

# Tracking in Sparse Multi-Camera Setups using Stereo Vision

Gwenn Englebienne\*, Tim van Oosterhout<sup>†</sup>, Ben Kröse\*<sup>†</sup>

\*IAS Group, University of Amsterdam, Amsterdam, Netherlands

Email: {G.Englebienne,B.J.A.Krose}@uva.nl

<sup>†</sup>Hogeschool van Amsterdam, Amsterdam

**Abstract**—Tracking with multiple cameras with non-overlapping fields of view is challenging due to the differences in appearance that objects typically have when seen from different cameras. In this paper we use a probabilistic approach to track people across multiple, sparsely distributed cameras, where an observation corresponds to a person walking through the field of view of a camera. Modelling appearance and spatio-temporal aspects probabilistically allows us to deal with the uncertainty but, to obtain good results, it is important to maximise the information content of the features we extract from the raw video images. Occlusions and ambiguities within an observation result in noise, thus making the inference less confident.

In this paper, we propose to position stereo cameras on the ceiling, facing straight down, thus greatly reducing the possibility of occlusions. This positioning also leads to specific requirements of the algorithms for feature extraction, however. Here, we show that depth information can be used to solve ambiguities and extract meaningful features, resulting in significant improvements in tracking accuracy.

## I. INTRODUCTION

Wide-area multi-camera tracking is applied in a variety of situations, ranging from strongly constrained situations such as traffic monitoring [14], to highly unconstrained situations such as crowd flow analysis, intelligent homes [5, 12] or surveillance and security applications [1, 2]. In order to be successful, such applications require accurate tracking of individuals across cameras which, in practical applications, are typically sparsely distributed. Reliable tracking of objects or individuals in such a context is especially challenging, for a number of reasons. First, the objects visit the cameras irregularly and exhibit inhomogeneous motion outside of the cameras' view. Second, the absence of overlap between cameras precludes the use of continuous motion models such as Kalman filters or particle filters. Finally, different light conditions and viewing angles affect the object's appearance. Solving the association problem between different observations must therefore take into account travel distance, travel speed, possible paths between different cameras, object appearance and camera-specific lighting conditions.

Most importantly, accurate tracking over sparsely distributed cameras requires an accurate representation of the objects' appearance. In particular, we show that a global description of a person's colour distribution is suboptimal, and that features that include geometric information to describe an individual's appearance substantially improve tracking accuracy.

In this work, we focus on features for camera to camera associations. We therefore consider a full pass of a person through the field of view of a camera as a single observation, and build a probabilistic model of the observations and the associated people's identities. As we will see, these features are partially based on the RGB values of all the pixels associated with a person's travel through the camera's field of view. The quality of the observations affects the performance of the associations; it is therefore important to avoid, as much as possible, contaminating the features with pixel values from the background or other persons.

In this work, we use ceiling-mounted cameras, which are directed straight down. This layout has two main advantages:

- It is very rare, in such a set-up, for an individual to be constantly occluded while travelling through the camera's field of view. With wall-mounted camera's, occasional occlusions are commonplace and total occlusions, where no useful information can be extracted at all, do happen. This is far rarer with the current setup.
- As people travel through the camera's field of view, they are seen from widely varying angles. This provides us with a more robust estimate of the person's appearance: in the case of wall-mounted cameras, one camera may provide us with a frontal view of a person, while another provides us only with a view of the person's back. Depending on the clothing, these may result in very different appearances, thus making the association problem more difficult.

However, if ceiling-mounted cameras make the segmentation of individual people easier, it also makes structural information about each individual (such as the person's height, the colour of different regions along the vertical axis, *etc.*) harder to obtain. We show that the overall colour is a poor summary of a person's appearance, and that such structural information improves the tracking substantially. Such information is more difficult to obtain in the current setup, because, depending on the person's position, it is hard to evaluate which pixels are high on the body and which are low. We therefore use stereo vision to solve this problem.

The contributions of this work are twofold:

- 1) We argue for the value of using ceiling-mounted, straight-down cameras
- 2) We demonstrate that information about the structure of

appearance is important to reliable tracking and describe a method for extracting such features

## II. RELATED WORK

There is a substantial body of work on tracking of objects with cameras. Most past research has focussed on tracking within a camera's field of view and tracking across cameras with overlapping fields of view. In such situations, models of motion such as Kalman filters [11], particle filters [13] and the corresponding smoothing algorithms are possible. By contrast, the problem that this paper focusses on precludes the use of motion models across the cameras. The resulting association problem scales exponentially with the number of observations, so that approximations are required. A number of different approximations have been proposed in the past, such as heuristic approximations [8] and MCMC sampling [14]. Javed *et al.* approximate the posterior probability of the person's identity by finding the most likely identity associated with each observation [10], which can be computed in polynomial time.

In this work, we use a probabilistic model similar to the one proposed by Zajdel *et al.* [17] to infer the associations between observations, and compare how different appearance features compare under this model. In the original work, the cameras were positioned such that the people were observed from the side. In his approach, a single frame where the object was clearly visible was selected manually from the sequence of images of the pass through the field of view. Then, also manually, the image pixels were segmented into foreground and background classes. This procedure was followed to obtain a data set that is independent from various inaccuracies introduced by automated segmentation.

The foreground object was divided into three regions, to obtain a compact description of the appearance features  $\mathbf{a}$ . These regions (see Fig. 1 for details) are a heuristic choice based on the assumption that the observed objects are upright, walking people. For each region, a three-dimensional vector is computed, containing the mean colour (in RGB space) of the pixels in that region, yielding a 9D appearance feature  $\mathbf{a}$ . The resulting features provide a simple way to summarise colour while preserving partial information about geometrical layout. The resulting features are to a certain extent invariant to variations in person's pose.

Previous researchers have argued for the value of feature extraction from plan-view. Different researchers have proposed to use ceiling-mounted monocular cameras for tracking [4, 9]. The advantages of a top-down view are such that Harville proposes to use stereo cameras in traditional wall-mounted locations and to compute the top-down projection from the range information for tracking [5] and activity recognition [6].

Both monocular top-down views and stereo-vision-based plan-view reconstructions are, however, suboptimal, since in the former case the height information is missing, and in the latter case the advantage of having fewer occlusions is reduced. We show that by using ceiling-mounted stereo

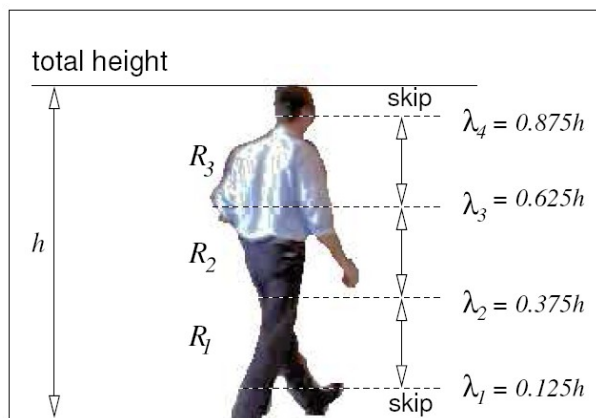


Fig. 1. Depiction of the segmentation of a silhouette, in 2D and lateral view. The same segmentation can be obtained from a top-down perspective if height information is available, taking advantage of the fact that as a person walks under the camera, their front, top and back are visible.

cameras, substantial improvements in tracking accuracy can be obtained.

## III. FEATURE EXTRACTION

In the current approach we use ceiling-mounted stereo cameras. The foreground is segmented automatically from the background, and we want to include information from all images of the pass. This means that the foreground pixels must be grouped in a "blob" per person, and that this blob must be tracked while it is in the camera's field of view. We first describe how the segmentation and tracking is performed, and then describe how the features are extracted from the results.

### A. Segmentation

The stereo cameras consist of two sensors, left and right. From each stereo camera we obtain at time step  $k$  a colour image,  $\mathbf{c}_k$ , obtained from one of the sensors, and a disparity map,  $\mathbf{r}_k$ , obtained by comparing the images from both sensors. The disparity map provides us with an estimate of the distance from the camera, which is transformed using the camera's calibration and stored as a vector of elevations. Since the disparity map is computed based on local texture information, its resolution is lower than the original colour images. Similarly, the disparity map is only valid in the area where the two cameras overlap. This information is part of the camera's calibration, and is used to fit the colour image onto the disparity map, so that the combined image  $\mathbf{j}_k$  contains both elevation and colour. For ease of exposition below, we denote  $\mathbf{c}_k(x, y)$  the elements of vector  $\mathbf{c}_k$  corresponding to the red, green and blue pixel intensities with coordinates  $(x, y)$  in the image. An example of the colour-augmented disparity map and the original colour image is given in figure 2. From this representation we extract a representation similar to Zajdel's representation.

First we segment the person from the background. This is done in two steps, based on the range and the colour image. First an adaptive segmentation algorithm [16] was used

on the disparity map  $r_i$  to segment foreground pixels from background pixels. An extension of this algorithm [18] is used for segmentation based on colour, which is combined with shadow suppression as described in [7]. The set of foreground pixels is then defined as the intersection set of the foreground pixels found by both algorithms. This approach deals with the issue of “foreground fattening” that occurs in the stereopsis-based approach, and has the advantage that objects that are connected in the colour image can often be easily segmented in the disparity map. This is illustrated in Fig. 2, where all three foreground objects are connected in the colour image, but can be easily segmented based on height.

### B. Tracking

The resulting set of foreground pixels is then grouped into blobs using the connected-components labelling algorithm described in [15]. A blob  $i$  is described by the corresponding set of connected pixel coordinates, denoted  $\mathcal{B}_i$ , and is tracked over consecutive frames. The robustness of the segmentation combined with the lack of ambiguities (resulting from the advantageous positioning of the camera) allows us to use a simple, ad-hoc tracking algorithm. The algorithm works as follows: a list of “active” objects is maintained as we step through the frames. Both the objects and the blobs are described in terms of their “position”, which is defined as the highest point in the object, and their bounding box.

A blob in the new frame is associated with an active object if:

- 1) The position of the blob is contained in the bounding box of the active object,
- 2) The position of the active object is contained in the bounding box of the new blob, or
- 3) The positions of the blobs are not distant by more than a heuristically chosen threshold (27cm in our case).

If multiple matches are possible, the closest match is chosen. When an active object reaches the edge of the field of view and disappears for more than a small number of frames (set to 10 in our case), the features are extracted, an observation is generated and the active object is removed from the list. In the discussion below, the index of the blob in frame  $k$  that is associated with observation  $j$  is denoted  $i_k^{(j)}$

### C. 3D features

For every blob  $i$  in frame  $k$  we calculate an appearance vector  $\mathbf{a}_{ik}$  in the following way. First the maximal height of the pixels in the blob is computed:  $r_i^{\max} = \max r_k(x, y) \mid x, y \in \mathcal{B}_i$ . The lower and upper limits are then shrunk by 12.5%, similarly to the approach of [17], to reduce the inclusion of non-informative pixels from the head and feet. The limits are therefore  $h_i^{\min} = \frac{1}{8}r_i^{\max}$  and  $h_i^{\max} = \frac{7}{8}r_i^{\max}$ . The pixels in frame  $k$  are then grouped into three groups:

$$\begin{aligned} S_i^1 &= \{(x, y) \in \mathcal{B}_i \mid h_i^{\min} \leq r_k(x, y) < \frac{1}{3}h_i^{\max}\} \\ S_i^2 &= \{(x, y) \in \mathcal{B}_i \mid \frac{1}{3}h_i^{\min} \leq r_k(x, y) < \frac{2}{3}h_i^{\max}\} \\ S_i^3 &= \{(x, y) \in \mathcal{B}_i \mid \frac{2}{3}h_i^{\min} \leq r_k(x, y) \leq h_i^{\max}\} \end{aligned} \quad (1)$$

For every group of pixels the average  $r, g$  and  $b$  colour values are calculated. After a full pass of a person through the field of view of the camera, the appearance feature of that person is the mean pixel value, over all pixels in the group, over all frames. Thus, the feature vector of observation  $j$ ,

$$\mathbf{a}_j = \begin{bmatrix} \mathbf{a}_{j1} \\ \mathbf{a}_{j2} \\ \mathbf{a}_{j3} \end{bmatrix} \quad (2)$$

consists of a 9D vector describing the appearance of a person as seen by that camera, which is computed as follows:

$$\mathbf{a}_{jn} = \frac{\sum_k \sum_{(x,y) \in S_{i_k^{(j)}}^n} \mathbf{c}_{i_k^{(j)}}(x, y)}{\sum_k |S_{i_k^{(j)}}^n|}, \quad (3)$$

where we sum over all frames  $k$  that affect observation  $j$ , *i.e.*, from the frame the person entered in the camera’s field of view to the frame they left, and  $|S_i^n|$  denotes the number of pixels in set  $S_i^n$ .

This feature extraction method deals with the occlusions that occur due to the positioning of the cameras, since only visible pixels are taken into account when constructing the appearance vector, and all pixel values that are observed during the person’s transit through the camera’s field of view are taken into account. The resulting feature vector is used for the camera to camera tracking using the model described below.

### D. 2D features

In the results section, we compare how features that include height information compare to features based on a person’s global colour. We therefore extract global features without including elevation information. These features are extracted based on the same tracking as described above (which does use stereo vision for the segmentation), but consist of the average RGB values of all the pixels in the tracked object, without the segmentation into three areas described above.

## IV. THE ASSOCIATION MODEL

Each observation  $\mathbf{y}_i = \{\mathbf{a}_i, t_i^e, t_i^l, l_i\}$  consists of the following variables: the location  $l_i$  (that is, the identifier of the camera that recorded the observation), the time of entry in the camera’s field of view  $t_i^e$ , the time of leaving its field of view,  $t_i^l$ , and a vector of appearance features  $\mathbf{a}_i$  obtained as described above. Let  $z$  be a categorical variable indicating the identity of a person. If we know the set of observations associated with a person  $z$ , the likelihood of that set of observations  $\mathbf{y}_{1:N_z}$  factorises as depicted in Fig. 3 and is given by:

$$p(\mathbf{y}_{1:N_z} | z) = \prod_{i=1}^{N_z} p(\mathbf{a}_i | z, l_i) p(l_i | l_{i-1}) p(t_i^e | l_i, t_{i-1}^l, l_{i-1}) \quad (4)$$

Notice that we do not include a term for  $p(t_i^l)$ , since it does not affect the model: indeed, in a more complete model,  $t_i^l$  would depend on  $t_i^e$ , but since the two are observed jointly,  $p(t_i^l | t_i^e) = 1$  for the observed  $t_i^l$  and  $t_i^e$ , so that we omit it from our description.

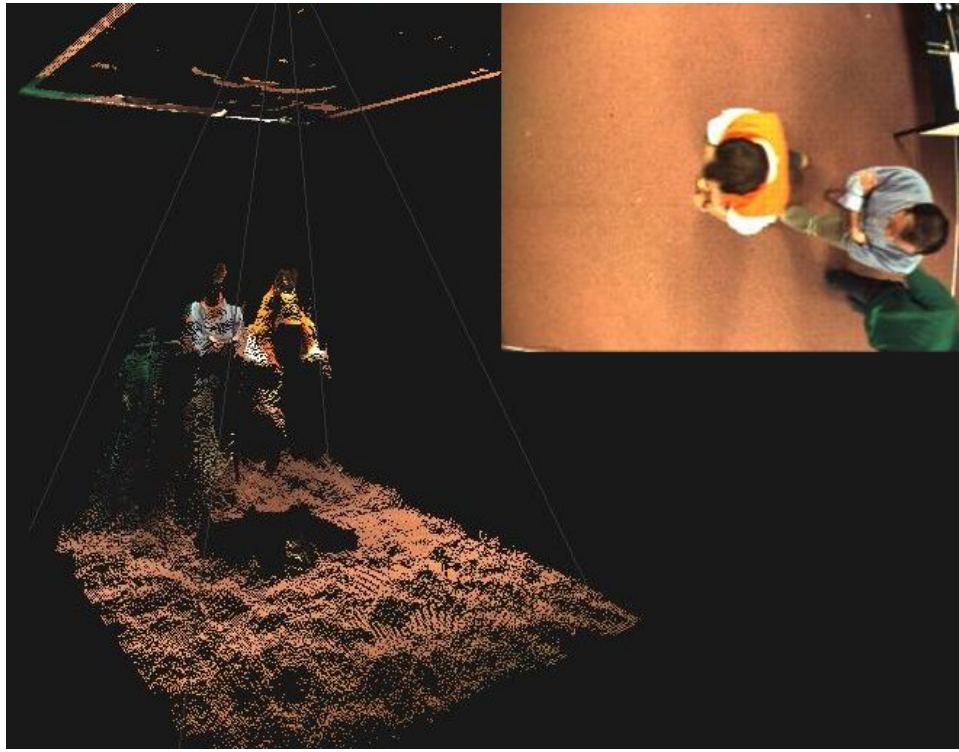


Fig. 2. Combined colour image and disparity map. The inset shows the view obtained by the camera, while the main image shows the 3D reconstruction of the same image, based on the disparity map. This illustrates how objects which are connected in the 2D image (all three people, in this case) can be easily segmented based on the elevation. The 3D reconstruction shows how many areas of a person’s body are hidden from the camera’s perspective; yet different portions become visible as the person moves across the camera’s field of view, and so a complete model of the appearance is constructed.

The appearance features are modelled with an additive model:

$$\mathbf{a}_i = \boldsymbol{\mu}_z + \mathbf{v}_i + \mathbf{w}_i, \quad (5)$$

where  $\boldsymbol{\mu}_z$  is the mean appearance vector for person  $z$ ;  $\mathbf{v}_i \sim \mathcal{N}(0, \boldsymbol{\Sigma}_z)$  is zero-mean, normally distributed noise, accounting for the spread in colour in the individual’s clothing; and the final noise term,  $\mathbf{w}_i$ , accounts for the difference in lighting conditions associated with each camera, differences in the camera’s settings *etc.*, and is distributed as  $\mathbf{w}_i \sim \mathcal{N}(\boldsymbol{\mu}_{l_i}, \boldsymbol{\Sigma}_{l_i})$ . The resulting distribution of  $\mathbf{a}_i$  is also normal, so that

$$p(\mathbf{a}_i | z, l_i) = \mathcal{N}(\mathbf{a}_i; \boldsymbol{\mu}_z + \boldsymbol{\mu}_{l_i}, \boldsymbol{\Sigma}_z + \boldsymbol{\Sigma}_{l_i}) \quad (6)$$

During training, we can trivially separate the two noise terms,  $\mathbf{v}_i$  and  $\mathbf{w}_i$ , since  $\mathbf{w}_i$  depends only on the camera (which is observed), while  $\mathbf{v}_i$  depends only on the person’s identity (which is jointly optimised using the EM algorithm).

The probability of transiting from one camera to the next,  $p(l_i | l_{i-1})$  depends on the environment and is set beforehand. Finally, the probability distribution of travel time between cameras, given by  $p(t_i^e | l_i, l_{i-1}, t_{i-1}^l)$  is difficult to model in practice, as it is typically strongly multimodal. Learning a precise parametric representation of the distribution would therefore require large amounts of data, which are not available in our application. We therefore choose to model it with a

uniform distribution:

$$p(t_i^e | l_i, l_{i-1}, t_{i-1}^l) \propto \begin{cases} 0 & \text{if } l_i = l_{i-1} \text{ and } t_{i-1}^l > t_i^e \\ 1 & \text{otherwise.} \end{cases} \quad (7)$$

where the limits of the uniform distribution are chosen such that the same person cannot be seen twice simultaneously by the same camera. Additional limits could be added to limit the search space of the algorithm, but in our case the sequences were short enough that we allowed for transitions between any observations, and the range of the uniform was thus defined by the length of the sequence. The simultaneous observation of the same person by different cameras has non-zero probability, to allow for the potential overlap of the cameras’ fields of view.

The conditional independences between consecutive observations of the same person are depicted in Fig. 3; observations stemming from different people are assumed to be independent. The complexity of this model stems from the fact that subsequent observations need not be from the same person: exact inference becomes intractable if multiple individuals are considered, as shown in Fig. 4. Inference and learning within a single partition is trivial, but the number of possible partitionings grows exponentially with the number of observations. In [17] this problem was solved by approximate marginalisation of the different partitionings of the data. This method, based on a combination of the EM algorithm [3] and a greedy search algorithm, was also used in this work.

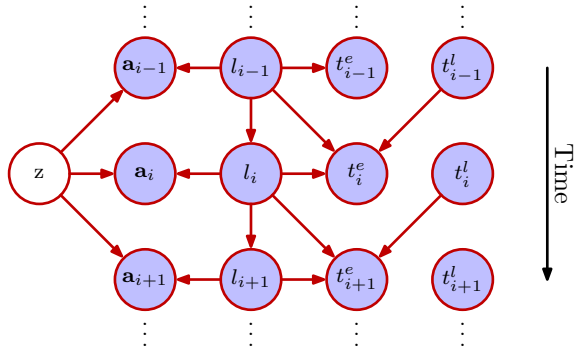


Fig. 3. Graphical model depicting the independence assumptions between two consecutive observations of the same person. See the text for details.

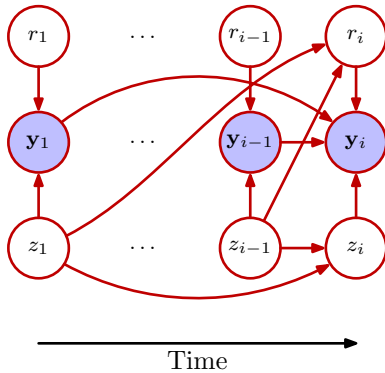


Fig. 4. Graphical model of the probabilistic model of observation  $y_i$ . Since the observations are ordered in time,  $y_i$  only depends on the identity of the person  $z_i$  and the previous observation of that person. However, since we do not know when the last observation of that particular person was,  $y_i$  depends on all past observations  $y_{1:i-1}$  and on an extra variable indicating what the last observation of the person was. The densely connected structure of the graph makes exact inference intractable.

In the experiments reported here, the transition probabilities between cameras was kept fixed, and only the parameters modelling the appearance were learnt, by maximum likelihood. The generalised EM algorithm was used, where the person’s appearance parameters,  $\mu_z$  and  $\Sigma_z$ , and the camera’s noise parameters,  $\mu_l$  and  $\Sigma_l$ , are optimised iteratively during the M step.

## V. RESULTS

We investigated the effect of the feature extraction on the tracking accuracy of our model on four different sequences, of approximately 1000 frames each. The sequences were

TABLE I  
ACCURACY OF THE TRACKING RESULTS

Seq.	Obs.	People	3D	2D
1	10	4	1.00	1.00
2	12	3	0.92	0.92
3	9	4	0.89	0.67
4	11	4	0.64	0.45

recorded with two *Point Grey bumblebee 2*<sup>1</sup> stereo cameras (*i.e.*, four sensors), and four volunteers were asked to walk randomly under the cameras. The fields of view of the cameras overlap slightly, but the overlap was not used in tracking, so that the only consequence of this overlap is that people can appear in one camera before they disappeared from the other. However, due to the way we modelled the travel time between cameras, this does not affect the results either.

The resulting sequences contain stretches with no observations, interspersed with short bursts of activity, where up to three people walk through the camera’s field of view simultaneously. The walking patterns were not constrained, and the sequences contain many different walking speeds, directions, people walking in close proximity, sudden changes of direction, *etc.*

Table I depicts the tracking accuracy, *i.e.*, the number of observations that are correctly associated with one identity over the total number of observations in the sequence, for four different sequences. The ground truth for these experiments was obtained by manually labelling the observations with the identity of the corresponding person. In this table, the column labelled ‘Obs.’ indicates the number of observations present in that sequence, ‘People’ lists how many different individuals were present in the sequence; ‘3D’ lists the results with the 3D features and ‘2D’ the association results when using the 2D features. The sequences contain few observations of the same person, making the problem more difficult, as less information is available about each person.

We can see that the use of geometric information leads to improved tracking accuracy. This difference becomes more marked as the sequences become more difficult and the individuals become harder to distinguish (mostly due to people walking along the edge of the camera’s field of view, resulting in less of the person being visible).

## VI. CONCLUSION

We have described a simple feature-extraction algorithm for the tracking of people over sparsely distributed cameras. We take advantage of the positioning of the camera to avoid occlusions, thus allowing us to use a simple and fast, heuristic tracking algorithm.

The location of the camera leads to ambiguities about the structure of the observed object: it is not easily possible to determine what pixels correspond to low and high elevations, and we show that this information is important for robust tracking across cameras. We use stereo vision to obtain this missing information, and show that the resulting tracking is improved markedly.

## REFERENCES

- [1] Qin Cai and Jake K. Aggarwal. Tracking human motion in structured environments using a distributed-camera system. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(11):1241–1247, 1999.

<sup>1</sup><http://www.ptgrey.com>

- [2] Robert T. Collins, Alan J. Lipton, Hironobu Fujiyoshi, and Takeo Kanade. Algorithms for cooperative multisensor surveillance. *Proceedings of the IEEE*, 89(10):1456–1477, 2001.
- [3] Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.
- [4] Michael Greiffenhagen, Visvanathan Ramesh, Dorin Comaniciu, Heinrich Niemann, and Visualization Department T. Statistical modeling and performance characterization of a Real-Time dual camera surveillance system. In *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, volume 2, pages 335–342, 2000.
- [5] Michael Harville. Stereo person tracking with adaptive plan-view templates of height and occupancy statistics. *Image and Vision Computing*, 22(2):127–142, 2004.
- [6] Michael Harville and Dalong Li. Fast, integrated person tracking and activity recognition with plan-view templates from a single stereo camera. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 2, pages II–398–II–405 Vol.2, 2004.
- [7] Thanarat Horprasert, David Harwood, and Larry S Davis. A statistical approach for real-time robust background subtraction and shadow detection. In *ICCV*, pages 1–19, September 1999.
- [8] Timothy Huang and Stuart J. Russell. Object identification: A bayesian analysis with application to traffic surveillance. *ARTIFICIAL INTELLIGENCE*, 103:1–17, 1998.
- [9] Stephen S. Intille, James W. Davis, and Aaron F. Bobick. Real-time closed-world tracking. In *Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on*, pages 697–703, 1997.
- [10] Omar Javed, Khurram Shafique, Zeeshan Rasheed, and Mubarak Shah. Modeling inter-camera space-time and appearance relationships for tracking across non-overlapping views. *Comput. Vis. Image Underst.*, 109(2):146–162, 2008.
- [11] R. E. Kalman. A new approach to linear filtering and prediction problems. *Journal of basic Engineering*, 82(1):35–45, 1960.
- [12] Rafael Muñoz-Salinas, Eugenio Aguirre, and Miguel García-Silvente. People detection and tracking using stereo vision and color. *Image and Vision Computing*, 25(6):995–1007, 2007.
- [13] Rafael Muñoz-Salinas, Rafael Medina Carnicer, Francisco J. Madrid-Cuevas, and Ángel Carmona-Poyato. Multi-camera people tracking using evidential filters. *International Journal of Approximate Reasoning*, 2009.
- [14] Hanna Pasula, Stuart J. Russell, Michael Ostland, and Yaacov Ritov. Tracking many objects with many sensors. In *International Joint Conference on Artificial Intelligence*, volume 16, pages 1160–1171. Lawrence Erlbaum Associates Ltd., 1999.
- [15] Linda G. Shapiro and George C. Stockman. *Computer Vision*. Prentice Hall, 2001.
- [16] Chris Stauffer and W. Eric L. Grimson. Adaptive background mixture models for real-time tracking. In *Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on.*, volume 2, 1999.
- [17] Wojciech Zajdel, A. Taylan Cemgil, and Ben J. A. Krose. Online multicamera tracking with a switching state-space model. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 4, pages 339–343 Vol.4, 2004.
- [18] Zoran Zivkovic. Improved adaptive gaussian mixture model for background subtraction. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 2, pages 28–31 Vol.2, 2004.