Fixed Point Probability Field for Occlusion Handling

(EPFL Technical Report ID: IC/2004/87)

François Fleuret Richard Lengagne Pascal Fua

Computer Vision Laboratory
Swiss Federal Institute of Technology (EPFL)
1015 Lausanne, Switzerland
Email: {François.Fleuret, Richard.Lengagne, Pascal.Fua}@epfl.ch

November 4, 2004

Abstract

In this paper, we show that in a multi-camera context, we can effectively handle occlusions at each time frame independently, even when the only available data comes from the binary output of a fairly primitive motion detector.

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.

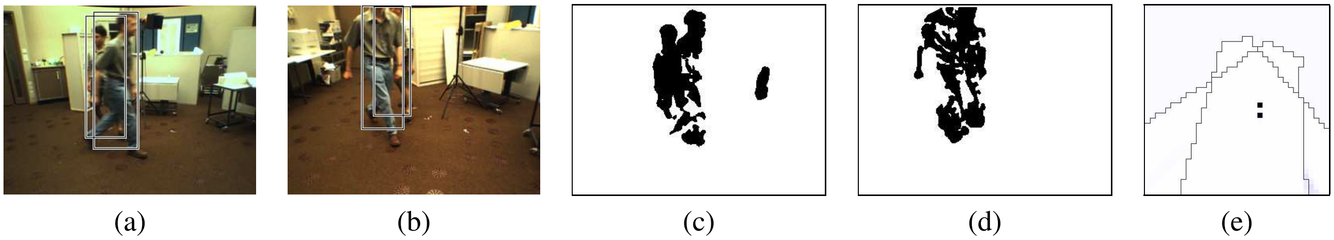


Figure 1: (a), (b): 2 images of 2 people occluding each other; (c), (d): binary motion blobs used as input to our algorithm; (e): resulting top view probability field. Note how well the two people are located in spite of the poor quality of the blobs. The corresponding locations are depicted by rectangles in (a) and (b).

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, such as the one depicted by Fig. 1, we can effectively handle occlusions at each time frame independently, even when the only data available comes from the output of a standard motion detector. 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 [13].

Given binary images such as those of Fig. 1, 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 probability 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 person'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 computation is the opposite of the feed-forward one performed by most state-of-the-art approaches [1, 4, 11, 12, 14] 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 [2] 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 motion blobs, such as [6], 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 [8], 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 [3], 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 [9], 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 [15], 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 [13], multiple people are detected and tracked in complex backgrounds using mixture particle filters guided by people models learnt by an Adaboost algorithm. In [5], 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[10] 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 [11], Kalman filtering is applied on 3D points obtained by fusing in a least-squares sense the image-to-world projections of points belonging to motion blobs. In [1], a Kalman filter is used to simultaneously track in 2D and 3D, and object locations are estimated through trajectory prediction during occlusion.

In [4], 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 [14], 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

[12] proposes a system that segments, detects and tracks multiple people in a scene using a wide-baseline setup of 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.

V	the set of binary views $(V_1, \dots, V_C)$ generated from the C video streams by the motion-detector.
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 3.
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

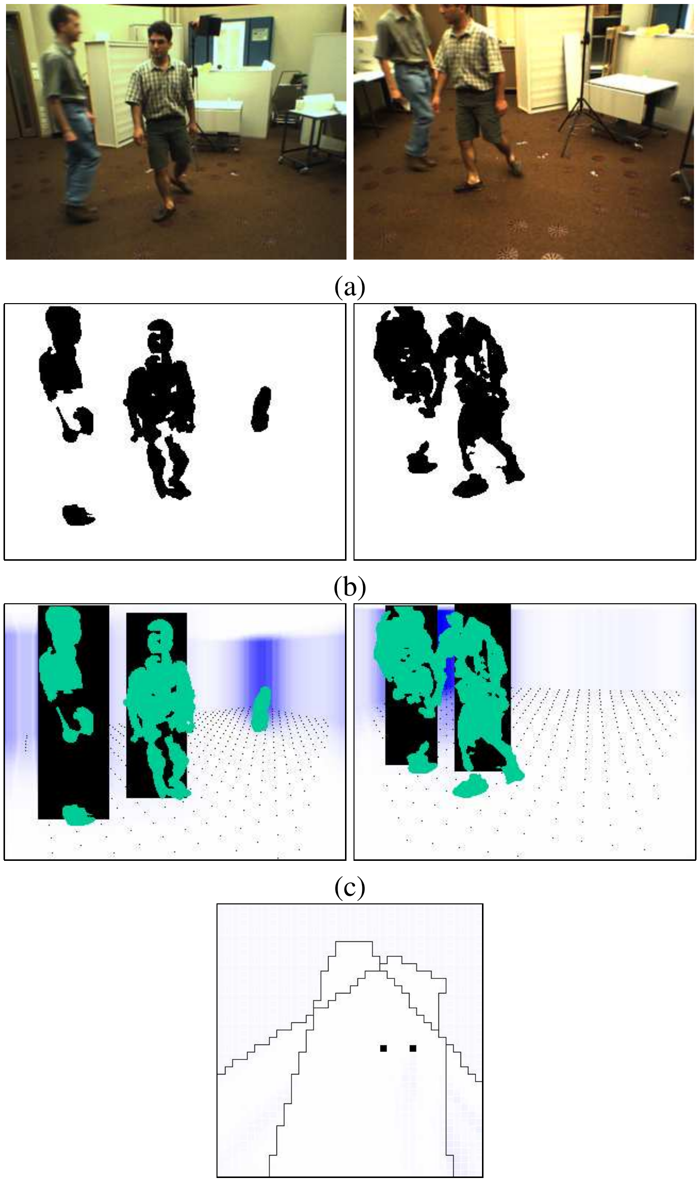
Secondly, evidence is gathered across cameras to compute a presence likelihood map on the ground plane which accounts for the visibility of each ground plane point in each view. Ground plane locations are then tracked over time using a Kalman filter.

In [7], 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 and lead to impressive results. 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

In practice, we create a virtual top view that we partition using a regular grid, as shown in figures 2(c) and 3. Our goal is to estimate the probabilities of occupancy $\rho_i = P(X_i = 1 \mid V)$, 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 point of an iterative process that repeatedly solves a large system of coupled equations.



(a)

(b)

(c)

(d)

Figure 2: Original views from two different cameras (a), the motion-based segmentation (b), the average images after convergence (c) and the corresponding probabilities in the top-view (d).

3.1 Probability of Occupancy

In Appendix A, we prove that

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

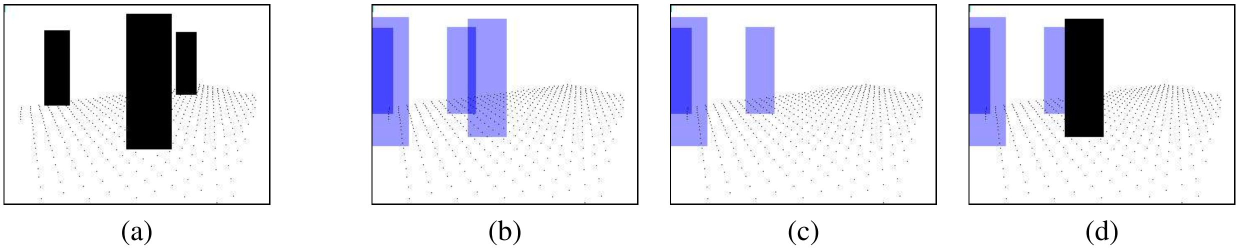

(a) (b) (c) (d)

Figure 3: 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 the notations are given in Table 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 3, 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. As explained in Appendix B, we do not actually generate those binary images, but compute directly the average image under an assumption of conditional independence of the X_i given V, and use integral images to compute quickly $\Delta(V_c, \overline{A}^c_{i,\xi})$.

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 Fig. 3 (c) or present with probability 1, as in Fig. 3 (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.

3.2 Algorithm

Note that, in Eq. 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 of Appendix B, at each iteration, we perform the following set of operations:

$$\begin{aligned}
\overline{A}^c &= 1 - \otimes_j \left(1 - \rho_j \mathcal{A}_j^c\right) \\
\sigma\left(\overline{A}^c_{i,\xi}\right) &= \sigma\left(\overline{A}^c\right) + \tfrac{\xi-\rho_i}{1-\rho_i}\,\sigma\left(\left(1-\overline{A}^c\right)\otimes \mathcal{A}_i^c\right) \\
\sigma\left(V\otimes\overline{A}^c_{i,\xi}\right) &= \sigma\left(V\otimes\overline{A}^c\right) + \tfrac{\xi-\rho_i}{1-\rho_i}\,\sigma\left(V\otimes\left(1-\overline{A}^c\right)\otimes \mathcal{A}_i^c\right) \\
\Delta(V,\overline{A}^c_{i,\xi}) &= \tfrac{1}{\sigma\left(\overline{A}^c_{i,\xi}\right)}\left(\sigma\left(V\right) - 2\,\sigma\left(V\otimes\overline{A}^c_{i,\xi}\right) + \sigma\left(\overline{A}^c_{i,\xi}\right)\right) \\
\rho_i &\leftarrow \tfrac{1}{1+\exp\left\{\lambda_i + \sum_c g(\Delta(V_c,\overline{A}^c_{i,0})) - g(\Delta(V_c,\overline{A}^c_{i,1}))\right\}}
\end{aligned}$$

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.

4 Results

In this section, we show several results we obtained from two 280-frame sequences shot by two cameras in a room where either 2 or 3 people were walking. We focus on difficult cases and show that our algorithm can still accurately detect and locate people in spite of sometimes severe occlusions. For each result, we display

- The original images with the reprojection on each view of each top view location where a person has been found standing. The detection criterion is defined as follows: A person is detected at a given location if this location a local maximum of the probability field inside a 1×1, or 3×3 or 5×5 neighborhood, and such that the sum of probabilities inside this neighborhood is larger than 0.75.
- The motion detection blobs used as input to our algorithm
- The top view with, for each location, the probability of occupancy. The darker the location, the higher the probability.

Note that, when displaying the probability of occupancy on the top view, no thresholding or post-processing of any kind has been applied.

In figure 4, the 2 people are seen distinctly in each view. In figure 5, the 2 people are seen distinctly in one view and occlude each other in the other view. In figure 1, the 2 people occlude each other in both views. In figure 6, each person is visible in only one view. In figures 7 and 8, there are 3 people. They are seen distinctly in one view and 2 people occlude each other in the other view. In figure 9, we display the true number of people in the room in each frame of both sequences versus the number of people detected by our algorithm.

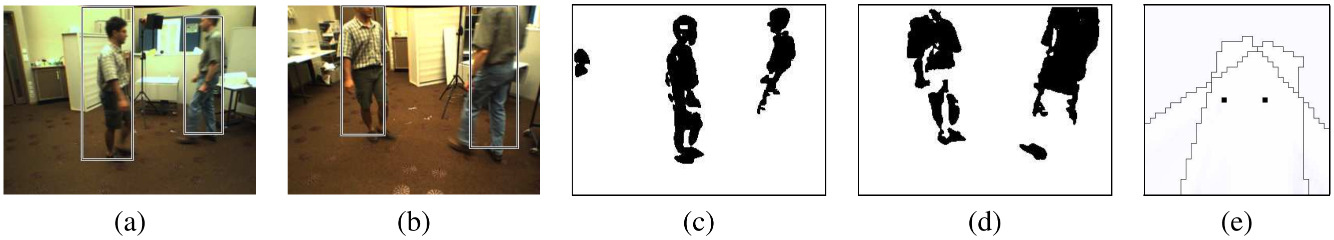


Figure 4: (a),(b): 2 people distinctly seen in each view; (c),(d): binary motion blobs; (e): top view probability field.

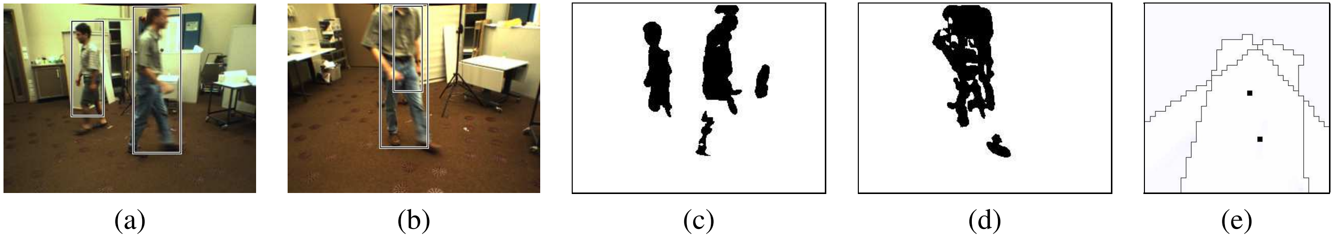


Figure 5: (a),(b): 2 people distinctly seen in one view and one occluding the other in the other view; (c),(d): binary motion blobs; (e): top view probability field

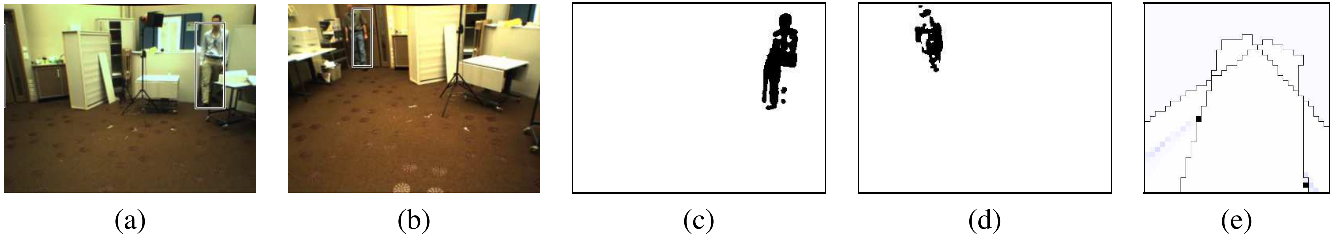


Figure 6: (a),(b): Each person visible in only one view; (c),(d): binary motion blobs; (e): top view probability field

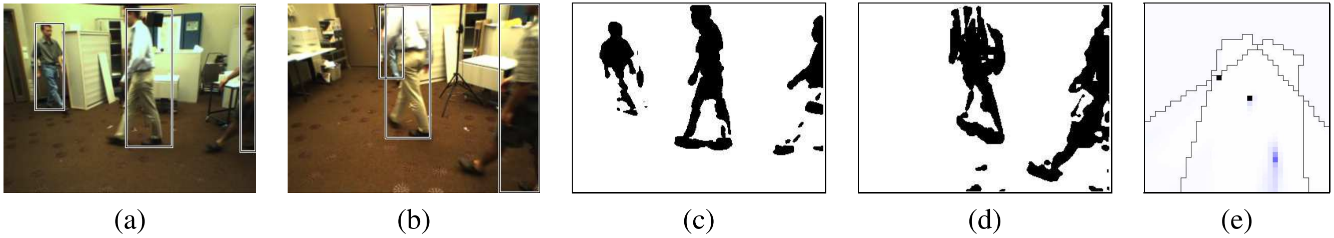


Figure 7: (a),(b): 3 people distinctly seen in one view and 2 people occluding each other in the other view; (c),(d): binary motion blobs; (e): top view probability field

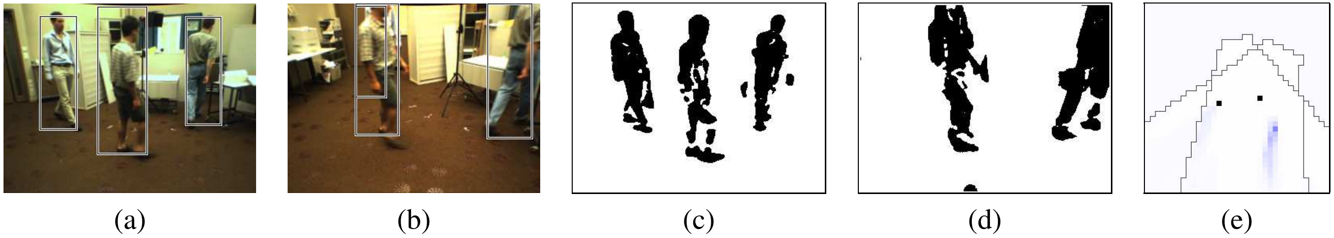


Figure 8: (a),(b): 3 people distinctly seen in one view and 2 people occluding each other in the other view; (c),(d): binary motion blobs; (e):top view probability field

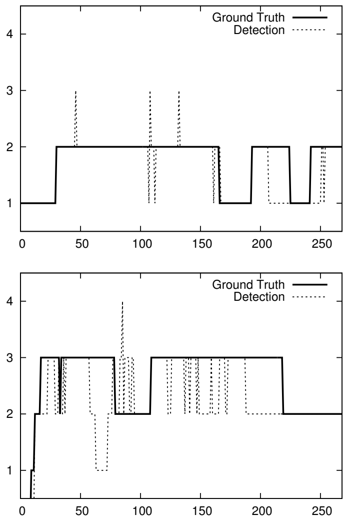


Figure 9: True number of people (ground truth) vs. the number of detected people over the 280-frame 2-people (top) and 3-people (bottom) sequences

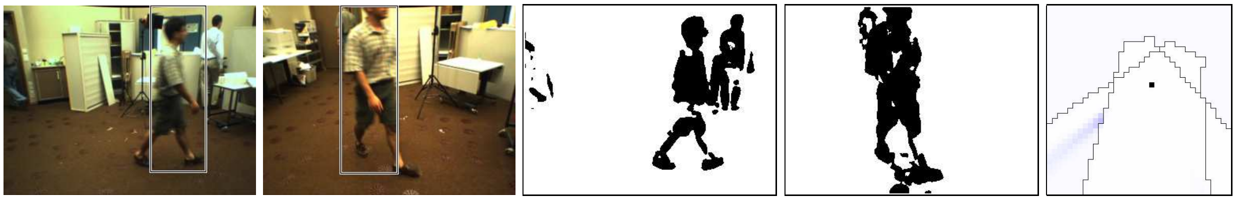


Figure 10: Failure mode: 3 people are distinctly seen in one view and 2 people occlude each other in the other view

We define the false positive rate as the ratio of false positives over the total number of persons detected throughout the sequence, and the misdetection rate as the number of false negatives over the total number of persons that should have been detected throughout the sequence. In the 2-people case, the false positive rate is 0.89% and the misdetection rate is 6.56 %. In the 3-people case, the false positive rate is 1.66% and the misdetection rate is 14.38%. Failures, such as the one shown in figure 10, can usually be attributed to one of two cases:

- The motion detection blobs are severely inaccurate, mostly because of lack of contrast between the person's texture and the background.
- One of the persons is visible in only one view, and the probability field too much spread around its true location to meet the above-mentioned detection criterion.

In other words, in absolute terms, our error rates are not sufficient for a complete system but they are excellent considering that we have only two cameras and use only a small proportion of the available image information. Our formalism is generic enough to incorporate other image clues that would definitely decrease the error rates.

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

The attached video files display the full sequences. Each file contains the original video streams with detected people, the motion blobs superposed on the average images after convergence and the resulting probability field on the top view. Also attached are 2 video files showing, for 2 single frames, the sequence of average images and the evolution of the probability field during the convergence process.

Using a simple criterion to terminate the process when it has converged, the average computation time is 3.4s per frame on a 1.4Ghz Pentium-M, given a room with 1332 locations (which corresponds to an accuracy of 20 cm for the ground plane coordinates), two cameras, and pictures of size 320×240. The motion detector is a real-time implementation designed for industrial applications and provided by VisioWave S.A [1].

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 a 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 motion blobs 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.

6 Acknowledgements

This work has been funded by the Swiss CTI/KTI project 6172.1 ENS-ET.

[1] http://www.visiowave.com

A Relation between the ρ_i

In this appendix, we give the derivation of equation (1). Let $X_{[i]}$ denote $(X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_N)$.

Note that if A, B and C are three finite random variables, we have

$$\begin{aligned}
&P(A = a \,|\, B) \\
&\quad= \sum_c P(A = a,\, C = c \,|\, B) \\
&\quad= \sum_c \frac{P(A = a,\, B,\, C = c)}{P(B,\, C = c)} \, \frac{P(C = c,\, B)}{P(B)} \\
&\quad= \sum_c P(A = a \,|\, B,\, C = c)\, P(C = c \,|\, B) \\
&\quad= E(P(A = a \,|\, B, C) \,|\, B)
\end{aligned}$$

From this, we get

$$\begin{aligned}
&P(X_i = 1 \,|\, V) \\
&\quad= E\left(P\left(X_i = 1 \,|\, V,\, X_{[i]}\right) \,|\, V\right) \\
&\quad= E\left(\frac{1}{1 + \frac{P(X_i{=}0 \,|\, V, X_{[i]})}{P(X_i{=}1 \,|\, V, X_{[i]})}} \,\middle|\, V \right) \\
&\quad= E\left(\frac{1}{1 + \frac{P(V \,|\, X_i{=}0, X_{[i]})}{P(V \,|\, X_i{=}1, X_{[i]})} \frac{P(X_i{=}0, X_{[i]})}{P(X_i{=}1, X_{[i]})}} \,\middle|\, V \right) \\
&\quad= E\left(\frac{1}{1 + \left(\prod_c \frac{P(V_c \,|\, X_i{=}0, X_{[i]})}{P(V_c \,|\, X_i{=}1, X_{[i]})}\right) \frac{P(X_i{=}0)}{P(X_i{=}1)}} \,\middle|\, V \right) \\
&\quad= E\left(\frac{1}{1 + \left(\prod_c \frac{\mu\left(V_c, \oplus_{j \neq i} X_j \mathcal{A}_j^c\right)}{\mu\left(V_c, \oplus_{j \neq i} X_j \mathcal{A}_j^c \oplus \mathcal{A}_i^c\right)}\right) \frac{P(X_i{=}0)}{P(X_i{=}1)}} \,\middle|\, V \right) \\
&\quad\simeq \frac{1}{1 + \left(\prod_c \frac{\mu\left(V_c, E(\oplus_{j \neq i} X_j \mathcal{A}_j^c \,|\, V)\right)}{\mu\left(V_c, E(\oplus_{j \neq i} X_j \mathcal{A}_j^c \oplus \mathcal{A}_i^c \,|\, V)\right)}\right) \frac{P(X_i{=}0)}{P(X_i{=}1)}}
\end{aligned}$$

The conditional expectation $E(\;.\; | \; V)$ above corresponds to an averaging over all possible configurations of presence of individuals $X_1, \dots, X_N$ in the room, given the current images acquired by the cameras.

This derivation involves two assumptions and a linearization. The first assumption is that individuals in the room do not take into account the presence of other individual in their vicinity when moving around, which is true as long as avoidance strategies are ignored. This leads to an independence assumption of the prior joint law $P(X_1, \dots, X_N) = \prod_n P(X_n)$. The second assumption is more subtle and implies that all statistical dependencies between views are due to the presence of individuals in the room. This can be understood as if the views were functions of the vector of presence $X_1, \dots, X_N$ and some independent noises: as soon as the presence of individuals is fixed, the views become independent. This is true as long as we ignore other hidden variables which influence several views at the same time (such as morphology, skin color or garment textures). We can formulate this assumption as $P(V_1, \ldots, V_C \mid X_1, \ldots, X_N) = \prod_c P(V_c \mid X_1, \ldots, X_N)$.

In the last line, we further assume that the conditional synthetic images $E(\oplus_{j \neq i} X_j \mathcal{A}_j \, \oplus x_i X_i \mid V)$ are highly concentrated, which legitimates the linearization of the expression. We swap the expectation and the functional to obtain a function of the average images. This conditional concentration is true by definition: only values of $X_1, \ldots, X_n$ leading to synthetic images compliant with V are likely, others have negligible probability. Even if those likely configurations in terms of locations of individuals can have huge variation, the corresponding synthetic pictures are similar.

We make the assumption that $\mu(v, \, a) = \exp g \left(\Delta(v, \, a)\right)$ where $\Delta(v, a) = \frac{\sigma(v \otimes (1-a) + (1-v) \otimes a)}{\sigma(a)}$ is a pseudo-distance between pictures, normalized with respect to the second operand and g the log of a normal law.

B Average images

Note that $\mathcal{A} \oplus \mathcal{B} = 1 - (1 - \mathcal{A}) \otimes (1 - \mathcal{B})$. Let $\overline{A}$ be current estimate of the average image. Since conditioning with V creates only a weak dependency between the X_i, which is not inconsistent with the coupling between the $P(X_i \mid V)$, we compute $\overline{A}$ under an independence assumption:

$$\begin{aligned}
\overline{A} &= E\left(\oplus_j X_j \mathcal{A}_j \mid V\right) \\
&= E\left(1 - \otimes_j \left(1 - X_j \mathcal{A}_j\right) \mid V\right) \\
&= 1 - \otimes_j (1 - E(X_j \mid V) \, \mathcal{A}_j) \\
&= 1 - \otimes_j (1 - \rho_i \, \mathcal{A}_j)
\end{aligned}$$

Under the same assumption, we have $\overline{\mathcal{A}}_{i,\xi} = 1 - \otimes_{j \neq i} \left(1 - \rho_j \mathcal{A}_j\right) \otimes \left(1 - \xi \mathcal{A}_i\right)$. We can write:

$$\Delta\left(V, \, \overline{A}_{i,\xi}\right) = \frac{\sigma\left(V\right) - 2\,\sigma\left(V \otimes \overline{A}_{i,\xi}\right) + \sigma\left(\overline{A}_{i,\xi}\right)}{\sigma\left(\overline{A}_{i,\xi}\right)}$$

Since $(1 - \rho_i)\left(1 - \frac{\xi - \rho_i}{1 - \rho_i}\right) = 1 - \xi$ we have $1 - \overline{A}_{i,\xi} = (1 - \overline{A}) \otimes \left(1 - \frac{\xi - \rho_i}{1 - \rho_i} \mathcal{A}_i\right)$, and we can easily show that

$$\sigma\left(V \otimes \overline{A}_{i,\xi}\right) = \sigma\left(V \otimes \overline{A}\right) + \frac{\xi - \rho_i}{1 - \rho_i}\,\sigma\left(V \otimes (1 - \overline{A}) \otimes \mathcal{A}_i\right)$$

The first term is the sum of all the pixels of $V \otimes \overline{A}$ and the second term is the sum of the pixels of $V \otimes \left(1 - \overline{A}\right)$ over the rectangle $\mathcal{A}_i$. This last computation can be done in constant time using integral images.

References

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

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

[3] 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.

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

[5] 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.

[6] 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.

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

[8] 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.

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

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

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