Gradient Based Approach to Mouth Shape Recognition

J.C. Wojdeł L.J.M. Rothkrantz P.S. Szczepaniak*

Knowledge Based Systems,
Faculty of Information Technology and Systems,
Delft University of Technology,
Zuidplatsoen 4, 2628DB Delft, The Netherlands
J.C.Wojdel@cs.tudelft.nl L.J.M.Rothkrantz@cs.tudelft.nl Piotr@ics.p.lodz.pl

Keywords: Recognition of facial expressions, neural networks, fuzzy systems, pattern recognition

Abstract

At this moment there is an ongoing project at the Delft University of Technology aimed at the development of the Integrated System for Facial Expression Recognition (ISFER). The ISFER project is being developed in the Knowledge Based Systems group at the Faculty of Technical Mathematics and Informatics, TU Delft. In this project the works about image preprocessing, feature extraction and the interpretation of the results are bound in a single system. One of the most important parts of the facial expression recognition is recognizing the mouth expression. This paper describes a new approach to this task. Contrary to the works already done in this field, it is rather knowledge- than graphically-based. We use here the fuzzy-system and the artificial neural network techniques to obtain the desired results.

1 Introduction

The part of the ISFER project which is the basis for further development is the multimedia workbench for the system [1]. In this workbench the modules performing different tasks of the facial expression recognition can be connected together to form the complex system. The facial expression recognition task can then be divided into sub-tasks such as preprocessing, feature extraction and recognition, or interpretation of the results. All those tasks can be performed by different modules written independently by people working in the project. The workbench gives also the opportunity to test different paths of recognition. Decision about which modules should perform which tasks can be done in an interactive manner by the end-user of the system. Some modules of the above system performing tasks such as preprocessing, facial features segmentation or template matching of the facial features are already developed. For example the interpretation part of the system was presented in [2, 3]. Yet the system as a whole is still under development. Such systems for facial expression recognition are being developed not only at TU Delft, but also for example at MIT Media Laboratory [4].

One of the parts of facial expression recognition is detection of exact expression of the mouth region in face. In fact, this region is the most changing in all the face and so should be considered as very important. Apart from the applications of the whole facial expression recognition, the recognition of mouth expression alone may be used to develop an automated lip-reading system. Such a system can then be used in a bimodal speech recognition to improve the effectiveness of the speech recognition [5] or to support just a segmentation of the speech signal in a noisy environment.

2 Problems with finding the mouth contour

Apart from the problems common to any of the image recognition tasks, such as e.g. the presence of noise, the recognition of the mouth shape puts the researcher to some additional difficulties.

Firstly, the mouth has to be located in the image. We can do that by finding the dark area in the image. Of course we may find different regions with this property and each of them might be taken as "mouth location hypothesis". All the hypothesis must then be evaluated, and the proper one must be chosen. The evaluation of hypothesis is based on such properties of found regions as its size, its relative position to other regions (as e.g. eyes) and face symmetry axis. Description of a system performing such a task can be found in [6].

After locating the mouth region one has to rec-

*Institute of Computer Science, Technical University of Łódź, ul. Sterlinga 16/18, 90–217, Łódź, Poland

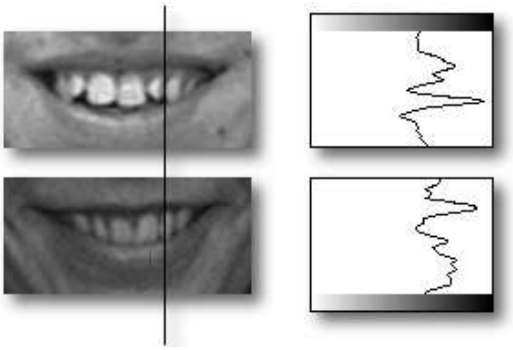

Figure 1: Two examples of mouth pictures. On the right hand side the intensity of the image along the vertical line is shown. Important thing is that there is no visible change in image intensity on the lower lip contour.

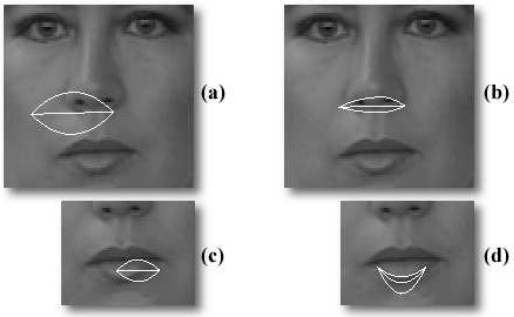

Figure 2: Two examples of improper template matching caused by bad initial parameters: (a) bad template position and (b) resulting final state of template, (c) too small initial size and (d) template after matching [6].

ognize correctly its shape features. So far most of the works in recognizing mouth shape concerns finding the mouth contour [7, 6]. As for the first sight it may look like a simple edge-detection problem, in a real life it isn't so. The first obstacle on the way to find mouth contour is that for most cases (except for women using lipstick) this contour is not really visible. It may be well seen on the following pictures (Fig. 1).

In fact, it is also hard to define correctness of contour which was found. It seems that the only criterion that can be used is whether effect corresponds to human perception of the mouth image. That makes measuring effectiveness of contour-finding methods a very difficult task.

Another problem arises when we consider video sequences some time-based representation of the natural process. It is obvious that this process is continuous with respect to the time. We should therefore demand from the contour found in the image sequence to change continuously too. The problem is whether contour representation can be easily interpolated between frames or not. For some methods that use parametric models (as template matching [6]) this can be achieved by interpolating parameters. For other ones (as "snakes" with varying number of points [7]) such interpolation isn't that easy and should be done very carefully. Even when the above is achieved, there is still the problem of the robustness of the method. For time-based cases such as lip-reading or image-sequence coding we must finally obtain real-time processing. This makes methods which are non-iterative or with *potential parallelism* more appropriate for this task.

3 Previous approaches

In the recent works the approaches such as deformable templates or active contours were used to obtain the mouth contour from the image of the mouth. Both of these methods are iterative and based on minimization of specially defined energy function.

The *snakes* optimize the closed polygon so that in the end of the iteration process its points lay on the edge of the mouth contour. This approach is in fact quite robust, because the energy function is based on very local properties of the image and therefore computationally inexpensive [7]. The drawback of this method is that there are no additional constraints on the shape of the resulting contour. It can even shrink to a single point in an unfortunate case.

The templates are based on more sophisticated approach to the problem. In this case the shape of the contour is given by the template itself, so the found mouth contour is always in acceptable shape. Yet it may happen that because of the bad initial parameters of the template it doesn't really represent the mouth (Fig. 2). The energy function for the templates is based more on global features of the image than in the case of *snakes*. It is then also more complicated and more computationally expensive. Still it seems that the obtained results are better in comparison to the *snakes*.

Still both approaches presented above are based on the mouth contour as the way for representing the mouth state. We may see that a human observer (the best so far system for facial expression recognition) does not use such a strict description of mouth shape. Even from the relatively big distance when exact tracing of the mouth contour is impossible we can see whether the observed person is smiling or not. We can be thus almost sure that the type of the information used in such recognition process is of the more global nature. We would like to present in the next section such a type of a more intuitive mouth shape representation.

4 A new representation of the mouth shape

The first thing that has to be understood is the nature of the data that is processed in mouth feature recognition. As it was previously stated, the process that generates mouth image-sequence is continuous and smooth in time. Every recognition system can be described as a function which for a given state of the human face gives an appropriate response concerning the expression shown by this face. It is obvious that such a function should be continuous and smooth. The reason for this is that small variations in facial states should result in small variations of the system responses if those responses are to be meaningful.

Yet the system never gets directly a state of the face as an input. We can present some steps through which information about the face must go before it becomes input for the system (Fig. 3).

As it can be seen the system should be able to construct a smooth and continuous function on the basis of data that is neither smooth nor continuous. Another problem is the size of the input data for the system. Of course, the less this size is the more robust and the easier to develop the recognition system is. Therefore we have to find some way to smooth information derived from a sampled image in respect to both time and space of facial-states, and to reduce its dimension while maintaining useful information.

The easiest way to smooth data and reduce its size is averaging image information. In an exaggerated example, an average of all pixels from an image is changing smoothly along time and facial-state space. Of course such a global averaging makes all significant data disappear and therefore is not useful at all. Yet it gives us a clue to a proper solution of the problem: we should average something in order to achieve smoothness and data size reduction. Still, we should also find in the image some information that is both significant and well approximable so it will not be completely lost after averaging.

It seems that the edge information in the image carries most of the information, so the intuitive term of edge must be transferred into some mathematical substitute. This substitute is the gradient of intensity of the image. When this gradient achieves some significantly big value we intuitively label it as an edge. The question that we have to answer is whether this can be well approximated with some average values taken from an image and if it holds enough information for proper recognition of a mouth expression.

All the information on a mouth expression can be shown using only a single drawing line and it still remains perfectly well recognizable for all the people. How does it come ? After a closer look we can observe that edge information (gradient) in a simple drawing is just an exaggerated approximation of that

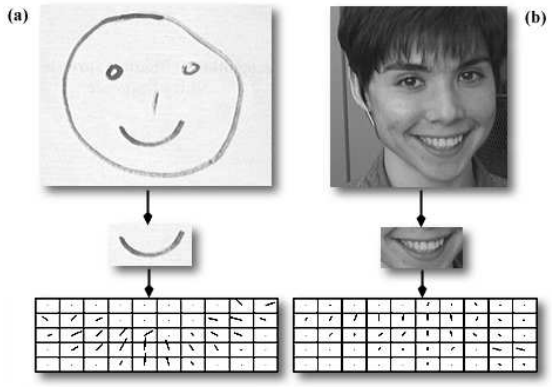

Figure 4: Simple drawing (a) and the real picture (b). Mouth region was cut off, then the gradient was drawn. It can be seen that the direction of the gradient changes in the similar way while its value is simply exaggerated on the drawing.

derived from real-life photo (Fig. 4). Even that rough approximation of the gradient information is enough to proper recognition of mouth expression. It seems therefore that some automatic approximation of that will be enough for our task.

We will now take a closer look at the gradient distribution in the image of the mouth region. First, the edge of the mouth contour in this image is nearly horizontal in all points. That means that there are no big and rapid changes of the gradient in most parts of the image. That is very good for approximating with average values. The next thing we can see is that for example, the information about whether a mouth is 'smiling' or 'sad' is passed through the average edge direction in the corners of the mouth. If the edge is on average 'going up' mouth is interpreted as 'smiling' if it 'goes down' mouth is interpreted as 'sad'. So, it seems that, after all information about average edge direction is more important than information about an edge in some specific point.

5 Fuzzy-system for edge information acquisition

We present now a new tool for obtaining information about the average edge direction in a mouth image. The basic idea of this tool comes from [8]. In this article a system based on fuzzy-reasoning for edge detection was proposed.

The proposed system performs reasoning based on two linguistic labels:

- local level of symmetry in the image in a specific point
- local image steepness in the direction of the symmetry axis

which correspond to two features of the gradient:

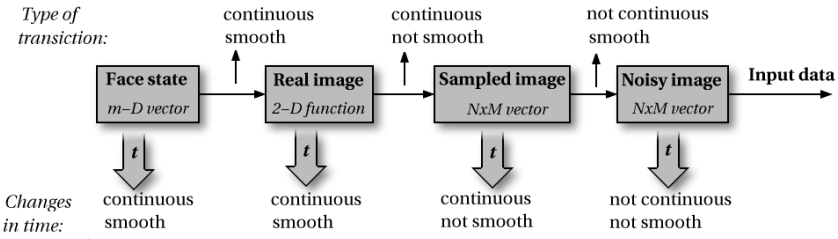

Figure 3: Data flow from the state of the face to the recognition system.

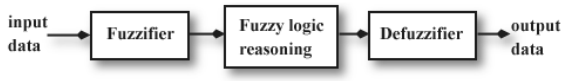

Figure 5: Standard fuzzy-system architecture and an example flow of data.

- the function is locally symmetrical along the gradient direction
- the gradient value corresponds to local steepness of the function

Based on those two features we can perform reasoning about the edges in the image (e.g. if the symmetry level and steepness are high then the gradient is high and can be accounted for an edge). A difference in our approach in comparison to the referred article [8] is that our main information is rather the direction of the gradient than its value (contrary to normal edge detection problem).

As a typical fuzzy-system it consists of three distinct parts: *fuzzifier*, *reasoning part* and *defuzzifier* (Fig. 5). In the *fuzzifier* part the fuzzy-labels must be derived from the incoming data. In our case they are symmetry level and steepness level. The *reasoning part* of the system performs reasoning on base of the statements passed from the *fuzzifier* and some set of rules. Then in the end the *defuzzifier* produces the numerical output from the results of the reasoning process.

To obtain the symmetry level in each point of the image, we check whether the image is symmetrical with respect to each of some different directions. For each direction an average difference between intensities of the symmetrically placed pixels is calculated. The direction with the minimal difference is assumed to be the symmetry axis in the specific point. The corresponding average difference subtracted from 1 gives us the symmetry level in that point.

The steepness of the image intensity is calculated as a difference between the average pixel intensities on three lines perpendicular to the symmetry axis.

After obtaining those numerical values, the *fuzzifier* converts them into six linguistic labels: SYMMETRY_HIGH, SYMMETRY_MEDIUM, SYMMETRY_LOW, STEEPNESS_HIGH, STEEPNESS_MEDIUM, STEEPNESS_LOW.
To obtain the above labels we utilize a typical approach of filtering the input values with the trapezoid filter functions. The parameters of those functions were adjusted during the experiments. Those six statements are then passed to the *reasoning part* of the system.

The *reasoning part* produces three output statements about edge intensity in a given point: EDGE_HIGH, EDGE_MEDIUM, EDGE_LOW. This part of the fuzzy-system is based on nine rules. Those rules correspond to our assumptions about the edges (e.g. the edge is where the symmetry level and steepness of the image intensity are both high). Therefore for example the following rules are part of the reasoning engine:

$$\begin{array}{lcll} \text{SYMMETRY_LOW} & \wedge & \text{STEEPNESS_LOW} & \Rightarrow \text{EDGE_LOW} \\ \text{SYMMETRY_LOW} & \wedge & \text{STEEPNESS_MEDIUM} & \Rightarrow \text{EDGE_LOW} \\ \text{SYMMETRY_HIGH} & \wedge & \text{STEEPNESS_HIGH} & \Rightarrow \text{EDGE_HIGH} \end{array}$$

The greatest advantage of using the fuzzy system is that we can freely use common sense to design the reasoning rules and then use them in a fuzzy system to obtain the tool which deals with the numerical values.

The three output statements must then be converted by the *defuzzifier* into some numerical output. It is simply done by performing the following calculation:

$$e = \frac{1.0\text{EDGE_HIGH}+0.5\text{EDGE_MED}+0.0\text{EDGE_LOW}}{\text{EDGE_HIGH}+\text{EDGE_MED}+\text{EDGE_LOW}}$$

To obtain the information about both: average edge intensity and direction (which is the basis of the presented method), we have to use the information about the direction of the symmetry axis which was computed in an intermediate step of the processing. The vectors representing that direction are used to obtain the vector of the average edge direction. The length of this vector is then set to the value of the average edge intensity. As our approach is based on analyzing the average properties of an edge vector-field, there is no need to trace edges, joint points or corners which must be done in a classical edge-detection task. There is also no need to distinguish between the edges of the mouth contour and other edges which may appear in the image because most of the *fake* edges are local enough to be lost in the process of averaging.

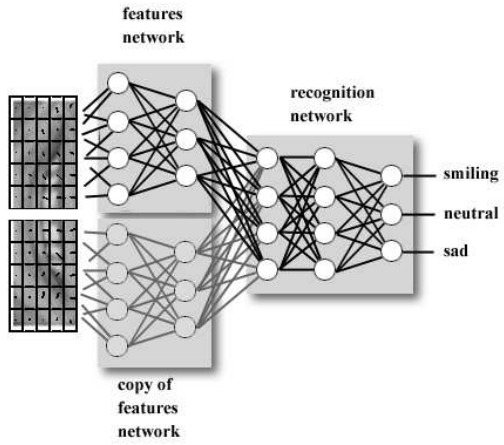

Figure 6: The architecture of the implemented ANN.

6 Neural network for mouth expression classification

This part of the system is just an implementation of a simple feed-forward, back-propagation trained artificial neural network (ANN). Only the network architecture is a little changed in order to reflect the specific features of the task.

This architecture is shown in Fig. 6. One of the important things is that both features networks should perform the same tasks and so they can be implemented as two copies of the same network. The reason for this is that although the two sides of the mouth act independently but with the same possible movements.

The network has three outputs which are trained so that only one, representing a correct class of mouth expressions has a value close to 1, while two other should have values close to 0. Of course it may happen that even after training the network response for some pictures will not be that clear (with two outputs set to 0 and one to 1) and therefore the output with the biggest value is assumed to be the result of the classification.

7 Experimental results

To test the presented approach we first selected the training/testing set of about 100 images. At first from the larger set of frontal head-shoulder images we selected those which could be easily labelled (by a human) as *smiling*, *neutral* or *sad*. We used images from various sources such as CCD video camera, scanned photos and images of unknown origin found on the Web. To make this set representative we chosed no more than 5 images of the same person. Later the mouth area of the image was manually cut from those images. So that in the end we go set of greyscale images each with size about 200 per 100 pixels.

Four different neural network architectures were tested for learning speed and accuracy. We will present two of them here. At first, quite a big architecture was chosen to check whether the task is solvable by a BP, feed-forward network. This network architecture was:

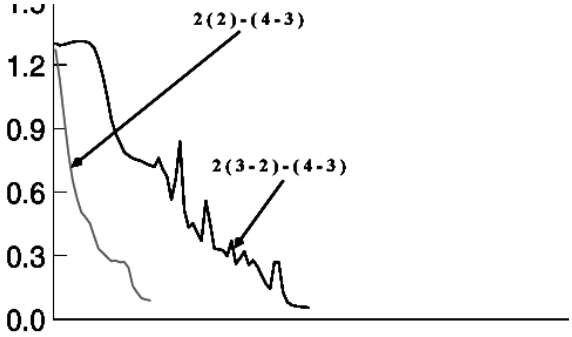

Figure 7: Average error in epoch during training process for two network architectures.

	(2-3)-(4-3)	(2)-(4-3)
Training time (epochs)	60	20
Recognition of:		
smiling mouths	100%	100%
neutral mouths	100%	100%
sad mouths	100%	100%
average recognition	100%	100%
Average response error:		
smiling samples	0.05	0.09
neutral samples	0.17	0.23
sad samples	0.02	0.02
overall	0.08	0.11

Table 1: Comparison of different network architectures.

(1)	Features network		Recognition net.	
	1st layer	2nd layer	1st layer	2nd layer
neurons	3	2	4	3

The above means that each copy of the features network had two layers, and respectively three and two neurons in each layer. The recognition network had also two layers with four neurons in the first layer, and three neurons in the second one.

This network achieved full 100% recognition level for both training and the test sets. The training took about 60 epochs. Changes of the average error of the network response during the training process can be seen in the Fig. 7 An average response error of this network on the testing set was 0.08.

The next network which was tested had only one layer of neurons in the features network:

(2)	Features network		Recognition net.	
	1st layer	2nd layer	1st layer	2nd layer
neurons	2	–	4	3

The results obtained with this one were comparable, but the training took only about 20 epochs. The progress of the training is also displayed in Fig. 7.

noise level	network response for smiling mouth		network response for neutral mouth		network response for sad mouth	
0	smiling	0.9810	smiling	0.0032	smiling	0.0050
	neutral	0.0097	neutral	0.9749	neutral	0.0036
	sad	0.0322	sad	0.0138	sad	0.9855
10	smiling	0.9815	smiling	0.1203	smiling	0.0159
	neutral	0.0091	neutral	0.5913	neutral	0.0098
	sad	0.0333	sad	0.0242	sad	0.9479
20	smiling	0.9820	smiling	0.0152	smiling	0.0221
	neutral	0.0091	neutral	0.0111	neutral	0.0036
	sad	0.0328	sad	0.9437	sad	0.9689
30	smiling	0.7573	smiling	0.0423	smiling	0.0302
	neutral	0.0016	neutral	0.0099	neutral	0.0024
	sad	0.4724	sad	0.9484	sad	0.9691
40	smiling	0.3202	smiling	0.0312	smiling	0.0223
	neutral	0.0015	neutral	0.0022	neutral	0.0034
	sad	0.8137	sad	0.9699	sad	0.9696

Table 2: The outputs of the network for noisy images.

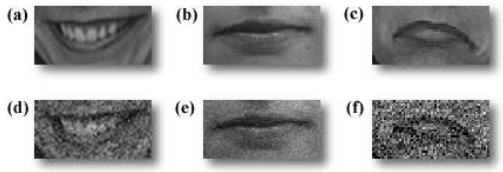

Figure 8: Noiseless images (a), (b), (c), and noisy images which were correctly recognized: (d) noise level 30, (e) noise level 10, (f) noise level 40.

Another thing which was tested was the sensitivity of the system to the noise. On three representative images the Gaussian noise was applied with ascending noise level. The level of the noise is expressed in units used by Adobe Photoshop 3.0. The Tab. 2 presents results obtained from the network (2). Above results seem to be very good; the presented method is not sensitive for the noise in a wide range.

8 Conclusions

The proposed method proved to be valid at least in some basic cases. As it was derived on the basis of the nature of a video signal rather than special properties of the mouth image, it can be easily applied to other face regions. In such a case only some special considerations about the nature of the gradient field in a specified region must be made in order to construct an appropriate tool for the shape classification.

Thanks to the small size of the data used for classification this method is robust and can be easily used in real-time applications. It seems to be very promising for all time-based applications.

It was not yet proved whether the proposed method is sufficiently sensitive. Some broader research should be done to check whether the method is sensitive enough to allow for example the recognition of the Action Units from the FACS set [9]. The method uses only some average properties of the image and therefore it may destroy subtle differences between mouth expressions which are crucial for proper recognition of all the AUs.

References

[1] L. J. M. Rothkrantz, M. R. van Schouwen, F. Ververs, and J. C. M. Vollering. A multimedia workbench for facial expression analysis. In *Proceedings of Euromedia-'98* [10].

[2] M. Pantic and L. J. M. Rothkrantz. Automatic recognition of facial expressions and human emotions. In *Proceedings of the 3rd annual conference of the Advanced School for Computing and Imaging*, Heijen, The Netherlands, 1997.

[3] M. Pantic, L. J. M. Rothkrantz, and H. Koppelaar. Automation of non-verbal communication of facial expressions. In *Proceedings of Euromedia-'98* [10].

[4] I. A. Essa and A. P. Pentland. Facial expression recognition using a dynamic model and motion energy. In *Proceedings of the Fifth International Conference on Computer Vision*, Cambridge, Massachusetts, 1995. IEEE Computer Society Press.

[5] S. Nakamura, R. Nagai, and K. Shikano. Improved bimodal speech recognition using tied–mixture HMMs and 5000 word audio–visual synchronous database. In *Proceedings of ECSA, Eurospeech97*, Rhodes, Greece, 1997.

[6] M. J. T. Reinders. *Model Adaptation for Image Coding*. PhD thesis, Delft University of Technology, Delft, The Netherlands, Dec. 1995.

[7] P. Juran and P. Krsek. Mouth segmentation using active contours. Isfer report, Delft University of Technology, Delft, The Netherlands, Oct. 1996.

[8] T. Law, H. Itoh, and H. Seki. Image filtering, edge detection and edge tracing using fuzzy reasoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(5), 1994.

[9] P. Ekman and W. V. Friesen. *Facial Action Coding System (FACS): A Technique for the Measurement of Facial Action*. Consulting Psychologists Press, Palo Alto, CA, USA, 1978.

[10] *Proceedings of Euromedia-'98*, Leicester, United Kingdom, Jan. 1998.