submitted to GW'99 - The 3rd Gesture Workshop,
Gif-sur-Yvette, France

Person Localization & Posture Recognition for Human-Robot Interaction

Hans-Joachim Boehme, Ulf-Dietrich Braumann, Andrea Corradini*, and Horst-Michael Gross

Department of Neuroinformatics, Technical University of Ilmenau,
D-98684 Ilmenau, Germany
email: {hans,ulf,andreac,homi}@informatik.tu-ilmenau.de

Abstract The development of a hybrid system for (mainly) gesture-based human-robot interaction is presented, thereby describing the progress in comparison to the work shown at the last gesture workshop (see [2]). The system makes use of standard image processing techniques as well as of neural information processing. The performance of our architecture includes the detection of a person as a potential user in an indoor environment, followed by the recognition of its gestural instructions. In this paper, we concentrate on two major mechanisms: (i), the contour-based person localization via a combination of steerable filters and three-dimensional dynamic neural fields, and (ii), our first experiences concerning the recognition of different instructional postures via a combination of statistical moments and neural classifiers.

Keywords: Human-Robot Interaction, Neural Networks, Dynamic Neural Fields

1 Introduction

Nowadays, taking into account the "come as you are" constraint, seeing and hearing machines offer a much broader range for natural human-machine interaction than common input devices like keyboard, mouse or data glove can do. In our group we are especially interested in novel techniques for interaction with mobile service systems in indoor environments. Such service systems should be able to observe their operation area in an active manner, to localize and contact a potential user, to interact with their users immediately and continuously, and to offer their services (transport, information presentation, or simply entertainment) in the context of the actual situation. Our robot platform MILVA (Multisensoric Intelligent Learning Vehicle in a neural Architecture) serves as the testbed for natural human-robot interaction. A two-camera system with 7 degrees of freedom (for each camera pan, tilt and zoom, additional pan for both cameras) will both capture the robot's environment and all interactional details expressed by persons. An additional camera in front of the robot provides the visual information for navigation.

Several systems for gesture-based human-machine interaction have been developed recently (e. g. see [5, 15, 16, 13, 18, 6]). A comprehensive collection of

* supported by the TMR Marie Curie Research Training Grant # ERB FMBI CT 97 2613

video-based gesture recognition systems can be found in [12]. Most of these approaches require certain constraints concerning the environmental conditions (lighting, distance between camera and person, etc.). During interaction with a mobile service system operating in an unconstrained indoor area one cannot assume such predefined circumstances. Therefore, the service system has to deal with highly varying environmental conditions which can neither be estimated nor be influenced. Taking into account this fact, we developed a robust saliency system for person localization (sec. 2). This saliency system integrates different visual cues into the localization process. Furthermore, acoustic information (estimation of source direction) is used to support the visual detection (see sec. 2.3).

After the detection of a person which is aligned towards the robot, a gesture recognition process is to be carried out to transmit the behavioral instructions from the user to the robot (sec. 3). Currently, we use a posture alphabet (see fig. 1), i.e. means we recognize a set of gestural symbols. In our future work we want to overcome this limitation and develop a system capable of continuously recognizing dynamic gestures.

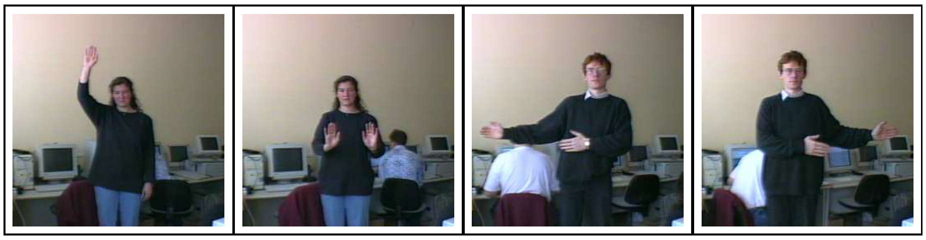

Figure 1. Gestures (postures) to be recognized; from left to right they carry the following meanings for the robot: come to me, stop, move left, move right

2 Saliency System for Person Localization

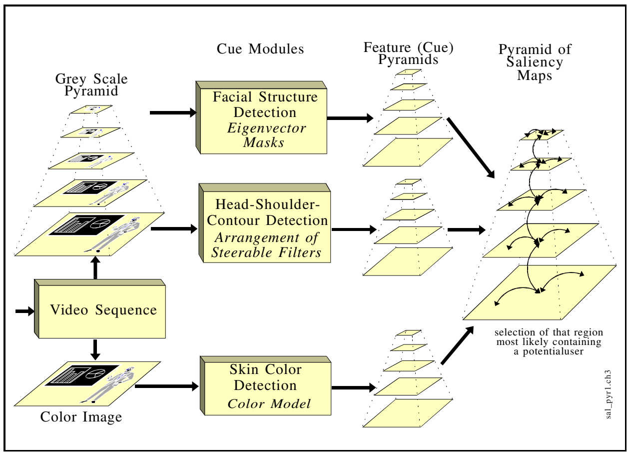

Figure 2. Components of the saliency system for person localization

Fig. 2 provides a coarse sketch of the saliency system for user localization. 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. Two cue modules sensitive to *facial structure* and *structure of a head-shoulder contour*, respectively, operate at all levels of a grayscale pyramid. That cue module processing for *skin color* detection uses the original color image. Its segmentation result is transformed into a pyramid representation, too, to obtain an uniform data structure for the different cues.

The utility of the different parallel processing cue modules is to make the saliency system robust and 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], too).

The output of the cue modules serves as the input for the saliency pyramid at each resolutional level. The maps are topographically organized neural fields containing dynamic neurons interacting among each other (see [1, 14]). In the saliency maps *all that regions* are to become prominent that cover most likely *the upper part of a person.*

2.1 Cues for Person Specific Saliency

In our previous work (see [2]) the three cues were assumed to be of equal importance. After a period of practical experiences we had to face that the shape-based approach supplies contribution to the localization process much more confident in contrast to the skin color and facial structure cues. The reasons are quite obvious: Skin color detection is highly influenced by illumination. Despite we use an additional color adaptation method (see [17]) to yield constant color sensation, robust skin color detection cannot ensured in general. Further, solving a localization problem is particularly of interest if a person is rather distant. Necessarily, relevant features should appear even on rather coarse resolutional scales so that details, as facial structures, are less appropriate. Facial structure can be detected confidently only if the distance between person and camera is not too large. Otherwise, the region covered by the face becomes to small to be localized.

Against this background, the method for head-shoulder contour detection was improved significantly. The actual method is described in more detail in the following subsection. Since the other cues can only support the person localization, but cannot ensure the localization alone, their methods were reduced to rather simple, but computationally efficient algorithms.

2.1.1 Head-Shoulder Contour

The contour which we refer to is that of the upper body of frontally aligned persons. Our simple contour shape prototype model consists of an arrangement of oriented filters doing a piecewise approximation of the upper shape (head, shoulder) of a frontally aligned person. The arrangement itself was learned based on a set of training images. Applying such filter arrangement in a multi-resolution manner, this leads to a robust localization of frontally aligned persons even in depth.

Arrangements of Steerable Filters – Motivation and Related Work The idea of this method refers just to a description of the outer shape of head and shoulders and is based both on some physiological considerations as well as on psychophysical effects.

The visual cortex consists in several parts of cells with oriented receptive fields. A lot of investigations have exposed that the profile of receptive fields of simple cells in the mammalian primary visual cortex can be modeled by some two-dimensional mathematical functions. Gaborian [10] and Gaussian (incl. low order derivatives) [11] appear to provide the typical profiles for visual receptive fields. So, local operations decomposing the visual information with respect to the frequency space are made.

Psychophysical aspects for the contour-shape based approach as, e. g., good continuation or symmetry (both belonging to the Gestalt laws), obviously describe effects which necessitate grouping mechanisms. Against this background, we conceptualized the approach of an *arrangement* of oriented filters.

Because each section of the contour should be approximated by a special oriented filter, localizing a person would require possibly as much *differently oriented* filters as orientations belong to the arrangement. Since it is computationally very costly we turn to use steerable filters.

Determining the Course of Contour Steerable filters have the nice property that an a priori limited number of convolutions is sufficient to derive any orientation information within an image. Thus, their use provides an extended set of orientations, avoids the necessity of numerous additional filters, and enables a more accurate computation of the course of contour.

Our complete data set consists of images showing ten persons in front of a homogeneous background under three different viewing angles ($0°, +10°$ and $-10°$, where $0°$ corresponds to an exactly frontal aligned face). All these images have been recorded on identic conditions (position, illumination, distance). Additionally, in order to have a symmetrical response the whole data set was vertically mirrored extending the data set to 60 images. Subsequently, the 256×256-images (grayscale) were low-pass filtered and scaled down resulting in the size 16×16. Then, we applied a Sobel operator to the images enhancing the edges of each image. Next, all of those edge-marked intermediate images were averaged, since the contour to be determined *on average* should match the real outer contour. After this we thresholded to find *that* edge representing the typical contour shape.

Now, we have the course of the contour of interest resulting in a 16×16 binary matrix where the elements along the contour are set to 1, the others remain 0. We refer to this contour matrix, our template, as $\boldsymbol{\Lambda}^{\star}$. The local orientation of each contour element is determined by means of the steerable filters (see below). These are applied to the binary contour shape so that for each element of $\boldsymbol{\Lambda}^{\star}$ an angle of orientation can be determined resulting in a matrix $\boldsymbol{\Lambda}$ (see fig. 3).

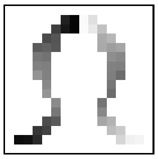

Figure 3. The determined shape of contour $\boldsymbol{\Lambda}$: orientation angles coded with gray values (0°: black; 90°: medium gray; 180°: white). Note that around the parting transitions 180° to 0° occur. The contour shape is symmetric since the original data set was mirrored.

Applying Steerable Filters After determining the binary contour, we measure the local orientation by means of a set of filters which are oriented in every direction. We take the powerful approach of *steerable filters* (see [7]) for orientation estimation. It provides an efficient filtering output by applying a few *basis filters* corresponding to a few angles and then interpolating the basis filter responses in the desired direction. Steerable filters are computationally efficient and do not suffer from the orientation selection problem.

In general, a function $f(\cdot)$ is considered to be steerable if the following two conditions are satisfied. First, its basis filter set is made up of their M rotated copies $f^{\alpha_1}(\cdot)\ldots f^{\alpha_M}(\cdot)$ on any certain angles $\alpha_1 \ldots \alpha_M$. Second, a rotated copy $f^{\vartheta}(\cdot)$ of it on some angle ϑ has to be obtained by a super-imposition of its basis set times interpolation functions $k_j(\vartheta)$ as

$$f^{\vartheta}(\cdot) = \sum_{j=1}^{M} k_j(\vartheta) f^{\alpha_j}(\cdot) \tag{1}$$

In our work we take a quadrature pair by using the second order derivative of a Gaussian and an approximation of its Hilbert transform by a third-order polynomial modulating a Gaussian. From the steering theorem [7] these functions are steerable and need $M = 7$ basis functions. To measure the orientation along the contour, we use the phase independent squared sum of the output of the quadrature pair. This squared response as a function of the filter orientation at a point (x, y) represents an *oriented energy* $E^{(x,y)}(\vartheta)$. Because of the symmetry of the functions, the energy at every pixel is periodic of period π. To accurately estimate the *dominant* local orientation one could *pointwise* maximize the orientation energy by taking $\vartheta_{MAX}^{(x,y)} = \arg\max\{E^{(x,y)}(\vartheta) \mid \vartheta \in [0, \pi)\}$. However, to find this maximum value we do not search degree-wise for the maximum because there already exists an analytical solution for the maximization [7]. We further refer to the matrix of all these angular values $\vartheta_{MAX}^{(x,y)}$ corresponding to the image as $\boldsymbol{\Theta}$. Besides, there exists a separable basis set in Cartesian coordinates which considerably lowers the computational costs.

Computing the Neural Field Input The previous section describes the theory and use of steerable filters. By means of those filters we calculate both the matrix $\boldsymbol{\Lambda}$ describing a typical course of the head-shoulder-portrait and that matrix $\boldsymbol{\Theta}$ (corresponding to the image wherein a person is to be found) containing the dominant local orientation values.

Subsequently, we are going to search for the presence of the *visual cue* head-shoulder-portrait, represented by the kernel $\boldsymbol{\Lambda}$, within the matrix $\boldsymbol{\Theta}$. To do this, we utilize a matching technique based on a *similarity measure* $m^{(x,y)}$. Due to the π-periodicity of the outcome of the steerable filters and in order to properly describe the likeness between two elements of $\boldsymbol{\Lambda}$ and $\boldsymbol{\Theta}$, the similarity function requires the same periodicity.

$$m^{(x,y)} = \frac{\sum_{i=0}^{I-1} \sum_{\substack{j=0 \\ \lambda_{i,j} \neq 0}}^{J-1} \frac{1}{2} \left[\cos\left(2 \left| \lambda_{i,j} - \vartheta_{MAX}^{(x+i-\frac{I}{2}, y+j-\frac{J}{2})} \right| \right) + 1 \right]}{\operatorname{card}\left(\operatorname{supp}\left(\boldsymbol{\Lambda}\right)\right)} \qquad (2)$$

Herein, $\lambda_{i,j}$ refers to the element of $\boldsymbol{\Lambda}$ at position (i, j) and $\vartheta_{MAX}^{(x+i-\frac{I}{2}, y+j-\frac{J}{2})}$ to the one of $\boldsymbol{\Theta}$ at $(x + i - \frac{I}{2}, y + j - \frac{J}{2})$. $I = J = 16$ represent the dimensions of the matrix $\boldsymbol{\Lambda}$. The normalization to the cardinality of the support of $\boldsymbol{\Lambda}$ (the support of a matrix considers only nonzero elements) ensures $m^{(x,y)} \in [0, 1]$ for the further processing. Fig. 4 summarizes the processing steps.

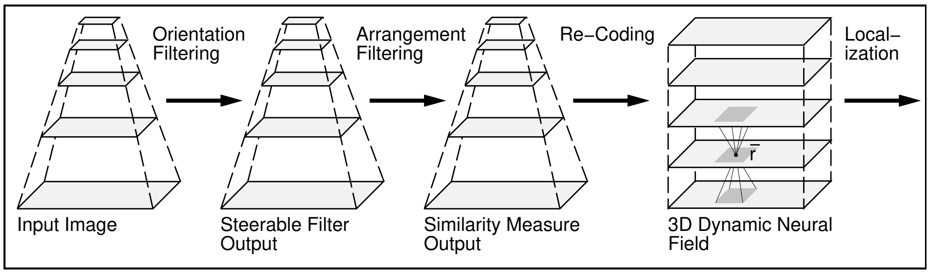

Figure 4. Processing steps for person localization via head-shoulder contour detection. Starting from a multi-resolution representation of the image, each level is treated by steerable filters. Applying the filter arrangement we determine a distance measure which is taken as input of a three-dimensional field of dynamic neurons. The resulting blob (locally delimited pattern of active neurons) is used to localize a person.

2.1.2 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). The skin color detection uses the original color image. 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 [17]). Then, we return into the RGB color space, use the chromatic projection $r = \frac{R}{R+G+B}$ and $g = \frac{G}{R+G+B}$, and define a bimodal 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, after a person (face region) could be localized, 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 of special importance because the hand regions cannot be segmented by structural information (see [9], and sec. 3.1). A Mahalanobis-based distance measure is employed to compute the similarity between the color value of each pixel and the color model. To achieve an appropriate input for the 3D dynamic neural field the resulting similarity map is recoded into an activity map, where the highest activity stands for the highest similarity. A more detailed description of our skin color investigations can be found in [2].

2.1.3 Facial Structure

We assume that a person can considered to be a user in those cases if its face is oriented towards the robot.
In our previous work, the detection of facial structure employed eigenfaces (see [2, 4]). The disadvantage of that method is their computational complexity, resulting in time consuming calculations. Due to real-time constraints a new, similar method was implemented. First, a prototype (mean) pattern of a frontally aligned face (15 x 15 pixels) was created by means of the images contained in the ORL data set (`http://www.cam-orl.co.uk/facedatabase.html`). Then we calculate the similarity between each image region and the prototype pattern via normalized convolution. The higher the convolution result, the higher the similarity, and the convolution result can be used directly as the input for the saliency pyramid.

2.2 The Saliency Pyramid as a 3D Nonlinear Dynamic Field

To achieve a good localization, a *selection mechanism* is needed to make one definite choice among those regions within the pyramid where rather high similarity measures concerning the different cues are heaped. Since dynamic neural fields are powerful for dynamic selection and pattern formation using simple homogeneous internal interaction rules, we adapted them for our purposes. Because we use five fine to coarse resolutions in our scale space (see fig. 2) we actually can localize persons even in different distances. Therefore, a neural field for selecting the most salient region should be three-dimensional. That field F can be described as recurrent nonlinear dynamic system. Regarding to the selection task, we need a dynamic behavior which leads to *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 equation describes the system:

$$\tau \frac{\mathrm{d}}{\mathrm{d}t} z(\boldsymbol{r},t) = -z(\boldsymbol{r},t) - \mathrm{c}_h h(t) + \mathrm{c}_i x(\boldsymbol{r},t) + \mathrm{c}_l \int_N w(\boldsymbol{r}-\boldsymbol{r}\,')y(\boldsymbol{r}\,',t)\mathrm{d}^3\boldsymbol{r}\,' \qquad (3)$$

Herein $\boldsymbol{r}$ denotes the three-dimensional coordinate of a neuron position in the field, $z(\boldsymbol{r},t)$ is the activation of a neuron $\boldsymbol{r}$ at time t, $y(\boldsymbol{r},t)$ is the output activity of this neuron computed as a sigmoidal function of $\boldsymbol{r}$ alone, $x(\boldsymbol{r},t)$ denotes the external inputs (corresponding to the re-coded similarity measures for the different cues, combined by a Min-Max fuzzy operator), $h(t)$ is the global inhibition at time t gathering the activity from each neuron over the entire field $F \subseteq \mathbb{R}^3$. $w(\boldsymbol{r}-\boldsymbol{r}')$ denotes the Mexican-hat-like function of lateral activation of neuron $\boldsymbol{r}$ from the surrounding neighborhood $N \subseteq \mathbb{R}^3$. For one $\boldsymbol{r}$, N is symbolically marked as dark regions in fig. 4 (right). The constants c_h, c_l and c_i represent parameters of the system.

As also illustrated in fig. 4, to use a three-dimensional neural field, we have to consider the local correspondences between the resolution levels. Therefore, we do a re-coding into a cuboid structure. One side effect is that the coarser a pyramid level is the less we can locate something by means of the similarity measure. However, without particularly treating this effect we just noticed that those levels z of the neural field activated from the rather coarse pyramid levels take little more steps to develop a blob (or a part of a blob, respectively).

Results for Person Localization The results of the saliency system are qualitatively illustrated in fig. 5.

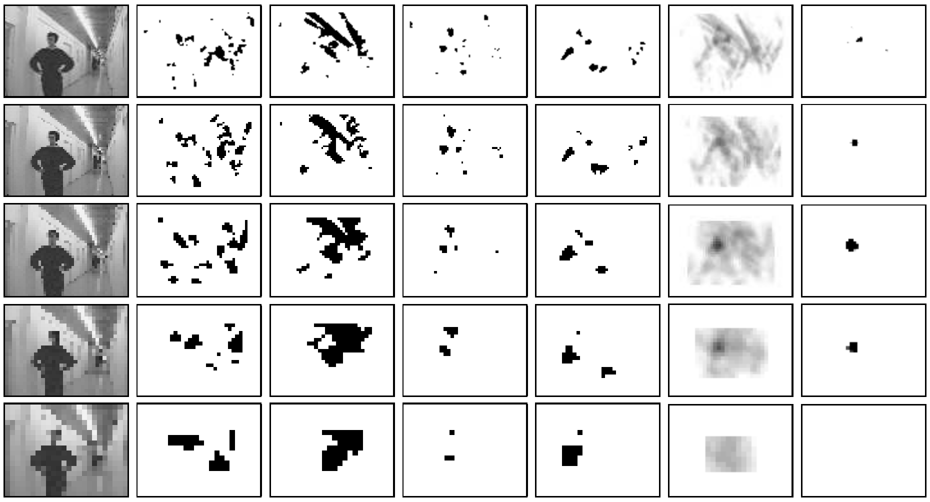

Figure 5. Localization results in an indoor environment: The localization of a person occurs not sharply at one of the pyramidal planes, the originating spatial blob (rightmost column) is most strongly developed on the central of the five planes. Each row contains the results of one of the five (distance $1/\sqrt{2}$) computed resolution steps. The seven columns depict the following: input, results of the orientation filtering for selected angles $0°$, $45°$, $90°$ and $135°$, the result of the filtering with the filter arrangement and finally the result of the selection within a three-dimensional field of dynamic neurons.

The images of the rightmost column show the state of the five layers of the dynamic neural field in a snapshot at that moment when the activity change of the most active neuron became less than 1%. On average, the system takes 11 iteration steps using a time-discrete Euler method. The range of the blob is not restricted to one plane. To get a more precise specification of the distance of a person one could interpolate the z-coordinate of the blob center within the field.

Our presented results are exemplary, the usage of the shape of contour and the additional cues skin color and facial structure provide a robust solution for the person localization problem, even under quite different conditions. Other results unfortunately cannot be shown here for briefness reasons. The novel approach with a three-dimensional dynamic neural field can be assessed as robust method for the selection process.

2.3 Auditory Saliency

Additionally to the visually-based saliency system a model for selective auditory attention was developed in our department (see [19]). This model was already implemented on MILVA and is to support the user localization. For example, the user can attract MILVA's attention by clapping her hands, i.e. MILVA will align her active-vision system towards that direction in which an auditory signal source was recognized.

3 Posture Recognition

In this section we describe the processing steps to be carried out to recognize the postures shown in fig. 1. The first step consists of a camera control procedure. The second camera of the active vision head is aligned towards the selected person and acquires the "posture images". An additional zoom control ensures that the person emerges in an approximately constant, predefined scale.

3.1 Posture Segmentation

The segmentation of face and hands as the gesture relevant parts is exclusively based on skin color processing. From the face region we take color values to construct a specific color model for the skin type and the illumination at hand. Via a simple distance measure (Mahalanobis based) each pixel is classified to be a member of the skin class or not (see fig. 6) to obtain a binarized image.

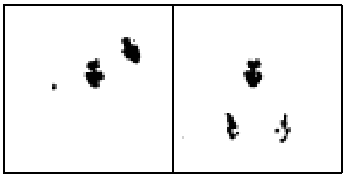

Figure 6. Skin color segmented "posture images"

3.2 Moment-based Posture Description

From the skin color segmented posture image, sub-sampled to a dimension of 64 times 64 pixels, we compute a feature vector $\boldsymbol{v}$ containing 9 statistical moments (central moments). For these moments the corresponding equation is given by equ. 4, a graphical illustration can be found in fig. 7.

$$\mu_{pq} = \sum_{x,y} (x - \bar{x})^p \cdot (y - \bar{y})^q \cdot f(x,y) \tag{4}$$

Herein μ_{pq} denotes the moment, x and y are the image coordinates, $\bar{x}$ and $\bar{y}$ describe the center of gravity. $f(x,y)$ is a Gaussian-like weighting function.

Before classifying the feature vectors into the four posture classes, in the next step we investigated the alteration of the feature vectors when the person (the posture, respectively) is slightly shifted. The result for one posture image is shown in fig. 8.

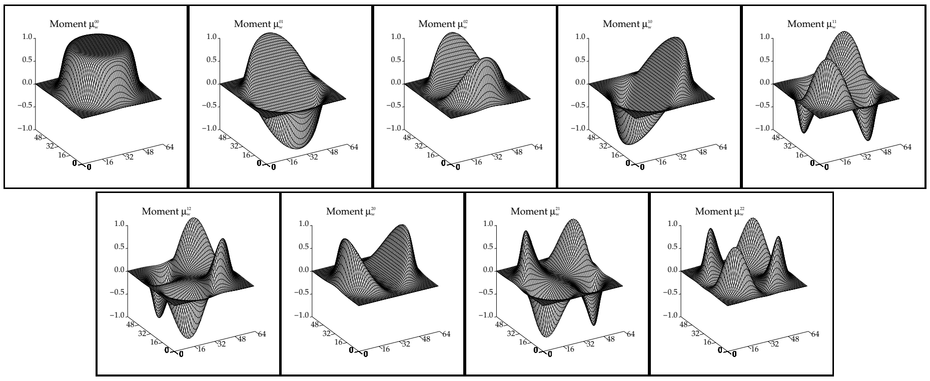

Figure 7. Graphical illustration of the moments used for the description of the postures; the moments were computed from an image of equal gray level; the sketch of moment μ_{oo} shows clearly the shape of the weighting function $f(x, y)$

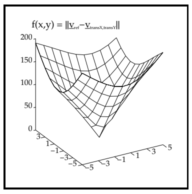

Figure 8. Fluctuation of the feature vector $\boldsymbol{v}$ depending on the deviation of the same posture. We can see, that the Euclidean distance between the reference feature vector $\boldsymbol{v}_{\mathrm{ref}}$, obtained at the central position, and the feature vectors calculated under deviations up to 5 pixels, grows smoothly. The shown distance values up to 200 are small in relation to the number of 9 moments and their value intervals.

3.3 Posture Recognition with Neural Classifiers

For the training of the neural classifiers we used a data set containing 360 feature vectors. These vectors were computed from the four postures, 90 examples of each. We used 180 vectors for the training and 180 vectors for the test of the networks.

First, we used a Radial Basis Function (RBF) network containing 9 input nodes, 20 hidden (RBF) nodes, and 4 output nodes. The RBF layer was trained first with a Neural Gas algorithm to approximate the input data distribution. Then, the second weight layer was trained via the standard delta rule.

Second, a modified Counterpropagation (CP) network was employed for posture recognition. The network had the same topology as the RBF network. The hidden layer was trained first with a Neural Gas algorithm, too. Then, the second weight layer was trained by a learning rule similar to Grossberg's outstar model (see [8]).

The table summarizes the achieved performance concerning the two networks. The RBF network yielded robust performance, and the number of false classified patterns was rather slow, whereas the CP network suffers from a large number of misclassifications and additionally from a slow recognition rate. Concluding this results, currently we implement the RBF based approach on our mobile robot.

Our future work will concern a dynamic approach for continuous gesture recognition. More precisely, we try to describe different space-time gestures via

Network	Topology	# of Trainingspatterns	# of Testpatterns	# of false classified patterns	# of unclassified patterns	Recognition Rate
RBF	9-20-4	180	180	6	10	91.2 %
CP	9-20-4	180	180	40	48	75.6 %

the observed trajectory in the moment feature space. The more crucial problem consists in the "behavioral grounding" of such dynamic gestures, whereas the correspondence between the postures used in our example and the behavioral meanings for the robot is rather simple. One possible way could be the parallel utilization of speech and gesture to find out coherences between these two information channels in order to teach the robot to use gesticulated or spoken commands alternatively.

4 Conclusion and Outlook

The results concerning person localization by means of the usage of multiple cues and their integration into a selection process via 3D dynamic neural fields led to a robust person specific saliency system. Furthermore, the integration of auditory saliency makes it easy for the user to attract the attention of the robot.

The work for posture and gesture recognition is still ongoing. Therefore, the presented approach is just the begin of the investigations. Nevertheless, the already implemented method is appropriate to transmit several gesticulated commands to the robot, but is, of coarse, still far away from really natural human-robot interaction.

References

1. Amari, S. Dynamics of pattern formation in lateral-inhibition type neural fields. *Biological Cybernetics*, 27:77–87, 1977.
2. Boehme, H.-J., Brakensiek, A., Braumann, U.-D., Krabbes, M., and Gross, H.-M. Neural Architecture for Gesture-Based Human-Machine-Interaction. In *Gesture and Sign-Language in Human-Computer Interaction*, pages 219–232, 1998.
3. Boehme, H.-J., Braumann, U.-D., Brakensiek, A., Krabbes, M., Corradini, A., and Gross, H.-M. Neural Networks for Gesture-based Remote Control of a Mobile Robot. In *International Joint Conference on Neural Networks*. IEEE, 1998.
4. 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 *Proceedings International Conference on Automatic Face- and Gesture Recognition*, pages 486–491. IEEE Computer Society, 1998.
5. Darrell, T., Basu, S., Wren, C., and Pentland, A. Perceptually-driven Avatars and Interfaces: active methods for direct control. In *SIGGRAPH'97*, 1997. MIT Media Lab Perceptual Computation Section, TR 416.
6. Fjeld, M., Bichsel, M., and Rauterberg, M. BUILD-IT: An Intuitive Design Tool Based on Direct Object Manipulation. In *Gesture and Sign-Language in Human-Computer Interaction*, pages 297–308, 1998.

7. Freeman, W.T. and Adelson, E.H. The Design and Use of Steerable Filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 13(9):891–906, 1991.
8. Grossberg, S. Some Networks That Can Learn, Remember, and Reproduce Any Number of Complicated Space-Time Patterns. *Journal of Mathematics and Mechanics*, 19(1):53–99, 1969.
9. Hunke, M.H. Locating and Tracking of Human Faces with Neural Networks. Technical report, Carnegie Mellon University, 1994.
10. Jones, J.P. and Palmer, L.A. An Evaluation of the Two-Dimensional Gabor Filter Model of Simple Receptive Fields in Cat Striate Cortex. *Journal of Neurophysiology*, 56(8):1233–1258, 1987.
11. Koenderink, J.J. and van Doorn, A.J. Receptive Field Families. *Biological Cybernetics*, 63:291–297, 1990.
12. Kohler, M. Vision Based Gesture Recognition Systems. `http://ls7-www.informatik.uni-dortmund.de/html/englisch/gesture/vbgr-table.html`.
13. Kohler, M. Special Topics of Gesture Recognition Applied in Intelligent Home Environments. In *Gesture and Sign-Language in Human-Computer Interaction*, pages 285–296, 1998.
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 (AAAI-96)*, 1996.
16. Maggioni, C. and Kämmerer, B. GestureComputer – History, Design, and Applications. In *Proceedings of the Workshop on Computer Vision in Man-Machine Interfaces*, 1996.
17. Pomierski, T. and Gross, H.-M. Biological Neural Architectures for Chromatic Adaptation resulting in Constant Color Sensations. In *ICNN'96, IEEE International Conference on Neural Networks*, pages 734–739. IEEE Press, 1996.
18. Rigoll, G., Kosmala, A., and Eickeler, S. High Performance Real-Time Gesture Recognition Using Hidden Markov Models. In *Gesture and Sign-Language in Human-Computer Interaction*, pages 69–80, 1998.
19. Zahn, T., Izak, R., Trott, T., and Paschke, P. A paced analog silicon model of auditory attention. In *Neuromorphic Systems: Engineering Silicon from Neurobiology. 1st European Workshop on Neuromorphic Systems*, pages 99–112. World Scientific, 1997.