Detection of facial microexpressions

Gabor Revy
Department of Measurement and Information Systems
Budapest University of Technology and Economics
Budapest, Hungary
revy.gabor@gmail.com

Gabor Hullam
Department of Measurement and Information Systems
Budapest University of Technology and Economics
Budapest, Hungary
hullam.gabor@mit.bme.hu

Daniel Hadhazi
Department of Measurement and Information Systems
Budapest University of Technology and Economics
Budapest, Hungary
hadhazi@mit.bme.hu

***Abstract*—Facial microexpressions are instantaneous features signaling various details regarding the emotional and mental state of human beings. A key property of such features is that their interpretation as signals is the same or closely similar for all people. Currently, their detection requires a human expert. The automation of this task would allow a more widespread use. In this paper, we propose a hybrid solution, which is based on a framework of landmark points identified by a machine learning-based method. Upon this, we designed an expert system which utilizes image processing and signal processing algorithms such as homomorphic filtering, RANSAC parabola fitting, Hessian based shape analysis and change detection in order to identify microexpression features such as gaze detection and eyebrow raising. We evaluate these algorithms in real videos and pictures, and examine their applicability in practical scenarios. Our long-term goal is to detect complex facial expressions and emotions with the help of the detected microexpressions.**

***Index Terms*—microexpression, image processing, landmark points, expert system, facial expressions**

I. Introduction

Microexpressions are the visible features of emotions appearing on the face for a very short time, e.g. an involuntary reaction to a question. Automating the detection of facial expressions would allow a wide range of uses, e.g. to study reactions to an advertisement or to assist in the diagnosis of mental disorders. Experts usually distinguish 7 different basic emotions: anger, disgust, fear, happiness, sadness, surprise and contempt. In order to categorize reactions in videos, proper detection of microexpressions is required. In this paper, we propose algorithms to detect some of them. The first step is to detect the motion of the muscles, the so-called action units on the face. These parts of the face are described in detail in the FACS system [1]. Emotions can be determined based on the activated action units. The proposed methods estimate gaze direction, detect blinking and eyebrow raising by utilizing the output of an open-source landmark detection method [3].

This research was supported by the ÚNKP-20-5-BME-92 New National Excellence Program of the Ministry for Innovation and Technology from the source of the National Research, Development and Innovation Fund, and the János Bolyai Research Scholarship.

Fig. 1: Layout of the landmark points

A. Detection: machine learning vs expert system approach

There are two main approaches to detect microexpressions: machine learning based and expert image processing based systems. Machine learning based solutions can be robust if the training data is properly annotated and has adequate diversity. However, in the current scenario such a dataset is not available, and the available emotion detection models are not accurate enough. In an expert systems, several image processing algorithms are utilized to make a prediction from an image or a video. Such methods require a lot of fine tuning, most of which are arbitrary. We propose a hybrid solution: using a pre-trained, machine learning-based system, landmark points are established as shown in Figure 1. Based on the landmark points, image processing methods are applied to detect facial features and muscle movements.

II. Gaze detection

The main parts of estimating gaze detection are the localization of the eye and the iris. It also includes blink detection which can be an essential feature in the identification

of surprise, disgust or fear. The proposed solution for gaze detection was based on an open source gaze estimator called EyeTab [8] which was modified in order to increase robustness to changes in illuminations, to the variance in the visible size of the eyes.

A. Eye localization

In the first step, the regions of interest (ROIs) corresponding to eyes are determined. This was based on OpenCV Haar-like feature based cascade classifiers [7] in the EyeTab method. However, our initial analysis indicated that the ROIs defined by the bounding box of eye specific landmark points, provide more accurate results. Thus we modified EyeTab to utilize the polynom defined by the eye landmark points.

B. Pupil detection

The pupil is detected by combining two algorithms: the first is a gradient based [5] while the second is an isophote based [6] approach. In addition, the original input (the red channel of the RGB image) is replaced with the homomorphic filtered grayscale image, which resulted in more accurate pupil predictions.

1) Homomorphic filter: Homomorphic filtering can be used to compensate inadequate lighting, for example a shaded eye socket. Only the energy of low frequency components are reduced, therefore fast intensity changes (e.g. wrinkles) are preserved. It can be formulated as follows:

$$\mathbf{Im}_{homomorphic} = T(\exp(\log(\mathbf{Im}) - \log(\mathbf{Im}) * \mathbf{G}_\sigma))$$
$$T(x) = \min(x, 1)$$

Here, $\mathbf{G}_\sigma$ is the 2 dimensional, isotropic Gaussian function with σ parameter and $*$ means convolution. T cuts the values under 0 and over 1.

C. Blink detection

The pupil detection algorithm proposes a region of interest not only when the eyes are open, but also when the eyes are closed or are blinking. In order to detect blinking, a local Hessian based image analysis is applied to distinguish true and false ROIs proposed by the pupil detection method.

A typical pupil in a grayscale intensity image, after utilizing homomorphic luminance compensation method, is a round region, which is darker compared to its neighboring pixels. These regions can be highlighted by local Hessian based filtering method, which is defined by:

$$\mathbf{H}_\sigma = \alpha(\sigma) \cdot \nabla^2(\mathbf{I} * \mathbf{G}_\sigma)$$

where ∇ is the gradient operator and $*$ denotes the operator of the convolution. $\alpha(\sigma)$ is a normalization scalar compensating the multiplicative dependency of the norm of the matrix on the σ parameter. Based on the scale space theory [4] $\alpha(\sigma) = \sigma^2$. The value of σ depends on the size of the blob which is proposed as a pupil candidate, which is determined by the solution of the optimization problem:

$$\sigma(r) = \arg\max_\sigma \Big\{ (\mathbf{D}_r * (\alpha(\sigma) \cdot \mathbf{G}_\sigma))(0,0) \Big\} = r/\sqrt{2}$$

where r denotes the radius of the blob, $\mathbf{D}_r$ denotes the binary image of the $(0,0)$ centered homogeneous sphere with radius r:

$$\mathbf{D}_r(x,y) = \begin{cases} 1, & \text{if } \Big\|[x\ y]\Big\|^2 \leqslant r \\ 0, & \text{if } \Big\|[x\ y]\Big\|^2 > r \end{cases}$$

The local Hessian operator assigns a 2×2 matrix to each pixel of the examined image. Since the shape of the visible pupil is highly dependant on the direction of the gaze and the distance between the lower and the upper eyelids, its radius can not be precisely estimated by the pupil proposal algorithm, therefore a set R of possible r-s is defined. An ellipse is fitted to the limbus points using the RANSAC [2] method, the axes of which are used as radius proposals. Furthermore we found, that the eye landmark points are stable in the two corners of the eye. Based on these fixed ratios ($\frac{1}{2}, \frac{5}{12}, \frac{1}{3}$) of the half distance between the corners of the eye are added to the proposals.

The best fitting $r \in R$ is defined by its corresponding scale, in which the largest amplitude curvature of the surface defined by the intensities of the image is the minimal (which corresponds to our observation, that the pupil can be approximated by a dark blob):

$$r^* = \arg\max_r \Big\{ \lambda_{\max}\{\mathbf{H}_{\sigma(r)}\} \Big\}$$

where λ_{max} denotes the maximal amplitude eigenvalue of the Hessian matrix, and (x_0, y_0) denotes the proposed center of the pupil. Eigenvalues of the Hessian are calculated by:

$$\lambda_1\{\mathbf{H}\} = \frac{\operatorname{trace}\{\mathbf{H}\} + \sqrt{\operatorname{trace}\{\mathbf{H}\}^2 - 4\det\{\mathbf{H}\}}}{2}$$
$$\lambda_2\{\mathbf{H}\} = \operatorname{trace}\{\mathbf{H}\} - \lambda_1\{\mathbf{H}\}$$

The first equality can be derived based on the following observations:

$$\begin{aligned} \operatorname{trace}\{\mathbf{H}\} &= \operatorname{trace}\{\mathbf{Q}^T\mathbf{\Lambda}\mathbf{Q}\} = \operatorname{trace}\{\mathbf{\Lambda}\mathbf{Q}\mathbf{Q}^T\} \\ &= \operatorname{trace}\{\mathbf{\Lambda}\} = \lambda_1 + \lambda_2 \\ \det\{\mathbf{H}\} &= \det\{\mathbf{Q}^T\mathbf{\Lambda}\mathbf{Q}\} = \det\{\mathbf{Q}^T\} \cdot \det\{\mathbf{\Lambda}\} \cdot \det\{\mathbf{Q}\} \\ &= \det\{\mathbf{\Lambda}\} = \lambda_1 \cdot \lambda_2 \end{aligned}$$

where $\mathbf{H} = \mathbf{Q}^T\mathbf{\Lambda}\mathbf{Q}$ is the eigenvalue–eigenvector decomposition of the Hessian matrix, which always exists, because $\mathbf{H}$ is symmetric. Please note that $\mathbf{Q}$ is an orthonormal matrix, therefore $\mathbf{Q}^T\mathbf{Q} = \mathbf{I}$.

After we get the scale, the only remaining task is to examine the circularity of the examined blob, which can be measured by:

$$C_{(r)}(x_0, y_0) = \lambda_{\max}\Big\{\mathbf{H}_{\sigma(r)}(x_0, y_0)\Big\} \cdot \lambda_{\min}\Big\{\mathbf{H}_{\sigma(r)}(x_0, y_0)\Big\}$$

Since the Hessian matrix is symmetric, it can be diagonalized by an orthonormal matrix, which means that the product of the eigenvalues is equal to the determinant of the Hessian, which

can be computed faster than the value of the lower amplitude eigenvalue, therefore computed by:

$$C_{(r)}(x_0, y_0) = \det\left\{\mathbf{H}_{\sigma(r)}(x_0, y_0)\right\}$$

From there, the openness of the examined eye is calculated by:

$$\mathrm{Ind}\left(C_{(r)}(x_0, y_0) > c\right) \cdot \mathrm{Ind}\left(\lambda_{\max}\left\{\mathbf{H}_{\sigma(r)}(x_0, y_0)\right\} > 0\right)$$

Here, Ind denotes the indicator function and c is the "roundness/non-prolongation" set to 0.3. The first term is true if the portion of the eigenvalues is not too high, therefore the proposed pupil region is circular, and the second term is true, if the region is darker than its neighboring pixels. The eye is considered open if both terms are true.

III. Eyebrow-raising detection

The movement of the eyebrows is also a relevant feature in case of detecting facial expressions, as it can be used to discriminate between fear, anger and surprise. We detail our proposed eyebrow-raising detection algorithm in this section. These methods require video as input, since the change in eyebrow position is examined. Based on our empirical observations, the localization of the landmark points of the eyes and the corresponding eyebrows are accurate enough to detect eyebrow-raising based on the distances between the centroid of the eye mask and the middle of the landmark points of the eyebrow (19th and 26th points in Figure 1) respectively. The video is initially processed frame-by-frame to extract landmark points. The next step is to extract the distances of particular points.

To detect the eyebrow-raising, eyebrow-nose bottom and eyebrow-eye distances are examined. Another input for this method is the output of the blink detection.

Based solely on the eyebrow-eye distances (shown in Figure 2a) on each of the frames, the detection of corresponding microexpressions is not possible (the person in the image may squirm, lean forward, nod, etc.). Therefore, the proposed method examines the local trends of these distances by fitting a Gaussian distribution based on the previous frames. Using a 20 wide sliding window, mean and variance are calculated. In the next step, the relative likelihood is calculated for the next 3 points (one by one) using the probability density function of the normal distribution defined by the observed distances in the window. By multiplying the 3 conditional likelihoods, the joint probability is calculated assuming conditional independence. The null hypothesis is that if the records come from the same distribution, then they are independent given the window:

$$p(x_{win+3}, x_{win+2}, x_{win+1} \mid x_{win}) = \prod_{i=1}^{3} p(x_{win+i} \mid x_{win})$$

The joint probability is defined by:

$$\prod_{i=1}^{3} p(d_{win+i} \mid win) = \prod_{i=1}^{3} \frac{1}{\sigma_{win}\sqrt{2\pi}} \cdot e^{-\frac{(d_{win+i}-\mu_{win})^2}{2\sigma_{win}^2}},$$

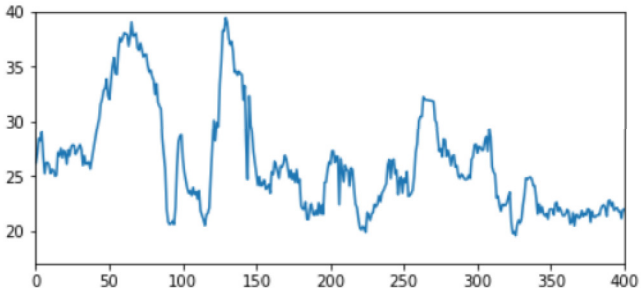
where d_{win+i} denotes the distance after the window with i frames; σ_{win} and μ_{win} denotes the standard deviation and the mean of the window respectively. The joint probability is calculated after each window in the time series. For ease of use, we work with the logarithm of the values. In the next, step eyebrow movements are detected. First, a binary threshold is applied. Using a maximum filter with a width of 5, movements close to each other are merged into one "sequence". In these sequences, local minimum search is performed to find the timestamp belonging to the most salient movement. These are points, which belong with low probability to the distribution defined by the window, thus are considered as outliers. Many of these candidates are false positives, which is detected because of the eye lowering and contraction. These proposals are eliminated by comparing the window mean and distance values belonging to that timestamp. Until this point, time series belonging to each eye are treated separately. However, while investigating the landmark points on videos, we found, that moving the eyebrow on only one side (left) affects the landmark points on the other side (right), thus "generating movement". This movement is much smaller on the unraised side (right). Thus pairing the movements of the two sides is necessary. Detections on the two sides with a distance in time less than $0.1\,\mathrm{s}$ are considered as the same movement. If the distance on one side belonging to such a movement is less than half of that on the other side, then it is a "generated movement", else it is a bilateral eyebrow-raising.

The majority of this process is also performed on the eyebrow-nose tip distance (see Figure 2b) time series. Joint probabilities are computed and are filtered using a harder threshold. Filtering for only raising movement is also applied.

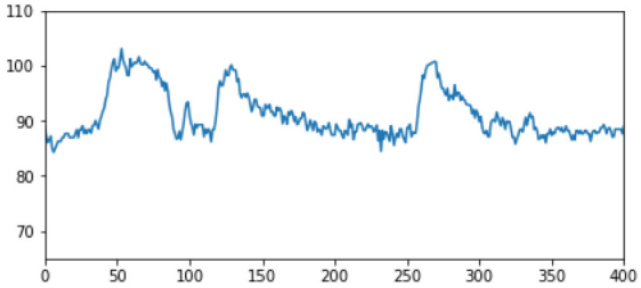
Blink timestamps resulting from the gaze detection part are used to mark possibly false positive detections in a window. If such a movement can also be found in the eyebrow-nose tip distance time series, then it is considered as a real movement, else as a false positive movement (see Figure 2d).

IV. Results

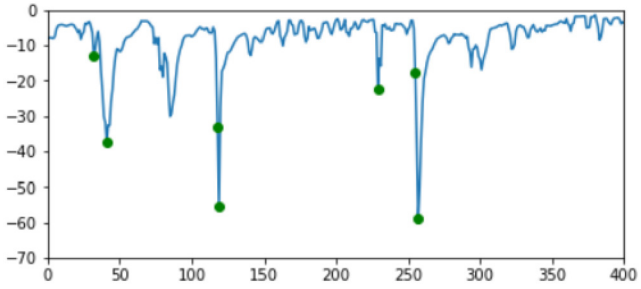
Figure 3 shows results of the eyebrow-raising detection method on adjacent video frames. Figure 3a shows an example where there is no blinking during the eyebrow raising. In case of blinking, the decision is made based on the eyebrow-nose distance. In Figure 3b a blink was detected during the eyebrow raising. However, in Figure 3c - based on the eyebrow-nose distance - it was established, that there was no eyebrow raising. The proposed algorithms for gaze and eyebrow raising detection may serve as components of an expert system aiming the detection of microexpressions. Along with other detected features, such as mouth shape and angle or wrinkles appearing on the face, these microexpressions can be utilized to identify basic emotions.



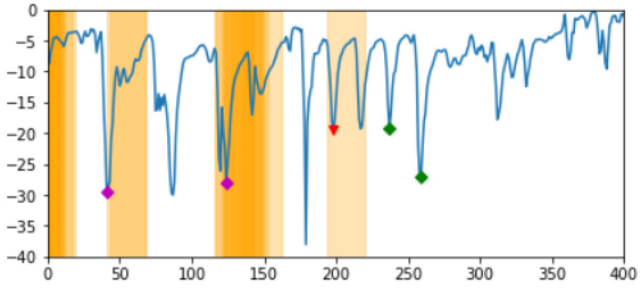
(a) Eyebrow-eye distance.



(b) Eyebrow-nose distance.



(c) Filtered detections calculated based on the eyebrow-nose distance.



(d) Final filtered detections calculated based on the eyebrow-eye distance. Orange zones are the blinking timestamps with a window around them. Green diamonds mark the detections out of the blink-windows. Magenta diamonds are the detections that fall into the orange zone but can be detected based on the eye-nose distance, thus are retained. Red triangles denote the detections that fall into the orange zone and can not be detected based on the eye-nose distance thus are permanently removed.

Fig. 2: Steps of eyebrow raising detection (right eye). Time frames are shown on the X-axis, whereas Y-axis shows pixel distances in Figues 2a and 2b and the logarithm of the probabilities in Figures 2c and 2d.

(a) Eyebrow-raising without blinking. First green diamond (at around 38) in Figure 2d.

(b) Eyebrow-raising during blinking. First purple diamond (at around 124) in Figure 2d.

(c) Blinking without eyebrow-raising. First red triangle (at around 199) in Figure 2d.

Fig. 3: Eyebrow-raising detections shown in Figure 2d.

References

[1] Paul Ekman, Joseph C. Hager, and Wallace V. Friesen. *Facial action coding system: the manual*. Research Nexus, 2002.

[2] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[3] Davis E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758, 2009.

[4] Tony Lindeberg. Scale-space theory in computer vision, 1993.

[5] Fabian Timm and Erhardt Barth. Accurate eye centre localisation by means of gradients. *Visapp*, 11:125–130, 2011.

[6] Roberto Valenti and Theo Gevers. Accurate eye center location and tracking using isophote curvature. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008.

[7] Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001*, volume 1, pages I–I. IEEE, 2001.

[8] Erroll Wood and Andreas Bulling. Eyetab: Model-based gaze estimation on unmodified tablet computers. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, pages 207–210, 2014.