Mixed Fuzzy-system and Artificial Neural Network Approach to the Automated Recognition of Mouth Expressions.

J.C. Wojdeł L.J.M. Rothkrantz

Knowledge Based Systems,

Faculty of Information Technology and Systems,

Delft University of Technology

Delft, The Netherlands

(J.C.Wojdel, L.J.M.Rothkrantz)@cs.tudelft.nl

Abstract

One of the most important parts of the automatic recognition of facial expression is recognition of the mouth expression. This paper describes a new approach to this task. Contrary to the works already done in this fields, it is knowledge- rather than graphically- based. We use here a fuzzy-system in combination with an artificial neural network to obtain the recognition of the mouth shape.

1 Introduction

At this moment there is an ongoing project in the Knowledge Based Systems group at the Delft University of Technology, which is aimed at the development of the Integrated System for Facial Expression Recognition (ISFER). The multimedia workbench [1] is currently the basis for the development of the system. Some modules of the ISFER system are already developed. For example the interpretation part of the system was presented in [2]. Such systems for facial expression recognition are being developed not only at TU Delft, but also for example at MIT Media Laboratory [3].

One of the parts of facial expression recognition is the detection of the exact expression of the mouth region in the face. In fact, this region is the most changing part of the face and so should be considered as a very important one. Apart from the applications of the whole facial expression recognition, the recognition of the 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 application to improve the effectiveness of the speech recognition [4] 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. Each of the regions matching some properties be taken as *mouth*

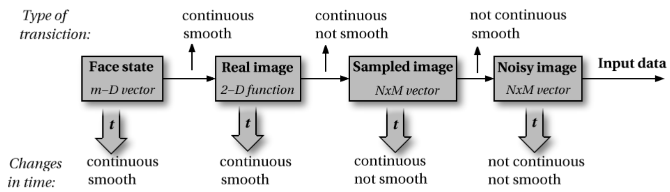

Figure 1: Data flow from the state of the face to the recognition system.

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 [5].

After locating the mouth region one has to recognize correctly its shape features. So far most of the works in recognizing mouth shape concerns finding the mouth contour [6, 5]. Unfortunately, there is no significant change in the image intensity at the contour of the mouth (except for women wearing lipstick). In fact, it is also hard to define correctness of the contour which was found. It seems that the only criterion that can be used is whether effect corresponds to a human perception of the mouth image.

3 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. The process that generates mouth image-sequence is continuous and smooth in time. Yet the system never gets directly a state of the face as an input (Figure 1). As it can be seen the system should be able to construct a smooth and continuous response on the basis of data that is neither smooth nor continuous.

The information from 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 the edge information (gradient) in a simple drawing is just an exaggerated approximation of that derived from real-life photo. 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. The above considerations led us to the conclusion that the appropriate representation of the mouth shape may be the information about the average edge intensity and direction in an image.

4 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 [7]. 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:

- 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 [7] is that our main information is rather the direction of the gradient than its value (contrary to normal edge detection problem).

The numerical values representing symmetry and steepness level are converted by a *fuzzifier* part of the fuzzy-system into the labels. Those labels describe whether each of the values is *low*, *medium* or *high*. Such statements are then passed to the *reasoning engine*. This part of the fuzzy-system is based on nine rules. Those rules correspond to our assumptions about the edges (e.g. if the steepness is *high* and the symmetry level is *high* then the edge intensity in this point is *high*). The reasoning part produces then also three statements (*low*, *medium* and *high*) about the edge intensity in a given point .

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.

To obtain the information about both: edge intensity and direction , we use the information about the direction of the symmetry axis which was computed in an intermediate step of the processing. Combining the information about the direction and the intensity of the edge in a given point we obtain a vector. This vector-field is then averaged and only 50 average vectors are the output of this stage of image processing.

5 Classification of mouth expression by ANN

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 bit changed in order to reflect the specific features of the task.

The thing which is reflected in the network layout is that the mouth has the vertical symmetry axis. Of course it may happen that the left-hand side of the mouth is different than the right-hand one, as both sides may act almost independently. The symmetry means in this case that both sides can perform the adequate actions. There is no possible action of the one side of the mouth which couldn't be performed by the other one. That means that there are some shape features of the mouth which are the same for both of the sides. The network which was implemented in this work is therefore divided into two parts:

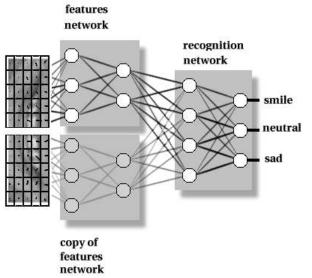

Figure 2: The architecture of the implemented ANN.

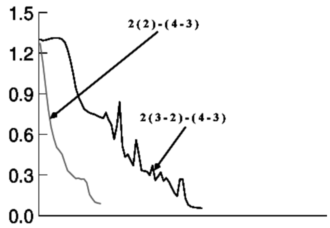

Figure 3: Average error in epoch during training process for two network architectures.

- shape features extraction part, which is built of two symmetrically responding networks (features networks),
- reasoning part which on the basis of outputs from both of the features networks produces some network output.

This architecture is shown in Figure 2. 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. Of course then the error must be propagated not only within the single network but also from the recognition network to both of the features networks. It can be seen that in this case, for each training sample the features network is updated twice – once for the left side and once for the right side. Such a construction of the network should reflect in both: the speed-up of the training process, and better generalization properties

6 Experimental results on single pictures

To evaluate the method we chose a set of about 100 pictures of the mouth area which were manually classified to one of the three categories: *smiling*, *neutral* and *sad*. In each experiment, the 10 examples were randomly chosen as a test set and the network was trained on the remaining 90 samples. We show here the results obtained with two different network architectures. Both architectures are shown in the Table 1.

In the first case, 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 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 Figure 3. An

Number of	Features network		Recognition network	
neurons in	1st layer	2nd layer	1st layer	2nd layer
Network 1	3	2	4	3
Network 2	2	–	4	3

Table 1: Network architectures

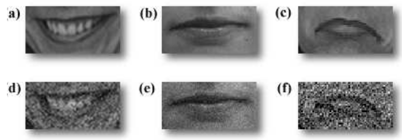

Figure 4: Noiseless images (a), (b), (c), and noisy images which were correctly recognized (d), (e), (f).

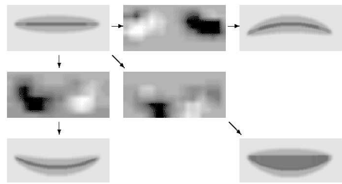

Figure 5: The image sequences used and the visualized output of the fuzzy-system.

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. The results obtained with this one were similar (the average response error 0.11), but the training took only about 20 epochs. The progress of the training is also displayed in Figure 3.

Another thing which was tested, was the sensitivity of the system to noise. On some representative images Gaussian noise was added with ascending noise level. The achieved results are at least satisfactory (see Figure 4); the presented method is not sensitive for the noise in a sufficiently wide range.

7 Experimental results on artificially made image sequences

The method described in the previous sections is aimed at the processing of a sequence of images. Therefore we decided to visualize the changes in the output of the fuzzy-system in some example sequences. For the beginning we chose the artificially made sequences of the simplified mouth actions. Three sequences of the mouth changing from *neutral* to *smile*, *open smile* and *sad* (Figure 5) were investigated.

The areas in which the edges are rotating clockwise are depicted in dark grey contrary to the areas with the edges rotating counter-clockwise which are bright. The areas with no significant change of the edge direction are in neutral grey color. For this visualization we took the average change of the fuzzy-system output in three following frames of the sequence. The patterns obtained from such a visualization can be seen in Figure 5. In Figure 5 we can see also the *neutral mouth* image which was the beginning of the sequence and three images to which this first one was morphed.

It can be well seen that the obtained patterns are very easily recognizable. The sequences with the same “emotional contents” (*smile* and *open smile*) give also similar patterns with bright areas in the right-hand, bottom part of the image and dark areas in the left-hand bottom part of it. This may help the neural network which will classify those patterns to generalize the output.

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.

Both steps (fuzzy-system and ANN) of the presented method can be well parallelized. The first one because it repeatedly and independently uses the same fuzzy-system for all the pixels of the image. The second because the ANN consists of independently working neurons and even in this case full networks. This property of the proposed method may be crucial in a real-time applications such as e.g. lip-reading. Also the small size of the data passed from the fuzzy-system to the neural network results in robustness of this approach.

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 [8]. 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*.

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

[3] 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.

[4] 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.

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

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

[7] 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.

[8] 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.