A Novel Multi-Planar Homography Constraint Algorithm for Robust Multi-People Location with Severe Occlusion

Paper ID:086

Abstract

Multi-view approach has been proposed to solve occlusion and lack of visibility in crowded scenes. However, the problem is that too much redundancy information might bring about false alarm. Although researchers have done many efforts on how to use the multi-view information to track people accurately, it is particularly hard to wipe off the false alarm. Our approach is to use multiple views cooperatively to detect objects and use objects silhouette on planes of different height to remove false alarm. To achieve this we adopt a novel multi-planar homography constraint to resolve occlusions and false alarm. Experimental results show that our algorithm is able to accurately locate people in crowded scene maintaining correct correspondences across views. Moreover, the false alarm rate is obviously reduced.

1 Introduction

Occlusion between people in crowded scenes is very common. Sometimes, the foreground region belongs to more than one people. Therefore, it is particularly difficult to determine which target the foreground region belongs to even using the color distribution, shape and orientation. Although many studies were done to track people under the occluded scenes, no algorithm is able to track people robustly in the case of occlusion. This problem is partially induced by the information loss in the 2D imaging process. Reversely, we human beings could track target accurately even in the case of very severe occlusion. One of the reasons is that we all have two eyes and can locate the people in 3D space. Illuminated by this binocular system, researchers try and use multiple views of the same scene in an effort to recover scene information that might be missing in the single view. In this paper we propose a multi-view approach to detecting and tracking people in crowded scenes. The problem we want to resolve is to accurately locate and track people even in sever occluded case. Obviously, under this condition, color distribution and shape information are of no use. Our method is based on geometrical constructs proposed by S. M. Khan and M. Shah[1]. Core idea is to find the pixels locating on each plane using a planar constraint. That is to find people's section of different height. All these sections constitute the 3D target distribution. If there is a target on the ground of a certain position, the corresponding point in each section image belongs to foreground. Multiplying these section figures, we can finally locate targets. The only requirement for our algorithm is that the targets and the scene are distinguishable so that we can segment the foreground approximately. Finally, the location information is propagated to each view.

One of preparations in our method is to get the homography mappings of different-height plane between different views. This is easy to achieve in a common multi-camera system. Only using at least four point pairs, we can calculate the mappings between two views. By this way, we can map the view without calibration information which is not easy to get in outdoor scenes. The rest of the paper is structured as follows. In section 2 we discuss related work. Section 3 details multi-planar homography constraint. In section 4 we present our algorithm using the multi-planar homography constraint to locate people in the overlooking field. Experimental results are provided in section 5. Finally, we conclude this paper in section 6.

2 Related work

Multi-view methods are mostly based on monocular methods, which can't resolve occlusion in the tracking system. In [2] multiple people are tracked with a Kalman filter in a single camera using 3D shape models of people that were projected back to image space to aid in segmentation and resolving occlusions. In [3], multiple people are also detected and tracked in front of complex backgrounds using mixture particle filters guided by people models learnt by boosting. Finally [4] proposes a particle-filtering scheme with a MCMC optimization which handles naturally entrances and departures, and introduces a finer modeling of

interactions between individuals as a product of pair wise potentials. These and other similar algorithms[5, 6, 7] are challenged by occluding and partially occluding objects, as well as appearance changes. Connected foreground regions may not necessarily correspond to one object, but might have parts from several of them.

In order to track people in crowded scenes even when occlusion happens, many researchers have dedicated to multi-view method. In [8] multiple people are tracked using 'color' model constructed from multi-view foreground color information. J. Berclaz, F. Fleuret, Pascal Fua[9, 10] use a probabilistic occupancy map to depict the probability of target existing in the space position. This method is not sensitive to foreground detection result. However, calibration information is required, which restricts the application of this method. What's more, one of the difficult problems is the false alarm due to redundancy information of multi view. Especially when the targets are more than the cameras, false alarm is serious. In [11], a multi-view method without foreground segmentation is presented. This method is based on the assumption that the foreground area belonging to the same target in different view is consistent in color distribution. Obviously, this assumption is not usually correctly. RecentlySaad M. Khan and Mubarak Shah[1] proposed a multi-view approach to tracking people in crowded scenes using a planar homography constraint. Although this method requires no calibration information, it is sensitive to the foreground segmentation result. Moreover, the less cameras, the higher false alarm rate the algorithm brings about.

Even though these methods attempt to resolve occlusions, problems like false alarm are brought about. Our method tracks multiple people in crowded scene basing on multi-planar constraint with low false alarm rate and low requirement for foreground segmentation accuracy and no calibration information.

3 Multi-planar Homography Constraint

The planar constraint proposed in [1] is described as follows.

Let $p = (x, y, 1)$ denote the image location (in homogeneous coordinates) of a 3D scene point in one view and let $p = (x', y', 1)$ be its coordinates in another view. Let H denote the homography of the plane $\prod$ between the two views and H_3 be the third row of H. When the first image is warped toward the second image using the homography H, then the point p will move to p_w in the warped image:

$$p_w = (x_w, y_w, 1) = \frac{Hp}{H_3 p} \tag{1}$$

For 3D points on the plane $\prod$, $p_w = p'$. For 3D points off $\prod$, $p_w \neq= p'$. The misalignment $p_w - p'$is called the plane parallax. Geometrically speaking warping pixel p from the first image to the second using the homography H amounts to projecting a ray from the camera center through pixel and extending it till it intersects the plane $\prod$ at the point often referred to as the 'piercing point' of pixel p with respect to plane $\prod$. The ray is then projected from the piercing point onto the second camera. The point in the image plane of the second camera that the ray intersects is p_w. In effect p_w is where the image of the piercing point is formed in the second camera. As can be seen in figure 1, 3D points on the ground plane have no plane-parallax while those off the plane have considerable plane-parallax.

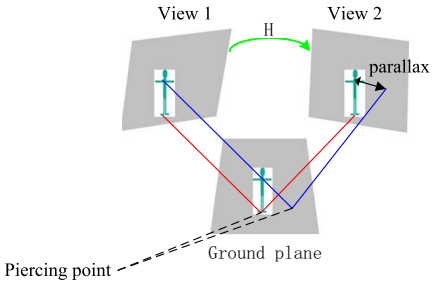

Figure 1 Mapping between two views

Figure 1 shows a person standing on the ground. The scene is being viewed by two cameras. is the homography of the plane ground from view 1 to view 2. Warping a pixel from view 1 with amounts to projecting a ray on to the ground plane at the piercing point and extending it to the second camera. Pixels that are image locations of scene points off the plane have plane parallax when warped. This can be observed for the blue ray in the figure.

Suppose a scene containing a ground plane is being viewed by a set of wide-baseline stationary cameras. The background models in each view are available and when an object appears in the scene it can be detected as foreground in each view using background difference. Any 3D point lying inside the foreground object in the scene will be projected to a foreground pixel in every view. The same is the case for 3D points inside the object that lie on the ground plane, except however that the projected image locations in each view will be related by homographies of the ground plane. Now we can state the following proposition.

Proposition 1. Let ϕ be the set of all pixels in a reference view and let H_i be the homography of plane $\prod$ in the scene from the reference view to view i. If $\exists p \in \phi$ such that the piercing point of p with respect to $\prod$ lies inside the volume of a foreground object in the scene then $\forall i, p'_i \ni \psi_i$ where $p'_i = H_i p$ and ψ_i is the foreground region in view.

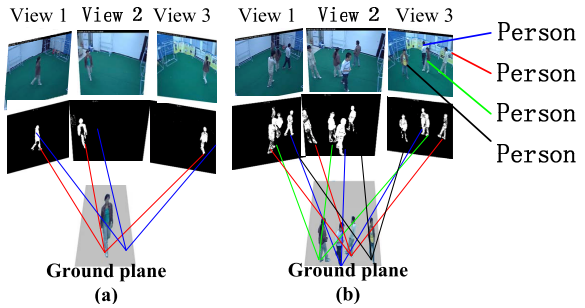

Figure 2 The first row shows people viewed by a set of cameras. The second row shows the foreground detected in each view.

However, false alarm might occurs by only using ground plane when some foreground pixels belonging to one people mapping with foreground pixels belonging to another people. What's more, if the feet in some view are not segmented completely, targets might be undetected. To resolve these, we propose a multi-planar constraint on the basis of proposition 1. Our proposition is described as follows.

Proposition 2. Let plane $\prod_j$ be the jth plane of all planes. Let ϕ_j be the set of all pixels that are foreground pixels in a reference view on $\prod_j$(according to Proposition1). A true target in the space should have foreground pixels on each plane $\prod_j$.

By this proposition, more stereo information is known about an object and more constraints are used so that we can eliminate false alarm effectively and lower undetected rate could be achieved due to the multi-planar information.

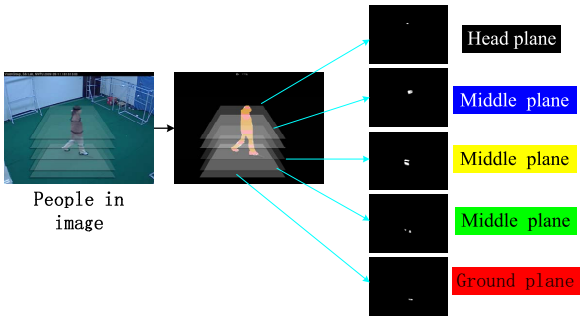

Figure 3 This figure illustrates Proposition 2.

A true person in the space should have section image on all the planes from ground plane to the head plane. If there is a false alarm in space, it might not have silhouette on all the planes. According to this principle, we can eliminate the false alarm. Besides, undetected rate can be reduced for we can get more information than only the feet on the ground plane. If foreground is not segmented completely, especially when one's feet are undetected, then the person would be undetected when only using ground planar constraint.

4 Using the Multi-planar Homography Constraint to Locate People

Our algorithm for locating people is easy to understand. First, we segment the foreground area using Gaussian mixture model-based background suppression[12] in each view. Then, on each plane of different height, the foreground pixels in all the other views are warped to the reference view and the warped results are multiplied to produce a section map. Finally, all these section maps are mapping to the overlooking scene and multiplied to locate people and wipe off the false alarm. The results would be propagated to each view.

Our algorithm is shown in Fig. 4. As we can see, it contains four key steps. The first step is to calculate the mapping homography between different views and between each view and the overlooking view. The second step is to obtain the section view on each plane. The next step is to combine all the section images on different planes to get the final object occupancy probabilistic field. Finally, we back propagate the result to each view.

The first key point is how to get the section image on a plane using four warped foreground images. Let's suppose we have n cameras and the mapping parameters θ between views are known. Under this suppose we can believe that different view images are independent of each other since the cameras do not have to do with each other. Let e_i denote the point i in space and I_k denote the foreground image of view k. then we can have formula (2).

$$\begin{aligned} p(e_i=1|I,\theta) &= \tfrac{p(I|e_i=1,\theta)p(e_i=1|\theta)}{p(I|\theta)} \\ &= \prod_{k=1}^{n} p(I_k|e_i=1,\theta)\, \tfrac{p(e_i=1|\theta)}{\prod_{k=1}^{n} p(I_k|\theta)} \\ &= \tfrac{p(e_i=1|\theta)}{\prod_{k=1}^{n} p(I_k|\theta)} \prod_{k=1}^{n} \tfrac{p(e_i=1|I_k,\theta)p(I_k|\theta)}{p(e_i=1|\theta)} \\ &= \tfrac{1}{(p(e_i=1|\theta))^{n-1}} \prod_{k=1}^{n} p(e_i=1|I_k,\theta) \end{aligned} \tag{2}$$

In formula (2), $(p(e_i = 1|\theta))^{n-1}$ is only dependent on the mapping parameter . So we can obtain the inference that

$$p(e_i=1|V) \propto \prod_{k=1}^{n} p(e_i=1|I_k,\theta) \tag{3}$$

According to formula (3), we can calculate the probability of a point being a foreground point in space by formula (4).

$$p(e_i=1|V) = \prod_{k=1}^{n} p(e_i=1|I_k,\theta) = \begin{cases} 1 & \forall k, I_k(i)=1 \\ 0 & else \end{cases} \tag{4}$$

In this way, we get the section image on a plane by calculating the probability map using formula (4).

The second key point is how to get the final object occupancy probabilistic field using the section images on each plane. Let N denote the plane number and S denote the section images on all planes. As we all know, the more cameras we have, the more accurate we can locate target in the scene and the more effective we can wipe off false alarms. The reason is that more cameras bring in more constraint to target so that we can get accurate target position. Our multi-planar constraint is exactly similar to multi cameras in this sense. So we combine the section images on different planes using formula (5).

$$p(e_i = 1|V) = \prod_{k=1}^{N} p(e_i = 1|S_k, \theta) = \begin{cases} 1 & \forall k, S_k(i) = 1 \\ 0 & else \end{cases} \tag{5}$$

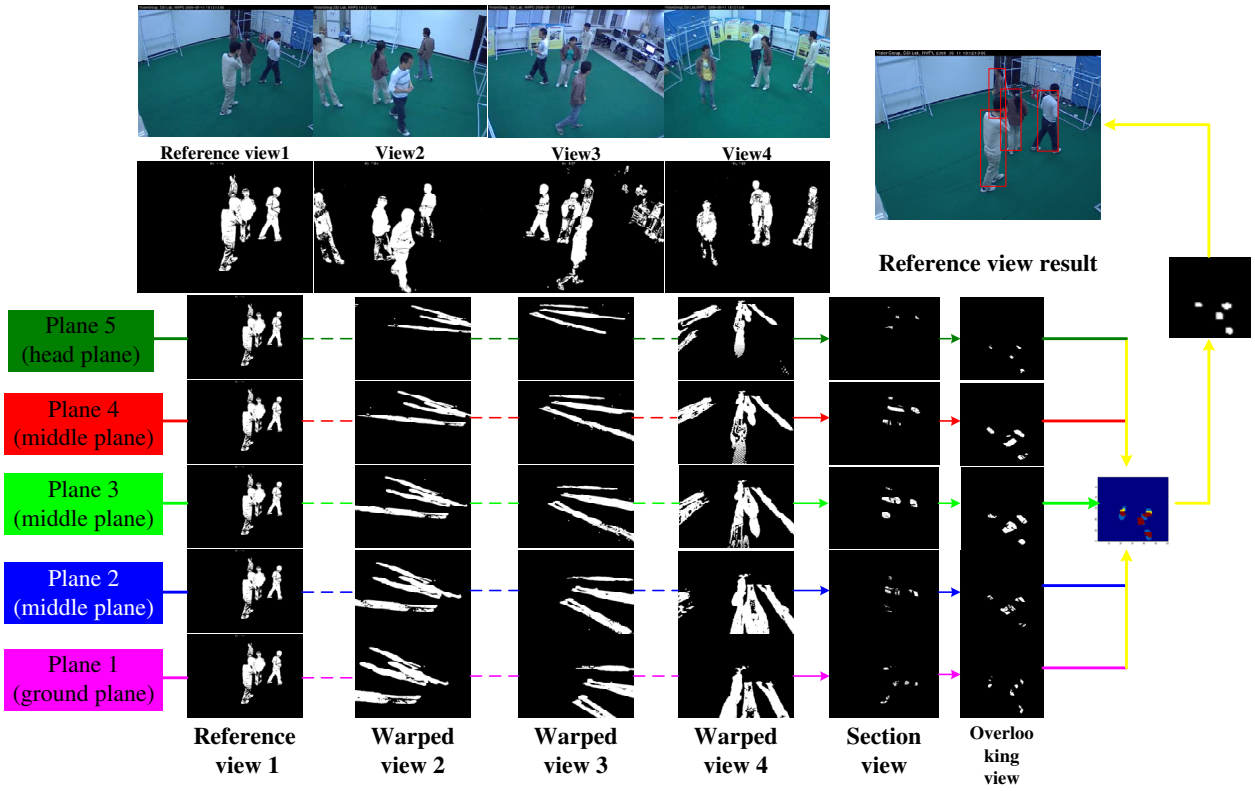

Figure 4 This figure shows how our method works in the actual scene. On the top row are the four views captured by four stationary cameras and the second row is foreground regions obtained using background suppression based on mixing gauss model. First four columns of the third row are the warped result of head plane. The fifth column is the head silhouette on the head plane and the sixth column shows the silhouette mapping to the overlooking view. The following rows are the result on the middle plane and the ground plane. The binary image on the right bottom is the final result using multi-planar constraint. The detecting result is finally back propagated to the reference view as in the image on the top right of Fig. 4.

Obviously, false alarm can be eliminated using the multi-planar constraint by formula (5). However, it often happens that foreground regions are not segmented completely and the section image is not accurate. Under this condition, true targets might be missed. Since we have information on different planes, we now consider use the multi-plane information complementary to reduce undetected rate.

Suppose we have Ni planes. We can calculate a probability map by formula (5). The bigger Ni is, the more authentic is the probability map. By changing Ni from 1 to N, we can get N probability maps. To make full use of the N probability maps, we sum them use different weights $w_{Ni}(w_{Ni} > w_{Ni-1})$, as in formula (6).

$$p(e_i = 1|V) = \sum_{Ni=1}^{N} w_{Ni} \prod_{k=1}^{Ni} p(e_i = 1|S_k, \theta) \tag{6}$$

By connection analysis and local maximum, targets are finally located. In this way, we can eliminate false alarm and reduce undetected rate by our multi-planar constraint. Table 1 shows the whole algorithm processes.

Tabel 1 Algorithm processes

- In the four views, choose view 0 as the reference view.
- Calculate each mapping h_j between the overlooking view and the view 0 on plane $\prod_j$
- Calculate homography H_{ij} between the reference view and view i on plane $\prod_j$
- for each frame
  - segment foreground regions using Gaussian mixture model-based background suppression
  - for each plane $\prod_j$

    for each camera view i

    warp the foreground result to the reference view using homography H_{ij}

    end

    Combine the warped results to get the section image on plane $\prod_j$ using formula (4)

    mapping the section image to the overlooking view using h_j to get the object occupancy probabilistic field
  - end

  Combine object occupancy probabilistic field on all planes to get the final target location by formula (6)

  Locate target by connection analysis and local maximum

  Back propagate the result to each view
- end

5 Experimental Results

We conducted several experiments on the basis of actual data and compared the results of our algorithm with the results of the algorithm using the one planar constraint. The configuration of the computer used for experiments is CPU Intel(R) Core(TM) 2 Duo 2.66GHz, RAM 2.0 G. Our experimental site is indoor space about 10mx10m and we have 4 cameras in the scene.

To evaluate our method,we conducted several experiments with varying number of people and cameras in the scene. Our purpose is to confirm that using the multi-planar constraint amounts to increasing the number of cameras. That is to say, we can detect much more people than just the number of cameras even in crowded scene, which is impossible for traditional multi-camera detecting methods. We compare our method's results with the traditional methods'. Part of the experiment result is displayed as follows.

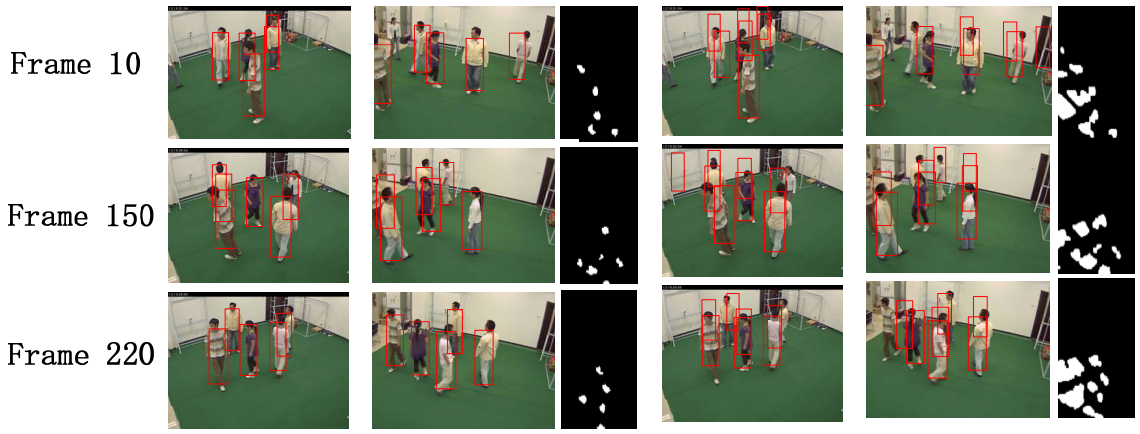

Figure 5 This figure shows the comparisons of our method and the traditional method using two cameras. The first three columns are the detection results of frame 10,150,220 and the last three columns are the detection results of overlooking view by our method. The third and fourth columns are results by the traditional method using a planar constraint.

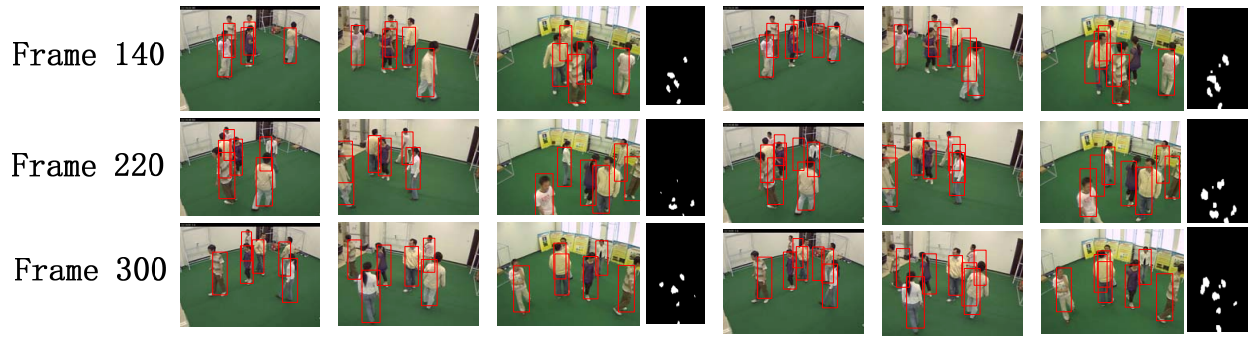

Figure 6 This figure shows the comparisons of our method and the traditional method using three cameras. The first four columns are the detection results of frame 140,220,300 and the last four columns show the detection result of overlooking view by our method. The third and fourth columns are results by the traditional method using a planar constraint.

From Fig. 5 and Fig.6 we can see both our method and the traditional method could resolve occlusion even in crowded scene. However, many false alarms occur due to too few constraints when there are more targets than cameras in traditional method. Usually, cameras need to be increased to increase the constraint to targets so that the false alarm could be wiped off. Here, we adopt a multi-planar constraint to replace increasing cameras. In this way, we can use the fewest cameras, even only two, to detect almost all the people (up to 6 people) with few false alarms in our 10mx10m scene. To confirm our method is more effective to wipe off false alarm and reduce undetected rate, we obtain the error rate ($ErrRate$) using statistical analysis. False alarm rate is calculated as formula (7).

$$ErrRate = \frac{falsealarmnumber + undetectednumber}{truetargetnumber} \tag{7}$$

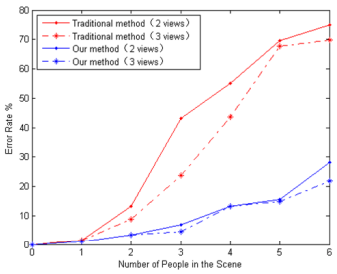

Figure 7 This figure shows the false alarm per frame of our method and the traditional method with different number of people in the scene.

Fig.7. shows the false alarm rate (statistic result from 1000 frames) of our method and the traditional method.We can see that our method can wipe off false alarm effectively even there are many targets in the scene compared with the traditional method by Fig.7. Besides, it is obvious that the traditional method could not work well if there are too few cameras. Our method is not sensitive to camera numbers for it could performance well even using only two cameras.

6 Conclusion

We propose a novel multi-camera method to detect people in crowded scene. The major contribution of our work is elimination of false alarm using a multi-planar constraint instead of the increasing cameras. For the plane of a height, we combine the foreground image from all views into a reference view to get the target section on the plane using a planar constraint. Using our multi-planar constraint, we mapping the section image on all planes to the overlooking view and combine them to get the final overlooking view. The last step is to locate people by simply clustering the overlooking map. In the future, we plan to do research on tracking people in the overlooking view as some people are too close to distinguish.

References

[1] S. M. Khan and M. Shahr. A multiview approach to tracking people in crowded scenes using a planar homography constraint. *Lecture Notes in Computer Science,Computer Vision - ECCV 2006*, 3954 LNCS(7):133–146, January 2006.

[2] T. Zhao and R. Nevatia. Tracking multiple humans in crowded environment. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2(7):II406–II413, January 2004.

[3] K. Okuma, A. Taleghani, N. Freitas, J. Little, and D. Lowe. A boosted particle filter: Multitarget detection and tracking. *European Conference on Computer Vision*, LNCS 3021(7):28–39, May 2004.

[4] K. Smith, D. Gatica-Perez, and J. M. Odobez. Using particles to track varying numbers of interacting

people. *Conference on Computer Vision and Pattern Recognition*, I(7):962–969, January 2005.

[5] S. J. McKenna, S. Jabri, Z. Duric, A. Rosenfeld, and H. Wechsler. Tracking groups of people. *Computer Vision and Image Understanding*, 80(1):42–56, October 2000.

[6] R. Rosales and S. Sclaroff. 3d trajectory recovery for tracking multiple objects and trajectory guided recognition of actions. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2(7):117–123, January 1999.

[7] H. Sidenbladh, M. J. Black, and D. J. Fleet. Stochastic tracking of 3d human figures using 2d image motion. *ECCV 2000*, LNCS 1843(7):702–718, January 2000.

[8] J. Orwell, S. Massey, P. Remagnino, D. Greenhill, and G. A. Jones. A multi-agent framework for visual surveillance. *ICIP 1999*, 99(7):1104–1107, January 1999.

[9] J. Bercla, F. Fleuret, and P. Fua. Robust people tracking with global trajectory optimization. *Proceedings - 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2006*, 1(7):744–750, January 2006.

[10] F. Fleuret, J. Berclaz, R. Lengagne, and P. Fua. Multicamera people tracking with a probabilistic occupancy map. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2):267–282, January 2008.

[11] R. Eshel and Y. Moses. Homography based multiple camera detection and tracking of people in a dense crowd. *CVPR.2008*, 2(ISBN: 978-1-4244-2242-5):1–8, June 2008.

[12] N. M. Brisson and A. Zaccarin. Moving cast shadow detection from a gaussian mixture shadow model. *CVPR 2005*, 2(7):643–648, January 2005.