Exploiting 3D Geometric Primitives for Multicamera Pedestrian Detection

Muhammad Owais Mehmood[1], Sebastien Ambellouis[1] and Catherine Achard[2]
[1]LEOST, IFSTTAR, Villeneuve D'Ascq, France
[2]Sorbonne Universites, UPMC Univ Paris 06 and CNRS UMR 7222, ISIR, F-75005, Paris, France
{owais.mehmood, sebastien.ambellouis}@ifsttar.fr, catherine.achard@isir.upmc.fr

Abstract

In this paper, we present an approach for multicamera pedestrian detection exploiting the concepts of multiview geometry and the shapes of 3D geometric primitives. Multicamera occupancy maps provide peak responses corresponding to the object detection but suffer from several false detections known as ghosts. The novelty of this paper is the introduction of shape patterns which can model the objects, such as pedestrians, by defining a kernel function in the projected occupancy space. This kernel depends upon the geometry of the 3D primitives and also varies in relation to their position with respect to the cameras in the real world configuration. For multiple objects visible across several cameras, we define a formation model which is the convolution of this spatially varying kernel with the set of possible object locations. The locations corresponding to detections can thus be obtained through a deconvolution process. For efficient computations, we further propose an estimated deconvolution process specific to our kernel responses which can also be heavily parallelized. We show the application of this process towards pedestrian detection by studying various 3D cylindrical primitives. Experiments on two public dataset sequences, including comparison with another approach, show the efficiency of the proposed method in terms of pedestrian detection and ghost pruning, including in adverse and challenging conditions.

1. Introduction

Intelligent automated visual surveillance is an active area of research in signal, image processing and computer vision communities. Pedestrian detection is a well-studied issue which also finds several applications in the domain of visual surveillance. Extensive research has been done in the area of monocular pedestrian detection. However, these methods remain restricted in the handling of occlusion, clutter, scale, people density [3].

Research community has focused on using multicamera systems for improvements in pedestrian detection and thus the visual surveillance. The ubiquitous presence of cameras with the increase in computational resources has also fueled the development of multiview research. Sensor fusion, multiview visual analysis are some of the challenges faced in this area.

Multiview object detection can be achieve with the aid of proper image registration of views covered by various cameras present in the scene. Multiview geometric techniques such camera homographies and the use of camera calibration have been employed to project cameras to a common search space such as the ground plane. Khan and Shah [7] use the camera homography constraint to generate occupancy maps which is the fusion of multiple scene planes. Eshel and Moses [4] use multiplanar projections for top-view camera topologies in order to perform head detection and eventually the pedestrian detection. Probabilistic methodologies have also been utilized for multiview detection. Probabilistic occupancy map method models the pedestrian with a rectangle of average human height placed on a discretized ground plane [6].

All the methods presented so far suffer from a high false detection rate. Besides focusing on the rather complete detection or tracking systems, this phenomenon of false detections has been studied in the literature as the ghost pruning problem. False detections occurring due to the intersection of non-corresponding regions, based on the camera and object positions, are referred to as ghosts in the literature (see Fig. 1). An approach based on color matching across the camera views has been proposed in [12]. Evans et al. compute the probability of a ghost detection based on a spatiotemporal relationship of the objects present in the scene with the camera positions [5]. Mehmood et al. [8] define a background occupancy map to gather image evidence across all camera views in order to remove the ghosts.

We present a novel approach for pedestrian detection using multicamera occupancy maps and by modeling the object shapes as 3D geometric primitives (see Fig. 1). We define a spatially varying kernel which depends upon the shape and geometric characteristics of the primitives and the camera calibration. We propose an analytical formation model for object detection by performing convolution of the proposed kernel with the object location map. Our spatially varying kernel is able to perform suppression of false detection through multiview reasoning. This specific kernel allows us to define sharp peak responses corresponding to the object detections. These detections can be localized by a deconvolution process. We also propose an efficient approximative deconvolution using a modified version of watershed transform specific to our kernel.

The proposed algorithm is able to account for challenging situations such as lighting, color, weather variations, and is able to robustly localize the pedestrians handling occlusion and projective shadows from a high density crowd. The proposed algorithm is not limited to a specific camera topology [4, 8], requires no temporal information [5], performs multicamera reasoning rather than the concatenation of monocular primitive detections [2], and has lower number of parameters with no optimization requirements as in [14, 15]. The quantitative analysis shows the efficiency of our approach on two public dataset sequences. We also provide time complexity analysis of the algorithm and propose a multicore implementation for obtaining optimal runtime efficiency.

The rest of the paper is organized as follows. The proposed algorithm is presented in Section 2. We present the evaluation methodology, quantitative comparison and analysis of our approach in Section 3. Finally, the paper concludes in Section 4.

2. Primitive Detection in Multicamera Occupancy Maps

Occupancy maps are well known in the multicamera context to exhibit peak responses corresponding to the object locations in the scene [7]. Multicamera occupancy maps assign a probability that is based on the normalized sum of the image evidence, binary or probabilistic, gathered from all the cameras and projected to a common search space such as the ground plane. Fig. 1(c) shows an example of occupancy map generated using two camera views and multiplanar projections parallel to the ground. Multicamera occupancy maps also exhibit artifacts depending upon the relative position between the camera and the objects [5]. These artifacts are not due to the noise or a random phenomenon, they are explained by multiview geometry reasoning and can even be predicted. Therefore, this work applies the concepts of multiview geometric reasoning in order to overcome such artifacts.

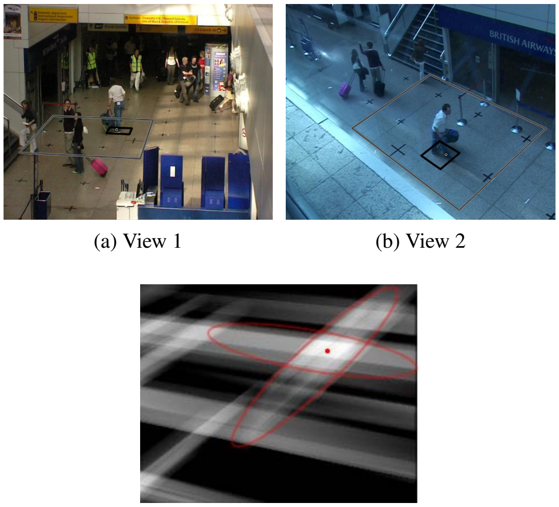
2.1. Primitive Formation

Let $\mathbf{O}(X,Y)$ be an occupancy map that defines the probability of the presence of an object at each location (X,Y) on the ground plane $Z=0$. The strong distinct peak responses in the occupancy map can be modeled as a sum of 2D dirac delta functions

$$\mathbf{D}(X,Y)=\sum_{i=1}^{n}\delta(X-X_i,Y-Y_i). \qquad (1)$$

where (X_i,Y_i) are the coordinates of the n objets. The goal of this work is to localize a set of detections $\mathcal{D}=\{(X_1,Y_1),(X_2,Y_2),\ldots,(X_n,Y_n)\}$ using a response similar to that of the delta function in the occupancy maps. As presented earlier, this work applies the concepts of multiview geometric reasoning to overcome the artifacts that are not due to noise or to a random phenomenon and can be perfectly explained by a multi-view analysis. In this context, if we model each object as a 3D geometric primitive then it is possible to formulate an analytical shape approximation in the corresponding occupancy map. We propose a 2D kernel to achieve this, and the object localization can be considered as a deconvolution process.

We assume an arbitrary primitive shape of diameter $\varnothing$ visible by two cameras (see Fig. 2). For convenience, let us study the primitives in the cylindrical coordinate system (ρ,φ,Z) originating at the ground location of the camera $(L_{c_X},L_{c_Y},0)$. The vertical cross sectional view of this cylinder will be a polygon comparable to a rectangle of


(a) View 1

(b) View 2

(c) Multicamera Occupancy Map

Figure 1. Illustration of the multicamera occupancy map and primitive based detection using two views of the PETS 2007 dataset. The occupancy map contains an ‘X’-shape pattern corresponding to the person and it’s visibility in two two cameras at different heights. This shape pattern is modeled using geometric primitives. Application of a threshold to the occupancy map will create several false detections referred to as ghosts. The rectangular bounding box represents the Area of Interest (AOI). Difficulties such as those arising from the errors in the background subtraction process, and the projection/presence of people outside AOI in the occupancy are present in this figure.

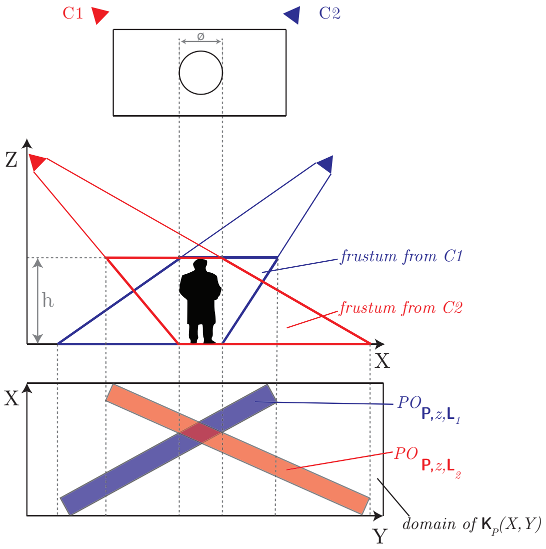

C1
C2
ø
Z
h
frustum from C1
frustum from C2
X
X
$PO_{\mathbf{P},z,\mathbf{L}_1}$
$PO_{\mathbf{P},z,\mathbf{L}_2}$
Y
domain of $\mathsf{K}_p(X,Y)$

Figure 2. Illustration of the projected kernel profiles. Primitive placed in the scene, viewed by two cameras, and as visible from: (Top) the top view in the real world, (Middle) the cross sectional view in the projected kernel domain, (Bottom) the top view in the projected kernel domain. fr is a frustum obtained in the cross sectional view of the primitive, defined in the cylindrical coordinate system for convenience. po is a polygon obtained in the top view formation of the primitive, defined in the rectilinear coordinate system.

height h and width $\varnothing$ while the primitive's projected cross sectional view will form a collection of frustums bounded by the intersection of the camera rays to the ground and the elevated parallel planes [2]. The projected profile of the kernel is the combination of these cross sectional frustums, along the parallel planar heights, $Z = 0$ and $Z = h$, or the combination of polygons as viewed from the top

$$\mathbf{K}(X, Y; \mathbf{P}, \mathbf{L}_C, Z) = \frac{1}{|C|} \sum_c \sum_z po(\mathbf{P}, \mathbf{L}_c, z). \qquad (2)$$

where $\mathbf{L}_c = (L_{c_X}, L_{c_Y}, L_{c_Z})$ represents the location of the camera $c \in C$ and $|C|$ is the cardinality. $\mathbf{P}$ is the set of points belonging to a specific geometric primitive. This spatially varying asymmetric kernel defines a local spread in the occupancy map that depends upon the shape, size and position of the primitive, the projection heights, all in relation to the position of the camera.

The geometric primitives can include cylinders, cubes, pyramids or other shapes depending upon the object to be detected. For this particular work focusing on pedestrians, primitives are selected as cylinders. Pedestrians occupy a combinatorial rectangular profile in the multicamera based kernel profile, and occupancy maps (see Fig. 3). The particular of selection of top view introduces robustness against the requirement of modeling using multiple cross sectional views.

2.2. Model Formation and Matching

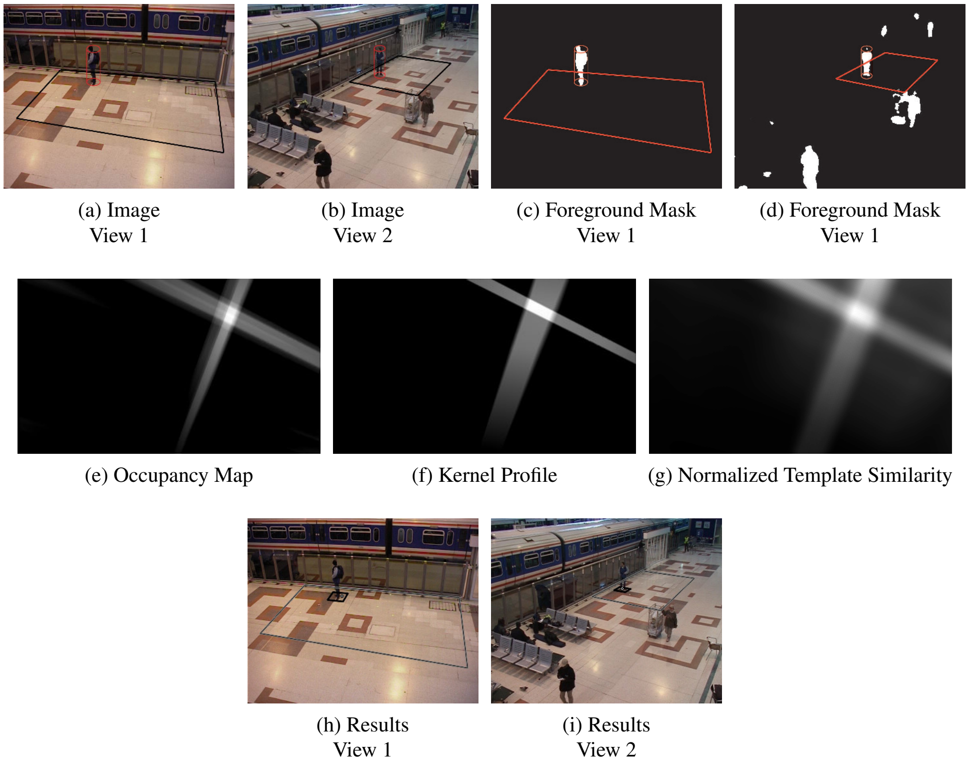
The corresponding occupancy map specific to the kernel $\mathbf{K}$ is a convolution of its spatially varying response with the corresponding set of object locations $\mathbf{D}_\mathbf{K}(X, Y)$ in addition to the noise ϵ observed due to the presence of multiple objects

$$\mathbf{O}(X, Y) = \mathbf{D}_\mathbf{K}(X, Y) * \mathbf{K}(X, Y; \mathbf{P}, \mathbf{L}_C, Z) + \epsilon. \qquad (3)$$

Pedestrian detection can now be achieved by the deconvolution of $\mathbf{O}$ followed by a peak extraction process. However, deconvolution with multiple kernels is a computationally expensive step and sensitive to noise such as that from the background subtraction process, imprecisions of camera calibration, time synchronization errors. Assuming that the scene is not of overly dense crowds, template similarity can be utilized as an estimated deconvolution

$$\hat{\mathbf{D}}_\mathbf{K}(X, Y) = \frac{1}{\|\mathbf{K}(X, Y)\|_{max}} \sum_X \sum_Y min(\mathbf{K}(X, Y), \mathbf{O}(X, Y)). \qquad (4)$$

where $\|\mathbf{K}(X, Y)\|_{max}$ is the max-norm [13]. $\hat{\mathbf{D}}_\mathbf{K}(X, Y)$ searches for any evidence of local matching and proceeds further by normalizing it with respect to the global kernel space for a better context in terms of the difference between $\mathbf{K}(X, Y)$ and $\mathbf{O}(X, Y)$. There exists a trade off-between the detection accuracy and the computational processing defined by the number of samples over which the similarity is computed. Higher number of samples will produce a sharper response at the cost of time required. However, we use a multi-core implementation in order to compute the template similarity score between the occupancy map and the kernel profiles.


(a) Image View 1

(b) Image View 2

(c) Foreground Mask View 1

(d) Foreground Mask View 1

(e) Occupancy Map

(f) Kernel Profile

(g) Normalized Template Similarity

(h) Results View 1

(i) Results View 2

Figure 3. Illustration of the various stages of the proposed algorithm. (Top) Two views of PETS 2006 dataset, and their corresponding foreground masks. The boundaries represent the Area of Interest (AOI). The pedestrian is modeled by a 3D cylinder. (Middle) The occupancy map obtained from the foreground masks, the kernel profile of the cylinder, and the corresponding estimated deconvolution. (Bottom) The results obtained. The bounding box around the person represents the ground truth, and the circle marks the estimated detection. The proposed analytical model induces a maximum response for the object center and the estimated detection is the result of this maxima selection in the estimated deconvolution.

2.3. Watershed Based Maxima Selection

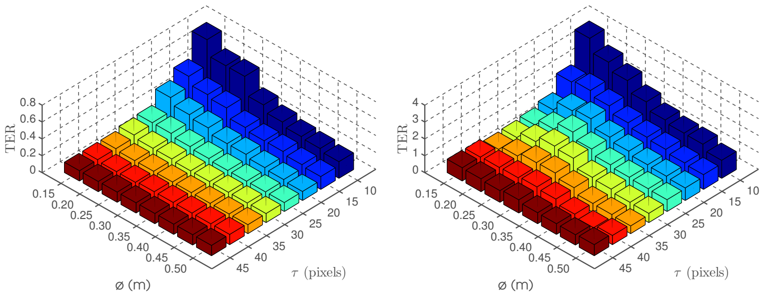
The template similarity does not resemble the required combination of dirac delta responses (see Fig. 3(g)). The result is a distribution consisting of several modes for which it is necessary to estimate the maxima to obtain the number of objects or pedestrians, and their locations on the ground plane. For this purpose, watershed based maxima selection algorithm is applied. This algorithm performs the extraction of the centralized location in the overlapping areas obtained in the kernel domain (see Fig. 2). Local maxima are extracted in 8×8 pixel blocks. Watershed transform with markers [1] is applied to these local maxima such that their topological prominence is greater than the tolerance threshold τ. It computes the geometric centre if several local maxima fulfill the criteria [9]. τ gauges the closeness of the two detections, and intuitively decides the size of the overlapping areas in the kernel domain (see Fig. 2).

3. Experiments

We have compared our approach to the MSPL algorithm [7]. In the MSPL approach, multiple scene planes are selected for occupancy map generation. For detection, we apply watershed based maxima selection to a similar selection of planar heights Z. Binary foreground masks are generated using the default parameters of the publicly available implementation of multi-layer background subtraction method [16]. The occupancy map $\mathbf{O}(X, Y)$ is calculated by averaging the foreground scores across all camera views (see Fig. 3(e)).

We use two public datasets: PETS 2006 [10] and PETS 2007 [11]. For PETS 2006, we selected the cameras 3 and 4 of the S1 sequence. PETS 2006 is recorded in an indoor environment. For PETS 2007, we selected the cameras 2 and 3 of the S8 sequence. This scene is a combination of both indoor and outdoor scenarios, including sunlight variations across the sequence. PETS 2007 has a higher density of pedestrians compared to PETS 2006. For both sequences, the AOI is defined such that it is visible from all cameras. The cameras above are selected to cover these AOI from relatively varying positions and topologies.

We annotated a total of 159 frames for PETS 2006 and 120 frames for PETS 2007. We use the camera calibration data which is available for both the datasets. If absent, cam-



(a) PETS 2006

(b) PETS 2007

Figure 4. Evaluation of the proposed method with different parameter settings. Total Error Rate (TER) as a function of the τ and $\varnothing$ parameters for: (a) PETS 2006 and, (b) PETS 2007 dataset.

Name	Valid Range	Description
FDR	$[0, \infty)$	False Detections Rate: ratio of estimates not corresponding to ground truth to the total number of objects in groundtruth
MDR	$[0, 1]$	Missed Detections Rate: ratio of ground truths not found to the total number of objects in groundtruth
MIR	$[0, 1]$	Multiple Instances Rate: ratio of multiple estimates assigned to a ground truth to the total number of objects in groundtruth
TER	$[0, \infty)$	TER=FDR+MDR+MIR

Table 1. Summary of the evaluation measures used. Maximum matching is applied across all camera views as defined in [15].

era homography can be used [7]. The tools and criteria as defined in [15] are used for annotation purposes[1]. We notice presence of errors in the foreground masks (see Fig. 3(d)), errors of camera calibration (see the slight variation of the estimated position in Fig. 3(h) and Fig. 3(i)), and significant effect of the projections of pedestrian outside our AOIs (see Fig. 1(c)). The evaluation measures used are projected position error metrics as defined in [15] and summarized in Table 1.

The cylindrical primitives are 1.75m high [6, 14, 15]. The planar heights used are 211 planes, one plane for each centimeter between 0–2.1m, covering the range of possible human heights. We evaluate the algorithm for two parameters: diameter of the cylinder $\varnothing$ and tolerance threshold τ.

Quantitative Comparison: The evaluation results are presented in Table 2. If we consider TER, it can be observed that we obtain improvements over MSPL both in PETS 2006 and PETS 2007. For the proposed MGP algorithm, Fig. 4 shows TER plotted as a function of the τ and $\varnothing$ parameters for both datasets. $\varnothing$ performs 3D reasoning whereas τ performs local analysis in the kernel space. The $\varnothing$ parameter fits to the specific radius of the pedestrians in the dataset, and our values resemble those in [2]. We can observe that our novel application of watershed based maxima selection and the related selection of τ eliminates MIR for both algorithms and with the two datasets.

The proposed algorithm is influenced by the height of the cameras. If we study the two extreme cases: (a) camera at extreme low height provides imprecise detection, precise height estimation, (b) camera at extreme top provides precise detection, imprecise height estimation. Thus, the results can further be improved with increased height of the cameras, such as PETS 2007 View 1 (see Fig. 5(c)), or introduction of another camera with more height. Fig. 5 shows examples of the results obtained.

Runtime: The spatially varying kernel and formation model exhibits negligible linear dependence to the number of camera views and image resolution. The template similarity module has a linear dependency on the image resolution. The maxima selection stage has a constant runtime. The proposed algorithm is scalable and runs efficiently employing a multi-core implementation.

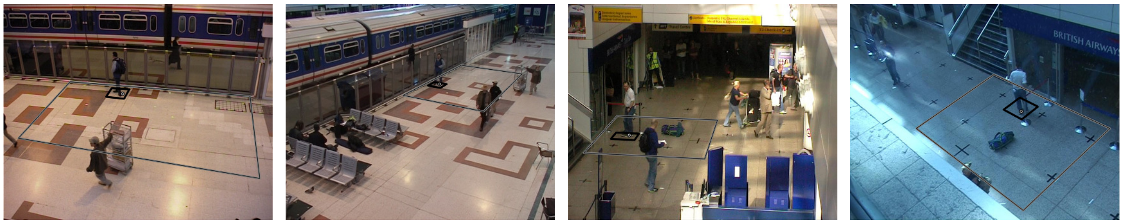
4. Conclusions

In this paper, we have proposed an efficient approach for performing pedestrian detection using multiview reasoning in the multicamera occupancy maps. These maps exhibit the problem of ghosts. We propose a spatially varying kernel in the projective space which analyzes the shape patterns

Sequence	Method	TER	FDR	MDR	MIR
PETS 2006	MGP	0.10	0.00	0.10	0.00
	MSPL [7]	0.28	0.18	0.10	0.00
PETS 2007	MGP	0.36	0.08	0.28	0.00
	MSPL [7]	1.08	0.89	0.19	0.00

Table 2. Comparison of the proposed Multicamera Geometric Primitives (MGP) method with the Multiple Scene Planes Localization (MSPL) method [7]. The parameter set is such that the TER is minimized. For MSPL: $\tau = 35$. For MGP and PETS 2006: $\varnothing = 0.50, \tau = 35$, and PETS 2007: $\varnothing = 0.40, \tau = 40$.

[1] The sequences, with their foreground masks, ground truth annotation and calibration data are available at http://www.owaismehmood.com/codes.



(a) PETS 2006 View 1 (b) PETS 2006 View 2 (c) PETS 2007 View 1 (d) PETS 2007 View 2

Figure 5. Examples of the estimated pedestrian locations in the PETS 2006 and 2007 datasets. The algorithm correctly distinguishes between the pedestrians and bag, trolley, ambiguities of presence in Area of Interest (AOI) across views. The algorithm is also able to handle indoor, outdoor situations, variations of intensity such as sunlight vs interior lighting, and the projections of pedestrians outside the user-defined AOI.

in the occupancy map. This kernel depends upon the properties of an assumed 3D geometric primitive and the camera parameters. Moreover, it allows us to propose an analytical formation model, the deconvolution of which provides the object locations. We further introduce a novel parallelized estimated deconvolution approach specific to our kernel responses. We show that our approach is able to recover the pedestrian locations in two different datasets despite the challenging conditions. For future work, we plan to explore other geometric primitives such as studying the 3D cuboids for vehicle detection.

5. Acknowledgements

The research leading to these results has received funding from the French FUI framework DeGIV.

References

[1] S. Beucher and F. Meyer. The morphological approach to segmentation: the watershed transformation. Mathematical morphology in image processing. *Optical Engineering*, 34:433–481, 1993. 4

[2] P. Carr, Y. Sheikh, and I. Matthews. Monocular object detection using 3d geometric primitives. In A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, editors, *Computer Vision ECCV 2012*, volume 7572 of *Lecture Notes in Computer Science*, pages 864–878. Springer Berlin Heidelberg, 2012. 2, 3, 5

[3] P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: An evaluation of the state of the art. *PAMI*, 34, 2012. 1

[4] R. Eshel and Y. Moses. Tracking in a dense crowd using multiple cameras. *International Journal of Computer Vision*, 88(1):129–143, 2010. 1, 2

[5] M. Evans, L. Li, and J. Ferryman. Suppression of detection ghosts in homography based pedestrian detection. In *Advanced Video and Signal-Based Surveillance (AVSS), 2012 IEEE*, pages 31–36, Sept 2012. 1, 2

[6] F. Fleuret, J. Berclaz, R. Lengagne, and P. Fua. Multicamera people tracking with a probabilistic occupancy map. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(2):267–282, Feb 2008. 1, 5

[7] S. Khan and M. Shah. Tracking multiple occluding people by localizing on multiple scene planes. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(3):505–519, March 2009. 1, 2, 4, 5

[8] M. O. Mehmood, S. Ambellouis, and C. Achard. Ghost pruning for people localization in overlapping multicamera systems. In *VISAPP (2)*, pages 632–639, 2014. 1, 2

[9] M. O. Mehmood, S. Ambellouis, and C. Achard. Launch these Manhunts! Shaping the Synergy Maps for Multi-Camera Detection. In *VISAPP, International Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, page 8p, Berlin, Mar. 2015. 4

[10] PETS. Performance evaluation of tracking and surveillance dataset 2006. http://www.cvg.reading.ac.uk/PETS2006/data.html, 2006. [Online]. 4

[11] PETS. Performance evaluation of tracking and surveillance dataset 2007. http://www.cvg.reading.ac.uk/PETS2007/data.html, 2007. [Online]. 4

[12] J. Ren, M. Xu, and J. Smith. Pruning phantom detections from multiview foreground intersection. In *Image Processing (ICIP), 2012 19th IEEE International Conference on*, pages 1025–1028, Sept 2012. 1

[13] N. Srebro and A. Shraibman. Rank, trace-norm and max-norm. In P. Auer and R. Meir, editors, *Learning Theory*, volume 3559 of *Lecture Notes in Computer Science*, pages 545–560. Springer Berlin Heidelberg, 2005. 3

[14] A. Utasi and C. Benedek. A 3-d marked point process model for multi-view people detection. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3385–3392, June 2011. 2, 5

[15] A. Utasi and C. Benedek. A bayesian approach on people localization in multicamera systems. *Circuits and Systems for Video Technology, IEEE Transactions on*, 23(1):105–115, Jan 2013. 2, 5

[16] J. Yao and J. Odobez. Multi-layer background subtraction based on color and texture. In *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pages 1–8, June 2007. 4