User Localisation for Visually-based Human-Machine-Interaction

Hans-Joachim Boehme, Ulf-Dietrich Braumann, Anja Brakensiek*, Andrea Corradini**, Markus Krabbes*, Horst-Michael Gross
Department of Neuroinformatics, Technical University Ilmenau, Fed. Rep. of Germany

*Supported by the Thuringian Ministry of Science, Research and Culture (TMWFK, GESTIK-Project). **Supported by the TMR Marie Curie Research Training Grant (EU) # ERB FMBI CT 97 2613

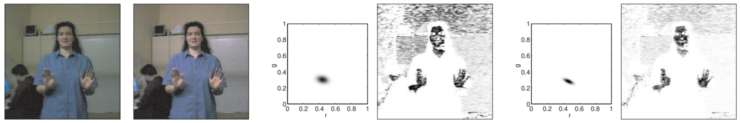
Presented at the Third International Conference on Automatic Face and Gesture Recognition (FG'98), April 14 - 16, 1998, Nara, Japan

Skin Colour

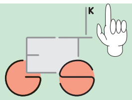
For the generation of a skin colour training data set, portrait images of different persons under appropriate lighting conditions were manually segmented. The excursion of the cluster formed by the colour values of the image within a physiologically motivated colour space (see [15]) corresponds to the deviation in illumination from typical daylight. By a *colour adaptation* process, the cluster is transformed so that its elongation will be along the achromatic axis. So we can ensure equal colour sensations under different lighting conditions (see 1st and 2nd figure below). The skin colour classification takes place within the *intensity normalised* RGB-colour-space (r, g). To model the skin colour distribution, we define a two-dimensional Gaussian by calculating mean and covariance of the skin colour data set (3rd and 4th fig.), see [10,17].

Colour Model Adaptation

If a face region could be verified by other cues, a new Gaussian model is created more specific for the current illumination and skin type. So the detection of skin coloured regions, especially of hands, can be improved (5th and 6th fig.).



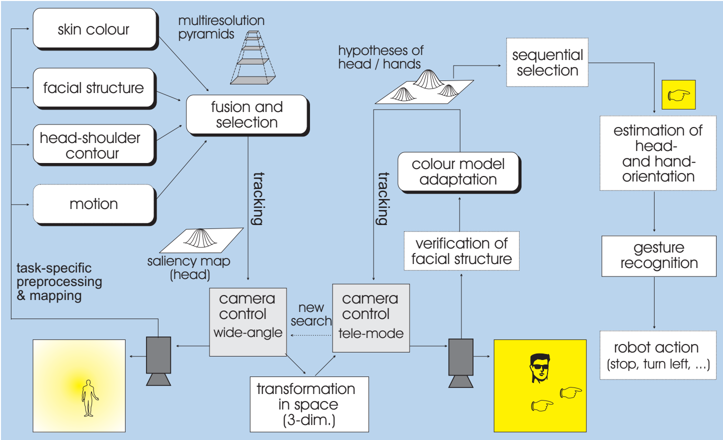
Scenario / Abstract



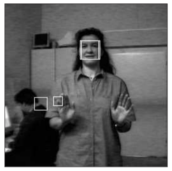
We present a neural architecture for user localisation, embedded within a complex system for visually-based human-machine-interaction (HMI). User's localisation is an absolute prerequisite to video-based HMI, see [5,6,8,12].
Due to the main objective, the *highest possible robustness of the localisation* under widely varying environmental conditions, we propose a *multiple cue approach*. It treats the features facial structure, head-shoulder contour, skin colour, and motion within a multiscale representation. The following selection of that image region most likely containing a user is realised by a WTA-process within the multiscale representation.
The two-camera system of our robot MILVA 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 environment.
The use of our system as an intelligent luggage carrier, for instance at railway stations or airports, was chosen as a hypothetic scenario for the following reasons: First, we must take into account the capabilities of our robot which does not have manipulators and can only move itself. Second, that scenario naturally motivates a *gesture-based dialogue ('hello', 'stop', 'turn left', ...)* between the user and the serving system [3,4].

Architecture

Initially, both cameras of the two-camera system operate in wide-angle mode in order to cover the broadest possible area of the environment. Multiresolution pyramids transform the images into a multiscale representation. Four cue modules which are sensitive to *skin colour*, *facial structure*, *structure of a head-shoulder contour* and *motion* [2] respectively, operate at all levels of the pyramids. Dashed boxes depict those modules currently under construction. All other parts have been realised.



Facial Structure



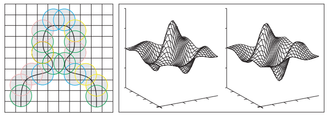
The image regions used for the *PCA* were manually extracted, cover a region of 15 x 15 pixels, and were normalised by their mean and standard deviation, see also [13,16].
Then, the input image is processed with 3 eigenfaces (according to the largest eigenvalues). The classification of the obtained fit values is realised by a supervised *Growing-Neural-Gas-Network* (GNG, see [7]). For the training of the GNG a data set of face and non-face (bootstrapping) examples was created.

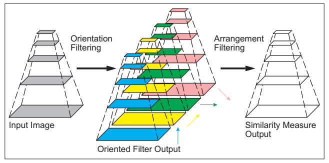
Head-Shoulder Contour

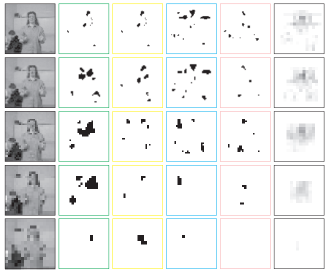
The idea is to use a spatial arrangement of oriented filters representing a special contour model, see [14]. Since it can be supposed that a user is frontally aligned towards the robot, the outer contour shape of head and shoulders can serve as another visual cue to localise the user. Each oriented filter (e.g. Gabor filter, steerable filter) is approximating one line element of the contour (see sketches right). The depicted Gabor filters exemplarily show a quadrature pair oriented at 135°. Just altogether four filters already suffice to made up a simple but effective contour model. To get a more precise model using more different orientations it is much more efficient to use steerable instead of Gabor filters.
To remain flexible concerning the distance of the person we expand the grey scale image into a multiscale representation, which enables a robust localisation of frontally aligned persons even in depth.
According to the righthand side sketch, at each resolutional level the oriented filters are applied doing a decomposition of the visual stimuli with respect to the frequency space. These certain directional pyramids are then combined at each level by applying the contour model Λ (arrangement filtering) onto corresponding subregions centered at the coordinates (x,y). Thus, we can determine a *local similarity measure* $m^{(x,y)}$ for the existence of persons based on local orientational measurements Θ_{MAX}:

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

This similarity measure is symmetric arround 0 and π-periodic.
The right grid of figures correspondingly to the above sketch gives an example how an input image is processed. Each column gives the content of one pyramid: input image, orientational filter results 0°, 45°, 90°, 135°, similarity measure.

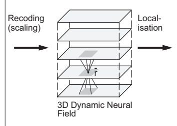






Fusion and Selection

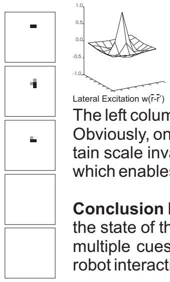
The output of the cue modules serves as the input for the primary saliency maps at each level of the multiresolution pyramid. The maps are topographically organised neural fields containing mutually interacting dynamic neurons, see [1,9,11]. In those primary saliency maps *all these regions* are to become prominent that cover *gesture-relevant* parts such as faces and hands. Because of the features *facial structure* and *head-shoulder contour*, faces will become the most prominent or salient regions. *That* saliency map containing the overall highest local activity determines the further processing steps. The fusion and selection of the maps is implicitly done by means of one three-dimensional neural field. To ensure the local correspondencies one must recode the pyramidal representation into a cuboid structure.



The dynamic neural field can be described as a non-linear differential equational system. What is of interest is the blob of remaining active neurons in an equilibrium state of the field. An equation of the system is of the following type:

$$\tau \frac{d}{dt} z(\vec{r},t) = -z(\vec{r},t) - c_h h(t) + c_i x(\vec{r},t) + c_l \int_N w(\vec{r}-\vec{r}')y(\vec{r}',t)d^3\vec{r}'$$

- $\vec{r}$ denotes the three-dimensional coordinate of a neuron position in the field
- $z(\vec{r},t)$ is the activation of the neuron $\vec{r}$ at time t
- $y(\vec{r},t)$ is the output activity (sigmoidal function of $z(\vec{r},t)$)
- $h(t)$ is a global inhibition
- $x(\vec{r},t)$ denotes the external input consisting of the similarity measures of all cues
- $w(\vec{r}-\vec{r}')$ denotes the Mexican-hat-like function of lateral activation of neuron $\vec{r}$ from the surrounding neighbourhood $N \subseteq \boldsymbol{R}^3$
- c_h, c_l and c_i are parameters of the dynamic system



The left column depicts an exemplary result for the selection within the dynamic neural field. Obviously, one gets a blob of active neurons out of more than one level. This is due to a certain scale invariance of our two head-specific cues, so that the neural field finds a solution which enables some interpolation based on the centre of gravity.

Conclusion It should be noted that the development of the system is in progress. However, the state of the current implementation and the obtained results underline the necessity of multiple cues for person localisation, so that the gesture recognition system for human-robot interaction can further be implemented based on a real robust localisation.

References

[1] Amari, S.: Dynamics of Pattern Formation in Lateral-Inhibition Type Neural Fields. *Biological Cybernetics*, 1977.
[2] Bobick, A.: Computers Seeing Action. *British Machine Vision Conference*, 1997.
[3] Boehme, H.-J. et al.: Neural Architecture for Gesture-Based Human-Machine-Interaction. *Gesture-Workshop Bielefeld*, 1997.
[4] Boehme, H.-J. et al: Visually-Based Human-Machine-Interaction in a Neural Architecture. *SOAVE'97 - Selbstorganisation von adaptivem Verhalten*, 1997.
[5] Maggioni, C. et al.: GestureComputer – History, Design, and Applications. *Computer Vision in Man-Machine Interfaces*, 1996.
[6] Darrell, T. et al.: Perceptually-driven Avatars and Interfaces: Active Methods for Direct Control. Technical Report TR 416, MIT, 1997.
[7] Fritzke, B.: A Growing Neural Gas Network Learns Topologies. *Advances in Neural Information Processing Systems*, 1995.
[8] Goodridge, S.: *Multimedia Sensor Fusion for Intelligent Camera Control and Human-Computer-Interaction*. PhD thesis, North Carolina State University, 1997.
[9] Gross, H.-M. et al.: A Neural Network Hierarchy for Data and Knowledge Controlled Selective Visual Attention. *Int. Conf. on Artificial Neural Networks*, 1992.
[10] Hunke, M.: Locating and Tracking of Human Faces with Neural Networks. Technical Report, CMU-CS-94-155, 1994.
[11] Kopecz, K.: Neural Field Dynamics provide Robust Control for Attentional Resources. *Aktives Sehen in Technischen und Biologischen Systemen*, 1996.
[12] Kortenkamp, D. et al.: Recognizing and Interpreting Gestures on a Mobile Robot. *13th National Conf. on Artificial Intelligence*, 1996.
[13] Moghaddam, B. et al.: Maximum Likelihood Detection of Faces and Hands. *Int. Workshop on Automatic Face- and Gesture-Recognition*, 1995.
[14] Oren, M et al.: Pedestrian Detection Using Wavelet Templates. *Proc. of the IEEE Comp. Soc. Conf. on Comp. Vision and Patt. Recogn.*, 1997.
[15] Pomierski, T. et al.: Biological Neural Architecture for Chromatic Adaptation Resulting in Constant Color Sensations. *Int. Conf. on Neural Networks*, 1996.
[16] Rowley, H. A. et al.: Human Face Detection in Visual Scenes. Technical Report, CMU-CS-95-158R, 1995.
[17] Schiele, B. et al.: Gaze Tracking Based on Face-Color. *Int. Workshop on Automatic Face- and Gesture-Recognition*, 1995.