Proc. 1998 IEEE Int. Conf. on Face and Gesture Recognition, Nara, Japan 1998, pp. 486-491

User Localisation for Visually-based Human-Machine-Interaction

Hans-Joachim Boehme, Ulf-Dietrich Braumann, Anja Brakensiek, Andrea Corradini,
Markus Krabbes, Horst-Michael Gross
Technical University of Ilmenau, Department of Neuroinformatics
98684 Ilmenau, Germany, hans@informatik.tu-ilmenau.de

Abstract

Recently there is an increasing interest in video based interface techniques, allowing more natural interaction between users and systems than common interface devices do.

Here, we present a neural architecture for user localisation, embedded within a complex system for visually-based human-machine-interaction (HMI).

User's localisation is an absolute prerequisite to video-based HMI. Due to the main objective, the greatest possible robustness of the localisation as well as the whole visual interface under highly varying environmental conditions, we propose a multiple cue approach. This approach combines the features facial structure, head-shoulder-contour, skin color, and motion, with a multiscale representation. The selection of that image region most likely containing a possible user is then realised via a WTA-process within the multiscale representation.

Preliminary results show the reliability of the multiple cue approach.

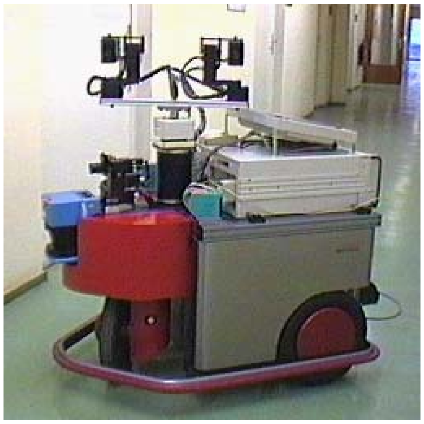

Figure 1. The mobile robot MILVA, provided with 68040-VME-system, 2 PC-systems, CNAPS-board, framegrabbers, and several sensors (3 cameras, laserscanner, ultrasound and infrared distance measures, bumpers).

1. Introduction

A considerable number of approaches for the design of intelligent and adaptive human-machine-interfaces have been proposed (see for instance [7, 8, 15]).

In our group, a project named GESTIK (supported by the Thuringian Ministry of Science, Research and Culture) was started to develop a neural network architecture for video-based HMI between a user and the robot MILVA (Multisensory Intelligent Learning Vehicle in neural Architecture). A two-camera-system with 7 degrees of freedom (for each camera pan, tilt and zoom, additional pan for both cameras) serves for the interaction with a possible user and actively observes its operational environment. An additional camera, mounted at the front of the robot, provides the visual information for navigation. MILVA is shown in figure 1 and serves as the test bed for interaction with a user, at the moment based on the transmission of behavioral commands via static gestures.

The use of our system as an intelligent luggage carrier, for instance at a railway station or an airport, was chosen as a hypothetic scenario for the following reasons: First, we must take into account the capabilities of our robot which does not have manipulators and can only move itself. Second, the scenario is to naturally motivate a gesture-based dialogue between the user and the serving system. At a railway station with a lot of people and a high amount of surrounding noise a gesture-based dialogue seems to be the only possible way for interaction.

In this article, we concentrate on a subsystem responsible for the localisation of a possible user, which is the absolute prerequisite to any kind of HMI. Due to the fact, that MILVA has to operate under highly variable environmental conditions (scene content, illumination), the requirements on the user localisation system are very hard. Consequently, we designed a neural architecture, integrating several cues to make the localisation robust and most possible independent of variable environmental conditions.

2. Neural architecture for user localisation and gesture recognition

Figure 2 provides a coarse sketch of the whole neural architecture for user localisation and gesture recognition.

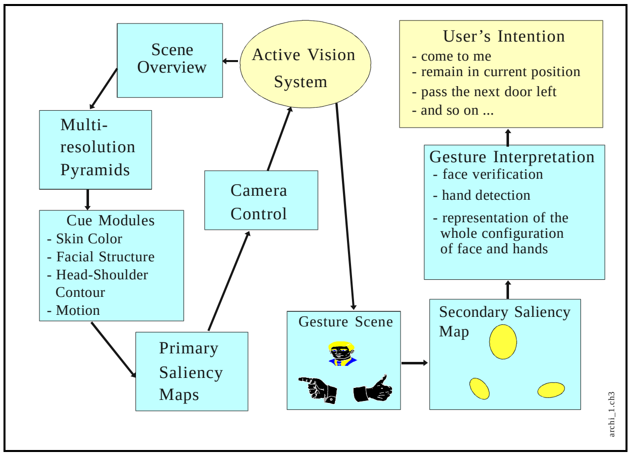

Figure 2. Building blocks of the neural architecture for user localisation and gesture recognition

That components of the architecture responsible for user localisation are described in the following section.

3. User localisation

3.1. Cue modules

Initially both cameras of the two-camera-system operate in wide-angle-mode in order to cover the greatest possible area of the environment. Multiresolution pyramids transform the images into a multiscale representation. Four cue modules which are sensitive to *skin color*, *facial structure*, *structure of a head-shoulder-contour* and *motion*, respectively, operate at all levels of the two pyramids. The utility of the different, parallel processing cue modules is to make the whole system robust and more or less independent of the presence of *one certain* information source in the images. Hence, we can handle varying environmental circumstances much easier, which, for instance, make the skin color detection difficult or almost impossible. Furthermore, high expense for the development of the cue modules can be avoided (see [4, 3, 11], too).

a) Skin color
For the generation of a skin color training data set, portrait images of different persons (of our lab) were segmented manually. The images were acquired under appropriate lighting conditions.
A linear transformation maps the RGB-values into a physiologically motivated fundamental color space (see [17]), which is formed by a Red-Green(RG)-, Yellow-Blue(YB)-, and Black-White(BW)-dimension. The pixels (color values) of an image form a certain cluster within this color space. The whole cluster will be elongated from the WB axis (achromatic axis) depending on the illuminative conditions during image acquisition. The elongation of this cluster characterizes the deviation in illumination from the typical daylight condition regardless of the image contents. By means of a color adaptation process, the cluster is transformed in such a way that its elongation will be along the BW axis. So we can ensure equal color sensations under different lighting conditions (see [17]).

Whereas the color adaptation is carried out within the described fundamental color space, the skin color classification takes place within the RGB color space. To reduce the influence of varying intensities, the projection formed by the normalized R- and G-achses (r', g') is utilised. To model the skin color distribution roughly, we define a 2-dimensional Gaussian function via calculation of the mean and the covariance (see figure 3, too) of the skin color data set.

If a face region could be verified, a new Gaussian model is created, more specific for the illumination and the skin type at hand. Via this model the detection of skin colored regions, especially hands, can be improved. This is very important because the hand regions cannot be segmented by structural information (see [13], too).

The different skin color models are necessary for the following two reasons: In the beginning, the system observes its operational area and the skin color segmentation has to operate with the coarse model because no face verification is available. Just after the first successful face verification, the fine-tuned model, based on actual skin color and illumination, can be used. Figure 3 gives a segmentation example using the coarse as well as a fine-tuned skin color model.

A detailed description of our skin color investigations can be found in [6].

b) Facial structure
Because of the unknown distance between the camera and the user to be localised, the detection of facial structure has to be carried out at each level of the multiresolution pyramids (see also figure 2). In our scenario we assume that a person is an intended user if its face is oriented towards the

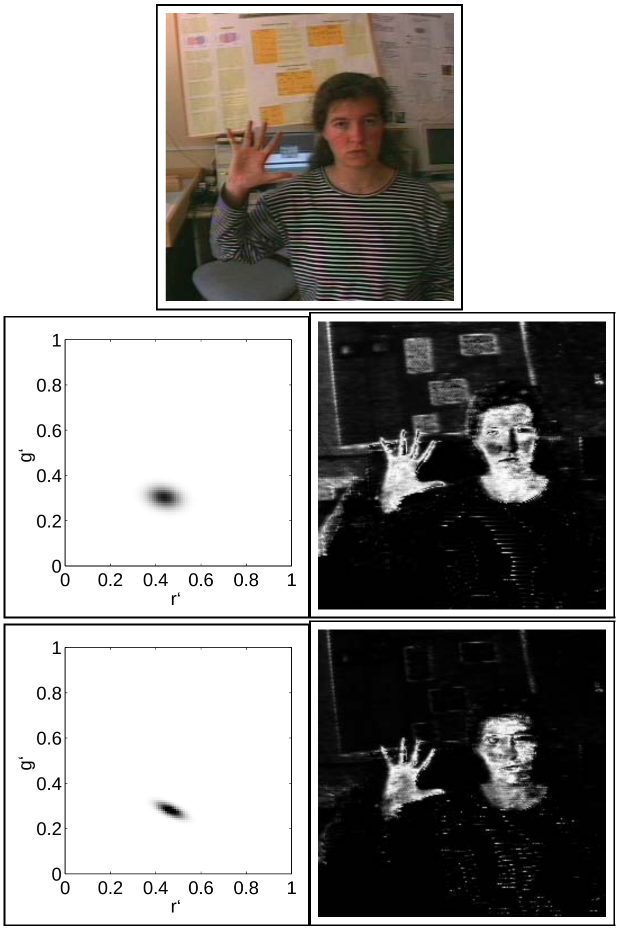

Figure 3. Top: original imageMiddle: coarse Gaussian skin color model and segmentation result; Bottom: fine-tuned Gaussian skin color model after face verification and segmentation result

robot.

The detection of facial structure uses the gray value image and employs eigenfaces generated by a principal component analysis (PCA) of the images contained in the ORL data set (http://www.cam-orl.co.uk/facedatabase.html; see [16], too). The image regions used for the PCA were extracted manually, cover a region of 15 x 15 pixels, and the regions were normalized by their mean and standard deviation (see also [20, 19]). Then, the input image is processed with 3 eigenfaces (according to the largest eigenvalues). Besides the preprocessing steps, the classification of the obtained fit values remains a difficult problem. The best results we achieved with a Growing-Neural-Gas-Network (GNG, [10]) performing a mapping from the fit values to 2 classes (face, no face). For the training of the GNG a data set of 174 positive (face) and 174 negative (no face) examples was created. To improve the generalization ability of the network, we implemented a bootstrap algorithm [19] which encloses false classified image regions into the set of the negative examples automatically. Besides the preprocessing steps explained above, we use no further transformations as, for instance, histogram equalization. The remaining uncertainties of the detection of facial structure can be compensated by the parallel use of all different cue modules (see also [5]).

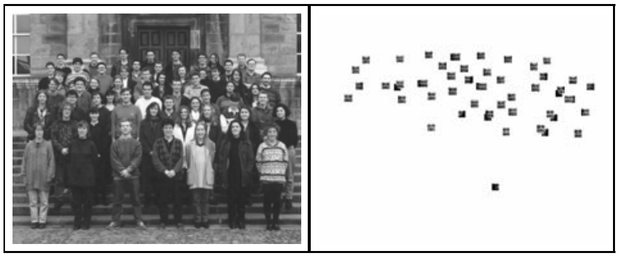

Figure 4. Detection of faces using eigenvector masks and GNG as neural classifier. The detected faces are marked in the right image (likelihood higher than 0.7).

An example for the face detection is demonstrated in figure 4, where an image taken from [19] was processed. False positive detected regions cannot be avoided entirely, but such regions very likely cover no skin color, and therefore, by combining skin color and facial structure such mislocalisations can be rejected.

c) Head-shoulder-contour

Similar to the detection of facial structure, the localisation of a head-shoulder-contour operates on the gray level image of each level of the multiresolution pyramids. The basic idea is to use an appropriate spatial configuration of Gabor filters (see figure 5) and to classify the obtained filter outputs by a specially tuned distance measure between the actual filter outputs and a prototype.

d) Motion

Our favoured approach was proposed in [2] and [9], and is demonstrated in figure 6. Based on image differentiation motion is detected in the first step, leading to a binary motion energy image. The second step accumulates this motion energy over a certain period of time resulting in a motion history image. This approach is reliable especially for the following reason: The detection as well as the accumulation of motion could be realized via dynamic neural fields, and by means of different sets of parameters of such

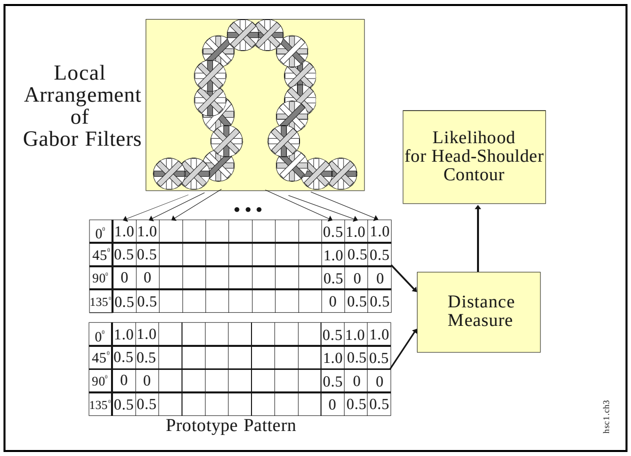

Figure 5. Method for detection of a head-shoulder-contour, based on a specially fitted grid of Gabor filters and a task specific distance measure

fields, different task specific aspects of motion information can be obtained.

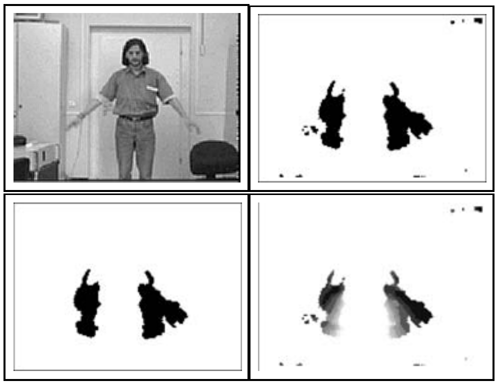

Figure 6. Top Left: original image from the sequence; Top Right: binary motion image; Bottom Left: thresholded binary motion image; Bottom Right: motion history image

3.2. Generation of primary saliency maps

The output of the cue modules serves as the input for the primary saliency maps at each level of the multiresolution pyramid. The maps are topographically organised neural fields containing dynamic neurons interacting with each other (see [1, 14, 21, 12]). In the primary saliency maps *all that regions* are to become prominent that cover *gesture-relevant* parts such as faces and hands. Because of the features facial structure and head-shoulder-contour, faces become the most prominent or salient regions. The saliency map containing the overall most salient activity blob determines the further processing steps.

Currently, the pyramid containing the primary saliency maps (see figure 7) is under construction.

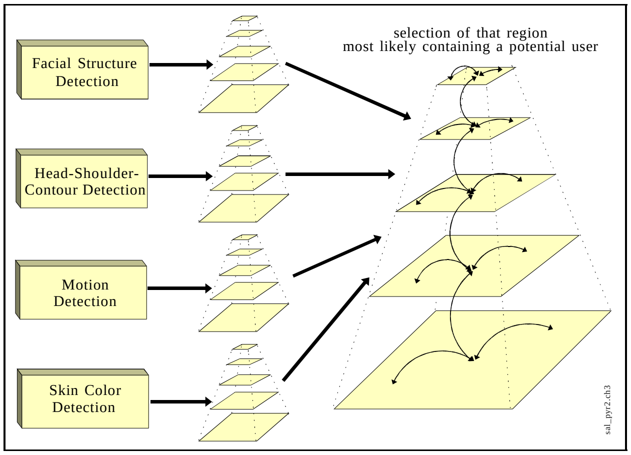
Facial Structure Detection
Head-Shoulder-Contour Detection
Motion Detection
Skin Color Detection
selection of that region most likely containing a potential user
sal_pyr2.ch3

Figure 7. Generation of a scale space pyramid of primary saliency maps

As a preliminary result concerning the primary saliency maps, figure 8 shows the selection of the most likely head-shoulder-contours at all levels of the multiresolution pyramid. Here, dynamic neurons interact inside each level and between adjacent levels of the pyramid. The neurons receive their input from the head-shoulder-contour detector. Due to the fact, that the head-shoulder-contour detector supplies a strong output at adjacent levels of the pyramid, the selection becomes much more robust, and numerous false positive detections can be rejected.
The same principle is to be extended to the whole saliency pyramid, integrating all cue modules.

3.3 Control of the two-camera-system

After the localisation of a possible user the final step envolves the control of the second camera.

As soon as a possible user (face region) is detected in one of the camera images, this camera serves as *general-view-camera*, whereas the second camera becomes the *gesture-camera*. The necessary distance estimation is provided by the cue modules detecting structural information (face and head-shoulder-contour). The gesture-camera is controlled

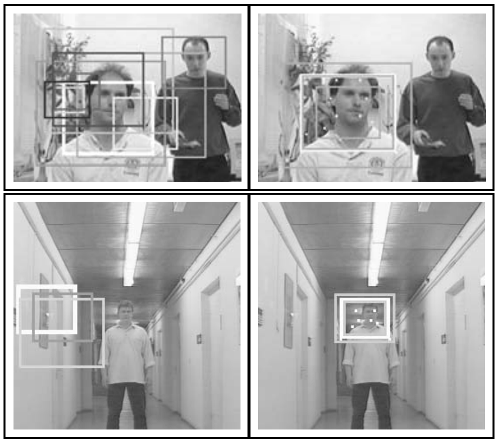

Figure 8. Input images with marked head-shoulder-contours, obtained at the different levels of the multiresolution pyramid by the proposed method; the left images of each example show the result without dynamic selection, whereas the right images contain the selected contours obtained by means of dynamic selection (white rectangles mark the highest likelihood).

such that the expected face region will appear on a predefined position in the image with a predefined scale, too.

First, a camera control module for a single camera was implemented based on a neural approach proposed in [18], where a Kohonen-Map is used that learns an input-output-mapping between the actual target position and the corresponding pan/tilt angles.

Recently, this method was extended for the control of the two-camera system. The basic idea is that a definite configuration of the cameras is assumed, which is necessary to use that mapping method of the single camera system for the two-camera-system, too.

Therefore, after a possible user (face region) was located in either camera image, the second camera is directed towards this user. This is realized by means of controlling the pan/tilt of this camera as well as the additional pan for both cameras. Therefore, the initial camera configuration (especially the base distance) remains stable. The resulting *gesture-scene* should contain all gesture-relevant parts of the intended user. By means of the control of the gesture-camera we can ensure that faces and hands will always have approximately the same size, so we do not have to ensure scale invariance by the further processing steps.

4. Conclusion and outlook

4.1. User localisation

Depending on the environmental conditions (illumination, image content, distance between robot and user) which can neither be influenced nor be estimated a priori, the different cue modules provide more or less confident results. Our preliminary results concerning user localisation clarify, that only the parallel utilization of different methods leads to appropriate localisation results. Hence, the system becomes much more robust, can handle highly varying environmental conditions and is less dependent on the presence of one certain feature.

Furthermore, we concentrate on the final implementation of the pyramid containing the primary saliency maps. Only when the whole primary saliency system is stable running, we can estimate the sufficiency of the developed cue modules. The cue module for motion analysis has to be realized and integrated into the saliency system.

4.2. Work in progress

Generation of the secondary saliency map A secondary saliency map is created for the gesture-scene, which determines the sequential processing of this scene. Similar to the primary saliency map we utilize topographically organised neural fields, too.

To simplify the task, we use only the skin color information as the input for this field, thereby assuming that the skin color segmentation is robust enough.
Because of the camera control, the prominent position and size of a hypothetic face region is known. So, by means of specially tuned field parameters (coupling width and strength) the emergence of an activity blob covering the face region is highly supported. Therefore, the face region will be the first area to be analysed in detail (see the following paragraph). The hand regions become salient, too.

Face verification and estimation of face orientation The next processing step must provide a face verification, that means we have to decide if there is a face at all, and if it is oriented towards the robot.

Detection and interpretation of gestures For complexity reasons, we have predefined a gesture alphabet and have assumed only static gestures (poses), which are stable for a certain period of time. The mapping between the gestures to be recognized and the associated actions of the robot is predefined, too (see also [15]).

Further, we assume that the content of a gesture can be extracted only by taking into account the whole configuration of face and hands, whereas the orientation of face and

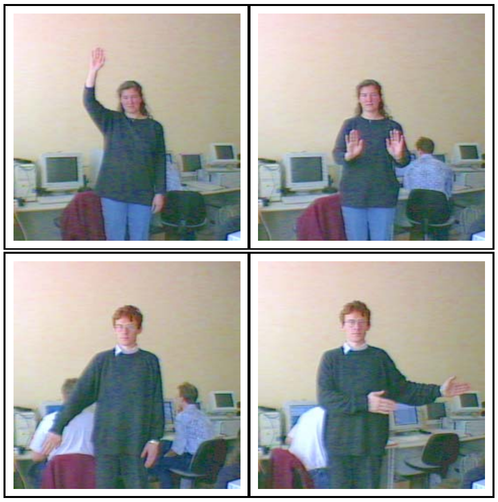

Figure 9. Possible intuitive gestures (poses); from left to right they could carry the following meanings for the robot: hello, stop, come to my left, move right

hands is not important at the moment (see fig. 9). These restrictions are only introduced to handle the ongoing problems and they shall be put away step by step.

Acknowledgements

The authors thank Rolf Nestler, Thomas Kleemann and Carsten Rieck for helpful comments and discussions as well as the technical support.

References

[1] S. Amari. Dynamics of pattern formation in lateral-inhibition type neural fields. *Biological Cybernetics*, 27:77–87, 1977.

[2] A. Bobick. Computers seeing action. In *British Machine Vision Conference*, 1997.

[3] Boehme, H.-J., Brakensiek, A., Braumann, U.-D., Krabbes, M., and Gross, H.-M. Neural Architecture for Gesture-Based Human-Machine-Interaction. In *Gesture-Workshop Bielefeld*. Springer Verlag, 1997. to appear.

[4] Boehme, H.-J., Brakensiek, A., Braumann, U.-D., Krabbes, M., and Gross, H.-M. Visually-Based Human-Machine-Interaction in a Neural Architecture. In *SOAVE'97 - Selbstorganisation von adaptivem Verhalten*, pages 166–175. VDI Verlag, 1997.

[5] Brakensiek, A., Braumann, U.-D., Boehme, H.-J., Rieck, C., and Gross, H.-M. Farb- und strukturbasierte, neuronale Verfahren zur Lokalisation von Gesichtern in Real-World-Szenen. In *Mustererkennung 1997*, pages 113–120, 1997.

[6] Brakensiek, A., Braumann, U.-D., Corradini, A., Boehme, H.-J., and Gross, H.-M. Neuronale Verfahren zur Lokalisation und Bewertung von Handgesten in Real-World-Szenen. In *Neuronale Netze in der Anwendung*. Technische Universitaet Magdeburg, 1998.

[7] C. Maggioni and B. Kaemmerer. GestureComputer – History, Design, and Applications. In *Computer Vision in Man-Machine Interfaces*. Cambridge University Press, 1996.

[8] Darrell, T., Basu, S., Wren, C., and Pentland, A. Perceptually-driven Avatars and Interfaces: active methods for direct control. Technical report, MIT Media Lab Perceptual Computation Section, 1997. TR 416.

[9] Davis, J.W. and Bobick, A.F. The Representation and Recognition of Action Using Temporal Templates. Technical report, MIT Media Lab Perceptual Computation Section, 1997. TR 402.

[10] B. Fritzke. A Growing Neural Gas Network Learns Topologies. In *Advances in Neural Information Processing Systems*, volume 7, pages 625–632, 1995.

[11] S. Goodridge. *Multimedia Sensor Fusion for Intelligent Camera Control and Human-Computer-Interaction*. PhD thesis, North Carolina State University, 1997.

[12] Gross, H.-M., Franke, R., Boehme, H.-J., and Beck, C. A Neural Network Hierarchy for Data and Knowledge Controlled Selective Visual Attention. In *International Conference on Artificial Neural Networks*, pages 825–828, 1992.

[13] M. Hunke. Locating and Tracking of Human Faces with Neural Networks. Technical report, Carnegie Mellon University, 1994. CMU-CS-94-155.

[14] K. Kopecz. Neural field dynamics provide robust control for attentional resources. In *Aktives Sehen in technischen und biologischen Systemen*, pages 137–144. Infix-Verlag, 1996.

[15] Kortenkamp, D., Huber, E., and Bonasso, P.R. Recognizing and interpreting gestures on a mobile robot. In *Thirteenth National Conference on Artificial Intelligence (AAI-96)*, 1996.

[16] Moghaddam, B. and Pentland, A. Maximum Likelihood Detection of Faces and Hands. In *International Workshop on Automatic Face- and Gesture-Recognition*, pages 122–128, 1995.

[17] Pomierski, T. and Gross, H.-M. Biological Neural Architecture for Chromatic Adaptation Resulting in Constant Color Sensations. In *International Conference on Neural Networks (ICNN'96)*, pages 734–739, 1996.

[18] Ritter, H., Martinetz, T., and Schulten, K. *Neuronale Netze*. Addison-Wesley, 1991.

[19] Rowley, H. A., Baluja, S, and Kanade, T. Human Face Detection in Visual Scenes. Technical report, Carnegie Mellon University, 1995. CMU-CS-95-158R.

[20] Schiele, B. and Waibel, A. Gaze Tracking Based on Face-Color. In *International Workshop on Automatic Face- and Gesture-Recognition*, pages 344–349, 1995.

[21] Stephan, V. and Gross, H.-M. Formerhaltende sequentielle visuelle Aufmerksamkeit mittels strukturierter neuronaler Felder. In *Mustererkennung 1997*, pages 411–418, 1997.