A FLEXIBLE MULTIMODAL OBJECT TRACKING SYSTEM

Harald Breit and Gerhard Rigoll

Institute for Human-Machine-Communication
Munich University of Technology, Germany
http://www.mmk.e-technik.tu-muenchen.de/

ABSTRACT

In this paper we present a flexible multimodal object tracking system. It is based on a particle filter which combines the outputs of different measurement methods (also called modes or cues) in a flexible manner. The modes for locating the desired object can be selected depending on the specific object which is to be tracked and the environment of the object. By combining multiple modes we aim to add the strengthes while at the same time to overcome the specific disadvantages of the different modes. Thus the robustness of the tracking system will be increased and a successful tracking will be possible in critical situations where a system using only a single mode would fail. We have used this approach for tracking persons in different environments. The measurement modes which we have implemented for this purpose are a pseudo 2-dimensional hidden Markov model (P2DHMM), a color based skin finder, and a motion detector. We describe the theory and the architecture of this tracking system and finally depict some exemplary results.

1. INTRODUCTION

The tracking of moving objects in video sequences is a major problem in the area of visual surveillance and vision-based man-machine-interfaces. We have proposed approaches where the main goal was the possibility to track persons in front of moving backgrounds. For this we used a combination of a pseudo 2-dimensional hidden Markov model (P2DHMM) and a Kalman filter (see e. g. [1, 5]). This combination delivered good results, and so the question arose how this approach could further be improved with regard to robustness and the possibility of handling occlusion effects.

Because it seems that each method for locating a desired object has its specific advantages and disadvantages, one could try to combine the advantages of different measurement methods and at the same time to overcome their special disadvantages. This leads to the idea of so-called *multimodal tracking methods*, where several modes are exploited in order to increase the robustness of a tracking algorithm under real-world conditions.

When developing multimodal tracking algorithms, basically two problems have to be solved: Firstly, the choice and implementation of the different modes that are selected in order to support the tracking process with an appropriate measurement signal, and secondly, the successful combination of the different modes using a selected paradigm. The approaches that we investigated here are a combination of a P2DHMM with a skin finder or a motion detector for person tracking. The motivation for this choice was to sustain our proven P2DHMM system as one of the modes in our new multimodal system. As second mode a color based skin finder has been considered to be a good complementary information source, since skin and face information is not explicitly considered in the P2DHMM (which operates on gray level images) and especially since this second mode would be suitable to recover the tracking process in case of occlusions of the lower body. As a third mode a motion detector has been used, which robustly works on image sequences with a constant background. The tracking modes are merged in a probabilistic way using a particle filter. A particle filter is based on a sequential random-sampling framework and also known as Monte Carlo filter. This type of filter has been found to be especially interesting for multimodal fusion since it offers flexible methods for a stochastic combination of the conditional measurement probabilities which are generated by the different tracking modes. We used a variant of this filter which is known as condensation algorithm and which we will describe in the following section.

2. PRINCIPLES OF THE CONDENSATION ALGORITHM

The condensation algorithm shall be described only briefly here. A more detailed description can be found in [2]. The name is an artificial construction and stems from the conditional density propagation which is an important aspect of the algorithm. The purpose of this algorithm is to describe the temporal propagation of conditional densities, which can be decomposed into three temporal consecutive steps, namely a deterministic drift, a stochastic diffusion and a re-

From: ICIP 2003, IEEE

active effect of a measurement. This is also done e. g. by a Kalman filter, but the condensation algorithm has the advantage that it is simpler from a mathematical point of view and therefore allows an uncomplicated and flexible combination of several measurement modes, as will be shown later.

In the following text we denote the state of the modelled object at the discrete time k as $\mathbf{x}_k = \mathbf{x}(t_k)$ and its history as $\mathbf{X}_k = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k\}$. In an analogous manner a set of image features is gathered in a measurement or observation vector $\mathbf{z}_k$ with the history $\mathbf{Z}_k = \{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_k\}$. Using these symbols and Bayes' rule the tracking problem can be formulated in terms of conditional probabilities:

$$p(\mathbf{x}_k|\mathbf{Z}_k) \propto p(\mathbf{x}_k|\mathbf{Z}_{k-1}) \cdot p(\mathbf{z}_k|\mathbf{x}_k) \tag{1}$$

The condensation algorithm uses a set of samples of the state vector to approximate its conditional probability density function $p(\mathbf{x}_k|\mathbf{Z}_k)$. This sample set consists of N samples $\mathbf{s}_k^{(n)}$, each weighted with the probability $\pi_k^{(n)}$ which is obtained from the measurement $p(\mathbf{z}_k|\mathbf{x}_k = \mathbf{s}_k^{(n)})$. Now the conditional state density can be represented by the weighted sample set $(\mathbf{s}_k^{(n)}, \pi_k^{(n)}, n = 1 \ldots N)$.

For a description of how this sample set can be obtained recursively from the previous sample set and for further details see e. g. [2].

3. COMPUTATION OF THE CONDITIONAL PROBABILITIES

The conditional probabilities $\pi_k^{(n)}$ have to be acquired by a measurement within the current image of the tracking sequence. Our approach is currently able to utilize three methods for acquiring this measurement data, namely a P2DHMM, a skin finder, and a motion detector.

The problem is now to evaluate a measurement vector $\mathbf{z}_k$ which results from one of the measurement modes (delivering e. g. a bounding box) in such a way that we can compute the conditional probability of this measurement under the condition of a given sample, expressed as $p(\mathbf{z}_k|\mathbf{x}_k = \mathbf{s}_k^{(n)})$. The relation between $\mathbf{z}_k$ and $\mathbf{x}_k$ is expressed by the measurement equation $\mathbf{z}_k = \mathbf{H} \cdot \mathbf{x}_k + \mathbf{v}_k$, where $\mathbf{H}$ is the measurement matrix and $\mathbf{v}_k$ is the measurement noise. If $\mathbf{v}_k$ is white noise, it is a reasonable assumption that the variable $\mathbf{z}_k$ is a stochastic process that can be characterized by a Gaussian distribution where $\mathbf{H}\mathbf{x}_k$ can be considered as mean value of the process. In this case the above mentioned Gaussian distribution can be interpreted as the probability of the measurement vector $\mathbf{z}_k$ under the assumption that the sample $\mathbf{s}_k^{(n)}$ is the correct state vector, resulting in

$$p(\mathbf{z}_k|\mathbf{x}_k = \mathbf{s}_k^{(n)}) \propto \exp(-\frac{1}{2}(\mathbf{z}_k - \mathbf{H}\mathbf{x}_k)^T \mathbf{C}^{-1} (\mathbf{z}_k - \mathbf{H}\mathbf{x}_k)). \tag{2}$$

In this function $\mathbf{C}$ denotes the covariance matrix which has to be chosen appropriately. The resulting probabilistic values are subsequently normalized so they will sum up to 1.

The state vector $\mathbf{x}$ (and each sample vector $\mathbf{s}$) consists of the components $\mathbf{x} = [x_c, y_c, v_x, v_y, w, h]^T$, where x_c and y_c describe the center of a bounding box with the width w, the height h and the velocity components v_x and v_y.

The functionality of this approach can be confirmed easily by the following assumptions: If the current measurement vector $\mathbf{z}_k$ is almost identical to $\mathbf{H}\mathbf{x}_k$, then measurement and sample must be located very closely together (i. e. $\mathbf{z}_k$ confirms $\mathbf{x}_k$ very well) and thus (2) will yield a very high probability for this sample. It is therefore a suitable equation for the probabilistic interpretation of the output $\mathbf{z}_k$ of our various modes.

3.1. P2DHMM

The abbreviation P2DHMM stands for pseudo 2-dimensional hidden Markov model. We will describe this method only very briefly here; for further details see e. g. [4, 1, 3]. The model which we used consists of 20 states which are arranged in 4 superstates (modeling columns) with each of them containing 5 normal states. The model has been trained to several hundred images that each show just one person surrounded by some arbitrary complex background. After this training has been accomplished, an image containing a person can be presented to the P2DHMM, and by means of the Viterbi algorithm the most probable state sequence and assignment of states to image areas can be calculated. In this way one obtains a segmentation of the image into person and background blocks. From this segmentation a bounding box (the smallest rectangle with horizontal and vertical edges that contains all pixels classified as person) and its center can be extracted.

Furthermore, the velocity of this bounding box can be calculated as the difference of the position of the center of the bounding box in the current frame and its position in the previous frame. Because this value can be very volatile, we smooth it by calculating a weighted mean value of the current velocity (70 %) and the previous velocity (30 %). Thus the result of the measurement of the P2DHMM will be a measurement vector of the form

$$\mathbf{z}_{\mathrm{P2D}} = [x_\mathrm{c}, y_\mathrm{c}, v_x, v_y, w, h]^T, \tag{3}$$

and the appropriate measurement matrix is a unity matrix.

3.2. Skin finder

As a second method for acquiring measurement data we use a simple implementation of a skin finder. The intention here was not to optimize this skin finder, but to demonstrate how a second measurement can be integrated into our condensation based tracking approach. As will be shown later,

this measurement can have a strong positive influence on the tracking results, even if it is not always very accurate.

The skin finder is based on an approach using color histograms and conditional probabilities as it is described e. g. in [6]. The result of this measurement will be a two dimensional vector which describes the center of gravity of the skin colored pixels and has the form

$$\mathbf{z}_{\mathrm{skin}} = \left[x_{\mathrm{cog,skin}}, y_{\mathrm{cog,skin}}\right]^T. \tag{4}$$

Because this point is expected to indicate the position of the face of a person, it will be positioned higher than the center of the bounding box of the whole person by an amount which can be estimated to be approximately 30% of the height of the bounding box. Therefore, for the measurement matrix of the skin finder we use

$$\mathbf{H}_{\mathrm{skin}} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & -0.3 \end{bmatrix}. \tag{5}$$

3.3. Motion detector

As a third method for acquiring measurement data we use a motion detector. Again here the intention was to demonstrate how another measurement can be integrated into our condensation based tracking approach and thus improve the tracking results. The motion detector bases on a calculation of differences d between pixels $\mathbf{i}(x, y)$ in the current image and corresponding pixels in a reference image according to

$$d_k(x, y) = \|\mathbf{i}_k(x, y) - \mathbf{i}_{\mathrm{ref}}(x, y)\| \tag{6}$$

and a subsequent thresholding. For those pixels with a difference exceeding the threshold, a bounding box will be calculated, and its parameters (center, width, height) are combined in a motion measurement vector with the components

$$\mathbf{z}_{\mathrm{m}} = \left[x_{\mathrm{cobb,m}}, y_{\mathrm{cobb,m}}, w_{\mathrm{bb,m}}, h_{\mathrm{bb,m}}\right]^T. \tag{7}$$

4. COMBINING MULTIPLE MODES

A very interesting aspect of the condensation algorithm is the possibility to very flexibly integrate the data of several measurements. As mentioned in the introduction, such a combination can make it possible to overcome disadvantages of a single method and to combine the strong points of several methods.

The point where we merged our measurements into the condensation algorithm is the calculation of the weights $\pi_k^{(n)}$ for the sample vectors $\mathbf{s}_k^{(n)}$. Here one has many possibilities to combine the probabilities $p(\mathbf{z}_i|\mathbf{s}_k^{(n)})$ which are generated by the different measurement modes. The method that we used here is to calculate a weighted product of those probabilities according to the equation

$$\pi_k^{(n)} = \prod_i p(\mathbf{z}_i|\mathbf{s}_k^{(n)})^{w_i} \tag{8}$$

and a subsequent normalization to ensure that the sum will be 1. Here the weights w_i are chosen manually according to the reliability of each measurement mode.

These modified sample weighting probabilities will have a strong impact on the tracking result, which is now the result of a multimodal fusion of different information channels.

5. RESULTS

Some interesting results of our tracking algorithm are depicted in Fig. 1 and Fig. 2, where the bold white bounding box indicates the expectation value of the samples.

In Fig. 1 a typical outdoor surveillance scenario with a non moving background is depicted (data from PETS 2001). For this sequence we used a combination of a P2DHMM and a motion detector. In the upper row we can see a case where the system with the P2DHMM mode alone loses the track after a while (see the third frame in this row), whereas in the lower row it can be seen that after integration of the motion detector mode the system keeps the track. In the last frame in the upper row a detailed result of the motion detector with the detected motion area and its bounding box can be seen. Also here, the use of the motion detector as single measurement mode will fail because other moving objects (see the passing car in the second frame) are severely disturbing this measurement.

In Fig. 2 the tracking results on a difficult outdoor sequence with occlusion of the lower body of the person and intensive camera operations are shown. For this sequence we used a combination of a P2DHMM and a skin finder. When using only the P2DHMM the tracking system failed, because the P2DHMM was trained to fully visible persons, while in this sequence the lower part of the person is occluded when the person walks behind a car. After adding the skin finder however as an additional measurement mode (indicated by a cross near the head of the person) it can be seen that the system now is capable of tracking this sequence.

6. CONCLUSION

In this paper we presented an approach for a multimodal tracking system based on a particle filter, a P2DHMM, a skin finder, and a motion detector. The architecture of this system has been described and implemented, and some exemplary results have been shown. The major innovation of our approach is the computation of conditional probabilities from the measurement vectors and the probabilistic mode fusion based on these values. Tests have shown that the combination of several tracking modes is a suitable approach to increase the performance of a tracking system in critical scenarios where a single mode alone fails.

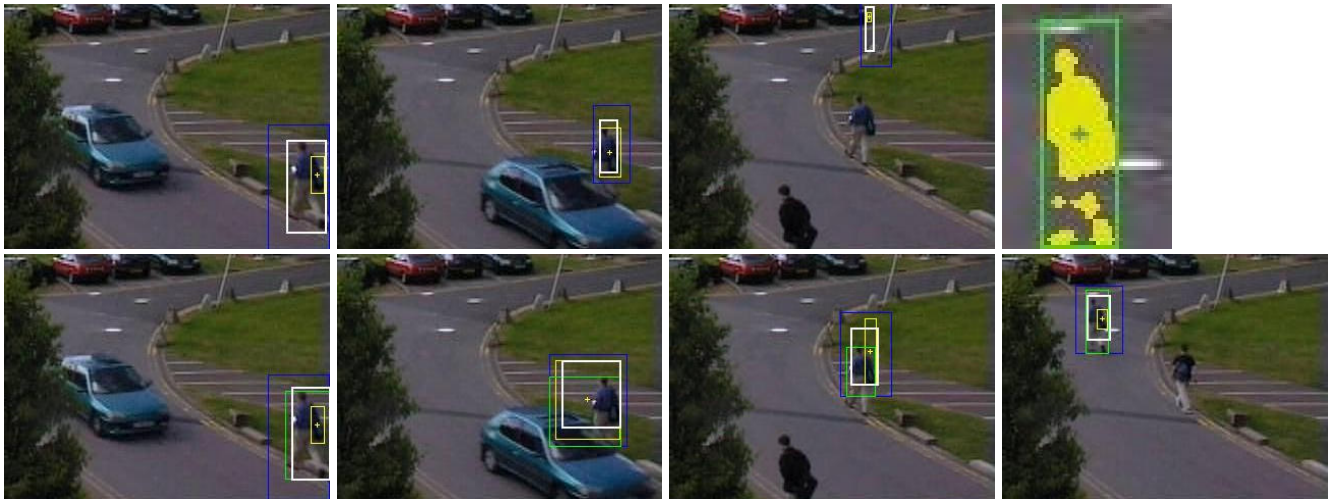

Fig. 1. Tracking results on a realistic outdoor surveillance sequence. Upper row: Only P2DHMM. Lower row: P2DHMM combined with the motion detector (indicated by an additional bounding box).

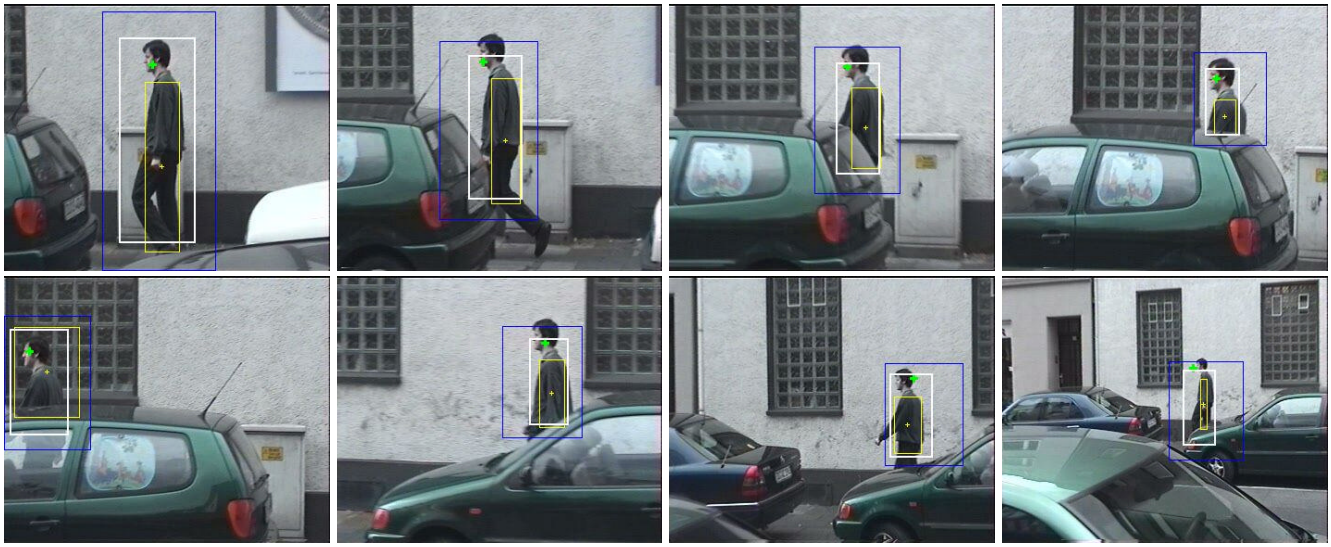

Fig. 2. Tracking results on a difficult outdoor sequence with occlusion of the lower body of the person and intensive camera operations.

References

[1] H. Breit and G. Rigoll. Improved Person Tracking Using a Combined Pseudo-2D-HMM and Kalman Filter Approach with Automatic Background State Adaptation. In *Proc. ICIP*, Thessaloniki, Greece, Oct. 2001.

[2] M. Isard and A. Blake. Condensation – Conditional Density Propagation for Visual Tracking. *International Journal of Computer Vision*, 29(1):5–28, Aug. 1998.

[3] L. R. Rabiner and B. H. Huang. An Introduction to Hidden Markov Models. *IEEE ASSP Magazine*, Jan. 1986.

[4] G. Rigoll, S. Eickeler, and S. Müller. Person Tracking in Real-World Scenarios Using Statistical Methods. In *IEEE Int. Conf. on Automatic Face and Gesture Recognition*, Grenoble, France, Mar. 2000.

[5] G. Rigoll, S. Eickeler, and I. K. Yalcin. Performance of the Duisburg Statistical Object Tracker on Test Data for PETS2000. In *IEEE Int. Workshop on Performance Evaluation of Tracking and Surveillance*, pages 1–7, Grenoble, France, Mar. 2000.

[6] K. Schwerdt. *Appearance-Based Video Compression*. PhD thesis, Inst. National Polytechnique de Grenoble, May 2001.