A HYBRID APPROACH TO MOUTH FEATURES DETECTION

MAJA PANTIC, MILAN TOMC and LEON J.M. ROTHKRANTZ

Delft University of Technology, Faculty of Information Technology and Systems
Department of Media, Knowledge Engineering and Mathematics,
P.O. Box 356, 2600 AJ Delft, The Netherlands

Abstract

This paper presents a novel, robust and flexible method for extracting four mouth features (top of the upper lip, bottom of the lower lip, left and right mouth corners) from facial image sequences. While robustness is referred to subject variability, pose, and image quality, flexibility is begotten by efficient fusion of several information sources and expounding the certainty of the generated results.

Keywords

Edge detection, fuzzy reasoning, ANN, rule-based reasoning, reasoning with uncertainty.

1 Introduction

Mouth feature extraction forms a fundamental stage of face image analysis for numerous application areas like man-machine interaction based on the observed human behavior, affective computing, videoconferencing, video-telephony, face and person identification, bimodal speech recognition, face and visual speech synthesis. In all of those applications there is a need of efficient and fully automated mouth feature extraction methods, characterized by high robustness to speaker and facial image characteristics variability.

In general, extraction of the prominent facial features (eyebrows, eyes, nose, mouth, and chin) from input facial images may be divided into at least three dimensions:

- are the features extracted in an automatic way,
- is temporal information (image sequence) used,
- are the features holistic (spanning the whole prominent facial feature) or analytic (spanning salient points of the prominent facial feature).

Given this glossary, most of the approaches to feature extraction in facial images are directed towards automatic, static, analytic extraction of the prominent facial features [7].

Likewise, most mouth feature extraction techniques rely on the localization of salient feature points of the mouth [7, 8]. Generally these feature-based (analytic) approaches have the advantage of being simple but lack in robustness; they work well only under strong environmental constraints like controlled illumination conditions, constant mouth appearance (i.e. pose, scale, and orientation), etc.

Model-based (holistic) mouth feature extraction techniques like Deformable Templates [12], Active Shape Models [5], and Snakes [3], approximate lips contours with spline functions, that is, with a given mouth model. Generally these methods achieve high robustness if reliable constraints on the mouth model deformation are included in the cost function, which is usually defined based upon heuristic considerations or through (supervised) learning from examples. Yet, the minimization of the cost function commonly involves computationally expensive algorithms and the model definition often employs (semi-automatic) user-guided procedures or a priori knowledge of environmental constraints under which the examined image has been obtained.

Thus both, feature-based and model-based mouth feature detectors have their vantages and limitations. Generally, their performance is constrained by mouth image characteristics for which the employed image processing technique performs well. This observation motivated our research on a robust mouth detector, which combines several distinct extraction techniques fashioning a hybrid, knowledge-based approach to mouth feature extraction from facial images.

Four salient mouth feature points, namely, top of the upper lip, bottom of the lower lip, left and right mouth corners are the most important mouth features. They provide:

- measures of the height and width of the outer contour of the mouth;
- measures for normalizing mouth images with respect to translation, rotation, and scaling; and
- a reference for extraction of other prominent facial features [2, 4, 11] and for recognition of most lower-face muscle actions [6] occurring frequently in speech and (affective/attitudinal) facial expressions.

These considerations motivated our effort to develop a robust and flexible algorithm for detection of four mouth features listed above.

The proposed approach is founded upon a multi-phase multi-detector processing of input facial image sequences coupled with a rule-based reasoning with

uncertainty about mouth features' location. The three phases of the proposed algorithm comprise: coarse mouth detection, fine mouth detection, and mouth features extraction. The paper begins by describing each of these phases. It presents further the experimental results and, finally, discusses vantages and limitations of the proposed method.

2 Coarse mouth detection

Most of the existing approaches to mouth feature extraction assume that the presence of the mouth in the input image is ensured [3, 12, 5, 8]. Nevertheless, in most of the real-life situations where an automated mouth detector is to be employed (e.g. man-machine interaction, videoconferencing, etc.) even the location of the face in the scene is not known a priori, let alone the location of the mouth. Hence, the first step in automatic mouth feature extraction is to ascertain the presence of a face in the scene and determine the Region Of Interest (ROI) in the input facial image (i.e. a more or less large area around the mouth).

The presence of a face in input images can be ensured either by employing an existing method for automatic face detection in arbitrary scenes (for a small review on such methods, readers are referred to [1]) or by using a camera setting that will ascertain the assumption at issue. The algorithm proposed here does not perform face detection in an arbitrary scene; it operates on facial image sequences obtained from typical audiovisual speech databases or acquired by the mounted-camera device illustrated in Fig. 1.

Fig. 1: Mounted-camera device

To determine the ROI in an input frame, we apply color-based segmentation technique. Given that the natural color of the lips is red, the input image is transformed into the HSV color space (in order to separate color from intensity) and then into red domain in order to segment the ROI from the rest of the image. Image in red domain is obtained from the hue component of the input frame by applying the following function (Fig. 2):

$$f(\boldsymbol{h}) = \begin{cases} 1 - \dfrac{(\boldsymbol{h} - \boldsymbol{h}_0)^2}{\boldsymbol{w}^2} & |\boldsymbol{h} - \boldsymbol{h}_0| \le \boldsymbol{w} \\ 0 & |\boldsymbol{h} - \boldsymbol{h}_0| > \boldsymbol{w} \end{cases}$$

where $\boldsymbol{h}$ is the shifted hue value of each pixel so that the $\boldsymbol{h}_0$=1/3 corresponds to red color, and $\boldsymbol{h}_0$ and $\boldsymbol{w}$ control the positioning and the width of the filter.

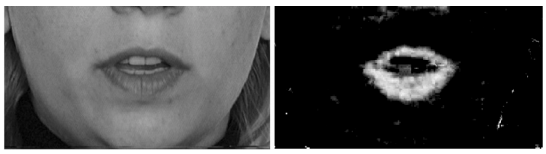

Fig. 2: Image and its corresponding red domain

Fig. 3: Image and the corresponding extracted face

Since there is a high variability in intra- and inter-person appearance as to the redness of the skin (intra-person variability is usually caused by illumination conditions variability), filter parameters $\boldsymbol{h}_0$ and $\boldsymbol{w}$ should be computed separately for each frame of the examined image sequence. To achieve this in a fully automatic way, the following procedure is executed:

1. Extract the face from the input image (Fig.3) as the biggest object in the scene having the hue value $\boldsymbol{H} \in [-40 < \boldsymbol{H}_{avg} - 20, \boldsymbol{H}_{avg} + 20 < 60]$, where $\boldsymbol{H}_{avg}$ is the average hue in the box containing the horizontal middle of the face. This box is determined based on a simple analysis technique of image histograms (the vertical image histogram shows the color-differences between the successive rows, pixel-wise). A similar approach to face extraction based upon the relative RGB model is presented in [10].
2. Based upon the result of the algorithm executed in step-**1** define box ***BF*** that bounds the face.
3. Define box ***BM*** (Fig. 2), bounding a more or less large area around the mouth, so that it lies on the line going through the mouth (4th peak of box ***BF*** vertical histogram), below the line going through the nostrils (3rd peak of box ***BF*** vertical histogram) and above the line representing the border between the chin and the neck (5th peak of box ***BF*** vertical histogram).
4. Optimize filter parameters $\boldsymbol{h}_0$ and $\boldsymbol{w}$ so that the maximal separability of ***BM*** and ***BF*** is achieved.

Once the values of the parameters $\boldsymbol{h}_0$ and $\boldsymbol{w}$ are computed for the given input image $\boldsymbol{I}$, the image in

red domain $\boldsymbol{I_{red}}$ is obtained by the formula given above. A threshold value determined as 90% of the maximal filter response in $\boldsymbol{I_{red}}$ is employed for the binarization of the image. Mathematical morphology is used further to fill small noisy holes inside the lips. The width and the height of the ROI are determined as the double width and height of the shrunken image (i.e. without empty rows and columns). The center of the ROI is set to the mean value of the red pixels coordinates in the shrunken image. To account for translations, in- and limited out-plane rotations, the horizontal boundaries of the ROI are positioned parallel to the mouth-through line, which is a distinct valley in the horizontal integral projection of the binarized image. To account for scale variations, a multi-resolution representation of the input frame is formed and the same procedure of the coarse mouth detection is executed at different resolutions. Typical examples of the ROI positioning are given in Fig. 4.

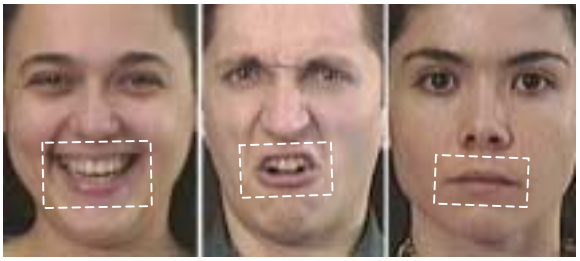

Fig. 4: Positioning of the ROI

3 Fine mouth detection

In its second phase, the proposed algorithm performs spatial sampling of the mouth contour and classifies the horizontal and the vertical movements of the mouth. To accomplish this, four procedures are concurrently applied to the previously determined ROI. This section explains these algorithms in detail.

3.1 Spatial sampling of the mouth

Two algorithms, namely *Curve Fitting of the Mouth* and *Mouth Template Matching*, localize the mouth contour in the input ROI.

Curve Fitting of the Mouth reduces the input image color depth of 24 bits to 256 gray levels and applies two-dimensional Gaussian low-pass filter. In the binarized image, the lowest highlighted pixel is then selected as the boundary-following starting point. A pixel directly connected to the current pixel, representing a zero crossing of the 2nd derivative function of the mouth image, continues the mouth boundary. The points, where the conjunction of the lips ends and changes in a disjunction, are marked as mouth corners. A refined estimate of the mouth shape is then obtained by fitting two 2nd-degree parabolas on the upper lip and a 2nd-degree parabola on the lower lip. The 2nd order least square model algorithm is used to find the best relation between the points of the extracted mouth contour and the parameters of each of the parabolas. Typical results of the *Curve Fitting of the Mouth* procedure are shown in Fig. 5.

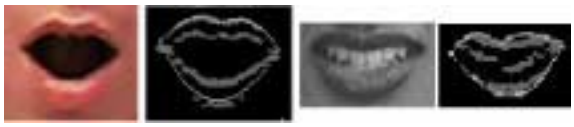

Fig. 5: Results of the *Curve Fitting of the Mouth*

Mouth Template Matching localizes the contour of the mouth in the ROI by fitting a two-dimensional model of the lips to the mouth. The utilized model consists of 8 parabolas (Fig. 6), which provide a parametric description of the mouth shape sufficient for inferring visual speech dynamics [8]. The eleven parameters of the utilized model are: four vertices (***v1*** - ***v4***), four apices (***a1*** - ***a4***), the center of the mouth ($\boldsymbol{C_x}$, $\boldsymbol{C_y}$), and the mouth slant $\boldsymbol{\phi}$. Since the aim is to extract merely four salient mouth feature points listed above, only four parameters (i.e. ***v1***, ***v2***, ***a1***, ***a2***) of the fitted model are considered in a further processing of the proposed algorithm.

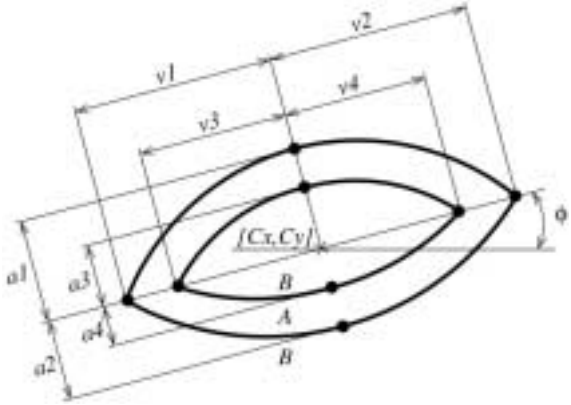

Fig. 6: 2D parameterized model of the lips

The initial positions of the outer vertices (***v1***, ***v2***) are defined by the outermost highlighted pixels in the filtered and binarized input ROI image. The inner vertices (***v3***, ***v4***) are initially set to zero. The initial positions of the apices are extracted from the input ROI image in red domain. Two vertical rectangles (***r1***, ***r2***) are positioned first in the input image (Fig. 7a). A vector ***d*** is then computed as the mean of two vectors obtained by summing up ***r1*** and ***r2*** along their rows. If ***d*** has a single maximum, the mouth is closed (Fig. 7b) and only the outer apices (***a1***, ***a2***) are initialized. Otherwise the mouth is open (Fig. 7c) and all apices are initialized. The points where the values of ***d*** are 60% of its maximum(s) define the initial positions of the apices. Fig 8 illustrates typical results of the *Mouth Template Matching* model initialization.

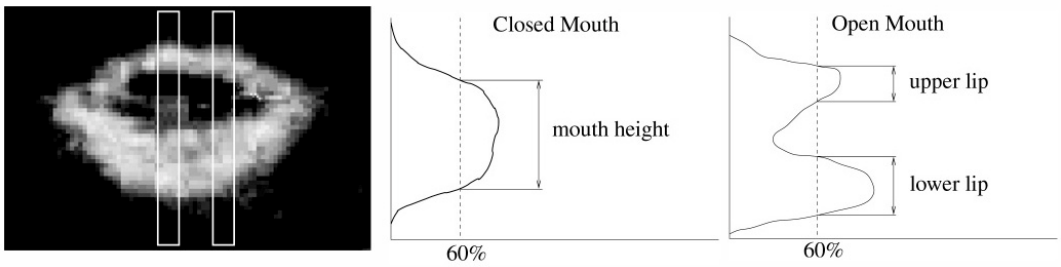

Fig. 7: (a) Regions for apices estimation in red domain image; (b) Closed mouth profile; (c) Open mouth profile

The utilized cost function is the sum of the frequencies of non-red pixels inside (region ***A*** in Fig. 6) and outside the lips (region ***B*** in Fig. 6) decreased by the sum of frequencies of red pixels inside and outside the lips. Optimization of the cost function is done in two epochs. In the first epoch, the center of the mouth, the slant angle, and the parameters of the outer contour of the mouth are change so that the cost function is minimized. In the second epoch, changing the parameters of the inner contour of the lips obtains further optimization of the cost function. Typical results of the *Mouth Template Matching* procedure are shown in Fig. 9.

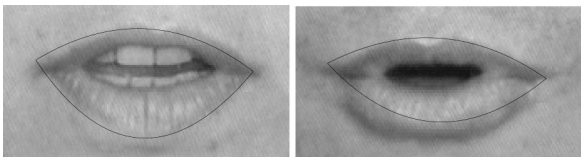

Fig. 8: Initial positioning of the model

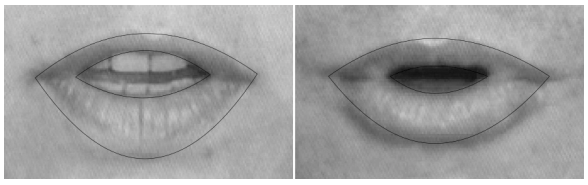

Fig. 9: Results of the *Mouth Template Matching*

3.2 Classification of the mouth movements

Examining children's or caricature drawings may lead to an interesting conclusion. The mouth shape can be shown using only a single drawing line that still perfectly reflects the intention of the drawer (Fig. 10). This leads further to the conclusion that an appropriate representation of the mouth shape may be the information about the average edge intensity and direction in the corners of the mouth. If the edge is on average "going up", mouth could be interpreted as "smiling". If the edge is on average "going down", mouth could be interpreted as "sad". This idea has been implemented in a form of an ANN classifier of "vertical" mouth movements [9]. If the edge is on average "protruding", mouth could be interpreted as "stretched". If the edge is on average "shrinking", mouth could be interpreted as "puckered". This idea has been implemented in a form of a rule-based classifier of "horizontal" mouth movements. Both, *Vertical ANN Mouth Classifier* and *Horizontal Rule-based Mouth Classifier* are based on fuzzy reasoning about the edge information.

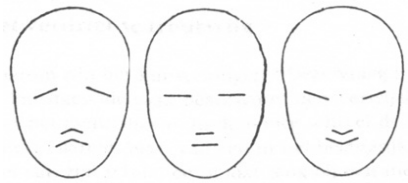

Fig. 10: Line drawings of faces

The fuzzy reasoning for edge detection is based on two characteristics of the gradient, namely, the gradient value corresponds with local steepness of the function and function is locally symmetrical along the gradient direction. It proceeds as follows. The numerical values typifying symmetry and steepness level are first fuzzified into the labels low, medium, or high, and then passed to the reasoning part of the process. The reasoning part is based upon nine rules such as "if the steepness and the symmetry level are high, then edge intensity in this point is high". It depicts the edge intensity in a given point in terms of the labels low, medium, and high. The information on the direction of the mouth symmetry axis is used to obtain the information about both, the intensity and the direction of the edge in a given point. Combined intensity and direction of the edge in a given point forms a vector representation of that point. The obtained vector field for the whole mouth region is then averaged and sampled in a fixed number of regions. We used a rectangular-grid decomposition of the mouth region with 10 columns and 5 rows of average edge direction vectors. The resulting 50 vectors (100 values) are further classified by an ANN in the case of the *Vertical ANN Mouth Classifier* and by a rule-based method in the case of the *Horizontal Rule-based Mouth Classifier*.

Vertical ANN Mouth Classifier utilizes a back-propagation ANN layout (Fig. 11) that reflects the vertical symmetry of the mouth. The implemented

architecture contains two 50×3×2 "features" ANNs set in parallel (one for each side of the mouth) whose output is passed further to a 4×3 "recognition" ANN. The output of the network is a singular classification of "vertical" mouth movements into one of smile, neutral, and sad categories. Both features networks perform the same task and they are implemented, therefore, as two copies of the same network. In such a case, the error is propagated within the single ANN as well as from the recognition ANN to both of the features ANNs. This speeds up the training process and results in better generalization properties of the network [9].

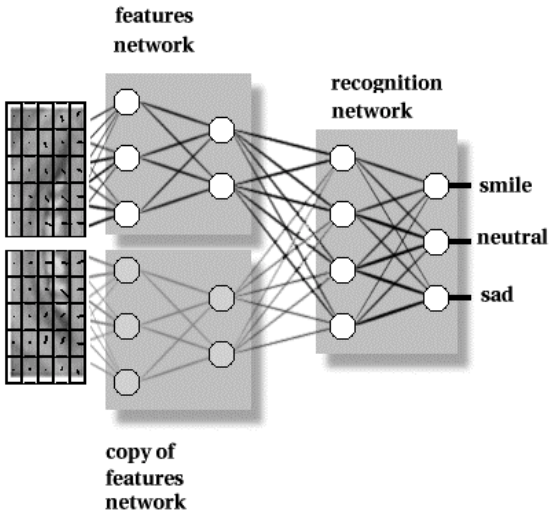

Fig. 11: ANN architecture utilized by the *Vertical ANN Mouth Classifier*

The result of *Vertical ANN Mouth Classifier*, 50 vectors resulting from the fuzzy reasoning about the edge information in the ROI of the current frame, and 50 vectors resulting from the fuzzy reasoning about the edge information in the ROI of the first frame (expressionless mouth), form the input to *Horizontal Rule-based Mouth Classifier*. The procedure employs further four rules such as "if the result of *Vertical ANN Mouth Classifier* is *smile* or *sad*, then the mouth is *stretched*" and "if the result of *Vertical ANN Mouth Classifier* is *neutral* and the vectors found in the first and the last two columns of the decomposed ROI of the current frame are longer than the vectors found in the same columns of the decomposed ROI of the first frame, then the mouth is *stretched*". This results in a singular classification of mouth movements into one of stretched, neutral, and puckered categories.

4 Mouth features extraction

In its last phase, the proposed algorithm extracts the four salient mouth feature points from the mouth contours sampled by the *Curve Fitting of the Mouth* and *Mouth Template Matching* procedures. In the case of the mouth contour sampled by *Curve Fitting of the Mouth*, the top of the upper lip $\boldsymbol{T}$, the bottom of the lower lip $\boldsymbol{B}$, the left mouth corner $\boldsymbol{L}$, and the right mouth corner $\boldsymbol{R}$ are extracted as follows: $\boldsymbol{T}$ = (x, ½(y_{max1} + y_{max2})), $\boldsymbol{B}$ = (x, y_{min}), $\boldsymbol{L}$ = (x_{min}, y), and $\boldsymbol{R}$ = (x_{max}, y). In the case of the mouth contour sampled by *Mouth Template Matching*, the salient mouth feature points are extracted as follows: $\boldsymbol{T}$ = $\boldsymbol{a1}$, $\boldsymbol{B}$ = $\boldsymbol{a2}$, $\boldsymbol{L}$ = $\boldsymbol{v1}$, and $\boldsymbol{R}$ = $\boldsymbol{v2}$.

To compute the data certainty of the determined salient mouth feature points, the proposed algorithm performs an intra-solution consistency check for each of the two applied mouth-contour-spatial-sampling procedures. This consistency check is based on the knowledge about the facial stability of the medial point of the mouth. Namely, independently of the facial muscles' action affecting the facial appearance of the mouth (horizontal action like mouth stretching, up-down action like jaw drop, oblique action like smile, or orbital action like tightening the lips), the (imaginary) medial point $\boldsymbol{M}$ of the mouth, computed as $\boldsymbol{M}$ = center (center ($\boldsymbol{L}$, $\boldsymbol{R}$), center ($\boldsymbol{T}$, $\boldsymbol{B}$)), remains stable (Fig. 12).

Fig. 12: Medial point of the mouth

To each salient mouth feature point $\boldsymbol{P}$ extracted from the mouth contour sampled by *Curve Fitting of the Mouth*, the proposed algorithm assigns an equal data certainty measure $\boldsymbol{CM} \in [0,100]$ according to the calculated deviation of the medial point $\boldsymbol{M_c}$ at issue (computed based on the salient mouth feature points $\boldsymbol{P}$) from the medial point $\boldsymbol{M_n}$ localized in the first frame (expressionless mouth). This mapping is defined as $\boldsymbol{CM}$ = $\boldsymbol{S(x)}$ * 100, where $\boldsymbol{S(x)}$ = $\boldsymbol{sigm}$ ($\boldsymbol{d}(\boldsymbol{M_c}, \boldsymbol{M_n})$; 2, 7, 15), $\boldsymbol{d}(\boldsymbol{p1}, \boldsymbol{p2})$ is the block distance between points $\boldsymbol{p1}$ and $\boldsymbol{p2}$ (maximal difference in x and y direction), and $\boldsymbol{sigm}$ (y; α, β, γ) is the Sigmoid membership function. In the same way the proposed algorithm assigns a data certainty measure to each salient mouth feature point extracted from the mouth contour sampled by *Mouth Template Matching*.

To select the best of the available solutions, the algorithm performs an intra-solution consistency check for both, the salient mouth feature points extracted from the mouth contour sampled by *Curve Fitting of the Mouth* and the salient mouth feature points extracted from the mouth contour sampled by *Mouth Template Matching*. This consistency check is based upon six rules such as "if the result of *Vertical ANN Mouth Classifier* is *smile* and $\boldsymbol{L_y} < \boldsymbol{LN_y}$ and $\boldsymbol{R_y} <$

RN_y, where L and R are currently generated mouth feature points and LN and RN are the pertinent points detected in the first frame, then increase a variable i" and "if the result of *Horizontal Rule-based Mouth Classifier* is *stretched* and $L_x < LN_x$ and $R_x > RN_x$, then increase a variable i". The best solution is the one for which the value of the assigned CM as well as the value of i is the largest.

5 Experimental results

The proposed algorithm has been tested on several typical audiovisual speech databases as well as on several facial image sequences generated in our lab by the mounted camera device shown in Fig. 1.

For the first 100 frames of the CIF 30Hz "Miss America" sequence, the salient mouth feature points have been correctly estimated in 99% of the cases ($CM > 0.7$). For the first 100 frames of the CIF 30Hz "Claire" sequence, 98% correct estimates ($CM > 0.7$) have been accomplished.

For 24 sequences (70 to 100 PAL video frames) generated on our own (six subjects of both sexes, 20 to 45 years old, of 3 different ethnic backgrounds, and recorded wile showing various facial expressions under different lightning conditions), the salient mouth feature points were correctly detected in 96% of the cases ($CM > 0.7$). Typical results obtained during this test are illustrated in Fig. 13.

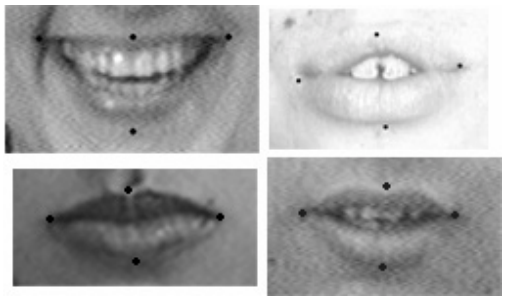

Fig. 13: Mouth feature points found by the algorithm

6 Conclusions

A robust and flexible algorithm for salient mouth feature points detection from facial image sequences is introduced. The originality of this contribution is the hybrid approach that allows integrating different sources of information and concurrently utilizing them to decide the best estimate of the salient mouth feature points. The proposed algorithm outperforms the existing ones by eliminating the drawbacks imposed by employing a single image-processing technique. Furthermore, unlike most of the existing approaches to mouth feature extraction, it does not assume that the presence of the mouth in the input image is ensured. It determines Region Of Interest (i.e. a more or less large area around the mouth) in the input facial image in a fully automatic way. By a high number of experiments we demonstrated the method's robustness with respect to variations in subjects, their poses, and illumination conditions.

Although enhancing the state-of-the-art in mouth feature extraction from facial image sequences, the algorithm does have limitations. It can deal only with limited out-of-plane head rotations and sequences starting with an expressionless mouth appearance.

References

1. R. Feraud, O.J. Bernier, J.E. Viallet and M. Collobert, A fast and accurate face detector based on neural networks, IEEE Trans. PAMI, vol. 23, no. 1, pp. 42-53, 2001.
2. C.L. Huang and Y.M. Huang, Facial expression recognition using model-based feature extraction and action parameters classification, J. Visual Comm. & Image Represent., vol. 8, no. 3, pp. 278-290, 1997.
3. M. Kass, A. Witkin, D. Terzopoulos, Snake: active contour model, Proc. IEEE ICCV, pp. 259-269, 1987.
4. S. Kimura and M. Yachida, Facial expression recognition and its degree estimation, Proc. IEEE CVPR, pp. 295-300, 1997.
5. J. Luettin, N.A. Thacker and S.W. Beet, Visual speech recognition using active shape models and hidden Markov models, Proc. IEEE ICASSP, vol. 2, pp. 817-820, 1996.
6. M. Pantic and L.J.M. Rothkrantz, Expert system for automatic analysis of facial expressions, Image & Vision Compu. J., vol. 18, no. 11, pp. 881-905, 2000.
7. M. Pantic and L.J.M. Rothkrantz, Automatic analysis of facial expression: State of the art, IEEE Trans. PAMI, vol. 22, no. 12, pp. 1424-1445, 2000.
8. D.G. Stork and M.E. Hennecke, Eds, Speech-reading by Man and Machine: Data, Models and Systems, New York: Springer-Verlag, 1996.
9. J. Wojdel and L.J.M. Rothkrantz, Mixed fuzzy-system & ANN approach to automated recognition of mouth expressions, Proc. ICANN, pp. 833-838, 1998.
10. J. Yang and A. Waibel, A real-time face tracker, Proc. Workshop on Applications of Computer Vision, pp. 142-147, 1996.
11. M. Yoneyama, Y. Iwano, A. Ohtake, K. Shirai, Facial expression recognition using discrete Hopfield ANN, Proc. IEEE ICIP, vol. 3, pp. 117-120, 1997.
12. A.L. Yuille, D.S. Cohen and P.W. Hallinan, Feature extraction from faces using deformable templates, Proc. IEEE CVPR, pp. 104-109, 1989.