Proceedings of the 5th WSEAS Int. Conf. on SIGNAL, SPEECH and IMAGE PROCESSING, Corfu, Greece, August 17-19, 2005 (pp236-241)

Simultaneous Localization and Tracking of Persons in a Cluttered Scene with a Single Camera

HOLGER FILLBRANDT, KARL-FRIEDRICH KRAISS
Chair of Technical Computer Science
RWTH Aachen University
Ahornstr. 55, 52074 Aachen
GERMANY

Abstract: Visual surveillance systems often require exact localization of tracked persons on the ground plane of the monitored scene. This becomes a challenging task in a cluttered environment, where persons not only overlap each other in the image plane but are also at times partially occluded by furniture or similar objects. In contrast to many multi-camera or stereo approaches to overcome these difficulties, we present a framework to solve the surveillance task using only one stationary monocular color camera. Our approach is to reconstruct the full silhouette from the visible parts, to merge several position assumptions and to incorporate relevant a priori knowledge about the monitored scene into the system. The paper will discuss this approach and present special algorithms developed for this task.

Key-Words: Video Surveillance, People Tracking, Localization, Segmentation, Camera Model, Silhouette Alignment

1 Introduction

In recent years, the interest in automatic visual surveillance of sensitive areas is continuously growing. The general goals of such vision systems are to detect and track people in the monitored scene, to determine their exact positions on the ground and often also to evaluate their behaviour and generate an alarm if something suspicious happens. These image processing tasks become very challenging if the observed room is narrow with many people inside and additionally cluttered. The problems increase if only one monocular camera is to be used.

Our work is part of the development of a visual surveillance system for civil aircraft cabins. The system is supposed to determine the exact position of each passenger on the floor map of the cabin at any time using one stationary monocular camera per cabin section. In this environment, the tracked passengers do not only occlude each other in the image plane, but are also often partially occluded by stationary objects in the scene like seating rows or overhead compartments. Therefore the floor position calculation cannot solely be inferred from the head or feet coordinates of a tracked person. In this paper, we will address these difficulties and propose a framework to combine silhouette reconstruction, various position assumptions and a priori knowledge for robust person tracking and localization under said circumstances.

One group of common approaches towards people tracking work entirely in the image plane, i.e. they keep track of the 2D image areas and image coordinates of the persons in the scene but do not attempt to extract the real world positions and trajectories [3][5][7]. Occlusions between persons are often handled by exploiting the merging and splitting of tracked foreground regions. Elgammal and Davis [2] propose a probabilistic framework to segment people separately during occlusion using color distributions of the clothes in combination with assumptions about the depth positions.

A second approach to tracking systems employs camera models to perform transformations between image and real world coordinates [8]. Besides the possibility to determine the ground position of each person, this approach has the advantage that more real world knowledge can be incorporated into the system, e.g., the maximum walking speed or the correlation between position and size of the tracked human shape. Most of these tracking systems deal with the occlusion problem by using multiple cameras with overlapping fields of view [1][4][6].

The majority of publications about people tracking adress the problem of inter-person overlapment in various ways but do not deal with the equally demanding problem of partial or full occlusion by objects in the scene. The latter becomes important as soon as exact ground localization is required in a complex situation like public transport or office buildings.

In the following, we will first give an overview of our approach to this task. Thereafter we will present the tracking system in detail.

2 Overview

A human observer solves the task of visual tracking and locating people in a video sequence of a complex environment by three-dimensional understanding of the

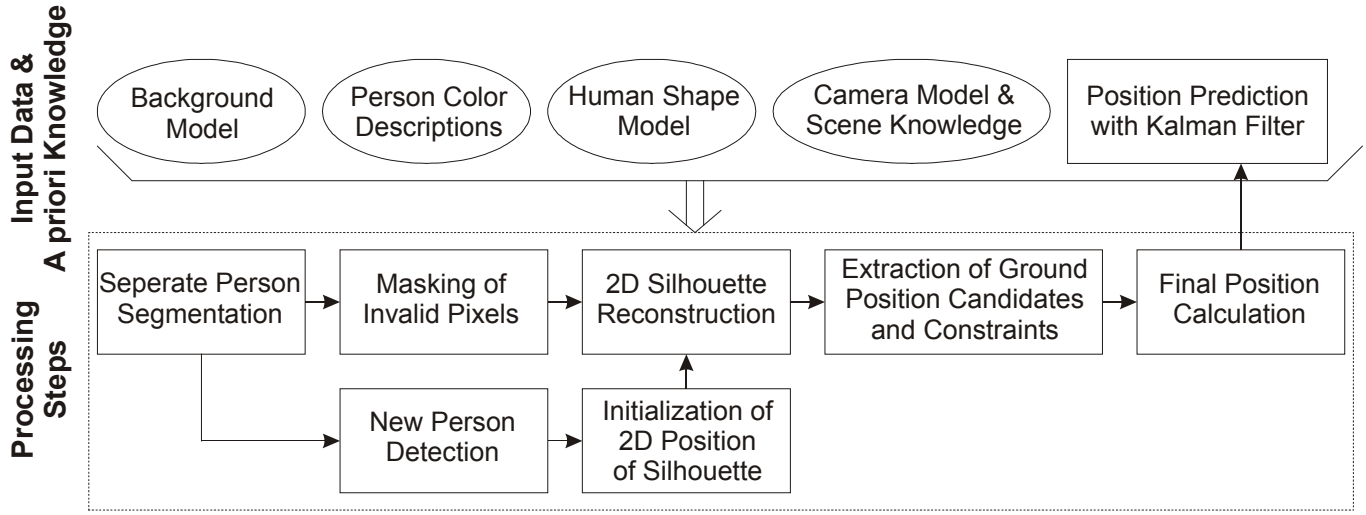

Fig. 1: Overview of the tracking system

scene in its entirety. Not only the persons in the scene are recognized and localized but also every single object as well as the whole structure of the room. This is combined with vast knowledge about human behaviour, body structure and physical laws. Needless to say that present computer vision is far from that level of scene understanding, therefore, what is the relevant a priori knowledge and image information used to solve this task?

a b c

Fig. 2: Examples of different situations

Fig. 2 shows three examples which make obvious, that robust ground position extraction has to rely on multiple observations. In Fig. 2a, the position of the feet on the floor plane is the most reliable information, while in Fig. 2b it is the head position in combination with an assumption or knowledge of the person's body height. In the case of occluded upper and lower end of a person's silhouette (Fig. 2c), the human observer does basically two important things to determine the person's position in the scene: First, he is always aware of the complete silhouette of the person, even if parts of it are not visible. The full silhouette is reconstructed in the mind's eye, resulting in a hypothesis about where the feet of the person are on the ground. Secondly, he looks at the lower end of the visible body part to find out, which seats occlude the person's shape and also, which are occluded by the person. In this specific situation, the latter information is needed to determine the row the person stands in.

Fig. 1 presents an overview of the tracking system. Basically, it attempts to realize all the components mentioned above and combines them into a framework that automatically takes advantage of the most reliable information available.

The first processing step for a new frame in an image sequence is to classify each pixel either as background, as belonging to a specific person, or as unrecognized foreground. Besides a background model, a color description of each tracked person is used as well as a prediction of the current ground positions using a Kalman Filter.

In the next step, the predicted depth positions of all tracked persons in combination with the known location of scene elements are used to define, separately for each person, by which image areas the silhouette is occluded. In these areas, the segmented foreground does not contain any valid information, whether a pixel is inside the projected shape of the person or not.

This occlusion mask together with a coarse model of the human shape are then employed to reconstruct the full silhouette from the visible part. For a robust detection of the head and the feet, this step is performed even if the person is completely visible, because the segmentation can have holes or additional areas due to similar background colors as the person's clothes, not fully removed shadows or raised arms of the person.

At this point, there are three assumptions about a person's position: 1. the prediction, 2. the position calculated from the feet coordinates in the image, and 3. the position calculated from the head in the image. These assumptions are described as weighted probability distributions and are constrained by various observations and a priori knowledge. Finally, the most likely position is calculated.

3 A priori Knowledge

3.1 Camera Model

The camera model transforms 3D world coordinates (x, z, h) into the 2D image plane (x_i, y_i) and vice versa, with z being the depth position in the scene. Since one dimension is missing in the camera image, additional information is needed for the reverse transformation, e.g. the height h above the ground. In this application, h is either equal 0, if the feet coordinates in the image are used, or equal to the height of the person H_P, if the head coordinates are used.

Using the camera model, the ground position of a person can be derived from the feet coordinates in the image or from the head coordinates, if the person's height is known. The height is calculated whenever the head and the feet position in the image are detected simultaneously and averaged over time.

The greater the distance of a person from the camera and the flatter the angle of depression towards the considered image coordinate, the less accurate is the calculated ground position. Since the maximum accuracy of the image coordinates is one pixel, the theoretical errors e_x and e_z of the resulting ground position can be estimated as

$$e_x = |x(x_i - 0.5, y_i, h) - x(x_i + 0.5, y_i, h)|, \quad (1)$$

$$e_z = |z(x_i, y_i - 0.5, h) - z(x_i, y_i - 0.5, h)|. \quad (2)$$

These error estimations will be used to evaluate the reliability of the ground positions calculated from the head and the feet positions of a person.

3.2 Scene Knowledge

As stated above, a priori knowledge about the structure of the monitored area becomes necessary in cluttered environments. One important information is provided by a *binary map of the floor* (Fig. 3b). It defines the valid ground positions where a person can walk and stand in the scene. The requirement that the connecting line between two subsequent positions of one person lies entirely inside the valid floor area prevents the system from tracking a person through objects by mistake. Additional information can be included in the floor map, e.g., the possible enter/exit points, where a person can enter or leave the field of view.

The second a priori knowledge used is a *depth template of the camera background image*. This is a matrix of the same size as the camera image which contains the z-position of each pixel in the empty back-ground scene. (Fig. 3c). The depth template serves two purposes:

1. The full silhouette of a person can be reconstructed from the visible part by predicting which image parts are located in front of that person.
2. From image areas that occlude or are occluded by a person, constraints for the detected depth position can be derived.

There are multiple possibilities of how to create the depth template. The optimum would be to derive the template from a three-dimensional CAD representation of the room. We use a 2½-D representation, that is a floor map with additional height data. This data is converted using the camera model.

Fig.3: a) Example Scene, b) Floor Map, c) Depth Template

3.3 Human Shape Model

Knowledge about the human shape is necessary to detect the head and feet coordinates of a person robustly and to reconstruct the full silhouette under partial occlusion. The current version of the system uses an average human silhouette for this purpose (Fig. 5a) that is compressed or stretched in x-direction according to the varying width of a person if viewed from the side or from the front. Since this description of the human shape has only four parameters (x- and y-position, scale and relative width), a fast implementation of the alignment to the segmented foreground is possible.

4 Silhouette Segmentation and Reconstruction

The goal of the first three processing steps is to get the full silhouette of each tracked person in the camera image, whether it is entirely visible or not.

4.1 Segmentation

The most common method to segment an image sequence into static background and moving foreground areas is adaptive background subtraction. The weakness of this principle is, that the silhouettes of overlapping people melt together into one foreground blob. We therefore developed a technique to use a color description of each tracked person and the background to separate people from each other during occlusion. For related approaches see [2] and [6]. The appearance model is based on local color histograms and is continously updated while a person is not occluded.

First, the color similarity between each pixel of the current frame and the background scene is calculated. A pixel is classified as unrecognized foreground if its

similarity value is below a certain threshold. Then, the predicted image areas of all tracked persons sorted from front to back are processed, classifying each pixel that has a higher similarity with the respective appearance model as belonging to this person.

A sufficiently large area of unrecognized foreground is initialized as a new person, if the initial silhouette alignment and position calculation provide plausible results.

4.2 Masking of Invalid Areas

In a new frame of the image sequence, the ground positions of all tracked persons are predicted using a second order Kalman Filter. The predicted depth position of each person is compared to each pixel of the depth template of the monitored scene. If a scene object is closer towards the camera than the person, the corresponding pixels will be marked invalid for this specific person in a binary occlusion mask (Fig. 4). In these areas, the foreground segmentation from the previous step provides no valid information, therefore it is important for the silhouette alignment in the next processing step to know these areas. In a similar way, occlusion by other people is handled using the output of the preceeding segmentation module and all predicted depth positions.

Fig. 4: Occlusion mask for the person marked by an ellipse in the left image

4.3 2D Silhouette Reconstruction

In this processing step, the human shape model is aligned to the visible part of a person's body, thus reconstructing the full silhouette. Input data are the segmented foreground and the occlusion mask of the current person. Since the model alignment has to be performed for every tracked person in the scene, the choice of a convenient algorithm is time critical. We propose a fast method that uses only Boolean variables and operations to adapt the model parameters in few iterations. This method therefore allows high speed implementation.

The silhouette adaptation is initialized with the predicted position and size of the person's shape in the image plane, calculated from the predicted floor position using the camera model. The initial relative width of the silhouette is set to the same value as has been found in the previous frame.

Let $O(x_i, y_i)$ be the Boolean array of occluded pixels at each image position (x_i, y_i), $F(x_i, y_i)$ the segmented foreground, and $S(x_s, y_s)$ the average human shape (Fig. 5a). A relevant difference $d_r(x_s, y_s)$ between each pixel (x_s, y_s) of this reference silhouette and the segmented foreground at the corresponding image position (x_i, y_i) is given by

$$d_r(x_s, y_s) = \text{NOT}\,(O(x_i, y_i))\ \text{AND}\ (S(x_s, y_s)\ \text{XOR}\ F(x_i, y_i))\,. \quad (3)$$

Pixel positions outside the image plane are treated as occluded pixels as well. Each relevant pixel difference is caused by the reference silhouette not perfectly fitting on the valid foreground and therefore requires parameter changes. To this end, AND-operations are performed between d_r and a set of precalculated templates Δ_{right}, Δ_{left}, Δ_{up} and Δ_{down} (Fig. 5b-e). These templates define, to what translation of the reference silhouette a pixel difference at a certain position relates. Because the silhouette border includes the most important information for exact alignment, a small area (usually 5 or 10 pixels) around suffices. For the scaling of the silhouette, similar templates can be derived by Boolean operations between the given templates and the average silhouette.

Fig. 5: a) Average Human Shape;
Displacement templates: b) left, c) right, d) up, e) down

For each of these templates, the number of arguments for the corresponding parameter change is counted and from these sums, the necessary translation and scaling are derived.

Furthermore, the templates serve a second purpose: The percentage of not-occluded pixels inside each area associated with a certain parameter change is defined as the reliability measure $r \in [0,1]$ for this parameter. If for example the reliability for adjusting the width is lower than a certain threshold, because e.g. the whole left side of a person is occluded, the width is not changed at all.

This alignment is repeated in a few iterations (usually about 5). In the last iteration, the reliabilities of the identified relevant coordinates of the full silhouette are calculated from the parameter reliabilities: The reliability r_X of the x-position of the full silhouette, of the y-coordinate of the upper silhouette edge (head) r_{YH} and those of the lower edge (feet) r_{YF}.

5 Ground Position Calculation

The calculation of the final ground position has to take into account the varying reliabilities of the several position assumptions as well as certain constraints that limit the possible position range. In the following, we will describe the framework used for this purpose.

Resulting from the steps described above, we finally have three assumptions about the ground position of the tracked person:

1. the position calculated from the feet coordinates in the image (x_F, z_F),
2. the position calculated from the head coordinates in the image using the average height of the person (x_H, z_H),
3. the predicted position from the Kalman Filter (x_P, z_P).

Each assumption is represented as a two-dimensional uncorrelated Gaussian probability distribution on the floor plane, e.g., for the position resulting from the feet coordinates:

$$P_F(x,z) = \frac{1}{2\pi\sigma_{xF}\sigma_{zF}} e^{-\frac{\sigma_{zF}^2(x-x_F)^2+\sigma_{xF}^2(z-z_F)^2}{2\sigma_{xF}^2\sigma_{zF}^2}} \tag{4}$$

The variances σ_{xF}^2, σ_{zF}^2, σ_{xH}^2 and σ_{zH}^2 are set equal to the squared theoretical position errors due to the coordinate transformation of the head or the feet coordinates (eq. 2).

$$\sigma_{xF}^2 = e_{xF}^2 \tag{5}$$

Therefore these variances are directly correlated with the camera perspective and position in the scene: The greater the distance of a person from the camera and the smaller the angle of depression towards the feet or the head position, the greater is the variance of the corresponding position calculation.

The variances σ_{xP}^2 and σ_{zP}^2 of the predicted position are extracted directly from the covariance matrix of the a priori error of the Kalman Filter.

The Mahalanobis distance between a ground position (x, z) and each Gaussian distribution is given by

$$d_{M_F}(x,z) = \frac{(x-x_F)^2}{\sigma_{xF}^2} + \frac{(z-z_F)^2}{\sigma_{zF}^2} \tag{6}$$

$$= d_{M_{Fx}}(x) + d_{M_{Fz}}(z) \tag{7}$$

The distance metric to all three distributions is defined as the weighted sum of the three Mahalanobis distances with the reliabilities of the corresponding image coordinates used as weights. Because of the uncorrelated nature of the x- and z-distributions and different reliabilities, we describe the distance metrics D_x and D_z of both axes separately:

$$D_x(x) = \frac{r_X d_{MFx}(x) + r_X d_{MHx}(x) + r_P d_{MPx}(x)}{2r_X + r_P} \tag{8}$$

$$D_z(z) = \frac{r_{YF} d_{MFz}(z) + r_{YH} d_{MHz}(z) + r_P d_{MPz}(z)}{r_{YF} + r_{YH} + r_P} \tag{9}$$

The reliability $r_P \in [0,1]$ of the prediction is a user defined parameter that defines the influence of the prediction on the final position. A good value has proven to be 0.3. At the most likely position $(\tilde{x}, \tilde{z})$, the distance to the three distributions is minimal. Together with eq. 6 we get:

$$\tilde{x} = \frac{\frac{r_X}{\sigma_{xF}^2}x_F + \frac{r_X}{\sigma_{xH}^2}x_H + \frac{r_P}{\sigma_{xP}^2}x_P}{\frac{r_X}{\sigma_{xF}^2} + \frac{r_X}{\sigma_{xH}^2} + \frac{r_P}{\sigma_{xP}^2}} \tag{10}$$

$$\tilde{z} = \frac{\frac{r_{YF}}{\sigma_{zF}^2}z_F + \frac{r_{YH}}{\sigma_{zH}^2}z_H + \frac{r_P}{\sigma_{zP}^2}z_P}{\frac{r_{YF}}{\sigma_{zF}^2} + \frac{r_{YH}}{\sigma_{zH}^2} + \frac{r_P}{\sigma_{zP}^2}} \tag{11}$$

Additionally, the final position is limited by a number of rules and observations and corrected if necessary:

1. It has to be located in the valid depth area for the tracked person. This depth interval is extracted from the shape of the reconstructed human silhouette by analyzing the areas the person occludes as well as those by which the person is occluded using the depth template of the scene.
2. The maximum possible speed and acceleration of a person can be defined and therefore constrains the radius of possible positions around the prediction.
3. The resulting position has to lie onto the valid floor as defined in the floor map.
4. The resulting position has to have a certain minimum distance from another person.

With the final position derived with this method, the measurement update of the Kalman Filter is performed.

6 Experimental Results and Discussion

In addition to numerous live experiments, test sequences were recorded from different camera perspectives in an office environment with stationary foreground objects like tables or monitors. A camera positioning tool was developed to get the necessary extrinsic camera parameters as well as the a priori scene knowledge. The recorded sequences represent typical challenging situations for people tracking systems:

1. One person walking around in the scene, being occasionally occluded by scene objects or leaving partially the camera field of view (Fig. 6a).
2. Two persons performing various types of occlusions (Fig. 6b,c):
   a) Passing each other in the image plane,
   b) Walking towards each other and stopping close-up,
   c) Walking around each other in small circles,
   d) Walking along the viewing axis of the camera.
3. Up to six people moving at random in the scene.

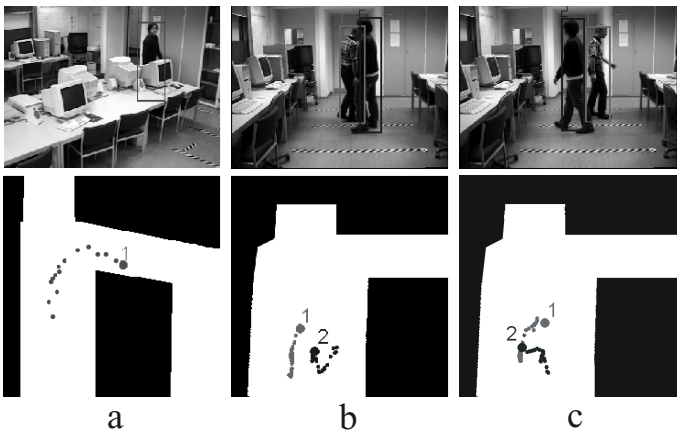

Fig. 6: Example Results: a) Occlusion by scene objects; b) Two people during and c) after occlusion

Most of the experiments of category 1 or 2 generate the desired results: The ground positions of the tracked persons are located robustly, even if varying parts of their bodys are not visible. The algorithm chooses dynamically, which image information is most reliable. This is, in addition to the ability to deal with occlusion by scene objects, the main advantage over previously published methods for ground plane tracking that rely either on the visibility of the heads or of the feets of the tracked persons to calculate their position [1][4][6][8].

The accuracy of the ground position depends on the camera perspective: The steeper the camera angle, the more accurate are the results. In our experiments, the observed accuracy lies within the human diameter, using a camera height of above 2 meters.

During the majority of interactions between persons, the system provides a reliable localization as a result of seperate segmentation and full silhouette reconstruction. However, person swapping occured in some experiments of category 2d and 3 with people wearing similar colored clothes. Here, tracking stability was increased by an additional framework to re-identify the persons after complex encounters and to correct their recorded trajectories.

The current implementation, which is not speed-optimized in any way, runs at 5-10 fps on a Pentium 4 processor, depending on how many people are tracked simultaneously.

7 Conclusion and Future Work

The presented framework to track and localize several persons in a cluttered scene using only one stationary camera represents a viable approach to the realization of cost-effective surveillance systems. The same methods can also be applied easily in a multi-camera setting, thus increasing the reliability of the detected ground positions.

In future work, various enhancements will be implemented and tested. A more detailed human shape model could improve the silhouette reconstruction and the detection of the feet and head coordinates. The tracking logic has to be enhanced to deal with special situations like one person in the front covering others for a long time span or two persons entering the scene together.

One drawback of the current system is its restriction to a stationary environment, so e.g. moved chairs will cause a problem. A possible solution could be an adaptive depth template of the scene that is automatically updated by analyzing the segmented shapes of moving persons.

Acknowledgements
This work was supported in part by the Federal Ministry of Economics and Labor of Germany under grant no. 20K0302Q. The authors are responsible for the contents of this publication.

References:

[1] J. P. Batista, “Tracking Pedestrians Under Occlusion Using Multiple Cameras”, *Int. Conference on Image Analysis and Recognition*, Springer LNCS 3212, 2004, pp. 552-562.

[2] A. M. Elgammal, L. S. Davis, “Probabilistic Framework for Segmenting People Under Occlusion”, *8th ICCV*, Vol. 2, 2001, pp. 145-152

[3] I. Haritaoglu, D. Harwood, L. S. Davis, “W4: Real-Time Surveillance of People and Their Activities”, *IEEE PAMI,* Vol. 22, No. 8, 2000, pp. 809-830

[4] J. Kang, I. Cohen, G. Medioni, “Tracking People in Crowded Scenes across Multiple Cameras”, *Proc. Asian Conf. on Computer Vision,* Korea, 2004

[5] S. J. McKenna, S. Jabri, Z. Duric, H. Wechsler, A. Rosenfeld, “Tracking Groups of People”, *CVIU,* Vol. 80, No. 1, 2000, pp. 42-56

[6] A. Mittal, L. S. Davis, “M2-Tracker: A Multi-View Approach to Segmenting and Tracking people in a Cluttered Scene”, *Int. Journal on Computer Vision*, Vol. 51, No. 3, 2003, pp. 189-203

[7] K. Sato, J. K. Aggarwal, “Recognizing and Tracking Two-Person Interactions in Outdoor Image Sequences”, *IEEE Workshop on Multi-Object Tracking*, 2001, pp. 87-94

[8] T. Zhao, R. Nevatia, “Tracking Multiple Humans in Complex Situations”, *IEEE PAMI*, Vol. 26, No. 9, 2004, pp. 1208-1221