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, PF 10 05 65, 98684 Ilmenau, Germany
`http://cortex.informatik.tu-ilmenau.de/`
*Supported by the European Union: TMR Marie Curie Research Training Grant # ERB FMBI CT 97 2613

Summary

This poster introduces a way to **locate persons in visual images** of cluttered scenes using a **shape-of-contour approach**. We think, especially in real-world scenes a person´s outer contour shape represents a sufficient and appropriate invariant. Our simple contour shape prototype model consists of an arrangement of oriented filters doing a piecewise approximation of the upper shape (head, shoulders) 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.
However, the application of the filter arrangement in a real-world environment typically yields a coarse distributed outcome. To select the most promising location, we apply a **dynamic pattern formation** performing a winner-take-all (WTA) process (AMARI 1977) within a **three-dimensional dynamic neural field** to get the location 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 solution. The activity of the field-neurons can be considered as internal state enabling a permanent localization helpful for tracking the person.
In a framework of **human-machine interaction** (HMI) the vision-based gesture recognition system on board of our mobile robot platform MILVA for the localization of a user´s head has essential importance, since it is prerequisite for any further gesture-related analyses. Further, solving the localization task 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, are rather inappropriate.

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 et al. 1997 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 in the cells in the mammalian primary visual cortex can be modeled by some two-dimensional mathematical functions. Gaborian (see JONES 1987, MELLER 1997), and Gaussian (see YOUNG 1991, KOENDERINK 1990), include low order derivatives, 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.
Psychophysical aspects for the contour-shape based approach as, e.g., good continuation or symmetry belonging to the **Gestalt laws** (see KOFFKA 1935), 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 different oriented filters as orientations belonging to the arrangement. Since it is computationally very costly we turn to use **steerable filters** (see FREEMAN 1991).

References

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

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

JONES, JUDSON P. und PALMER, LARRY A.: *An Evaluation of the Two-Dimensional Gabor Filter Model of Simple Receptive Fields in Cat Striate Cortex.* Journal of Neurophysiology 58, pp. 1233-1258, 1987

KOENDERINK, JAN J. und DOORN, ANDREA J.: *Receptive Field Families.* Biological Cybernetics 63, pp. 291-297, 1990

KOFFKA, KURT: *Principles of the Gestalt Psychology.* Harcourt, Brace & World, New York, 1935

MELLER, SEBASTIAN: *Segmentierungsfreie Personendetektion.* Master´s Thesis, Friedrich-Alexander-Universität Erlangen-Nürnberg, 1997

OREN, MICHAEL; PAPAGEORGIOU, CONSTANTINE; SINHA, PAWAN; OSUNA, EDGAR und POGGIO, TOMASO: *Pedestrian Detection Using Wavelet Templates.* Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'97), pp. 193-199, 1997

YOUNG, RICHARD A.: *Oh Say, Can You See? The Physiology of Vision.* TR GMR-7364, General Motors Research Laboratories, Warren (MI), 1991

Acknowledgments The authors would like to thank Markus Krabbes, Heiko Kempe and Andrea Corvino for their help and critical suggestions.

Steerable Filters

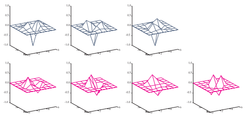
In general a function $f(\cdot)$ is considered to be steerable if the following two conditions are satisfied. First there exists a basis filter set of their M rotated copies $f^{\alpha_1} \ldots f^{\alpha_M}$ on some certain angles $\alpha_1 \ldots \alpha_M$. Second, a rotated copy $f^{\theta}(\cdot)$ of it on any angle θ can be obtained by a superimposition of ist basis set times interpolation function $I(\theta)$ as following: $f^{\theta}(\cdot) = \sum_{j=1}^{M} I^{(\alpha_j)}(\theta) f^{\alpha_j}(\cdot)$

We take a quadrature pair by using the second order derivative of a Gaussian and an approximation of its Hilbert transform with a total of $M=7$ basis function. The squared sum of the quadrature pair´s output represents an oriented energy function $E(\theta)$ which has to be maximized. $E^{(x,y)}(\theta) = \left(G^{(x,y)}(\theta)\right)^2 + \left(H^{(x,y)}(\theta)\right)^2$

Because the energy function looks like a Fourier series in angle where only even frequencies are present because of the squaring operation, we use only the lowest terms to maximize it. Therefore, the function to maximize can be written in the form: $E^{(x,y)}(\theta) \approx a_0 + b_1 \sin(2\theta) + c_1 \cos(2\theta)$

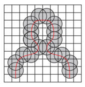
By doing that it is straightforward to show that the energy function can be analytical maximized and the dominant orientation is given by:

$$\theta^{(x,y)}_{MAX} = \arg\max \left\{ E(\theta), \quad \forall\theta \quad \middle| \quad \theta \in \left\{ \pm\frac{1}{2} \arccos(\pm c_1 / \sqrt{b_1^2 + c_1^2}) \right\} \right\}$$

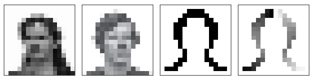


On the right, the filters making up the quadrature pair (three for the Gaussian and four for its Hilbertransform, here with wavelength $k=0.22$) are depicted.

Filterarrangements



In order to have a more precise prototype of the contour related to real data we derive our model from a set of images. Our complete data set consists of 60 images showing persons in front of a homogeneous background under viewing angles in the range $-10^{\circ} \ldots 10^{\circ}$. All these images have been recorded under identic conditions, vertically mirrored, lowpass-filtered and downscaled. Then we apply a Sobel operator to enhance the edges of each image. Next all these edge-marked images were averaged and thresholded resulting in a $16x16$ binary matrix Λ. What is further of interest is the local orientation along the contour. It is achieved by means of the steerable filters. Subsequently, to search the presence of the visual cue head-shoulder-portrait we utilize a matching-technique based on a similarity measure $m^{(x,y)}$ which, due the π-periodicity of the filters´s outcome, requires the same periodicity:



$$m^{(x,y)} = \frac{\sum_{\substack{i=0 \\ \lambda_{i,j} \neq 0}}^{I-1} \sum_{j=0}^{J-1} m^{*}\left(\lambda_{i,j} - \theta_{MAX}^{(x+i-\frac{I}{2},\, y+j-\frac{J}{2})}\right)}{\mathrm{card}\left(\mathrm{supp}\left(\Lambda\right)\right)}$$

Herein $I=J=16$ is the size of the template Λ, $\lambda_{i,j}$ an element of it, $\theta^{(\cdot)}_{MAX}$ the value of the pointwise dominant orientation within the input image. Between both these values we define an elementary measure $m^{*}(\gamma)$:

$$m^{*}(\gamma) = \frac{\cos\left(2\left|\gamma\right|\right) + 1}{2}$$

The cardinality of the support of the matrix Λ exactly determines the number of its non-zero elements. This normalization provides $m^{(x,y)}$ to be included in the range $[0,1]$ for further processing steps.

3D Dynamic Fields

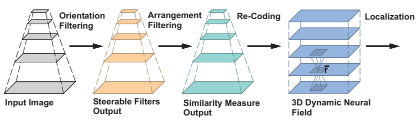
To achieve a good localization, a selection mechanism is needed to make one definite choice among those regions within the pyramid where rather high similarity measures are heaped. Since dynamic neural fields are powerful for dynamic selection and pattern formation using simple homogeneous internal interaction rules, we adapted them for our pourposes. Because we use five fine to coarse resolutions in our scale space we actually can localize a person even at different distances. Therefore, a neural field for selecting the most salient region should be three-dimensional. That field can be described as a recurrent nonlinear dynamic system by means of the following equation:

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

Herein r denotes the three-dimensional coordinate of a neuron position in the field, z is the activation of a neuron at location r, x denotes the external input corresponding to the re-coded similarity measure m, and the constants c_h, c_l und c_i represent parameters of the system. The function w ist the Mexican-hat-like function of lateral activation of neuron r from the surrounding neighbourhood, y is its output activity computed as a sigmoidal function and finally, h is the global inhibition gathering the activity from each neuron over the entire field.

$$w(\vec{r} - \vec{r}\,') = 2\exp\left(\frac{-3|\vec{r} - \vec{r}\,'|^2}{2\sigma^2}\right) - \exp\left(\frac{-|\vec{r} - \vec{r}\,'|^2}{\sigma^2}\right) \qquad h(t) = \int_{\mathbf{F}} y(\vec{r}\,'', t) d^3\vec{r}\,'' \qquad y(\vec{r}, t) = \left(1 + \exp(-z(\vec{r}, t))\right)^{-1}$$

Here on the right, the processing steps for person localization are depicted. Starting from a multi-resolution representation of an input image, each level is treated by steerable filters. Applying the filter arrangement we determine a distance measure wich is taken as input for the 3D field of dynamic neurons. The resulting blob (locally delimited pattern of active neurons) is used to localize a person.



Results

The localization results in an indoor environment are qualitatively illustrated. In this case the person sharply appears at one of the pyramidal planes, the originating spatial blob (rightmost column) was just developed on the topmost of the five planes. However, the blob is not necessarily restricted to a single plane. Each row of the figure displays the result of one of the five computed resolution steps. In the seven columns are depicted: input image, result of the orientation filtering for the angle ranges $(-22.5^{\circ}..22.5^{\circ}]$, $(22.5^{\circ}..67.5^{\circ}]$, $(67.5^{\circ}..112.5^{\circ}]$, $(112.5^{\circ}..157.5^{\circ}]$, similarity measure outcome m, and the result of the selection task. The images of the rightmost column show the state of the five layers of the dynamic neural field 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.

Presented at the IAPR Workshop on Machin Vision Applications (MVA'98), November 17 - 19, 1998, Chiba, Japan