1998 IEEE World Congress on Computational Intelligence Proceedings (WCCI '98), IJCNN, vol. 1, pp. 372-377

Neural Networks for Gesture-based Remote Control of a Mobile Robot

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

Abstract— **We present a neural network architecture for gesture-based interaction between a mobile robot and its user, thereby spanning a bridge from the** *localisation of the user* **over the** *recognition of its gestural instruction* **to the** *generation of the appropriate robot behavior*. **Since this system is applied under real-world conditions, especially concerning the localisation of a human user, some proper techniques are needed which have an adequate robustness. Hence, the combination of several components of saliency towards a multi-cue approach, integrating structure- and color-based features, is proposed. At the moment, the gestures itself are very simple and can be described by the spatial relation between face and hands of the person. The organisation of the appropriate robot behavior is realised by means of a mixture of neural agents, responsible for certain aspects of the navigation task. Due to the complexity of the whole system, above all we use "standard neural network models", which are modified or extended according to the task at hand. Preliminary results show the reliability of the overall approach as well as the sufficient functionality of the already realised submodules.**

I. Introduction and scenario

Fig. 1. The mobile robot MILVA. Provided with the necessary on-board equipment (68040-VME-system, 2 PC-systems, CNAPS-board, framegrabber) and different sensors (3 cameras, laserscanner, ultrasound and infrared distance measures, bumpers) MILVA serves as the testbed for the human-machine-interaction.

Figure 1 shows our robot platform 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.

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.

Recently there is an increasing interest in video based interface techniques, allowing more natural interaction between users and systems than common interface devices do. A considerable number of approaches for the design of intelligent and adaptive human-machine-interfaces have been proposed (see for instance [8], [15], [7]).

The superior goals of our research concerning the proposed architecture (GESTIK-project[1]) are the highest possible robustness of the intelligent visual interface under highly varying environmental conditions as well as the sufficient organisation of the appropriate robot behavior, achieved by continuous interaction between robot and human user.

II. Neural architecture for user localisation and gesture recognition

Figure 2 provides a coarse sketch of the neural architecture for user localisation and gesture recognition. The several components of the architecture will be described in the following subsections.

[1] Supported by the Thuringian Ministry of Science, Research and Culture (TMWFK, GESTIK-Project)

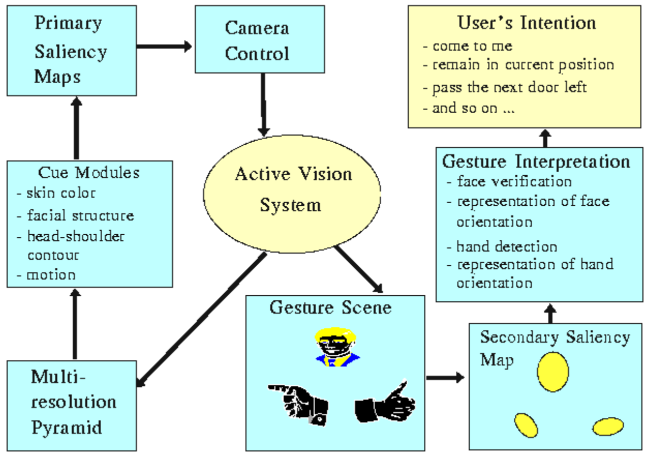

Fig. 2. Building blocks of the neural architecture for user localisation and gesture recognition

A. Multi-cue approach for user localisation

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 (typical for our lab environment).

In order to obtain almost constant color sensation, first we map the RGB color space into a fundamental color space and employ a color adaptation method (see [21]). Then, we return into the RGB color space and define a 2-dimensional Gaussian function via calculation of the mean and the covariance of that skin color data set to model the obtained skin color distribution roughly. Furthermore, 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).

A more detailed description of our skin color investigations can be found in [5] and [6].

b) Facial structure

In our scenario we assume that a person is an intended user if its face is oriented towards the 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 [19], too). The image regions (15 x 15 pixels) used for the PCA were extracted manually and were normalised by their mean and standard deviation (see also [24], [23]). Then, the input image is processed with 3 eigenfaces (largest eigenvalues). Besides the preprocessing steps, the classification of the obtained fit values remains a difficult problem. The best results we achieved with a supervised 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 generalisation ability of the network, we implemented a bootstrap algorithm [23] which encloses false classified image regions into the set of the negative examples automatically.

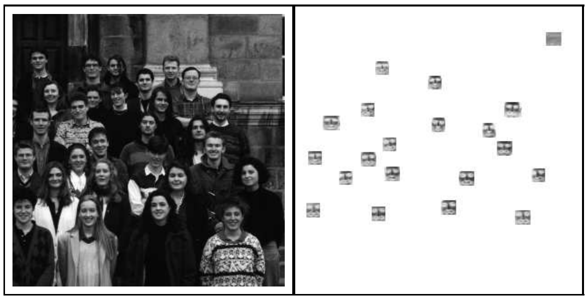

Fig. 3. Detection of frontally aligned faces. The detected faces are marked in the right image (likelihood higher than 0.7).

The performance of the face detection is demonstrated in figure 3, where an image taken from [23] was processed. False positive detected regions cannot be avoided entirely (top right), but this region very likely covers 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 (filter arrangement, see figure 4) and to classify the obtained filter outputs by a specially tuned distance measure (Hamming distance) between the actual filter outputs and a prototype.

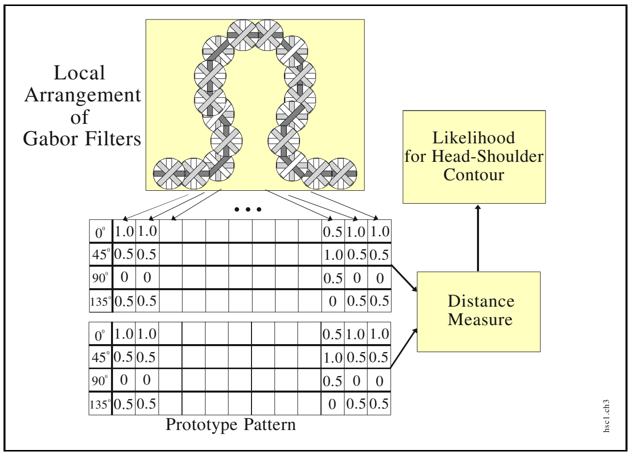

Fig. 4. Processing scheme for detection of a head-shoulder-contour.

d) Motion

Our favoured approach was proposed in [2] and [9]. 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 realised via dynamic neural fields, and by means of different sets of parameters of such fields, different task specific aspects of motion information can be obtained.

e) Dynamic neural fields for generation of primary saliency maps

All cue modules supply input for the primary saliency maps at each level of the multiresolution pyramid, as shown in figure 5.

To achieve a good localisation a selection mechanism is needed to make a definite choice. This is not limited to a two-dimensional position. Since we use five resolutions (fine to coarse) we actually can localise persons even in different distances. Therefore, a neural field (array) for selection of the most salient region should be three-dimensional.

Those fields can be described as recurrent nonlinear dynamic systems (cf. [1], [14]). Regarding to the selection task we need a dynamic behavior which leads to

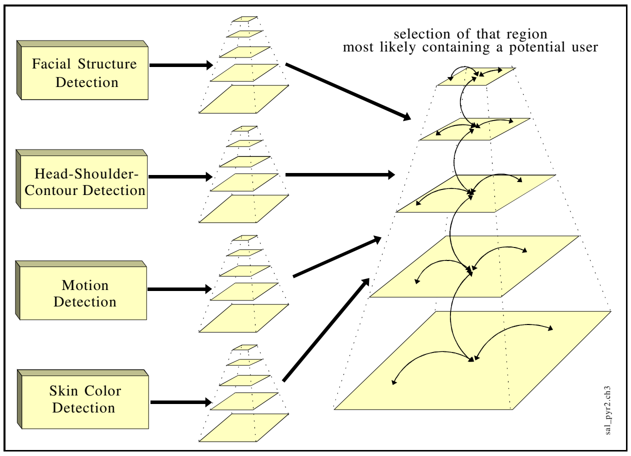

Fig. 5. Generation of a scale space pyramid of primary saliency maps

one local region of active neurons successfully competing against the others, i.e. the formation of one single blob of active neurons as an equilibrium state of the field. The following equations describe the system:

$$\tau \frac{d}{dt} z(\vec{r},t) = -z(\vec{r},t) - c_h h(t)$$

$$+c_l \int_R w(\vec{r}-\vec{r}\,')y(\vec{r}\,',t)d^2\vec{r}\,' + c_i x(\vec{r},t) \qquad (1)$$

$$\text{with} \quad w(\vec{r}-\vec{r}\,') = 2\exp(\frac{-3|\vec{r}-\vec{r}\,'|^2}{2\sigma^2})$$

$$-\exp(\frac{-|\vec{r}-\vec{r}\,'|^2}{\sigma^2}) \quad , \qquad (2)$$

$$y(\vec{r},t) = \frac{1}{1+exp(-z(\vec{r},t))} \quad \text{and} \qquad (3)$$

$$h(t) = \int_R y(\vec{r}\,'',t)d\vec{r}\,'' \qquad (4)$$

Herein $\vec{r} = (x,y,z)^T$ denotes the coordinate of a neuron, $z(\vec{r},t)$ is the activation of a neuron $\vec{r}$ at time t, $y(\vec{r},t)$ is the activity of this neuron, $x(\vec{r},t)$ denotes the external input, $h(t)$ is the activity of a global inhibitory interneuron, $w(\vec{r}-\vec{r}')$ denotes the function of lateral activation of neuron $\vec{r}$ from the surrounding neighbourhood R. Further, τ is the time constant of the dynamical system and σ is the deviance of the gaussians determining the function of lateral activation. For the computation we used the following values for the constants: $c_h = 0.025$, $c_l = 0.1$, $c_i = 0.1$, $\sigma = 2$ (halved z-direction), $\tau = 10$ with $\Delta T = 1$ (ΔT: sampling rate). The range R of the function of lateral activation reachs over 5 pixels and 3 pixels in z-direction, respectively (anisotropic neighbourhood).

The results of the systems shall be qualitatively illustrated in figure 6. The presented results are exemplary, the usage of the shape of contour provides one solution for the person localisation problem, even under quite different conditions. In our ongoing work, the same principle is extended to the whole saliency pyramid, integrating all cue modules. The novel approach with a three-dimensional dynamic neural field can be assessed as robust method for the selection process, very reliable for the task at hand.

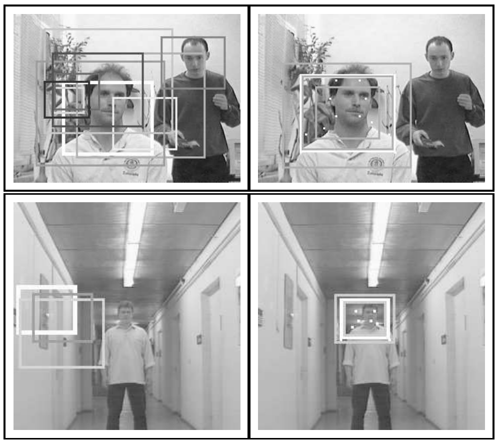

Fig. 6. Input images with marked head-shoulder-contours, obtained at the different levels of the multiresolution pyramid by the proposed method; *left:* without dynamic selection; *right:* by means of dynamic selection (white rectangles mark the highest likelihood).

B. Control of the two-camera-system

A camera control module, based on a neural approach proposed in [22], was extended for the control of the two-camera system. The basic idea is that a definite configuration of the cameras is assumed. Therefore, after a possible user (face region) was located in either camera image, the second camera is directed towards this user. This is realised 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.

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 resulting *gesture-scene* should contain all gesture-relevant parts of the intended user. Furthermore, the gesture-camera is controlled such that the expected face region will appear on a predefined position in the image with a predefined scale, too. Hence, we do not have to ensure scale invariance by the further processing steps.

Fig. 7. Possible intuitive gestures (poses); from left to right they could carry the following meanings for the robot: hello, stop, move right

C. Detection and interpretation of gestures

a) Definition of a gesture set

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 (see fig. 7). The mapping between the gestures to be recognized and the associated actions of the robot is predefined, too (see also [15]).

b) 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 utilise topographically organised neural fields.

To simplify the task, we mainly employ 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 section). The hand regions become salient, too.

c) Face verification and representation of gestures

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.

To obtain this information, a very simple method for direct mapping of grey value image parts to the corresponding object orientation (both, faces and hands), based on a MLP network, was tested. Preliminary results show the sufficient functionality of such an approach under certain constraints (unstructured background). At present, the approach is examined under real world conditions. Furthermore, the detailed analysis of faces and hands via a regular grid of Gabor filters and a following classification of the Gabor filter outputs with a neural classifier (see also [17], [18]), will be taken into account, too.

If there is no face at the assumed position, the face verification fails. In that case the gesture-camera turns towards the next salient region of the primary saliency maps or returns to the wide-angle-mode.

Besides the face, hand regions become prominent in the secondary saliency map, mostly due to their skin color, but we do not know whether the skin colored regions are hands or skin colored regions of the background. For simplicity reasons, actually we describe the static gestures by means of the spatial relation between face and hands of the person. So, a graph-like data structure is obtained. To decide, whether there is made a relevant gesture at all and what gesture it is, a distance measure between the actually obtained graph and the prototype gesture graphs will be implemented and tested.

III. Neural architecture for navigation behavior

The superior direction of the research in our department concerns the organisation of adaptive behavior. A lot of projects deal with the different aspects of behavioral organisation, such as direct mapping from sensory information to motor commands, organisation of sensorimotor representations, integration of different sensors, organisation of reactive as well as globally planning behavior and so on (see [12]), in order to extend the behavioral performance of the robot MILVA continuously.

A. *Multi-agent approach*

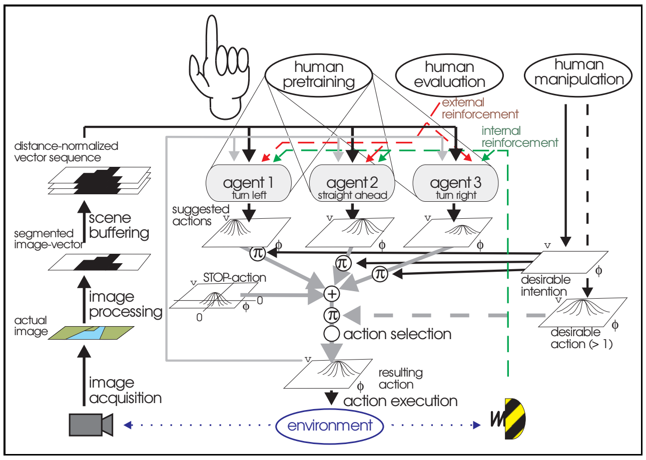

Fig. 8. Neural architecture for organisation of complex navigation behavior; see text for explanation

The second part of the GESTIK-project deals with the mapping from sensory situations to articulated behavior. Figure 8 outlines the corresponding architecture. The basic idea is that different neural agents are trained to become experts for certain navigation tasks or actions, respectively (turn left, straight ahead, turn right). The design of the neural agents is grounded on the ALVINN-approach proposed by Pomerleau [20], well known as a kind of imitation learning. The ALVINN-approach was extended in different directions: i) The input contains a path history, including the last perceived sensory situations, because the last sensory situations have equal importance for navigation, especially for turnings (depending on the robot geometry). ii) The output consists of a two-dimensional neural field, coding steering angle as well as as velocity value in a topological manner. iii) The selection of the appropriate action is done via a WTA process inside the two-dimensional output field.

To superimpose the action proposals coming from the different neural agents ("action selection" in figure 8), we utilise a dynamic neural field approach, too. The selected action is assumed as the best one, corresponding to the actual sensory situation.

B. *Integration of user's gestural instruction*

The interaction between the purely sensory based navigation behavior and the intention of the user, coming from the gesture recognition module, is realised by means of an additional neural agent ("human manipulation" in figure 8). The desired robot behavior could be, for instance, "come to me", articulated via the "hello"-pose. Then this submitted intention leads to a modulation of the action selection process, such that the moving towards the user is supported. However, the finally selected action will always be determined by the actions possible in the current sensory situation at all. In other words, if the user's intention supports the straight forward action, but there is an obstacle in front of the robot, the selection of this action will be inhibited. A detailed description of the multi-agent based organisation of robot behavior can be found in [16].

IV. Conclusion and outlook

The investigations described above, concerning the navigation behavior, were carried out using a miniature robot KHEPERA. Currently, we realize and test the submodules of the proposed neural architecture on the MILVA robot for real world human-robot-interaction.

Our preliminary results concerning user localisation clarify, that only the parallel utilisation 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. 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. For

Fig. 9. Skin color (middle) and face structure (right) hypotheses obtained from the input image (left)

example, the uncertainties concerning skin color segmentation are demonstrated in figure 9 (middle). However, the detection of facial structure supplies a reliable result (right).

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 realised and integrated into the saliency system.

Much effort is needed for the integration and test of the whole system, running on MILVA. Besides the parallel implementation of the methods detecting facial structure and head-shoulder contours to fulfil real-time requirements, the continuous interaction between robot and user remains a still difficult problem. Here, an interaction regime has to be developed, which allows the user to understand the current interpretation state of the robot.

Additionally to the visually-based interaction scheme a model for selective auditory attention was developed in our department (see [25]). This model was already implemented on MILVA and is to support the user localisation. For example, the user can attract MILVA's attention by clapping her hands.

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] Boehme, H.-J., Braumann, U.-D., Brakensiek, A., Krabbes, M., Corradini, A., and Gross, H.-M. User Localisation for Visually-based Human-Machine-Interaction. In *International Conference on Automatic Face- and Gesture Recognition*. IEEE, 1998. to appear.

[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. and Boehme, H.-J. Verhaltensorganisation in interaktiven und lernfaehigen mobilen Systemen – das MILVA-Projekt. In *SOAVE'97 – Selbstorganisation von adaptivem Verhalten*, pages 1–13. VDI Verlag, 1997.

[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] Krabbes, M., H.-J. Boehme, V. Stephan, and H.-M. Gross. Extension of the ALVINN-Architecture for Robust Visual Guidance of a Miniature Robot. In *EUROBOT'97*, 1997.

[17] Kubisch, R. and Ritter, H. Erkennung menschlicher Kopfhaltungen mittels kuenstlicher neuronaler Netze. In *Mustererkennung 1996*, pages 109–117, 1996.

[18] Littmann, E. and Ritter, H. Neural and Statistical Methods for Adaptive Color Segmentation - A Comparison. In *Mustererkennung 1995*, pages 84–93, 1995.

[19] 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.

[20] D. Pomerleau. *Neural Network Perception for Mobile Robot Guidance*. Kluwer Academic Publishers, 1993.

[21] 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.

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

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

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

[25] Zahn, T., Izak, R., Trott, K., and Paschke, P. A Paced Analog Silicon Model of Auditory Attention. In *Neuromorphic Systems – Engineering Silicon From Neurobiology*. World Scientific, 1997. in press.