International Gesture Workshop Bielefeld 1997, pp. 219-232, Springer

Proc. Bielefeld Gesture Workshop, September 1997, pp. 219-232, IEEE Computer Society Press

Neural Architecture for Gesture-Based Human-Machine-Interaction *

Hans-Joachim Boehme, Anja Brakensiek, Ulf-Dietrich Braumann, Markus Krabbes and Horst-Michael Gross

Department of Neuroinformatics, Technical University of Ilmenau, D-98684 Ilmenau, Germany

Abstract. We present a neural architecture for gesture-based interaction between a mobile robot and human users. One crucial problem for natural interface techniques is the robustness under highly varying environmental conditions. Therefore, we propose a multiple cue approach for the localisation of a potential user in the operation field, followed by the aquisition and interpretaion of its gestural instructions. The whole approach is motivated in the context of a reliable operation scenario, but can be extended easily for other applications, such as videoconferencing.

1 Introduction and scenario

The field of intelligent interfaces covers a broad range of applications in which systems are to be controlled by an external user or in which system and user should immediately interact. 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 (see [8, 16, 20]).

For the mobile service robot domain this could be systems to support disabled people or driverless transport systems for industrial application. Especially for the interaction between a user and a mobile system the visual communication may be important to give the system the capability to observe its operational area in an active manner, whereas without visual communication the system has to wait passively for a message via a special input device.

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 observes actively its operational environment. In the following sections we will outline how the detection of intended users and the detection and interpretation of gestures are realized.

The use of our system as an intelligent luggage carrier, for instance at a railway station (see figure 1) or an airport, was chosen as a hypothetic scenario for the following reasons: First, we must take into account the capabilities of our

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

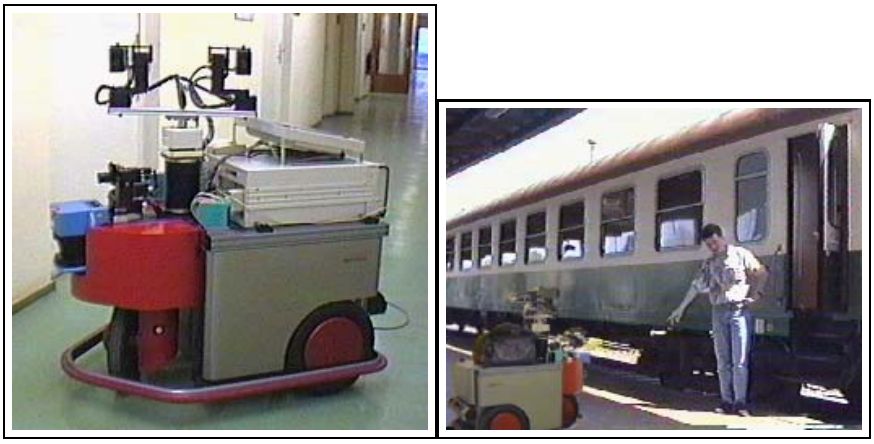

Fig. 1. Left: Our mobile robot MILVA. Provided with highly developed on-board equipment (68040-VME-system, 2 PC-systems, CNAPS-board, framegrabber) and different sensors (3 cameras, laserscanner, ultrasound and infrared distance measures, bumpers) MILVA serves as the testbed for the gesture-based interaction with a user.
Right: MILVA in human-machine-interaction at the Ilmenau railway station

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. On 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, that means that everybody has to be able to understand the gestures as well as to carry out the gestures.
- The whole system has to operate in real-time.

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

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

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

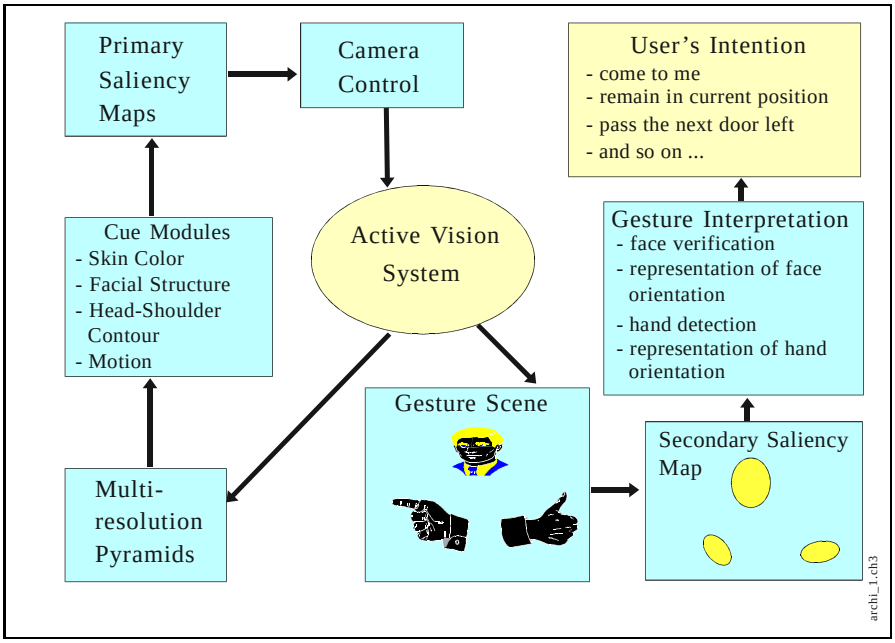

Fig. 2. Coarse sketch of the overall architecture

2.1 Cue modules

In the beginning both cameras of the two-camera-system operate in wide-angle-mode in order to cover the most possible area of the environment. Multiresolution pyramids transform the images into a multiscale representation. 4 cue modules sensitive for *skin color*, *facial structure*, *structure of a head-shoulder-contour* and *motion*, respectively, operate at all levels of the two pyramids.
The usage of the various in parallel processing cue modules is to make the whole system robust and more or less independent on the presence of *one certain* information source in the images.

2.2 Generation of primary saliency maps

The output of the cue modules serves as the input for the primary saliency maps at each level of each multiresolution pyramid. The maps are topographically organised neural fields containing dynamical neurons interacting with each other (see [1, 15]). 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. That saliency map containing the overall most salient activity blob determines the further processing steps.

2.3 Control of the two-camera-system

As soon as a possible user's face region is detected in one of the camera images, this camera serves as *general-view-camera*, whereas the second camera is assigned as *gesture-camera*. The gesture-camera is controlled in such a way that the expected face region will appear on a predefined position in the image as well as with predefined scale. The necessary distance estimation is provided by the cue moduls detecting structural information (face or head-shoulder-contour). 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 always have approximately the same size, so we do not have to ensure scale invariance by the further processing steps.

2.4 Generation of the secondary saliency map

For the gesture-scene a secondary saliency map is created, which determines the sequential processing of this scene. Similar to the primary saliency map we also utilize topographically organised neural fields of dynamical neurons.
Here, in particular we will use skin color and facial structure information as the input for this field.
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 analysed in detail (see the following section). The regions of the hands should become salient, too.

2.5 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 whether it is directed towards the robot. To get 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 turns back to the wide-angle-mode. If the orientation estimation (and therefore the face verification) was successful, the further processing steps have to be gesture detection and gesture interpretation.

2.6 Detection and interpretation of gestures

Definition of a gesture set
Because of 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 fig. 3 and [16]). These restrictions are only be made to handle the ongoing problems and they should be put away step by step.

Fig. 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

Localization of hands
Besides the face, in the secondary saliency map hand regions become prominent, mainly because of their skin color, but we do not know if 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.

Representation of hand orientation – gesture recognition
As well as for the head orientation, we generate a continuous representation of the hand orientations for both hands, assuming that if no orientation estimation is possible the just processed region does not cover a hand. For simplicity reasons, the meaning of a certain gesture should be determined by the orientation of one hand or of both hands (see also figure 3) and their spatial relation to the face.

2.7 From gesture recognition to the generation of behaviour

The main direction of the research in our department concerns the organisation of adaptive behaviour. A lot of projects deals with the different aspects of behavioural 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 global planning behaviour and so on. Hence, the behavioural performance of the robot MILVA is to be extended step by step (see [13]).

The second part of the GESTIK-project deals with the mapping from sensory situations to articulated behavior. Figure 4 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 [22], well known as a kind of imitation learning.

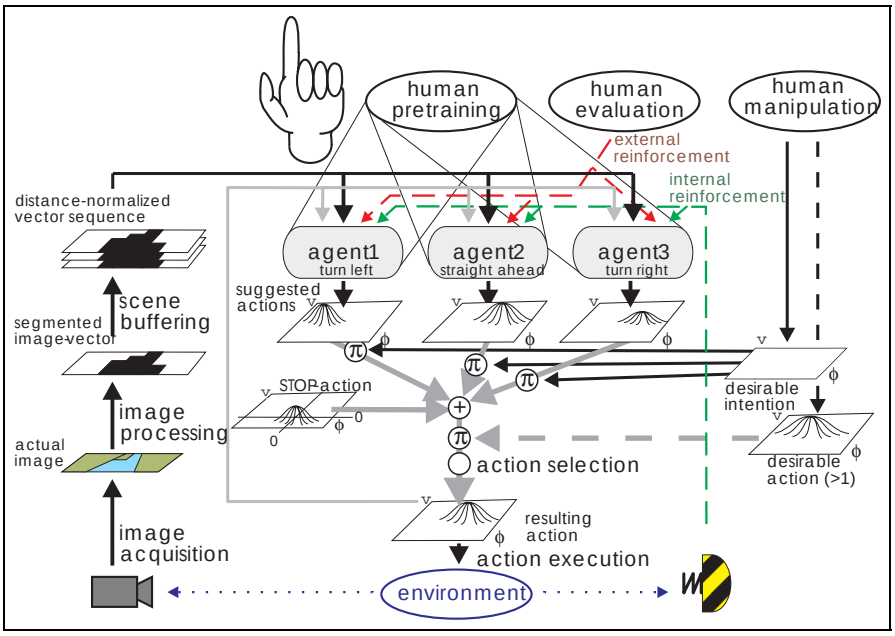

Fig. 4. Neural architecture for organisation of complex navigation behavior; see text for detailed explanation

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 4), we utilize a dynamic neural field approach, too. The selected action is assumed as the best one, corresponding to the actual sensory situation.

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 4). The desired robot behavior could be, for instance, "come to me", articulated via the "call for help" pose. Then this submitted intention leads to a modulation of the action selection process, such that the moving towards the user is supported. But 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 [17].

3 Close-ups of submodules and preliminary results

3.1 Cue modules

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 [3, 4, 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 (and almost constant) lighting conditions.
The RGB-values are transformed into a physiologically motivated fundamental color space (see [23]), 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 [23]).
However, the color adaption described there requires that the content of an image contains all or at least many different colors. This, in fact, cannot be provided by our indoor environment in general. Hence, we 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 according to the actual face is to improve the segmentation of the hand regions based on skin color (see [14], 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 [23] 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.

To estimate the likelihood of one pixel to be skin colored, we use an unsupervised Growing-Neural-Gas-Network (GNG, [10]). The GNG is trained with

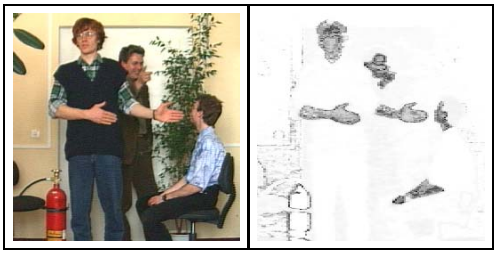

Fig. 5. *Left:* input image; *Right:* segmentation result, whereby the darkest pixels mark the highest likelihood for skin color

the manually extracted color values of our data set (see [5, 6] for details, too) and serves as our skin color model. In contrast to our previous work, wherein the WB-dimension of the color space was neglected to become robust against differing light intensities, now we utilize the whole color space. The drawback of this approach is the increasing number of GNG-neurons required to sample the color cluster in a sufficient manner. However, the advantage is that in almost all cases the "real" skin regions contained in an image can be segmented correctly (see [19] for a discussion of this problem). A detailed description of our skin color investigations can be found in [5] and [6]. 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. 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 5. Such results are not always possible but this is not necessary at all, because the skin color segmentation provides only *one* contribution to the localisation process.

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 pramids (see also figure 2). In our scenario we assume that a person is an possible user if its face is oriented towards the robot.
The detection of facial structure uses the grey 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 [21] and [5], too). The image regions used for the PCA were extracted manually, cover a region of 15×15 pixels, and the regions were normalised by their mean and standard deviation (see also [24, 25]). 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 GNG network performing a mapping from the fit values to 2 classes (face, no face). For the training of the GNG 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

[24] 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.

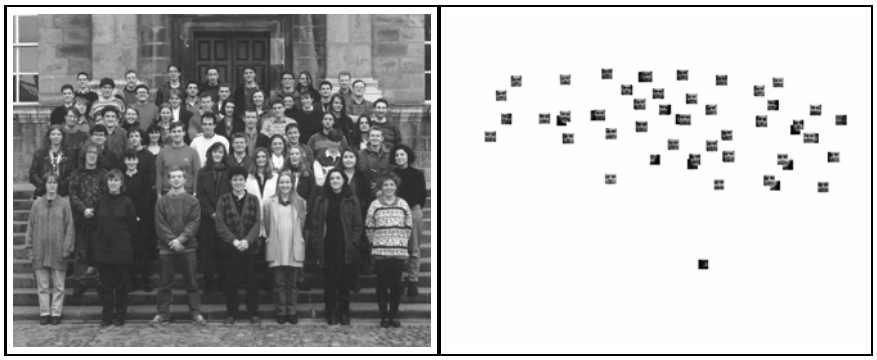

Fig. 6. Detection of faces using eigenvector masks. The detected faces are marked in the right image (likelihood higher than 0.7).

The performance of the face detection is demonstrated in figure 6, where an image taken from [24] was processed. False positive detected regions cannot be avoided entirely (right), but the utilization of the different cues will lead to reliable localisation results and therefore, such mislocalisations can be avoided.

c) Head-shoulder-contour
Similar to the detection of facial structure, the localisation of a head-shoulder-contour operates on the grey 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 pattern (see [7], too).

d) Motion
Our favoured approach was introduced by [2] and [9] and is demonstrated in figure 8. 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 fields, different task specific aspects of motion information can be obtained.

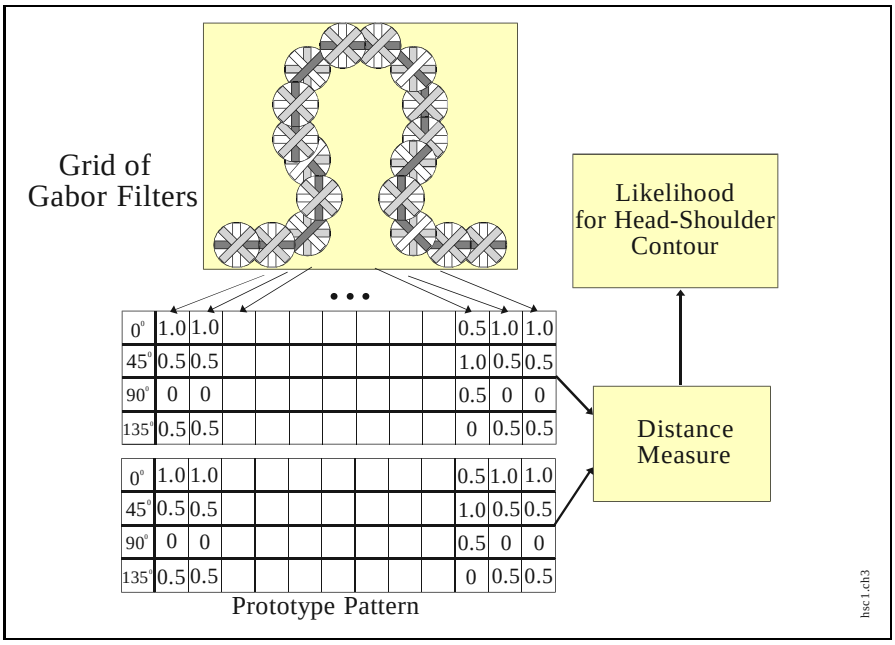

Fig. 7. Method for detection of a head-shoulder-contour, based on a specially arranged grid of Gabor filters and a task specific distance measure

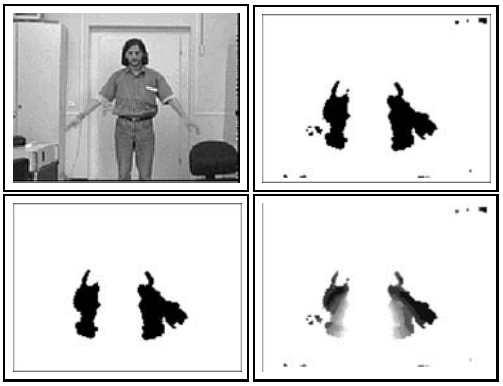

Fig. 8. Top Left: original image from the sequence; Top Right: binary motion image; Bottom Left: thresholded (concerning the size) binary motion image; Bottom Right: motion history image, where the latest motion results in the darkest image regions

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 mutually interacting dynamic neurons (see [1, 12, 15]).
The basic system equation is given as:

$$\tau \frac{d}{dt} z(\mathbf{r},t) = -z(\mathbf{r},t) - c_h h(t) + c_l \int_R w(\mathbf{r}-\mathbf{r}\,')y(\mathbf{r}\,',t)d^2\mathbf{r}\,' + c_i x(\mathbf{r},t) \qquad (1)$$

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

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

$$h(t) = \int_R y(\mathbf{r}\,'',t)d\mathbf{r}\,''$$

The activation state of the neuron at position $\mathbf{r}$ and timestep t is symbolized by $z(\mathbf{r},t)$. $h(t)$ describes a global inhibition function, $w(\mathbf{r}-\mathbf{r}\,')$ is a mexican-hat like interaction kernel, $y(\mathbf{r}\,',t)$ is the output activity of the neuron, $x(\mathbf{r},t)$ is the input, and c_i are only weighting coefficients. The given equations are prerequisite to implement a WTA-process.

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.

Currently, the pyramid containing the primary saliency maps (see figure 2) is under construction.
As a preliminary result concerning the primary saliency maps, figure 9 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 Face verification, estimation of face and hand orientation

The detailed analysis of faces and hands will be realized via a regular grid of Gabor filters and a following classification of the Gabor filter outputs with a neural classifier (see also [18, 19]). Furthermore, a very simple method for direct mapping of the image regions onto a continuous representation of 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.

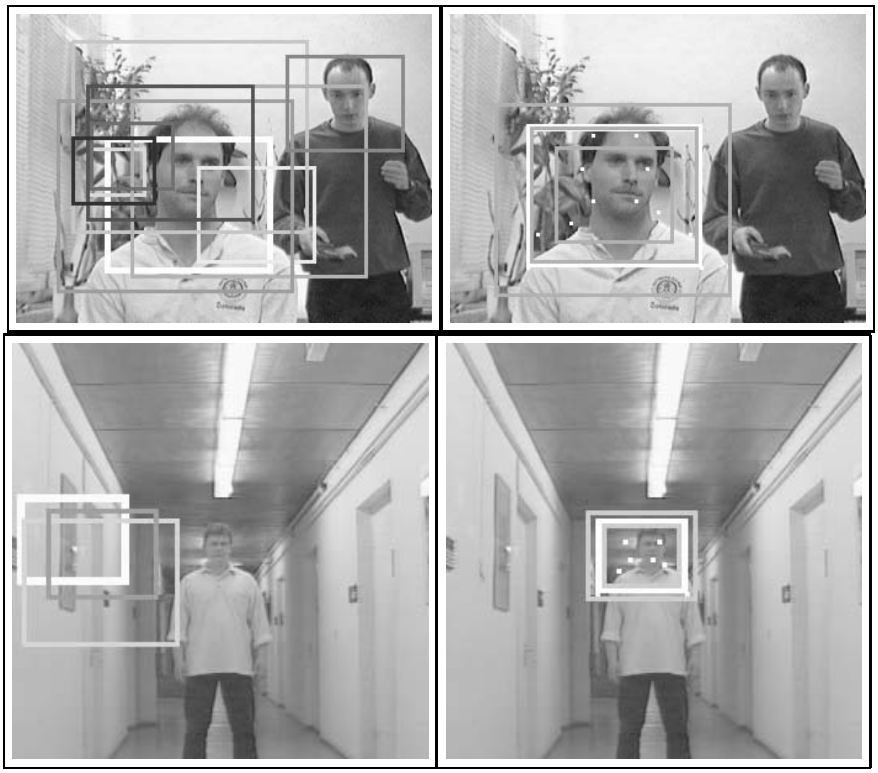

Fig. 9. Input images with marked head-shoulder-contours, obtained at the different levels of the multiresolution pyramid by the proposed method; the left images of each example show the result without dynamic selection, whereas the right images contain the selected contours obtained by means of dynamic selection (white rectangles mark the highest likelihood).

4 Summary and outlook

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.

At the moment, we concentrate on the final 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 for user localisation.

The represenation and interpretation of the aquired gestural instruction remains still an open problem. Our idea is to use a graph like structure for coding the relation between the detected face and hands, and to map this representation to the gestural instruction directly, which was shown by the examples of figure 3, too.

Furthermore, 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.

Acknowledgement

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

References

1. **Amari, S. (1977).** Dynamics of Pattern Formation in Lateral-Inhibition Type Neural Fields. In: *Biological Cybernetics No.27*, pp. 77-87
2. **Bobick, A.F. (1996).** Computers Seeing Action. In: *Proc. of British Machine Vision Conference '97*, MIT Media Lab. Perc. Comp. Sec. TR No. 394
3. **Boehme, H.-J., Brakensiek, A., Braumann, U.-D., Krabbes, M., Gross, H.-M. (1997).** Visually-Based Human-Machine-Interaction in a Neural Architecture. In: *Proc. SOAVE'97-Workshop, Ilmenau, Germany, VDI Verlag*, pp. 166-175
4. **Boehme, H.-J., Braumann, U.-D., Brakensiek, A., Gross, H.-M., Krabbes, M. (1998).** Neural Approach to Video-Based User Localisation for Human-Machine-Interaction. submitted to: *IJCNN'98*
5. **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. *Mustererkennung 1997, Braunschweig*, pp. 113-120
6. **Braumann, U.-D., Brakensiek, A., Boehme, H.-J., Gross, H.-M. (1997.)** Ein neuronales Verfahren zur Segmentierung hautfarbener Bildregionen. *3.Workshop Farbbildverarbeitung, Erlangen, Germany, IRB Verlag Stuttgart*, pp. 67-72
7. **Braumann, U.-D, Brakensiek, A., Boehme, H.-J., Gross, H.-M. (1998).** Gaborfilterbasierte visuelle Personenlokalisation mit dreidimensionalen Feldern dynamischer Neuronen. To appear in: Tagungsband des 3. Internationalen Workshops "Neuronale Netze in der Anwendung" (NN'98, Magdeburg)
8. **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*
9. **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*
10. **Fritzke, B. (1995).** A Growing Neural Gas Network Learns Topologies. In: *Advances in Neural Information Processing Systems 7*, pp. 625-632

11. **Goodridge, S.G. (1997.)** Multimedia Sensor Fusion for Intelligent Camera Control and Human-Computer-Interaction. *PhD Thesis, North Carolina State University*
12. **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
13. **Gross, H.-M., Boehme, H.-J. (1997).** Verhaltensorganisation in interaktiven und lernfaehigen mobilen Systemen – das MILVA-Projekt. *Proc. SOAVE'97-Workshop, Ilmenau, Germany, VDI Verlag*
14. **Hunke M.H. (1994).** Locating and Tracking of Human Faces with Neural Networks. *CMU-CS-94-155*
15. **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
16. **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)*
17. **Krabbes, M., H.-J. Boehme, V. Stephan, H.-M. Gross (1997).** Extension of the ALVINN-Architecture for Robust Visual Guidance of a Miniature Robot. *Proc. EUROBOT'97*, to appear
18. **Kubisch, R., Ritter, H. (1996).** Erkennung menschlicher Kopfhaltungen mittels kuenstlicher neuronaler Netze. In: *Mustererkennung 1996*, pp. 109-117
19. **Littmann, E., Ritter, H. (1995).** Neural and Statistical Methods for Adaptive Color Segmentation - A Comparison. In: *Mustererkennung 1995*, pp. 84-93
20. **Maggioni, C., Kaemmerer, B. (1996).** GestureComputer - History, Design, and Applications. In: *Proceedings of the Workshop on Computer Vision in Man-Machine Interfaces, R.Cippola (ed.), Cambridge University Press*
21. **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
22. **Pomerleau, D.A. (1993).** Neural Network Perception for Mobile Robot Guidance. *Kluwer Academic Publishers*
23. **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
24. **Rowley, H. A., Baluja, S. Kanade, T. (1995).** Human Face Detection in Visual Scenes. In: *Technical Report CMU-CS-95-158R, Pittsburgh*
25. **Schiele, B., Waibel, A. (1995).** Gaze Tracking Based on Face-Color. In: *Proc. Int. Worksh. on Aut. Face- and Gesture-Recog.*, pp. 344-349

This article was processed using the LaTeX macro package with LLNCS style