Hybrid Neural Networks to recognize Body Postures in Human-Computer Interaction

A. Corradini*, U.-D. Braumann, H.-J. Boehme, H.-M. Gross
Department of Neuroinformatics
Ilmenau Technical University
PF 10 05 65, 98684 Ilmenau, Germany
http://cortex.informatik.tu-ilmenau.de/
*supported by the EU TMR Marie Curie grant # ERBFMBICT972613

Summary
In this poster a novel approach to human posture analysis and recognition using standard image processing techniques as well as hybrid neural information processing is presented.
Since we think a reliable and robust saliency system for person localization is the prerequisite for a further gesture analysis we first cope with the head localization task using a multi-cue localization approach.
Then we focus on the user´s static gestural instructions based on both a skin color model and statistical moment invariants.
The segmentation of face and hands as the gesture relevant parts occours by using a skin color model. After determining the location of the user´s head we update that parametric skin model in order to fit more specific for the illumination and the skin at hand. By means of the Mahalanobis distance each pixel is classified to be or not a member of the skin class according to a threshold value.
After extracting from the binary image the 'greatest' connected regions we model them as circles around their centre of gravity. Finally we extract 14 translation and size invariants which we use as input for two hybrid supervised-unsupervised neural classifiers whose results are then compared.
To train and test the neural classifiers we gathered the data from five people performing 18 repetitions of each of the five postures from a predefined vocabulary which consists of the 'terms': stop, go left, go right, waving left and waving right.
The system uses input from a color video camera and is user independent. It achieves a recognition rate up to 93%.

References
AMARI, S.-I.: *Dynamics of Pattern Formation in Lateral-Inhibition Type Neural Fields.* Biological Cybernetics 27, pp. 77-87, 1977

FREEMAN, W.T. and ADELSON, E.H.: *The Design and Use of Steerable Filters.* IEEE Transactions on Pattern Analysis and Machine Intelligence 13, pp. 891-906, 1991

HECHT-NIELSEN, R.: *Counterpropagation Networks.* Proceedings of the IEEE Int. Conference on Neural Networks, vol.2, pp.19-32, 1987

HU, K.: *Visual Pattern Recognition by Moment Invariants.* IRE Transactions in Information Theory, pp. 179-187, 1962

JÄHNE, B.: *Practical Handbook on Image Processing for Scientific Applications.* CRC Press LLC, Boca Raton, 1997

KOENDERINK, J.J. and DOORN, A.J.: *Receptive Field Families.* Biological Cybernetics 63, pp. 291-297, 1990

MARTINETZ, T. and SCHULTEN, K.: A Neural Gas Network learns Topologies. Proc. of Int. Conf. on Artificial Neural Networks, pp. 379-402, 1991

YANG, J. and WAIBEL, A.: *A Real-Time Face Tracker.* Proc. IEEE Workshop on Appl. of Computer Vision, pp. 142-147, 1996

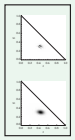
ZIMMERMANN, H.-J.: *Fuzzy Set Theory - and its Applications.* Kluwer Academic Press, 2nd Edition, 1991

Acknowledgments The authors would like to thank Anja Brakensiek, Markus Krabbes, Peter Paschke, Lara Neumann and Fabian Braumann for their criticism, help and encouragement.

Multi-Cue Person Localization
Finding persons within cluttered visual scenes is a non-trivial task as long as certain conditions cannot be tightly constrained.For this work a frontal alignment of the person related to the camera view is assumed. This allows the use of the following person specific cue components:
- outer shape of **contour** of head and shoulders
- **color** of visible skin (face, hands)
- frontal **face**

Contour

This component relates to a quite typical invariant of a person. The frontal shape of head and shoulders can appear almost independently of the lighting intensity or contrast. A measure to evaluate a part of an image for the presence of such contour can be determined using a **shape-adapted arrangement of oriented filters** in the position space. By each filter it is just determined whether the respective dominant local orientation matches with some prototype template (see sketch). Steerable filters (FREEMAN 1991) or the 2D interia tensor approach (JÄHNE 1997) both turned out to be elegant solutions for the necessary orientation filtering.

Skin Color


Human skin color might appear to have a special color, but this color is not unique! For the generation of a skin color model a set of images containing skin regions were segmented manually. To become more independent from illumination changes in all further processing steps the RGB_{EBU}-tripels are projected onto a chromatic plane rg (cf. YANG 1997). Analyses yielded a good approximation of this data set (upper diagram) by means of a **2D Gaussian function** as a parametric model (lower diagram).

Face

Reliable face detection is the most challenging cue within this framework since it necessitates fine resolutional levels requiring high computational effort. In search of an algorithm to detect frontal faces a simple but both robust and easy to handle algorithm based on cosine metrics is used, i.e. a **normalized scalar product** of an `average´ face (see figure) and image tiles.

Fusion and Selection
The idea behind multimodality is to have certain partly independent components providing sufficient redundancy. Therefore, the fusion of the cues cannot be simply a superposition. A good solution turned out to be the **fuzzy MIN-MAX γ-operator** (ZIMMERMANN 1991) which allows to do a compromise between superimposition and a selection of the strongest component. All cue components are calculated in five resolutional levels (pyramidal representation) which provides information to determine the person´s distance.
In order to have a selection for the localization, the **dynamic neural field approach** (AMARI 1977) was extended towards a three-dimensional topology. That field can be described as recurrent non-linear dynamic system with a dynamic behaviour which leads to *one* local region of active neurons succesfully competing against others, i.e. a formation of a single blob of active neurons as an equilibrium state of the field.

Posture Description
In our work the segmentation of face and hands as the gesture relevant parts is exclusively based on **skin color processing**.
After detecting the location of the head, we consider a window subregion around it and we characterize the pixel value distribution inside it by a multidimensional Gaussian function. By doing that we adapt the skin color model to fit more specific for the illumination and the skin type at hand. The **chromatic projections** of each RGB pixel inside that subregion are used.
After that by means of the Mahalanobis distance **each** pixel of the current image is classified to be or not a member of the skin class according to an empirically determined threshold value.
Now we apply to the resulting binary image a winner-take-all algorithm (AMARI 1977) to obtain the regions corresponding to the hands and head (which we assume to be the three greatest ones) and we determine their centers of gravity (COG). We model each of these regions as a circle around their COGs. That avoids the problems deriving by the shape of each region due to the choice of the color threshold.

From that binary image we compute a 14-dimensional feature vector containing translation and scale invariant elements characterizing the shape of the segmented scene (**view-based description**).
As first ten vector elements we choose the **scale normalized moments** (HU 1962) up to the third order. The remaining four elements were computed as follows.
To compensate the shift variation of the person gesticulating in front of the camera we define for each image a suitable coordinate system by fixing its origin point at the current determined head´s center of mass. It allows to calculate a feature vector related to the head position and regardless to the user´s position within the image. In this new coordinate system in order to ensure invariance also with respect to image size change, we use the polar coordinates of both hands´ COGs.

The neural classifiers
For the posture recognition we use two different neural classifiers trained with a data set containing 450 feature vectors. These vectors were computed from a set of 90 examples for each posture performed by five different persons. We use 225 vectors for training and the others for testing.
A first feedforward network (14-20-5) was taught in the hidden layer with an unsupervised **Neural Gas** (NG) algorithm (MARTINETZ 1991) and in the output layer with the standard delta rule (DR). In its simplest form the NG layer functions in a **winner-take-most** fashion. Unlike to other self-organizing methods in the NG the adaptation steps are not determined by the location of the neural input within a topologically predefined lattice, but instead by the relative distance between the neurons in the input space.
The adaption step for an arbitrary weight $\vec{w}_j$ occours according to the following rule:

$$\vec{w}_j^{(t+1)} = \vec{w}_j^{(t)} + \varepsilon\, e^{-k_j/\lambda}\,(\vec{v} - \vec{w}_j^{(t)})$$

The two costants ε and λ describe the overall extend of the weight adaptation and the number of neural units mostly changing at each step their weights, respectively. Each time an input signal $\vec{v}$ is presented, the adjustment depends on the position k_j of $\vec{v} - \vec{w}_j$ within the set $\{|\vec{v} - \vec{w}_i| \text{ for each } h\}$ sorted in ascending order.
The second network relies on the counterpropagation network (HECHT-NIELSEN 1987). It has the same topology as the previous one and was trained by a hybrid combination of the unsupervised NG paradigm in the hidden layer and by the **Grossberg Outstar** (GO) algorithm in the output layer. To train an outstar neuron, its synaptic weights are adjusted to be like a desired target vector.
The training equation that follows was used:

$$\vec{w}_j^{(t+1)} = \vec{w}_j^{(t)} + \beta(t)\,(\vec{y}_j - \vec{w}_j^{(t)})$$

Results and Future Work
The first network yields a robust performance, and the number of false classified patterns is rather low, whereas the counterpropagation-like network suffers from both a large number of misclassifications and a low recognition rate. Therefore we decide to first use the first network which converges very speedy and further once the convergence is achieved to continue the training with the second slower network. By this, we got more accurate results. The table below summarizes the achieved preformances concerning the networks.

Network	Topology	% of false training patterns	% of not classified patterns	Recognition rate
NG+DR (a)	14-20-6	3.3	5.4	91.3 %
NG+GO (b)	14-20-6	8.5	10.2	81.3 %
(b) then (a)	14-20-6	2.8	4.6	92.6 %

Up to now we used a limited posture alphabet but we are currently extending the system with the aim to both overcome this limitation and deal with continuously dynamic gestures. More precisely we want to describe different gestures via their trajectory in the input feature space. Taking the time into account means the introduction of a new degree of freedom which permits to bypass the problem of single posture overlapping.

Presented at the Workshop "Neural Computation in Science and Technology" (NCST'99), Oct. 10 - 13, 1999, Kibbutz Maale Hachamisha, Israel