Matthias Zobel, Joachim Denzler, Heinrich Niemann

Coupling Rays – Probabilistic Modeling of Spatial Dependencies

appeared in:
International Conference on Imaging Science, Systems, and Technology (CISST'99)
Las Vegas, Nevada
p. 416–422
1999

International Conference on Imaging Science, Systems, and Technology (CISST'99), Las Vegas, Nevada, p. 416–422

Coupling Rays – Probabilistic Modeling of Spatial Dependencies

M. Zobel, J. Denzler, H. Niemann
Chair for Pattern Recognition (Computer Science 5)
University Erlangen–Nuremberg
Martensstr. 3, 91058 Erlangen, Germany
{zobel,denzler,niemann}@informatik.uni-erlangen.de

Abstract *In this paper we show, how modeling of spatial dependencies between single parts can be used to improve the robustness of the localization of multi–part objects. Spatial dependencies are described by a probabilistic modeling of the features' locations, and connecting them by "coupling rays" into a so called "coupled structure". The approach is embedded into a completely probabilistic framework which allows generalization to multi–part objects of any kind. We describe how the localization process can be mapped onto a corresponding energy minimization problem. An outline is sketched for the tracking of coupled structures in image sequences over time. Finally, the approach is applied to the problem of localizing facial features and experimental results are presented.*

Keywords: coupled features, probabilistic model, MAP based localization, energy minimization, facial features

1 Motivation

Localization and tracking of objects is one major problem in computer vision. Examples are video surveillance, multi media application, autonomous driving and — for a few years — augmented reality. Despite the fact that most objects can be divided into different parts, object localization and tracking is mostly done in a holistic manner. This means that primitives are extracted in the image (for example, edges, corners, or regions) which are taken to model the whole object. Such an approach neglects the fact that a couple of important and significant parts of the object could be more easily detected than the whole object itself in one step. For a multi–part approach it is more natural to define so called *belief sensors* that are specialized on localizing a certain part of the object. One example is to find a face in an image. This can be done by looking for the two eyes and the mouth whose positions are not independent from each other. The problem that needs to be solved now in such a multi–part approach is how to make use of the a priori known spatial relationships between the different parts.

In this contribution we show that the localization of an object consisting of multiple parts that have known spatial interpart relationships, can be done by solving an optimization, i.e. energy minimization problem. The main point is a *probabilistic model* that represents the spatial dependencies. For finding the locations of the features, one has to determine those parameters of the model that maximize the *a posteriori* probability (MAP) of the model conditioned by the current data.

The work which has inspired us mostly, is the one on feature networks in [4]. There, the coupling of certain features as well as the composition of higher level geometric constraints is used to improve the accuracy of tracking. But in contrast to [4], we use a concrete model that is completely embedded into a probabilistic framework. It is shown that the probabilistic model is strongly associated to the elastic, deformable contour model

This work was supported by the "Deutsche Forschungsgemeinschaft" under grant SFB603/TP B2. Only the authors are responsible for the content.

in the field of active contours [6]. The elastic coupling of features was introduced in [3] for facial feature tracking by means of springs, and it was later used in [9] in the context of deformable templates.

Our work reduces the whole estimation process to an energy minimization problem. It can also be compared with active, elastic contours, if the contour points are substituted by higher level features; to localize faces, these features may represent the two eyes and the mouth (cf. Section 3). The values of the model parameters, representing the spatial dependencies, can be estimated in a training step. In our current work, this is done by using a labeled training set. For this, the probabilistic framework is advantageous because of the rich theory already available for parameter estimation, and the possibility of handling uncertainty, caused by noisy data.

This paper is organized as follows: first, the probabilistic model, called *coupled structure*, is introduced in Section 2.1, together with a maximum a posteriori approach to localize a multi–part object. It is shown how the model can be build up from single so called *coupling rays*. Also a short outlook is given on how tracking of coupled structures over time can be embedded into the general probabilistic framework, too. In Section 3 the presented approach is applied to localizing facial features. Finally, we present experimental results from a large set of face images and manually high distorted images in Section 4. The results show the accuracy and reliability of such a probabilistic coupled structure, even for the case of very noisy images.

2 Coupled Structure for Object Localization

2.1 Probabilistic Model

The model that is described in the following, is based on the active rays approach that has been successfully used for contour based object tracking [1]. There, a 2–D contour is represented by different 1–D rays, which originate from one reference point that lies inside the contour. Now, instead of interpreting a point on a ray as a candidate for a contour point, it can be generally seen as the location of any given feature. The concept of a contour in the image plane, which is represented by a given set of rays, is therefore replaced by a general concept that we call *coupled structure*.

The position of a certain feature is given by a *coupling ray* $\boldsymbol{\varrho}_i = (\lambda_i, \phi_i)^T$ with length λ_i and angle ϕ_i. The pose of the ray is determined by the angle ϕ_i measured with respect to a given reference line in the image (usually the horizontal line). All coupled rays originate in a common point called the *coupling center* $\boldsymbol{m} = (m_x, m_y)^T$ with its image coordinates m_x and m_y (s. Figure 1). So the model, i.e. the coupled structure $\boldsymbol{s}$ is defined by the n coupling rays and the coupling center

$$\boldsymbol{s} = (\boldsymbol{\varrho}_1, \ldots, \boldsymbol{\varrho}_n, \boldsymbol{m})^T.$$

Because of the fact that the locations of the features of the objects under consideration often change slightly (think of a non-rigid motion of a face) and that the detection of features is distorted by noise, it is reasonable to regard the important quantities of the model in a probabilistic way. This can be done by modeling the variations in the concrete values of the lengths λ_i and angles ϕ_i of a ray $\boldsymbol{\varrho}_i$ by an appropriate probability density function

$$p_{\boldsymbol{\varrho}_i}(\lambda_i = l, \phi_i = \varphi | \boldsymbol{\varrho}_i).$$

This representation is intended to show explicitly the generality of the approach. For example, it can be thought of features that have more than one plausible location along a certain ray. So the necessity may arise to use multi-modal probability density functions. It is worth noting that $\boldsymbol{s}$ may have more than one coupling center $\boldsymbol{m}$ and that the description can be extended to the 3–D case by using 3–D rays. Here, the description is restricted to the case of only one coupling center and to features lying in one plane.

2.2 MAP Based Localization

Now, we treat the coupling structure $\boldsymbol{s}$ as a random vector in $\mathbb{R}^{2n+2}$. Then, a maximum a posteriori estimation for localizing the object can be applied. Spoken in different words, one has to seek for a

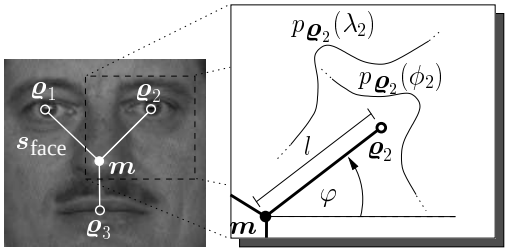

Figure 1: The coupled structure with three coupling rays is shown as it was used for modeling the spatial relations between facial features. The right side shows a magnification of one ray to explain the quantities.

parameter set $\boldsymbol{s}^* = (\boldsymbol{\varrho}_1, \ldots, \boldsymbol{\varrho}_n, \boldsymbol{m})^T$ which maximizes the posterior distribution $p(\boldsymbol{s}|\boldsymbol{f})$ of $\boldsymbol{s}$ conditioned on the image $\boldsymbol{f}$. Using Bayes' rule one gets

$$p(\boldsymbol{s}|\boldsymbol{f}) = \frac{p(\boldsymbol{f}|\boldsymbol{s})p(\boldsymbol{s})}{p(\boldsymbol{f})}, \tag{1}$$

where $p(\boldsymbol{f}|\boldsymbol{s})$ denotes the sensor model and $p(\boldsymbol{s})$ the prior of observing a certain configuration of the model. In a given reference coordinate system we can calculate $p(\boldsymbol{s})$ by

$$p(\boldsymbol{s}) = p(\boldsymbol{\varrho}_1) \cdot p(\boldsymbol{\varrho}_2) \cdot \cdots \cdot p(\boldsymbol{\varrho}_n) \cdot p(\boldsymbol{m}). \tag{2}$$

The independence assumption in (2) is valid, since the dependencies between different rays are implicitly given by the common coupling center $\boldsymbol{m}$. The joint probability

$$p(\boldsymbol{\varrho}_i) = p(\lambda_i|\phi_i)p(\phi_i)$$

must be estimated from the data in the model generation process.

If the model undergoes a transformation $\mathcal{T}$, for example, a rotation in the image plane, the corresponding density $p(\mathcal{T}\boldsymbol{s})$ is given by

$$p(\mathcal{T}\boldsymbol{s}) = |\det(J_{\mathcal{T}^{-1}}(\boldsymbol{s}))|\, p(\mathcal{T}^{-1}\boldsymbol{s}) \tag{3}$$

with $J_{\mathcal{T}^{-1}}$ being the Jacobian of the transformation $\mathcal{T}^{-1}$. A simple and useful transformation may also be a global scaling operation, which influences only the length λ_i of the ray $\boldsymbol{\varrho}_i$.

To model the sensor characteristic $p(\boldsymbol{f}|\boldsymbol{s})$, a common method is applied. We express the correspondence of the model $\boldsymbol{s}$ with the sensor data $\boldsymbol{f}$, i.e. the probability of observing $\boldsymbol{f}$ given the model, by a Gibbs distribution of the form

$$p(\boldsymbol{f}|\boldsymbol{s}) = \frac{1}{z_{\text{ext}}} \exp\left[-E_{\text{ext}}(\boldsymbol{f}, \boldsymbol{s})\right] \tag{4}$$

with z_{ext} being a normalizing constant. The term $E_{\text{ext}}(\boldsymbol{f}, \boldsymbol{s})$ can be interpreted as an *external energy* and needs to be specified dependent on the application. It should return high positive values for image data which do not correspond to the model, and low positive values for good matches.

Now, the estimation of the unknown parameter $\boldsymbol{s}^*$ can be described as an MAP estimation

$$\boldsymbol{s}^* = \underset{\boldsymbol{s}}{\operatorname{argmax}} \frac{p(\boldsymbol{f}|\boldsymbol{s})p(\boldsymbol{s})}{p(\boldsymbol{f})}. \tag{5}$$

2.3 MAP Based Tracking

For tracking a coupled structure $\boldsymbol{s}_t$ with time index t a MAP based approach can be applied again. For that we assume that the object dynamics can be described as a temporal Markov chain, i.e.

$$p(\boldsymbol{s}_t|\boldsymbol{s}_{t-1}, \ldots, \boldsymbol{s}_0) = p(\boldsymbol{s}_t|\boldsymbol{s}_{t-1}).$$

We also assume the image data $\boldsymbol{f}_t$ to be independent, both mutually and with respect to the dynamical process , i.e.

$$p(\boldsymbol{f}_{t-1}, \ldots, \boldsymbol{f}_0, \boldsymbol{s}_t|\boldsymbol{s}_{t-1}, \ldots, \boldsymbol{s}_0) = \\ = p(\boldsymbol{s}_t|\boldsymbol{s}_{t-1}) \prod_{i=0}^{t-1} p(\boldsymbol{f}_i|\boldsymbol{s}_i).$$

Now, equation (1) becomes

$$p(\boldsymbol{s}_t|\boldsymbol{f}_t, \ldots, \boldsymbol{f}_0) = \frac{1}{z_t} p(\boldsymbol{f}_t|\boldsymbol{s}_t) p(\boldsymbol{s}_t|\boldsymbol{f}_{t-1}, \ldots, \boldsymbol{f}_0)$$

where

$$p(\boldsymbol{s}_t|\boldsymbol{f}_{t-1}, \ldots, \boldsymbol{f}_0) = \\ = \int_{\boldsymbol{s}_{t-1}} p(\boldsymbol{s}_t|\boldsymbol{s}_{t-1}) p(\boldsymbol{s}_{t-1}|\boldsymbol{f}_{t-1}, \ldots, \boldsymbol{f}_0).$$

The term z_t is a normalizing constant, which does not depend on $\boldsymbol{s}_t$. The treatment of the dynamical process looks quite complicated. One way to handle this is to make use of the CONDENSATION algorithm [5], which allows an efficient propagation of the conditional density $p(\boldsymbol{s}_t|\boldsymbol{f}_t,\ldots,\boldsymbol{f}_0)$ over time.

The dynamical case is not considered here any further. In the following subsection we give concrete examples of the model $\boldsymbol{s}$ in the area of localizing facial features as well as concrete terms for the prior $p(\boldsymbol{s})$ and the sensor model $p(\boldsymbol{f}|\boldsymbol{s})$.

3 Application to Localizing Facial Features

To localize the facial features eyes and mouth, it is intuitive to model their spatial dependencies by a coupled structure $\boldsymbol{s}_{\text{face}}$ that consists of *three* coupling rays with the coupling center being the tip of the nose. There is one coupling ray for each eye and one for the mouth (cf. Figure 1).

Since there is only one reasonable position for each facial feature in a face, the length and the angle of each ray are regarded as Gaussian distributed random variables, i.e.

$$p_{\boldsymbol{\varrho}_i}(\lambda_i = l) \sim \mathcal{N}({}^{\lambda}\mu_i, {}^{\lambda}\sigma_i^2), \text{ and}$$

$$p_{\boldsymbol{\varrho}_i}(\phi_i = \varphi) \sim \mathcal{N}({}^{\phi}\mu_i, {}^{\phi}\sigma_i^2).$$

Therefore it is sufficient to specify the two means ${}^{\lambda,\phi}\mu_i$ and the two variances ${}^{\lambda,\phi}\sigma_i^2$ of this distributions for each ray $\boldsymbol{\varrho}_i$. They are obtained by segmentation of a sample set of images taken from frontal views of different persons.

For the prior $p(\boldsymbol{s}_{\text{face}})$ in equation (2) it is necessary to specify explicitly $p(\boldsymbol{\varrho}_i)$. For the joint probability density function $p(\lambda_i, \phi_i)$ we write

$$p(\boldsymbol{\varrho}_i) = p(\lambda_i)p(\phi_i).$$

This independence assumption was verified by applying the χ^2 test to data from 339 face images. Thus, we get for the prior $p(\boldsymbol{s}_{\text{face}})$ of our model parameters

$$p(\boldsymbol{s}_{\text{face}}) = p(\boldsymbol{m}) \prod_{i=1}^{3} p(\lambda_i)p(\phi_i).$$

Assuming a Gaussian distribution of the two parameters λ_i and ϕ_i as mentioned earlier and an uniform distribution $p(\boldsymbol{m})$ over the image plane, i.e. no knowledge is used about the position of the face in the image, we get a distribution of the form

$$p(\boldsymbol{s}_{\text{face}}) = \frac{1}{z_{\text{int}}} \exp\left[-E_{\text{int}}(\boldsymbol{s}_{\text{face}})\right],$$

where z_{int} is a normalizing constant and

$$E_{\text{int}}(\boldsymbol{s}_{\text{face}}) = \sum_{i=1}^{3} \frac{({}^{\lambda}\mu_i - \lambda_i)^2}{{}^{\lambda}\sigma_i^2} + \frac{({}^{\phi}\mu_i - \phi_i)^2}{{}^{\phi}\sigma_i^2}.$$

The term $E_{\text{int}}(\boldsymbol{s}_{\text{face}})$ can be interpreted as an *internal energy* of the model [7], that is low for configurations that are similar to the modeled mean and high for large deviations.

Thus the MAP approach can be seen as an energy minimization problem, with a term E_{int} describing the deformation ability of the model and a second term E_{ext} (cf. Eq. (4)) given by the image data conditioned on the model.

In the following we use a straight forward approach for the external energy definition because of the observation, that high vertical energies in an image can be used to identify the unknown positions of the facial features. One can think of more sophisticated features, but this is beyond the scope of this paper.

The energies in the image are computed by using the DCT (discrete cosine transformation) that is supported in hardware by many of todays frame grabbers. To get the vertical energies $b_v(j,k)$ from each 8×8 DCT block (j,k) the entries of the first and second column of each DCT block are summed [8]. Applied to the coupled structure for each ray $\boldsymbol{\varrho}_i$ a certain rectangular area $\mathcal{A}_i(\boldsymbol{\varrho}_i)$ with its center at (λ_i, ϕ_i) is defined, for which the vertical energies $b_v(j,k)$ of DCT blocks are summed up; that results in an total external energy

$$E_{\text{ext}}(\boldsymbol{f}, \boldsymbol{s}_{\text{face}}) = \sum_{i=1}^{3} \frac{1}{\sum\limits_{(j,k)\in\mathcal{A}_i(\boldsymbol{\varrho}_i)} b_v(j,k)}. \quad (6)$$

With the prior of the model (2) and the sensor model (4) defined by the external energy (6) the unknown parameter set $\boldsymbol{s}^*_{\text{face}}$ can be determined using (5).

Determining $\boldsymbol{s}^*_{\text{face}}$, i.e. localizing the facial features, was implemented by means of a scalable search algorithm. The algorithm works directly on the positions of the facial features in the energy map. From these positions the parameters of the coupled structure are determined afterwards.

The coarse structure of the search algorithm can be outlined as follows. First the algorithm creates a list $\mathcal{L}$ of the entries in the energy map. Each entry $l = (j, k)$ in the list stores the indices j and k of the corresponding entry in the energy map. For all triples $(l_1, l_2, l_3) \in \mathcal{L}\times\mathcal{L}\times\mathcal{L}$ the total energy of the corresponding coupled structure can be computed using l_1 as the location of the left eye, l_2 as the location of the right eye and l_3 as the location of the mouth. The best triple represents directly the locations of the facial features in the energy map, i.e. the coupled structure with the lowest total energy. It is clear that this global search is not applicable. But fortunately in our case, the search space can be restricted drastically.

First, the list $\mathcal{L}$ can be sorted by decreasing energy. Since entries with high energy values are good candidates for presenting the location of a facial feature, these entries come first in $\mathcal{L}$. Second, not the whole list is used to build the triples, only the first n entries of the list are used. Already the selection of the 50 best entries is sufficient to perform a good, but maybe suboptimal, localization. As it can be seen in Table 1, for $n = 50$, a good trade–off between computational effort and accuracy is achieved. By using knowledge about the task domain, here localization in frontal views of faces, the search can be accelerated further. So we do not need to examine triples where the right eye is on the left of the left eye, or the mouth is above the eyes, etc.

4 Experimental Results

To demonstrate the applicability of the proposed approach a sample set of 335 face images was used. The positions of the eyes and the mouth were manually labeled in each image of the sample set. For each of the sample images we created an energy map containing the vertical energies $b_v(j, k)$ as they are needed to compute the external energy in (6). Since the vertical energies result from 8×8 DCT blocks, the spatial resolution in localizing the facial features in the original image is also limited to 8×8 pixel blocks.

best n	3	5	10
$\mu_{\boldsymbol{s}}$	171.51	147.84	62.54
runtime [ms]	16	28	38
best n	20	50	100
$\mu_{\boldsymbol{s}}$	7.74	3.96	3.90
runtime [ms]	354	8762	75337
best n	150	200	
$\mu_{\boldsymbol{s}}$	3.66	3.45	
runtime [ms]	260640	637040	

Table 1: Mean error $\mu_{\boldsymbol{s}}$ of the coupled structure (in 8×8 blocks), depending on the number n of the selected first entries in the sorted list $\mathcal{L}$. Runtimes are measured on a Pentium II with 333 MHz.

The accuracy of the coupled structure approach was tested by dividing the sample set into a training part that was used to estimate the parameters, and into a test part that was used for evaluation. To judge the quality of the results depending on the number of training images, the whole sample set was divided randomly into five subsets of equal size. These subsets were systematically combined to build training and test sets of different sizes. First, we performed experiments with one subset for training and four subsets for evaluation. Second, we used two subsets for training and the remaining three subsets for testing, and so on. All these experiments were done twice with different partitions of the whole sample set. Thus, a total number of 10050 localizations were conducted.

The quality of the facial feature extraction by the coupled structure was judged by computing the distances between the estimated position of the two eyes and the mouth, and the true position, obtained from the labeled sample set.

The experiments were performed by using the scalable search algorithm, described in Section 3. Some statistics of the experiment with the best mean error $\mu_{\boldsymbol{s}}$ of the whole structure of 3.52 blocks is given in Table 2. The maximal total mean error was 4.94 blocks. The mean error over all experi-

	μ_i	σ_i	$\min_i$	$\max_i$
Left eye	0.99	0.70	0.00	2.83
Right eye	1.03	0.91	0.00	4.24
Mouth	1.50	1.13	0.00	6.32
$\mu_s = \sum_{i=1}^{3} \mu_i$	3.52			

Table 2: Euclidean error for coupled localization using 268 training and 67 test images. For each facial feature the mean, standard deviation, minimal and maximal error in units of 8×8 blocks is given.

ments was 4.04 blocks with a standard deviation of 0.27 blocks. In Figure 2 we show some example results from localizations of facial features from images of the sample set.

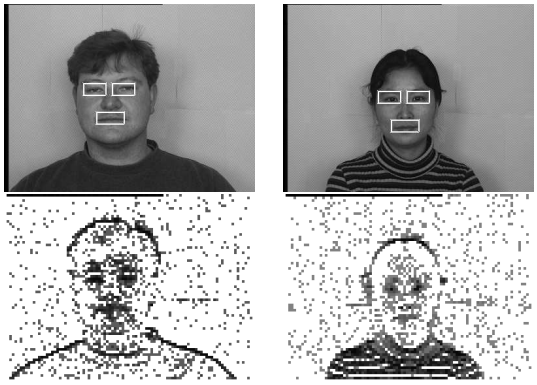

Figure 2: Two example images from the sample set with their energy maps. The localized facial features are marked by white boxes.

Although, the search algorithm yields good results, it is not applicable to situations where the feature detection is highly distorted by noise, because the distortion can cause the corresponding DCT block indices to not appear in the first part of $\mathcal{L}$, and so it is not considered as a potential candidate for a facial feature location. An alternative to handle high distortions is to use a random based global search procedure, like the *adaptive random search* (ARS) algorithm [2].

The robustness of our coupled approach is demonstrated by applying ARS to manually highly distorted face images. The results show, that because of the use of an internal energy in our coupled model, the distortions can mostly be neutralized, so they do not affect the localization process (Figure 3).

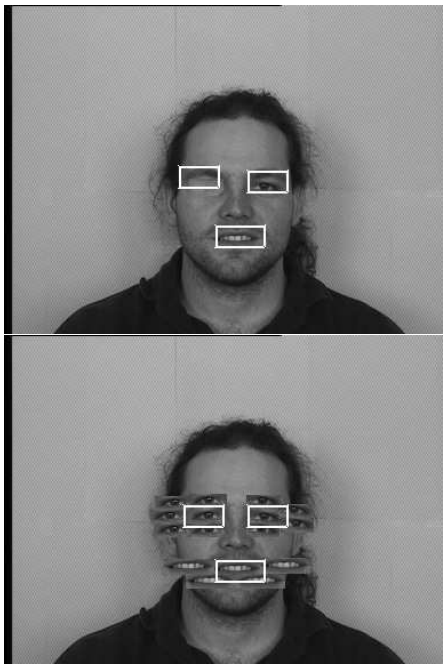

Figure 3: Results for artificially highly distorted face images. No left eye visible (top), more than one mouth and more than two eyes visible (bottom).

5 Conclusion

In the paper we describe a probabilistic method for modeling the spatial dependencies between multiple parts of objects. This leads to a robustness in the localization of the whole object in the case of distortions, wrong measurements, or uncertainty in the feature computation.

The experiments show, that each facial feature can be localized with an error less than 1.5 blocks. The advantage of the spatial modeling becomes obvious in the case of missing features due to occlusions or noisy data. The result itself is quite promising, just because the external energy is simple and one can think of a more specific one.

Summarizing the approach, we like to emphasize that the idea of coupling different features of an object is natural and not new — as mentioned while giving the bibliography review in Section 1.

Nevertheless, a complete formalization of this idea in a probabilistic framework, as given in the paper, has not been done until now. The main advantages arise from

1. the abstract description of the coupled structure, which will include 3–D objects in our future work; the position in 3–D can be estimated by integrating the transformation $\mathcal{T}$ (cf. Eq. 3) in the parameter estimation process (5).

2. the possibility to use multi–modal densities for describing the position of a certain feature,

3. the possibility to define different sensor models for each feature. In our case, this is demonstrated by the size of the rectangular area $\mathcal{A}_i(\boldsymbol{\varrho}_i)$, which differs between the two eyes and the mouth.

In our future work we will focus on the integration of 3–D information, to track rotating faces, too. There, we expect some problems with the computational effort in the practical realization of the MAP estimation by energy minimization. Additionally, we will apply more sophisticated sensor models to identify the facial features. Finally, the approach has to be applied to a different domain, to show its generality.

References

[1] J. Denzler, B. Heigl, and H. Niemann. An efficient combination of 2d and 3d shape description for contour based tracking of moving objects. In H. Burkhardt and B. Neumann, editors, *Computer Vision - ECCV 98*, pages 843–857, Berlin, Heidelberg, New York, London, 1998. Lecture Notes in Computer Science.

[2] S. M. Ermakov and A. A. Zhiglyavskij. On random search of global extremum. *Probability Theory and Applications*, 28(1):129–136, 1983.

[3] M.A. Fischler and R.A. Elschlager. The representation and matching of pictorial structures. *IEEE Transactions on Computers*, 22(1):67–92, 1973.

[4] G.D. Hager and K. Toyama. X vision: Combining image warping and geometric constraints for fast visual tracking. In A. Blake, editor, *Computer Vision - ECCV 96*, pages 507–517, Berlin, Heidelberg, New York, London, 1996. Lecture Notes in Computer Science.

[5] M. Isard and A. Blake. Contour tracking by stochastic propagation of conditional density. In A. Blake, editor, *Computer Vision - ECCV 96*, pages 343–356, Berlin, Heidelberg, New York, London, 1996. Lecture Notes in Computer Science.

[6] M. Kass, A. Wittkin, and D. Terzopoulos. Snakes: Active contour models. *International Journal of Computer Vision*, 2(3):321–331, 1988.

[7] D. Terzopoulos and R. Szeliski. Tracking with Kalman snakes. In A. Blake and A. Yuille, editors, *Active Vision*, pages 3–20. MIT Press, Cambridge, Massachusetts, London, England, 1992.

[8] H. Wang and S.-F. Chang. A Highly Efficient System for Automatic Face Region Detection in MPEG Video. *IEEE Transactions on Circuits and Systems for Video Technology*, 7(4):615–628, August 1997.

[9] A. Yuille and A. Blake. Deformable templates. In A. Blake and A. Yuille, editors, *Active Vision*, pages 21–38. MIT Press, Cambridge, Massachusetts, London, England, 1992.