Ghost Pruning for People Localization in Overlapping Multicamera Systems

Muhammad Owais Mehmood[1], Sebastien Ambellouis[1] and Catherine Achard[2]
[1]*LEOST, French Institute of Science and Technology for Transport, Spatial Planning, Development and Networks, Villeneuve D'Ascq, France*
[2]*Institute for Intelligent Systems and Robotics, Universite Pierre et Marie Curie, Paris, France*

Keywords: People Localization, Ghost Pruning, Multicamera Surveillance.

Abstract: In this paper, we propose a novel ghost pruning technique for multicamera people localization in overlapping scenarios. First, synergy map is obtained from multiplanar projections across multiple overlapping cameras. Second, occupancy map is generated by back projection from the synergy map across various image layers. This back projected occupancy map is combined with constraints to remove ghosts. The novelty of this paper is the introduction of an intuitive ghost pruning technique, which does not require any temporal information. Experiments on a sequence of the PETS 2009 dataset show significant reduction in the number of ghosts. The purpose and novelty of this paper is focused to the ghost pruning module but detection metrics show results comparable to those of the complete, state-of-the-art multicamera object detection systems.

1 INTRODUCTION

Advances in the imaging technology have enabled ubiquitous presence of cameras allowing multicamera setups to be used in scenarios like visual surveillance, 3D reconstruction or human modeling. As humans are an integral part of the environment, their localization, tracking and analysis is an important aspect of research. Significant research has been done on single camera people localization but the approach remains restricted for example at small scales, under occlusion (Dollar et al., 2012). Multicamera localization is thus a straightforward extension in order to improve the localization accuracy and robustness.

Fusion of information across multiple views is a challenging aspect of multicamera systems requiring some level of consistency across all the views of the object of interest, including its presence or not. Homography and planar projection techniques provide a reasonable degree of success when addressed to solve this issue (Eshel and Moses, 2010; Fleuret et al., 2007; Khan and Shah, 2009). However, the multiview homography constraint suffers from false detections known as "ghosts". Here it is important to point out that even if temporal information and tracking techniques are known to reduce ghosts, this papers tries to solve the ghost problem at its fundamental and uses only spatial information. The technique introduced in this paper is simple, intuitive yet effective because the detections are verified across the multiple camera views by using back projected occupancies. Hence, the novelty of this paper is to perform effective ghost pruning without temporal information.

The remainder of this paper proceeds with a discussion on the state-of-the-art methods. Section 3 introduces the concepts of multiplanar projections and synergy map. Section 4 addresses the back projected occupancies. Experimental setup and the results on PETS 2009 dataset along with comparisons are presented in Section 5 with conclusion in the end.

2 RELATED WORK

Since the last few decades, people localization has been an active area of research. There has been a lot of development in the single camera localization, a recent literature survey is (Dollar et al., 2012). However, as summarized by Dollar et al., single camera algorithms are limited in terms of scale, handling occlusion and for performance in the dense and cluttered environments. Introducing tracking, as presented in (Yilmaz et al., 2006), may alleviate the problem but it does not completely solve it. Compared to single camera, multicamera systems can inherently provide more information and thus have been exploited

Mehmood M., Ambellouis S. and Achard C..
Ghost Pruning for People Localization in Overlapping Multicamera Systems.
DOI: 10.5220/0004741306320639
In *Proceedings of the 9th International Conference on Computer Vision Theory and Applications* (VISAPP-2014), pages 632-639
ISBN: 978-989-758-004-8
Copyright © 2014 SCITEPRESS (Science and Technology Publications, Lda.)

for this problem. This section specifically focuses on the multicamera surveillance applications for overlapping cameras and literature based on the ghost pruning methods.

Registration of an object present across multiple camera views can be used to estimate its location. One common approach in these systems constraints the search space to the ground plane using the planar world assumption (Eshel and Moses, 2010; Fleuret et al., 2007; Khan and Shah, 2009). Therefore, assuming that the objects do not float in the air, planar homographies are calculated for the ground plane. Recent approaches extend this by using multiplanar homographies combined with the ground plane but this is not robust for several reasons such as the bad foreground detections or the occlusion of the lower part of the body.

Khan and Shah introduce the planar homographic constraint at multiple planes and combine it with graph cut segmentation to track people (Khan and Shah, 2009). No calibration information is required but planar references must be present in at least one of the views and affine homography must be manually computed by the user for each sequence. Their proposed solution suffers from false positives or ghosts due to the limitations of the homography constraint. Khan and Shah account for ghosts using the space-time occupancies. Eshel and Moses perform people tracking in a dense, crowded environment using homography constraints at the top layers combined with the pixel intensity correlation and motion direction, velocity constraints (Eshel and Moses, 2010). This method requires the use of partial calibration data. Temporal information is used to reduce phantoms. But, the algorithm is limited to those sequences in which heads are visible in a top view configuration. Different from the first two techniques, Fleuret et al. define a probabilistic occupancy map based on a quantized ground plane along with a distance measure in relation to the multiview projections (Fleuret et al., 2007). They further integrate it with Hidden Markov Model (HMM) for joint color, motion and occupancy modeling to perform tracking. However, this algorithm is limited to tracking up to a maximum of six people, performs poorly in dense situations and fails to account for height variations like the detection of children. More recently, Utasi and Benedek introduce novel features, a 3D configuration model and its optimization in order to perform multicamera people detection (Utasi and Benedek, 2011; Utasi and Benedek, 2012).

In parallel with the complete detection or tracking systems, research also focuses on resolving more fundamental issues such as ghosts. Ren et al. define ghosts as the false positives due to the intersections of non-corresponding regions (Ren et al., 2012). They propose to use color template matching for ghost pruning. But, as we will show later, their method is unable to account for views with high variations in the color constancy. Moreover, their equations are limited to only two views. Unlike (Ren et al., 2012), our proposed algorithm has no limitation in the number of views, number of planes used and is able to account for views which lack color constancy. Evans et al. introduced a suppression map technique which is able to predict the possible location of the ghosts based on the scene geometry but it requires prior information about the location of the objects of interest which is obtained from the previous frames (Evans et al., 2012). Unlike this method, our proposed technique does not require any temporal information.

The novelty of our method is to perform ghost pruning without using temporal information. We also account for color constancy variations and our algorithm can work across more than two camera views by taking into account the planes at several heights of the body, not just the top. Our algorithm has been tested on the *City Center* sequence of the PETS 2009 dataset using three overlapping camera views (PETS, 2009). The results show significant reduction in the number of ghosts, including a comparison with (Ren et al., 2012). Besides this, we achieve detection rates which are better than the *Probability Occupancy Map* (POM) detector module of (Fleuret et al., 2007) and results comparable to one of the more recent multicamera people detector in (Utasi and Benedek, 2012).

3 MULTIPLANAR PROJECTIONS AND SYNERGY MAP

The multiplanar projection algorithm as proposed in (Utasi and Benedek, 2012) is used. The inputs for the algorithm are the foreground masks $F_v(x,y)$ of each view v. Instead of using the Mixture of Gaussians (MoG), as employed by Utasi and Benedek, our foreground masks are obtained using the more robust multilayer background subtraction method as proposed in (Yao and Odobez, 2007). Next, the multiplanar projections are used to create the synergy map as explained in the following two sections.

3.1 Multiplanar Projections

The camera calibration model is used to project the silhouettes obtained by background subtraction to the ground plane and the planes parallel to it. If

(x_c, y_c) represent the ground coordinates of any camera placed at height h_c then as presented by (Utasi and Benedek, 2012):

$$x_h = x_0 - (x_0 - x_c)h/h_c$$
$$y_h = y_0 - (y_0 - y_c)h/h_c \quad (1)$$

where, (x_0, y_0) are the ground coordinates of an arbitrary point and (x_h, y_h) are the coordinates of the same point projected at a height h parallel to the ground.

If camera calibration model or homography information is not available then the method proposed by Khan and Shah in section 4.2 of their paper can be used (Khan and Shah, 2009). Figure 1 illustrates the projections obtained on a randomly selected frame of the PETS 2009 sequence.

(a) (b) (c)

Figure 1: Multiplanar projections on a frame of PETS 2009. Projections at (a) ground (b) 100 cm (c) 190 cm.

3.2 Synergy Map

Khan and Shah define synergy map as the 2D grid of object occupancy likelihoods (Khan and Shah, 2009). Synergy map is obtained by the fusion of the multiplanar projections obtained in the last section. Let us assume that $P_{v,p}(x,y)$ is the map corresponding to the projection of the camera view $v \in V$ on the plane $p \in P$ and n_v is the total number of views. Then, the synergy map $S(x,y)$ and normalized synergy map $S_n(x,y)$ are generated as follows:

$$S(x,y) = \frac{1}{n_v} \sum_P \prod_V P_{v,p}(x,y)$$
$$S_n(x,y) = \frac{S(x,y)}{max(S(x,y))} \quad (2)$$

For the rest of the paper we will use the normalized synergy map and refer to it as $S(x,y)$ or simply S. Figure 2 shows an example of the synergy map generated.

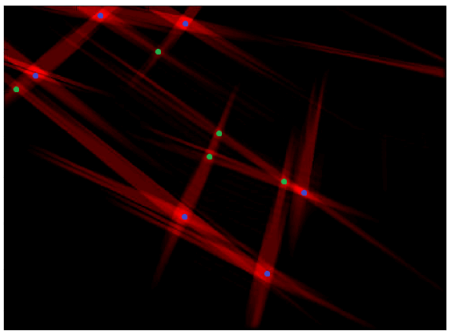
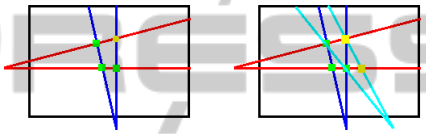


Figure 2: Illustration of a normalized synergy map obtained by using Equation 2. Brighter red colors indicate higher probabilities. The blue dots represent the ground truth that is the location of the people. The green dots represent the ghosts.



(a) (b)

Figure 3: Illustration of ghosting phenomenon in an arbitrary scenario. Here, three real objects are known as a priori. Black boundary represents a physical area. Red, blue and cyan are lines drawn from 3 cameras to the known objects. Green points are the accurate detections and yellow points are ghosts. (a) illustrates ghost generation using two cameras. (b) is generated using an additional camera with a limited Field of View (FOV).

4 BACK PROJECTED OCCUPANCY MAPPING

Object detection from the synergy map is possible by identifying the peaks. However, as shown in Figure 2, some false peaks are also present. A simple illustration in Figure 3 shows how the intersection of a line from the camera center to the object of interest generates a ghost detection, a concept discussed in detail in (Evans et al., 2012). Moreover, any errors in the camera calibration, time synchronization and background subtraction stage is likely to generation ghosts, a realistic scenario in the current systems. As we are focusing on localization therefore we further constraint ourselves not to use prior knowledge about the number of people or their location.

Our idea is simple and intuitive. We propose to verify if an object is present at each high probability location of the 2D synergy map by intersecting its back projected occupancy map, denoted by $Moc_{v,h}(x,y)$, and its corresponding foreground mask $F_v(x,y)$. Because a person can be occluded in one

view therefore the back projected occupancy map is computed across several views v. Moreover, to take into account the inaccurate foreground detections, the back projected occupancy map calculations are done with respect to a set H of planes including the ground plane and the planes parallel to the ground at different heights h.

The back projected occupancy map is calculated as follows. First, the high probability locations map (HPL) is computed by thresholding the 2D synergy map. Second, a segmentation step is applied to HPL to yield an image containing regions that may correspond to people. This segmented synergy map is back projected on each view v leading to several occupancy maps $Moc_{v,h}(x,y)$. Two back projected occupancy maps are illustrated in Figure 4. All the occupancy maps are studied to finally decide for each region of the segmented synergy map if it is a ghost or a person. These regions are further constrained to account for the size of minimum human projection in the synergy map, $Area_{TH}$. This threshold can be calculated a priori by the camera setups and calibration.

Moreover, as it is difficult to properly fix a threshold to binarize the synergy map, we use hysteresis thresholding and obtain two sets of regions: R_0 and R_s. R_0 is the set of regions obtained with zero or no threshold and R_s is the set of regions obtained at a higher threshold denoted by TH_s. Some quick operations are performed to clean up the map regions R_0: when the region bounding boxes overlap more than 60%, only the largest region is retained. The map regions R_s is often over-segmented due to the high threshold TH_s but the corresponding regions in R_0 are not affected by this. So, if multiple regions are detected as real objects in R_s, in a close spatial proximity, and if all of them correspond to a single region in the set R_0, then, only one detection is retained at the location of the region in R_0.

Now, we have to find a criterion to decide if a region r belonging to R_s corresponds to a real object or to a ghost. This criterion, similar to an occupancy rate, uses the generated back projected occupancy maps and is defined by:

$$O_{r,v,h} = \frac{\sum_{x\in X}\sum_{y\in Y} F_v(x,y) Moc_{v,h}(x,y)}{\sum_{x\in X}\sum_{y\in Y} Moc_{v,h}(x,y)} \qquad (3)$$

where $F_v(x,y)$ is the foreground mask in the image view v, X and Y are the coordinates of the pixels belonging to the projection of the region r in the view v and h is the assumed height of the synergy map according to the ground plane.

And the region r is detected as a real object if,

$$\forall v \in V,\ \exists h \in H, \qquad O_{r,v,h} > T(h) \qquad (4)$$

So, there is an assumed height of the synergy map such as the region r of the synergy map corresponds to binary detections in all of the views v.

The parameter $T(h)$ is slightly increased with the height, accounting for the assumption that if a projection is missing in one view for an assumed height then, there is a higher probability that it is also missing in the other heights, and if it is present then $T(h)$ must be higher as well. The specific combinations of heights, views and the values of the thresholds are discussed in the next section. The process is summarized in Algorithm 4.1.

Algorithm 4.1: Back Projected Occupancy Mapping.

clean up the map regions R_0
for $\forall\ h \in H$ **do**
  compute the occupancy maps for each view v, $Moc_{v,h}(x,y)$
  for $\forall\ r \in R_s$ **do**
    compute the criterion $O_{r,v,h}$ in each view
    if $\forall\ v,\ O_{r,v,h} > T(h)$ & $Area(r) > Area_{TH}$ **then**
      add region r to the detection list and remove region r from R_s
    end if
  end for
end for
remove over-segmentation using R_0

5 EXPERIMENTAL RESULTS

For the evaluation of our algorithm, we used a public sequence *City Center* of PETS 2009 dataset (PETS, 2009). For comparison purposes, the dataset, ground truth, and annotations are processed in the same fashion as described by (Utasi and Benedek, 2012). The evaluation sequence is an outdoor dataset containing 400 frames across three views with a large FOV (View_001, View_002, View_003). As shown in Figure 7, the Area of Interest (AOI) used is of the size $12.2m \times 14.9m$ and is selected on the basis of visibility from all the views. There are a maximum of 8 people present simultaneously in the scene with cases of occlusion and cluttering. It is also important to mention the inaccuracies in time synchronization and camera calibration. Even if they are not significant they cause noticeably different projections in the three views.

For evaluation, we use the metrics introduced by (Utasi and Benedek, 2012):

- The *Missed Detections Rate* (MDR) corresponds to the cases where no detection is assigned to the ground truth.

- The *False Detections Rate* (FDR) corresponds to the occurrences when a detection is assigned to an area without ground truth.
- The *Multiple Instances Rate* (MIR) corresponds to the situations where multiple detections match with the same ground truth area.
- The *Total Error Rate* (TER) is a combination of the last three measures defined by $TER = MDR + MIR + FDR$.

All of these evaluations are performed in the 2D image domain. These measures are expressed in percent of the number of all objects annotated in the dataset. The ground truth takes into account the inaccuracies due to time synchronization and calibration inaccuracies and is relaxed accordingly. Similarly, it is also adjusted for the ambiguities present at the borders of the AOI. Therefore, this popular sequence is challenging in many aspects.

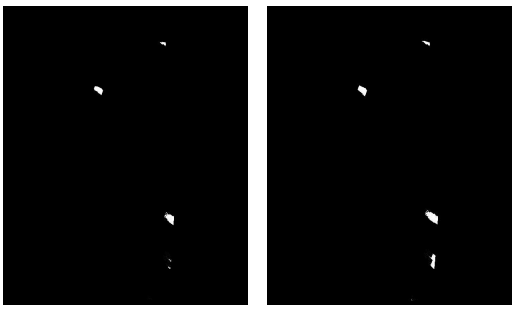
The synergy map is generated using all the planes, $p \in P$, between 100 to 150cm, at a distance of 1 cm each. We use this specific combination, because the torso region is present in all views without undergoing too many occlusions. Moreover, as the heights of people can vary and the system has to be robust, the top layers are not used. For back projection, we use two heights h: 0cm and 5cm. Figure 4 illustrates a case of varying information present across the two different heights. For PETS 2009, we find this selection of two heights as sufficient with respect to the pruning efficiency and computational performance of the algorithm, as the computation performance is negatively impacted with the increase in the number of heights. The synergy map is calculated from all three views. We decided not to use the *View*_003 for back projection step because (1) this view suffers from a significant perspective effect, and (2) the AOI in this view is represented by lower number of pixels than in the two other views. With high perspective effect, the resolution of projected area might not be sufficient for the threshold requirement of Equation 4.

The other parameters to set include B (threshold for background subtraction), TH_s (threshold to binarize the synergy map), $Area_{TH}$ (a threshold on the human size in the synergy map) and $T(h)$ used for the occupancy rate. For background subtraction, we use the default parameters of the algorithm suggested by (Yao and Odobez, 2007). The output is a probability map that is the likelihood of each pixel to belong to the foreground. This map is binarized using the threshold B. For comparison purposes, TH_s and B are optimized in a fashion similar to that of (Utasi and Benedek, 2012). For $T(h)$, the criteria is to assign a low value for the lowest height and gradually increase it for the subsequent height, as discussed in the previous section. The precision of this low value is not of paramount importance because the information is complemented with the other heights. Thus, we use $T(0) = 0.05$ and $T(5cm) = 0.1$. The parameter $Area_{TH}$ is set to 25. For the multiplanar projections (synergy map), a constant 2 cm grid resolution is used. The parameters B and TH_s are optimized for minimizing the TER. Figure 6(b) shows that the optimal value for TH_s is 0.8. This is a good compromise between a loss of detection and overdetection due to the first threshold set to 0. Figure 6(a) shows that the best foreground masks are obtained with $B = 0.1$. This result is confirmed by visually looking at the foreground images, with almost no foreground for $B > 0.4$.

Table 1: Ghost pruning performance of the proposed algorithm and comparison with the two color template matching algorithms for ghost pruning proposed by Ren (Ren et al., 2012). All parameters have been optimized so that the TER is minimized.

Method	TER	FDR	MDR	MIR
No Pruning	0.362	0.291	**0.047**	0.024
Pruning	**0.127**	**0.035**	0.092	**0.000**
Ren Eq. (6)	0.877	0.197	0.678	0.002
Ren Eq. (7)	0.885	0.202	0.681	0.002

For ghost pruning comparisons with (Ren et al., 2012), both of the color feature equations were implemented. Their algorithm is limited to two camera views so it was tested on *View*_001 and *View*_002; besides, any color comparison with *View*_003 will give significantly bad results due to the lack of color constancy. The same threshold (a ghost has at least three times greater joint likelihood than that of a real object) and three planes in the torso region as suggested by (Ren et al., 2012) are used. The planes used are: 100 cm, 124 cm and 149 cm. It was not possible to test our algorithm on their dataset as it is not public.



(a) (b)

Figure 4: Back projected occupancy maps for a frame of PETS 2009. (a) Generated at $T(0cm)$, (a) Generated at $T(5cm)$. These occupancy maps are obtained for *View*_001 and $TH_s = 0.8$. Notice the difference in the foreground masks at the two heights.

Table 2: Performance of the proposed ghost pruning algorithm compared to the full detection systems: POM (Fleuret et al., 2007) and Utasi (Utasi and Benedek, 2012). All parameters have been optimized so that the TER is minimized.

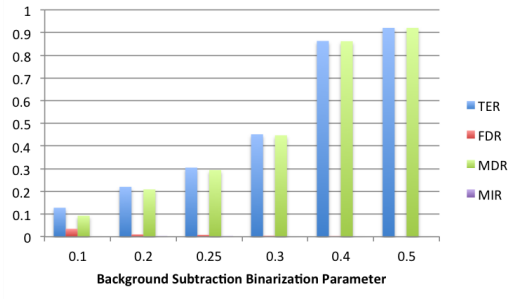
Method	TER	FDR	MDR	MIR
POM	0.205	0.150	**0.055**	**0.000**
Utasi	0.107	**0.014**	0.087	0.006
Proposed	0.127	0.035	0.092	**0.000**
Proposed + Boundary	**0.100**	0.021	0.079	**0.000**

Similarly, a comparison with (Evans et al., 2012) is not possible since they use tracking information.

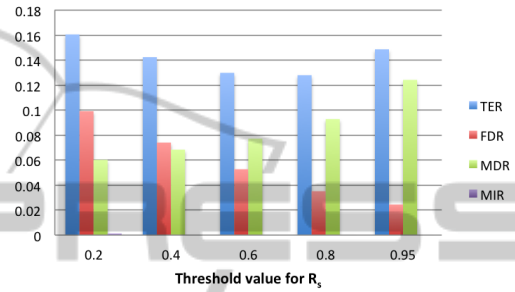
Table 1 shows the ghost pruning performance of our algorithm. There is a significant decrease in the number of false detections or the number of ghosts. The FDR improves by 25.6% and TER improves by 23.5% with a negligible increase in MDR. The proposed algorithm also performs significantly better than (Ren et al., 2012). We believe the failure of Ren algorithm is due to (a) the template matching that does not take into account the difference in the geometry of the scene and planar projections, more specifically in comparison to a less challenging Ren's dataset (b) the threshold selection criteria that is difficult to fix.

Next, we compared our proposed ghost pruning algorithm to the complete detection systems of (Utasi and Benedek, 2012) and (Fleuret et al., 2007) as in Table 2. Our results show significant improvement over (Fleuret et al., 2007) and a comparable performance to that of (Utasi and Benedek, 2012). The reason we don't surpass Utasi algorithm is the absence of a 3D model and its subsequent parameter optimization. Figure 5 shows a frame among many in which a detection is identified as false, also giving a missed detection, even though it is present on the boundary of the ground truth. Utasi designs the 3D model and ground truth for a detection to be precisely in the center of the person's body. If we relax the evaluation metric to include the boundary of the ground truth in the calculation of the different detection rates then, as

Figure 5: Illustration of a case in which a false detection is present on the boundary box. False detection is indicated by a cross (the estimated projected position) without a rectangle. The thin white rectangle represents the missed detections. Notice that the false detection also introduces a missed detection. Image is cropped for illustration purposes.



(a)



(b)

Figure 6: Curves between optimized algorithm parameters and the evaluation metrics where (a) Curve for the binarization threshold B, generated for a constant $TH_s = 0.8$ (b) Curve for the threshold TH_s, generated for a constant $B = 0.1$.

in the last row of Table 2, we can show that our algorithm, without any 3D model, can give, not only comparable, but identical and a little better performance to Utasi in terms of TER.

6 CONCLUSIONS

In this paper, we have presented a technique for ghost pruning in the context of people localization. First, multiplanar projections are generated to produce a synergy map. The synergy map is then complemented with the proposed back projected occupancy mapping technique. Experiments performed on a sequence of PETS 2009 dataset show significant reduction in the number of ghosts and improvement over other methods without the use of temporal information. The detection metrics also show results comparable to the state-of-the-art on complete detection schemes. We further propose to introduce our ghost pruning module in a complete detection system, for example with a 3D model or in a tracking system.

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

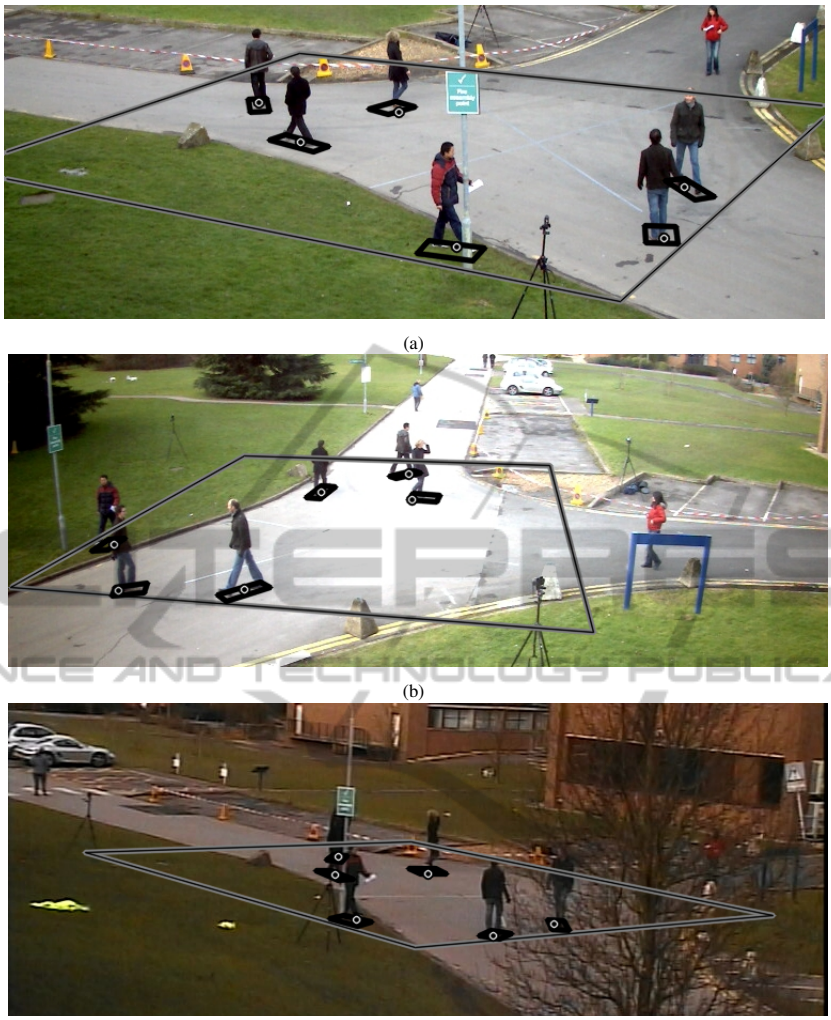


Figure 7: Detections in the three camera views of the *City Center* sequence of PETS 2009 dataset. (a) View_001 (b) View_002 (c) View_003. The black rectangle indicates a correct detection, the boundary of which is the ground truth and circle defines the estimated projected location. All people are correctly detected in all views of this frame including those in (c) despite occlusion and clutter. The AOI is illustrated by a black rectangle with gray outline. Images are cropped for illustration purposes.

SCITEPRESS
SCIENCE AND TECHNOLOGY PUBLICATIONS

REFERENCES

Dollar, P., Wojek, C., Schiele, B., and Perona, P. (2012). Pedestrian detection: An evaluation of the state of the art. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(4):743–761.

Eshel, R. and Moses, Y. (2010). Tracking in a dense crowd using multiple cameras. *Int. J. Comput. Vision*, 88(1):129–143.

Evans, M., Li, L., and Ferryman, J. M. (2012). Suppression of detection ghosts in homography based pedestrian detection. In *AVSS*, pages 31–36. IEEE Computer Society.

Fleuret, F., Berclaz, J., Lengagne, R., and Fua, P. (2007). P.: Multi-camera people tracking with a probabilistic occupancy map. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Khan, S. and Shah, M. (2009). Tracking multiple occluding people by localizing on multiple scene planes. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(3):505–519.

PETS (2009). Pets dataset: Performance evaluation of tracking and surveillance. http://www.cvg.rdg.ac.uk/PETS2009/a.html. [Online].

Ren, J., Xu, M., and Smith, J. S. (2012). Pruning phantom detections from multiview foreground intersection. In *ICIP*, pages 1025–1028.

Utasi, Á. and Benedek, C. (2011). A 3-D marked point pro-

cess model for multi-view people detection. In *Proceedings of The IEEE Conference on Computer Vision and Pattern Recognition*, pages 3385–3392, Colorado Springs, CO, USA.

Utasi, Á. and Benedek, C. (2012). A bayesian approach on people localization in multi-camera systems. *IEEE Transactions on Circuits and Systems for Video Technology*.

Yao, J. and Odobez, J. (2007). Multi-layer background subtraction based on color and texture. In *IEEE Conference on Computer Vision and Pattern Recognition, 2007. CVPR'07*, pages 1–8.

Yilmaz, A., Javed, O., and Shah, M. (2006). Object tracking: A survey. *ACM Comput. Surv.*, 38(4).

SCITEPRESS
SCIENCE AND TECHNOLOGY PUBLICATIONS