In: Proceedings of the X Italian Workshop on Neural Nets (WIRN'98), pp. 201-206, May 21-23, 1998, Vietri Sul Mare (SA), Italy

Visual Person Localization with Dynamic Neural Fields: Towards a Gesture Recognition System

Andrea Corradini*, Ulf-Dietrich Braumann, Anja Brakensiek**, Markus Krabbes**, Hans-Joachim Boehme, Horst-Michael Gross

Department of Neuroinformatics, Technical University of Ilmenau

D-98684 Ilmenau, Fed. Rep. of Germany

{andreac,ulf}@informatik.tu-ilmenau.de

Abstract

For any *visually-based interaction* between persons and acting systems within a real-world environment the *localization* of a user by the system is a necessary condition. The presented work deals with this visual localization problem of a user concretely referred to the *autonomous mobile robot system MILVA* of our department. Since this system is applied under real-world conditions especially for the localization some proper techniques are needed which have an adequate robustness. In our opinion, this requires the combination of several components of saliency towards a multi-cue approach, consisting of structure- and color-based features [2].

This paper introduces one of them: the localization based on a typical *shape of contour*. A 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. Applying such filter arrangement in a multiresolution manner, this leads to a robust localization of frontally aligned persons even in depth. The central problem of *selecting* the most promising (salient) image region is treated by means of a *three-dimensional dynamic neural field* performing a dynamic winner-take-all process (WTA, [1, 6]).

After a successful localization of a person one can start a more detailed analysis of the *gesture's* meaning: besides the recognition of static gestures we also concentrate on the acquisition and later the recognition of *dynamic gestures.*

1 Introduction

The autonomous mobile robot system MILVA of our department serves not only as an experimental platform for investigations of autonomous robot navigation in real-world environments, but also for work concerning gesture-based interaction between user and robot. It's getting more and more important for both navigation and interaction to tolerate a rising level of complexity of the surroundings the robot system is operating in. The goal should be to operate even without special preparations so that e. g. no remarkable restrictions should appear concerning lighting conditions etc. Our *superior goal* is that MILVA *develops an intelligent behavior* both concerning navigation and interaction using

*supported by: TMR Marie Curie EU Fellowship # ERB FMBI CT 97 2613

**supported by: Thuringian Ministry of Science, Research and Culture: GESTIK-Project

an architecture consisting of elements with neural or neurally inspired mechanisms.

MILVA can develop an active and therewith intelligent detail analysis (attention-based perception, e. g. for the analysis of a gesture) only if there are a priori proper saliency mechanisms available enabling a preselection of a visual image. Consequently, for the interaction with persons we need saliency mechanisms to localize them.

After the introduction of one of the used saliency mechanisms and its selection process with dynamic neural fields we show what follows the localization of a person: a detailed visual analysis of that person. Besides the recognition of static gestures [2] we also concentrate on dynamic gestures. So, the second part of the paper introduces our present results of segmenting regions of motion in visual images by means of the Poggio-Reichardt motion computation model and a following dynamic neural WTA process.

2 Arrangements of Oriented Filters for Feature Extraction

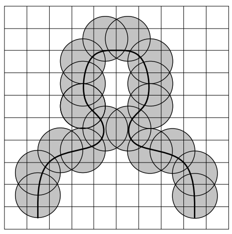

Figure 1: *Schematic sketch of the arrangement of oriented filters.*

This paper introduces a contour shape based approach for saliency based on some physiological considerations as well as on psychophysical effects. Recently, Oren [9] independently developed an interesting similar approach. While our idea refers just to the shape of head and shoulder, his approach considers the complete body of pedestrians.

The visual cortex consists in several parts of cells with oriented receptive fields. Thus, referred to a retinal position broad ranges of the frequency space are covered by a set of oriented filters. A lot of investigations [5, 7] have exposed that there is a high similarity between the profiles of the mentioned receptive fields and two-dimensional Gabor-functions. So, local operations decomposing the visual information with respect to frequency space are made. Psychophysical aspects as good continuation or symmetry (both belonging to the Gestalt laws, [8]) obviously describe effects which necessitate grouping mechanisms. Against this background we conceptualized the approach of *arrangements* of oriented filters. While fig. 1 shows the positions of the filters it does not explicitly show the respective directions. What is important is that we only use four orientations. Since we do not want to discriminate an orientation in the range of $[0, \pi)$ we should cover 45°-sectors of the frequency space. That's why the filters belonging to the arrangement do have exactly that orientation which is the most prominent caused by the contour shape within the filter range[1]. The arrangement is applied by determining

[1]The positioning and orientation selection of the filters is in general an optimization problem. However, since we use a rather coarse discrete grid and very few different orientations it is not as complex so that one would be forced to use a special algorithm. If one defines an energy function there would be on principle no reason against an adaptive neural algorithm.

whether the dominating orientation at each coordinate fits the dominating orientation of a prototype shape. As distance measure we use simply the relative Hamming distance.

Gabor filters are ideal if one considers their sharpness both in position and frequency space, but have high computational costs. There is a way providing a good compromise between computational cost and exactness. This is the approach of inertia tensors [4]. The direction of such tensor corresponds to that of the most dominant oriented visual structure within a certain range. The computational costs are magnitudes less than those of the Gabor filters.

3 Dynamic Neural Fields for Selection

To achieve a good localization a selection mechanism is needed to make a definite choice. This is not limited to a two-dimensional position. Since we use five fine to coarse resolutions we actually can localize persons even in different distances. Therefore, a neural field for selecting the most salient region should be three-dimensional.

What do we understand by dynamic neural fields? Those fields can be described as recurrent nonlinear dynamic systems. 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:

$$\tau \frac{d}{dt} z(\vec{r}, t) = -z(\vec{r}, t) - \mathrm{c}_h h(t) + \mathrm{c}_i x(\vec{r}, t) + \mathrm{c}_l \int_N w(\vec{r} - \vec{r}\,') y(\vec{r}\,', t) d^3 \vec{r}\,' \quad (1)$$

Herein $\vec{r}$ denotes the three-dimensional coordinate of a neuron position in the field, $z(\vec{r}, t)$ is the activation of a neuron $\vec{r}$ at time t, $y(\vec{r}, t)$ is the output activity of this neuron computed as a sigmoidal function, $x(\vec{r}, t)$ denotes the external input, $h(t)$ is the global inhibition at time t gathering the activity from each neuron over the entire field and $w(\vec{r} - \vec{r}\,')$ denotes the Mexican-hat function of lateral activation of neuron $\vec{r}$ from the surrounding neighborhood $N \subseteq \mathbb{R}^3$. The constants c_h, c_l and c_i represent parameters of the system.

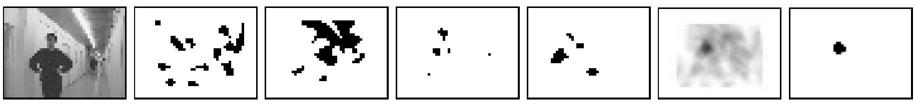

Figure 2: *These pictures illustrate the localization results in an indoor environment of one pyramidal plane* (48×36 *pixels*)*. From left to right: input image, orientation filtering* ($0°$, $45°$, $90°$ *and* $135°$)*, filter arrangement result, selection within a 3D field of dynamic neurons.*

The results of the system shall be qualitatively illustrated in fig. 2. All the images show the state of the system in a snapshot at that moment when the activity change Δy of the most active neuron became less than 1%. On average, the system takes 11 steps. The expansion of the blob is not restricted to one plane. To give a more precise specification of the distance of a person one should determine the center of the blob and interpolate the distance.

Our presented results are exemplary, the usage of the shape of contour provides one solution for the person localization problem, even under quite different conditions. The novel approach with a three-dimensional dynamic neural field can be assessed as robust method for the selection process.

4 Motion-based segmentation

Our goal is to build a motion-based recognition framework which can detect motion from a sequence of monocular images of a scene in which objects are in motion.

The knowledge of research in neurophysiology and psychophysics has influenced the design of vision systems particularly in recognition, interpretation and description of motion from a time-varying image sequence. In our paper we present an approach for the recognition of motion from image sequences modeling biological vision. Many psychophysical experiments have shown that the human visual system is able to detect, localize and isolate a moving object against a surround on the basis of motion information alone. The relative movement between an object and the ground permits both to identify its boundaries and to detect its presence [10].

Based on behavioral experiments and neurophysiological knowledge of the visual system of the fly *Musca domestica*, Poggio and Reichardt [10] proposed a neural circuitry for the figure-ground discrimination by relative motion. In our work we used a neural circuitry with forward shunting inhibition like the Poggio-Reichardt's one adapted for monocular images. We computed the motion region $R(t)$ between each sequential pair of images at time t and $t+1$ of a certain interval and then we summed over t in order to obtain an cumulative image named binary motion region BMR [3]. It describes those spatial regions within the sequence where motion occurred starting from the beginning of the recognition task.

In order to remove the noisy regions from the detected ones we then applied to the BMR image a process of extraction of relevant regions by a neurally-plausible, iterative competition/cooperation system of artificial neurons; a WTA mechanism.

Because the BMR image is two-valued, say on and off, the task the WTA has to aim at is to find that *connected* region consisting of the largest number of on-values. Because of the conditions of the images to process and the goals to aim at, selecting the actually largest region within the image cannot be accomplished by the neural scheme employed for the localization task[2] but only by a different approach.

This another WTA circuit, referred to as *largest region WTA* (LRWTA),

[2]The dynamic neural field introduced in the previous section cannot treat this problem because the blobs (fig. 2) cannot exceed some certain size (equilibrium) but the motion regions we want to segment now might be quite different depending on the kind of motion. Moreover we need a different approach dealing with binary inputs.

is composed of a single layer of N excitatory neurons of identical type each corresponding with one single pixel of the image. The neurons are mutually connected within their neighborhood defined by excitatory synaptic strength depending on the definition of connected region[3]. We consider pixels as neighbored if they have a common edge, i. e. we used a 4-neighborhood.

In order to find the actually largest region within the image each neuron of the correspondent neural field has to collect the excitatory activity of its neighboring neurons. For the synaptic strength of each neuron we chose a $n \times n$ separable kernel satisfying the property of symmetry, separability and unimodality. To avoid attenuation or amplification the sum of the elements of the kernel is required to be 1. Because a 4-neighborhood cannot be defined through a separable kernel we circumvented this problem with the introduction of a proper function $f(\cdot)$ of the activity. For a 3×3 separable kernel originated by $\vec{w} = [a,\, b,\, a]$ and under the condition $4a < b$ we defined it as follows:

$$f(z_i(t)) = \begin{cases} 0: & z_i(t) \leq 4a^2 + b^2 \\ z_i(t): & z_i(t) > 4a^2 + b^2 \end{cases} \tag{2}$$

The dynamic of the field can be described by the following equation:

$$z_j(t+1) = \frac{\sum\limits_i w_{ji} f(z_i(t))}{\max\limits_{j \in \text{Field}} \{z_j(t)\}} \cdot I_j \tag{3}$$

where $z_j(t)$ is the inner state of the neuron j at time t, I_j is the correspondent initial input value of the BMR image, w_{ij} are the synaptic strengths and $f(\cdot)$ is a linear function which depends on the choice of the neighborhood. In addition, each element receives the original binary input I_j. It avoids that original cells without activity can contribute to the activity of the bounded region. The iterative computation of eq. 3 is stopped as soon as the change of two following states is below a certain threshold (stopping criterion).

Figure 3: *Demonstration of the functioning of the LRWTA. From left to right: initial and final image of a sequence, BMR image from the sequence, result of the first step of the competition/cooperation task (largest region was selected, others are temporary suppressed), time courses of the activities of all neurons until the selection of the largest region was finished.*

Beginning from the winning neuron the underlying region is segmented and its area is computed. When the measure of this area is less than a fixed percentage of the largest hitherto segmented region, it is considered a noisy region and the selection task is stopped. Otherwise the region is inhibited and a new

[3] In our paper a region is considered connected when we can reach any pixel within it only by moving to neighboring pixels.

competitive process begins. The selected motion region has to be permanently inhibited in order to avoid that it is again selected in the next competition task. Without inhibition the currently selected region would be always selected again (it is the largest!) and would not permit smaller regions to win the competition.

5 Conclusions and Future Work

Our contour-shape-based approach is one of the saliency components besides others like facial structure and skin color. In a series of tests under several conditions (indoor, outdoor) we obtained a quite robust person localization. For computational reasons, the used image resolution was highly reduced so that the distances of the persons had to remain rather close.

What does the results of the localization task mean for the development of a gesture recognition system? In our opinion, the analysis of the motion region could provide a promising way for the analysis of dynamic gestures. From the motion template we can first extract a feature vector and then recognize it by means of a statistical classifier (e. g. NN or HMM). A good head localization task is therefore essential for a further feature extraction step. It allows to calculate a feature vector relating to the relative head position and regardless to the user's position within the image (translation invariance).

References

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

[2] Boehme H.-J., Braumann U.-D., Brakensiek A., Corradini A., Krabbes M. and Gross H.-M.: User Localization for Visually-based Human-Machine-Interaction, In *Proc. of the Third IEEE Int. Conf. on Automatic Face and Gesture Recognition*, pp. 486–491, 1998

[3] Davis, J. W. and Bobick, A. F.: The Representation and Recognition of Action Using Temporal Templates. In *Proc. of the IEEE Computer Society Conference on Comp. Vis. and Patt. Rec.*, pp. 928–934, 1997.

[4] Jähne, B.: *Digital Image Processing* Springer-Verlag, 1995.

[5] Jones, J. P. and Palmer, L. A.: An Evaluation of the Two-Dimensional Gabor Filter Model of Simple Receptive Fields in Cat Striate Cortex. *J. Neurophysiol.*, 58(6):1233–1258, 1987.

[6] Kaski, S. and Kohonen, T.: WTA Networks for Physiological Models of Competitive Learning. *Neural Networks*, 7(6/7):973–984, 1994.

[7] Koenderink, J. J. and v. Doorn, A. J.: Receptive Field Families. *Biol. Cyb.*, 63:291–297, 1990.

[8] Koffka, K.: *Principles of the Gestalt Psychology.* Brace & World, 1935.

[9] Oren, M., Papageorgiou, C., Sinha, P., Osuna, E., and Poggio, T.: Pedestrian Detection Using Wavelet Templates. In *Proc. of the IEEE Computer Society Conference on Comp. Vis. and Patt. Rec.*, pp. 193–199, 1997.

[10] Reichardt, W., Poggio, T., and Hausen, K.: Figure-Ground Discrimination by Relative Movement in the Visual System of the Fly – Part II: Towards a Neural Circuitry. *Biol. Cyb.*, 46(suppl.):1–30, 1983.