Multi-camera Pedestrian Detection with a Multi-view Bayesian Network Model

Peixi Peng[1]
pxpeng@jdl.ac.cn

Yonghong Tian[1*]
yhtian@pku.edu.cn

Yaowei Wang[2]
ywwang@jdl.ac.cn

Tiejun Huang[1]
tjhuang@pku.edu.cn

[1] National Engineering Laboratory for Video Technology, School of EE & CS, Peking University, China (* Corresponding Author)

[2] Department of Electronic Engineering, Beijing Institute of Technology, China

Abstract

In this paper, we propose a novel method with the multi-view Bayesian network (MBN) model to detect pedestrians from multi-camera surveillance videos. In our method, the ground plane is discretized in a predefined set of locations and our aim is to estimate the occupancy probability of each location that can be then used to predict the occurrence of pedestrians. To reduce the possible phantoms, we use MBN to model the potential occlusion relationship of all locations in all views, and the "subjective supposing" node states (SSNS) as a set of Boolean parameters of MBN to denote whether a pedestrian occurs at the corresponding location. Thus a learning algorithm is proposed to estimate the SSNS parameters, by finding such a configuration that the final occupancy possibility can best explain the image observations (i.e., foreground masks) from different views. The experimental results on the APIDIS and PETS09 S2L1 benchmark datasets show that our method can obtain at least 10% performance gain compared with several state-of-the-art algorithms.

1 Introduction

In recent years, more and more cameras are widely deployed for video surveillance in a cooperative manner. In such scenarios, multiple-pedestrian detection has become an essential technology for many applications such as crowd behaviour analysis. Often, occlusions among pedestrians will complicate the detection process and make it difficult for the system to accurately detect the pedestrians after heavy occlusion. In this sense, the availability of multi-view information will make pedestrian detection easier and more accurate.

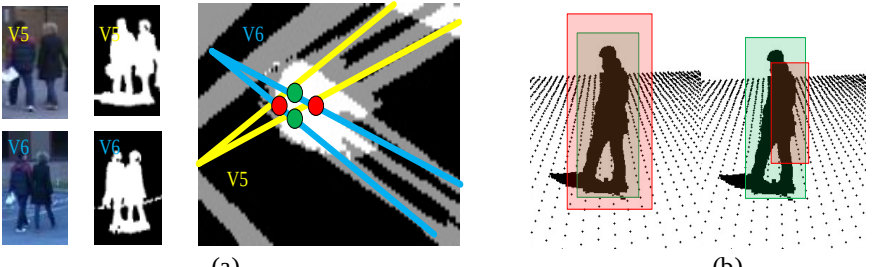
To exploit multi-view information for pedestrian detection in multi-camera surveillance video, we should estimate the occupancy possibility of each location in each view and integrate the possibilities of all views together to obtain the final occupancy possibility on the ground plane, which could then be used to predict the occurrence of a pedestrian in this location. Towards this end, Sankaranarayanan et al. [2] presented a general framework to utilize the geometric constraints (e.g., intersecting the view lines from multi-cameras on the ground plane) for object detection and tracking. In practice, their approach can yield satisfying results when people are well-separated in multiple views. However, if there are heavy occlusions among people, it will generate many "phantom" phenomena in the multi-

http://dx.doi.org/10.5244/C.26.69

view projection due to the intersections of viewing rays at locations that are not occupied by any pedestrian (as shown in Figure 1(a)) [6].

To address this problem, we first classify the phantoms in a single view into two categories. The *first-class phantoms* are those who occlude some pedestrians (e.g., the left one in Figure 1(b)). Often, the phantoms of this kind are generated due to the projection of inaccurate foreground extraction results on the ground plane. In this case, if these phantoms are directly treated as detection results, the matching degree with the foreground masks should be much less than the pedestrians which are occluded by them. In order to reduce the first-class phantoms in the multi-view projection, the key point is to make the detection results best match the foreground masks using the occlusion relationship among phantoms and pedestrians. On the other hand, the *second-class phantoms* denote those that are occluded by pedestrians, despite they can also match the foreground masks well (e.g., the right one in Figure 1(b)). The reason for generating the phantoms of this kind is usually due to the non-invertible mapping from 3D world coordinates to 2D image coordinates. These phantoms always be occluded by pedestrians mostly. Thus to reduce the second-class phantoms, we need to estimate the non-occluded parts for each phantom.

By summarizing the two cases above, we can conclude that the key problem to reduce the possible phantoms in the multi-view projection is to effectively model and utilize the occlusion relationship among potential pedestrians at different locations in all views. It is notable that a 1st phantom in one view may be the 2nd phantom in another view and vice versa. Considering it, a multi-view Bayesian network (MBN) is proposed in this paper. In general, a MBN is constructed with the locations on the ground plane and several single Bayesian networks (SBNs), where each SBN is used to characterize the potential occlusion relationship of all locations in a single view, while the locations on the ground plane is used to establish the correspondence among all SBNs through the geometric constraints among cameras (See Figure 2). Moreover, we also model the "subjective supposing" node states (SSNS) as a set of Boolean parameters of MBN, which are then used to denote whether a pedestrian occurs at the locations. To calculate the occupancy possibility of a location, we can estimate the part of the pedestrian at this location which are not occluded by other pedestrians using SSNS. A learning algorithm is then proposed to estimate the SSNS parameters of the MBN, by finding such a configuration that the final occupancy possibility can best explain the image observations (i.e., foreground masks) from different views. The overall framework of our method is shown in Figure 2.



(a) (b)

Figure 1: (a) An example of phantoms: locations (red circles) are occupied by phantoms and locations (green circles) are occupied by pedestrians; (b) The first-class (left) and second-class (right) phantoms, where the red ones denote phantoms.

In our method, two classes of phantoms can be effectively handled: For a first-class phantom, the real locations occluded by it will have a small occupancy possibility during the inference of MBN, consequently making the final occupancy possibility not explain the

foreground mask well. While for a second-class phantom, since its most parts are occluded by real locations (which can be judged through SSNS), its final occupancy possibility will also be small during the inference of MBN. In both cases, the phantom will not be treated as a pedestrian after the SSNS learning process.

We evaluate our method on two challenging benchmark datasets for multi-view pedestrian detection, namely, the APIDIS and PETS09 S2L1 datasets. The experimental results show that our method can obtain at least 10% performance improvement compared with several state-of-the-art algorithms.

Our contributions can be summarised as follows:

1. A MBN model, together with SSNS, is proposed to characterize the occlusion relationship among pedestrians at different locations in all views. Using this model, two classes of phantoms can be effectively processed.

2. A learning algorithm is developed to estimate the SSNS parameters by finding such a configuration that the final occupancy possibility can best explain the image observations (i.e., foreground masks) from different views.

The rest of the paper is organized as follows: Some related works are simply reviewed in section 2. Section 3 presents the MBN model. Section 4 describes the learning method for MBN. Section 5 shows our experiment results in PETS2009 S2L1 and APIDIS datasets. Finally, Section 6 concludes the paper.

2 Related work

Pedestrian detection plays an important role in many computer vision applications. In last decades, monocular approaches for detection and tracking have made a great progress. An extensive review of state-of-the-art single camera methods has been given in [1]. However, monocular approaches often suffer poor performance in the occlusion scenes, mainly due to insufficient information available from a single camera. In such scenes, multi-views approaches have a great advantage, since some geometric constraints of real world, such as the presence of the ground plane and the mapping between the 3D real-world with camera views, can provide some useful information for multi-view approaches [2].

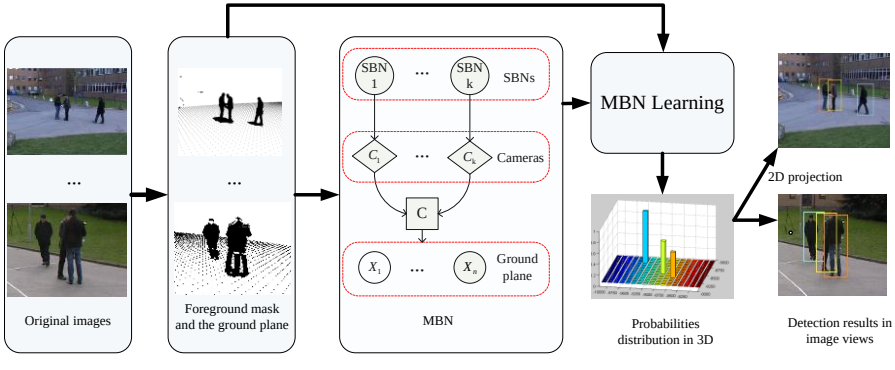


Figure 2. The framework of our method

In [3], Eshel and Moses placed the cameras at a high elevation and synthetized head information through multi-view fusion to avoid occlusion. It was capable of tracking up to

twenty people walking in a small area. However, their method seems inadequate for handling the field views near a low elevation. Using the holography map, Khan and Shah [4], and Arsić et al. [5] utilized multi-view foreground fusion to localize pedestrians on multiple parallel planes. Among them, Khan and Shah [4] removed some phantoms by some prior geometry assumption, while Arsić et al. [5] utilized temporary information to omit the phantoms. Nevertheless, the two methods are both sensitive to some foreground information in some cases. For example, if the foreground information doesn't include pedestrians' feet, pedestrian localization will be failed.

The MPP model [7], which was originally developed to detect and count crowds in a single view, has been promoted from 2D individual camera views to 3D real-world space [6][8]. Among them, Ge and Collins [6] handled phantoms through "sliding" the detection results in depth along a viewing ray. Fleuret. et al. proposed a probabilistic occupancy map to detect target from multi-view video in [9][10]. Alahi [11] formulated multi-view pedestrian detection as an inverse problem of deducing an occupancy vector from the noisy binary silhouettes observed as foreground pixels in each camera. These methods both aim to use the different optimization models to best explain the foreground. Compared with them, our method focuses on modelling and utilizing the potential occlusion relationship in the optimization process. This makes our model discriminate the phantoms and the pedestrians more effectively.

In [14], Mittal and Davis modelled the appearance (colour) and locations of pedestrians for segmentation from multiple camera views. These features are indeed helpful for the separation of foreground regions belonging to different objects. The similar assumption is also used in this paper. However, different with [14], our method does not utilize any prior knowledge such as the number of pedestrians and the occlusion relationships among pedestrians from temporary information. This makes our method generalizable to a wide range of surveillance applications where the number of pedestrians is unknown in advance.

3 The MBN Model

Our system focuses on the surveillance video in which multiple calibrated and fixed cameras monitor the same area simultaneously. Specifically, the monitored area is divided as a grid of n locations. For a given location i on the ground plane, we generate a rectangle of roughly the same size and aspect ratio as the motion blob that a pedestrian standing at location i in view k (denoted by r_i^k). Our goal here is to estimate the occupancy probability of each location, which can then be used to predict the occurrence of a pedestrian at this location. Before modelling, we use the RSS projection method proposed in [11] to omit some locations where are occupied by pedestrians impossibly. It is a useful pre-processing step to reduce much computation cost, though this method will also choose some locations which are actually occupied by pedestrians or phantoms. Some symbols in our method are listed in Table 1.

$x = \{x_1, \ldots, x_n\}$	The locations on the ground plane
n	The number of locations on the ground plane
X_i	The Boolean random variable standing for the presence of an pedestrian at location i.
r_i^k	The rectangle corresponding to location i in camera k
R_i^k	The Boolean random variable standing for the presence of an pedestrian at r_i^k from view k

$an_k(R_i^k)$	The set of ancestor nodes of R_i^k in SBN k.
$C = \{C_1, \dots, C_K\}$	The set of cameras used in our system
K	The number of cameras in our system
$\delta = \{\delta_i, \dots, \delta_n\}$	The set of SSNS for all locations
$D = \{D_1, \dots, D_k\}$	The set of images from background subtraction in all cameras
$\|D_k\|$	The number of pixels in D_k.
$(w,h)_k$	The pixel whose image coordinate is (w, h) in camera k.

Table 1. Some symbols used in our method

3.1 Bayesian network in a single camera (SBN)

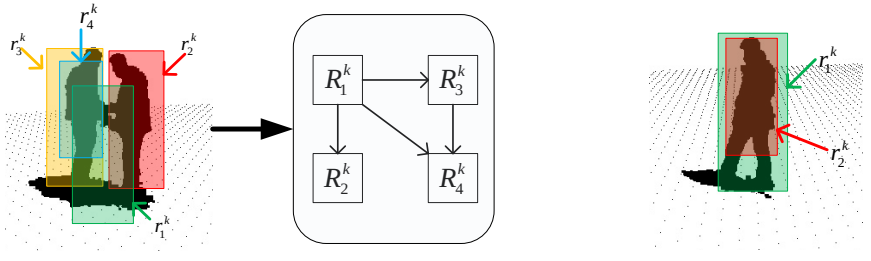
On the ground plane, $\{X_1, \dots X_n\}$ are independent from each other. However, in a single camera view, a pedestrian may be occluded by others. Here, we built a single camera Bayesian network (SBN) for each camera to indicate the potential occlusion relationship in a single view. Then a SBN is constructed, where R_i^k is the i^{th} node in SBN k, and R_j^k is a parent node of R_i^k in SBN k if r_j^k occludes r_i^k in camera k (including the part occlusion case). Figure 3(a) show an example of SBN. Based on the SBN, we have

$$P(R_i^k) = \sum_{j,\tau_j^k \in \{0,1\}} P\left(R_i^k \mid \{R_j^k = \tau_j^k\}_{R_j^k \in an_k(R_i^k)}\right) P\left(\{R_j^k = \tau_j^k\}_{R_j^k \in an_k(R_i^k)}\right). \quad (1)$$

Considering $P\left(R_i^k \mid \{R_j^k = \tau_j^k\}_{R_j^k \in an_k(R_i^k)}\right)$, the presence of pedestrians at the locations expect x_i is known as prior knowledge $\{R_j^k = \tau_j^k\}_{R_j^k \in an_k(R_i^k)}$. This makes possible to calculate the occupancy possibility through estimating the parts for location x_i where are not occluded by other pedestrians :

$$P\left(R_i^k \mid \{R_j^k = \tau_j^k\}_{R_j^k \in an_k(R_i^k)}\right) = \frac{\sum_{(w,h)_k \in r_i^k} \left(\prod_{R_j^k \in an_k(R_i^k) \cap (w,h)_k \in r_j^k} \left(1 - \tau_j^k\right)\right)}{\sum_{(w,h)_k \in r_i^k} 1}; \quad (2)$$

where $\tau_j^k = \{0,1\}$ and $(w,h)_k \in$ foreground. An typical example of Eq. (2) is shown in Figure 3(b), the occupancy possibility of R_2^k is influenced by the state of R_1^k. If R_1^k is treated as a real person, the occupancy possibility of R_2^k is 0, otherwise, it will be 1.



(a) (b)

Figure 3. (a) The left is the foreground information from a single view with four candidate locations, and the right is its corresponding SBN. (b) A typical example of Eq. (2) where $P(R_2^k = 1 | R_1^k = 1) = 0$, $P(R_2^k = 1 | R_1^k = 0) = 1$.

3.2 The Multi-view Bayesian network (MBN)

In order to estimate the occupancy possibility $P(X_i)$, we integrate all SBNs together as a multi-view Bayesian network (MBN). In fact, R_i^k corresponds to X_i in camera k, that is, $P(X_i|C_k) = P(R_i^k)$. Hence, we have

$$P(X_i) = \sum_{k=1}^{K} P(C_k) P(X_i \mid C_k) = \sum_{k=1}^{K} P(C_k) P\left(R_i^k\right), \tag{3}$$

where $P(C_k)$ is the weight value of camera k ($P(C_k) = 1/K$ in our experiments).

Considering $O(2^n)$ terms in the summation formula Eq. (1), Eq. (3) is a NP-hard problem, if there is no prior knowledge about the presentence of pedestrians. In order to cope with this difficulty, we further introduce a set of independent Boolean variable $\delta = \{\delta_1, \ldots, \delta_n\}$, which is called *"subjective supposing" nodes state* for MBN, where δ_i indicates whether the location x_i is occupied by a pedestrian.

Based on MBN with SSNS, the total probability is calculated by

$$P(X_i|\delta) = \sum_{k=1}^{K} P(C_k) \left(\sum_{\tau_j^k \in \{0,1\}} P\left(R_i^k \mid \{R_j^k = \tau_j^k\}_{R_j^k \in an_k(R_i^k)}, \delta\right) P\left(\{R_j^k = \tau_j^k\}_{R_j^k \in an_k(R_i^k)} \mid \delta\right) \right). \tag{4}$$

Then we can easily derive some Bayesian properties on MBN:

1) δ_j and R_i^k $(k = 1,2 \ldots K; i \neq j)$ are independent with each other;

2) δ_i is the stationary state of X_i and R_i^k, thus

$$P\left(R_i^k = \delta_i | \delta_i\right) = 1; P\left(R_i^k \neq \delta_i | \delta_i\right) = 0. \tag{5}$$

Hence, Eq. (4) is equal to

$$\begin{aligned} P(X_i|\delta) &= \sum_{k=1}^{K} P(C_k) P\left(R_i^k \mid \{R_j^k = \delta_j\}_{R_j^k \in an_k(R_i^k)}, \delta_i\right) \\ &= \sum_{k=1}^{K} P(C_k) P(R_i^k \mid \delta_i) P\left(R_i^k \mid \{R_j^k = \delta_j\}_{R_j^k \in an_k(R_i^k)}\right). \end{aligned} \tag{6}$$

This can be further simplified as

$$P(X_i = 1|\delta) = \delta_i \sum_{k=1}^{K} P(C_k) P\left(R_i^k \mid \{R_j^k = \delta_j\}_{R_j^k \in an_k(R_i^k)}\right). \tag{7}$$

Combining Eq. (7) and Eq. (2), we can estimate the occupancy possibility for each location on the ground plane if SSNS have been known.

4 MBN Learning

This section introduces our learning method of SSNS to best explain the image observations (i.e., foreground masks) from different views. Without loss of generality, suppose pixel $(w,h)_k$ belongs to $\{r_j^k\}$. In the ideal situation, foreground pixels all come from the pedestrians and pedestrians only appear in the foreground area in each view. With this assumption, we define the loss function $\Psi(w,h)_k$ to each foreground pixel $(w,h)_k$ in camera k, which quantifies the difference between the final results estimated from Eq. (7) and the ideal situation to pixel $(w,h)_k$.

$$\Psi\left((w,h)_k|\delta\right)=\begin{cases}\prod_{\mathrm{j}}(1-\mathrm{P}(X_j=1|\delta)), & if\ (w,h)_k\in foreground\\ 1-\prod_{\mathrm{j}}(1-\mathrm{P}(X_j=1|\delta)), & if\ (w,h)_k\in background\end{cases}. \tag{8}$$

Given the occupancy probability of all locations on the ground plane estimated from Eq. (7), we model the conditional probability $\mathrm{P}(D_k|\mathrm{X},\delta)$ of the image observations in camera k with the loss function:

$$\mathrm{P}\left(D_k|\mathrm{X},\delta\right)=\exp\left(\frac{-\sum_{(w,h)_k}\gamma(w,h)_k\Psi\left((w,h)_k|\delta\right)}{|D_k|}\right), \tag{9}$$

where $\gamma(w,h)_k$ is the weight value of pixel $(w,h)_k$. Note that $\gamma(w,h)_k$ are different for foreground and background pixels in our experiments. We suppose various background subtraction images from different views are independent from each other, so the likelihood function is defined as follow

$$F\left(\delta_1,\ldots\ldots\delta_n\right)=\ln\left(P\left(D\mid X,\delta\right)\right)=\sum_{k=1}^{K}\ln(P\left(D_k\mid X,\delta\right)), \tag{10}$$

where $\forall i,\delta_i=\{0,1\}$. In order to make $(\delta_1,\ldots\ldots\delta_n)$ generalizable well to the view observation, we have the following optimization problem

$$\left(\delta_1,\ldots\ldots\delta_n\right)=\arg\max F(\delta_1,\ldots\ldots\delta_n), \tag{11}$$

where $\forall i,\delta_i=\{0,1\}$. Eq. (11) gives the exact discrete formulation of the optimization of the likelihood function. It is a NP-hard problem because δ_i has a discrete constraint $\delta_i\in\{0,1\}$. As in the combinatorial optimization, we relax this discrete formulation to the continuous domain. Typically, here a set of auxiliary, real-valued, and continuous variable $\varepsilon=\{\varepsilon_i\}\in\{-\infty,+\infty\}^n$ are used to replace the original discrete variable $\forall$i, $\delta_i\in\{0,1\}$ with the sigmoid function

$$\delta_i=\frac{1}{1+\exp(-\varepsilon_i)},\ \varepsilon_i\in\left(-\infty,+\infty\right). \tag{12}$$

Thus, from Eq. (12), we obtain the following continuous formulation

$$\left(\varepsilon_1,\ldots\ldots\varepsilon_n\right)=\arg\max F(\varepsilon_1,\ldots\ldots\varepsilon_n). \tag{13}$$

It is obvious that $\mathrm{F}(\varepsilon_1,\ldots\ldots\varepsilon_n)$ is a derivable function. But it is difficult to derive its gradient formulation. Fortunately, we can get the value of gradient from Eq. (14) approximately

$$\frac{\partial F}{\partial\varepsilon_i}=\frac{F\left(\varepsilon_i+\Delta\varepsilon_i\right)-F\left(\varepsilon_i\right)}{\Delta\varepsilon_i}. \tag{14}$$

In the following, we use ∇F to denote the gradient vector. Then our algorithm for Eq. (13) is summarized in Alg.1.

Alg.1

1. Given $F_0=0,F_1=1$, initialize $\forall$i ,$\varepsilon_i^0=0$,t=0;

3. while $\left|\frac{F_1-F_0}{F_0}\right|>\epsilon$ do

$$F_0=\mathrm{F}(\varepsilon^t)$$

$$\Delta\tau\rightarrow\text{Line search }\varepsilon$$

$$\varepsilon^{t+1} = \varepsilon^t + \Delta\tau \frac{\nabla F^t}{|\nabla F^t|_\infty}$$

$$F_1 = \mathrm{F}(\varepsilon^{t+1})$$

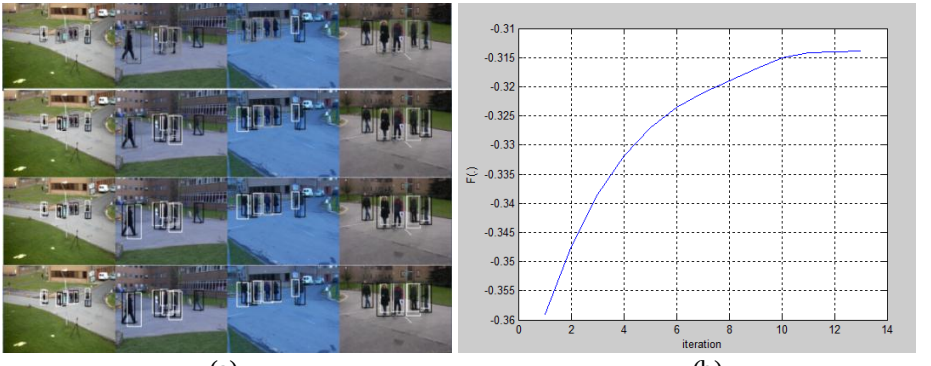
t=t+1

end

4 return $\forall \mathrm{i}, \varepsilon_i = \varepsilon_i^t$

Finally, putting ε into Eq. (2) and Eq. (7) to replace δ, we can get the final result $\mathrm{P}(X_i = 1|\delta)$.

Figure 4 illustrates an example of the learning process. In our experiments, Alg.1 will terminate mostly after 15~25 iterations. The processing time of our method is about 3s per frame on a PC with 3.1 GHz CPU and 4 GB memory.



(a) (b)

Figure 4. An example of the iterative process of Alg. 1 on a single frame. (a). From the top row to the end row, there are detection results after 1, 6, 11, 16 iterations of Alg. 1. The locations with a higher probability have a darker rectangle to correspond. At the 1st iteration, every location has a similar occupancy possibility. In the following iterations, the results can be distinguished clearly. The white rectangles stand for phantoms, and the black ones stands for pedestrians. (b). The curve of the likelihood function F(ε) under the iterations. The horizontal axis stands for the number of iterations of Alg.1.

5 Experiments

5.1 implementation

We generated foreground masks using an adaptive background subtraction algorithm [18]. For Eq. (9), the weight values of foreground and background pixels are equal to 2 and 1, respectively. When calculating the gradient using Eq. (14), the step size $\Delta\varepsilon_i$ is set to 10^{-4}. For Alg.1, we assign the terminate threshold as $\epsilon = 10^{-3}$.

5.2 Experiments results

We evaluated our method on the PETS2009 S2L1 and APIDIS datasets and used MODA and MODP [16] as the evaluation metrics. Usually, MODP measures the localization quality of the correct detections, while MODA measures the detection accuracy by taking into account both false and true correspondence. For comparison with state-of-the-art results in recent literature, we also uses the precision and recall measures calculated by the

ratios TP/(TP + FP) and TP/(TP +FN), where TP, FP and FN are the numbers of True Positive, False Positive and False Negative, respectively. Note that a true positive is counted only when a person is correctly located on the ground plane [11]. For all metrics, the larger their values, the better the performance. Several sample detection results for both datasets are shown in Figure 5(a) and (c).

PETS2009 S2L1 dataset. PETS2009 S2L1 dataset contains seven outdoor sequences from seven cameras, with 795 frames per every sequence. It is widely recognized that this dataset is a popular yet challenging benchmark dataset for multi-view pedestrian detection. In our experiments, we used four camera views (view 1,5,6,8) and compared our detection results with POM [10], which is one of the top-performers in Winter-PETS2009 [15] (the evaluation results of POM on the PETS09 S2L1 dataset come from [6]). We also compare our method with the method in [6], one of the latest results on this dataset. These experimental results show that our method obtains a much better performance both in terms of MODA and MODP. In addition, the Recall and Precision of our method are equal to 91% and 92%, respectively.

	Camera 1		Camera 5		Camera 6		Camera 8	
	MODA	MODP	MODA	MODP	MODA	MODP	MODA	MODP
Ours	**0.7931**	**0.7404**	**0.8183**	**0.7026**	**0.8392**	**0.7054**	**0.7812**	0.6989
Wu.G[6]	0.7532	0.6805	0.6998	0.6872	0.8162	0.6953	0.6941	**0.7004**
POM [10]	-0.1037	0.5806	0.2630	0.6071	0.3354	0.6467	0.2188	0.6344

Table 2: MODP and MODA in each view of PETS2009 S2L1, at an overlap threshold of 0.5

APIDIS dataset. APIDIS dataset contains seven cameras monitoring a basketball game. Compared with PETS09 S2L1 dataset, APIDIS dataset is more complex, such as more frequent severe collusions, non-standard standing gesture, the reflection of the players on the floor and strong shadows. Like [11], we tested our algorithm in camera 1,2,4,7 on the left-half of the basketball court. In this settings, we compared our method with POM [10] and Alahi [11], where the evaluation results of POM on the APIDIS dataset come from [11]. As far as our knowledge, Alahi [11] is the latest results on the APIDIS dataset. Note that since the work [11] does not provide the evaluation results in terms of MODA and MODP, we only compare their results in terms of Precision and Recall. Our method also performs very well on the APIDIS dataset, although there many noises present in the foreground (as shown in Figure 5(b)).

	Camera 1		Camera 2		Camera 4		Camera 7	
	MODA	MODP	MODA	MODP	MODA	MODP	MODA	MODP
Ours	0.77	0.67	0.75	0.66	0.78	0.67	0.74	0.75

(a)

	Recall	Precision
Ours	**0.84**	**0.90**
Alahi A [11]	0.69	**0.90**
POM [10]	0.63	0.51

(b)

Table 3: (a) MODP and MODA in each view of APIDIS, at an overlap threshold of 0.5. (b) Recall and Precision evaluation results.

It is interesting to note that our method can obtain the best performance even though we do not use any temporal information or post-processing on a single view. This is because by using MBN with SSNS, we can reduce the influence of phantoms effectively.

Note that the parameters in the experiments keep constantly on two different datasets. This makes our method applicable in different surveillance scenes.

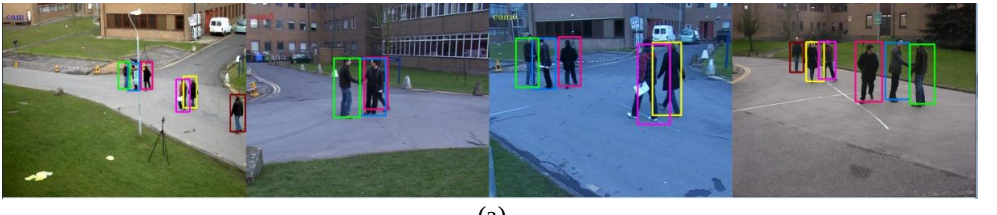
However, it is notable that the performances of our method in PETS view 1 and 8 are not as good as view 5 and 6. The reason is mainly because the presence of a slope in the scene makes homograph mapping not work well. Some examples of failure detection results are shown in Figure 5(d). In fact, the MODA of our method in camera 1 will be up to 0.84 if the slope area is ignored.

6 Conclusion

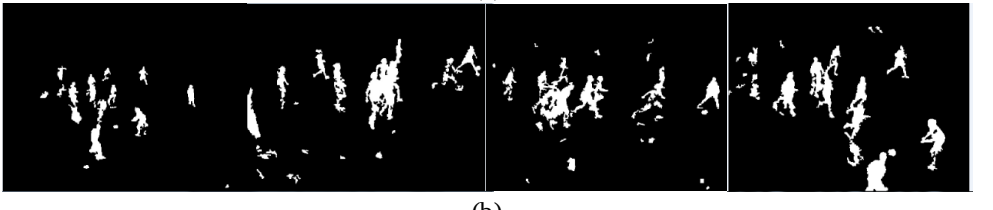
In this paper, we present a novel method that utilizes the multi-view Bayesian network (MBN) model to detect pedestrians from multi-view surveillance videos. Using MBN, our method can discriminate phantoms and pedestrian well. The experimental results on two challenging benchmark datasets demonstrate the effectiveness of our method compared with three state-of-the-art methods.

Acknowledgments

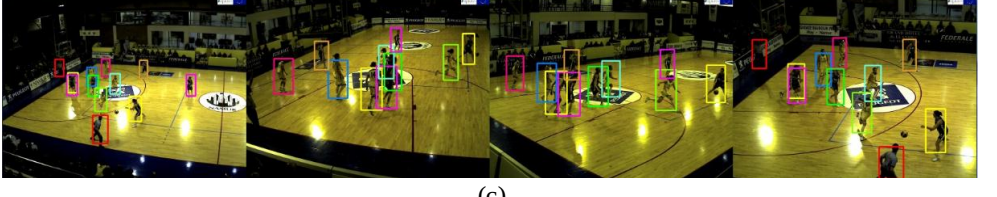
This work is partially supported by grants from the Chinese National Natural Science Foundation under contract No. 61035001 and No. 61072095, National Basic Research Program of China under contract No. 2009CB320906, and Fok Ying Dong Education Foundation under contract No. 122008.



(a)



(b)



(c)

(d)

Figure 5. Some detection results on PETS09 S2L1 and APIDIS datasets. (a) Sample detection result on the PETS09 S2L1 dataset. (b) and (c) An example for the foreground information and detection results on the APIDIS dataset. (d) A failure case due to the presence of the slope (black circle in view 1 and view 8).

References

[1] A. Yilmaz, O. Javed, and M. Shah. Object tracking: A survey. *ACM Computing Surveys*, 38(4):13, 2006.

[2] A. C. Sankaranarayanan, A. Veeraraghavan, and R. Chellappa. Object detection, tracking and recognition for multiple smart cameras. *Proc. of the IEEE*, 96(10):1606–1624, 2008.

[3] R. Eshel and Y. Moses. Homography based multiple camera detection and tracking of people in a dense crowd. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2008.

[4] Khan, S.M., Shah, M.: Tracking multiple occluding people by localizing on multiple scene planes. *IEEE Trans. Pattern Anal. Mach. Intell.*31 (2009) 505-519

[5] D. Arsić, N. Lehment, E. Hristov, B. Hrnler, B. Schuller,and G. Rigoll，"Applying multi-layer homography for multi camera tracking," in *Proceeedings Second ACM/IEEE International Conference on Distributed Smart Cameras*, ICDSC2008, Stanford, CA, USA, sep 2008.

[6] W. Ge and R. T. Collins. Crowd detection with a multiview sampler. In *Proc. of* In *European Conference on Computer Vision*, pages 324–337, 2010

[7] W. Ge and R. T. Collins. Marked point processes for crowd counting. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 2913–2920, 2009.

[8] A. Utasi and C. Benedek, "A 3D Marked Point Process Model for Multi-View People Detection," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 3385-3392, 2011.

[9] Fleuret, F., Lengagne, R., Fua, P.: Fixed point probability field for complex occlusion handling. In: *Proc. of the International Conference on Computer Vision* ,2005

[10] F. Fleuret, J. Berclaz, R. Lengagne, and P. Fua. Multicamera people tracking with a probabilistic occupancy map. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(2):267–282, 2008.

[11] Alahi A., Jacques L., Boursier Y., and Vandergheynst P., "Sparsity Driven People Localization with a Heterogeneous Network of Cameras,"*Journal of Mathematical Imaging and Vision*, vol.41,pp.39-58, 2009.

[12] Wei Qu,Dan Schonfeld,and Magdi Mohamed. Distributed Bayesian Multiple-Target Tracking in Crowded Environments Using Multiple Collaborative Cameras. *Journal on Advances in Signal Processing* Volume 2007

[13] Kuo-Chin Lien and Chung-Lin Huang Multi-view Based Cooperative Tracking of Multiple Human Objects. *EURASIP Journal on Image and Video Processing* Volume 2008

[14] A. Mittal and L. S. Davis. M2tracker: A multi-view approach to segmenting and tracking people in a cluttered scene usingregion-based stereo. *Int. J. of Computer Vision*, 51(3):189–203, 2002.

[15] Ellis, A., Shahrokni, A., Ferryman, J.M.: PETS2009 and Winter-PETS 2009 results:A combined evaluation. In: *Winter-PETS*. (2009)

[16] R. Kasturi, D. Goldgof, P. Soundararajan, V. Manohar,J. Garofolo, R. Bowers, M. Boonstra, V. Korzhova,and Jing Zhang. Framework for performance evaluation of face, text, and vehicle detection and tracking in video: Data, metrics, and protocol. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(2):319–336, Feb. 2009.

[17] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In I*EEE Comp. Vision and Pattern Rec.*, 2005.

[18] Zhipeng Hu, Yaowei Wang, Yonghong Tian, Tiejun Huang. *Selective eigenbackgrounds method for background subtraction in crowed scenes*. ICIP 2011: 3277-3280.