Fixed Point Probability Field for Complex Occlusion Handling*

François Fleuret Richard Lengagne Pascal Fua
École Polytechnique Fédérale de Lausanne
CVLAB
CH – 1015 Lausanne, Switzerland
{francois.fleuret,richard.lengagne,pascal.fua}@epfl.ch

Abstract

In this paper, we show that in a multi-camera context, we can effectively handle occlusions in real-time at each frame independently, even when the only available data comes from the binary output of a simple blob detector, and the number of present individuals is a priori unknown.

We start from occupancy probability estimates in a top view and rely on a generative model to yield probability images to be compared with the actual input images. We then refine the estimates so that the probability images match the binary input images as well as possible.

We demonstrate the quality of our results on several sequences involving complex occlusions.

1. Introduction

In recent years, many people detection and tracking systems have been proposed, whether from monocular images or multiple cameras. They usually try to estimate the location of the people either in the image plane or on the ground plane from visual clues such as color, silhouettes, wide-baseline stereo and binary masks obtained from a preliminary image segmentation.

Most of them strongly rely on temporal information and use a Bayesian framework such as Hidden Markov Models (HMMs) to combine a motion model, that is a probability distribution of state transitions over time, and an appearance model, usually the conditional probability of the locations of individuals given the image data. However, imposing temporal continuity is a mixed blessing: When all goes well, it definitely increases robustness but, if errors start creeping in, the estimator may begin to diverge and eventually fail. Furthermore, in many real systems such as those where images are only acquired at much less than 25 frames per second, it simply does not apply.

In this paper, we show that in a multi-camera context, with a realistic and relatively small number of cameras such as depicted by figures 1 and 2, we can effectively handle occlusions at each time frame independently, even when the only data available comes from the output of a simple blob detector based on background subtraction and when the number of individuals is unknown a priori. This is of course not to say that in a complete system temporal continuity or more sophisticated texture measures should not be used. Instead, we would argue that our proposed algorithm should be combined with more traditional approaches to enforcing motion models to simultaneously perform tracking and detection, as was so convincingly demonstrated in [15].

Figure 1. Images from four surveillance cameras with superposed boxes representing detected individuals.

Given binary images such as those of figure 2 (c and d), our algorithm estimates the conditional probability of presence of people at a particular top view location as follows: We first define a generative model that, given a particular probability distribution in the top view, computes probabil-

*This work was supported in part by the Federal Office for Professional Education and Technology and conducted in collaboration with Visiowave, S.A.

(a) (b) (c) (d) (e)

Figure 2. (a), (b): 2 images of 2 people occluding each other; (c), (d): result of background subtraction used as input to our algorithm; (e): resulting top view probability field (the jagged lines stand for the limits of the camera fields of view). Note how well the two people are located in spite of the poor quality of the blobs. The corresponding locations are depicted by rectangles in (a) and (b).

ity images in which each pixel is assigned its likelihood of belonging to the union of all binary blobs. This generative model is an analytical average of pictures drawn by putting human-sized rectangles at occupied locations. Those rectangles are automatically computed from the approximate knowledge of a persons's height and the homography mapping the ground plane in the camera view and the top view. We then iteratively optimize the probability field so that the generated probability images match the binary input images until the field converges towards a fixed point. This estimation is performed in less than 150ms per frame on a standard 2.4Ghz PC.

This computation is the opposite of the feed-forward one performed by most state-of-the-art approaches [2, 6, 13, 14, 16] that rely on matching segmented regions and deriving a probability estimate in the top view from those matches. We start from the probability estimate in the top view and generate probability images to be compared with the actual input images. In spirit, this is close to the analysis by synthesis approaches to human motion modeling that have become popular because they allow the algorithms to take into account very subtle effects that would be difficult to handle any other way. Our method could also be related to Space-Carving approaches for multi-view 3D scene reconstruction such as [3, 1] where each voxel in 3D space is assigned its probability of belonging to the scene, given a set of images. By shifting the inversion process — that is, the estimation of the probability field as a function of the observations — onto an iterative algorithm, we avoid the requirement of an analytical expression of the inverse mapping and thus can handle a complex and stochastic dependency between the presences of individuals and the binary segmented images.

We have verified experimentally that the mathematically sound framework we propose has the following desirable emergent behavior:

- The algorithm can deal with potentially low quality binary input images, which are the norm rather than the exception in real-world applications, and with a very primitive human shape model.
- Complex occlusions between individuals are correctly handled. For instance, if the estimated probability of presence at a certain location is high, the system considers it has little or no information about any other location that is occluded by it.
- Perspective effects with regards to blob size and apparent feet and head location are correctly taken into account.

Furthermore, the probability of occupancy in the top view that our algorithm computes has a clear probabilistic semantic. It would therefore be natural to incorporate this algorithm into a complete HMM framework.

2. Related Work

Related work can be divided into monocular and multi-view approaches that we briefly review in this section. Some monocular approaches have proved efficient but, when the density of people increases, it becomes essential to merge information from several cameras in order to ensure a people detection and tracking strategy that are both accurate and robust against occlusions.

2.1. Monocular approaches

2.1.1 Blob-based methods

Some algorithms rely on binary blobs, such as [8], where a single-camera system combines shape analysis and tracking to locate people and maintain appearance models in order to track them even in presence of occlusions.

Approaches that perform tracking in a single view prior to computing correspondences across views also fall into this category. In [10], the limits of field of view of each camera are computed in every other camera from motion information. When a person gets visible in one camera, it is automatically found in other cameras where it is visible. In [4], a background/foreground segmentation is performed on calibrated images, followed by human shape extraction

from foreground objects and feature point selection extraction. Feature points are tracked in a single view and the system switches to another view when the current camera no longer has a good view of the person.

2.1.2 Color-based methods

In [11], the images are segmented pixel-wise into different classes, thus modeling people by continuously updated Gaussian mixtures. A standard tracking process is then performed using a Bayesian framework, which helps keep track of people under occlusion. When such a case occurs, models of persons in front keep being updated, but the update of occluded ones stops, which may cause trouble if their appearances have changed noticeably when they reemerge.

More recently [17], multiple humans are simultaneously detected and tracked in crowded scenes. Markov chain Monte-Carlo-based methods are used to compute the Maximum A Posteriori estimate of the number of objects in the scene, their positions and their correspondences in previous frames. In [15], multiple people are detected and tracked in complex backgrounds using mixture particle filters guided by people models learnt by an Adaboost algorithm. In [7], multi-cue 3D object tracking was addressed by combining particle-filter based Bayesian tracking and detection using learnt spatio-temporal shapes. This approach leads to impressive results but requires shape, texture and stereo information as input.

These last two approaches demonstrate that an efficient detection algorithm can dramatically improve tracking performances, especially when dealing with a varying number of objects. Moreover, combining detection and tracking can help robustly recover object location in complicated situations where tracking alone would tend to drift and fail.

2.2. Multi-view approaches

Despite the effectiveness of such methods, the use of multiple cameras soon becomes necessary when one wishes to accurately detect and track multiple people and compute their 3D locations in a complex environment. Occlusion handling may be facilitated by the use of 2 sets of stereo color cameras[12] but, in most approaches that only take a set of 2D views as input, occlusion is mainly handled using the temporal consistency brought by a motion model, whether from Kalman filtering or Markov models.

2.2.1 Blob-based methods

In [13], Kalman filtering is applied on 3D points obtained by fusing in a least-squares sense the image-to-world projections of points belonging to binary blobs. In [2], a Kalman filter is used to simultaneously track in 2D and 3D, and object locations are estimated through trajectory prediction during occlusion.

In [6], a best-hypothesis and a multiple-hypothesis approaches are compared to find people tracks from 3D locations obtained from foreground binary blobs extracted from multiple calibrated views. In [16], silhouette-based visual angles are obtained from motion blobs. In case of occlusion ambiguities, multiple occlusion hypotheses are generated given predicted object states and previous hypotheses. A Bayesian framework is applied to test multiple hypotheses using a state transition model, a dynamics model for transitions between occlusion structures and the measurements.

2.2.2 Color-based methods

[14] proposes a system that segments, detects and tracks multiple people in a scene using a wide-baseline setup of up to 16 synchronized cameras. Intensity information is directly used to perform single-view pixel classification and match similarly labeled regions across views to derive 3D people locations. Occlusion analysis is performed in two ways. Firstly, during pixel classification, the computation of prior probabilities takes occlusion into account. Secondly, evidence is gathered across cameras to compute a presence likelihood map on the ground plane which accounts for the visibility of each ground plane point in each view. Ground plane locations are then tracked over time using a Kalman filter.

In [9], individuals are tracked both in image planes and top view. The 2D and 3D positions of each individual are computed so as to maximize a joint probability defined as the product of a color-based appearance model and 2D and 3D motion models derived from a Kalman filter.

Those approaches propose complete tracking systems incorporating many visual cues as well as temporal consistency. Our contribution addresses a sub-part of such systems and proposes a way of handling occlusions at the pure detection level.

3. Finding a Fixed Point

The fundamental idea of our method is to estimate for every location of a virtual top-view the conditional probability of presence of an individual, given the detected blobs in every camera view. It appears that this is tantamount to finding probabilities such that corresponding synthetic images match the actual binary inputs.

In practice, we partition the top view with a regular grid as shown in figures 3(c) and 4. Our goal is to estimate the probabilities of occupancy $\rho_i = P(X_i = 1 \,|\, V)$, where X_i is the Boolean random variable standing for the presence of an individual at grid location i and V the set of input binary images. We show here that the ρ_i can be found as the fixed

V	the set of binary views $(V_1, \ldots, V_C)$ generated from the C video streams by the background subtraction.
X_i	the boolean random variable standing for the presence of an individual at location i.
ρ_i	the conditional marginal probability $P(X_i = 1 \mid V)$ that an individual is standing at location i, given the current observations.
$\overline{A}^c_{i,\xi}$	with $\xi \in \{0, 1\}$ the average synthetic image obtained when ρ_i is forced to ξ. It is a function of the other ρ_j, see figure 4.
λ_i	is $\log \frac{P(X_i=0)}{P(X_i=1)}$, the log-ratio of the prior probabilities.
Δ	is a pseudo-distance between images.
$g(.)$	is the log of a normal density.
$\mathcal{A}^c_i$	the image composed of 1s inside a rectangle standing for the silhouette of an individual at location i seen from camera c, and 0s elsewhere.

Table 1. Notations

point of an iterative process that repeatedly solves a large system of coupled equations.

3.1. Probability of Occupancy

As shown in [5], with the notations given in Table 1, we have

$$\rho_i = \frac{1}{1 + \exp\ \lambda_i + \sum_c g(\Delta(V_c, \overline{A}^c_{i,0})) - g(\Delta(V_c, \overline{A}^c_{i,1}))} \quad (1)$$

In particular, $\overline{A}^c_{i,\xi}$ is the average synthetic image from camera c where the value of ρ_i is forced to $\xi \in \{0, 1\}$. Average images such as the ones of figure 4, can be understood as the average of a large number of binary images generated randomly by putting human-sized rectangles at the various locations according to the ρ_i. Those rectangular shapes are automatically generated from the computation of the homography between the ground plane and the top-view and the approximate knowledge of a person's height. We do not actually generate those binary images, but compute directly the average image.

When we force one of the ρ_i to either 0 or 1, we make the corresponding rectangular shape in the average image either absent, as in figure 4 (c) or present with probability 1, as in figure 4 (d). The expression (1) therefore makes sense: If the fit with the rectangular shapes for position i present in the images improves, the score $g(\Delta(V_c, \overline{A}_{i,1}))$ grows, $g(\Delta(V_c, \overline{A}_{i,0}))$ becomes smaller and the sum in the exponential is negative, leading to a larger ρ_i. Moreover, occlusion is taken into account naturally: If a rectangular shape at position i is occluded by another one whose presence is very likely, the value of ρ_i does not influence the average image and only λ_i remains in the exponential. Thus the resulting ρ_i remains equal to the prior.

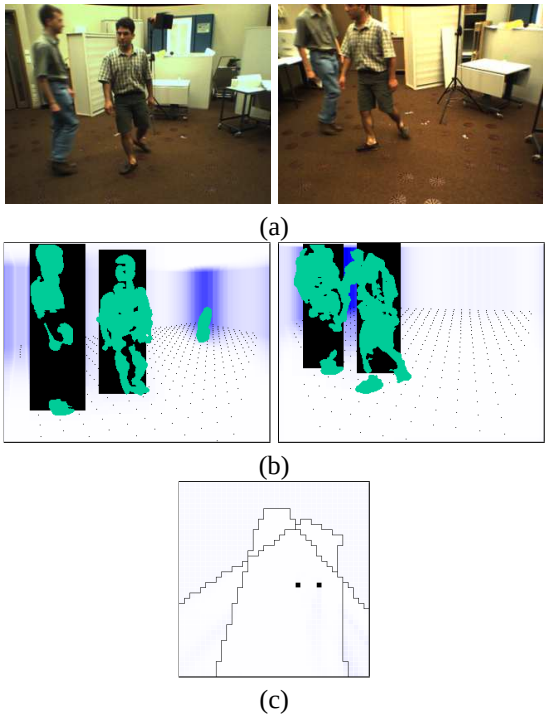


Figure 3. Original views from two cameras (a), average images after convergence and binary masks (b), the corresponding probabilities in the top-view (c). The shades of blue in figure (b) corresponds to pixel values in the average image (see §3.1). The jagged lines in (c) stand for the limits of the camera fields of view.

3.2. Algorithm

Note that, in equation 1, $\overline{A}^c_{i,\xi}$ depends on the $\rho_j, j \neq i$. We can therefore estimate the ρ_i as follows: We first give them a uniform value and use them to compute the average synthetic images $\overline{A}^c_{i,\xi}$. We then re-estimate every ρ_i with equation (1) and iterate the process until a stable solution is reached, which typically takes about 100 iterations.

More formally, let $\oplus$ denote the pixel-wise disjunction operator between binary images (the "union" image), $\otimes$ the pixel-wise product (the "intersection" image), $\sigma\,(I)$ the sum of the pixels of an image I and let 1 be the constant image whose pixels are all equal to 1. Given the expression of the average synthetic images [5], we perform the following set of operations at each iteration:

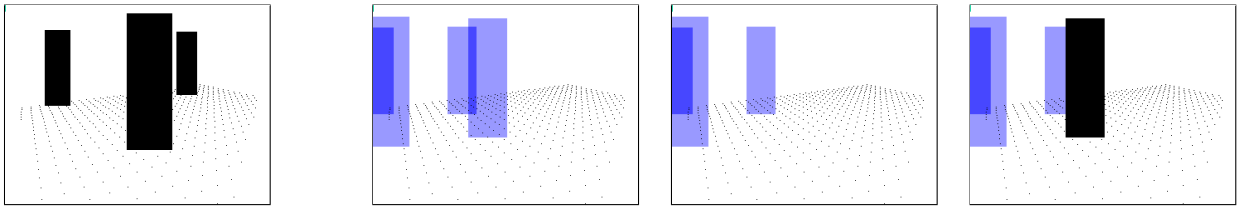
$$\overline{A}^c = 1 - \otimes_j \left(1 - \rho_j \mathcal{A}^c_j\right)$$

$$\sigma\left(\overline{A}^c_{i,\xi}\right) = \sigma\left(\overline{A}^c\right) + \tfrac{\xi - \rho_i}{1 - \rho_i}\, \sigma\left(\left(1 - \overline{A}^c\right) \otimes \mathcal{A}^c_i\right)$$

$$\sigma\left(V_c \otimes \overline{A}^c_{i,\xi}\right) = \sigma\left(V_c \otimes \overline{A}^c\right) + \tfrac{\xi - \rho_i}{1 - \rho_i}\, \sigma\left(V_c \otimes \left(1 - \overline{A}^c\right) \otimes \mathcal{A}^c_i\right)$$

$$\Delta(V_c, \overline{A}^c_{i,\xi}) = \tfrac{1}{\sigma(\overline{A}^c_{i,\xi})} \left(\sigma\left(V_c\right) - 2\,\sigma\left(V_c \otimes \overline{A}^c_{i,\xi}\right) + \sigma\left(\overline{A}^c_{i,\xi}\right)\right)$$

$$\rho_i \leftarrow \tfrac{1}{1 + \exp\left\{\lambda_i + \sum_c g(\Delta(V_c, \overline{A}^c_{i,0})) - g(\Delta(V_c, \overline{A}^c_{i,1}))\right\}}$$



(a) (b) (c) (d)

Figure 4. Picture (a) shows a synthetic picture corresponding to a deterministic presence of people at three locations. Picture (b) shows the average image $E(A|V)$ where all ρ_j are null but four of them equal to 0.2. Pictures (c) and (d) show $\overline{A}^c_{i,0}$ and $\overline{A}^c_{i,1}$ respectively, where i is the location corresponding to the black rectangle in (d).

where $\lambda_i = \log \frac{P(X_i=0)}{P(X_i=1)}$. Note that by using integral images, given any image I and any rectangular shape $\mathcal{A}$, the cost of the computation of $\sigma\left(I \otimes \mathcal{A}\right)$ does not depend on $\mathcal{A}$, apart from a pre-computation whose cost is proportional to the area of I. At each iteration and for every c, the first step cost is proportional to the sum of the areas of the $\mathcal{A}_i$. The 4 other steps have a cost proportional to the number of positions. Except for the exponential in the last step, which has to be repeated for every location, the computation only involves sums and products and is therefore fast.

Finally, one can also estimate the number of people actually present in the scene. If we denote by N this unknown quantity, we have trivially $\hat{E}(N \,|\, V) = \hat{E}\left(\sum_i X_i \,|\, V\right) = \sum_i \rho_i$.

4. Results

In this section, we show experimental results obtained on two sequences shot by 4 cameras in a 20 square meter room. The cameras are located at every corner of the room, two of them being mounted at a height of 2.2m, while the two others are at a height of 1.7m. In the first 4000 frame-one, 4 people enter the room successively. In the second 1000 frame-sequence, six people are moving about in the room.

Figures 5 and 6 depict a few frames of each sequence. In each case, we display the original images with the reprojection of each top view location where a person has been found standing, and the top view with the locations of detected people. The detection criterion consists of selecting the $\hat{N}$ locations with the highest probabilities, where $\hat{N} = \hat{E}(N|V)$ is the number of people the system automatically estimates (see §3.2).

In figure 6 we also display two isolated frames from the 6 people-sequence. The first one shows a successful detection for all 6 people in terms of localization, whereas the second one shows a good localization for 5 people and a failure for the last one. This second situation is actually very common in the 6 people-sequence and leads us to believe that incorporating temporal consistency upon our framework will improve the stability of the results and enable us to handle such a configuration.

To summarize, in our setting, using 4 cameras enables us to detect up to 4 people with very good accuracy, both in terms of the estimation of the number of people and the precision of people localization, as illustrated in the video. When dealing with a larger number of people, the averaged estimation of the number of people remains close to the true one, which may be of great interest when the target application focuses more on density estimation than on accurate people localization. Note that the upper limit of the number of people for a given camera setup is related to the area of the scene. The small size of our room leads to a high density of persons when more than four of them are present, and gives rise to frequent complete camera occlusions and unusable blobs such as the ones depicted by the last row of figure 6.

We can compute precise error rates by counting in each frame the number of actually present individuals, the number of detected individuals, and the number of false detections. We define a correct detection as one for which the reprojected boxes intersect the individual on at least three camera views out of four. Such a tolerance accommodates for cases where, due to optical deformations, while the estimated location is correct, the reprojection does not match the real location on one view. With such a definition, the estimated false negative error rate on a continous $2\text{min}30s$ video is 6.14% and the false-positive error rate is 3.99%. In other words, in absolute terms, our detection performance is excellent considering that we have used only a small proportion of the available image information. In effect, our formalism is generic enough to incorporate other image clues that would definitely improve the performances.

We checked the influence of the size of the rectangular shapes we use as models: The results are almost unchanged for model sizes between 1.7 m and 2.2 m. The performance tends to decrease for sizes noticeably smaller. This can be explained easily: if the model is shorter than the person, the algorithm will be more sensitive to spurious binary blobs

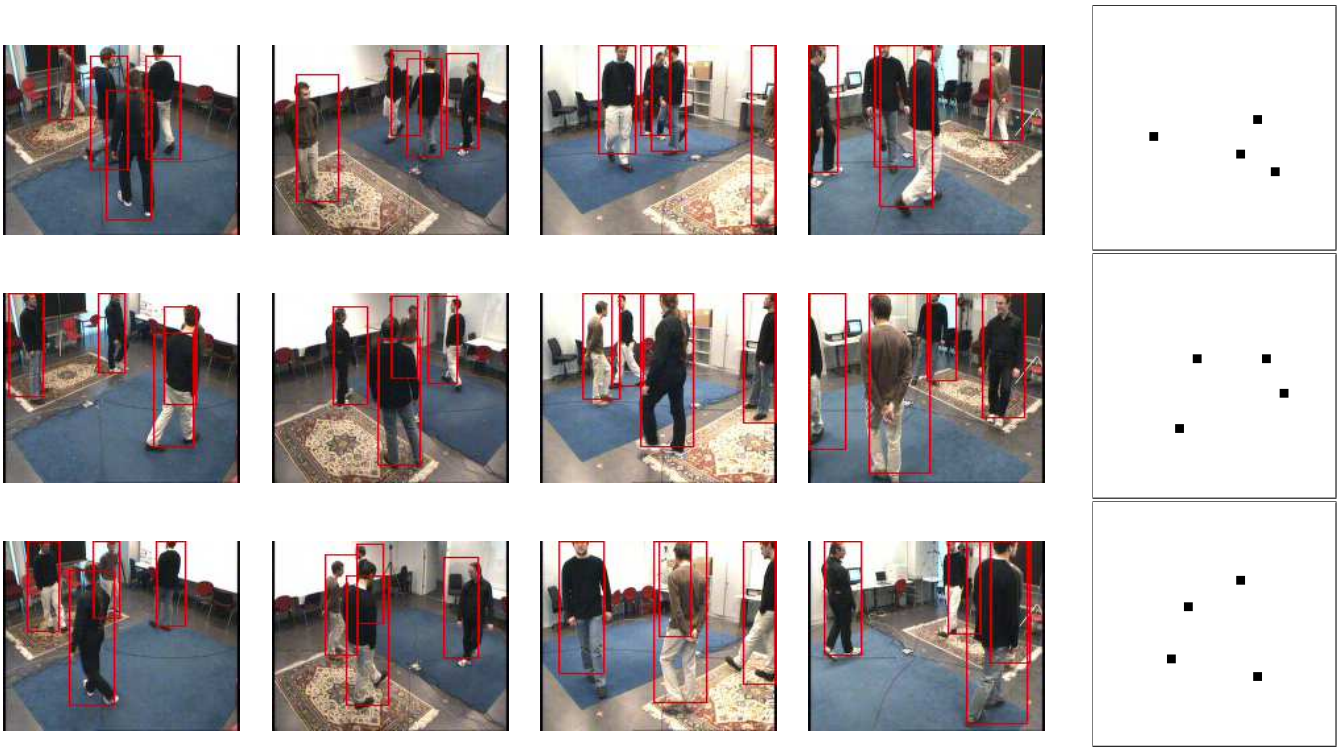


Figure 5. The result on each frame is illustrated by the original pictures from the 4 cameras with boxes representing detections (left) and localization in the top-view (right).

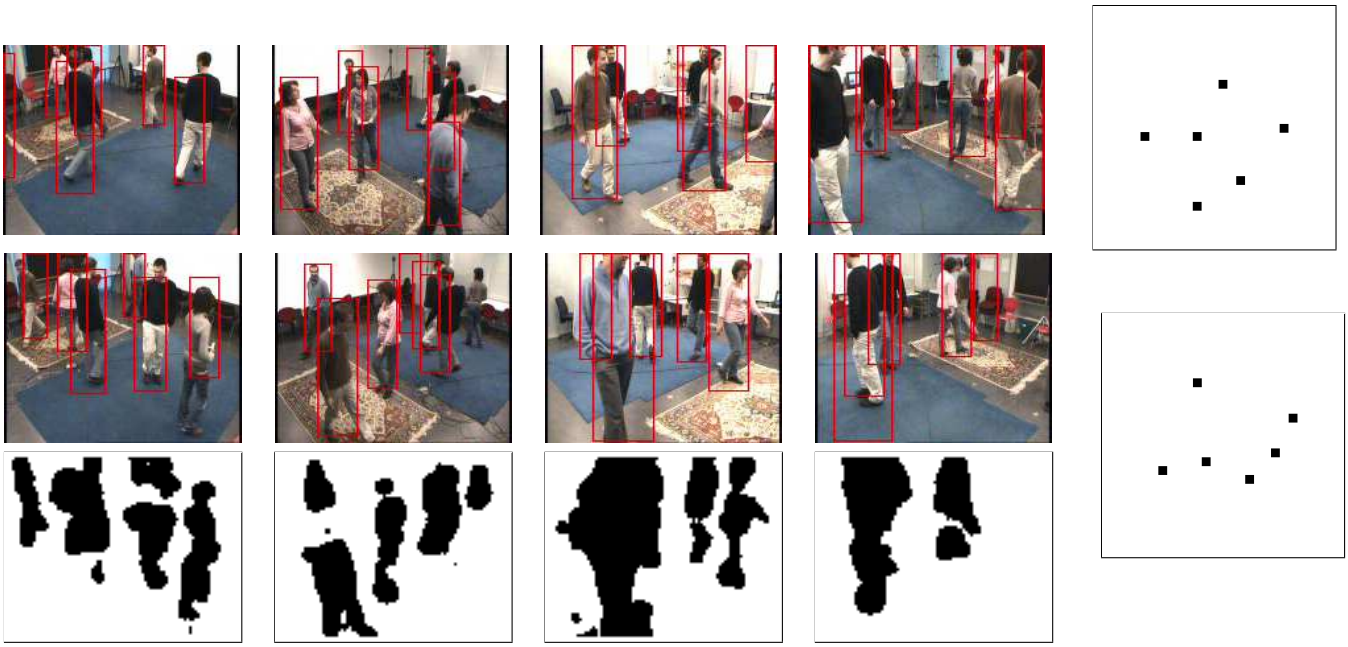


Figure 6. The result on each frame is illustrated by the original pictures from the 4 cameras with boxes representing detections (left) and localization in the top-view (right). The figure also shows the input binary blobs for the lower frame for a better understanding of the misdetection (second view from the left, individual on the right, due to the degraded corresponding blob)

that it may explain by locating a person in the scene, which is less likely to happen with taller models.

Using a simple criterion to terminate the process when it has converged, the average computation time is less than 150ms per frame on a 2.4Ghz Intel Pentium-4, given a room with 784 locations (which corresponds to an accuracy of 25 cm for the ground plane coordinates), 4 cameras, and pictures of size 90×72. The whole process can therefore be done at about 5 frames per second. Background subtraction is performed using a real-time implementation of a change detector that extracts foreground objects from a reference image that is automatically generated and updated over time. It is designed for industrial applications and provided by an industrial partner.

5. Conclusion

In this paper, we have presented a novel approach for multi-view multi-people detection. Our main concern was to design a frame-by-frame detection algorithm able to deal as efficiently as possible with the occlusions that inevitably occur in a surveillance context. This algorithm achieves excellent performance without using either temporal continuity or sophisticated texture measures, which is attributable to the mathematically well-founded generative model we use. We therefore view it as potentially very useful component of a more complete system that would also use these additional sources of information.

It does not rely on either accurate result from the background subtraction or complex human shape models. Therefore, the low quality of the binary blobs we can deal with, as well as the very primitive human shape model we use should make our algorithm well-suited for typical surveillance applications where the binary blob may not have a standard human shape, for example because the person is carrying a suitcase or wearing a hat.

Moreover since the underlying generative model is a mapping between a set of top view locations and a set of binary shapes in each camera view, the approach is guaranteed to be versatile. It is in no way specific to people detection and could be used with few changes for other object detection applications that require a powerful occlusion handling capability.

References

[1] R. Bhotika, D.J. Fleet, and K. N. Kutulakos. A Probabilistic Theory of Occupancy and Emptiness. In *European Conference on Computer Vision*, pages 112–130, 2002.

[2] J. Black, T.J. Ellis, and P. Rosin. Multi-view image surveillance and tracking. In *IEEE Workshop on Motion and Video Computing*, 2002.

[3] A. Broadhurst, T.W. Drummond, and R. Cipolla. A probabilistic framework for space carving. In *IEEE Proceedings on International Conference on Computer Vision*, 2001.

[4] Q. Cai and J.K. Aggarwal. Automatic tracking of human motion in indoor scenes across multiple synchronized video streams. In *IEEE Proceedings on International Conference on Computer Vision*, 1998.

[5] F. Fleuret, R. Lengagne, and P. Fua. Fixed point probability field for occlusion handling. Technical Report IC/2004/87, EPFL, October 2004.

[6] D. Focken and R. Stiefelhagen. Towards vision-based 3d people tracking in a smart room. In *IEEE International Conference on Multimodal Interfaces*, 2002.

[7] J. Giebel, D.M. Gavrila, and C. Schnorr. A bayesian framework for multi-cue 3d object tracking. In *Proceedings of European Conference on Computer Vision*, 2004.

[8] I. Haritaoglu, D. Harwood, and L. Davis. Who, when, where, what: A real time system for detecting and tracking people. In *Automated Face and Gesture Recognition*, pages 222–227, 1998.

[9] J. Kang, I. Cohen, and G. Medioni. Tracking people in crowded scenes across multiple cameras. In *Asian Conference on Computer Vision*, 2004.

[10] S. Khan, O. Javed, and M. Shah. Tracking in uncalibrated cameras with overlapping field of view. In *2nd IEEE Workshop on Performance Evaluation of Tracking and Surveillance*, 2001.

[11] S. Khan and M. Shah. Tracking people in presence of occlusion. In *Asian Conference on Computer Vision*, 2000.

[12] J. Krumm, S. Harris, B. Myers, B. Brummit, M. Hale, and S. Shafer. Multi-camera multi-person tracking for easy living. In *Third IEEE Workshop on Visual Surveillance*, 2000.

[13] I. Mikic, S. Santini, and R. Jain. Video processing and integration from multiple cameras. In *Proceedings of the 1998 Image Understanding Workshop, Morgan-Kaufman, San Francisco*, 1998.

[14] A. Mittal and L. Davis. M2tracker: A multi-view approach to segmenting and tracking people in a cluttered scene. *International Journal of Computer Vision*, 51(3):189–203, 2003.

[15] K. Okuma, A. Taleghani, N. de Freitas, J.J. Little, and D.G. Lowe. A boosted particle filter: multitarget detection and tracking. In *European Conference on Computer Vision*, Prague, Czech Republic, May 2004.

[16] K. Otsuka and N. Mukawa. Multi-view occlusion analysis for tracking densely populated objects based on 2-d visual angles. In *IEEE Proceedings of Computer Vision and Pattern Recognition*, 2004.

[17] T. Zhao and R. Nevatia. Tracking multiple humans in crowded environment. In *IEEE Proceedings of Computer Vision and Pattern Recognition*, 2004.