Available at
www.ElsevierComputerScience.com
POWERED BY SCIENCE DIRECT®

Image and Vision Computing 22 (2004) 571–582

www.elsevier.com/locate/imavis

Robust tracking of persons in real-world scenarios using a statistical computer vision approach

Gerhard Rigoll*, Harald Breit, Frank Wallhoff

Munich University of Technology, Institute for Human Machine Communication, Arcisstr. 16, 80333 München, Germany

Abstract

In the following work we present a novel approach to robust and flexible person tracking using an algorithm that combines two powerful stochastic modeling techniques: the first one is the technique of so-called Pseudo-2D Hidden Markov Models (P2DHMMs) used for capturing the shape of a person within an image frame, and the second technique is the well-known Kalman-filtering algorithm, that uses the output of the P2DHMM for tracking the person by estimation of a bounding box trajectory indicating the location of the person within the entire video sequence. Both algorithms are cooperating together in an optimal way, and with this cooperative feedback, the proposed approach even makes the tracking of persons possible in the presence of background motions, for instance caused by moving objects such as cars, or by camera operations as e.g. panning or zooming. We consider this as a major advantage compared to most other tracking algorithms that are mostly not capable of dealing with background motion. Furthermore, the person to be tracked is not required to wear special equipment (e.g. sensors) or special clothing. Additionally, we show how our approach can be effectively extended in order to include on-line background adaptation. Our results are confirmed by several tracking examples in real scenarios, shown at the end of the article and provided on the web server of our institute.

Keywords: Person tracking; Hidden Markov models; Kalman-filter; Statistical object modeling; Background adaptation

1. Introduction

Tracking of objects in arbitrary complex environments is one of the key problems of visual surveillance. A good overview describing various approaches for surveillance is given in Ref. [8].

However, if one looks closer at most of the methods presented in this survey, it becomes obvious that many approaches still have severe limitations. This becomes more apparent if one considers the following classification of tracking algorithms: the most simple algorithms are the ones that use additional sensors for tracking, as e.g. bulbs or special clothing. In this case, the problem of tracking mainly reduces to the problem of locating the sensor signal in each frame of the image sequence. Such an approach is often taken in gesture recognition applications for tracking body parts that would be difficult to locate without any additional equipment. This can be still a difficult and challenging problem, but the limitations of that approach are obvious. A second class of tracking algorithms is mainly based on the evaluation of motion information (see [21]), either derived from difference images or from computation of the optical flow [12]. There are several examples for the successful use of this approach (see, e.g. [5,11,13,24]). For instance, the system presented in Ref. [3] uses almost exclusively the difference image for motion detection and accumulates the motion patterns of people in one single representation, mainly with the purpose of action recognition rather than tracking. However, as soon as there is other motion in the image besides the moving object, such approaches have severe difficulties. It is worthwhile to note that—unfortunately—this situation is very frequent for real scenarios, e.g. the surveillance of traffic, corridors or gas stations. If there are only other moving objects in the image, one could still hope to segment the desired object from the different optical flow fields caused by all moving objects, although this is another difficult problem. If one simply relies on tracking the largest optical flow field in the image, one cannot be sure to track the right object. Also, many of these approaches require a known static background (see e.g. [20]).

* Corresponding author. Address: Institute for Human-Machine Communication, Munich University of Technology, 80290 Munich, Germany. Tel.: +49-89-289-28541; fax: +49-89-289-28535.
E-mail address: rigoll@ei.tum.de (G. Rigoll).

0262-8856/$ - see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/j.imavis.2003.09.014

In Ref. [2] the optical flow computation is additionally combined with the propagation of probability distributions as first suggested in the Condensation algorithm (see [14]). The examples presented in Ref. [2] are relatively simple and the use of optical flow is still expected to be troublesome in the presence of other motion in the image.

Things are getting more complicated if typical camera operations such as zooming or panning are carried out, leading to motion information distributed over the entire image. In this case, it becomes basically impossible to track the object by evaluating the motion information. Another class of tracking algorithms can be identified as algorithms without explicit shape models. These algorithms are somehow similar to the first class of algorithms, with the major difference that they do not search for sensor signals in each frame, but they look instead for features that indicate the presence of a specific cue or object. Especially in case of tracking persons, they are often based on the use of colors (as e.g. in Ref. [22]). In this case, colors are mostly used for performing either a segmentation process or a block matching process, where each block of the image is classified to either contain the object to be tracked or part of it, or to belong to the background. Such a process is often supported by additional shape features. The major problem of such an approach is the fact that the block-by-block matching process is based on features that are globally present in the investigated block. In this way, it is e.g. investigated if the current block contains a sufficiently high amount of skin color information or typical frequencies indicating the shape of a body. This leads to the fact that the matching process can be still erroneous, especially if confusing shape or color cues are present in the image and frequent confusions occur. A typical example showing the difficulty of such an approach is the attempt to locate people in static images using a block matching technique leading to reported recognition rates of around 70% [18].

Alternatively, it is possible to use explicit shape models (see e.g. [1,6]), but the construction of such models is a quite tedious task, and due to the large flexibility of human body movements, it is very difficult to establish a shape model that is able to cope with all these varieties. In this case, the typical limitations and inflexibility of deterministic rule-based systems become apparent.

In Ref. [10], a powerful tracking system is presented that is even capable of tracking multiple persons. It makes use of a relatively simple image-processing scheme, mainly relying on background subtraction, background update and fundamental morphological operations. Additionally, it makes use of the previously mentioned explicit shape models for tracking specific body parts. However, since it relies heavily on background subtraction, it apparently seems to be dependent on a fixed camera position with no panning or zooming operations to assure a background that basically is static and may change only slightly due to changing illumination and other environmental conditions. This assumption seems to be confirmed by the experiments presented in that article.

The approach suggested in this article for tracking of people is one of the first attempts to use a statistical shape model for tracking. The statistical model is represented by a so-called Pseudo-2D Hidden Markov Model (P2DHMM) (see [15]). Additionally, this P2DHMM is combined with a Kalman-filter for motion prediction. As will be shown later in more detail, such an approach has the following advantages:

- Similar to the explicit shape model approach, the statistical shape model is able to exploit some apriori-knowledge about human body shapes (e.g. the rough decomposition into head, body, legs). It therefore retains some of the advantages of this approach while being able to expand its flexibility and robustness.
- At the same time, the advantage of the model-free approach can be exploited, to automatically learn the features which are relevant for the problem. Thus, it combines the advantages of model-based and model-free approaches.
- The fact that the system does not rely on any motion information has several important advantages. One of them is the capability of tracking people independent of the fact that they are moving or not.
- Another consequence is the advantage that tracking is possible in presence of other moving objects in the background.
- The approach even works for panning or zooming operations generating motion information throughout the entire image sequence. This is demonstrated at the end of this article in Figs. 5 and 6. There, it is also explained that another advantage of the P2DHMM approach is the exploitation of the automatic scaling capabilities of HMM's in general, that become important in zooming operations due to the fact that the tracked object may change its size drastically.
- Through HMM multistream technique, it is easily possible to combine various features, such as e.g. color or shape features and to give those features a different weighting. This is accomplished in the following way: different types of features are extracted from the image frame, e.g. one feature type based on gray values describing the shape of the object to be analyzed and another feature type using color features, e.g. derived in the RGB space. Both feature extraction methods lead to vectors, which are not combined into one single larger vector, but kept separately, leading to two separate vector sequences for the image frame. During training, for each vector stream a separate Gaussian distribution is estimated as emission probability for each HMM state. During recognition, in each state the two different probabilities resulting from the shape and the color stream are added in the log domain with a weighting factor. The relation of the weighting factors for the shape

and the color stream indicates the emphasis given to the system either for shape or for color evaluation. Of course, additional feature types (e.g. texture-based) can be extracted and handled in the same way, thus allowing for a flexible emphasis of different feature types in the HMM framework.

- Using the previously mentioned capability, it is possible to either design the system for person-independent mode (e.g. by emphasizing more general shape features) or for person-specific mode (e.g. by emphasizing more the color of the clothes). In the latter case, one could track a specific person in presence of other moving persons, e.g. in a pedestrian zone or a shopping mall. Fig. 5 at the end of the article shows such a case.
- Through specific capabilities of P2DHMMs, the system is capable of exploiting local rather than global information. This enables the system for instance to track a person with a red T-shirt and blue jeans in presence of a person wearing a blue T-shirt and red jeans, or vice versa. This would not be possible for many other systems that are evaluating global color features, e.g. color histograms in each frame.

2. Basic tracking algorithm

As mentioned already in Section 1, the key feature of our algorithm is the fact that it makes use of two powerful stochastic modeling techniques, namely Pseudo-2D HMMs and Kalman-filters. In this case, the input of the Kalman-filter relies on the information provided by a complex shape model of the person's body of which the structure has been automatically learned and acquired by the P2DHMM. The dynamic information needed for tracking is solely generated by the Kalman-filter. While the Kalman-filter obtains its input information from the P2DHMM, the filter itself feeds its output information back to the P2DHMM and improves in this way the shape detection procedure of the P2DHMM. This optimal feedback between these two modules is another reason for the powerful performance of the approach. By letting only the Kalman-filter be responsible for the dynamic information of the tracking process, and relying in the measurement process completely on shape and (optionally) color information, the tracking procedure becomes entirely independent of other disturbing motions in the background. In this way, it is e.g. possible to track a person in front of a street with moving cars, because for each frame of the image sequence the Kalman-filter will only react on the detected position of the person's shape and not on the motion of the person, the motion of the cars, or the shape of the cars. This confirms the statements made at the end of Section 1 concerning the advantages of our approach. In the following sub-sections, we describe the basic functionality of both major components of our tracking system.

2.1. Measurement vector generation with pseudo-2D hidden Markov models

The P2DHMM generates a measurement vector that is used as input to the Kalman-filter. The components of this vector are the center of gravity of the person detected in the image and the width and height of the bounding box. The following steps are carried out for that purpose: first, the image is processed with a DCT-based feature extraction method that is adopted from a face recognition system [7]. The image is scanned with a sampling window from top to bottom and left to right. The pixels in the sampling window of the size 8×8 are transformed using the DCT according to the following equation

$$C(u,v) = \alpha(u)\alpha(v)\sum_{x=0}^{7}\sum_{y=0}^{7} f(x,y)\cos\left(\frac{(2x+1)u\pi}{16}\right) \times \cos\left(\frac{(2y+1)v\pi}{16}\right) \quad (1)$$

A triangle shaped mask extracts the first 10 coefficients ($u + v \leq 3$), which are arranged in a vector. An overlap between adjacent sampling windows improves the ability of the HMM to model the neighborhood relations between the windows. The result of the feature extraction is a two-dimensional array of vectors with the dimensionality 10. This array is presented to a P2DHMM as shown in Fig. 1.

Such a P2DHMM can be considered as a 2D stochastic model of an object in an image. It models the occurrence of a feature vector sequence which can be derived from that object if the object is pre-processed in the same manner as described above (see [15,16]). The parameters of the P2DHMM consist of the transition and output probabilities of the various HMM states and can be learned in order to model different objects. The learning of person shapes can be accomplished in the following way: Several hundred images of persons with appropriate pre-processing are presented to the P2DHMM for learning the structure of a human body by applying parameter estimation methods, such as the Forward–Backward algorithm, to the P2DHMM (see e.g. [19]). Since the P2DHMM can be considered as an elastic model (see also [23] for other approaches to deformable models), it is capable of modeling the human body in various positions.

This is illustrated in Fig. 1 by the hand-drawn sketch above the P2DHMM, showing a person within a complex background. The scenario of modeling hand-drawn sketches has been investigated extensively in Ref. [17]. A direct relation to an equivalent segmentation procedure of a real image can be easily derived by assuming an image pre-processed by an edge detector. In this case, the edges in the image would be very similar to the strokes in the sketch. It is therefore taken here as an example with a theory in the background that is well described in Ref. [17]. If such

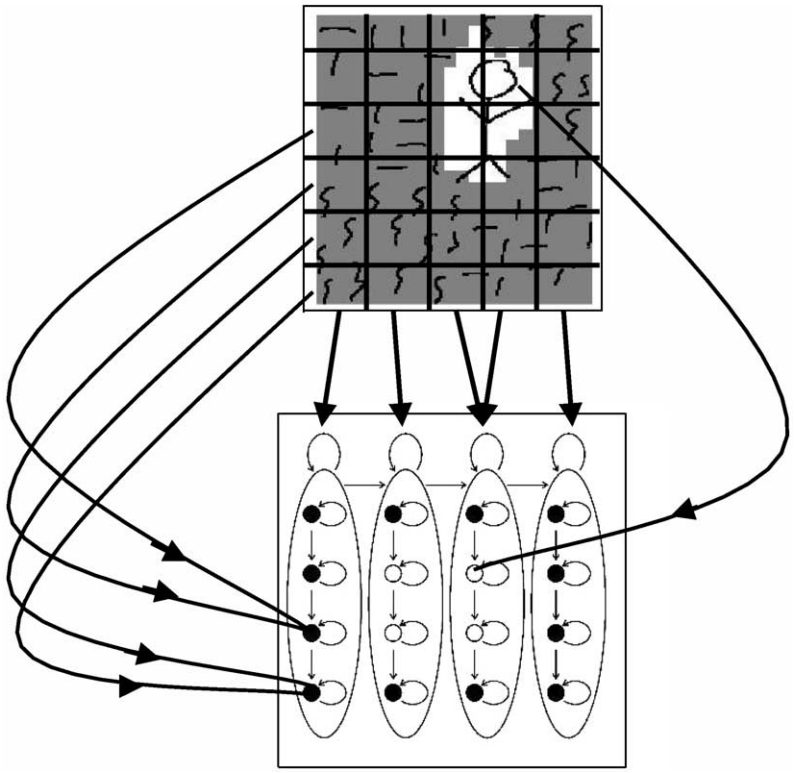

Fig. 1. Stochastic model of a two-dimensional object using a P2DHMM.

a sketch is modeled by a P2DHMM, each vertical stripe (column) of the above image will be assigned to one of the superstates of the P2DHMM, indicated by straight arrows directly leading to the vertical ellipses of the HMM in Fig. 1. Additionally, the blocks in each stripe will be aligned to the states within the superstates of the P2DHMM in vertical direction, as hinted by the curved arrows in Fig. 1. Thus, the P2DHMM has the basic capability of performing a non-linear 2D warping on the image.

In order to be capable of locating a body within a flexible environment with a complex background, the following important step is carried out: the above displayed P2DHMM is trained with static images that show a person within a complex environment, and not isolated or in front of a uniform background. In this way, while the HMM parameters of the system are learned successfully, several of the P2DHMM states will be assigned to the background, and other states will be assigned to the body regions of the person. In a subsequent step, the images are again presented to the learned P2DHMM and a Viterbi-alignment (see [19]) is carried out. In this way, it is possible to find out which states of the P2DHMM have been assigned to background features and which states typically represent the body parts. Then the states are marked accordingly as background states or person states, respectively. In Fig. 1, it is assumed that all states marked by a white circle represent body parts, and all states marked with dark circles are typical background states. This integrated segmentation procedure is illustrated by noting that all blocks within the white frame of the hand-drawn sketch in Fig. 1 have been aligned to the white states of the P2DHMM. It is important to note that this learning procedure is a preparatory action that has to be carried out only one single time and is not part of the actual tracking procedure. It is only required if the object to be tracked changes (e.g. from a person to a car) or the model has to be especially adapted to a specific person with special look and clothing. This stochastic model is capable of modeling persons within a complex background. As mentioned in Section 1, it combines the advantages of approaches with explicit shape models by exploiting the shape information learned from the body examples in the training images and of approaches without those shape models by maintaining the statistical learning and classification capabilities of these methods.

A popular paradigm used in tracking which can be considered to be most close to the P2DHMM approach is the use of blobs (see [22]), which makes use of similar statistical modeling techniques; however, blobs have a much simpler structure and seem to have less capabilities for the local modeling of features. For instance, a person with a red T-shirt and blue jeans might be difficult to be modeled with a blob, but can be effectively modeled by a P2DHMM by assigning probability density functions emphasizing red color features in the upper states of the P2DHMM and emphasizing the emission of blue color features in the lower states of the P2DHMM in Fig. 1.

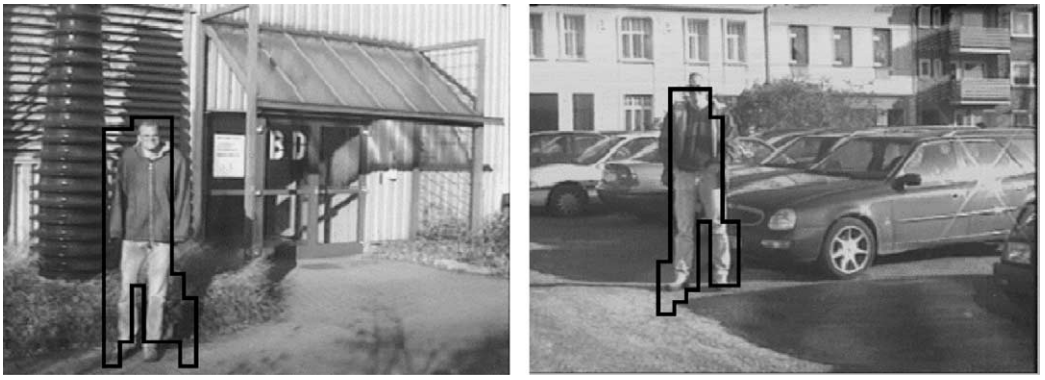

Fig. 2. Example for person segmentation in complex environments using the P2DHMM paradigm.

The actual tracking procedure starts with the presentation of the first frame of the tracking video sequence to the trained P2DHMM. If an image containing a person is presented to the above displayed especially trained P2DHMM, the Viterbi algorithm can be used again in order to compute the segmentation of the image into blocks assigned to the dark background states and blocks assigned to the white person states, thus obtaining the person's shape.

Fig. 2 shows examples for such a segmentation process on two static images. Because of the detailed and elastic matching capabilities of the P2DHMM, the algorithm is capable of locating the person even in complex environments, as shown in Fig. 2. Due to the special embedding process of the body states and the background states, very accurate person segmentation can be achieved in this way. The power of this approach can be explained by the fact that in this case the Viterbi algorithm performs an integrated segmentation and recognition process for the human body and thus follows the important principle that in many computer vision applications object segmentation is only possible if the object is recognized at the same time. This is especially true for the problem of locating people in complex backgrounds, where typical segmentation procedures based on histograms and color tables will fail.

It is interesting to analyze such a segmentation procedure in comparison to a discriminative segmentation that would be for instance provided by a neural classifier or a Support Vector Machine (SVM) trained on positive and negative examples for background and persons. The SVM will make use of a real recognition approach, by scanning the image frames for regions and classifying these regions as either 'foreground' or 'background'. The P2DHMM approach as presented in this article is mainly an alignment procedure rather than a classification procedure. Although trained according to the Maximum Likelihood (ML) principle, the P2DHMM obtains its discriminative power by forcing each image block to align either to background or foreground states rather than classifying the blocks into one of these categories.

It is further noteworthy that this approach also works very well if the background of the actual investigated image is different from the background of the training images. However, it is a logical consequence that this combined person recognition and segmentation procedure works especially well if the system operates in the same environment where also the training samples have been acquired and therefore the statistical properties of the non-stationary background are known to the system. This, however, may not be a very severe limitation, because such a condition comes naturally with most popular surveillance applications. For instance, if the system was used for surveillance of a corridor in a shopping mall, it could be trained on persons in that corridor where perhaps the background consists mainly of various shop-entrances or shop-windows with products. Even if the system would scan the entire corridor, resulting into a non-stationary background, the approach works well because the statistical properties of the background would be still stored in the various HMM states. Non-stationary background is therefore no limitation for the system. If the system would be reused, e.g. for surveillance of a traffic intersection, the background states could be easily adapted with the Forward–Backward algorithm to this new situation (for more details on the background adaptation see Section 3).

In the next step, the center of gravity (COG) of the person is computed from the segmentation result obtained from the Viterbi algorithm by simply calculating the appropriate moment from the blocks inside the black marked area indicating the person (as shown in Fig. 2). The coordinates of this COG, denoted as x_s and y_s, and the size of the bounding box of the segmentation, denoted as w and h, serve as the measurement input to the Kalman-filter.

2.2. Combination of P2DHMM output with Kalman-filter

In order to describe the moving person and to represent the result of the tracking procedure, a state vector $\mathbf{x}$ is

introduced, consisting of the following components:

$$\mathbf{x} = \begin{bmatrix} x_s : x\text{-coordinate of COG of person} \\ y_s : y\text{-coordinate of COG of person} \\ v_x : \text{horizontal velocity of COG of person} \\ v_y : \text{vertical velocity of COG of person} \\ w : \text{width of bounding box} \\ h : \text{height of bounding box} \end{bmatrix} \tag{2}$$

The motion of the person is described by a simple dynamic model, assuming that the person moves with a constant velocity between the sample points k and $k+1$. In this case, the dynamic behavior of the person can be expressed by the system equation

$$\mathbf{x}_{k+1} = \mathbf{A}\cdot\mathbf{x}_k + \mathbf{w_k} \tag{3}$$

with the system matrix

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{4}$$

and $\mathbf{w}_k$ as the process noise added to this process. It is assumed that only a part of the actual state variables can be directly measured from the actual input, resulting in the measurement equation

$$\mathbf{y}_k = \mathbf{H}\cdot\mathbf{x}_k + \mathbf{v}_K \tag{5}$$

with the measurement matrix $\mathbf{H}$ and the measurement noise $\mathbf{v}_k$ which is resulting from the measurement errors. The Kalman-filter computes the reconstruction of the state vector $\mathbf{x}$ from the measurement $\mathbf{y}$ according to the following equations (see e.g. [9,20]):

$$\hat{\boldsymbol{x}}_k^+ = \hat{\boldsymbol{x}}_k^- + \mathbf{K}(k)\cdot[\boldsymbol{y}_k - \mathbf{H}\hat{\boldsymbol{x}}_k^-] \tag{6}$$

$$\mathbf{K}(k) = \frac{\mathbf{P}^-(k)\mathbf{H}^{\mathrm{T}}}{\mathbf{H}\mathbf{P}^-(k)\mathbf{H}^{\mathrm{T}} + \mathbf{R}(k)} \tag{7}$$

$$\mathbf{P}^+(k) = [\mathbf{I} - \mathbf{K}(k)\cdot\mathbf{H}]\cdot\mathbf{P}^-(k) \tag{8}$$

$$\hat{\boldsymbol{x}}_{k+1}^- = \mathbf{A}\cdot\hat{\boldsymbol{x}}_k^+ \tag{9}$$

$$\mathbf{P}^-(k+1) = \mathbf{A}\cdot\mathbf{P}^+(k)\cdot\mathbf{A}^{\mathrm{T}} + \mathbf{Q}(k) \tag{10}$$

In this way, the gain matrix $\mathbf{K}(k)$ is updated for each discrete frame k according to the well-known equations of the Kalman-filter. $\mathbf{Q}$ and $\mathbf{R}$ are the covariance matrices of the stochastic processes $\mathbf{w}$ and $\mathbf{v}$, respectively. The measurement vector $\mathbf{y}$ is in this case:

$$\boldsymbol{y} = [x_s, y_s, w, h]^{\mathrm{T}} \tag{11}$$

resulting in the following choice for the measurement matrix $\mathbf{H}$:

$$\mathbf{H} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{12}$$

From the input information of the P2DHMM, contained in condensed form in the vector $\mathbf{y}$, the system estimates the state vector $\mathbf{x}$ and predicts in that way the information about the bounding box, contained in the last two dimensions of $\mathbf{x}$. The third and fourth dimension of $\mathbf{x}$ deliver the velocity of the person and mainly serve as variables supporting the mathematical model of the person's motion and the stability of the system.

2.3. Interaction between Kalman-filter and P2DHMM

An important point is the fact that—while the vector $\mathbf{x}$ is constructed from the vector $\mathbf{y}$ in the Kalman equations—the update of the vector $\mathbf{x}$ is used in return as input to the P2DHMM in order to improve the estimation of the vector $\mathbf{y}$, thus resulting into a cooperative feedback between the Kalman-filter and the P2DHMM and vice versa. This positive cooperation is realized in the following way: The bounding box estimated by the Kalman-filter is enlarged by a factor of 1.5, and the Viterbi search for the P2DHMM is concentrated on this specific region. This significantly improves the segmentation procedure provided by the Viterbi algorithm and results in a very good shape segmentation even if the person modifies his shape (e.g. by shrinking his arms) during the movement.

Such effective segmentation can only be achieved by the accurate shape modeling with the P2DHMM approach in combination with the reduction of the search region for the Viterbi algorithm provided by the bounding box estimation of the Kalman-filter. If a simpler shape model was used for this purpose, an accurate segmentation of the body (which will be mostly in presence of other moving objects or in cluttered environments) would be more difficult and the tracking algorithm would probably fail. In our approach, the previously mentioned integrated segmentation and recognition procedure guarantees a high quality segmentation result even under difficult conditions in real scenarios. In return, only this superior shape segmentation procedure enables the system to rely exclusively on static shape information of the person for providing the measurement signal to the Kalman-filter and thus makes the approach truly independent of any background motion. Therefore, the strength of the algorithm lies in the use of this special segmentation procedure plus the suitable feedback from the Kalman-filter output for optimizing the search space of the segmentation process.

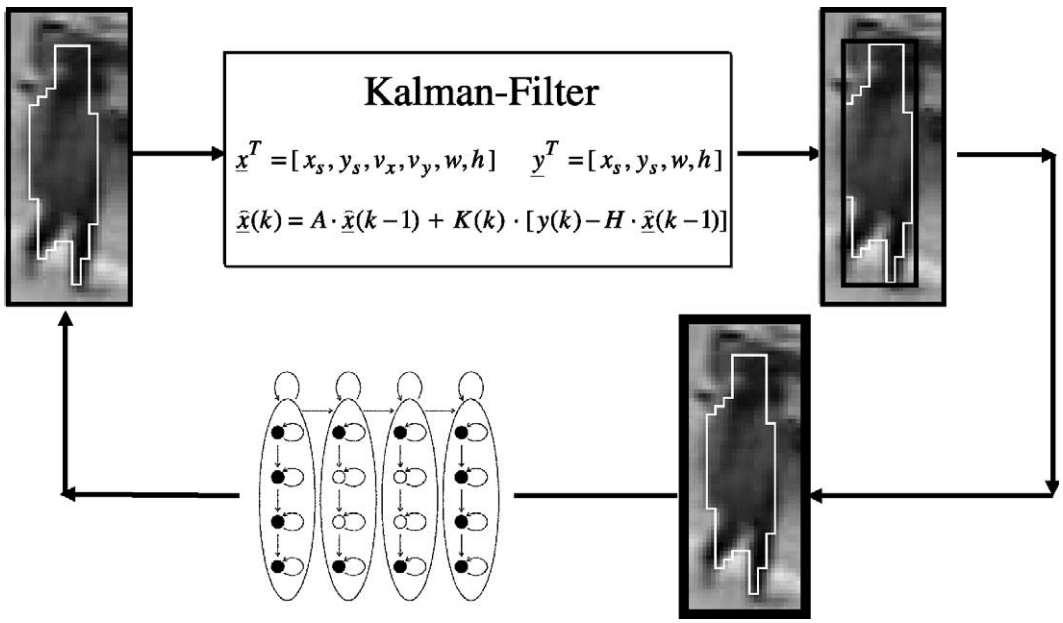

Fig. 3. Block diagram of the interaction between the P2DHMM and the Kalman-Filter.

In theory, it would be possible to aim for a more probabilistic combination between Kalman-filter and P2DHMM. One possibility for this would be to affect the measurement equation (5) by the production probability of the HMM, e.g. by increasing the value of the matrix **R** (indicating the measurement noise covariance) if the production probability is low (thus introducing less confidence in the current measurement). One problem with this idea is the fact that the production probability of the P2DHMM is mainly affected by the probability generated by the background according to its states which could be low if the current background does not match the model of the background states, although the segmentation could be very good due to a good person match for the foreground states. There are other probabilistic combination possibilities feasible and we plan to investigate this point in future activities.

Furthermore, the P2DHMM approach allows the elegant incorporation of additional features, such as e.g. colors or textures in the person segmentation procedure, by using multi-stream techniques. In this case, different features are derived from each frame of the image sequence (for instance DCT-based features, color features and texture features). Each different feature type leads to a different feature stream, if all frames of the sequence are processed. The states of the P2DHMM model the occurrence of each feature with a different probability density function, and the overall observation probability of the combined features in a certain state is computed as the product of the probabilities generated by each feature's density function. Weighting factors can be introduced in order to adjust the influence of the various feature streams. Consequently, the system can even be used to track a person in presence of other moving persons, if the person to be tracked has been acquired previously by the P2DHMM parameters which will automatically learn the shape cues from the person's body and color and texture cues from his clothes.

The complete interaction procedure between P2DHMM and Kalman-filter is illustrated in Fig. 3: On the left upper side, a moving person has been segmented, and the coordinates of the COG serve as measurement signal for the Kalman-filter which predicts a new state vector from this measurement input and the motion equation. On the right upper side, this leads to a new bounding box which can be derived from the updated state vector (inner black rectangle). This area is enlarged and thus yields an image fraction shown on the right lower side (black bold rectangle) which serves as search area for the P2DHMM. From there, the loop is closed by yielding a new segmentation which generates the new measurement signal in the upper left part of Fig. 3.

3. Background adaptation

As already mentioned at the end of Section 2.1, one advantage of our P2DHMM-based tracking system is that it can easily be adapted to different environments, even online during a tracking process.

Because we know which states of the P2DHMM are responsible for the background, it is feasible to adapt specifically those background states to changes in

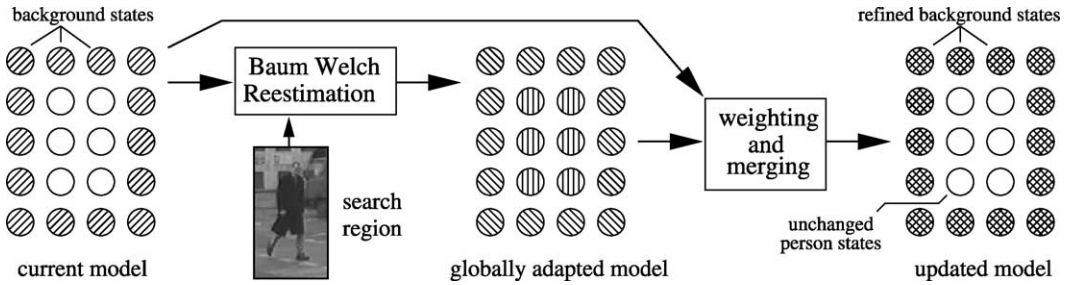

Fig. 4. Procedure of the adaptation of the P2DHMM to the background of the search region.

the background of the scenery [4]. This adaptation of the classifier can lead to an improved and more reliable person segmentation due to a better discrimination between person and background model. Therefore, we think that our system architecture is especially suitable for such an adaptive approach.

The adaptation procedure is illustrated in Fig. 4. The left part of this figure shows the states of the P2DHMM before the adaptation. To this model a Baum-Welch re-estimation algorithm is applied using the image section in the current bounding box as training data. This globally adapts the probability density functions of all states of the model to the current search region within a specific image frame. Since we apply only a few FB-iterations, starting with the parameters of the previous background information, this procedure corresponds to an adaptation rather than to a re-estimation and can also be carried out relatively quickly. The result is a globally adapted model to which subsequently a weighted merging process is applied that adopts the person states of the current model unchanged while updating the background states with a suitable weighting factor. The final result is an updated model with unchanged person states and adapted background states, as shown at the right part of Fig. 4. Background adaptation is thus possible because the spatially distributed structure of our model gives us the possibility to know exactly where the new background information is assigned to and which part of our model we have to update accordingly.

The person states are left unchanged to ensure that the original model is not changed too much which could impair the quality of the segmentation. This is reasonable because it is not to be expected that the person will change significantly during the tracking process and because the usual changes of the person can be taken into account by the elasticity of the model. With our adaptation procedure it is also possible to update the person states of the model with a small weighting factor, but it has to be considered that this could decrease the robustness of the tracking process because the adaptation is unsupervised. Section 4 will show some effects of our adaptation procedure on actual tracking examples.

4. Experimental results

Fig. 5 at the end of the article shows a tracking example for a moving person in a complex environment. It contains six frames selected from a complete tracking sequence consisting of approximately 200 frames. The complete sequence is available under http://www.mmk.e-technik.tu-muenchen.de/demo/tracking.html. The tracking result is visualized by the bounding box estimated from the Kalman-filter in each frame. In the first frame of this sequence, a person enters a corridor and passes through the door. In the second frame, another person is leaving from a door in that corridor. This second person passes the person to be tracked in the third and fourth frame. Note that between the first and the second frame as well as between the second and the third frame a fast zooming operation is performed. It can be seen that despite of the additional motion—caused by the moving door and the zooming operation—the person is tracked reliably, even when another person is passing through the scene. Special attention should also be given to the zooming operation with respect to another effect: Typically in a zooming operation, not only the object's location but also its size can change dramatically. It is well known that HMM's are especially suitable for scaling effects in the time or space domain, due to the self-transitions allowing the accumulation of an arbitrary number of feature blocks and thus processing objects of (almost) arbitrary size. In the image sequence, this effect of size variation and the way this is handled by the tracking algorithm can be seen very well.

Fig. 6 shows six frames of another complex scene, where a person passes a traffic light with moving cars in the background. This entire sequence is available under the same above web address. In the second frame, a van passes behind the tracked person. The most remarkable frames are in third and fourth position, where the person passes the pole of the traffic light (and therefore is partially occluded for a while) and a car is additionally passing through the scene. Also note that a panning operation is performed throughout the entire image sequence, generating additional motion in each frame. Also in this case, the tracking procedure has no

Fig. 5. Indoor tracking example with a passing person and zooming operations.

difficulties, neither with the disturbing motion nor with the occlusion by the pole.

In Fig. 7, a comparison of tracking results with background adaptation on and off is shown. In the upper row there are three frames from a video sequence as in Fig. 6 with tracking marks inserted by a tracking process with background adaptation switched off. These frames are especially interesting because a red van is passing behind the person who is to be tracked, so the background around the person will change fast and strongly. The markings of the tracking process consist of three rectangles and a cross. The outer large rectangle surrounds the search region, the inner bold rectangle describes the Kalman estimation (i.e. the prediction of the Kalman-filter) of the position of the person, the inner thin rectangle surrounds the person area which is segmented by the P2DHMM (that is the area within the search region that is classified as belonging to the person; see also Fig. 8), and the cross describes the COG of the person area. It can be seen from those markings that in the lower sequence (where adaptation is switched on) the person is captured significantly more precisely. This is especially clearly visible in the second and third frame of each sequence, where the white Kalman box is much smaller and more focused to the person in the lower sequence, whereas it is distracted by the passing van in the upper sequence.

In Fig. 8, magnifications of the search regions from the image frames shown in Fig. 7 are displayed, with the image pixels that have been classified as part of a person marked in white. It can be seen that without background

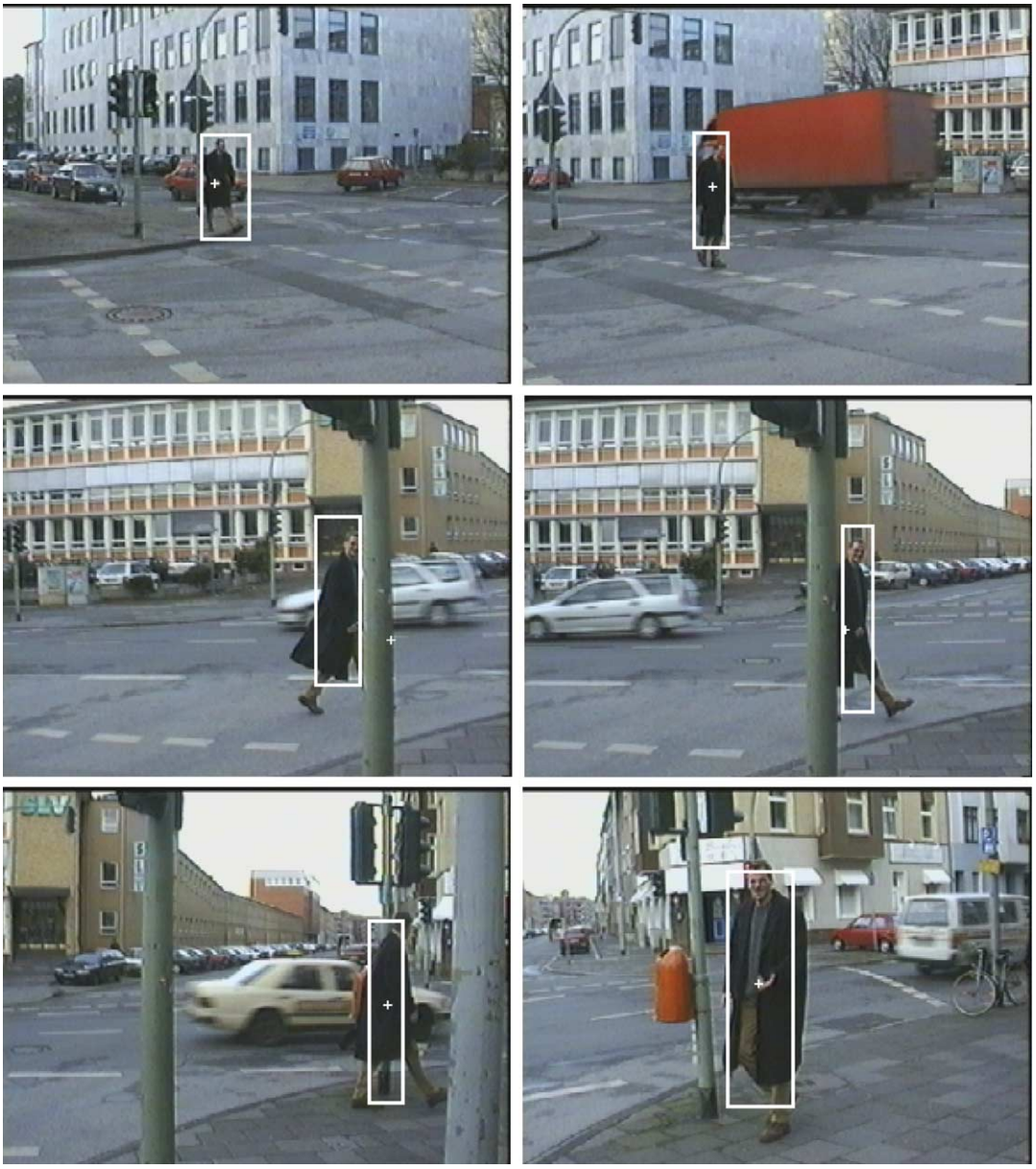

Fig. 6. Outdoor tracking example with passing cars, occlusion by a pole and panning operations.

adaptation the segmented person area is spread towards the passing van in the background (especially in the last frame), while this effect is strongly reduced when the background adaptation is active. The overall outcome is a much more precise tracking result if the background adaptation is switched on.

5. Summary and conclusion

In this article, a new approach to the tracking of people in arbitrary complex environments has been presented. The major novelty of this approach is the fact that through the use of P2DHMMs the advantages of model-based and of feature-based tracking approaches can be combined, leading to a very reliable module for person segmentation that can be combined effectively with a Kalman-filter, taking care of the dynamic aspects of the tracking procedure, and contributing to an effective limitation of the search region for the segmentation. The tracking examples presented in this article demonstrate that this approach indeed has the basic capability of tracking people in arbitrary environments, with arbitrary motion in the background that can be caused by other moving objects or zooming and panning operations. It has been demonstrated how our approach can be effectively extended to an adaptive procedure, with an on-line update of our background model, resulting in an improved

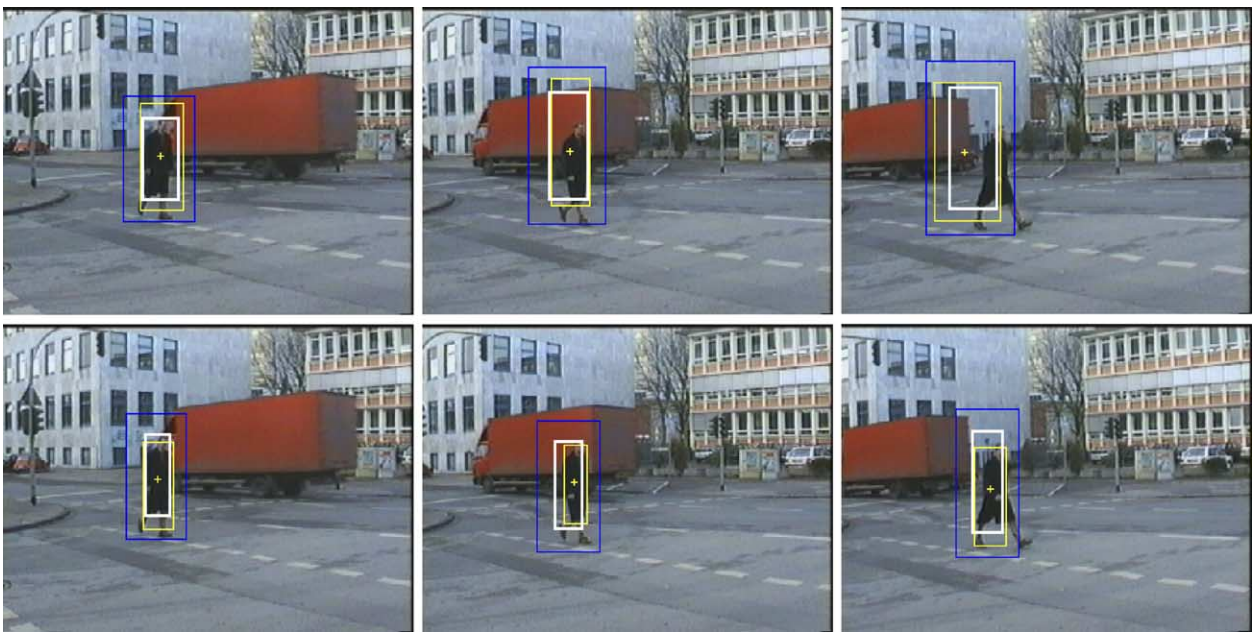

Fig. 7. Some exemplary results of our tracking algorithm with background adaptation off (upper row) and on (lower row).

(a) Background adaptation off (b) Background adaption on

Fig. 8. Visualization of the pixels within the search region that are classified as person (white area) with background adaption off (a) and on (b).

tracking performance. Further possible improvements include the investigation of alternative Kalman-filter structures and better interaction between the two stochastic modeling techniques P2DHMM and Kalman-filter. These improvements are currently under investigation. Other long term goals of this research include the previously mentioned aim of tracking specific persons in the presence of other moving people using highly sophisticated P2DHMMs with multistream techniques.

References

[1] A.M. Baumberg, D.C. Hogg, An Efficient Method for Contour Tracking Using Active Shape Models, Technical Report 94.11, School of Computer Studies, University of Leeds, April 1994.

[2] M.J. Black, D.J. Fleet, Probabilistic detection and tracking of motion discontinuities, Proceedings of IEEE International Conference on Computer Vision (1999) 551–558.

[3] A.F. Bobick, J.W. Davis, Real-time recognition of activity using temporal templates, in: Proceedings of the Workshop on Applications of Computer Vision, December 1996.

[4] H. Breit, G. Rigoll, Improved Person Tracking Using a Combined Pseudo-2D-HMM and Kalman Filter Approach with Automatic Background State Adaptation, in: Proceedings of the IEEE International Conference on Image Processing (ICIP), Thessaloniki, Greece, October 2001.

[5] Q. Cai, J. Aggarwal, Tracking Human Motion Using Multiple Cameras, in: Proceedings of the ICPR, Vienna, vol. 3, 1996, pp. 68–72.

[6] T.F. Cootes, C.J. Taylor, D.H. Cooper, J. Graham, Active shape models—their training and application, Computer Vision and Image Understanding: CVIU 61 (1) (1995) 38–59.

[7] S. Eickeler, S. Müller, G. Rigoll, Recognition of JPEG compressed face images based on statistical methods, Image and Vision Computing Journal, Special Issue on Facial Image Analysis 18 (4) (2000) 279–287.

[8] D. Gavrila, The visual analysis of human movement: a survey, Computer Vision and Image Understanding 73 (1) (1999) 82–98.

[9] M.S. Grewal, A.P. Andrews, Kalman Filtering Theory and Practice, Prentice-Hall, Englewood, NJ, 1993.

[10] I. Haritaoglu, D. Harwood, L.S. Davis, W4: who? when? where? what? a real time system for detecting and tracking people, in: Proceedings of IEEE International Conference on Face and Gesture, April 1998.
[11] B. Heisele, C. Wöhler, Motion-Based Recognition of Pedestrians, in: Proceedings of International Conference on Pattern Recognition, 1998, pp. 1325–1330.
[12] B.K.P. Horn, B.G. Schunck, Determining optical flow, Artificial Intelligence 17 (1981) 185–203.
[13] A. Iketani, A. Nagai, Y. Kuno, Y. Shirai, Detecting Persons on Changing Background, in: Proceedings of the ICPR, Brisbane, vol. 1, 1998, pp. 74–76.
[14] M. Isard, A. Blake, Condensation—conditional density propagation for visual tracking, International Journal of Computer Vision 29 (1) (1998) 5–28.
[15] S. Kuo, O.E. Agazzi, Keyword spotting in poorly printed documents using pseudo 2-D hidden Markov models, IEEE Transactions on Pattern Analysis and Machine Intelligence (1994) 842–848.
[16] S. Marchand-Maillet, 1D and Pseudo-2D Hidden Markov Models for Image Analysis—Applications and Results, Technical Report MMWP-99xx, Department of Multimedia Communications. EURECOM Institute, Sophia, Antipolis, 1999.
[17] S. Müller, S. Eickeler, C. Neukirchen, B. Winterstein, Segmentation and Classification Of Hand-Drawn Pictograms In Cluttered Scenes—An Integrated Approach, in: Proceedings of IEEE-ICASSP, Phoenix, 1999, pp. 3489–3492.
[18] M. Oren, C. Papageorgiou, P. Sinha, E. Osuna, T. Poggio, Pedestrian Detection Using Wavelet Templates, in: Proceedings of Computer Vision and Pattern Recognition, 1997, pp. 193–199.
[19] L.R. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition, Proceedings of the IEEE 77 (2) (1989) 257–285.
[20] J. Segen, S. Pingali, Camera-Based System for Tracking People in Real Time, in: Proceedings of the ICPR, Vienna, vol. 3, 1996, pp. 63–67.
[21] M. Shah, R. Jain, Motion-Based Recognition, Computational Imaging and Vision, Kluwer Academic Publishers, Dordrecht, 1997.
[22] C. Wren, A. Azarbayejani, T. Darrell, A. Pentland, Pfinder: real-time tracking of the human body, IEEE Transactions on Pattern Analysis and Machine Intelligence 19 (1997) 780–785.
[23] L.-Q. Xu, D.C. Hogg, Using Neural Network to Learn Spatial–Temporal Models for Moving Deformable Objects Training, in: Proceedings of the International Workshop on Neural Networks for Identification, Control, Robotics and Signal/Image Processing, 1996, pp. 145–153.
[24] T. Yamane, Y. Shirai, J. Miura, Person tracking by integrating optical flow and uniform brightness regions, in: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA-98), Piscataway, May 16–20 1998, IEEE Computer Society, 1998.