In: Proceedings of the "IAPR Workshop on Machine Vision Applications" (MVA'98), pp. 93-96, November 17-19, 1998, Chiba, Japan

Contour-based Person Localization by 3D Neural Fields and Steerable Filters

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

Abstract

This paper introduces a way to locate persons in visual images of cluttered scenes using a *shape-of-contour approach*. The contour which we refer to is that of the upper body of frontally aligned persons. After deriving an approximation of it using a set of example images we take a *spatial arrangement of steerable filters* to determine the pointwise orientation along the contour.

However, the application of the filter arrangement typically yields a coarse distributed outcome. To select the most promising location, we apply a dynamic pattern formation within a *three-dimensional dynamic neural field* to get a localization even considering the distance of a person. It turned out that by means of simple homogeneous internal interaction rules the dynamic neural field can find robust localization solutions. The activity of the field-neurons can be considered as internal state enabling a *permanent* localization helpful for tracking the person.

1 Introduction

In a framework of an image-based gesture recognition system on board of our mobile robot platform MILVA the *localization of a user's head* has essential importance, since it is a prerequisite for any further gesture-related analyses. In this context of gesture-based human-robot interaction, the present work deals with the visual localization of persons in real-world environments. 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, clearly are less appropriate.

We think, especially in real-world scenes a person's outer contour shape represents the most appropriate invariant. 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 learnt 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.

The central problem of selecting the most promising (salient) image region is treated by means of a *three-dimensional dynamic neural field* performing a winner-take-all (WTA) process (blob-like pattern formation, see [1]). In future work such neural field will be used for merging with further visual cues (e. g. skin color) appropriate for person localization.

2 Arrangements of Steerable Filters

2.1 Motivation and Related Work

Our idea refers just to a description of the outer shape of head and shoulders, whereas the interesting and independently developed approach of Oren and Papageorgiou [7] considers the complete body of persons (pedestrians) using *Haar wavelet templates*. The common aspect between the two approaches is a set of locally distributed oriented filters used to determine the strength of certain orientations of visual "structure" for a small region.

The idea of a contour-shape based approach for saliency is based on some physiological considerations as well as on psychophysical effects. 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 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 [4] and Gaussian (incl. low order derivatives) [5, 8] 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.

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

§Address: PF 10 05 65, D-98684 Ilmenau, Fed. Rep. of Germany. E-mail: {andreac,ulf}.informatik.tu-ilmenau.de

Psychophysical aspects related with the contour-shape based approach as, e. g., good continuation or symmetry (both belonging to the Gestalt laws [6]), 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 this is computationally very costly we turn to use steerable filters (see below).

2.2 Determining the Course of Contour

In our previous work [2] we used an heuristically defined model of the contour. It was based on a manual design restricting to just four filter orientations. Obviously, one would have a more precise model using more than four orientations, i. e. the contour model should be closely related to *real data.*

Steerable filters have the nice property that an initially 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 $0°$ correspond 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 subsampled 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 are 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. High threshold values result in gaps within the contour, whereas low ones yield too broad contours.

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 $\mathbf{\Lambda}$. What is further of interest is the local orientation of each contour element. It is achieved by means of the steerable filters (see below) applied to the binary contour shape so that for each element of $\mathbf{\Lambda}$ an angle of orientation can be determined. Fig. 1 illustrates both the course of the shape of contour and the local orientations.

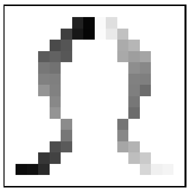

Figure 1: *The determined orientation angles along the shape of contour* $\mathbf{\Lambda}$*: the angles are coded as gray values from* $0°$ *(black) to* $180°$ *(white). Around the parting transitions from* $180°$ *to* $0°$ *occur.*

2.3 Applying Steerable Filters

The previous task provides a binary image representing an averaged head-shoulder-portrait but gives no information about the local orientation at a given contour point. After determining the contour, we measure the local orientation by means of a set of filters which are oriented in every direction. This again could be done, e. g., using the conventional *Gabor-type* filters, but it requires the choice of certain oriented (pair of) filters each of them differing from the others by a certain small rotation. In this case each filter pair corresponds to that angle the filter is tuned to. This also means that the orientation at a point of the contour is provided by the filter pair which has maximal response in this point. Unfortunately, by such an approach there is a trade-off between the required exactness of the orientation value and the number of filters. The more exactly the measure of the orientation has to be, the more filters (e.g. certain orientation) we have to choose. In this paper, we consider a different approach using *steerable filters* [3] for orientation estimation. This approach 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 efficiency and do not suffer from the orientation selection problem.

In general, a function $f(\cdot)$ is considered 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 superimposition of its basis set times interpolation functions $k_j(\vartheta)$ as

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

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 [3] these functions are steerable and need $M = 7$ basis functions (see fig. 2).

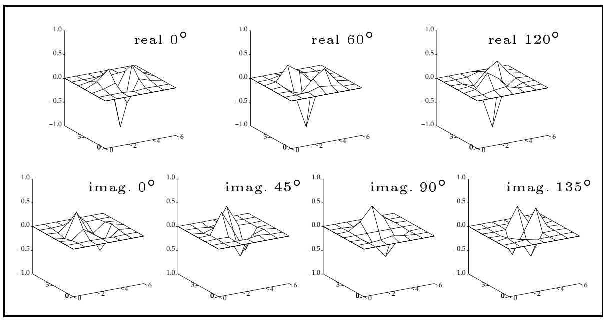

Figure 2: *The basis filter set: upper row with 3 basis functions (second order derivative of Gaussian), bottom row with 4 more basis functions (Hilbert transform of the second derivative of Gaussian).*

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 [3]. We further refer to the matrix of all these angular values $\vartheta_{MAX}^{(x,y)}$ corresponding to the image as $\mathbf{\Theta}$.

Unlike a Gabor-type filter approach, the processing scheme by steerable filters requires no additional convolution after the initial pass through the seven basis filters. Moreover, we choose these certain steerable filters because there exists a separable basis set in Cartesian coordinates which considerably lowers the computational costs.

3 Dynamic Neural Fields for Localization

3.1 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 $\mathbf{\Lambda}$ describing a typical course of the head-shoulder-portrait and that matrix $\mathbf{\Theta}$ (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 $\mathbf{\Lambda}$, within the matrix $\mathbf{\Theta}$. 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 $\mathbf{\Lambda}$ and $\mathbf{\Theta}$, the similarity function requires the same periodicity.

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

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

3.2 The 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 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. 3) 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:

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

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 of $\vec{r}$ alone, $x(\vec{r}, t)$ denotes the external input (corresponds to the re-coded similarity measure $m^{\vec{r}}$, cf. equation 2 and see fig. 3), $h(t)$ is the global inhibition at time t gathering the activity from each neuron over the entire field $F \subseteq \mathbb{R}^3$. $w(\vec{r} - \vec{r}\,')$ denotes the Mexican-hat-like function of lateral activation of neuron $\vec{r}$ from the surrounding neighborhood $N \subseteq \mathbb{R}^3$. For one $\vec{r}$, N is symbolically marked as dark regions in fig. 3 (right). The constants c_h, c_l and c_i represent parameters of the system.

As also illustrated in fig. 3, 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).

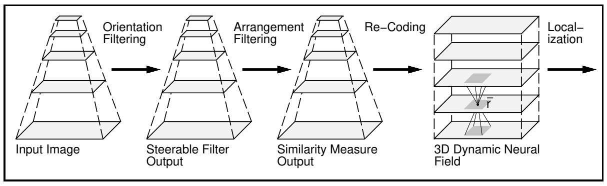

Figure 3: *Processing steps for person localization. Starting from a multi-resolution representation of the image, each level is treated by steerable filters. Applying the filter arrangement we determine a distance measure which is taken as input of a three-dimensional field of dynamic neurons. The resulting blob (locally delimited pattern of active neurons) is used to localize a person.*

3.3 Results

The results of the system are qualitatively illustrated in fig. 4. 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 provides one 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.

Acknowledgments

The authors especially thank Heiko Kempe and Andrea Corvino for helpful discussions and comments on this work.

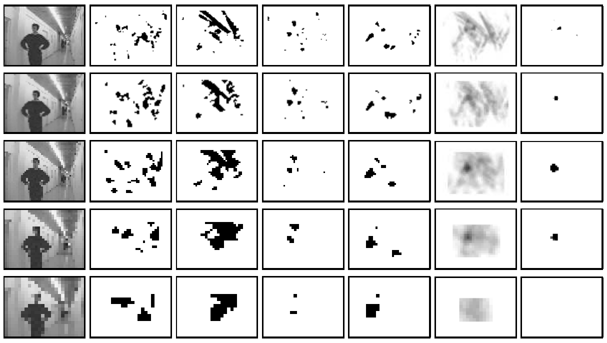

Figure 4: *Localization results in an indoor environment: The localization of a person does not sharply appear at one of the pyramidal planes, the originating spatial blob (rightmost column) is most strongly developed on the central of the five planes. Each row contains the results of one of the five (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.*

References

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

[2] H.-J. Boehme, U.-D. Braumann, A. Brakensiek, A. Corradini, M. Krabbes, and H.-M. Gross: "User Localisation for Visually-Based Human-Machine-Interaction". In *Proc. of the Third IEEE Intern. Conference on Automatic Face and Gesture Recognition (FG'98)*, pp. 486–491, 1998.

[3] W. T. Freeman and E. H. Adelson: "The Design and Use of Steerable Filters". *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 13(9):891–906, 1991.

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

[5] J. J. Koenderink and A. J. van Doorn: "Receptive Field Families". *Biological Cybernetic*, 63:291–297, 1990.

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

[7] M. Oren, C. Papageorgiou, P. Sinha, E. Osuna, and T. Poggio: "Pedestrian Detection Using Wavelet Templates". In: *Proc. of the IEEE Comp. Soc. Conference on Comp. Vision and Pattern Recognition (CVPR'97)*, pp. 193–199, 1997.

[8] R. A. Young: "Oh Say, Can You See? The Physiology of Vision". *General Motors Research, Technical Report 7364*, 1991.