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

Andrea Corradini*, Ulf-Dietrich Braumann,
Hans-Joachim Boehme, Horst-Michael Gross
Technical University of Ilmenau, Dept. of Neuroinformatics

D-98684 Ilmenau, Federal Republic of Germany
E-mail: {andreac,ulf}@informatik.tu-ilmenau.de

Abstract

In this paper a novel approach to human posture analysis and recognition using standard image processing techniques as well as hybrid neural information processing is presented. We first develop a reliable and robust person localization module via a combination of *certain salient cues* and *three-dimensional dynamic neural fields*. Then we focus on the view-based recognition of the user's static gestural instructions from a predefined vocabulary based on both a skin color model and statistical normalized moment invariants. The segmentation of the postures occurs by means of the skin color model based on the Mahalanobis metric. From the resulting binary image containing only regions which have been classified as skin candidates we extract translation and scale invariant moments.

These are used as input for two different *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 five postures (our vocabulary): stop, go left, go right, hello left and hello right. The system is currently under development with constant updates and new developments. It uses input from a color video camera and is user-independent. The aim is to build a real-time system able to deal with dynamic gestures.

1 Introduction

Human beings exploit the functions of the gesture already from the early childhood. Infants, long before being able to speak, gesticulate to convey their desires and needs and these abilities in gesticulating continuously improve and become natural and intuitive the more the person becomes adult. In processes acting as intermediary agents between humans and computers, people must be allowed to concentrate their attention and efforts on the content of the interaction. Therefore the optimal interaction does not require any remembrance and is similar to that they are familiar, thus the interaction with other people [7].

---

*supported by the European Union: Training and Mobility of Researchers – Marie Curie Grant # ERB FMBI CT 97 2613; any correspondence should be directed to A. Corradini

For this reason gestures with regard to *Human-Computer-Interaction (HCI)* purposes have become an intensive field of research. Although for many years the dominant input devices to capture gestural information have been *intrusive* dispositives like the keyboard, mouse or glove, the demand for more natural and body-centered applications increased the interest for *non-intrusive* camera-based input devices [9]. These methods are user-friendly and do not require the user to wear an additional instrument but suffer both from too high computational costs for real-time image processing and the difficulty of extracting information from 2D visual image. To sense gestures with a camera limits the user to face it and requires *highly constrained environments*. Taking into account this fact we have developed a robust saliency system for person localization integrating different visual cues. After the detection of a person aligned in front of the camera a gesture recognition process is to be carried out to capture the user's movements.

We propose to combine skin color-based image segmentation with shape analysis by means of invariant moments as input vector to a *hybrid unsupervised-supervised neural network*. The results obtained using two different neural classifier paradigms are presented.

2 Person Localization

2.1 System Overview

We think that a robust saliency system for person localization is the prerequisite for a further gesture analysis. Finding persons within cluttered visual scenes is a non-trivial task as long as certain conditions cannot be tightly constrained. However, even under normal indoor conditions, e. g. office or lab rooms or floors, the variety of possible viewpoints with quite different complexities for the localization task is surprisingly high. Thus, it may not be surprising to take a multi-cue approach as we did in order to widen the basis of feature modalities, and by this, to get a sufficient redundancy. Fig. 1 provides a coarse sketch of

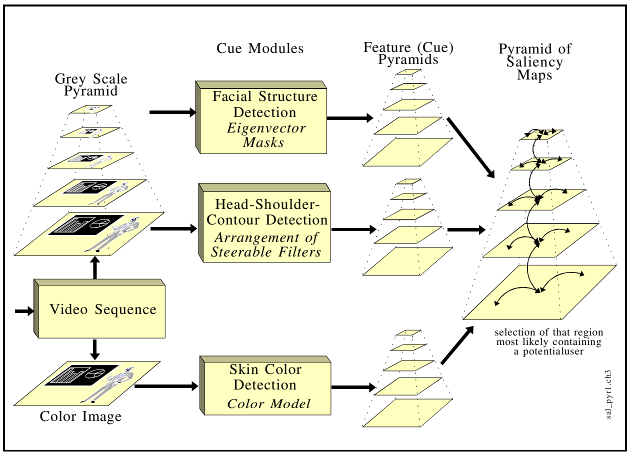

Figure 1: *Components of the saliency system for person localization*

the saliency system for user localization. Multiresolution pyramids transform the input images into a multiscale representation. Two cue modules sensitive to *facial structure* and *structure of the head-shoulder contour*, respectively, operate on all levels of a grayscale pyramid. The cue module for *skin color* detection uses the original color image and bases on a statistical parametric color model. The segmentation of the color module 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. Moreover, because of the multi-cue approach the algorithmical effort for one single cue can be kept rather low.

The output of the cue modules serves as the input for a 3D dynamic neural field. To achieve a good localization we need a *selection mechanism* to make a definite choice. So we actually can localize persons even in different distances, since we use five fine-to-coarse resolutions. Therefore, a neural field for selecting the most salient region should be three-dimensional.

2.2 Visual Cues

A nice property of persons getting in contact with a machine is their frontal alignment related to the system's camera view, which can be considered as some intuitive behavior if one visibly wants to communicate with another agent. This allows to consider the *outer shape of contour of head and shoulders*, *color of visible skin (face, hands)* and *frontal face* as reasonable person-specific cue components.

The first 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 determined whether the respective *dominant local orientation* matches with some prototype template. Steerable filters turned out to be an elegant implementation of oriented filters. However, similar results can be obtained by means of JÄHNE's inertia tensor approach for local orientation estimation [5].

Although human skin might appear to have a special color, this color is not unique! On one hand, this non-uniqueness is the reason why a color cue alone is reasonable only in very special constrained cases. On the other hand, the advantage of a color-based detection is its simplicity, whereas in this work color is taken as supplementary cue. For the generation of a skin color model we manually segmented a set of images containing skin regions. Because after some experiments we noted that the color distribution of human skin can be good approximated by a Gaussian distribution, we use it as parametric model. By this statistical model of the typical subspace of skin color within a color space all these regions can be segmented which appear like (caucasian) skin. Due to its simplicity, we adapted that one proposed by YANG and WAIBEL [10], which uses a projection ($\mathbb{R}^3 \rightarrow \mathbb{R}^2$) onto a plane through the upmost values of the $\mathrm{RGB}_{\mathrm{EBU}}$-cube (chromatic projection).

Finding faces in images is the most challenging part, since it necessitates fine resolutional levels requiring high computational effort. Since this is no face discrimination task the algorithms can be kept rather simple which at all allows a reasonable implementation of that cue. In search of an algorithm to detect frontal faces we decided to use a quite simple but robust algorithm based on cosine metrics (normalized scalar product of an 'average' face and image tiles).

The idea behind multimodality is to have independent components providing sufficient redundancy. Therefore, the fusion of the cue cannot be simply a superimposition. We propose a fuzzy MIN-MAX-operator which allows to do a compromise between superimposition and a selection of the strongest component. Note that all cue components are calculated in five resolutional levels (pyramid), which provides information to determine the person's distance.

Finally in order to have a localization selection mechanism yielding a unique clear solution, we extended AMARI's dynamic neural field [1] towards a three-dimensional topology. That field can be described as recurrent nonlinear dynamic system with 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 [1]. A more detailed description of the whole localization task can be found in [2].

3 Posture Segmentation

In our work the segmentation of face and hands as the gesture relevant parts is exclusively based on skin color processing therefore we assume skin color is always present within an image.

After detecting the location of the head as described above, we consider a window subregion around it which we call *head box* (Fig. 2,b). Then we characterize the distribution of the pixel values inside that subregion by a multidimensional Gaussian with centroid location and a covariance matrix describing the local distribution around the centroid. By doing that we adapt the skin color model to fit more specific for the illumination and the skin type at hand. Therefore the detection of skin colored regions can be improved. We handle multiple scales by choosing head boxes of different sizes according to the level in the pyramide.

By using the chromatic projection $r = \frac{R}{R+G+B}$ and $g = \frac{G}{R+G+B}$ of each pixel inside the head box the actual color model is uniquely determined by the multivariate normal density

$$p(\vec{x}) = \frac{e^{-\frac{1}{2}(\vec{x}-\vec{\mu})^T\Sigma^{-1}(\vec{x}-\vec{\mu})}}{(2\pi)|\Sigma|^{1/2}} \tag{1}$$

where the mean $\vec{\mu}$ is a two-dimensional vector, Σ is a 2×2 covariance matrix, and $|\Sigma|$ represent its determinant. Using the quantity appearing in the exponent of equ. 1 (also called Mahalanobis distance from $\vec{x}$ to $\vec{\mu}$) each pixel $\vec{x}$ of the image is then 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 (Fig. 2,c) a winner-take-all (WTA) algorithm [1] to obtain the regions corresponding to the hands and head (which we assume to be the three greatest regions) and we determine their centers of gravity (COG). Then we model each of these regions as a circle around their COGs with constant radius (Fig. 2,d). That avoids problems deriving by the shape of each region due to the choice of the color threshold.

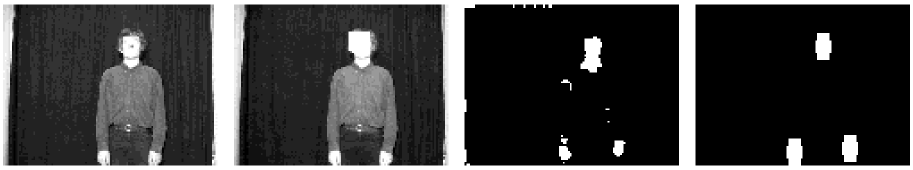

Figure 2: *From left to right: input image, head localization result with head box, thresholded skin classification by means of an adapted color model derived from the pixel distribution inside the head box, modelling of the three greatest regions as circle around their centers of mass.*

4 Posture Recognition

4.1 Moment-based Posture Description

From that binary image, sub-sampled to a dimension of 64×64 pixels we compute a feature vector $\vec{v}$ containing 13 translation and scale invariant elements characterizing the shape of the segmented scene.

Given a pixel distribution $f(x, y)$ its two-dimensional $(p+q)$th order central moments are defined by

$$\mu_{pq} = \sum_{x,y} x^p y^q f(x - \bar{x}, y - \bar{y}) \tag{2}$$

where $\bar{x}$ and $\bar{y}$ represent, respectively, the x and y coordinate of the image's COG. Applying the theory about algebraic invariants by Hu [4], it is straightforward to show that the values

$$\nu_{pq} = \frac{\mu_{pq}}{\mu_{00}^{(1+\frac{p+q}{2})}} \tag{3}$$

known as the *scale normalized moments*, remain unchanged under image translation and size changes. In our work we take them up to the third order yielding the first 10 invariant values of our feature vector. The computation of them for binary image yields theoretically an error-free estimate of the continuous moments which is also independent of illumination as opposed to the value deriving from grayvalue images.

To compute the remaining 4 feature vector elements we operate as follows. To compensate the shift variation of the person gesticulating in front of the camera we choose 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 relating 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's COG (Fig.3).

4.2 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 used 225 vectors for the training and 225 vectors for the test of the networks.

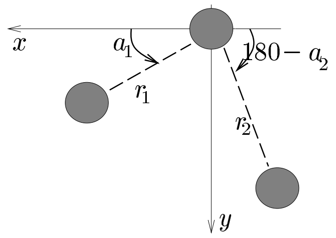

Figure 3: *The new defined coordinate system with origin centered in the COG of the head. If* (r_1, a_1) *and* (r_2, a_2) *represent the polar coordinates of the hands's COG as last feature vector elements we take the four values* $(\frac{r_1}{\max\{r_1,r_2\}}, a_1, \frac{r_2}{\max\{r_1,r_2\}}, a_2)$.

A first feedforward network (13 input, 20 hidden, and 5 output nodes) was trained in the hidden layer with a unsupervised Neural Gas (NG) algorithm by Martinetz and Schulten [6] and in the output weight layer via the standard delta rule (DR). In its simplest form the NG layer functions in a 'winner-take-most' fashion. Unlike to other self-organizing algorithms in the Neural Gas the adaptation steps are not determined by the location of the neural units within a topologically predefined lattice, but instead by the relative distances between the neurons in the input space. The adaptation step for an arbitrary weight $\vec{w}_j$ occurs according to the following Hebbian-like rule:

$$\vec{w}_j^{(t)} = \vec{w}_j^{(t-1)} + \epsilon e^{-k_j/\lambda}(\vec{v} - \vec{w}_j^{(t-1)}) \tag{4}$$

The two constants $\epsilon \in [0, 1]$ and λ define the overall extend of the weight adaptation and the number of neural units mostly changing at each step their synaptic weights respectively. Each time an input signal $\vec{v}$ is presented, the adjustment of the synaptic weight $\vec{w}_j$ depends on the position k_j of $\vec{v} - \vec{w}_j$ within the set $\{\|\vec{v} - \vec{w}_l\| \,\forall$ unit neuron $l\}$ sorted in ascending order.

The second network relies on the counterpropagation (CP) network developed by Hecht-Nielsen [3]. The network has the same topology as the previous one. It was trained by a hybrid combination of the unsupervised NG paradigm in the hidden layer and by the supervised Grossberg Outstar (GO) algorithm in the output one. To train an outstar neuron, its synaptic weights are adjusted to be like a desired target vector. The training equation that follows is:

$$\vec{w}_j^{(t)} = \vec{w}_j^{(t-1)} + \beta(t)(\vec{y}_j - \vec{w}_j^{(t-1)}) \tag{5}$$

where β is a training coefficient starting near 0.1 and gradually reducing to zero as training progresses and $\vec{y}_j$ is the desired output.

5 Results and Future Work

Tab. 1 summarizes the achieved performance concerning the two networks and their pipeline combination. The first network yields a robust performance, and the number of false classified patterns is rather slow, whereas the CP-like network suffers from both a large number of misclassifications and a slow recognition rate. Therefore we decided to use the first network which converges very speedy and further once the convergence is achieved to continue the training with the second slower but more accurate in the results network.

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 are describing different space-time

Network	Topology	Test & Training Patterns	False classified Patterns in %	Not classified Patterns in %	Recognition Rate in %
(a) NG+DR	13-20-5	225	3.3	5.4	91.3
(b) NG+GO	13-20-5	225	8.5	10.2	81.3
(b) then (a)	13-20-5	225	2.8	4.6	92.6

Table 1: *Summary of the achieved performances using the NG with two different paradigms for the output layer and a combination of both. Inputs are considered as not classified if after feeding it into the network in the output layer more than one neuron shows high activity. See text for details.*

gestures via the observed trajectory in the moment feature space using Hidden Markov Models (HMM) [8]. Taking time into account means the introduction of a new degree of freedom. That will permit to extend our vocabulary bypassing many problems deriving from the overlapping of single postures in the two-dimensional posture space.

References

[1] S.-I. Amari, "Dynamics of Pattern Formation in Lateral-Inhibition Type Neural Fields", *Biological Cybernetics*, 27:77–87, 1977.

[2] A. Corradini, U.-D. Braumann, H.-J. Boehme, H.-M. Gross, "Contour-based Person Localization by 3D Neural Fields and Steerable Filters", *Proc. of MVA '98, IAPR Workshop on Machine Vision Appl.*, pp. 93–96, 1998.

[3] R. Hecht-Nielsen, "Counterpropagation Networks", *Proc. of the IEEE First Int. Conference on Neural Networks, vol. 2*, pp. 19–32, 1987.

[4] K. Hu, "Visual Pattern Recognition by Moment Invariants", *IRE Transactions on Information Theory*, pp. 179–187, 1962.

[5] B. Jähne, "Practical Handbook on Image Processing for Scientific Applications", *CRC Press LLC, Boca Raton*, 1997.

[6] T. Martinetz, K. Schulten, "A Neural Gas Network Learns Topologies", *Proc. of the 1991 Int. Conf. on Artificial Neural Networks*, pp. 397–402, 1991.

[7] A. Mulder, "Hand Gestures for HCI", *Tech. Rep. NSERC Hand Centered Studies of Human Movement Project, Simon Fraser Univ., Burnaby*, 1996.

[8] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", *Proceedings of the IEEE*, 77(2):257–285, 1989.

[9] R. Watson, "A Survey of Gesture Recognition Techniques", *Trinity College, Dublin 2, Tech. Rep. TCD-CS-93-11*, 1993.

[10] J. Yang, A. Waibel "A Real-Time Face Tracker", *Third IEEE Workshop on Applications of Computer Vision (WACV´96)*, pp. 142–147, 1996.