Visually-Based Human-Machine-Interaction in a Neural Architecture *

H.-J. Boehme, A. Brakensiek, U.-D. Braumann, M. Krabbes, H.-M. Gross
Department of Neuroinformatics, Technical University of Ilmenau, 98684 Ilmenau (FRG)
e-mail: hans@informatik.tu-ilmenau.de http://cortex.informatik.tu-ilmenau.de

Abstract

The article presents a neural architecture for gesture-based interaction between a mobile robot and its user. One of the main goals of our research is the greatest possible robustness of the intelligent interface under highly varying environmental conditions. After a short introduction of our mobile robot, a discussion of a reliable operation scenario and the motivation for a gesture-based dialogue in section 1, in section 2 we consider related approaches from the literature. Section 3 gives an overview over the whole neural architecture and outlines *what we want*, not taking into account the current implementation state. *What we currently can* is described in section 4.

1 Introduction and operation scenario

Figure 1: *Our mobile robot MILVA. Provided with highly developed 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 gesture-based interaction with a user.*

Recently there have been strong efforts to develop intelligent, natural interfaces between users and systems which can be used easily and intuitively and which are able to substitute common interface devices (keybord, mouse, data glove etc.) and/or to extend their functionality.

The operational area of such intelligent interfaces covers a broad range of fields in which an arbitrary system is to be controlled by an external user or in which system and user have to interact immediately (see for instance [7, 1, 3, 20]).

Figure 1 shows our robot platform MILVA (Multisensoric Intelligent Learning Vehicle in a 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.

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

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

into account the capabilities of our robot which has no manipulators and *can only move itself*. Second, the scenario is to motivate a gesture-based dialogue between the user and the serving system naturally. 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.
From this scenario the following requirements can be derived, which determine the design of the neural architecture: All processing capabilities have to show their robust functionality under highly varying environmental conditions, which can neither be estimated nor be influenced. The gesture-based dialogue must be user-independent. The gestures to be used should be highly instructive, which means that everybody has to be able to understand as well as carry out the gestures. The whole system has to operate in real-time.

The development and test of our system cannot take place at a real railway station, but the indoor-circumstances of our lab are complicated and variable enough to serve as a reliable environment.

2 Related work

The work of Pentland's group at MIT (see for a survey [1]) have the greatest similarities with our work. The interaction between system and user takes place in a so called *smart room*, in which the user interacts with an *animated world* via a screen wall. The capabilities of the used visual interface include user's localization and tracking, the localization of user's head and hands (via an additional telecamera) as well as the detection and interpretation of gestures. However, in the smart room some assumptions can be made which simplify the different processing steps but which do not hold in our scenario. Neither do we have stationary background, which would simplify the detection of the user, nor we do have certain lighting conditions, which would simplify the color-based segmentation of skin regions (faces and hands).

The PERSEUS-architecture proposed in [3] was developed for a mobile robot in an ordinary lab environment to detect a user, to recognize a pointing gesture and to grasp an object the user is pointing to. The used object representations for ground (floor), background, lights and persons require certain assumptions concerning the environment and only one user may be in the scene.

Kortenkamp [20] uses a stereo-camera-system for the interaction between a mobile robot and a user via 3D-gesture recognition with coarse geometrical models for the upper part of the human body. These models are tracked in real-time. The system assumes static, predefined gestures (poses). Their detection associates different actions via a static input-output-mapping. There are no limitations concerning the environmental conditions.

3 Neural architecture for gesture-based human-machine-interaction

Figure 2 provides a coarse sketch of the whole neural architecture for gesture-based human-machine-interaction. The different components of the architecture will be described in the following sections.

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. 4 cue modules which are sensitive to *skin color*, *facial structure*,

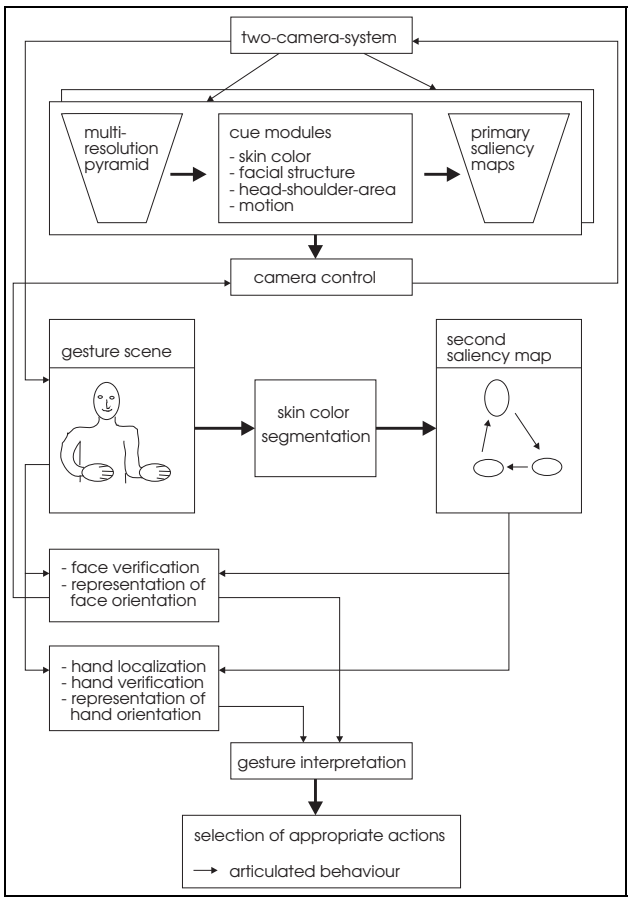

Figure 2: *Coarse sketch of the overall architecture*

structure of a head-shoulder-area 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.

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 organized neural fields containing dynamic neurons interacting with each other (see [15, 18, 19]). 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 fact that especially facial structure supplies a strong contribution to the primary saliency maps, we expect that faces become the most prominent or salient regions. The saliency map containing the overall most salient activity blob determines the further processing steps.

Control of the two-camera-system

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 gesture-camera is controlled such that the expected face region will appear on a predefined position in the image with an also predefined scale. The necessary distance estimation is provided by the cue modules detecting structural information (face or head-shoulder-area). The resulting *gesture-scene* is to contain all gesture-relevant parts of the intended user.

By means of the control of the gesture-camera we can assume 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.

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 organized neural fields containing dynamic neurons interacting with each other.

In order to simplify the task we use only the skin color information as the input for this

Figure 3: *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; the shown gesture examples suggest that the relevant information can be extracted only considering the orientation of one hand or both hands*

field, thereby assuming that the skin color segmentation is robust enough (see caption 4 for details).
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 regions of the hands should 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. To obtain this information, the assumed face region is analysed in detail by an additional module which merges the face verification and the estimation of the face orientation.

The output of this module consists of a continuous representation of the face (head) orientation. If there is no face at the assumed position, the orientation estimation 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.

If the orientation estimation (and therefore, the face verification) was successful, the following processing steps have to be gesture detection and gesture interpretation.

Detection and interpretation of gestures

Definition of a gesture set

For complexity reasons, we predefine a gesture alphabet and assume only static gestures or 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 [20]).

Further, we assume that the content of a gesture can be extracted only by using the orientation of one or both hands, whereas the whole configuration of face and hands is not important at the moment (see fig. 3). These restrictions are only introduced to handle the ongoing problems and they shall be put away step by step.

Localization of hands, representation of hand orientation – gesture recognition

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. So, similar to the detailed processing of the face region, we have to combine estimation of hand orientation and hand verification.

Here, too, we generate a continuous representation of the hand orientations for both hands, assuming that if no orientation can be estimated then the presently processed region does not contain a hand at all. As outlined before, the meaning of a certain gesture should only be determined by the orientation of one hand or of both hands (see also figure 3). Hence, for the present we avoid a detailed gesture modelling and can concentrate our efforts on the remaining problems which are sufficiently complex.

From gesture recognition to the generation of behaviour

The main direction of the research in our department concerns the organization of adaptive behaviour. A lot of projects deal with the different aspects of behavioural organization, such as direct mapping from sensory information to motor commands, organization of sensomotory representations, integration of different sensors, organization of reactive as well as globally planning behaviour and so on. Hence, the behavioural performance of the robot MILVA is to be extended step by step (see [17]).

One part of our GESTIK-project (see footnote of title) deals with the mapping from sensory situations to articulated behaviour. We utilize a dynamic neural field approach like that proposed in [26] to superimpose action proposals coming from different subsystems. These subsystems are responsible, for instance, for gesture interpretation, for local navigation or global planning, and selection of the appropriate action in a certain situation (see [21] for an extended description of this present work). The inputs to the neural field come from the local navigation system, from the global planner and from the gesture recognition system.
Currently, we investigate neural approaches for a direct mapping of sensory situations to appropriate actions such as proposed by Pomerleau [25]. As an extension of this work, we developed an multi-agent approach in which each neural agent is responsible for a certain action (moving straight forward, turning corners, passing a doorway etc.).

4 State of the implementation and preliminary results

Cue Modules

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 (and almost constant) lighting conditions.
The RGB-values are transformed into a physiologically motivated fundamental color space (see [8]), 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 [8]).
The color adaption described in [8] requires that the content of an image contains all or at least many different colors. This, in fact, generally cannot be provided by our indoor environment. Therefore, we want to extend the color adaptation in the following way: If a face region could be verified, its color values are projected into the fundamental color space. Then we calculate the relation of this cluster to the learned skin color model within the color space and carry out

a transformation moving the actual cluster into the learned model. Under the assumption that the color values are the same for both face and hand regions, this adaptation concerning the actual face is to improve the segmentation of the hand regions based on skin color (see [27], too).
The different color adaptation methods 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 images adapted by the method proposed in [8] because no face verification is available. Just after the first successful face verification, the extended adaptation method based on actual skin color can be used.

Figure 4: **Left:** *Distribution of the weight vectors of the GNG nodes within the two-dimensional input space (RG-YB-projection);* **Middle:** *input image;* **Right:** *segmentation result, whereby the darkest pixels mark the highest likelihood for skin color*

To estimate the likelihood of one pixel to be skin colored, we use an unsupervised Growing-Neural-Gas-Network (GNG, [9]). The GNG is trained with the manually extracted color values of our data set (see [24] for details, too) and serves as our skin color model. Figure 4 (left) outlines the position of the weight (reference) vectors of the GNG nodes in the input space. We utilize only the RG-YB-projection to become robust against differing light intensities. The disadvantage of such an approach is the increasing number of "false positive" segmented pixels, the advantage is that in almost all cases the "real" skin regions contained in an image can be correctly segmented (see [10] for a discussion of this problem). Because of the fact that we employ the statistics of the skin color distribution during training, in the area of highly frequent color values there is an increased density of weight vectors with a smaller variance (circles). Via a modified output function for the GNG nodes this effect is utilized to generate a higher skin color likelihood for color values activating GNG nodes having a small variance.
A very good skin color segmentation was achieved in the example of figure 4. Such results are not always possible but this is not necessary at all, because the skin color segmentation provides only *one* contribution to the localization process.

Facial structure
Because the distance between the camera and the user to be localized is not known, the detection of facial structure has to be carried out at each level of the multiresolution pramids (see also figure 2). In our scenario we assume that a person is an intended user if its face is oriented towards the camera and therefore towards the robot.
The detection of facial structure employs eigenfaces generated by a principal component analysis

(PCA) of the images contained in the ORL data set[1] (see also [11]). The approach proposed in [2], using an interpolated view method, was taken into account, too, but does not seem appropriate for our architecture because we require independence of a *certain* user. The image regions used for the PCA were extracted manually and cover a region of 15 x 15 pixels.
The regions were normalized by its mean and standard deviation (see also [12], [13]). The input image is processed with the 3 eigenfaces (with 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 MLP network performing a mapping from the fit values to 2 classes (face, no face). For the training of the MLP a data set of 100 positive (face) and 100 negative (no face) examples was created. To improve the generalization ability of the network we implemented a bootstrap algorithm [13] 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 such as proposed in [14] (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 [24]).

Figure 5: *Detection of faces using eigenvector masks. The detected faces are marked in the right image (likelihood higher than 0.7).*

The performance of the face detection is demonstrated in figure 5, where an image taken from [13] 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 mislocalizations can be rejected. Furthermore, faces which cannot be recognized by the human observer are not detected by our PCA method, either.

Combining skin color and facial structure

Our experience is that only the parallel utilization of different methods leads to appropriate localization 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) which can neither be influenced nor be estimated a priori, skin color segmentation and detection of facial structure provide more or less confident results. The uncertainties concerning skin color segmentation

[1] http://www.cam-orl.co.uk/face/database.html

Figure 6: *Skin color (middle) and face structure (right) hypotheses obtained from the input image (left), adapted by the method proposed in [8]; the used GNG consists of 49 nodes*

are demonstrated in figure 6 (middle). However, the detection of facial structure supplies a reliable result (right).

Head-shoulder-area

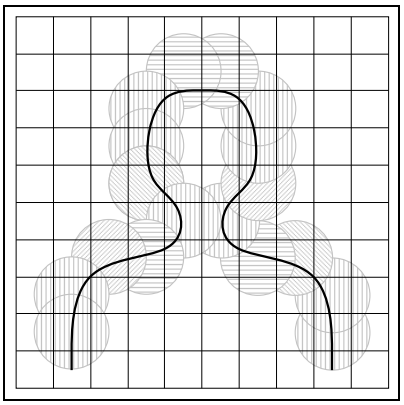

Figure 7: *Specially fitted grid of Gabor filters to detect a head-shoulder-area*

Similar to the detection of facial structure, the localization of a head-shoulder-area 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 7) and to classify the obtained filter outputs by a specially tuned distance measure between the actual filter outputs and a prototype.

Generation of saliency maps

The outputs of the cue modules supply the inputs for topographically organized dynamic neural fields, based on models proposed in [15], [16] or [18], respectively. Currently, the pyramid of primary saliency maps (see figure 2) is under construction.
As a preliminary result concerning the primary saliency maps, figure 8 shows the selection of the most likely head-shoulder-areas 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-area detector. Due tu the fact, that the head-shoulder-area detector supplies a strong output at adjacent levels of the pyramid, the selection becomes much more robust, and a lot of false positive detections can be avoided.
The same principle is to be extended to the whole saliency pyramid, inegrating all cue modules.

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

Control of the two-camera-system

First, a camera control module for a single camera was implemented based on a neural approach proposed in [28]. This module locates the target, estimates the size of the target and changes pan and tilt angles in such a way that the target emerges always on the same position within the image. A Kohonen-Map is used that learns an input-output-mapping between the actual

target position and the corresponding pan/tilt angles (see also [29]).
Currently this method is extended for the control of the two-camera system. The basic idea is that we assume a definite configuration of the cameras, which is necessary so that the mapping method used for the single camera can be employed for the two-camera-system, too. Therefore, after a possible user (face region) was located in one of the camera images, the corresponding camera is directed towards this user. This is realized by means of controlling the pan/tilt of this camera as well as the pan for both cameras (camera neck). Therefore, the initial camera configuration (especially the base distance) is always the same. So, we can train a Kohonen-Map to map the size and position of the target within the image of the general-view-camera to the corresponding pan/tilt/zoom parameters of the gesture-camera.

Face verification, estimation of face and hand orientation

The detailed analysis of faces and hands will be realized by a regular grid of Gabor filters and a following classification of the Gabor filter outputs with a neural classifier (see also [23, 10]). A similar approach was proposed in [22]. Here, the combination of Gabor filtering and PCA leads to a strong reduction of the pose space for faces. Currently, both methods are being implemented and investigated concerning their reliability for our architecture. For faces as well as for hands, we obtain continuous representations of the orientation. This kind of representation is necessary because we want to extend the currently used very simple gesture model into a more sophisticated one, involving the whole configuration of face and hands.

5 Conclusion and outlook

Currently we concentrate our efforts on the implementation of the pyramid containing the primary saliency maps. Only when the whole primary saliency system runs, we can estimate the sufficiency of the developed cue modules.
The cue module for motion analysis has to be realized. At the moment we have very limited practical experience concerning the detection and usage of motion, but there are some conceptional ideas to deal with the problem (see [4] for a survey of motion-based recognition). Our favoured approach was proposed by Pentland's group (see [5], [6]). Here, 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 seems to be 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 fields, different aspects of motion information could be obtained.

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

References

[1] **Darrell, T., Basu, S., Wren, C., Pentland, A. (1997).** Perceptually-driven Avatars and Interfaces: active methods for direct control. In: *MIT Media Lab Perc. Comp. Sec. TR No. 416*

[2] **Darrell, T., Essa, I., Pentland, A. (1994).** Correlation and Interpolation Networks for Real-time Expression Analysis/Synthesis. In: *Proceedings of NIPS'94*, MIT Media Lab. Perc. Comp. Sec. TR No. 284

[3] **Kahn, R.E. (1996).** PERSEUS: An Extensible Vision System for Human-Machine Interaction. In: *PhD-Thesis, University of Chicago, Dpt. of Computer Science*
[4] **Cedras, C., Shah, M. (1995).** Motion-Based Recognition: A Survey. In: *Image and Vision Computing 13(2)*, pp. 129-155
[5] **Bobick, A.F. (1996).** Computers Seeing Action. In: *Proc. of British Machine Vision Conference '97*, MIT Media Lab. Perc. Comp. Sec. TR No. 394
[6] **Davis, J.W., Bobick, A.F. (1996).** The Representation and Recognition of Action Using Temporal Templates. In: *MIT Media Lab. Perc. Comp. Sec. TR No. 402*
[7] **Maggioni, C. (1995).** GestureComputer - New Ways of Operating a Computer. In: *Int. Workshop on Automatic Face- and Gesture-Recognition, Zurich*, pp. 166-171
[8] **Pomierski, T., Gross, H.-M. (1996).** Biological Neural Architecture for Chromatic Adaptation Resulting in Constant Color Sensations. In: *Proc. of the Int. Conf. on Neural Networks (ICNN'96)*, pp. 734-739
[9] **Fritzke, B. (1995).** A Growing Neural Gas Network Learns Topologies. In: *Advances in Neural Information Processing Systems 7*, pp. 625-632
[10] **Littmann, E., Ritter, H. (1995).** Neural and Statistical Methods for Adaptive Color Segmentation - A Comparison. In: *Mustererkennung 1995*, pp. 84-93
[11] **Moghaddam, B., Pentland, A. (1995).** Maximum Likelihood Detection of Faces and Hands. In: *Proc. Int. Worksh. on Aut. Face- and Gesture-Recog.*, pp. 122-128
[12] **Schiele, B., Waibel, A. (1995).** Gaze Tracking Based on Face-Color. In: *Proc. Int. Worksh. on Aut. Face- and Gesture-Recog.*, pp. 344-349
[13] **Rowley, H. A., Baluja, S. Kanade, T. (1995).** Human Face Detection in Visual Scenes. In: *Technical Report CMU-CS-95-158R, Pittsburgh*
[14] **Sung, K.-K., Poggio, T. (1994).** Example-based Learning for View-based Human Face Detection. In: *Proc. Image Understanding Workshop*
[15] **Amari, S. (1977).** Dynamics of Pattern Formation in Lateral-Inhibition Type Neural Fields. In: *Biological Cybernetics No.27*, pp. 77-87
[16] **Gross, H.-M., Franke, R., Boehme, H.-J., Beck, C. (1992).** A Neural Network Hierarchy for Data and Knowledge Controlled Selective Visual Attention. In: *Proc. of ICANN'92, Brighton*, pp. 825-828
[17] **Gross, H.-M., Boehme, H.-J. (1997).** Verhaltensorganisation in interaktiven und lernfaehigen mobilen Systemen – das MILVA-Projekt. *this volume*
[18] **Kopecz, K. (1996).** Neural field dynamics provide robust control for attentional resources. In: *B. Mertsching (ed.), Aktives Sehen in technischen und biologischen Systemen*, Infix-Verlag, pp. 137-144
[19] **Stephan, V., Gross, H.-M. (1997).** Formerhaltende sequentielle visuelle Aufmerksamkeit mittels strukturierter neuronaler Felder. to be pblished in: *Mustererkennung '97, Braunschweig*
[20] **Kortenkamp, D., Huber, E., Bonasso, P.R. (1996).** Recognizing and interpreting gestures on a mobile robot. In: *Proc. of the Thirteenth Nat. Conf. on Art. Intell. (AAI-96*
[21] **Krabbes, M., H.-J. Boehme, V. Stephan, H.-M. Gross (1997).** Handlungsorganisation intentionaler neuronaler Agenten. *this volume*
[22] **Gong, S., McKenna, S., Collins, J. J. (1996).** An Investigation into Face Pose Distributions. In: *Proc. of the 2nd Int. Worksh. on Aut. Face- and Gesture-Recog., Killington*
[23] **Kubisch, R., Ritter, H. (1996).** Erkennung menschlicher Kopfhaltungen mittels kuenstlicher neuronaler Netze. In: *Mustererkennung '96*, pp. 109-117
[24] **Brakensiek, A., Braumann, U.-D., Boehme, H.-J., Rieck, C., Gross, H.-M. (1997).** Farb- und strukturbasierte, neuronale Verfahren zur Lokalisation von Gesichtern in Real-World-Szenen. to be published in: *Mustererkennung '97, Braunschweig*
[25] **Pomerleau, D.A. (1993).** Neural Network Perception for Mobile Robot Guidance. *Kluwer Academic Publishers*
[26] **Schoener, G., Engels, C. (1994).** Dynamic field architecture for autonomous systems. In: *Proc. of From Perception to Action conference*, pp. 242-253
[27] **Hunke M.H. (1994).** Locating and Tracking of Human Faces with Neural Networks. *CMU-CS-94-155*
[28] **Ritter, H., Martinetz, T., Schulten, K. (1991).** Neuronale Netze. *Addison-Wesley*
[29] **Kleemann, T. (1997).** Ein binokulares System fuer Szenenueberblicks- und Detailanalyse. *Diploma Thesis, Technical University of Ilmenau*