Post-print

Authors:	Gizatdinova Yulia, Surakka Veikko
Name of article:	Automatic Detection of Facial Landmarks from AU-Coded Expressive Facial Images
Name of work:	Proceedings of the 14th International Conference on Image Analysis and Processing (ICIAP'07)
Year of publication:	2007
Publisher:	IEEE Computer Society
Pages:	419-424
Discipline:	Natural sciences / Computer and information sciences
Language:	en

URN: http://urn.fi/urn:nbn:uta-3-986
DOI: http://dx.doi.org/10.1109/ICIAP.2007.4362814

Automatic Detection of Facial Landmarks from AU-coded Expressive Facial Images

Yulia Gizatdinova and Veikko Surakka
Research Group for Emotions, Sociality, and Computing
Tampere Unit for Computer-Human Interaction
Department of Computer Sciences
University of Tampere
{yulia.gizatdinova, veikko.surakka}@cs.uta.fi

Full citation:
Gizatdinova Y., Surakka V. (2007). Automatic detection of facial landmarks from AU-coded expressive facial images. 14th International Conference on Image Analysis and Processing (ICIAP'07), Modena, Italy, September 10-14, pp. 419-424.

DOI: 10.1109/ICIAP.2007.4362814

Abstract

The present aim was to develop a fully automatic feature-based method for expression-invariant detection of facial landmarks from still facial images. It is a continuation of our earlier work where we found that some certain muscle contractions made a deteriorating effect on the feature-based landmark detection especially in the lower face. Taking into account this crucial facial behavior, we introduced improvements to the method that allowed facial landmarks to be fully automatically detected from expressive images of high complexity. In the method, information on local oriented edges was utilized to compose edge maps of the image at two levels of resolution. The landmark candidates resulted from this step were further verified by edge orientation matching. We used knowledge on face geometry to find the proper spatial arrangement of the candidates. The results obtained demonstrated a high overall performance of the method while testing a wide range of facial displays.

Keywords: Computing Methodologies, Image Processing and Computer Vision, Segmentation, Edge and feature detection, Facial expressions.

1. Introduction

Human faces constitute a class of objects with rigid structure that does not vary significantly from person to person (i.e. nose is located between eyes and mouth). However, the problem of automatic detection of face and facial features has been challenging computer scientists already for several decades, and still needs further investigation. The difficulty comes from the fact that facial appearance varies noticeably with changes in environmental conditions (e.g. illumination, head pose, orientation, and occlusions), race, gender and facial expressions (e.g. emotional and social signals in the face). To solve the problem, a representation of the face is needed that remains robust with respect to variety of facial appearances. Following this idea, many techniques to face and facial feature detection have been proposed [1], [2].

In expressive facial behavior, muscle contractions produce skin displacements that change drastically the appearance of permanent (e.g. eyes, eyebrows, nose,

and mouth) and transient (e.g. wrinkles resulting from expressive and edge-specific face modifications) facial features. Facial expressions result in considerable changes of feature shapes and their locations on the face, presence/absence of teeth, out-of-plan changes (e.g. showing the tongue), and self-occlusions.

In the domain of behavioral science research, the Facial Action Coding System (FACS) [3], [4] is a well known linguistic description of all visibly detectable changes in the facial appearance. The FACS describes visible changes in the face as a result of single and joint muscle contractions in terms of action units (AUs). In other words, FACS represents an expressive image as a result of facial muscle activity without referring to emotional state of a person on the image.

Addressing the problem of expression-invariant facial landmark detection, Gizatdinova and Surakka [5] introduced a feature-based method that made use of local oriented edges extracted in still facial image. For this purpose, a set of multiorientation and multiresolution Gaussian filters was utilized. The detailed description of edge detection and grouping used can be found in Appendix A. Resulting from these stages, the final edge map of the image consisted of regions of connected edges presuming to contain facial landmarks. The existence of a landmark on the image was verified by matching candidates against the orientation model (for more details, see Appendix B).

The method was not fully automatic and required a manual classification of the detected edge regions. Besides that, the method was deteriorated by facial expressions, especially by those appeared in lower face [6]. The further analysis [7] revealed specific facial behaviors that influenced the performance of the method the most. It was found that incorrect nose and mouth detection was caused mainly by AUs activated during disgust (AU 9 and 10), happiness (AU 12), and some of their combinations with other AUs. Although the listed AUs have different effect on facial appearance, they commonly make the gap between nose and mouth smaller. The neighborhood distances between edges belonging to these landmarks became smaller than a threshold and caused erroneous grouping of nose and mouth into one region. In some cases, AUs 1 and 4 activated during sadness and anger caused eyes or eyebrows to draw up together resulting in incorrect upper face landmark detection.

In the present study, we extended the previous research. Taking into account the described facial behaviors interfering landmark detection, we improved the overall performance of the method. The method now allowed facial landmarks to be fully automatically detected from expressive images of high complexity.

2. Facial landmark detection

The method was improved in several respects. Instead of using an average contrast of the whole image to define thresholds for contrast filtering, we applied local contrast thresholding calculated in every filter neighborhood. This allowed more reliable edge detection. Further, edge grouping was improved as the method failed at this stage due to erroneous connection of edges belonging to different facial landmarks into one region. The top row of Figure 1 shows bounding box that includes merged eye regions on the left and merged nose and mouth on the right. To fix this problem, we applied the procedure of edge projection as follows. If a landmark candidate consisted of two or more regions of edge concentration, edge points were projected to x-axis for upper face landmarks and to y-axis for lower face landmarks. The projections were obtained from calculating the number of edge points along the corresponding (i.e. vertical or horizontal) rows of the final edge map for the given candidate. If the number of edge points was smaller than a threshold, edge points were eliminated (Figure 1). After each edge elimination step, if the region still was not separated the threshold was increased by 5 edge points. The initial threshold equaled a minimum number of edges in the column (row) of the given candidate.

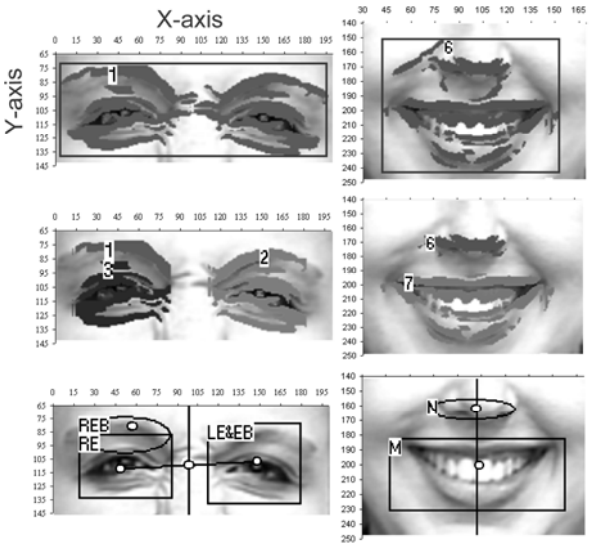

Figure 1. Landmarks grouped into one region (top), landmarks separated by edge projection (middle), and final detection result (bottom). Images are courtesy of the Cohn-Kanade AU-Coded Facial Expression database [8]. Reprinted with permission.

After the procedure of edge projection, the orientation portraits (i.e. the distribution of local oriented edges) of the received edge regions were matched against the orientation model. In this study, we allowed landmark candidates to have some deviations from the orientation model. It means that an orientation portrait of the candidate could slightly differ from the model, for example, it could have some orientations represented by zero number of edges. In further analysis, these edge regions were also considered in composing face-like constellations of the detected landmark candidates if there were missing landmarks. Figure 2,a shows the final edge map of the image with landmark candidates and discarded edge regions.

The final improvement of the method was the automatic classification of the detected landmark candidates. We formed constellations from a set of detected candidates and determined which constellations were the most face-like. The face model we used is shown in Figure 2,b. Due to side-by-side location of upper face landmarks, they guided the entire process of landmark classification, also those landmarks which were discarded by the orientation model. The search started with finding horizontal candidate pairs with approximately equal number of edge points and labeling them as eyes and eyebrows. If only one horizontal pair was found, it was labeled as eye region candidate (i.e. eye and eyebrow were detected as one region). The method then searched for

eyebrows above and eyes below the found pair location and if found any, relabeled the found candidates as eye and eyebrow, respectively. If there was not any pair found, it was assumed that eye regions were grouped together in one region and edge x-projection was applied to the candidate with maximum number of edges. The search for lower face landmarks was performed from top-to-bottom along the line of vertical symmetry that was drawn through the point that lied in the middle of the line connecting eye regions. If only one lower face candidate was found, the method assumed that nose and mouth were combined together and edge y-projection was applied to separate these landmarks. Although the method was allowed to miss landmarks, however, for efficient landmark detection at least one horizontal pair had to be found. As a measure of distances in the face model we utilized the dynamic parameter *D* calculated as a distance between mass centers of the eye region pair. Using this measure, the spatial constraints between locations of the rest of the candidates were verified. For example, nose is located between eyes and mouth not lower than one *D* from the middle point of the line connecting eye regions. At the same time, by utilizing geometrical relationships among the candidates, we verified the upper face landmarks. After the face-like constellation of landmarks was found, the location of the face in the image was also known.

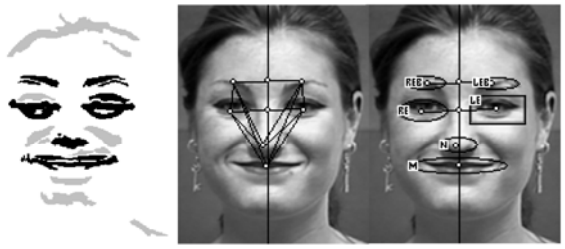

Figure 2. (a) Final edge map with landmark candidates (black) and discarded edge regions (grey), (b) face geometry model, and (c) final detection result. Image is courtesy of Cohn- Kanade AU-Coded Facial Expression database [8]. Reprinted with permission.

3. Database

The Cohn-Kanade AU-Coded Facial Expression Database [8] consists of image sequences taken from 97 subjects (65% female) of different skin color (81% Caucasian, 13% African-American, and 6% Asian or Latino) and ages varying from 18 to 30 years. There were no images with facial hair or eye-glasses. Each image sequence starts with neutral frame and ends up with an expressive frame labeled in terms of AUs. AUs occur alone or in combinations and are coded as numbers. The level of expression intensity can vary for images of different subjects and is coded as small letters. Capital letters L and R define left- and right-side expressions.

From the database we selected 468 neutral and 468 expressive images corresponding to the first and the last frames of the sequence. From this data we composed two datasets – "face only" dataset of cropped images including only facial region, and "face & hair" dataset of cropped images including both face and hair. "Face only" dataset served as a "baseline" to which we compared the robustness of the method with respect to such destructors as hair, decoration, and elements of clothe. All images were preset to the size of approximately 200-250 pixel arrays with 8-bit precision for grey scale values. No face alignment was performed.

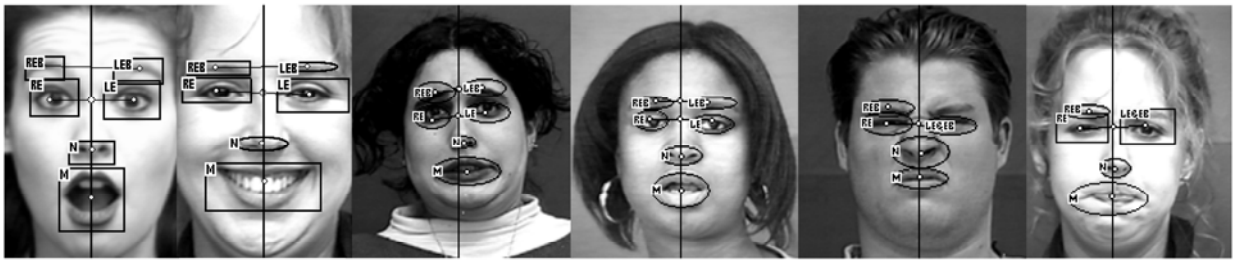

Figure 3. Examples of correctly localized facial landmarks in "face only" and "face & hair" datasets. Images are courtesy of the Cohn-Kanade AU-Coded Facial Expression database [8]. Reprinted with permission.

4. Results

The following facial landmarks were chosen to be detected: - right eye (RE), right eyebrow (REB), left eye (LE), left eyebrow (LEB), right eye and eyebrow (RE&EB), left eye and eyebrow (LE&EB), lower nose (N), and mouth (M). Figure 3 shows the final results of the landmark detection in both datasets. As figure shows, the size of the bounding box that contained a landmark was dynamic and varied according to the size of the detected edge region. The landmarks with orientation portraits slightly different from the orientation model were represented as ovals, and landmarks corresponded to the orientation model – as rectangles.

The final results were classified into one of the following classes: correct detection, wrong detection, and false detection. A correct detection was considered if the bounding box overlapped approximately at least 50% of the visible landmark, and edge region enclosed the area surrounding landmark less than the actual size of the landmark. In detecting eye regions, eyebrow together with a corresponding eye were localized as one region, or alternatively, eye and eyebrow were localized separately. If eyebrow was detected as a separate region, it was obligatory that a corresponding eye was also found. A wrong detection was considered if the bounding box covered several facial landmarks, excluding the case of eyes and eyebrows localized as one region. A false localization was considered when bounding box did not satisfy any of the two previous conditions. We defined the rate of the landmark detection as a ratio between a total number of landmarks correctly localized and the total number of images used in testing (as there was one face per image). A false positive was then defined as a number of noise regions (wrinkles, eyebrow localized without a corresponding eye, elements of face, ears, clothing and hair) which were misclassified as a facial landmark.

As it is seen from Table 1 and Table 2, there was no significant difference in the performance of the method on two datasets. Further, the facial landmarks were detected with nearly equal detection rates in both neutral and expressive images. Thus, the method achieved the average detection rates of 97.5% and 94% for neutral and expressive "face only" images, correspondently. The rates for "face & hair" dataset were 91.5% for neutral images and 90% for expressive images. A decrease in detection rates for lower face landmarks was observed; on the whole, however, the overall performance of the method was high.

We noticed that detection of lower face landmarks produced more errors than detection of upper face landmarks. For example, in some cases the method misclassified a chin as a mouth, (in the tables, the biggest number of false positives corresponding to mouth detection reflects this fact). Wrong detections

Table 1. Landmark detection rates (%) and false positives (FP) for "face only" dataset

Image	Right eye region	Left eye region	Nose	Mouth
Neutral	98 1 FP	99 1 FP	98 0 FP	95 12 FP
Expressive	97 6 FP	98 2 FP	92 1 FP	90 12 FP

Table 2. Landmark detection rates (%) and false positives (FP) for "face & hair" dataset

Image	Right eye region	Left eye region	Nose	Mouth
Neutral	95 2 FP	96 3 FP	89 1 FP	86 19 FP
Expressive	93 5 FP	94 6 FP	90 2 FP	81 25 FP

were observed mostly in detecting lower face landmarks. Thus, nose and mouth were detected as one region in 16 expressive images of "face only" dataset and in 18 expressive images of "face & hair" dataset. As it was expected, it occurred mainly due to the effect of lower face AUs 9, 10, and 12 occurring alone or in combinations with other AUs.

The eye region detection was high for all types of images showing expressions in upper and lower face, see Table 3 and Table 4, (note that AUs presented might occur singly or in conjunction with other AUs which are not represented in the tables). On the whole, the detection of lower face landmarks was more affected by AUs than the detection of eye regions. Lower face AUs 9, 10, 12, and AU combinations 9+25, 10+17, 10+20, 10+25, 12+16, 12+25, and 16+25 lowered down the nose and mouth detection average rates up to the range of 71-83%. Upper face AUs 5, 6, 7, and AU combination 6+7 also degraded the lower face landmark detection. These upper face AUs are usually activated during the lower face expression of anger when AUs 9 and 10 typically are also activated.

Table 3. Rates (%) of landmark detection in images showing upper face single AUs

AUs	Right eye region	Left eye region	Nose	Mouth
1	92	96	90	83
2	90	96	94	88
4	94	95	91	86
5	89	93	91	73
6	95	93	86	77
7	93	96	86	73
43/45	95	98	88	88

Table 4. Rates (%) of landmark detection in images showing lower face single AUs

AUs	Right eye region	Left eye region	Nose	Mouth
9	93	94	87	81
10	94	89	82	78
11	98	98	87	87
12	96	94	79	71
14	84	100	89	89
15	96	96	95	91
16	97	100	94	90
17	94	95	92	91
20	97	94	94	89
23	95	93	93	92
24	94	93	94	94
25	95	96	91	92
26	94	98	97	94
27	92	97	96	88

5. Discussion

A fully automatic method was designed for facial landmark detection in the expressive images of high complexity. The complexity of the expression was presented by closed/semi-closed eyes, variety of mouth appearances including open and tight mouth, visible teeth and tongue. The local oriented edges served as basic features for expression-invariant representation of facial landmarks. The results confirmed that in the majority of expressive images the landmark orientation portraits had the same structure as predefined by the landmark orientation model. The face geometry model further improved the overall performance of the method. Besides robustness to facial expressions, the method demonstrated robustness to skin color and noise like hair, ear-rings, and elements of clothe.

Comparing present results with previous ones, a significant improvement was achieved for detection of lower face landmarks, especially, in images showing AUs 9, 10, and 12. The landmark detection rates were comparable or superior to those presented in [9]-[11] while testing a wider range of facial displays.

Emphasizing simplicity of the method developed, we conclude that it can be used in preliminary localization of regions of facial landmarks for their subsequent processing where coarse landmark localization is followed by fine feature detection (e.g. local features like eye and mouth corners).

6. Acknowledgement

This work was financially supported by the Finnish Academy (project number 177857), the University of Tampere, and the Tampere Graduate School in Information Science and Engineering. The authors thank the creators of the Cohn-Kanade AU-Coded Facial Expression Database for the permission to reprint examples of the expressive images.

7. References

[1] E. Hjelmas, B. Low, "Face Detection: A survey", *Computer Vision and Image Understanding*, 83, 2001, pp. 235–274.

[2] M. Yang, D. Kriegman, and N. Ahuaja, "Detecting Face in Images: A Survey", *IEEE Transactions on Pattern Analysis and Image Understanding*, 24, 2002, pp. 34-58.

[3] Ekman, P., W. Friesen, *Facial Action Coding System (FACS): A Technique for the Measurement of Facial Action*, Consulting Psychologists Press, Inc., Palo Alto, California, 1978.

[4] Ekman, P., W. Friesen, and J. Hager, *Facial Action Coding System (FACS)*, A Human Face, Salt Lake City, Utah, 2002.

[5] Y. Gizatdinova, V. Surakka, "Feature-Based Detection of Facial Landmarks from Neutral and Expressive Facial Images", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28, 1, 2006, pp. 135-139.

[6] I. Guizatdinova, V. Surakka, "Detection of Facial Landmarks from Neutral, Happy, and Disgust Facial Images", *Proceedings of the 13th International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision (WSCG)*, Plzen, Czech Republic, 2005, pp. 55-62.

[7] Y. Gizatdinova, V. Surakka, "Edge Orientation Matching for Facial Landmark Localization in Images Showing Expressions in Upper and Lower face", submitted to *Computer Vision and Image Understanding.*

[8] T. Kanade, J. Cohn, and Y. Tian, "Comprehensive Database for Facial Expression Analysis", *Proceedings of the 4th IEEE International Conference on Automatic Face and Gesture Recognition*, Grenoble, France, 2000, pp. 46-53.

[9] P. Campadelli, R. Lanzarotti, G. Lipori, and E. Salvi, "Face and Facial Feature Localization", *Proceedings of the 13th International Conference on Image Analysis and Processing (ICIAP)*, Gagliari, Italy, 2005, 3617, pp. 1002-1009.

[10] D. Shaposhnikov, A. Golovan, L. Podladchikova, N. Shevtsova, X. Gao, V. Gusakova, and Y. Gizatdinova, "Application of the Behavioral Model of Vision for Invariant Recognition of Facial and Traffic Sign Images" (Применение поведенческой модели зрения для инвариантного распознавания лиц и дорожных знаков), *Journal of Neurocomputers: Design and Application*, 7(8), 2002, pp. 21-33.

[11] D. Cristinacce, T. Cootes, "Facial Feature Detection Using AdaBoost with Shape Constraints", *Proceedings of the 14th British Machine Vision Conference (BMVC)*, Norwich, England, 2003, pp. 231-240.

8. Appendix A: Edge detection

The grey scale image representation was considered as a two dimensional array $I = \{b_{ij}\}$ of the $X \times Y$ size. Each b_{ij} element of the array represented *b* brightness of the $\{i, j\}$ image pixel. If there was a color image, it was first transformed into the grey scale representation by averaging three RGB components. This allowed the method to be robust with respect to small illumination variations and skin color. To smooth a grey level image the recursive Gaussian transformation was used.

$$b_{ij}^{(l)} = \sum_{p,q} a_{pq} b_{ij}^{l-1}, \; b_{ij}^{(1)} = b_{ij}, \quad (1)$$

where a_{pq} was a coefficient of the Gaussian convolution; p and q defined the size of a filter, $p, q = -2 \div 2$; $i = 0 \div X - 1$; $j = 0 \div Y - 1$; l defined the level of image resolution. The smoothed low resolution image (*l=2*) was used to find all possible landmark candidates, and the original high resolution image (*l=1*) was used to analyse landmark candidates in detail.

Then the smoothed image was convolved with a set of ten-orientation Gaussian filters with shifted centres.

$$G_{\varphi_k}^{-} = \frac{1}{2\pi\sigma^2} e^{-\frac{(p-\sigma\cos\varphi_k)^2+(q-\sigma\sin\varphi_k)^2}{2\sigma^2}}, \quad (2)$$

$$G_{\varphi_k}^{+} = \frac{1}{2\pi\sigma^2} e^{-\frac{(p+\sigma\cos\varphi_k)^2+(q+\sigma\sin\varphi_k)^2}{2\sigma^2}}, \quad (3)$$

$$G_{\varphi_k} = \frac{1}{Z}(G_{\varphi_k}^{-} - G_{\varphi_k}^{+}), \quad (4)$$

$$Z = \sum (G_{\varphi_k}^{-} - G_{\varphi_k}^{+}), \; G_{\varphi_k}^{-} - G_{\varphi_k}^{+} > 0, \quad (5)$$

where σ was a root mean square deviation of the Gaussian distribution; φ_k was an angle of the Gaussian rotation, $\varphi_k = k \cdot 22.5°$; $k = 2 \div 6, 10 \div 14$; $p, q = -3 \div 3$.

The maximum response of all 10 kernels defined the contrast magnitude of a local edge at its pixel location. The orientation of a local edge was estimated with orientation of a kernel that gave the maximum response.

$$g_{ij\varphi_k} = \sum_{p,q} b_{i-p,\,j-q}^{(l)} G_{\varphi_k}, \quad (6)$$

The threshold for contrast filtering of the extracted edges was determined as an average contrast of the whole smoothed image. Edge grouping was based on neighborhood distances between edge points and limited by a radius consisting of possible neighbors for each edge point. Regions with small number of edge points were removed. The optimal thresholds for edge grouping were determined using small image set taken from the database. To get more detailed description of the extracted edge regions, edge detection and grouping were applied to high resolution image within the limits of these regions. In this case, the threshold for contrast filtering was determined as a double average contrast of the high resolution image.

9. Appendix B: Edge orientation matching

The procedure of orientation matching was applied to verify the existence of facial landmarks on the image. To do that, the detected regions were matched against the orientation model that was a specific distribution of the local oriented edges with two horizontal dominants (for example, see Figure 4). The following rules define the distribution of the orientation model: 1) horizontal orientations are represented by the greatest number of the extracted edges; 2) a number of edges corresponding to each of horizontal orientations is more than 50% greater than a number of edges corresponding to any other orientations; and 3) orientations cannot be represented by zero number of edges. Noise regions like, for example, elements of cloth and hair usually have an arbitrary distribution of the oriented edges and were discarded by the model.

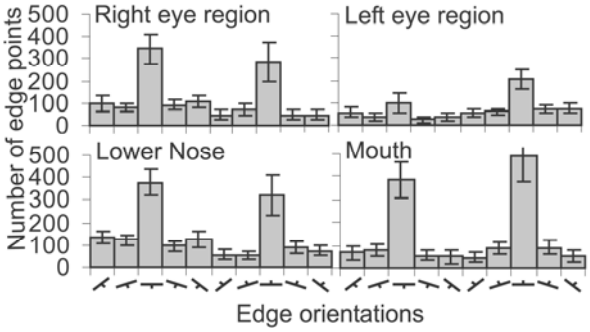

Figure 4. Examples of landmark orientation portraits averaged over "face only" datasets. The error bars show plus/minus one standard deviation from the mean values.