM will construct a separating hyperplane in that space that maximizes the margin between the two data sets. A good hyperplane will be the one that has the highest distance to different points in different classes [7]. We trained one SVM for detecting eye open or eye closed. In our research we experimented with linear kernels and compared the results of applying SVM with the results obtained from static Bayesian Networks and Dynamic Bayesian Networks.

**7. EXPERIMENTAL EVALUATION**

**7.1. Classification Results**

We have created a large database of images taken from spontaneous videos with an average duration of thirty minutes. The videos used for testing are from a sip study that Affective Computing at MIT Media Laboratory conducted in collaboration with major beverage company. Each participant is seated in front of a laptop (with a built-in webcam) and given a choice of two beverages that were located on the left and right of the laptop.

We used different images for testing than those used for training. The chosen images have head pitches which range from -9.2 to 12.23 degrees, head yaws which range from -60 to 65 degrees.
16.8 to 26.9 and head rolls which range from -5.0 to 7.0. We tested our methodology on 2100 images and Table 1 shows the accuracy of applying the three different classifiers on eye open and eye closed images. It is obvious that the accuracy of the DBN is the same like that of the BN. This is because the hidden node is dependent on 216 observed node in case of BN and 221 observed nodes in case of the DBN which includes the Gabor features and the five temporal nodes of the previous time slices. The effect of the extra five nodes, in case of the DBN, will have a minor effect on probability of the hidden node compared to the other 216 nodes.

To ensure that our approach is general and can be used on participants that the classifier is not trained on, we tested on three participants whose images are not used in the training. Fig. 5 shows the accuracy of the three classifiers for each participant. The images that were selected for training were extracted from the video of the first participant only. However, we used different images from the video of the first participant for testing.

We also, trained our classifiers on left eye images only and used 400 right eye images for testing on the same classifier to make sure that we do not need a separate classifier for the right eye. The results of applying the classifier on the right eye images shown in Fig. 6 shows that our classifiers work well even if they are trained on left eye images only.

7.2. Discussion

Our methodology depends mainly on the accuracy of the tracker. Since the center of the Gabor Jet is determined by one of the feature points generated by the tracker, any substantial drift in this feature point will result in misclassification of the eye images.

Our approach can be easily generalized to detect other AUs as shown in Fig. 7. For instance, we can easily detect mouth open (AU27) by making the center of the Gabor Jets in the center of the mouth. We can also, detect the presence of a frown (AU2) by making the center of the Gabor Jets at the center of the forehead.

8. CONCLUSION

This paper describes a methodology for differentiating between eye open (AU41) and eye closed (AU43). Detecting such AUs is important for different applications such as driver state monitoring and health application monitoring. We presented the results of applying three different machine learning classifiers on Gabor Jets features. We reached an average accuracy of 98% for eye closed and 93% for eye open. Our next step is to test our methodology on different AUs such as mouth closed and mouth stretch. And in order to account for the inaccuracy of the tracker feature point, we will work on increasing the size of Gabor Jets to 5x5 or 7x7 or 9x9. This will require using a feature extraction algorithm such as Adaboost in order to reduce the training and inference time and to be able to apply our approach in real-time.

9. ACKNOWLEDGMENTS

The authors would like to thank Hyungil Ahn and Rosalind W. Picard for making the corpus available for this work. The authors would also like to thank Ahmed Sameh, Joshua Gluckman for their help in this research and Google for making the face tracker available to our research.

10. REFERENCES