Dual-view Recognition of Emotional Facial Expressions

A. Wojdel J.C. Wojdel L.J.M. Rothkrantz

Delft University of Technology
Knowledge Based Systems Group
Zuidplantsoen 4, 2628 BZ Delft, the Netherlands
A.Wojdel@cs.tudelft.nl J.C.Wojdel@cs.tudelft.nl L.J.M.Rothkrantz@cs.tudelft.nl

Keywords: Image Interpretation, Multi–modal Data Analysis, Neural Networks, Recognition of Facial Expressions

Abstract

Non-verbal communication plays an important role in human communication. At the Delft University of Technology there is a project running on the automatic recognition of facial expressions. The developed system ISFER (Integrated System for Facial Expression Recognition) is based on the analysis of a frontal view of the face. In this paper we investigate the possible advantages of using an additional lateral view of the face. The experiments done with use of neural networks show the expected improvements when using the dual view approach. In this paper we describe the setup of the experiments and report the results.

1 Introduction

Most of the human–to–human communication is based on non–verbal information. People communicate with body movements, timbre of the speech etc. But the most valuable source of information is the human face. From all the signals that can be send by the facial expression the emotions are the most dominant. In the work of P. Ekman [1] six basic emotions are described from which he claims that they are universal.

On the other hand the facial expression may signal also other things such as e.g. punctuation marks and accents in spoken information. In order to describe all of those signals the Facial Action Coding System was developed (FACS) [2]. With use of Action Units (AUs) from this system any facial expression that gives a visible change on the human face can be described to some degree of accuracy.

The need to automate the facial expression recognition has several reasons. At first the automated systems for facial expression (or emotion) recognition may be used in the environments where the emotional condition of observed person has to be assessed [3]. For example in course of training the people that may face critical situations in their work (e.g. pilots, police officers etc.). The other application of facial expression recognition is highly efficient image coding (so called *model based coding*) used for transmitting the video–conferencing data through low bandwidth channels [4]. In this case at the sender side the image of the speaking person is divided into static description on how the person looks like and the traced changes in the facial expression. Then the full image of the face is transmitted only once in the beginning of the connection and the rest of the time only the updates of the expression are sent. On the receiver side the full motion picture is reconstructed from the static facial description and dynamically changing facial expressions. Finally the facial expression recognition can be used in designing human–computer interfaces that are more friendly and more intuitive for the user [5, 6].

At this moment the research in this field is done all over the world at many very well–known universities. At MIT the system for facial expression recognition was developed on the basis of tracking the energy of visual changes in the facial pictures [7]. Also the first prototypes of the intuitive human–computer interfaces (smart rooms) were constructed there [8]. As it comes to the *emotional computer* that not only registers the emotions of the human operator but also reacts to them in a proper manner, the most successful project is probably the Neuro–Baby'95 from the University of Kyoto [9].

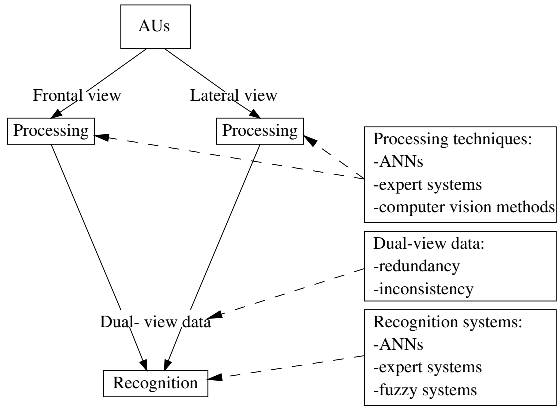

Figure 1: The dual–view recognition model

Since 1992 there is an ongoing research in Knowledge Based Systems Group on Technical University of Delft that is aimed at automated facial expression recognition and analysis. During this time several working prototypes for facial feature location, preprocessing facial expression recognition and analysis were built. All of this work was then gathered together in Integrated System for Facial Expression Recognition (ISFER) [10, 11].

At this moment all of the presented projects focus on the frontal view of the face. The reason for this is quite obvious; also in the human–to–human case, the communication is done usually when two persons sit in front of each other. This is therefore quite intuitive to design also the automatic systems on basis of frontal models. The lateral view of the face was already used for person identification [12]. In this paper we'll show that it can also be used for facial expression recognition especially in combination with the frontal one.

2 Model

At first we should describe what in fact is the dual–view recognition. In some way there is no qualitative difference between recognizing the facial expressions from frontal and lateral views. The space of AUs is used as a universal space to represent facial expressions. So, as it is depicted in the Fig. 1 the underlying process is the same and can be described with AUs that are not dependent on the way in which we project the face. Further, both views can be preprocessed in a similar way and using equivalent techniques. The difference is in the way in which the facial expression is coded in the picture. In the lateral view, half of the face is not visible, and so a lot of information is lost. But also in the frontal view of the face some features are missing. For example the forward–backward movement of the lips or chin is not directly visible. So the information in both of the cases is not complete but also not fully complementary to each other.

The recognition system that has to take an advantage of using both modalities of information can be built with different techniques. The possibilities here are e.g. Artificial Neural Networks (ANNs), expert systems, fuzzy systems, Bayesian networks and many combinations of them. But independently of chosen architecture they'll have to handle the same issues as redundancy of incoming data, contradictions in data coming from different modalities but referring to the same facial feature etc. This complexity of the data can be utilized by the recognition system in many ways. The first possibility is to use the redundancy to reduce the influence of the noise that comes with the data. The other is to replace the data that is missing in one stream with the results derived from the second one. Finally the diagnosis of the problems in the preprocessing stage can be concluded from the contradictions in the data streams. As it is well known the ANN–approach is a black–box approach. The system can be trained by just showing the labeled pictures. Any expert system requires expert knowledge about rules describing the relations between facial features and facial expressions.

In this paper our main goal is to compare the behavior of the automated systems for facial expression recognition in three cases:

- when the system is presented with frontal facial data
- when the system is presented with lateral facial data
- when the system combines both modalities of the information

To do this we decided to use ANNs as recognition technique. The choice is motivated by several factors. Firstly, there is a lot of powerful tools available for simulation of neural networks. We've chosen the Stuttgart Neural Network Simulator (SNNS) in version 4.1. The SNNS system gives a lot of possibilities in constructing the neural networks and in the same time provides very flexible output that can be further processed and analyzed.

The second reason to choose the ANN approach was that at this point we didn't know exactly how to combine those two data streams. There is a lot of issues that one has to cope with when combining two different data streams coming from the same information source. Such things as redundancy in data, reliability of different measurements of the same phenomenon etc. have to be taken into consideration. The ANNs with their "black–box" approach did this hard job

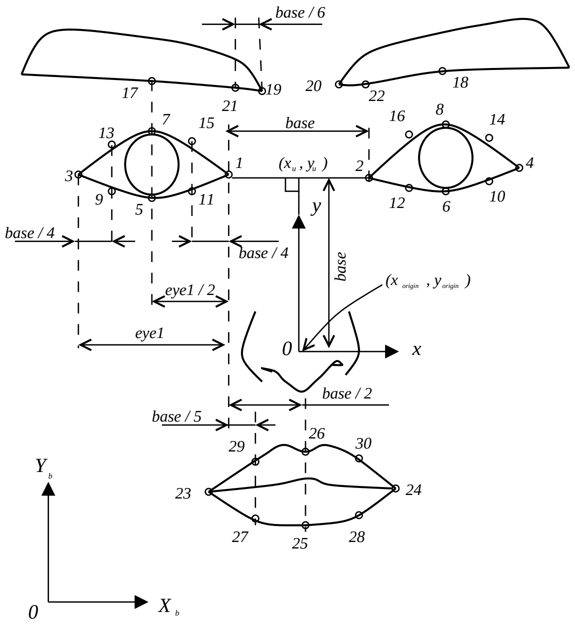

Figure 2: The model of facial characteristic points proposed by Kobayashi

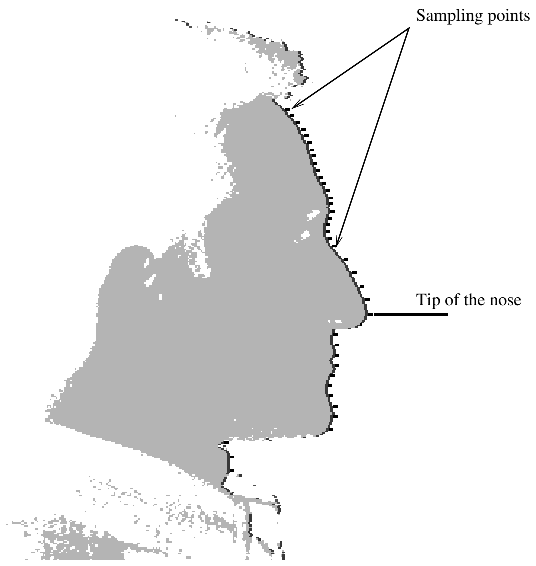

Figure 3: The model used for lateral view of the face

for us.

Before starting the experiments with ANNs we had to decide what kind of data should be given as input to the network. For the frontal part the choice was relatively easy, as the work in this field is already fairly advanced. We chose here for the model proposed by Kobayashi and Hara [13]. As depicted on Fig. 2 the model consists of 30 points placed in characteristic regions of the face. From those points the 21 specific distances are calculated and those distances form the input of the neural network. It is known that using those derived distances instead of raw positions of the 30 points gives usually higher recognition rate [13]. As our goal is to compare the results of dual–view recognition system with frontal–view systems, choosing the best one here seems reasonable.

Unfortunately there was not that much research in the field of lateral view of the face so far so we had to define our own model for this part of the experiment. We decided here to choose the holistic approach. Instead of defining the complicated model (which could prove to be inefficient or not complete enough) we decided to sample the profile line of the face in 40 points and to provide this data to the neural network. Choosing only the silhouette of the lateral view of the face allowed us to acquire this data completely automatically. The image was firstly thresholded in order to segment the face and the background. As a next step we scanned the binarized image horizontally in order to obtain the profile line. As the last step, the tip of the nose (the furthest to the right point of the profile line) was chosen as the reference point and the profile line was sampled in both directions from it (see Fig. 3).

Later in our research on lateral view of the face we proposed also a different model, which is described in the paper submitted to this conference [14].

3 Data acquisition

All pictures were taken with a CCD PAL video camera in front of the face of the respondents. The respondents were placed in front of a black background. Next to the camera a stimulus picture was presented with one of the six emotional facial expressions. The respondents were requested to imitate the facial expression. On a computer screen next to the camera they got a visual feedback.

Next to the head of the respondents there was a mirror positioned in such a way that both the frontal and lateral perspective were taken in one picture (see Fig. 4). We can see that in such an environment the optical distance between the camera and frontal or lateral view of the face is different. This required some preprocessing of the captured picture to compensate the differences in the two views. We performed scaling of the lateral part of the picture (Fig. 5). Although currently the scaling has to be done manually, the single calibration process has to be done only once per session.

In this environment we asked 8 respondents to present each of the six emotions and neutral face

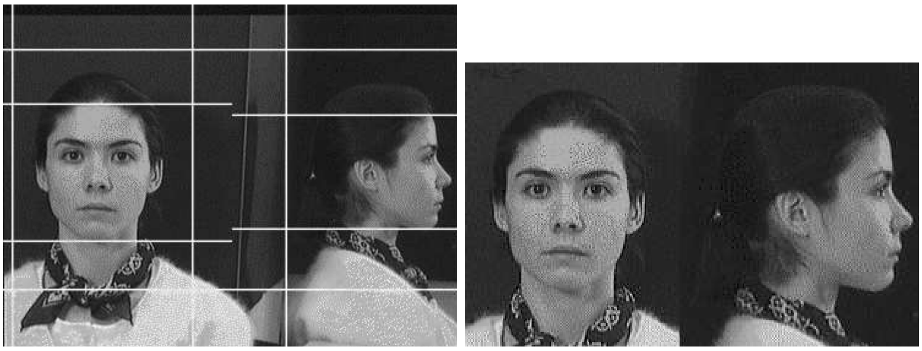

Figure 5: The calibration of the acquisition system and the resulting pictures

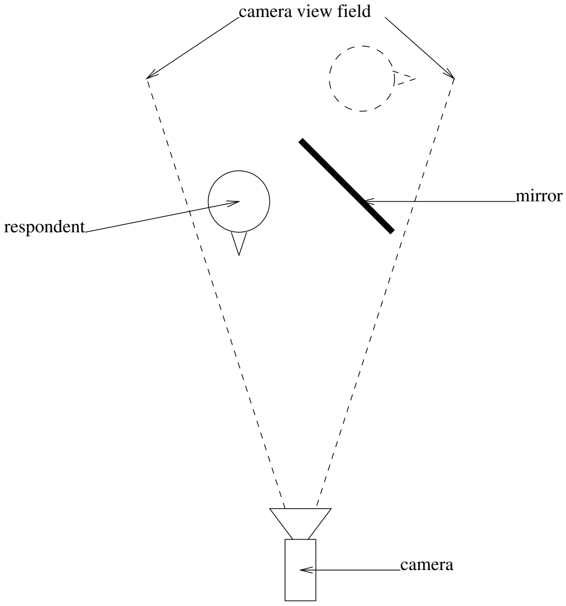

Figure 4: Set-up of dual-view image acquisition

couple of times (usually between 3 and 5). The pictures were then inspected by us and we chose a set of 140 pictures that were satisfying both in terms of technical correctness and the validity of emotional content.

As a result we obtained a dataset with pictures that satisfied the following conditions:

- The image shows only head and shoulders
- The background is solid and black
- The respondents have no glasses, beard or moustache
- No parts of the head are hidden by obstacles
- The respondent is looking straight into the camera (no rotation of the head)
- The emotion shown in the picture can be undoubtedly recognized even if the emotion isn't at full extent

In order to use this set to train the neural network we assessed the emotional content of the pictures. We used here a percentage scoring of the emotion expressed in the picture. From our independent assessments we did averaging to obtain the final percentage of the shown emotion for each of the pictures.

As we stated in Section 2, we used the model of Kobayashi for frontal view of the face. In order to obtain the measurements of the points from this model we used a semi–automated tool that supports setting the points according to constraints of the model. Still because of the lack of reliable and fully automated system for tracing those points, we had to do the measurements manually.

The data for lateral view recognition was obtained automatically using the method described in the previous section. In this case we profited from choosing a homogeneous black background; the threshold level could be set once and for almost all of the pictures the head was segmented properly. The profile detection and sampling of the silhouette was done in almost real–time.

The data obtained from the manual measurements and automatic profile detection was then saved in three different forms: *frontal only* (21 dimensional vectors), *lateral only* (40 dimensional vectors) and *combined* (61 dimensional vectors). After that we added the previously obtained labeling of expressed emotions to those samples. To prepare the data for training/testing of the neural networks we chose randomly 40 examples to form the test set and used the remaining 100 as training set. When the files were prepared in the form required by SNNS we were ready to start the experiments.

4 Tested network architectures

The first choice we made was to use only simple feed–forward, back–propagation networks. It seems reasonable, as we already know that such networks can handle the recognition of frontal

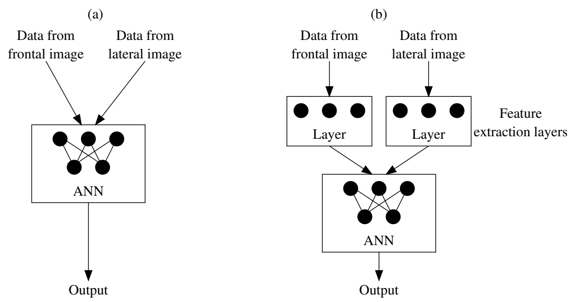

Figure 6: Possible architectures of the network for dual–view recognition

data [13, 15]. After some experiments we found out that the network with an input layer, two hidden layers and an output layer and quite small number of neurons in hidden layers is an optimal solution in terms of achieved recognition rate and network complication. The actual architecture parameters will be presented in the next section.

When incorporating two data streams we faced the problem of choosing the best network architecture. As it is depicted on Fig. 6 we decided to choose from two different levels of integration.

In the first case, the data coming from both streams are treated as a single source of information (Fig. 6a). We then have a single fully connected feed–forward network that processes the information without making the distinction between the two modalities of information. This approach with homogeneous network architecture led to the results worse than those obtained on basis of frontal only data.

In the second case there exists a part of the network that is splitted into two processing subnetworks (Fig. 6b). Such a network architecture results in faster and more reliable learning than the previous one and the best results obtained with this architecture were as good as for the *frontal only* case. Of course there remains still the questions about e.g. at which point the network should be splitted. In other words: should there be only one feature extraction layer or maybe more of them, how many layers should the network contain in the later part, what should be the number of neurons in the feature layers for each data stream etc. The network that achieved the best results is presented in the next section.

5 Experimental results

5.1 Frontal view recognition

When using neural networks for recognition of data prepared according to the model based on 21 characteristic distances, we obtained the results that are summarized in Tab. 1. The network

Table 1: Recognition rates for training and test sets for given emotions compared to recognition rates of human observers.

	training set	test set	human
frontal	90.0%	77.5%	80.0%
lateral	61.0%	55.0%	60.7%
dual	95.0%	77.5%	—

architecture that performed best in this case contained two hidden layers with only 7 neurons each The output layer consisted of 7 neurons; one for each emotion and one for neutral facial expression. The same output format was used for all other tested network architectures.

The recognition rate between 75% and 80% is similar to the recognition rate achieved by human observers. Also the cross–recognition table looks similar to the table obtained by visual inspection. The network has no problems with emotions such as happiness or anger. The neutral facial expression is recognized even in 100% of the cases. The emotions that cause wrong recognition are fear (often recognized as surprise) and disgust. The results similar to the above were already reported [13, 15], so we can conclude that the measurements, data preparation etc. were done correctly.

5.2 Lateral view recognition

For the lateral–only data the best results were achieved with the network with respectively 6 and 10 neurons in two following hidden layers The recognition rate in this case was, as one could expect, significantly lower than in case of frontal data.

The recognition rate of human observers in this case is 61% [16, 14] which is about the same as the recognition obtained by the neural network. That shows that using just a face profile is enough to recognize the emotional facial expressions with a satisfying rate.

5.3 Dual view recognition

To combine the two modalities of data we had to choose the appropriate model of networks' integration. After checking different possible network architectures we've found that the best results can be obtained with the network that has independent feature extraction layers and then one additional hidden layer that combines the output of them. The input layer of this network is divided into two groups - 21 neurons for the frontal part of the data and 40 neurons for lateral one. The feature extraction layers have in this case

Table 2: Cross-recognition results (in %)for three neural networks described in Section 5

Frontal	Anger	Disgust	Fear	Happiness	Neutral	Surprise	Sadness
Anger	**80**	0	0	0	20	0	0
Disgust	0	**50**	17	17	16	0	0
Fear	0	0	**40**	0	0	60	0
Happiness	0	0	0	**100**	0	0	0
Neutral	0	0	0	0	**100**	0	0
Surprise	0	0	17	0	0	**83**	0
Sadness	0	0	0	0	17	0	**83**
Lateral	Anger	Disgust	Fear	Happiness	Neutral	Surprise	Sadness
Anger	**20**	0	0	20	40	0	20
Disgust	0	**50**	0	0	33	17	0
Fear	0	20	**60**	0	20	0	0
Happiness	0	17	0	**83**	0	0	0
Neutral	33	0	0	17	**33**	0	17
Surprise	0	0	17	17	0	**66**	0
Sadness	17	0	17	0	0	0	**66**
Dual	Anger	Disgust	Fear	Happiness	Neutral	Surprise	Sadness
Anger	**60**	0	0	0	40	0	0
Disgust	0	**67**	0	0	33	0	0
Fear	0	0	**80**	0	0	0	20
Happiness	0	0	0	**100**	0	0	0
Neutral	0	0	0	17	**83**	0	0
Surprise	0	0	33	0	0	**67**	0
Sadness	0	0	0	0	17	0	**83**

respectively 7 and 6 neurons for frontal and lateral part of the data. The next hidden layer contains 10 neurons, and the output layer of course as usual consists of 7 neurons representing each of the 6 basic emotions and neutral expression of the face.

This network has similar performance on the test set and a little better on the training data set than the network that based the recognition only on frontal view. It seems that the lateral part of the data doesn't contain emotional information that couldn't be derivable from the frontal view and so adding the lateral view to the system does not significantly improve the recognition. This leads us to the conclusion that in fact all the visual information about emotional expression of the face is in the frontal view of the face. Of course that means that *combined* data is more redundant than *frontal only*. The question whether this redundancy can be in any way utilized by the recognition system remains open. The next experiment reported in this paper aims at proving that in fact it can.

5.4 Reliability of recognition

One of the possible ways to utilize the redundancy in the data is to use it for validation of the data when it is not reliable enough. In our case for example the measurements of the facial characteristic points and facial profile can be inaccurate. To test how the neural networks handle such case we applied a uniform noise with ascending level on the input data. We tested the *frontal only* and *combined* networks to see if dual–view approach gives any advantage over the single–view one.

The results of this experiment are depicted in Fig. 7. We can see that a *combined* network behaved here slightly better. With addition of noise the recognition rates of the test set degraded from 77.5% for both of the networks to respectively 42.5% and 50.0% for *frontal only* and *combined*. Although the above results are only indication of the possible advantages of dual–view approach, they show that it is worth to do further investigation in this field.

6 Conclusion

An expression shown on a human face can be represented as a point in the space of AUs. The visual changes caused by changing of the facial state are in fact three dimensional. The muscle movements that underly this process are not

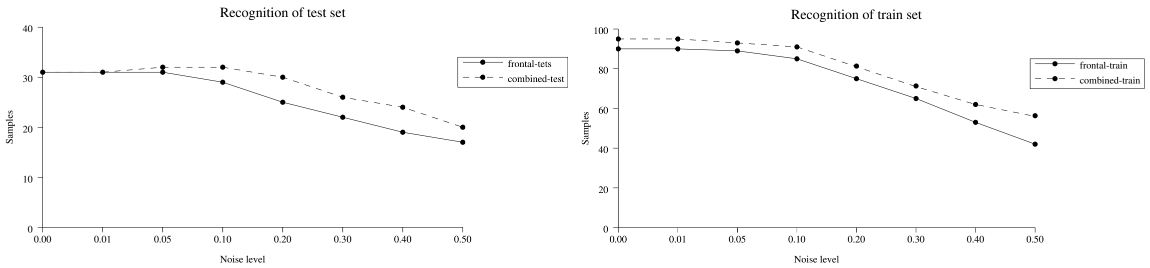

Figure 7: The number of correctly recognized training and testing samples with different levels of noise applied on the data.

aligned in any specific plane. This implies that the 3–D models in facial expression recognition are the most appropriate. The problem here is that the stereoscopic vision or 3–D geometry reconstruction are a broad fields of research for themselves only, so we needed a simpler approach. In this paper we presented that the depth information may be added to the recognition by utilizing a lateral view of the face. The dual–view approach is thus a way to combine frontal and lateral views.

We presented an approach in which the lateral view of the face is processed with very simple techniques. Combination of this “raw data” with the data extracted from the facial view according to well–established model was the input of our recognition system. Our reported system was build with use of ANNs using the SNNS tool. In this way without too much effort put into the design of the system itself, we could test how the dual–view data can be processed and recognized.

The results obtained in our experiments show that the recognition rates of emotional facial expressions in both frontal and combined views are about the same. That indicates that there is no additional information in the lateral view of the face that couldn’t be somehow concluded from the frontal view of the face. But we showed also that the redundancy of data introduced by a lateral view can be utilized by the recognition system. In our experiments the ANN trained *combined* data performed better in presence of noise than the network trained on *frontal only* data. The difference of almost 8% in recognition rate shows the potential hidden in the redundancy of dual–view information.

In the next future similar experiments should be conducted with different recognition systems (ES, fuzzy systems) and with more sophisticated models for the lateral view of the face. For example reasoning techniques could be used to determine errors in the information coming to the system or to make something like a “fall–back” mechanism in case when preprocessing of one part of the data failed. Expert systems seem to be the perfect choice for such goals and applying them on the dual–view recognition could be the next step in this research.

References

[1] P. Ekman and W. V. Friesen. *Unmasking the Face. A guide to recognizing emotions from facial clues.* PRENTICE-HALL, INC., Englewood Cliffs, New Jersey, 1975.

[2] 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.

[3] L. J. M. Rothkrantz, R. van Vark, and E. Kerckhoffs. General set-up of stress assessment through multimedia (SAM) system. In *Proceedings of MediaComm ’95*, Southampton, United Kingdom, Apr. 1995. Society for Computer Simulation.

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

[5] A. P. Pentland. Smart rooms: Machine understanding of human behavior. In *Computer Vision for Human-Machine Interaction* [17].

[6] K. Mase. Human reader: A vision-based man-machine interface. In *Computer Vision for Human-Machine Interaction* [17].

[7] 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.

[8] B. D. P. Maes, B. Blumberg, and A. Pentland. Full-body interaction with autonomous agents. In *Proceedings of Computer Animation'95*, 1995.

[9] R. Nakatsu and N. Tosa. *Neuro Baby*. ATR Media Integration & Communications Research Laboratories, Kyoto, Japan, 1996.

[10] 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* [18].

[11] M. Pantic, L. J. M. Rothkrantz, and H. Koppelaar. Automation of non-verbal communication of facial expressions. In *Proceedings of Euromedia-'98* [18].

[12] L. Harmon, M. Khan, R. Lash, and P. Ramig. Machine identification of human faces. *Pattern Recognition*, pages 97–110, 1981.

[13] H. Kobayashi and F. Hara. The recognition of basic facial expressions by neural network. In *International Joint Conference on Neural Networks, Singapore/Sydney*, pages 440–466, 1991.

[14] J. C. Wojdeł, A. Wojdeł, and L. J. M. Rothkrantz. Analysis of facial expressions based on silhouettes. In *ASCI'99 annual conference of Advanced School for Computing and Imaging*, 1999.

[15] M. Friedl and T. Svoboda. Recognition of facial expressions by neural network. Technical report, Delft University of Technology, Delft, The Netherlands, Sept. 1994.

[16] T. Chrpa, P. Hrubeš, P. Paclík, J. Prchal, P. Pták, and L. Vydra. Face profile segmentation. Technical report, Delft University of Technology, Delft, The Netherlands, Sept. 1998.

[17] R. Cipolla and A. P. eds. *Computer Vision for Human-Machine Interaction*. Cambridge University Press, 1998.

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