Matthias Zobel, Joachim Denzler, Heinrich Niemann
Tracking of Probabilistically Coupled Features

appeared in:
Vision Modeling and Visualization ’99 (VMV’99)
Erlangen, Germany
p. 133–140
1999

Tracking of Probabilistically Coupled Features

M. Zobel*, J. Denzler, H. Niemann

University of Erlangen, Chair for Pattern Recognition
Martensstraße 3, D-91058 Erlangen, Germany
Email: {zobel,denzler,niemann}@informatik.uni-erlangen.de

Abstract

In this paper an approach on tracking objects consisting of multiple parts is presented. Instead of tracking each part independently, all features are tracked simultaneously. Therefore, the spatial dependencies of the object's parts are described by a probabilistic model that is called "coupled structure". The tracking process is performed using an algorithm for propagating conditional probability density function over time, called CONDENSATION algorithm. The main advantage of our approach is, that object modeling and object tracking are embedded into a completely probabilistic framework, so uncertainty can be handled very powerful and elegant.

Finally, we demonstrate the applicability of our approach by tracking a moving human face in a difficult environment and give some experimental results.

1 Motivation

Localization and tracking of objects is one major problem in computer vision. Examples are video surveillance, multi media application, autonomous driving and robots and augmented reality. Most objects consist of different parts, for which specialized feature detectors exist being able to optimally localize those part. Dispite that fact, 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 used to model the whole object. Such an approach neglects the fact that sometimes a couple of important and significant parts of the object could be more easily detected than the whole object itself in one step. Thus, for a multi–part approach it is more natural to define specialized feature detectors that can better localize a certain part of the object than a general feature detector is able to localize the whole object. This strategy alone will of course not yield an improved result. In addition, the different parts must be coupled using a priori information about their spatial relationships. One example to support this statement 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 and tracking of an object consisting of multiple parts that have known spatial interpart relationships, can be solved completely in a probabilistic framework. 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. To track the features, not only a single state, but the whole probability density over the object's state conditioned on the measurements obtained while processing the image sequence is propagated over time. Compared with the well known Kalman filter, this has the advantage, that also multimodal densities can be used, i.e. mul-

*This work was supported by the DFG under grant SFB603

tiple hypotheses of the object's state are treated inherently.

For the coupling of multiple features, the most related work is the one on feature networks in [6]. 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 [6], we use a concrete model that is completely embedded into a probabilistic framework.

Our work reduces the whole MAP 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 4). 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. It is shown how the model can be build up from single so called *coupling rays*. Then, for the dynamic case, i.e. the tracking of such a coupled structure, the CONDENSATION algorithm for propagating the densities is summarized in Section 3. By applying our approach to track faces in an image sequence, the complete framework is illustrated in Section 4. There, the probabilistic models — for the object, for the object's motion, and for the measurement process — are presented, and brought together to track a face by coupling the facial features eyes and mouth. Finally, we present experimental results on a sequence of face images in Section 5. The results show the feasibility and the robustness of our probabilistic feature coupling in case of difficult environments.

2 Coupled Structures

In this section we introduce our model based on the active rays approach that has been successfully used for contour based object tracking [4]. 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.

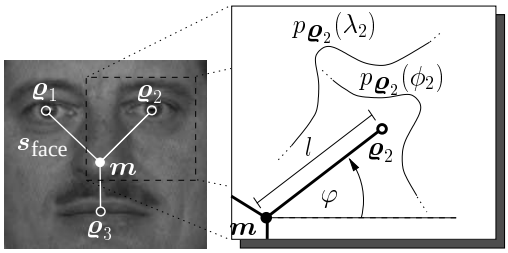

Figure 1: The coupled structure with three coupling rays is shown as it was used for modeling the spatial relations between facial features. The right side shows a magnification of one ray to explain the quantities.

3 Density Propagation

During tracking of a moving object a lot of information is collected about the a priori unknown trajectory of the object in 2–D or 3–D. This information should be used to the increase quality of the object localization at each time step. Usually, the position of the object in the next frame is predicted and this position is used as an initial starting point for object localization. In our case, the predicted position can be used as the initial parameters of an local optimization step during energy minimization.

For prediction, especially when dealing with moving objects, the Kalman filter is a well understood framework [1, 8]. It has been widely used in computer vision with focus on motion detection, motion computation and tracking [10, 3, 9]. To use the Kalman filter one needs a state transition model, describing the dynamics of a state $\boldsymbol{q}_t$ over time, and an observation model. Both are described as a deterministic linear or nonlinear function distorted by Gaussian noise. The Kalman filter is a linear predictor, which linearly combines at time step t the estimated state $\hat{\boldsymbol{q}}_t$ with the error between the estimated observation $\hat{\boldsymbol{f}}_t$ and the true one $\boldsymbol{f}_t$, weighted by the Kalman gain matrix. This linear prediction leads to an unimodal density $p(\boldsymbol{q}_t|\boldsymbol{f}_t, \ldots, \boldsymbol{f}_0)$ over the state space $\boldsymbol{q}$ conditioned on the observation made up to the actual time step t; the mean of this density corresponds to the estimated state.

In computer vision, there is often the need of handling more than one hypothesis for an unknown state. In most cases, due to sensor noise and uncertainty in the measurements more than one observation can be made that could be caused by the system whose state should be estimated. This is called the association problem. Thus, an unimodal density over the state is not appropriate. In [7] a new approach is presented, called conditional density propagation (CONDENSATION). A complete mathematical framework is provided which allows the propagation of a density with more than one mode over time. Each mode corresponds to one hypothesis for the unknown state. The main principle CONDENSATION is based on is *factored sampling* [5], which allows the computation of a density $p(\boldsymbol{x})$ that can be factored into two parts

$$p(\boldsymbol{x}) = p_1(\boldsymbol{x})p_2(\boldsymbol{x}).$$

If the density $p(\boldsymbol{x})$ cannot be given in analytical form, but there is the possibility to simulate $p_1(\boldsymbol{x})$, then $p(\boldsymbol{x})$ can be estimated by drawing N samples randomly with probability

$$p_i = \frac{p_2(x_i)}{\sum_{j=1}^{N} p_2(x_j)}$$

from a sample set of N samples x_i which were generated by $p_1(\boldsymbol{x})$. For $N \to \infty$ the resulting density converges weakly to $p(\boldsymbol{x})$ [5].

The correctness of the factored sampling theorem has also been proven for the dynamic case in [7]. The factor $p_2(\boldsymbol{x})$ corresponds to the observation density $p(\boldsymbol{f}|\boldsymbol{q})$, and $p_1(\boldsymbol{x})$ to the a priori density of the unknown state $p(\boldsymbol{q})$. Each sample, drawn from the sample set is first deterministically propagated over time by means of the state transition model. Then, the resulting state is stochastically diffused by the state transition noise and evaluated in the image by $p(\boldsymbol{f}|\boldsymbol{q})$. For more details we refer to [7].

As a result, one gets a multi modal probability distribution over the state space, i.e. a so called *belief state*. Note, the true state must not necessarily be the best rated state at time step t_i in terms of $p(\boldsymbol{f}|\boldsymbol{q})$ (for example, due to occlusion), but it does not disappear. At future time steps $t_j > t_i$ the likelihood may increase and so the true state will be the best, i.e. it will be found again.

4 Face Tracking with Coupled Structures

In this section we show how to apply the concepts described in the previous sections to a practical problem, the tracking of a human face, i.e. its facial features in an image sequence. The first thing we need to accomplish this task is a concrete model of the spatial relationships of the facial features represented in form of a coupled structure. With the coupled structure specified, we can apply the CONDENSATION algorithm to track the face over time. During the algorithm hypothetical instances of coupled structures are generated that need to be rated against the model and against the image data. It is shown in the last part of this section, how the application specific parts needed for the propagation algorithm can be modeled by appropriate energy terms.

4.1 The Model

It is intuitive to model the spatial dependencies of the eyes and the mouth of a face by a coupled structure $\boldsymbol{s}$ consisting of *three* coupling rays (cf. Figure 1). The three rays originate from a coupling center that is associated with the tip of the nose. There is one coupling ray for each eye and one for the mouth. The lengths and the angles of each ray can be modeled by Gaussian distributed random variables. This is a reasonable choice, because there is only one possible position for each feature on its coupling ray, i.e.

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

and

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

To characterize a certain face model completely it is sufficient to specify the two means ${}^{\lambda,\phi}\mu_i$ and the two variances ${}^{\lambda,\phi}\sigma_i^2$ of the distributions for each ray $\boldsymbol{\varrho}_i$. They may be set manually, or better, obtained by evaluating a sample set of frontal face images, or from physiological consideration. For the experiments described later they were chosen by hand, by taking a typical image from the sequence and determining the corresponding properties of the face in image.

4.2 Tracking

For tracking of a coupled structure one is especially interested in the timely development of the a posterior probability density function $p(\boldsymbol{s}_t|f_t, \ldots, f_0)$. Assuming that the object dynamics can be described as a temporal Markov chain, i.e.

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

and assuming the image data $\boldsymbol{f}_t$ to be independent, both mutually and with respect to the object's dynamics, i.e.

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

the a posteriori probability density function can be written as

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

where

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

with a normalizing constant z_t.

Because we model a face by three coplanar coupling rays and assume that the plane spanned by the three coupling rays moves only parallel to the image plane, the estimation of the prior at time index t (Eq. 1) can be split into two parts. The main advantage arises from the fact, that in our case, the coupled structure $\boldsymbol{s}$ can be subdivided into a part that varies with time and into a part that is independent from time. The dependent part is the coordinate vector of the coupling center $\boldsymbol{m}_t$. Because of the restricted motion, the model parameters that represent the coupling rays stay equal when time progresses. Especially, the parameters are also invariant under translation, i.e. they are independent on the position of the coupling center $\boldsymbol{m}_t$ (cf. Eq. 4).

So the prior can be rewritten as

$$p(\boldsymbol{s}_t|\boldsymbol{f}_{t-1},\ldots,\boldsymbol{f}_0) = \prod_{i=1}^{3} p(\boldsymbol{\varrho}_i) \cdot \cdot \int_{\boldsymbol{s}_{t-1}} p(\boldsymbol{m}_t|\boldsymbol{m}_{t-1}) p(\boldsymbol{s}_{t-1}|\boldsymbol{f}_{t-1},\ldots,\boldsymbol{f}_0)$$

whereby $\boldsymbol{m}_{t-1}$ is the coupling center from the coupled structure $\boldsymbol{s}_{t-1}$.

With this splitting we can write the a posterior $p(\boldsymbol{s}_t|\boldsymbol{f}_t,\ldots,\boldsymbol{f}_0)$ as

$$p(\boldsymbol{s}_t|\boldsymbol{f}_t,\ldots,\boldsymbol{f}_0) = \frac{1}{z_t} p(\boldsymbol{f}_t|\boldsymbol{s}_t) \prod_{i=1}^{3} p(\boldsymbol{\varrho}_i) \cdot \cdot \int_{\boldsymbol{s}_{t-1}} p(\boldsymbol{m}_t|\boldsymbol{m}_{t-1}) p(\boldsymbol{s}_{t-1}|\boldsymbol{f}_{t-1},\ldots,\boldsymbol{f}_0). \quad (2)$$

Now, we are able to apply the CONDENSATION algorithm to track the development of the a posterior $p(\boldsymbol{s}_t|\boldsymbol{f}_t,\ldots,\boldsymbol{f}_0)$ from Eq. 2. In contrast to the original work in [7] we modify the factored sampling process to take into account the splitting of a coupled structure into time dependent and independent parts. Instead of sampling from the posterior of the previous time step, predicting by the dynamic model, and evaluating the predicted state by the data, we sample first the time dependent part, i.e. the coupling center $\boldsymbol{m}$, from the posterior of the previous time step and predict it by the dynamical model. This leads to an incomplete coupled structure $\widetilde{\boldsymbol{s}}_t$. Then for each such $\widetilde{\boldsymbol{s}}_t$, we sample M times from each of the time independent priors of the coupling rays and evaluate now the $3M$ feature positions from $p(\boldsymbol{f}_t|\boldsymbol{\varrho}_i,\boldsymbol{m}_t)$ (cf. Eq. 3). Note, that it is $3M$ and not M^3, because of the mutually independence of the rays. The parameters of the $3M$ ray samples that maximize the fitness of their corresponding ray to the data, are taken as the parameters of the rays to complete $\widetilde{\boldsymbol{s}}_t$ to become a fully specified $\boldsymbol{s}_t$. This way, the parameters for the rays $\boldsymbol{\varrho}_i$ are chosen so that $p(\boldsymbol{\varrho}_i)p(\boldsymbol{f}|\boldsymbol{\varrho}_i)$ evaluates to the maximal value. Although the described technique for sampling and evaluating is intuitive and straight forward, unfortunately we cannot yet give a formal proof for its correctness, as it is given for the unsplit case in [7].

4.3 Measurements

During the propagation of the probability density functions with the CONDENSATION algorithm, three application dependent parts are used. These parts are given by the probability density functions $p(\boldsymbol{m}_t|\boldsymbol{m}_{t-1})$ (the object's dynamics), $p(\boldsymbol{f}|\boldsymbol{s})$ (the sensor model), and $p(\boldsymbol{s})$ (the object model). Here, the dynamic model is not considered further, because standard techniques can be applied. In the following we describe how the two remaining parts can modeled by mapping them onto equivalent energy terms.

External Energy To model the measurement process $p(\boldsymbol{f}|\boldsymbol{s})$ we use a common method. The correspondence of the model $\boldsymbol{s}$ with the image data $\boldsymbol{f}$ is expressed by a Gibbs distribution

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

with z_{ext} being a normalizing constant. The term $E_{\text{ext}}(\boldsymbol{f},\boldsymbol{s})$ should return high positive values if the image data does not correspond well to the data which is expected, given the coupled structure $\boldsymbol{s}$, and it should return low positive values for good matches. Therefore, this term can be interpreted as a kind of *external energy*.

For the experiments described later a somewhat simple approach is used. One method that directly supports the requirements above is that of template matching [11], i.e. compute the error ϵ between a template $\boldsymbol{T}$ of size $m \times n$ with an image area $\boldsymbol{f}$ at position (k,j) of equal size by

$$\epsilon_{k,j} = \sum_{\mu=0}^{m-1} \sum_{\nu=0}^{n-1} |f_{k+\mu,j+\nu} - T_{\mu,\nu}|$$

The smallest possible value of ϵ is zero in case of a perfect match between $\boldsymbol{f}$ and $\boldsymbol{T}$.

This leads to the external energy for the whole coupled structure, that consists of the sum of the external energies of each of the three coupling rays

$$E_{\text{ext}}(\boldsymbol{f},\boldsymbol{s}) = \sum_{i=1}^{3} E_{\text{ext}}(\boldsymbol{f},\boldsymbol{\varrho}_i,\boldsymbol{m}) = \sum_{i=1}^{3} \epsilon_{\boldsymbol{\varrho}_i} \quad (3)$$

with $\epsilon_{\boldsymbol{\varrho}_i}$ being the value of ϵ_{k_i,j_i} with image coordinates k_i and j_i obtained by the parameters of the ray $\boldsymbol{\varrho}_i$.

In our work, we defined the templates for the both eyes and the mouth by cutting off appropriate image areas from a typical face in the image sequence. It should be noticed, that it was not an aim of our work to develop new features for detecting eyes and mouths. Therefore, the quality of the chosen external energy may not be very sophisticated, but it was very fast to realize. The external energy based on vertical energies provided by the DCT coefficients (cf. [12]) was found to be not specific enough for tracking facial features in an cluttered environment, although it was successfully used for localization in the static case.

Internal Energy As we are working in an probabilistic context, the coupled structure model is described by a probability density function $p(\boldsymbol{s})$. This density function can be calculated in a given reference coordinate system by

$$p(\boldsymbol{s}) = p(\boldsymbol{\varrho}_1) \cdot p(\boldsymbol{\varrho}_2) \cdot \cdots \cdot p(\boldsymbol{\varrho}_n) \cdot p(\boldsymbol{m}), \quad (4)$$

where in general

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

Note that the independence assumption between the rays is reasonable, because the dependencies are implicitly given by the coupling center $\boldsymbol{m}$.

Especially, it was verified by a statistical test, that the joint probability density function $p(\lambda_i, \phi_i)$ can be written as

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

assuming independence of the length and the angle of a coupling ray i. Therefore, the prior $p(\boldsymbol{s})$ in (Eq. 4) is

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

Using the approach with the Gibbs distribution, the prior can be written as

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

with z_{int} being a normalizing constant and the term

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

Here, $E_{\text{int}}(\boldsymbol{s})$ can be interpreted as an *internal energy* that has low positive value in case of little deviation of the ray parameters from their mean values, and that is zero in case of a good match.

Using these two energy terms, a total energy for a coupled structure $\boldsymbol{s}$ can be defined simply as the sum of both. i.e.

$$E_{\text{total}}(\boldsymbol{s}) = E_{\text{int}}(\boldsymbol{s}) + E_{\text{ext}}(\boldsymbol{s}).$$

5 Experimental Results

The feasibility of the approach that was described in the previous sections was tested using an image sequence from a moving head in front of heterogeneous background. The sequence consists of 92 color frames of size 768×576 that were recorded at a frame rate of 12.5 frames per second. The person in front of the camera was only told to face the camera while moving.

The μ_i and σ_i of the model ray parameters were determined from evaluating the whole image sequence. The values of the model parameters are given in Table 1.

	$\boldsymbol{\varrho}_1$	$\boldsymbol{\varrho}_2$	$\boldsymbol{\varrho}_3$
${}^{\lambda}\mu_i$	45.79	41.74	33.76
${}^{\lambda}\sigma_i$	4.30	3.46	3.67
${}^{\phi}\mu_i$	2.48	0.83	4.71
${}^{\phi}\sigma_i$	0.12	0.18	0.07

Table 1: The parameter values of the three rays of the coupled structure. The lengths are given in pixels and the angles in radians.

To predict the coordinates of the coupling center from a time step to the next, a model of the object's dynamics is needed. In our case, we used a simple two dimensional linear second order dynamical model that is distorted by Gaussian noise.

The templates that are needed for the computation of the external energy of a given coupled structure are obtained from a typical frame of the image sequence (cf. Figure 2).

Figure 2: The templates of the facial features used for evaluating the external energy.

To evaluate the accuracy of the tracking process, the true positions of the facial features in the image sequence were manually labeled, so they can be compared to the positions provided by the tracking process. The results for each facial feature are listed in Table 2. The mean error for the combined features is therefore about 21 pixels. For the tracking experiments, we used 500 samples for the coupling centers and then 20 samples for each ray. Therefore, a total number of 30,000 samples were evaluated at each time step. One iteration of the CONDENSATION algorithm took about 50 s on a Pentium II with 300 MHz running Linux. In Figure 3 some example images from the sequence with the found positions of the facial features are depicted.

	μ_ε	σ_ε	$\min_\varepsilon$	$\max_\varepsilon$
Left eye	22.6	18.8	2.0	78.2
Right eye	20.2	19.6	1.0	85.0
Mouth	20.9	19.0	0.0	71.4

Table 2: Euclidean error for each feature over the whole sequence. For each feature, the mean, standard deviation, minimal and maximal error in pixel is given.

An important role in the tracking process plays the technique with which the obtained multimodal a posteriori probability density function $p(\boldsymbol{s}_t|\boldsymbol{f}_t, \ldots, \boldsymbol{f}_0)$ is evaluated at each time step. Taking the mean structure of the sample set, as it was proposed in [7], is not feasible because of the many occurring modes. So we decided to take that structure as the best one that provides the highest value of $p(\boldsymbol{s}_t|\boldsymbol{f}_t, \ldots, \boldsymbol{f}_0)$ in the sample set. This may sometimes lead to misdetection, because of outliers in the sample set.

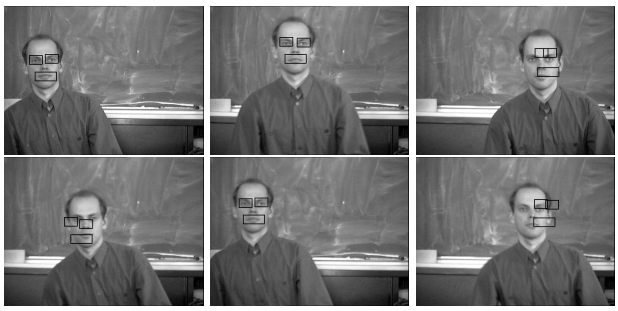

Figure 3: Every 18th image (from left to right) from the test sequence, with the found locations marked by black boxes.

6 Conclusion

We presented an approach for modeling of multi–part objects in a probabilistic framework, for extracting such objects in images by means of maximum a posteriori estimation, and for tracking objects over time by propagating the conditional density of the object's state over time. The experiment have proven the suitability and advantages of the proposed method. For highly distorted images as well as for cluttered background during tracking the probabilistic framework is the most appropriate one to handle uncertainty. This includes both the measurements in the image, and the models for motion and the object itself.

In this contribution no special effort has been spent on choosing the right or most appropriate motion model for the object's trajectory. Such an approach has been presented in [2]. As it was already stated in the text, the feature detectors themselves should be replaced by some more sophisticated and faster computable ones. In our future work, we will also concentrate on learning typical motion trajectories for a certain class of objects. Also, the approach will be applied to a different application, and an extension of the coupled structures to the 3–D case using 3–D rays is planned.

References

[1] Y. Bar-Shalom and T.E. Fortmann. *Tracking and Data Association*. Academic Press, Boston, San Diego, New York, 1988.

[2] A. Blake, M. Isard, and D. Reynard. Learning to track the visual motion of contours. *Artificial Intelligence*, 1995.

[3] C. Brown, H. Durrant-Whyte, J. Leonard, B. Rao, and B. Steer. Distributed data fusion using kalman filtering: A robotics application. In M.A. Abidi and R. C. Gonzalez, editors, *Data Fusion in Robotics and Machine Intelligence*, pages 267–310. Academic Press, 1992.

[4] 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.

[5] U. Grenandar, Y. Chow, and D.M. Keenan. *Hands: A Pattern Theoretic Study of Biological Shapes*. Springer, Berline, 1991.

[6] 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.

[7] M. Isard and A. Blake. Condensation — conditional density propagation for visual tracking. *International Journal of Computer Vision*, 29(1):5–28, 1998.

[8] R.E. Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, pages 35–44, 1960.

[9] D. Koller, K. Daniilidis, T. Thorhallson, and H. Nagel. Model–based object tracking in traffic scenes. In G. Sandini, editor, *Computer Vision - ECCV 92*, pages 437–452, Berlin, Heidelberg, New York, London, 1992. Lecture Notes in Computer Science.

[10] L. Matthies, R. Szeliski, and T. Kanade. Kalman filter–based algorithms for estimating depth from image sequences. *International Journal of Computer Vision*, 3(3):209–236, 1989.

[11] H. Niemann. *Pattern Analysis and Understanding* Second Edition. Number 4 in Springer Series in Information Sciences. Springer-Verlag, Berlin, 1990.

[12] M. Zobel, J. Denzler, and H. Niemann. Coupling rays – probabilistic modeling of spatial dependencies. In *International Conference on Imaging Science, Systems, and Technology (CISST'99)*, pages 416–422, Las Vegas, Nevada, 1999.