Yulia Gizatdinova and Veikko Surakka

Automatic localization of facial landmarks from expressive images of high complexity

DEPARTMENT OF COMPUTER SCIENCES
UNIVERSITY OF TAMPERE

D-2008-9

TAMPERE 2008

UNIVERSITY OF TAMPERE
DEPARTMENT OF COMPUTER SCIENCES
SERIES OF PUBLICATIONS D – NET PUBLICATIONS
D-2008-9, SEPTEMBER 2008

Yulia Gizatdinova and Veikko Surakka

Automatic localization of facial landmarks from expressive images of high complexity

DEPARTMENT OF COMPUTER SCIENCES
FIN-33014 UNIVERSITY OF TAMPERE

ISBN 978-951-44-7498-9
ISSN 1795-4274

Automatic Localization of Facial Landmarks from Expressive Images of High Complexity

Yulia Gizatdinova and Veikko Surakka

Abstract

The aim of this study was to develop a fully automatic feature-based method for expression-invariant localization of facial landmarks from static facial images. It was a continuation of our earlier work in which we found that lower face expressions deteriorated the feature-based localization of facial landmarks the most. Taking into account the found crucial facial behaviours, the method was improved so that it allowed facial landmarks to be fully automatically located from expressive images of high complexity. Information on local oriented edges was utilized to compose edge maps of the image at several levels of resolution. Landmark candidates which resulted from this step were further verified by matching them against the orientation model. The present novelty came from two last steps of the method which were edge projection that was used to enhance a search for landmark candidates, and structural correction of the found candidates based on a face geometry model. The method was tested on three facial expression databases which represented a wide range of facial appearances in terms of prototypical facial displays and individual facial muscle contractions. To evaluate the localization results, a new performance evaluation measure was introduced that represented a localization result as a rectangular box instead of a conventional point representation. The evaluation of the method by the proposed rectangular and conventional point evaluation measures demonstrated a high overall performance of the method in localization of facial landmarks from images showing complex expressions in upper and lower face.

Keywords - Image processing, computer vision, edge detection, landmark localization, action unit (AU), facial expression.

I. INTRODUCTION

Facial expressions are emotional, social, and otherwise meaningful cognitive and physiological signals in the face. Facial expressions result from contractions and relaxations of different facial muscles. These non-rigid facial movements result in considerable changes of facial landmark shapes and their location on the face, visibility of teeth, out-of-plan changes (showing the tongue), and self-occlusions (close eyes and bitted lips). In the domain of behavioural science research, facial expressions can be viewed according to two main approaches. The emotion-based approach proposed by Ekman [1] considers facial expressions as primary channels of conveying emotional information. Ekman defined six prototypical emotions and six corresponding prototypical facial expressions which are happiness, sadness, fear, anger, disgust, and surprise. In the contrast to emotion-based approach, the Facial Action Coding System (FACS) [2,3] proposes a descriptive approach to facial expression analysis. The FACS is an anatomically based linguistic description of all visibly detectable changes in the face. The FACS describes visible changes in the face as a result of single and conjoint muscle activations in terms of action units (AUs). In other words, FACS decomposes an expressive face into AUs and represents an expression as a result of facial muscle activity fully objectively, without referring to emotional, social, or cognitive state of the person in the image. However, some specific combinations of AUs can represent prototypical facial expressions of emotional.

It has been shown [1] that structural changes in the regions of prominent facial landmarks like, for example, eyebrows, eyes, nose, and mouth are important and in many cases sufficient for facial expression classification. In the automatic facial expression analysis, a manual preprocessing is typically needed to select a set of characteristic points as, for example, eye centres and mouth corners, in static images or initial frame of the video sequence. These characteristic points are further used to track changes in the face or to align an input image with a face model. Currently, there is a need for a system that can automatically detect facial landmarks in the image prior to the following steps of the automatic facial expression analysis.

The problem of automatic facial landmark detection has been generally addressed by modelling local texture information around landmarks and modelling shape information on spatial arrangement of the detected landmark candidates [4,5,6]. In practice, this process consists of selecting a feature representation of facial landmarks and designing a feature detector. Different features can be detected from the image, for example, edges, colours, points, lines, and contours. These features provide a meaningful and measurable description of the face as they represent specific visual patterns which can be used to identify corresponding structures between images. The main challenge is to find feature representations of facial landmarks which uniquely characterize a face and remain robust with respect to changes in facial appearance due to changes in the environmental conditions (illumination, pose, orientation, etc.), gender, race, and facial expressions. Facial landmark localization is a subtask of a more general detection problem and refers to finding true locations of facial landmarks, given that a face is shown in the image.

To detect facial landmarks, Burl, Leung and Perona [7] modelled a texture around the landmarks by employing a set of multi-scale and multi-orientation Gaussian derivative filters. The most face-like constellation of the found candidate locations was further captured by a statistical shape model. In Wiskott et al. [8], the locations of characteristic points around facial landmarks were first found using Gabor jet detectors. Further, the distribution of facial points was modelled by a graph structure. Feris et al. [9] proposed a two-level hierarchical landmark search using Gabor wavelet networks. In this method, the first level network represented the entire face and determined affine transformation used for a rough approximation of the landmark locations. The second level networks represented separate landmarks and were used to verify the precise landmark locations. Similarly, Cristinacce and Cootes [10] extended a well known face detector introduced by Viola and Jones [11] for the task of detecting individual facial landmarks. The local boosted classifiers were used to detect facial landmarks and statistical models of the landmark configurations were utilized to select the most suitable candidates.

Addressing the problem of facial landmark localization, Gizatdinova and Surakka [12] introduced a feature-based method in which the information on local oriented edges was utilized to compose edge maps of the image at several levels of resolution. Landmark candidates which resulted from this step were further verified by matching them against the orientation model. The method was not fully automatic and required a manual classification of the located edge regions. Besides that, the landmark localization was significantly deteriorated by the lower face expressions [13]. The further analysis [14] revealed specific facial behaviours which influenced the performance of the method the most. It was found that incorrect nose and mouth localization was caused mainly by the lower face AU 12 (lip corner puller) activated during happiness, AU 9 (nose wrinkler) and AU 10 (upper lip raiser) both activated during disgust, and AU 11 (nasolabial furrow deepener) that is usually activated in conjunction with all mentioned AUs.

Taking into account the described facial behaviours which deteriorated the landmark localization, we made a number of improvements to the method which allowed facial landmarks to be fully automatically located from expressive images of high complexity. The preliminary results [15] of the method testing on the AU-coded facial expression database showed a significant improvement of the method performance for images showing lower face AUs 9, 10, 11, and 12. In the present study, we continued improving the method and presented new results of the method testing on a wider range of facial displays classified in terms of prototypical facial expressions, single AUs, and AU combinations. A new performance evaluation measure was introduced in order to evaluate the results of landmark localization. The proposed evaluation measure represented a localization result as a rectangular box instead of a conventional point representation.

II. DATABASES

The first database we used was the Cohn-Kanade AU-Coded Facial Expression (Cohn-Kanade) database [16] - one of the most comprehensive collections of expressive images available. The database consisted of image sequences taken from 97 subjects (65% female) of different skin colour (81% Caucasian, 13% African-American, and 6% Asian or Latino) and ages varying from 18 to 30 years. Each image sequence started with a neutral frame and ended up with an expressive frame labelled in terms of AUs. AUs occurred alone or in combinations and were coded as numbers. The AU descriptors taken from the FACS manual [2,3] were as follows. Upper face AUs: 1-inner brow raiser, 2-outer brow raiser, 4-brow lowerer, 5-upper lid raiser, 6-cheek raiser and lid compressor, 7-lid tightener, 43-eye closure, 45–blink, and 46-wink. Lower face AUs: 9-nose wrinkler, 10-upper lip raiser, 11-nasolabial furrow deepener, 12-lip corner puller, 13–sharp lip puller, 14-dimpler, 15–lip corner depressor, 16–lower lip depressor, 17-chin raiser, 18-lip pucker, 20–lip stretcher, 22–lip funneler, 23–lip tightener, 24–lip presser, 25–lips part, 26–jaw drop, 27–mouth stretch, and 28–lips suck. Capital letters R and L in front of the numerical code indicated right and left face AUs. Small letters after the numerical code represented an intensity level of the expression. For each subject, we selected one neutral face and several expressive faces of the highest intensity which corresponded to the latest frames in each expressive sequence. In sum, a total of 97 neutral and 486 expressive images were selected. From this original data we composed datasets of cropped images with face, hair, and sometimes shoulders included (the background and the image indexes were cut out).

The images of the rest of the two databases were labelled in terms of neutral and six prototypical facial expressions which were happiness, sadness, fear, anger, disgust, and surprise. The Pictures of Facial Affect (POFA) database [17] consisted of 14 neutral and 96 expressive images of 14 Caucasian individuals (57% female). On average, there were 16 images per facial expression. The Japanese Female Facial Expression (JAFFE) database [18] consisted of 30 neutral and 176 expressive images of 10 Japanese females. There were about 30 images per facial expression in average. In POFA and JAFFE databases, a particular expression could vary in its intensity or facial configuration.

All images were preset to approximately 200 by 300 pixel arrays. No face alignment was performed. The potential impact of illumination change, facial hair or eye-glasses was controlled to some extent in all the databases and therefore ignored. The received datasets were used to examine the robustness of the method with respect to facial activity labelled in terms of single AUs, AU combinations, and prototypical facial expressions. The robustness of the method to such destructors as hair, decoration, and elements of clothing was also studied.

III. FACIAL LANDMARK LOCALIZATION

The feature-based method of facial landmark localization consisted of several stages which are illustrated in Figure 1. The image 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. On the preprocessing stage, the image was smoothed by the recursive Gaussian transformation (Equation 1) to remove noise and small details from the image. On the following stages of the method, the smoothed low resolution image was used to find all possible landmark candidates, and the original high resolution image was used to analyze the landmark candidates in detail.

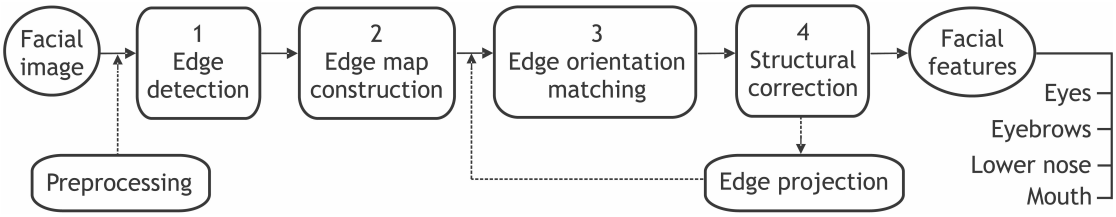


Figure 1. Block structure diagram of the facial landmark localization.

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

where a_{pq} is a coefficient of the Gaussian convolution; p and q define a size of the smoothing filter, $p, q = -2 \div 2$; $i = 0 \div X - 1$; $j = 0 \div Y - 1$; l define a level of resolution ($l = 2$).

If there was a colour image, it was first transformed into the grey scale representation by averaging three RGB components (Equation 2). This allowed the method to be robust with respect to small illumination variations and different skin colour.

$$b_{ij} = 0.299 \cdot R_{ij} + 0.587 \cdot G_{ij} + 0.114 \cdot B_{ij}\,. \qquad (2)$$

On the stage of edge detection, the smoothed low resolution image was filtered with a set of ten-orientation Gaussian filters (Equations 3-6) to extract local oriented edges.

$$G_{\varphi_k} = \frac{1}{Z}(G_{\varphi_k}^{-} - G_{\varphi_k}^{+})\,, \qquad (3)$$

$$G_{\varphi_k}^{-} = \frac{1}{2\pi\sigma^2}\exp\left(-\frac{(p-\sigma\cos\varphi_k)^2 + (q-\sigma\sin\varphi_k)^2}{2\sigma^2}\right), \qquad (4)$$

$$G_{\varphi_k}^{+} = \frac{1}{2\pi\sigma^2}\exp\left(-\frac{(p+\sigma\cos\varphi_k)^2 + (q+\sigma\sin\varphi_k)^2}{2\sigma^2}\right), \qquad (5)$$

$$Z = \sum(G^{-}_{\varphi_k} - G^{+}_{\varphi_k}),\ G^{-}_{\varphi_k} - G^{+}_{\varphi_k} > 0. \qquad (6)$$

where σ is a root mean square deviation of the Gaussian distribution; φ_k is an angle of the Gaussian rotation, $\varphi_k = k \cdot 22.5°$; $k = 2 \div 6,\ 10 \div 14$; $p,q = -3 \div 3$; $i = 0 \div X - 1$; $j = 0 \div Y - 1$.

The maximum response of all ten kernels (Equation 7) defined a contrast magnitude of the local edge at its pixel location. The orientation of the local edge was estimated by the orientation of the kernel that gave the maximum response.

$$g_{ij\varphi_k} = \sum_{p,q} b^{(l)}_{i-p,\,j-q} G_{\varphi_k}\ . \qquad (7)$$

On the stage of edge map construction, the extracted edge points were thresholded according to their contrast. The average contrast of the whole smoothed low resolution image was used to define a threshold for contrast filtering. Edge grouping was based on the neighbourhood distance (D_n) between edge points and limited by a minimum number of edge points in the region ($N_{\min}$). Thus, edge points were grouped into one region if the distance between them was less than D_n pixels and number of edge points inside the region was bigger than $N_{\min}$. Regions with small number of edge points were removed. This way, the final edge map of the image consisted of regions of connected edge points presuming to contain facial landmarks. The optimal thresholds for edge grouping were determined experimentally and summarized in Table I. To get more detailed description of the extracted edge regions, edge detection and edge grouping were applied to high resolution image ($l = 1$) within the limits of the found edge regions. In this case, the threshold for contrast filtering was determined as a double average contrast of the high resolution image.

TABLE I: SUMMARY OF THE THRESHOLDS FOR EDGE MAP CONSTRUCTION

Datasets	Neighborhood distance for edge grouping, D_n	Minimum number of edge points in the region, $N_{\min}$
Cohn-Kanade	1 pixel	100 pixels
POFA	3 pixels	150 pixels
JAFFE	2 pixels	160 pixels

On the stage of edge orientation matching, the existence of facial landmarks in the image was verified. To do that, a distribution of the local oriented edges inside the located regions, so called orientation portraits, was matched against the orientation model. The model specified a characteristic distribution of the local oriented edges with maximums corresponded to two horizontal orientations (dark-to-light and light-to-dark horizontal edges). Unlike facial landmarks, noisy regions as, for example, elements of clothing and hair usually had an arbitrary distribution of the oriented edges and were discarded by the orientation model. Figures 2*b*-*d* shows the stages 1-3 of the method and were described in more detail in [12].

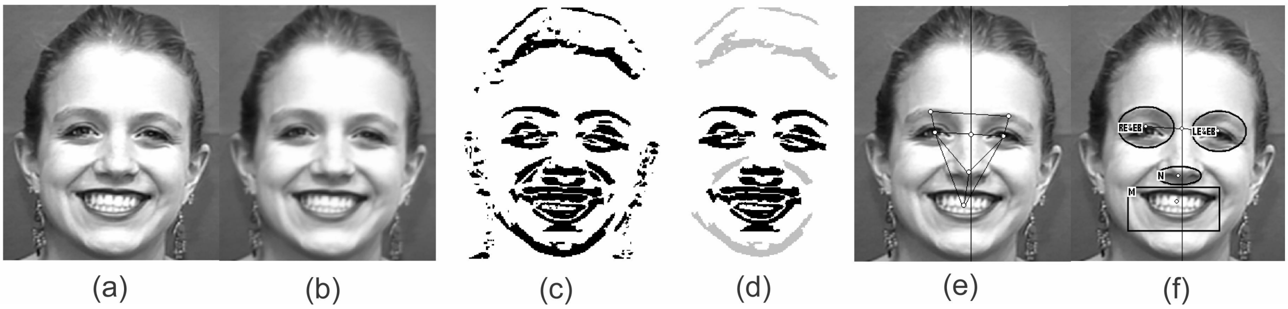


Figure 2. Facial landmark localization: (a) image of happiness; (b) smoothed image; (c) extracted local oriented edges; (d) edges grouped into regions representing landmark candidates (black) and noisy regions discarded by the edge orientation model (grey); (e) face geometry model; and (f) final localization result with "primary" landmarks (rectangles) and "secondary" landmarks (ovals). Image is a courtesy of the Cohn-Kanade database. Reprinted with permission.

A number of improvements were made to the earlier version of the method. First, instead of using a double average contrast of the whole high resolution image to define a threshold for contrast filtering of the located edge regions, we applied local contrast thresholding calculated in every filter neighbourhood. Second, the stage of edge map construction was improved as the method failed at this stage due to erroneous connection of edges belonging to different facial landmarks into one region. The one reason for that was a specific facial behaviour typically caused by AUs 9 (nose wrinkler), 10 (upper lip raiser), 11 (nasolabial furrow deepener), and 12 (lip corner puller). All the listed AUs result in deepening and pulling of the nasolabial furrow laterally up and raising of the upper lip up. Although there are marked differences in the shape of the nasolabial deepening and mouth shaping for these AUs, they all make the gap between nose and mouth smaller. In addition, nose as the most prominent feature in the face, often produces shadows and, thus, creates additional contrast in facial areas located below the nose. These changes in the lower face made the neighbourhood distances between edges extracted from nose and mouth smaller than a fixed threshold and caused a merging of nose and mouth into one region. Another reason for landmarks to be merged were AUs which are activated during anger, disgust, and sadness.

AU 4 (brow lowerer) pulls the eyebrows down and closer together producing vertical wrinkles between them. Further, AU 6 (cheek raiser) and AU 7 (lid tightener) often activated together with AU 4 also create additional contrast around eye regions. AU 9 (nose wrinkler) causes wrinkles to appear along the sides of the nose and across the root of the nose. These facial behaviours resulted in extracting of the noisy edges between the eyes and from the nose bridge region and, in some cases, caused a merging of eyes or eyebrows into one region.

To separate the neighbouring candidates merged into one region we proposed a simple but effective technique of x/y-edge projection. The schematic interpretation of the proposed technique is illustrated in Figure 3. It shows the merged facial landmarks and their separation by x/y-edge projection. 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 a number of edge points along the corresponding columns or rows of the edge map for the upper or lower face landmark candidate, respectively. If the number of projected edge points was smaller than a threshold (bold dashed line in the figures), edge points were eliminated. After each edge elimination step, if the region still was not separated the threshold was increased by 5 edge points. The initial threshold equalled a minimum number of edges in the column or row of the given candidate.

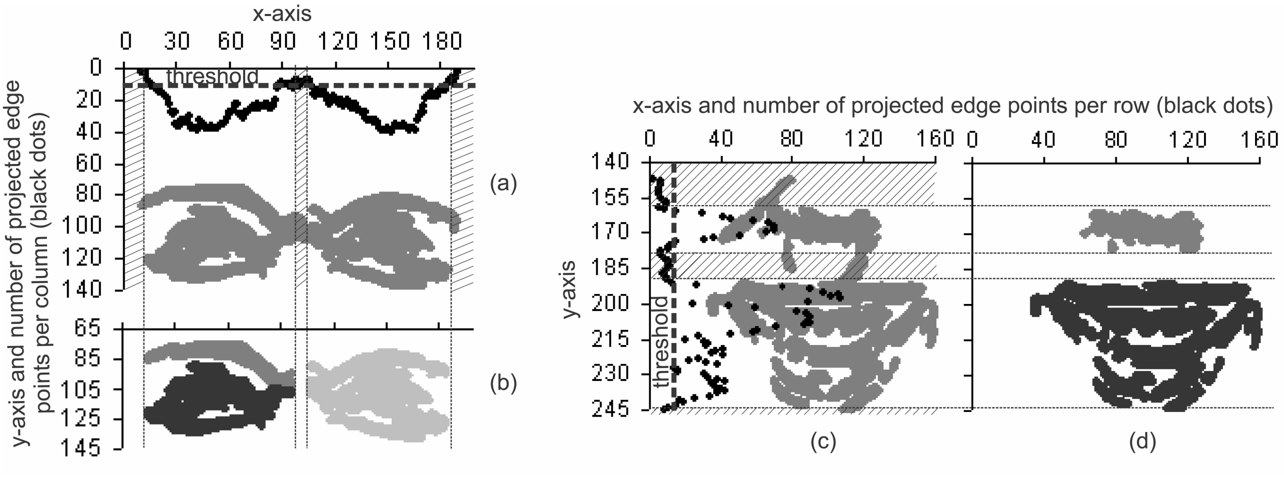


Figure 3. Edge maps of facial landmarks: (a) eye regions wrongly detected as one region; (b) eye regions separated by edge projection; (c) nose and mouth wrongly detected as one region; and (d) nose and mouth separated by edge projection. Black dots represent a number of projected edge points per columns or rows of the upper or lower face landmark candidates, respectively. Areas marked with upward diagonal lines show regions of the edge map where edges were eliminated.

Some changes were also made on the stage of edge orientation matching. The orientation portraits of the landmark candidates were allowed to have slight deviations from the orientation model. In particular, the orientation portrait of the candidate could have not strongly pronounced horizontal dominants in the edge orientation distribution and could have some orientations represented by zero number of edges. If orientation portrait of the candidate corresponded to the model, the candidate was labelled as “primary” candidate. On the contrary, if orientation portrait had slight differences from the orientation model, the corresponded candidate was labelled as “secondary” candidate. In further analysis, “secondary” candidates were also considered in the composition of face-like constellations of the candidates if there were missing landmarks.

Finally, the algorithm of fully automatic classification of the located candidates was applied on the stage of structural correction (see Figure 1). Upper face landmarks to be located were defined as right eye (RE), left eye (LE), right eyebrow (REB), left eyebrow (LEB), right eye region (RER), and left eye region (LER). In two latter cases, eyes and eyebrows were considered as one landmark. Lower face landmarks were represented by lower nose (N) and mouth (M). Due to side-by-side location of the upper face landmarks, their locations guided the entire process of the candidate classification, also those candidates which were discarded by the orientation model. After the face-like constellation of the candidates was found, the location of the face in the image was also known.

In order to find a proper spatial arrangement of the located candidates, the proposed algorithm applied a set of verification rules which were based on the face geometry model (see Figure 2,*e*). The knowledge on face geometry was taken from the anthropological study by Farkas [19]. This thorough study examined thousands of Caucasians, Chinese, and African-American subjects in order to determine characteristic measures and proportion indexes of the human face. The average distance between upper face landmarks (both eyes and eyebrows) *d(RER,LER)* and eyebrow-eye distance appeared to be useful facial measures for the purpose of structural correction of the located candidates. It has been demonstrated that these facial measures can slightly vary between subjects of deferent genders and races. Therefore we defined these measures as intervals between minimum and maximum values for the given database. The defined intervals were [35,85] and [10,35] for Cohn-Kanade, [60,90] and [10,40] for POFA, and [55,80] and [10,30] for JAFFE. The algorithm had several steps which are described in Figure 4.

1. The search started with finding all horizontally aligned pairs of upper face candidates at the distance *d(RER,LER)* with approximately equal number of edges. While searching for upper face candidates, the "primary" candidates were given the highest priority. Each candidate in the horizontal pair could have only one connection. In case of multiple connections, connections which had the longest length and the largest horizontal declination were eliminated until there was only one connection left.

2. Among the found upper face candidate pairs, there usually existed some noisy candidates, for example, elements of hair, closing, or decoration which needed to be eliminated. To do that, the distances between candidates and their relative locations in the face were verified.

   2.1. Although the algorithm was allowed to miss landmarks, however, at least one horizontal pair had to be found. If no pair was found, it was assumed that eye regions were merged together and edge x-projection was applied to the candidate with biggest number of edges located in the upper part of the image. After this step, edge map construction and edge orientation matching were applied to the received edge regions. After that the search started from the step 1.

   2.2. If one upper face candidate pair was found, it was labelled as eye region candidates. Using the eyebrow-eye distance, eyebrows above and eyes below the found pair location were searched for and if found any, the candidates were relabelled as eyes and eyebrows, respectively.

   2.3. If two upper face candidate pairs were found, they were verified by the eyebrow-eye distance and labelled as eyebrows and eyes, respectively.

   2.4. If more than two upper face candidate pairs were found, they were verified by the eyebrow-eye distance. If the distance between right or left candidates of the neighbouring pairs was less than the eyebrow-eye distance, the candidates were merged together. Otherwise, the candidate of the upper pair was eliminated. After there were one or two pairs left, the search started from the step 2.

   2.5. If the height of the upper face candidate was larger than a half of a dynamic parameter D_u that was calculated using a distance between mass centres of the upper face candidates, it was assumed that upper face candidate was merged with a hair. In this case, edge y-projection was applied in order to separate the candidate from the hair. After that the search started from the step 2.

3. At this step, the algorithm calculated a middle point between upper face landmark pair (eye or eye region pair) and built a vertical line of face symmetry called face axis.

4. The search for lower face landmarks was performed from top-to-bottom along the face axis. In order to verify spatial constraints between upper and lower face candidates and remove noisy candidates, the algorithm applied a dynamic parameter D_u. If the candidate mass centre was found down to the middle point between upper face landmark pair at a distance interval $[0.5D_u, 0.7D_u]$, it was marked as nose. If the candidate mass centre was found down to the midpoint at a distance interval $[1.2D_u, 1.7D_u]$, it was marked as mouth. This way, by utilizing geometrical constraints among the lower face candidates, the algorithm verified the upper face candidate locations.

   4.1. If only one lower face candidate was found, it was assumed that nose and mouth were combined together and edge y-projection was applied to separate these landmarks. After that the search started from the step 4.

Figure 4. Algorithm of structural correction of the located landmark candidates.

Fig. 2,*f* demonstrates the final result of the facial landmark localization. A localization result was defined as a rectangular box placed over the located region, not as a single point as typically has been the case. The location and size of the bounding box were calculated from the coordinates of the top, left, right, and bottom boundaries of the edge region. The mass centre of the located region indicated an estimate of the centre of the landmark.

IV. PERFORMANCE EVALUATION MEASURE

To our knowledge, the performance evaluation of different facial feature detectors proposed in the literature was given in terms of either visual inspection of the detection result or error measure calculated as a distance between manually annotated and automatically detected feature point locations. We do not consider visual inspection as an appropriate evaluation measure for any feature detector as it is a subjective decision and, therefore, can not give objective criteria of what to consider as a correct detection result. An error measure has been usually reported in terms of Euclidean pixel distance - the fewer pixels there were, the better the accuracy of the feature detector. In the work by Jesorsky et al. [20] a detection result was considered correct if the distance between manually annotated and automatically detected feature point location was less than 1/4 of the annotated intereye distance. This point measure is sufficient for all applications which can make use of a single pixel point result as an output of the feature detector. However, there is a number of applications which require a feature detector to find a region of facial landmark rather than to give a point solution. To our knowledge, there are no criteria for evaluation of the landmark detection result which is represented by a region in the image, not a single point.

In order to create a description of the landmark location in the image, we selected a set of characteristic points shown in Figure 5. Four points were selected to define locations of the eye, eye region, and mouth. These points defined the right, left, top, and bottom boundaries of these landmarks. Three points were used to describe locations of the eyebrow and nose. The eyebrow location was described by its top, bottom-left, and bottom-right points. Because it was unclear how to define the vertical dimensions of the lower nose, this region was defined by the centre point of the nose tip and locations of the nostrils. All the characteristic points were manually annotated in all the databases. Further, bounding boxes were built on the base of the selected characteristic points for each landmark in all the databases. The centre of the landmark was defined as the centre point of the bounding box [19].

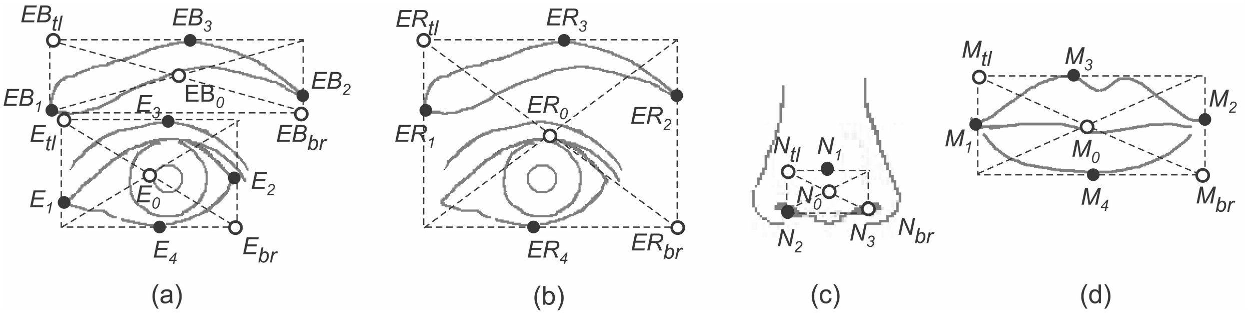


Figure 5. A set of characteristic points selected to define landmark location in the image: (a) eyebrow and eye; (b) eye region; (c) lower nose; (d) mouth. The bounding box that contains a landmark was defined by its top-left (*tl*) and bottom-right (*br*) bounding coordinates. The center point of the landmark was defined by the center point of the bounding box.

The centre point, top-left, and bottom-right coordinates of the bounding box were assumed to provide a good description of the landmark location in the image and be potentially useful for the purpose of the method performance evaluation. The pilot test was performed to verify this assumption. The characteristic points were manually annotated by 19 participants using a small dataset of images which well reflected the variations in facial expressions. The results of the pilot test supported our initial assumption; however, they also demonstrated that eyebrow characteristic points were difficult to define by humans and would not be as useful in the evaluation of the method performance as anticipated. In particular, some images contained subjects with very light eyebrows or eyebrows were covered by hair. For nearly all human subjects it was difficult to define boundary points for these landmarks. For some subjects, the difference in the annotation results was more than 15 pixels while the average length of the eyebrow was 38 pixels. As we aimed to compare the results of the automatic landmark localization to the human annotation results, the impact of poor definition of the eyebrow boundary points on the method performance evaluation was prevented by making the eyebrow a complementary landmark to locate. It means that the localization performance of the upper face landmarks depended only on the localization of the eyes.

The correctness of the localization result for nose was defined as a distance from manually annotated and automatically located centre of the nose tip. The correctness of the localization result for upper face landmarks and mouth was defined by a rectangular measure given in Equation 8:

$$\max(d(p_{tl}, \overline{p}_{tl}), d(p_{br}, \overline{p}_{br}) \leq R\,,\ R = N \cdot StDev\,. \qquad (8)$$

where R is a performance evaluation measure; p_{tl}, p_{br}, $\overline{p}_{tl}$, and $\overline{p}_{br}$ define the centre point, top-left and bottom-right coordinates of the manually annotated and automatically located landmark, respectively; N is a number that sets a desirable accuracy of the localization result; $StDev = 2$ pixels is a standard deviation of the manual annotation averaged over all the characteristic points in the pilot test. If $\overline{p}_{tl}$ and $\overline{p}_{br}$ were found inside the manually annotated landmark position, $\overline{p}_{tl}$ and $\overline{p}_{br}$ should be located in the top-left and bottom-right quadrants of the bounding box which includes the annotated landmark.

The rate of the landmark localization was defined as a ratio between a total number of correctly located landmarks and a total number of images used in testing (as there was one face per image). Eye region localization was counted correct in both cases – if bounding box included both eye and eyebrow, or if eye and eyebrow were located separately. If eyebrow was located as a separate region, it was obligatory that a corresponding eye was also found. Localization result was considered as a misclassification if landmark was located correctly but erroneously classified as another landmark. Localization result was classified as wrong if bounding box covered several neighbouring facial landmarks, excluding the case of eyes and eyebrows located as one region. Localization result was counted as a false localization if bounding box included a non-landmark, for example, wrinkles in the face, ears, clothing, hair, and eyebrow located without a corresponding eye.

V. RESULTS

Figure 6 demonstrates average orientation portraits of facial landmarks for all the databases. The results confirmed that individual orientation portraits of facial landmarks generally followed the rules predefined by the orientation model. On the contrary, the orientation portraits of noisy regions usually had an arbitrary distribution of local oriented edges and were discarded by the orientation model. Figures 7 and 8 demonstrate the final results of the landmark localization on the Cohn-Kanade and JAFFE databases, respectively. A bounding box that was built on the basis of the top-left and bottom-right coordinates of the edge region defined the location of the landmark in the image. The mass centre of the edge region indicated an estimate of the centre of the landmark. As figures show, the shape and size of the landmarks varied significantly with changes in facial expressions. Accordingly, the size of the bounding box was dynamic and varied in compliance with a size of the located edge region. This allowed, for example, to locate open and tight mouth as illustrated in Figures 8b and g.

Figure 9 demonstrates the results of the method performance evaluated by the conventional point measure calculated as a distance between manually annotated landmark centre and mass centre of the

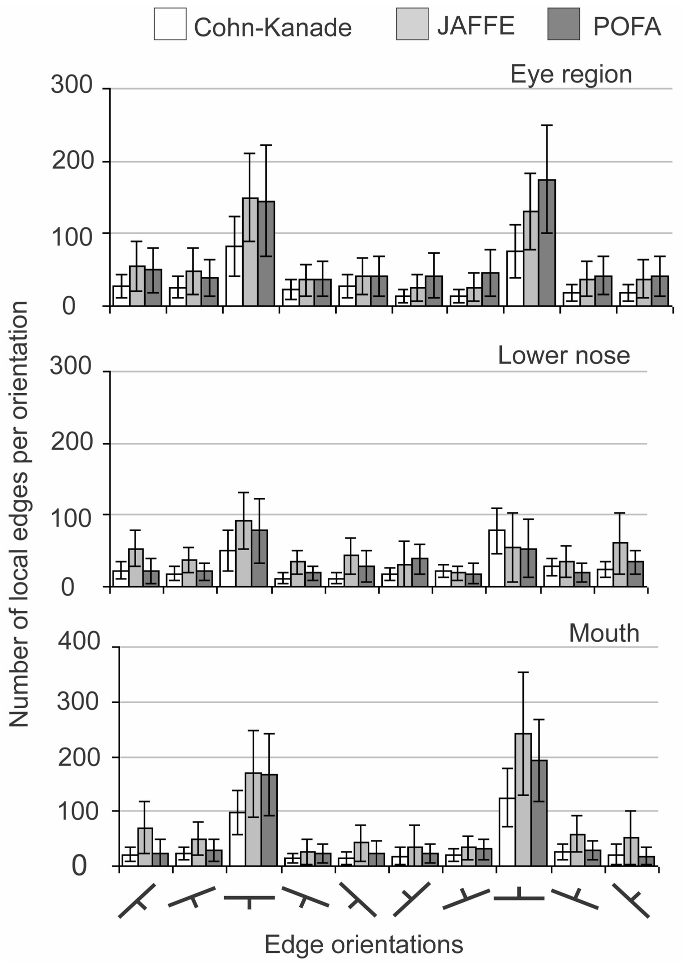


Figure 6. Examples of landmark orientation portraits averaged over all datasets. The error bars show plus/minus one standard deviation from the mean values.

automatically located edge region. The results of the method performance evaluated by the proposed rectangular measure are shown in Figure 10. The performance of the method in this case was evaluated by the rectangular measure that was calculated as a maximum distance between manually annotated and automatically located top-left and bottom-right corners of the rectangular box which contained a landmark (Equation 8). In both cases, as a measure of distance between annotated and automatically found landmark location we used a standard deviation of the manual annotation of the characteristic points averaged over all subjects in the pilot test.

Figures 9 and 10 show that for Cohn-Kanade datasets the performance of the method was higher for distance measure R_1 as compared to R_2 which equalled to 1/5 and 1/4 of the average intereye distance for a given dataset, respectively. For POFA and JAFFE neutral and expressive datasets, the landmark localization rates increased rapidly with the increase in the distance measures. In this case, the method located landmarks with relatively high rates for both distance measures.

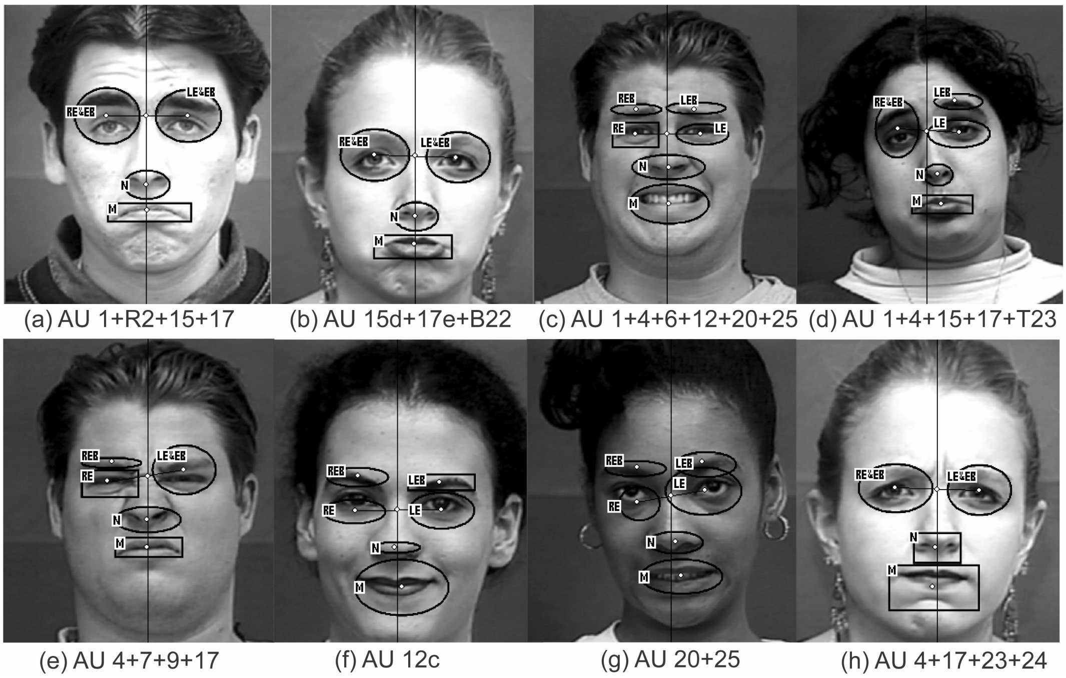


Figure 7. Examples of the final localization results with "primary" landmarks (rectangles) and "secondary" landmarks (ovals) of the landmark localization in the Cohn-Kanade images. Reprinted with permission.

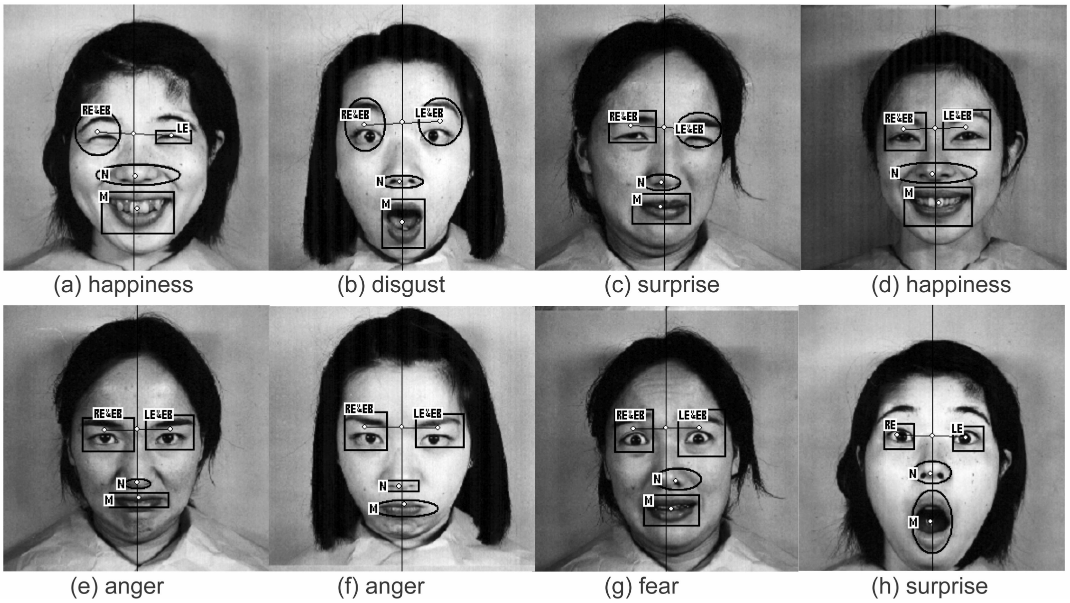


Figure 8. Examples of the final localization results with "primary" landmarks (rectangles) and "secondary" landmarks (ovals) of the landmark localization in the JAFFE images. Reprinted with permission.

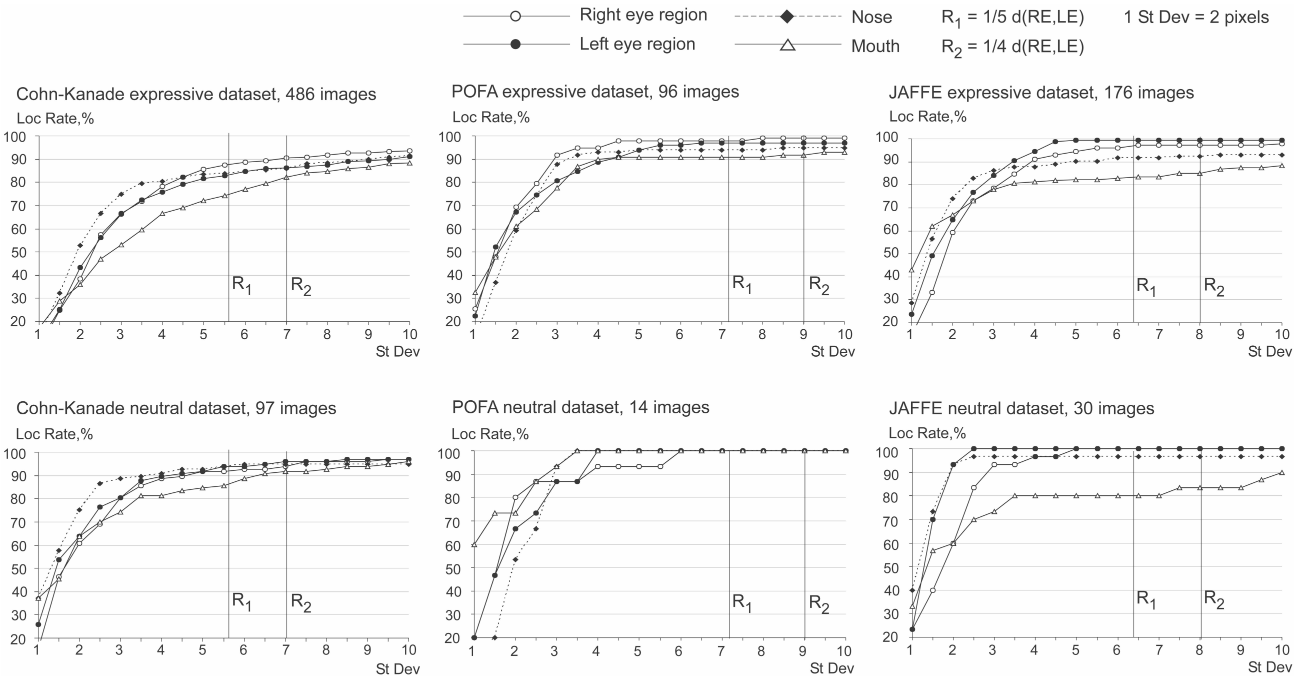


Figure 9. The performance of the method on Cohn-Kanade (first column), POFA (middle column), and JAFFE (right column) neutral and expressive datasets evaluated by the conventional point measure. The vertical lines indicate a distance measures which equaled to 1/4 and 1/5 of the average intereye distance for a given dataset.

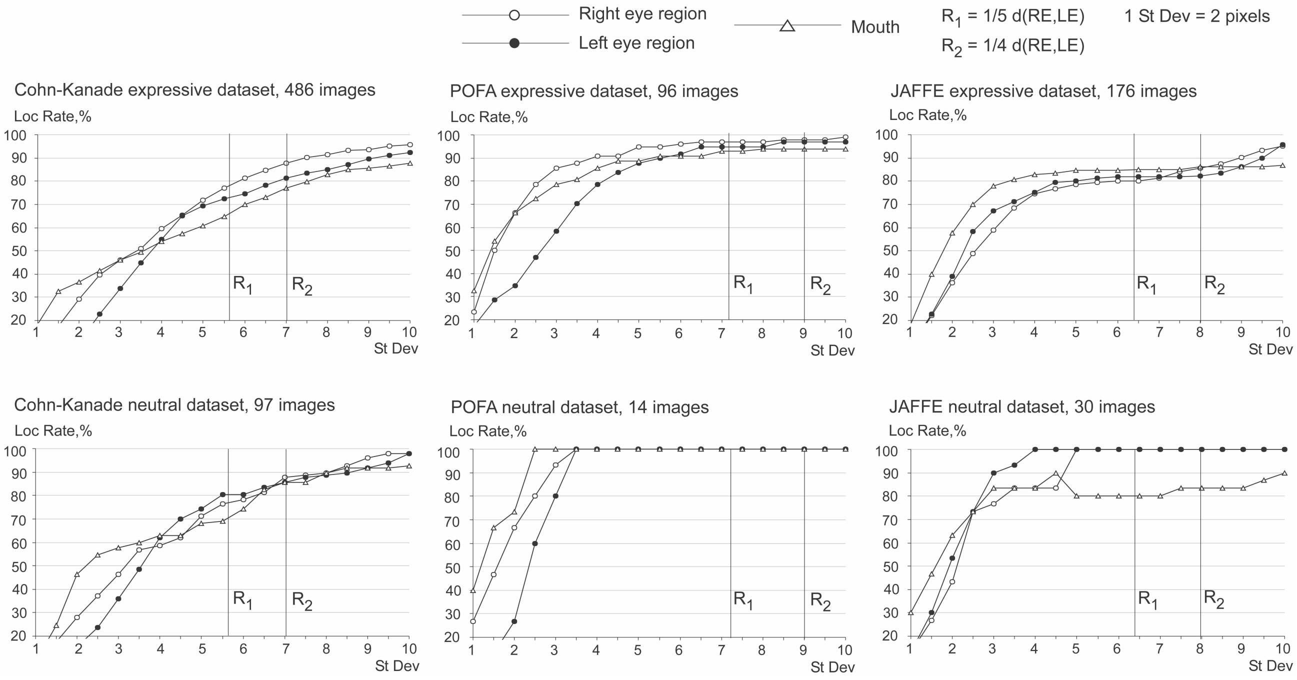


Figure 10. The performance of the method on Cohn-Kanade (first column), POFA (middle column), and JAFFE (right column) neutral and expressive datasets evaluated by the proposed rectangular measure. The vertical lines indicate a distance measures which equaled to 1/4 and 1/5 of the average intereye distance for a given dataset.

The performance of the method evaluated by the conventional and rectangular evaluation measures is summarized in Table II. For both types of the evaluation criteria the landmark localization was robust with respect to facial expressions as landmarks were located with nearly equal rates from neutral and expressive datasets. A decrease in the localization rates was observed for Cohn-Kanade and JAFFE expressive datasets; on the whole, however, the overall performance of the method as evaluated by the point and rectangular measures was high. A good method performance as evaluated by the rectangular measure reflected the fact that in most cases the method located landmark positions precisely meaning that inside the bounding box area surrounding landmark was less than the actual size of the landmark (for an opposite example refer to Figure 7h).

TABLE II: METHOD PERFORMANCE ON NEUTRAL AND EXPRESSIVE DATASETS, $R = {}^{1}\!/_{4}\,\mathrm{d}(\mathrm{RE},\mathrm{LE})$

Datasets	Method performance evaluated by point measure				Method performance evaluated by rectangular measure		
	L eye r	R eye r	Nose	Mouth	L eye r	R eye r	Mouth
Cohn-Kanade, expressive	90%	86%	86%	82%	88%	81%	77%
Cohn-Kanade, neutral	94%	96%	95%	92%	88%	86%	86%
POFA, expressive	99%	97%	95%	92%	98%	97%	94%
POFA, neutral	100%	100%	100%	100%	100%	100%	100%
JAFFE, expressive	97%	99%	92%	85%	86%	82%	86%
JAFFE, neutral	100%	100%	97%	83%	100%	100%	83%
Average	94%				91%		

The errors in the landmark localization were tracked manually and classified into misclassification, false localization, and wrong localization groups (Table III). Misclassification errors appeared on the last stage of the method due to errors in the algorithm of structural correction. Misclassification of the upper face landmarks was mainly due to the classification of eyebrows and eye regions as eyes. Lower face misclassification was due to the classification of nose as mouth and vice versa. Wrong localizations were mainly observed in the localization of lower face landmarks. As it was expected, it occurred mainly due to the effect of lower face AUs 9, 10, 11, and 12 activated alone or in combinations with other AUs, for example, in expressions of anger, disgust, and happiness. AUs 4, 6, and 7 sometimes caused the merging of the upper face landmarks into one region. Figure 11 illustrates some examples of the localization errors.

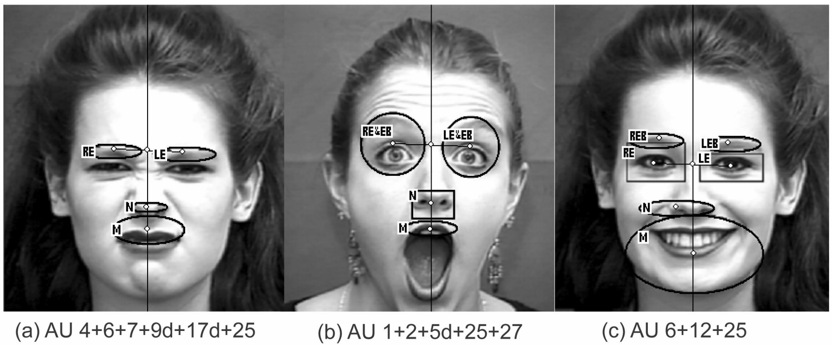
TABLE III: SUMMARY OF ERRORS OF THE METHOD PERFORMANCE ON NEUTRAL AND EXPRESSIVE DATASETS

Datasets	Misclassifications		Wrong localizations		False localizations	
	Upper face landmarks	Lower face landmarks	Upper face landmarks	Lower face landmarks	Upper face landmarks	Lower face landmarks
Cohn-Kanade, expressive	34	44 (7)	7	15	50	28
Cohn-Kanade, neutral	3	2 (1)	2	-	7	3
POFA, expressive	5	2	-	4	-	3
POFA, neutral	-	-	-	-	-	-
JAFFE, expressive	3	1 (3)	1	5	-	4
JAFFE, neutral	-	-	-	-	-	-

Note that numbers in brackets define wrong landmark localizations which resulted into the misclassification error.

It has been demonstrated that the method produced the lowest localization rates and biggest number of localization errors for Cohn-Kanade expressive dataset. This fact suggested a further analysis of the method performance on this database to reveal the influence of particular facial behaviours on the localization results as evaluated by the rectangular measure. Additionally, we were interested what kind of improvement in the method performance was achieved as compared to the earlier version of the method [14]. In that study we discovered that the method performance was especially deteriorated by the lower face AUs 9, 10, 11, and 12 and some combinations of these AUs with others. A comparison of the landmark localization rates of the present and earlier versions of the method was made for images with upper face AUs and AU combinations (Table IV), lower face AUs (Table V), and lower face AU combinations (Table VI).

As it is seen from the tables, localization rates were increased over 20% for 25 out of 47 different facial deformations. A significant improvement in the landmark localization was achieved for images with



(a) AU 4+6+7+9d+17d+25 (b) AU 1+2+5d+25+27 (c) AU 6+12+25

Figure 11. Examples of errors of the landmark localization: (a) eyebrows were misclassified as eyes; (b) upper lip was classified as mouth; and (c) mouth was merged with chin. Images are courtesy of the Cohn-Kanade and JAFFE databases. Reprinted with permission.

lower face AUs and AU combinations. For example, the improvement in the localization rates for lower face AUs 9, 9+17, 12+16, 16+20 was over 30%. The same improvement was achieved for upper face AU combinations 1+6, 4+6, 6+7.

Due to the fact that some AUs were not presented in the dataset or the number of images was too few (less than 6), only a limited number of AUs and AU combinations was used. This type of the result classification allowed the results to belong to more than one group. Moreover, AUs from the tables

TABLE IV: IMPROVEMENT OVER UPPER FACE AUS AND AU COMBINATIONS

Study / *AUs*	1	2	4	5	6	7	43/45	1+2	1+4	1+5	1+6	1+7	2+4	2+5	4+5	4+6	4+7	4+43/45	6+7
Ave Loc Rate, % Present study	84	83	85	86	77	81	91	83	82	84	92	80	80	84	85	82	82	84	76
Ave Loc Rate, % Study [14]	81	86	64	84	52	61	63	86	73	86	62	58	82	86	72	40	62	56	46
Improvement, %	2	-3	21	2	25	20	28	-3	9	-2	30	21	-2	-2	13	42	20	28	30

Note that AU43 (eye closure) and AU45 (blink) were combined together because they both have the same visual effect on the facial appearance and different durations of these AUs can not be measured from the static images.

TABLE V: IMPROVEMENT OVER LOWER FACE AUS

Study / *AUs*	9	10	11	12	14	15	16	17	20	23	24	25	26	27
Ave Loc Rate, % Present study	84	83	85	77	68	94	88	89	83	88	86	84	92	83
Ave Loc Rate, % Study [14]	30	58	66	60	77	70	67	64	68	78	79	68	71	90
Improvement, %	53	25	20	17	-9	24	21	25	16	9	7	16	21	-6

TABLE VI: IMPROVEMENT OVER LOWER FACE AU COMBINATIONS

Study / *AUs*	9+17	11+20	11+25	12+16	12+20	12+25	16+20	16+25	17+23	17+24	20+25	23+24	25+26	25+27
Ave Loc Rate, % Present study	83	88	88	86	63	74	94	86	87	86	83	86	91	84
Ave Loc Rate, % Study [14]	41	67	67	50	42	54	54	61	72	76	67	77	70	90
Improvement, %	42	21	21	36	21	19	40	25	14	9	16	10	21	-6

usually occurred singly or in conjunction with other AUs which are not represented in the tables. That is why the present results revealed an indirect effect of different AU and AU combinations on the landmark localization.

VI. DISCUSSION

A fully automatic method was designed for feature-based localization of facial landmarks in static grey-scale images. The local oriented edges served as basic image features for expression-invariant representation of facial landmarks. The results confirmed that in the majority of the images, the orientation portraits of facial landmarks had the same structure as predefined by the orientation model. This allowed to discard non-landmark regions like, for example, wrinkles in the face, ears, earrings, and elements of hair and clothing. Structural correction of the located candidates based on face geometry model further improved the overall performance of the method.

The performance of the method was examined on three databases of facial images showing complex facial expressions. The complexity of the expressions was presented by a variability of the deformations in soft tissues (wrinkles and protrusions), variety of mouth appearances including open and tight mouth, visible teeth and tongue, self occlusions (semi- and closed eyes and bitted lips). The results of the landmark localization were evaluated by the conventional point measure and the proposed rectangular measure that represented a localization result as a rectangular box, not a single point. Both types of evaluations demonstrated a sufficiently high overall performance of the method (94% as evaluated by the point measure and 91% as evaluated by the rectangular measure) that is comparable to some extent with performance of other facial feature detectors reported in the literature [9,10,21,22,23].

The comparison of the present results with the results of the method testing on the Cohn-Kanade dataset obtained in the earlier study [14] demonstrated that the modifications made to the method significantly improved the overall performance of the method. One has to be cautious when comparing the localization rates from Tables IV-VI because the implementation of the method, test setup, and evaluation criteria were different in these two studies. However, the general trend in the method improvement can be brought into the light.

As one significant problem in the earlier method was the deteriorating effect of expressions of happiness (when AUs 6 and 12 are usually activated), anger (when AUs 4, 6, and 7 are usually activated), and disgust (when AUs 9, 10, and 11 are usually activated). Occurring along or in conjunction, the listed

AUs caused a wrong localization error in the landmark localization. In the current version of the method, the applied procedure of x/y-edge projection in many cases allowed the merged landmarks to be separated. In particular, the merged nose-and-mouth landmark was successfully separated, especially in the images with AUs 10, 11, 12, 15, 16, 17, 21, 11+20, 11+25, 12+20, 16+25, and 25+26. Further, an essential improvement in locating landmarks was achieved for images with AUs 4, 6, 7, 43/45, 1+6, 1+7, 4+6, 4+7, 4+43/45, 6+7, and 9. These facial behaviours typically cause the whites of the eyes to become less or not at all visible in the face. These facial behaviours cause serious errors in the performance of eye detectors which rely particularly on the searching for pupils and whites of the eyes in the image. The present method can deal with self-occlusions in the eye regions by analyzing a local edge description of the whole eye region.

Generally, in all the databases the mouth region demonstrated a greater variability in its appearance than the regions of eyes or nose. For example, in the images of surprise and happiness the mouth appearance usually was represented by open mouth (when AU combination 25+26 is activated) sometimes with visible teeth and tongue. In the images of anger the mouth could be opened (AU 22+25+26) or closed with tightened lips (AU 23), pressed lips (AU 24), and even lips sucked inside the mouth (AU 28) so that the red part of the mouth became not visible in the face. These facial behaviours restrict the applicability of the colour-based methods of mouth detection in which colour is the key feature in the mouth representation. On the contrary, the proposed method was able to find the mouth regardless of whether the mouth was open or closed and whether the lips, teeth or tongue were visible or not (Figures 7-8).

Emphasizing the simplicity of the proposed method, we conclude that it can be used in the preliminary localization of regions of facial landmarks for their subsequent processing where coarse landmark localization is followed by fine feature detection. For example, eye and mouth corners can be searched for in the located regions. The method can be applied directly to the image without any image alignment given that a face takes the biggest part of the image. The method does not require intensive training either. These qualities gives the method an advantage over PCA, AdaBoost, or neural network based methods which require either facial alignment or intensive training made preliminary to the facial landmark localization.

Currently, the method was applied to static images where no temporal information was available, only structural information was used. The processing time of the method was slowed down by the stage of edge detection where each point of the image was checked for a local oriented edge. To increase the speed of the method for practical applications, face region estimation can be utilized as a preliminary step to edge extraction in order to discard those parts of the image which a priori can not contain facial landmarks. Our future plan is to utilize the developed method in real-time system of facial landmark detection and tracking for the purpose of facial expression analysis in human-computer interaction.

ACKNOWLEDGMENTS

This work was supported by the Finnish Academy (project number 177857 and 1115997), the University of Tampere, and the Tampere Graduate School in Information Science and Engineering. The authors thank the creators of Cohn-Kanade and JAFFE databases for the permission to reprint examples of the expressive images.

REFERENCES

[1] P. Ekman, The argument and evidence about universals in facial expressions of emotion, in: H. Wagner, A. Manstead (eds.), Handbook of Social Psychophysiology, Lawrence Associates Press, London, 1989, pp. 143-164.

[2] P. Ekman, W. Friesen, Facial action coding system (FACS): A technique for the measurement of facial action, Consulting Psychologists Press, Palo Alto, California, 1978.

[3] P. Ekman, W. Friesen, J. Hager, Facial action coding system (FACS), A Human Face, Salt Lake City, UTAH, 2002.

[4] E. Hjelmas, B. Low, Face detection: A survey, Computer Vision and Image Understanding 83 (2001) 235–274.

[5] M. Yang, D. Kriegman, N. Ahuaja, Detecting face in images: A survey, IEEE Trans. Pattern Analysis and Machine Intelligence 24 (2002) 34-58.

[6] M. Pantic, J.M. Rothkrantz, Automatic analysis of facial expressions: The state of the art, IEEE Trans. Pattern Analysis and Machine Intelligence 22, 12 (2000) 1424–1445.

[7] M. Burl, T. Leung, P. Perona, Face localization via shape statistics, in Proc. 1st Int. Workshop Automatic Face and Gesture Recognition, Zurich, Switzerland, June 1995, pp. 154-159.

[8] L. Wiskott, J.-M. Fellous, N. Kruger, C. von der Malsburg, Face recognition by elastic bunch graph matching, IEEE Trans. Pattern Analysis and Machine Intelligence 19, 7 (1997) 775–779.

[9] R. Feris, J. Gemmell, K. Toyama, V. Krüger, Hierarchical wavelet networks for facial feature localization, in Proc. 5th IEEE Int. Conf. Automatic Face and Gesture Recognition, Santa Barbara, CA, May 2002, pp. 118-123.

[10] D. Cristinacce, T. Cootes, Facial feature detection using AdaBoost with shape constraints, in Proc. 14th British Machine Vision Conf., Norwich, England, September 2003, pp. 231-240.

[11] P. Viola, M. Jones, Robust real-time face detection. Int. J. Computer Vision 57, 2 (2004) 137-154.

[12] Y. Gizatdinova, V. Surakka, Feature-based detection of facial landmarks from neutral and expressive facial images, IEEE Trans. Pattern Analysis and Machine Intelligence 28, 1 (2006) 135-139.

[13] I. Guizatdinova, V. Surakka, Detection of facial landmarks from neutral, happy, and disgust facial images, in Proc. 13th Int. Conf. Central Europe on Computer Graphics, Visualization, and Computer Vision, Plzen, Czech Republic, January 2005, pp. 55-62.

[14] Y. Gizatdinova, V. Surakka, Effect of facial expressions on feature-based landmark localization in static grey scale images, Int. Conf. Computer Vision Theory and Applications, Madeira, Portugal, January 2008, pp. 259-266.

[15] Y. Gizatdinova, V. Surakka, Automatic detection of facial landmarks from AU-coded expressive facial images, Int. Conf. Image Analysis and Processing, Modena, Italy, September 2007, pp. 419-424.

[16] T. Kanade, J. Cohn, Y. Tian, Comprehensive database for facial expression analysis, in Proc. 4th IEEE Int. Conf. Automatic Face and Gesture Recognition, Grenoble, France, March 2000, pp. 46-53.

[17] P. Ekman, W. Friesen, Pictures of facial affect, Consulting Psychologists Press, Palo Alto, California, 1976.

[18] M.J. Lyons, Sh. Akamatsu, M. Kamachi, J. Gyoba, Coding facial expressions with Gabor wavelets, in Proc 3d IEEE Int. Conf. Automatic Face and Gesture Recognition, Nara, Japan, April 1998, pp. 200-205.

[19] L. Farkas, Anthropometry of the head and face, (2 ed), Raven, New York, 1994.

[20] O. Jesorsky, K.J. Kirchberg, R.W. Frischholz, Robust face detection using the hausdorff distance, Lecture Notes in Computer Science, Proc. of the 3d Int. Conf. Audio- and Video-Based Person Authentication, Halmstad, Sweden, June 2001, pp. 90–95.

[21] P. Campadelli, R. Lanzarotti, G. Lipori, E. Salvi, Face and facial feature localization, in Proc. 13th Int. Conf. Image Analysis and Processing, Gagliari, Italy, September 2005, pp. 1002-1009.

[22] B. Fröba, C. Küblbeck, Orientation template matching for face localization in complex visual scenes, in Proc. Int. Conf. Image Processing, Vancouver, September 2000, pp. 251-254.

[23] D. Shaposhnikov, A. Golovan, L. Podladchikova, N. Shevtsova, X. Gao, V. Gusakova, Y. Gizatdinova, Application of the behavioural model of vision for invariant recognition of facial and traffic sign images, J. Neurocomputers: Design and Application 7-8 (2002) 21-33, (in Russian).