Pattern Recognition 48 (2015) 1760–1772

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Robust multiple cameras pedestrian detection with multi-view Bayesian network

Peixi Peng [a], Yonghong Tian [a,*], Yaowei Wang [b,a], Jia Li [c], Tiejun Huang [a]

[a] National Engineering Laboratory for Video Technology, School of Electronics Engineering and Computer Science, Peking University, China
[b] Department of Electronic Engineering, Beijing Institute of Technology, China
[c] State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, China

ARTICLE INFO

Article history:
Received 28 May 2014
Received in revised form
14 September 2014
Accepted 7 December 2014
Available online 17 December 2014

Keywords:
Pedestrian detection
Multiple cameras
Multi-view model
Bayesian inference
Height adaptive projection

ABSTRACT

Multi-camera pedestrian detection is the challenging problem in the field of surveillance video analysis. However, existing approaches may produce "phantoms" (i.e., fake pedestrians) due to the heavy occlusions in real surveillance scenario, while calibration errors and the diverse heights of pedestrians may also heavily decrease the detection performance. To address these problems, this paper proposes a robust multiple cameras pedestrian detection approach with multi-view Bayesian network model (MvBN). Given the preliminary results obtained by any multi-view pedestrian detection method, which are actually comprised of both real pedestrians and phantoms, the MvBN is used to model both the occlusion relationship and the homography correspondence between them in all camera views. As such, the removal of phantoms can be formulated as an MvBN inference problem. Moreover, to reduce the influence of the calibration errors and keep robust to the diverse heights of pedestrians, a height-adaptive projection (HAP) method is proposed to further improve the detection performance by utilizing a local search process in a small neighborhood of heights and locations of the detected pedestrians. Experimental results on four public benchmarks show that our method outperforms several state-of-the-art algorithms remarkably and demonstrates high robustness in different surveillance scenes.

1. Introduction

Pedestrian detection is a key step in many video surveillance applications, such as pedestrian tracking, crowd analysis and event detection. In existing studies on pedestrian detection, the most challenging task is to accurately locate multiple pedestrians in crowded scenes with heavy occlusions. In this case, it is often difficult to detect the occluded persons from the 2D view obtained from a single camera.

Intuitively, a feasible solution to detect the occluded persons in a crowded scene is to use multiple cameras that can provide complementary information for the same scene. Here a latent hypothesis is that two pedestrians overlapped in some views may become separable in other views. Following this hypothesis, Sankaranarayanan et al. [1] proposed to project the foreground of each view onto the same ground plane by homography. As shown in Fig. 1(a), such projections were then fused and their intersections were assumed to correspond to locations of the probable pedestrians. By assuming that all pedestrians have the same height [2–4], the candidate locations were projected back to each view for pedestrian detection (as shown in Fig. 1(b)). Similar to [1], Khan and Shah [5] projected the foreground from each view to a reference view. Kim [6] utilized lines to represent foregrounds in each view and projected those lines from each view to the ground plane.

Generally speaking, these approaches can achieve promising results in those scenes with weak occlusions. However, they may fail when the scenes become extremely crowded. In this case, the projection of one pedestrian may falsely intersect with the projections from some other pedestrians and such intersections may lead to phantoms (as shown in Fig. 1(c)). Moreover, the assumption that all pedestrians have the same height may not always hold. When projecting back to the camera views, the changing pedestrians' heights, as well as the synthesis noise from camera parameters, may lead to inaccurate detection results in all views (as shown in Fig. 1(d)).

To address these problems, we propose a novel approach for pedestrian detection in multiple cameras by using multi-view Bayesian network (MvBN). The overall framework of the proposed approach is shown in Fig. 2. We first obtain a set of preliminary detection results using the existing multi-view pedestrian detection methods with some predefined parameters (e.g., heights of pedestrians). Such results can be

* Corresponding author.
E-mail address: yhtian@pku.edu.cn (Y. Tian).

http://dx.doi.org/10.1016/j.patcog.2014.12.004
0031-3203/

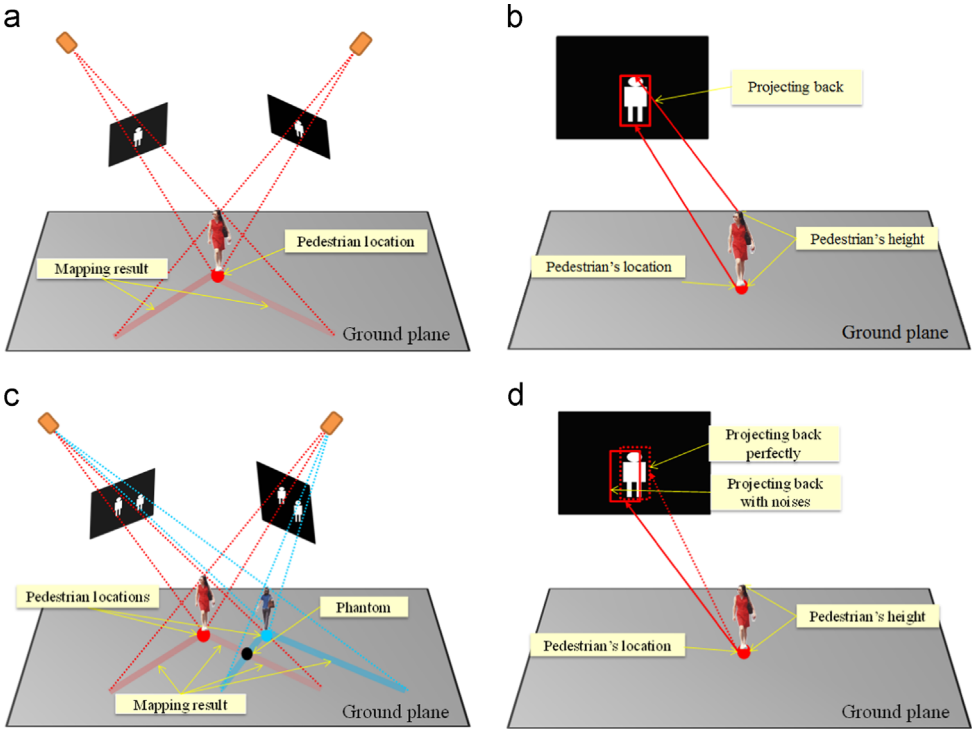

Fig. 1. (a) Homography-based multi-camera pedestrian detection method. (b) Projecting the location back with a predefined height to generate a detection result in the view. (c) The projection of one pedestrian falsely intersects with the projections of other pedestrians, consequently leading to phantoms. (d) An example of the synthesis noises.

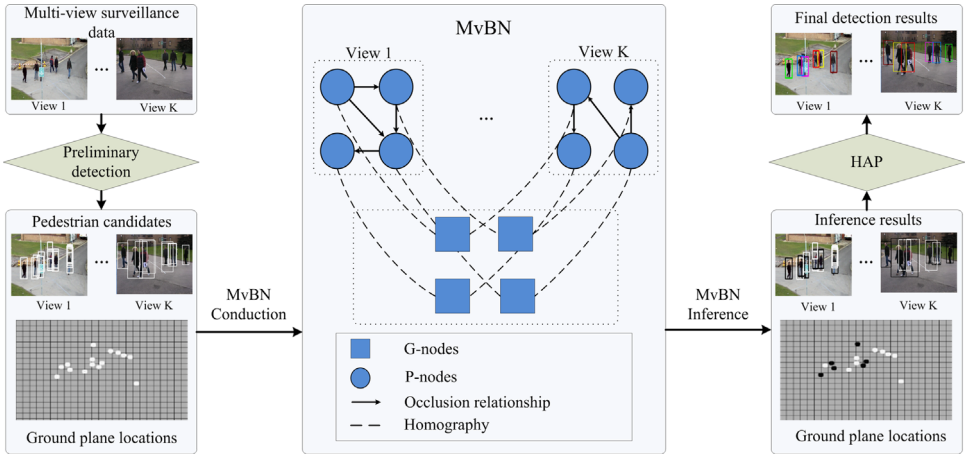

Fig. 2. The system framework of our approach. In the framework, the pedestrian candidates in all views and the corresponding locations on the ground plane are obtained by existing multi-view pedestrian detection methods with some predefined parameters (e.g., heights of pedestrians). Then, a Bayesian network is used to model the occlusion relationship between all candidates in each single view, while multiple Bayesian networks can be further combined to form a MvBN by considering the homography from the ground plane to the camera views. Thus phantoms (the white pedestrian candidates in Inference results) can be efficiently removed by inferring the G-nodes in the MvBN. Finally, the HAP is used to further refine the final detection results in each view, making the proposed method adaptive to diverse pedestrians' heights and robust to synthesis noises.

represented as the pedestrian candidates in all views and the corresponding locations on the ground plane. After that, a Bayesian network is used to model the occlusion relationship among all candidates in each camera view, and then multiple Bayesian networks can be further combined to form a MvBN by the homography from the ground plane to camera views. The MvBN includes two kinds of nodes that represent the pedestrian candidates (i.e., P-nodes) and the locations on the ground plane (i.e., G-nodes), respectively. The edges between P-nodes are used to model the occlusion relationship, while the edges between P-nodes and G-nodes represent the homography from the ground plane to different camera views. In other words, the MvBN is actually composed of G-nodes and several Bayesian networks, where G-nodes are used to combine the inference results from different Bayesian networks. Since phantoms are always concurrent with occlusions, such phantoms can be efficiently removed by inferring the G-nodes that demonstrate the highest probabilities of occluding the corresponding P-nodes.

Note that the preliminary work about MvBN have been published in [7]. But that version could not handle camera calibration noise and diverse pedestrians' heights. In order to solve this problem, a height-adaptive projection (HAP) method is proposed in this study. The HAP is used to further refine the detection results by utilizing a local search process in a small neighborhood of heights and locations of the detected pedestrians. In addition, more extensive experiments are conducted in this paper to demonstrate the effectiveness of the proposed approach. We test our approach on four public benchmarks,

including PETS09 S2L1[1], PETS09 City Center (CC)[2], APIDIS[3] and EPFL Terrace.[4] Experimental results show that our approach outperforms several state-of-the-art approaches (e.g., [3,4,8,9]) remarkably.

The main contributions of the proposed approach are summarized as follows:

1. We propose a multi-view Bayesian network to model pedestrian candidates and their occlusion relationship in all views. Through MvBN inference, phantoms from various views can be effectively removed.
2. We present a novel parameter learning algorithm for efficient MvBN inference. By incorporating a set of auxiliary, real-valued, and continuous variables, the MvBN inference process can be efficiently simplified.
3. A height-adaptive projection (HAP) method is proposed to obtain the final detection results in each view. Experiment results show that such method is robust to synthesis noises and calibration errors.

The rest of this paper is organized as follows: Section 2 reviews the related work and Section 3 presents the formulation of the MvBN model. A learning algorithm for MvBN inference is proposed in Section 4. Section 5 describes the height-adaptive projection method and Section 6 shows the experimental results on several benchmarks. Finally, the paper is concluded in Section 7.

2. Related work

In multi-view pedestrian detection, the main objective is to detect pedestrians in surveillance videos and estimate their 3D locations by fusing the visual cues from multiple viewpoints. Toward this end, existing approaches usually utilize the calibration information of multiple cameras and project the visual cues obtained from all views to the same reference plane. The reference plane, which is often selected as the ground plane, can thus be used to integrate the visual cues from multiple views for robust pedestrian detection [1]. Actually, the whole process of multi-view pedestrian detection can be summarized into three major steps: (1) extracting visual cues from all views and projecting them onto the reference plane; (2) fusing the cues to infer pedestrians' locations on the reference plane; and (3) projecting pedestrians' locations back to all views so as to obtain the final detection results. In this section, we will briefly review existing approaches mainly from these three steps.

In an early study of multi-view pedestrian detection, Kim and Davis [6] focused on refining the single-view pedestrian detection results with multi-view homography. They projected the detection results from all views to the same ground plane to find their intersection points, which were treated as the pedestrians' locations. Inspired by this idea, many approaches proposed to use the intersections on the ground plane to assist the pedestrian detection. Since it is often difficult to accurately detect pedestrians from each single view, most of these approaches only roughly detect the *foreground* regions from each view, while complicated analysis is conducted on the ground plane to locate pedestrians in these foreground regions. For example, Khan and Shah [5] projected multiple foreground masks (i.e., regions with motion) from all views to the reference plane by the planar homography. The overlapping regions of these projections were then selected as the pedestrian detection results. Franco and Boyer [10] also utilized homography to project foreground masks from multiple views to the ground plane and fuse them by a space occupancy grid. Generally speaking, these approaches are very efficient and outperform many single-view pedestrian detection approaches in the scenarios with weak occlusions. However, they may fail to process the scenarios with heavy occlusions. As stated in [11], the detection results generated by these approaches may contain many phantoms (i.e., fake pedestrians) in a crowded scene, which should be further distinguished from real pedestrians.

To solve this problem, the approaches in [12,13] extended the framework of [5] by projecting foreground masks to multiple parallel reference planes to distinguish phantoms and pedestrians. Although multiple reference planes can efficiently help to distinguish phantoms and real pedestrians, these approaches may be sensitive to inaccurate foreground masks (e.g., shadows). Eshel and Moses [14,15] proposed that the phantoms were tightly correlated with occlusions. To avoid occlusions, they placed the cameras at a high elevation, leading to much fewer phantoms. However, this scenario is quite different from many real scenarios where cameras are often fixed at a height of several meters.

Beyond these approaches, many recent studies proposed to formulate the task of multi-view pedestrian detection as an optimization problem. That is, they tried to integrate the visual cues from multiple views into an optimization framework to infer the real pedestrians by machine learning algorithms. For example, Fleuret et al. [2,3] proposed to construct a Probabilistic Occupancy Map (POM) for multi-view pedestrian detection by minimizing the Kullback–Leibler divergence between multi-view observations and the estimated probabilistic distribution. Alahi et al. [4] 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. Beyond these methods, some approaches proposed to extend the classical single view models to address the detection problem in the multi-view scenarios. For example, Ge and Collins [8] extended the classical Marked Point Process (MPP) model [16] by using a stochastic Gibbs sampling process on the ground plane. Akos and Benedek [17,9] also extended the MPP model, in which the sampling process was guided by the features extracted from pedestrians' heads and feet. In this manner, the samples obtained from all views become more accurate, leading to a better estimation of pedestrians' heights. However, this approach is sensitive to the number of cameras and may fail when the features from heads and feet are inaccurate.

To sum up, the existing multi-view pedestrian approaches can often outperform single view models by fusing the visual cues from multiple views. However, one main drawback of these multi-view approaches is that they may produce many phantoms since some crucial information may loss during the homography between camera views and the ground plane. Although some approaches such as [8] have tried to remove the phantoms independently in each view, their performances are still not very promising in scenes with heavy occlusions. Since phantoms are always concurrent with occlusions [18], a feasible solution to distinguish phantoms from real pedestrians is to analyze the occlusion relationship between various pedestrian candidates. Toward this end, we propose a multi-view pedestrian detection approach in this study, which adopts a multi-view Bayesian network to infer the phantoms by simultaneously using the occlusion relationship between pedestrian candidates and the homography relationship between the camera views and the ground plane. In the next section, we will introduce the technical details of the MvBN model.

3. Problem formulation

In this section, we present the formulation of the Multi-view Bayesian network (MvBN). To facilitate reading, the main notations are summarized in Table 1. The MvBN encodes the occlusion

[1] PETS09 S2L1: http://pets2009.net/
[2] PETS09 CC: http://pets2009.net/
[3] APIDIS: http://www.apidis.org/Dataset/
[4] EPFL Terrace: http://cvlab.epfl.ch/data/pom

Table 1
Notations.

N	The number of pedestrian candidates in each view
K	The number of cameras
$R_{k,i}$	A boolean variable standing for the ith pedestrian candidate in view k (1 for pedestrian and 0 for phantom)
$r_{k,i}$	The bounding box of the ith pedestrian candidate in view k
X_i	A boolean variable standing for the ith pedestrian candidate(1 for pedestrian and 0 for phantom)
$\mathbb{O}=\{O_k\}_{k=1}^{K}$	The occlusion relationship between pedestrian candidates in all views, where O_k is an $N\times N$ matrix with $O_k(i,j)=1$ if the ith pedestrian candidate occludes the jth one in view k, and $O_k(i,j)=0$ otherwise
H_k	The homography from the ground plane to view k, where H_k is an $N\times N$ matrix with $H_k(i,j)=1$ if the ith candidate location corresponds to the jth pedestrian candidate in view k by homography, and $H_k(i,j)=0$ otherwise
υ	A pixel in the foreground image
D_k	The foreground image in view k

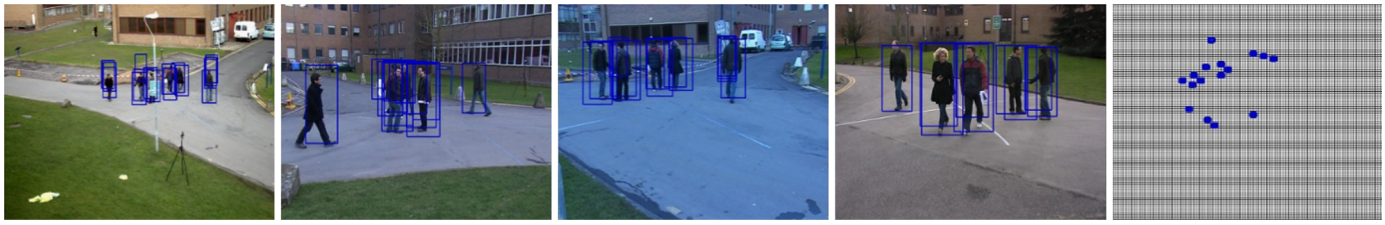

Fig. 3. Several detection examples of pedestrian candidates which include most of pedestrians and a lot of phantoms. The last column denotes the candidate locations on the ground plane.

relationship between pedestrian candidates (including pedestrians and phantoms) in K views as well as their homography relationship from the ground plane to each camera view. The pedestrian candidates can be produced by many existing methods and the Dimensionality Reduction method [4] is utilized in this study. It can obtain the locations of probable pedestrians (i.e., candidate locations) on the ground plane. After that, they use a fixed height (1.85 m) to generate pedestrian candidates in each view. In our method, the height is only a predefined parameter rather than the heights of final detected pedestrians. A typical detection example of pedestrian candidates is shown in Fig. 3.

3.1. Bayesian network in a single view

First, we will discuss a special case that aims to infer phantoms from a single view (i.e., $K=1$). Since all phantoms are concurrent with occlusions [18], we try to estimate the possibility that each candidate is a pedestrian in the view by analyzing the occlusion relationship (i.e., $P(R_{k,i}=1|O_k)$).

To model the pedestrian candidates and their occlusion relationship in the view, a Bayesian network $\mathcal{B}_k=\{\mathcal{V}_k,E_k\}$ is built. As shown in Fig. 4, each node (referred to as P-node) in $\mathcal{V}_k$ stands for a pedestrian candidate in view k and a binary variable $R_{k,i}$ is used to indicate whether it is a real pedestrian (i.e., $R_{k,i}=1$) or a phantom (i.e., $R_{k,i}=0$). In E_k, the edges between various P-nodes encode their occlusion relationship. $R_{k,i}$ is a parent node of $R_{k,j}$ if and only if the pedestrian candidate i occludes the pedestrian candidate j in view k (i.e., $O_k(i,j)=1$).

Since phantoms are always concurrent with occlusions, an intuitive way to infer whether $R_{k,i}=1$ is to measure how it is occluded by other P-nodes. For the sake of simplicity, we use $\mathcal{I}_{k,i}=\{i_1,\ldots,i_{|\mathcal{I}_{k,i}|}\}$ to represent the indices of P-nodes that occlude the ith P-node. Consequently, the probability that the ith P-node corresponds to a pedestrian has relationship with the prior probability of itself and its parent nodes:

$$P(R_{k,i}=1|O_k)=P(R_{k,i}=1|R_{k,i_1},\ldots,R_{k,i_{|\mathcal{I}_{k,i}|}})$$
$$=f(P(R_{k,i}=1),P(R_{k,i_1}=1),\ldots,P(R_{k,i_{|\mathcal{I}_{k,i}|}}=1)), \quad (1)$$

where $P(R_{k,i}=1)$ is the prior probability of that $R_{k,i}$ is a pedestrian. Intuitively, P-nodes with heavy occlusions have high probabilities of being phantoms, then $f(\cdot)$ can be defined as

$$f(P(R_{k,i}=1),P(R_{k,i_1}=1),\ldots,P(R_{k,i_{|\mathcal{I}_{k,i}|}}=1))$$
$$=\underbrace{P(R_{k,i}=1)}_{\text{prior probability}}\ \underbrace{\frac{1}{|r_{k,i}|}\sum[\upsilon\in r_{k,i}]_{\mathbf{I}}\left(\prod_{i_n\in\mathcal{I}_{k,i}}1-[\upsilon\in r_{k,i_n}]_{\mathbf{I}}P(R_{k,i_n}=1)\right)}_{\text{occlusion term}}. \quad (2)$$

where $\upsilon\in$ *foreground* and $|r_{k,i}|$ is the number of foreground pixels in $r_{k,i}$, $[\mathbf{e}]_{\mathbf{I}}=1$ if event $\mathbf{e}$ holds, otherwise $[\mathbf{e}]_{\mathbf{I}}=0$. As shown in (2), the probability of that $R_{i,k}=1$ is composed of two components: its prior probability and occlusion term. The occlusion term decreases $P(R_{i,k}=1|O_k)$ when the pedestrian candidate is occluded by others because phantoms are always occluded by real pedestrians. Although (1) can infer phantoms in a single view, yet it will not hold in the following situations:

1. Some phantoms may not be occluded by any pedestrian.
2. Some pedestrians are occluded by other pedestrians.

The case 1 will lead that some phantoms may be treated as pedestrians while the case 2 will cause that some pedestrians have a low probability by (2).

3.2. Multi-view Bayesian network

In order to address these two problems, we have to integrate the cues from multiple views. To this end, a group of virtual nodes $\{X_i\}_{i=1}^{N}$ (called as G-nodes) are introduced, where X_i denotes the ith candidate location on the ground plane. Then these Bayesian networks from multiple views can be combined together as a Multi-view Bayesian network (MvBN) $\mathcal{B}=\{\{X_i\}_{i=1}^{N},\{\mathcal{B}_k\}_{k=1}^{K}\}$. The MvBN is actually composed of K Bayesian networks $\{\mathcal{B}_k\}_{k=1}^{K}$ and N G-nodes $\{X_i\}_{i=1}^{N}$, where each G-node is used to combine the inference results for each

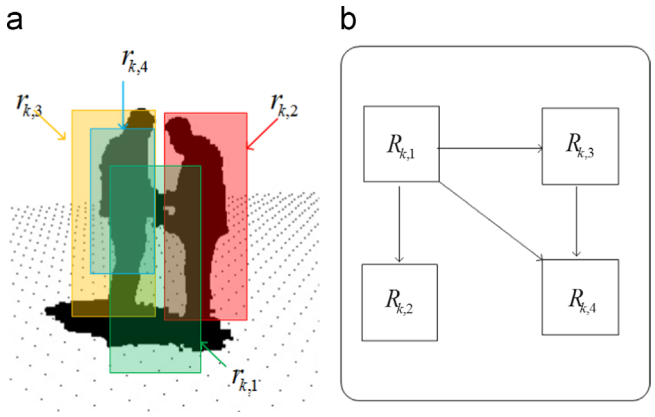

Fig. 4. An example of Bayesian network in a single view. (a) 4 pedestrian candidates in one view. $r_{k,2}$ and $r_{k,3}$ are pedestrians, while $r_{k,1}$ and $r_{k,4}$ are phantoms. Since phantoms are always concurrent with occlusions, we try to remove phantoms by analyzing the occlusion relationship. (b) The corresponding Bayesian network. Each node indicates a pedestrian candidate, while each edge encodes the relationship between two candidates. For example, $r_{k,3}$ occludes $r_{k,4}$, then $R_{k,3}$ is a parent node of $R_{k,4}$; $r_{k,1}$ occludes $r_{k,2}$, $r_{k,3}$ and $r_{k,4}$, hence $R_{k,1}$ is the parent node of all other nodes.

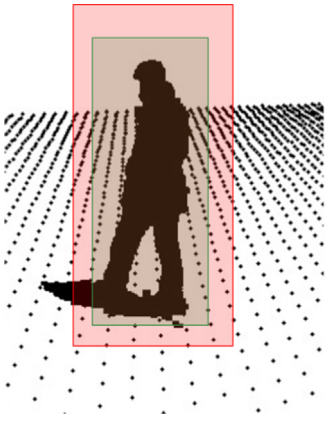

Fig. 5. The phantom (i.e., the large rectangle) always cannot match the foregrounds as well as the pedestrian (i.e., the small rectangle) who are occluded by them.

potential pedestrian from $\{\mathcal{B}_k\}_{k=1}^{K}$. Considering that $\{H_1,\ldots,H_K\}$ is a universal set about homography, the conditional probability of each G-node based on the occlusion relationship, which is the desired result for our model, can be estimated as (3) by the total probability formula:

$$P(X_i=1|\mathbb{O})=\sum_{k=1}^{K}P(H_k)P(X_i=1|\mathbb{O},H_k), \tag{3}$$

where $P(H_k)$ is the weight value of view k. In our study, all cameras are regarded as equally important:

$$P(H_1)=\cdots=P(H_K)=\frac{1}{K}. \tag{4}$$

Note that there are one-to-one correspondences between pedestrian candidates in camera k and locations on the ground plane through homography (i.e., $H_k(i,i)=1$ and $\forall H_k(i,j)=0(i\neq j)$). Therefore, we assume X_i is equivalent to $R_{k,i}$ when H_k is knowable:

$$P(X_i=1|\mathbb{O},H_k)=P(R_{k,i}=1|\mathbb{O},H_k). \tag{5}$$

Considering that $\forall O_l(l\neq k)$ is the occlusion relationship in the other views and H_k is the homography from the ground plane to the camera view, hence $\forall O_l(l\neq k)$ and H_k are independent with $R_{k,i}$. Then (5) can be further simplified as follows:

$$P(X_i=1|\mathbb{O},H_k)=P(R_{k,i}=1|O_k). \tag{6}$$

Then substituting (1), (4) and (6) into (3), the desired possibility of each G-node will be expressed as follows:

$$\begin{aligned}P(X_i=1|\mathbb{O})&=\sum_{k=1}^{K}P(H_k)P(R_{k,i}=1|O_k)\\&=\frac{1}{K}\sum_{k=1}^{K}f(P(R_{k,i}=1),P(R_{k,i_1}=1),\ldots,P(R_{k,i_{|\mathcal{I}_{k,i}|}}=1)).\end{aligned} \tag{7}$$

It is a challenge to solve (7) because the prior probabilities of P-nodes are unknown. Considering the homography from the ground plane to the camera views, we can simplify (7) from $K\times N$ unknown variables to N variables by the hypothesis:

$$P(R_{1,i}=1)=P(R_{2,i}=1)=\cdots=P(R_{K,i}=1)=\delta_i,\ \delta_i\in[0,1], \tag{8}$$

where $\{\delta_1,\ldots,\delta_N\}$ are a set of independent variables, which are presented as parameters of the MvBN. For the sake of simplicity, we utilize $\boldsymbol{\delta}$ to denote $\{\delta_1,\ldots,\delta_N\}$. Substituting the parameters that $P(R_{k,i}=1)=\delta_i$ to (7), the desired probability $P(X_i=1|\mathbb{O})$ can be future represented as

$$P(X_i=1|\mathbb{O})=\frac{\delta_i}{K}\sum_{k=1}^{K}\frac{1}{|r_{k,i}|}\sum[v\in r_{k,i}]_{\mathbf{I}}\left(\prod_{i_n\in\mathcal{I}_{k,i}}1-[v\in r_{k,i_n}]_{\mathbf{I}}\delta_{i_n}\right). \tag{9}$$

Recalling the two problems of the Bayesian network in the single view, these can be effectively solved by (9) because:

- To the phantoms which occlude pedestrian in some views, we can decrease the corresponding prior probabilities.
- To the pedestrians which are occluded by other pedestrians in a single view, it is impossible that he (or she) is occluded completely in all views.

Notice that the phantoms which occlude pedestrians always cannot match the foregrounds as well as the pedestrians who are occluded by them (as shown in Fig. 5). Based on this fact, we try to solve the prior possibilities of P-nodes by finding the parameters that make the MvBN inference results best explain image observations (foreground masks):

$$\delta^*=\arg\min_{\delta}L(P(X_1=1|\mathbb{O}),\ldots,P(X_N=1|\mathbb{O})), \tag{10}$$

where $L(\cdot)$ is the loss function. The calculation details and learning algorithm about (10) will be presented in the following section.

4. MvBN learning

After the MvBN inference procedure, the phantoms removal could be transformed to a parameter learning problem about δ. This section aims at solving two problems of (10): how to model the loss function and its learning algorithm.

4.1. Loss function

In the proposed approach, the loss function (10) is calculated by combining the loss functions of all foreground and background pixels from all views:

$$\begin{aligned}&L(P(X_1=1|\mathbb{O}),\ldots,P(X_N=1|\mathbb{O}))\\&\quad=\sum_{k=1}^{K}\frac{\sum_{v\in D_k}\gamma_v l(v,P(X_1=1|\mathbb{O}),\ldots,P(X_N=1|\mathbb{O}))}{|D_k|}.\end{aligned} \tag{11}$$

where γ_v is the weight value of a pixel (1.0 for a foreground pixel and 0.4 for a background pixel) and $l(\cdot)$ is the loss function for each pixel. Background subtraction is a commonly used technology to obtain foreground masks. Supposing that the background model is good

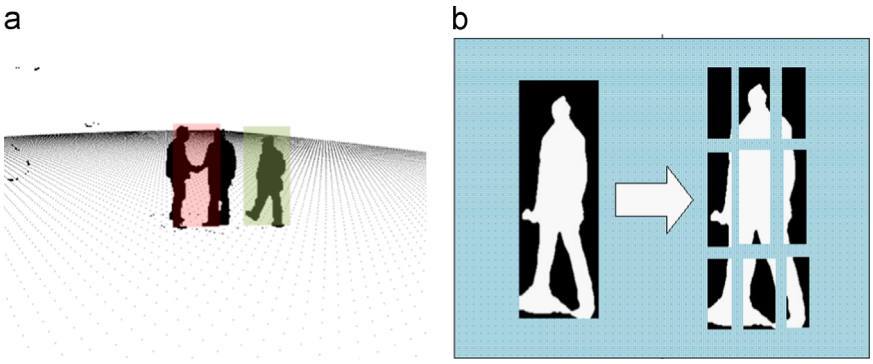

Fig. 6. (a) If a background pixel locates in the middle part of a bounding box of a pedestrian candidate (the left one), the pedestrian candidate is always fake; while a background pixel locates in the around part of a bounding box of a pedestrian candidate (the right one), the pedestrian candidate is likely a pedestrian. (b) Foreground pixels of a pedestrian do not distribute in the rectangle uniformly.

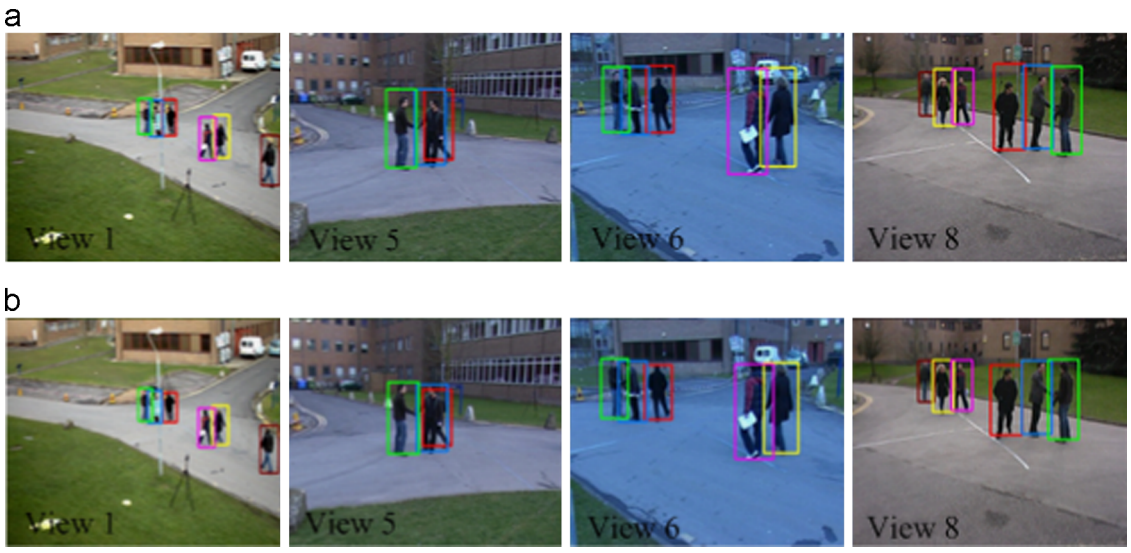

Fig. 7. (a) Before the robust height-adaptive detection method, all pedestrians have a same fixed size. (b) By the HAP method, the detection results are adaptive to diverse pedestrians' heights.

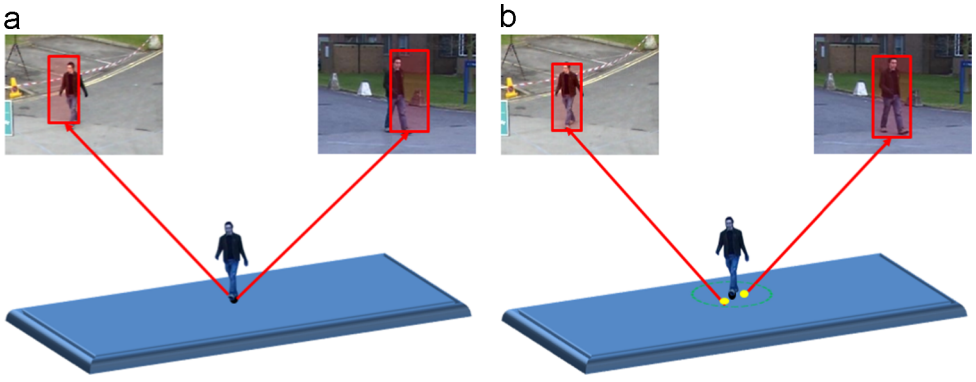

Fig. 8. (a) Projecting the original location directly will be influenced by synthesis noises easily. (b) An example of "projected locations". In our method, we generate detection results in single views by projecting the corresponding "projected locations," which are nearby the original location.

enough, all pedestrians should appear at the positions of foregrounds and all foreground pixels originate in pedestrians. Therefore, there are three clues about pixel and pedestrian:

1. If v is a foreground pixel, it means that there is at least one pedestrian appearing at the position of v.
2. If v is a background pixel, it is likely that there is no pedestrian at the position of v.
3. In the bounding box of a pedestrian, there are more foreground pixels distributing in the part close to the central axis because of the body, while less in the area far from the central axis (as shown in Fig. 6).

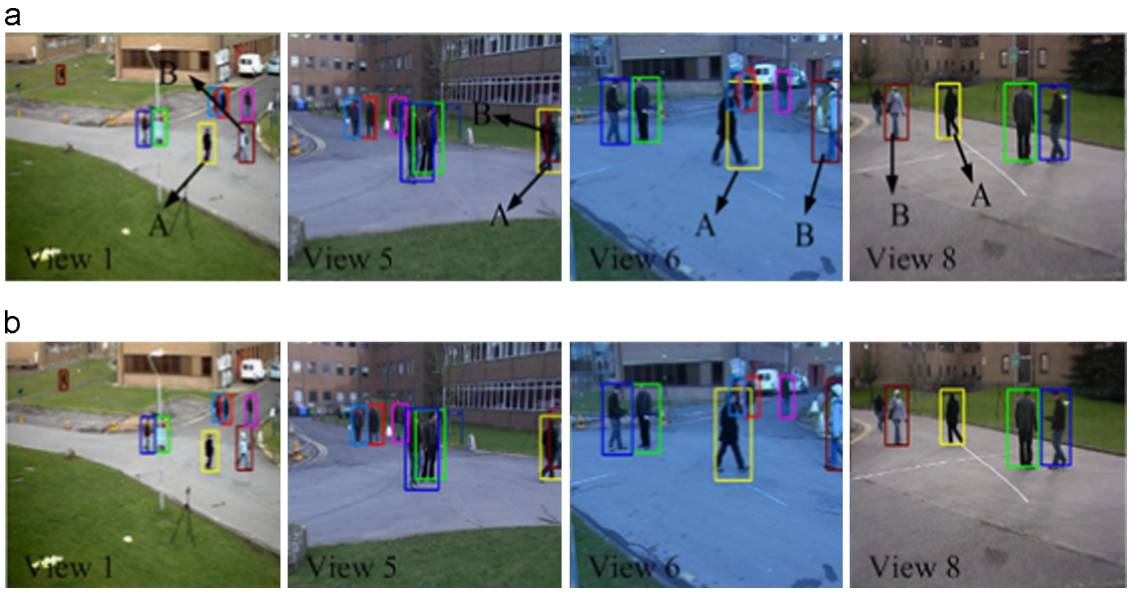

Fig. 9. (a) Synthesis noises influence detection results in each view independently. The detection results of pedestrian A are accurate in View 1, 5 and 8. But it has some deviation in View 6. The detection results of pedestrian B are accurate View 5 and 8, but have some deviations in Views 1 and 6. (b) Detection results refined by the proposed HAP method.

According to these clues, the loss function of a pixel could be defined as follows:

$$l(\upsilon, P(X_1=1|\mathbb{O}),\ldots,P(X_N=1|\mathbb{O})) = \begin{cases} 1-\max\limits_{\{i|\upsilon\in r_i^k\}}\{\phi(d_i)P(X_i=1|O_k)\}, & \text{if } \upsilon\in\text{foreground;} \\ \max\limits_{\{i|\upsilon\in r_i^k\}}\{\phi(d_i)P(X_i=1|O_k)\}, & \text{if } \upsilon\in\text{background.} \end{cases} \quad (12)$$

where d_i is the distance from υ to the middle vertical axis of r_i^k, $\phi(d_i)\in[0,1]$ and $\propto(1/d_i)$. In order to make (10) to be a derivable differentiable problem, we utilize the *softmax* function [19] (NOR in our experiments) as an approximation of the max function in (12).

4.2. Learning algorithm

The optimization problem of learning δ has been given by (10) and (12). Like in the combinatorial optimization, here a set of auxiliary, real-valued, and continuous variables $\varepsilon=\{\varepsilon_1,\ldots,\varepsilon_N\}$ are used to replace $\delta=\{\delta_1,\ldots,\delta_N\}$ with the sigmoid function:

$$\delta_i=\frac{1}{1+\exp(-\varepsilon_i)},\quad \varepsilon_i\in(-\infty,+\infty). \quad (13)$$

Thus, substituting these variables into (10) yields the optimization problem as follows:

$$\varepsilon^*=\arg\min_{\varepsilon} L(\varepsilon). \quad (14)$$

Obviously, $L(\varepsilon)$ is a derivable function, despite it is difficult to derive its gradient formulation. Thus we can estimate the gradient value of $L(\varepsilon)$ approximately by

$$\frac{\partial L}{\partial\varepsilon_i}=\frac{L(\varepsilon_i+\Delta\varepsilon_i)-L(\varepsilon_i)}{\Delta\varepsilon_i}, \quad (15)$$

where $\Delta\varepsilon_i$ is a very small number such as 0.001. In the following, we use ∇L to denote the gradient vector of $L(\varepsilon)$ about variables ε. Then the gradient descent method is utilized to solve (14) approximately. This algorithm is summarized in Algorithm 1.

Algorithm 1. MvBN learning.

Input: $L_0, L_1, \varepsilon_i=0 (i=1,\ldots n)$;

Output: ε^*

while $|L_0-L_1|>\Delta(\Delta$ is a constant) **do**

$L_0=L(\varepsilon)$;

$\nabla\tau\rightarrow Linesearch(L,\varepsilon)$;

$\varepsilon=\varepsilon+\nabla\tau\frac{\nabla l(\varepsilon)}{|\nabla L(\varepsilon)|_\infty}$;

$L_1=L(\varepsilon)$;

return $\varepsilon^*=\varepsilon$;

After the learning procedure, we put ε^* to (13) and (9) to obtain the final MvBN inference results of each G-nodes and choose the pedestrian candidates where $P(X_i=1|\mathbb{O})>threshold$ as final detected pedestrians. In our experiments, Algorithm 1 will be terminated after at most 15 iterations.

5. Height-adaptive projection (HAP)

As mentioned above, the MvBN can distinguish pedestrians from phantoms on the ground plane. The pedestrians' locations produced by MvBN inference are represented by $\{(x_i,y_i)\}_{i=1}^M$, where (x_i,y_i) is the world coordinate on the ground plane and m is the number of the detected pedestrians. Given by the predefined pedestrians' height h_0, the detection results in each view can be generated by projecting $\{(x_i,y_i)\}_{i=1}^M$ from the ground plane to the camera views. However, there are two problems which will cause detection errors in this process: First, the predefined height value h_0 does not handle the diverse pedestrians' heights (as shown in Fig. 7); Second, such projection procedure is easily influenced by the noisy inputs, e.g., errors in calibration and synchronization (as shown in Fig. 8(a)).

In order to solve these two problems, a novel height-adaptive projection (HAP) method is proposed here. In this method, each detected pedestrian is expressed as

$$\{(x_i,y_i),h_i,(x_{k,i},y_{k,i})_{k=1,\cdots,K}\}_{i=1,\cdots,M}, \quad (16)$$

where (x_i,y_i) denotes the pedestrian's original location and h_i is used to describe the pedestrian's height. $(x_{k,i},y_{k,i})$, which is in the

Table 2
Comparison of different datasets.

Dataset	View types	Indoor or outdoor	Detection targets
PETS09 CC	Far field	Outdoor	Pedestrians
PETS09 S2L1	Far field + eye-level	Outdoor	Pedestrians
APIDIS	Far field	Indoor	Basketball players
Terrace	Eye-level	Outdoor	Pedestrians

Table 3
Evaluation results on different datasets. MODA/MODP is the mean value of MODA/MODP for all used views in this dataset.

	The ground plane		Image views	
	RECALL	PRECISION	MODA	MODP
(a) PETS CC				
Pe-candidate	**0.93**	0.41	−0.11	0.77
MvBN only	0.90	**0.97**	0.82	0.76
MvBN+HAP	0.90	**0.97**	**0.87**	**0.78**
(b) PETS S2L1				
Pe-candidate	**0.96**	0.38	−0.59	0.71
MvBN only	0.95	**0.94**	0.81	0.73
MvBN+HAP	0.95	**0.94**	**0.87**	**0.75**
(c) APIDIS				
Pe-candidate	**0.94**	0.39	−0.53	0.69
MvBN only	0.87	**0.94**	0.75	0.70
MvBN+HAP	0.87	**0.94**	**0.83**	**0.75**
(d) TERRACES				
Pe-candidate	**0.86**	0.51	−0.51	0.69
MvBN only	0.81	**0.94**	0.71	0.68
MvBN+HAP	0.81	**0.94**	**0.82**	**0.73**

neighborhood of (x_i, y_i), is the "projected location" of the pedestrian for view k. In this method, detection results in a single view are generated by projecting the corresponding "projected locations" instead of the original locations (as shown in Fig. 8). Given $(x_{k,i}, y_{k,i})$ and h_i, the bounding box $r_{k,i}$ of the ith pedestrian in view k would be generated. Note that the "projected locations" are different among these views because the synthesis noises usually influence the detection results in different views independently. A typical sample is shown in Fig. 9.

Formally, the HAP method aims at finding $\mathbb{P} = \{(x_{k,i}, y_{k,i})_{k=1,\cdots,K}, h_i\}_{i=1,\cdots,M}$ which can make the final detection results $\{r_{k,i}\}_{i=1,\cdots,M}^{k=1,\cdots,K}$ explain image observations (foregrounds) best. Hence, it can be formulated as a constrained optimization problem:

$$\mathbb{P}^* = \arg\max_{\mathbb{P}} H(\mathbb{P}),$$
$$\text{s.t.}(x_{k,i}-x_i)^2+(y_{k,i}-y_i)^2 < C(\forall i,k), \tag{17}$$

where C is a distance constraint (1.0 m in our experiments setting). Here $H(\mathbb{P})$ is used to depict how well the final detection results explain foregrounds from different views, which is calculated by combining the loss functions of all pixels in all views:

$$H(\mathbb{P}) = -\sum_{k=1}^{K} \frac{\sum_{v\in D_k}\gamma_v L_h(v,\mathbb{P})}{|D_k|}, \tag{18}$$

where γ_v is the weight value of a pixel (1.0 for foreground pixels and 0.3 for background pixels). $L_h(v,\mathbb{P})$ is the loss function of the pixel v about $\mathbb{P}$, which is defined as

$$L_h(v,\mathbb{P}) = \begin{cases} \prod_{\{i|v\in r_{k,i}\}} (1-\phi(d_i)), & \text{if } v\in \text{foreground;} \\ 1-\prod_{\{i|v\in r_{k,i}\}} (1-\phi(d_i)), & \text{if } v\in \text{background.} \end{cases} \tag{19}$$

where d_i is the distance from v to the middle vertical axis of $r_{k,i}$ and $\phi(d_i)$ is same to (12).

Also it is difficult to derive the gradient formulation of (18). Similarly, we can estimate the gradient value approximately as (15). In the following, we use ∇H to denote the gradient vector of (18). Thus the algorithm for (17) is summarized in Algorithm 2. It will be terminated after approximately 8 iterations in our experiments.

Algorithm 2. Height adaptive projection.

Input: $H_0, H_1, \{x_{i,k}=x_i, y_{i,k}=y_i, h_i=h_0\}_{\forall i,k}$;
Output: $\mathbb{P}^*$
while $|H_0-H_1| > \Delta(\Delta \text{ is a constant})$ **do**
  $H_0 = H(\mathbb{P})$;
  if $(x_{i,k}-x_i)^2+(y_{i,k}-y_i)^2 \geq C, (\forall i,k)$ **then**
    $\frac{\partial H}{\partial x_{i,k}} = 0, \frac{\partial H}{\partial y_{i,k}} = 0$;
  $\nabla\tau \rightarrow Linesearch\{H, \mathbb{P}\}$;
  $\mathbb{P} = \mathbb{P} + \nabla\tau\frac{\nabla H}{|\nabla H|_\infty}$;
  $H_1 = H(\mathbb{P})$;
return $\mathbb{P}^* = \mathbb{P}$.

6. Experiments

6.1. Experimental settings

In this section, we evaluate the performance of our approach on four benchmark datasets, named *PETS2009 S2L1, City Center, APIDIS and EPFL Terrace*. They all contain crowded images from multiple calibrated cameras. The foreground masks are obtained by [20].

- *PETS09 S2L1* is one of the most popular challenging benchmark datasets to evaluate the performance of multi-view pedestrian detection algorithms. It contains seven sequences from seven outdoor cameras and each sequence consists of 795 frames. Four camera views are used in the experiments, including one far field view (*View 1*) and three eye-level views with frequent, severe occlusions (*Views 5*, *6* and *8*).
- The second dataset is also from the *PETS09* datasets. We selected the *City Center (CC)* images containing approximately 1 min of recordings (400 frames total) in an outdoor environment. Although this dataset comes from the same cameras as those in PETS S2L1, the main difference is that the chosen 2 cameras in PETS CC are all far field views (*Views* 1 *and* 2). An area-of-interest of size 12.2 m × 14.9 m is used in the experiments. It is visible from all two cameras as shown in Fig. 11.
- *APIDIS* comes from seven indoor cameras monitoring a basketball game. It is much more challenging compared with PETS2009. For example, there are more frequent severe occlusions and strong shadows caused by the reflection of the players on the floor. Persons in *APIDIS* are not "normal" pedestrians but basketball players with abrupt changes of behavior. They run, jump or change their motion pathes suddenly. The proposed algorithm was tested in *Views* 1, 2, 4 and 7 on the left-half of the basketball court.
- The fourth dataset is *EPFL Terrace*, which is 3 min and 20 s long (5000 frames total). It was recorded in a controlled outdoor environment. Several pedestrians walked in a small space on a terrace. As same as [9], two cameras (*Views* 1 *and* 2) which all have eye-level views were selected, and an area-of-interest as a 5.3 m × 5.0 m rectangle is used in the experiments. It is visible from all two cameras as shown in Fig . 11.

These sequences vary with respect to the viewpoints, types of pedestrian movement, surveillance environments and amount of

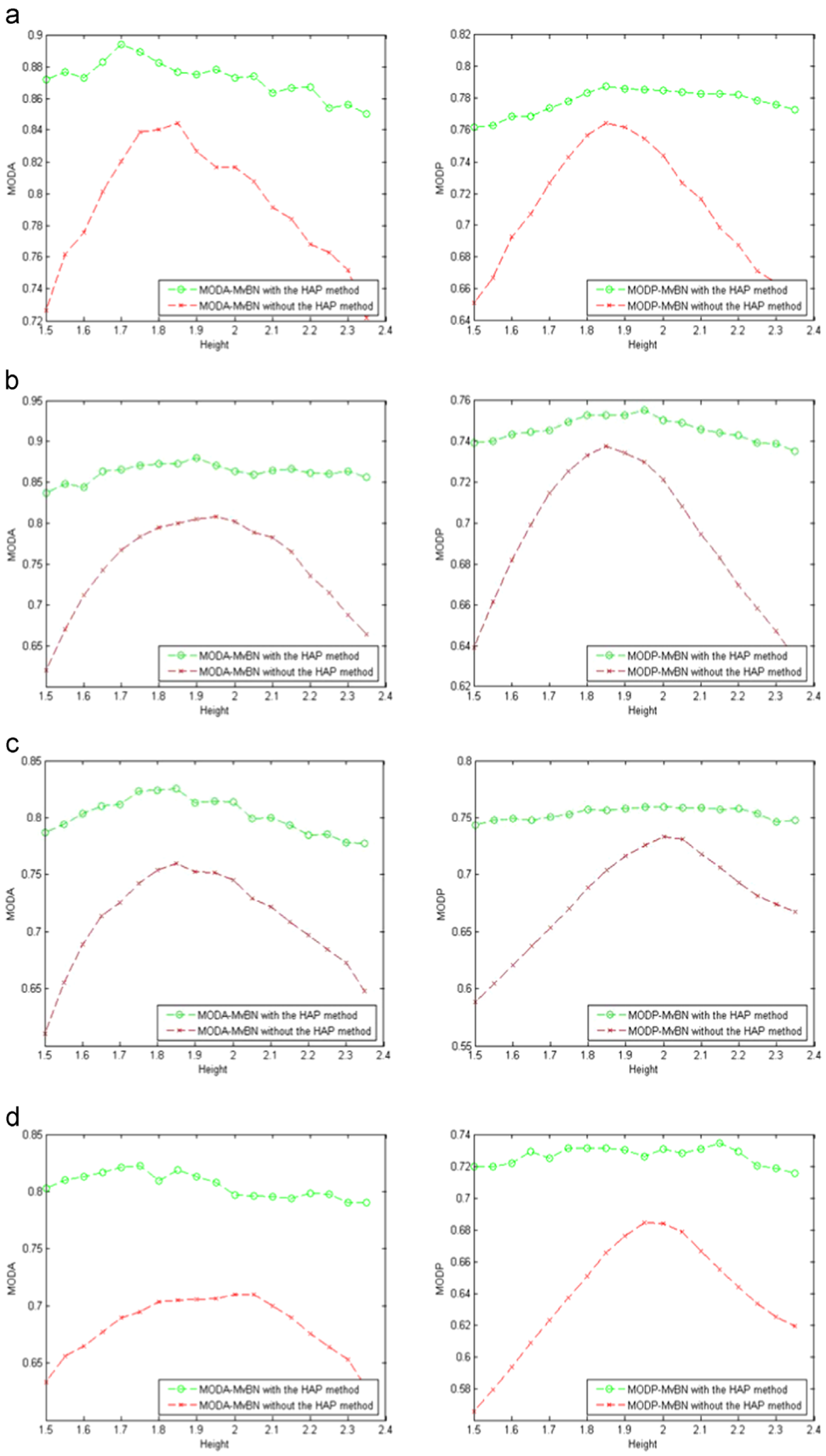

Fig. 10. Evaluation results on different datasets based on different prior pedestrians' heights. The horizontal axis indicates the prior pedestrians' height (m). The vertical axis is the mean value MODA/MODP for all used views in that dataset.

Table 4
Evaluation results on different datasets. MODA/MODP is the mean value of MODA/MODP for all used views in this dataset.

	The ground plane		Image views	
	TER		MODA	MODP
(a) PETS CC				
POM	0.28		0.70	0.55
3DMPP	0.31		–	–
Ours	**0.13**		**0.87**	**0.78**
	The ground plane		Image views	
	RECALL	PRECISION	MODA	MODP
(b) PETS S2L1				
POM	0.70	0.91	0.65	0.67
Multiview Sampler	–	–	0.72	0.69
Ours	**0.95**	**0.94**	**0.87**	**0.76**
	The ground plane		Image views	
	RECALL	PRECISION	MODA	MODP
(c) APIDIS				
POM	0.52	0.75	0.35	0.69
O-Lasso	0.70	0.90	–	–
Ours	**0.87**	**0.94**	**0.82**	**0.75**
	The ground plane		Image views	
	TER		MODA	MODP
(d) TERRACE				
POM	0.81		0.19	0.56
3DMPP	0.37		–	–
Ours	**0.24**		**0.82**	**0.73**

occlusions. The differences between these datasets are summarized in Table 2.

On these four datasets, three experiments are conducted. In the first experiment, the main objective is to demonstrate the effectiveness of different components of our approach. To be more specific, the experiment is used to validate the MvBN model and the HAP method respectively. Then, we will validate the robustness of the detection results against the prior pedestrians' height, which is a predefined parameter for MvBN. The last experiment will show our advantages by comparing with four state-of-the-art algorithms, including POM [3], Multiview Sampler [8], 3DMPP [9] and O-Lasso [4]. POM is a remarkable method of multi-view pedestrian detection and one of the top-performers as reported in Winter-PETS2009 [21]. Multiview Sampler is one of the latest and most effective methods which have evaluation results on *PETS S2L1*. The latest evaluation results of *APIDIS* were presented in O-Lasso to our knowledge. 3DMPP showed the latest results in the *PETS CC* and *EPFL Terrace*.

Recalling that the objective of multi-view pedestrian detection is to detect pedestrians in surveillance videos and estimate their 3D locations, we need to evaluate the detection results both on the camera views and the ground plane. Toward this end, we adopt two groups of evaluation criteria:

1. *RECALL/PRECISION*, which are used to evaluate the locations of the detected pedestrians on the ground plane. The PRECISION and the RECALL measures given by the ratios TP/(TP + FP) and TP/(TP +FN) respectively, where TP, FP and FN are the number of True Positive, False Positive and False Negative on the ground plane.
2. *MODA/MODP* [22](at an overlap threshold of 0.5), which are used to evaluate the detection results in different camera views. MODP measures the localization quality of the correct detections, MODA measures the detection accuracy taking into account both false and true correspondence. For both metrics, the larger value indicates a better performance.

6.2. Validation

This experiment aims at validating the two key developments of the paper: (1) the proposed MvBN can remove phantoms effectively while keeping most of pedestrians; (2) the HAP method can make our final detection results have a higher performance on different camera views.

The quantitative evaluation results are shown in Table 3. The "Pe-candidate" stands for pedestrian candidates produced by the base detection method (Reduction Dimension [4] in our experiments), which include pedestrians and phantoms. "MvBN only" stands for detection results by MvBN without the HAP method. In other words, the detection results of "MvBN only" in each view are got from projecting the each pedestrian's original location directly from the ground plane to each camera view with the predefined height (1.85 m). Compared with "MvBN only," the detection results of "MvBN +HAP" in camera views are generated by the HAP method.

As Table 3 shows, the RECALL of pedestrian candidates is really high, while the PRECISION is very poor. It means that pedestrian candidates contain most of pedestrians with a large number of phantoms. Compared with pedestrian candidates, MvBN has a much better performance in PRECISION while keeping a similar level in RECALL. It demonstrates that the proposed MvBN can remove phantoms effectively while keeping most of real pedestrians. In addition, "MvBN+HAP" has a same evaluation performance on RECALL and PRECISION with "MvBN only," because the

Fig. 11. There are several detection examples on different datasets and the last column is the detection results on the ground plane. The same pedestrian in different cameras has same color bounding box. (a) PETS CC, (b) PETS S2L1, (c) APIDIS, and (d) TERRACE.

HAP method does not change the detected pedestrians' locations on the ground plane. The performance gains on MODA and MODP indicate that the HAP method is more effective when generating detection results in camera views.

6.3. Robustness to the prior pedestrians' height

We continue discussion on the HAP method. All pedestrians are assigned to a fixed height as a predefined parameter of MvBN. In order to test the sensitivity of the detection results against the parameter, an experiment is conducted to show the performance variations of our approach using different predefined heights. During the experiment, we varied the height from 1.5 m to 2.4 m and series of tests were done with or without the HAP method. Fig. 10 shows the quantitative evaluation results, where x-axis stands for different predefined heights and y-axis is the mean MODA/MODP of all used views in the corresponding dataset. The green curve stands for evaluation results for MvBN with HAP method, while the red one is for MvBN without the HAP method. As is shown in Fig. 10, the detection results without the HAP method would be sensitive to the predefined height. The reason is that the predefined height is treated as the final height of all detected pedestrians. As expected, MvBN with the HAP method would get better evaluation performance and make the final detection results more robust to the predefined heights. Even if the height is unreasonable (such as 2.35 m), the HAP method could still obtain good evaluation performance. This is because the HAP method would adjust each pedestrian's height based on image observations rather than setting all pedestrians as the fixed predefined height.

6.4. Compare with the state-of-the-art algorithms

In this experiment, the proposed approach is compared with several state-of-the-art methods to demonstrate the effectiveness. As same as the above experiments, we use *MODA* and *MODP* to evaluate detection results in different camera views. To compare the proposed approach with the state-of-the-art methods on the ground plane, *RECALL/PRECISION* or *TER* is used to evaluate the detection results (O-LASSO only showed the RECALL/PRECISION evaluation results, while 3DMPP used TER [9] as the final evaluation metric). TER is used to measure the detection accuracy taking into account both false and true correspondence on the ground plane. For the TER metric, when value is less, performance is better. The comparison results are shown in Table 4. Fig. 11 shows several detection results of the proposed method.

In different datasets, all the evaluation results of the proposed method are reasonably high. From Table 4, we can see our approach outperforms all the other methods both in camera views (MODA and MODP) and 3D real world (TER or RECALL/PRECISION). Generally speaking, the main differences between our approach and other methods lie as follows:

1. The proposed method has a similar framework with POM [3] and O-Lasso [4]. These three methods model the multi-view pedestrian detection problem as different optimal problems and try to minimize the difference between the final detection results and the observed foreground masks. The main difference is that we utilize the occlusion relationship to model pedestrian candidates. The experimental results show that the occlusion relationship is effective to remove phantoms. In addition, we treat the input information with bias (i.e., diverse pedestrians' heights and calibration noise) and utilize the HAP method to refine detection results in camera views.
2. Multiview Sampler [8] also utilizes the occlusion relationship and considers the noisy input. Compared with [8], the proposed method depresses phantoms using multiple views information simultaneously rather than independently in each single view. Their method will make the final detection in a camera view lose the correspondence with other views.
3. Compare with 3DMPP [9], we utilize the statistical property of foreground pixels rather than the foregrounds in some crucial parts (feet and head), which makes our method more robustness to foreground errors.

7. Conclusion

This paper presents a novel approach for pedestrian detection in multiple cameras by removing phantoms from pedestrian candidates which can be produced by many existing methods. To remove phantoms, a Multi-view Bayesian network (MvBN) is constructed to model all pedestrian candidates and their occlusion relationship in all views. Such phantoms can be efficiently removed by inferring the nodes on MvBN that demonstrate the highest probabilities of occlusions. Moreover, we propose a height adaptive projection (HAP) method to generate final detection results in each view. By a local search process in a neighborhood of heights and locations of the detected pedestrians, our approach can make final detection results adaptive to pedestrians' heights and robustness to the noisy inputs. The experimental results have shown that our approach achieves a good performance on a variety of application scenarios, such as visual surveillance and sports analysis. Our approach has been shown to outperform other state-of-the-art algorithms. The influence of different components and the robustness to the inputs have been analyzed.

Conflict of interest

None declared.

Acknowledgments

This work is partially supported by grants from the National Natural Science Foundation of China under contract nos. 61035001 and 61390515.

References

[1] A. Sankaranarayanan, A. Veeraraghavan, R. Chellappa, Object detection, tracking and recognition for multiple smart cameras, Proc. IEEE 96 (10) (2008) 1606–1624. http://dx.doi.org/10.1109/JPROC.2008.928758.
[2] F. Fleuret, R. Lengagne, P. Fua, Fixed point probability field for complex occlusion handling, in: Tenth IEEE International Conference on Computer Vision, 2005, ICCV 2005, vol. 1, 2005, pp. 694–700.
[3] F. Fleuret, J. Berclaz, R. Lengagne, P. Fua, Multicamera people tracking with a probabilistic occupancy map, IEEE Trans. Pattern Anal. Mach. Intell. 30 (2) (2008) 267–282.
[4] A. Alahi, L. Jacques, Y. Boursier, P. Vandergheynst, Sparsity driven people localization with a heterogeneous network of cameras, J. Math. Imaging Vis. 41 (1–2) (2011) 39–58.
[5] S.M. Khan, M. Shah, A multiview approach to tracking people in crowded scenes using a planar homography constraint, in: Computer Vision-ECCV 2006, Springer, 2006, pp. 133–146.
[6] K. Kim, L.S. Davis, Multi-camera tracking and segmentation of occluded people on ground plane using search-guided particle filtering, in: Computer Vision-ECCV 2006, Springer, 2006, pp. 98–109.
[7] P. Peng, Y. Tian, Y. Wang, T. Huang, Multi-camera pedestrian detection with multi-view Bayesian network model, in: BMVC, 2012, pp. 1–12.
[8] W. Ge, R.T. Collins, Crowd detection with a multiview sampler, in: Computer Vision-ECCV 2010, Springer, 2010, pp. 324–337.
[9] A. Utasi, C. Benedek, A Bayesian approach on people localization in multi-camera systems, IEEE Trans. Circuits Syst. Video Technol. 23 (1) (2013) 105–115. http://dx.doi.org/10.1109/TCSVT.2012.2203201.
[10] J. Franco, E. Boyer, Fusion of multiview silhouette cues using a space occupancy grid, in: Computer Tenth IEEE International Conference on Vision 2005, ICCV 2005, vol. 2, 2005, pp. 1747–1753.

[11] M. Evans, L. Li, J. Ferryman, Suppression of detection ghosts in homography based pedestrian detection, in: 2012 IEEE Ninth International Conference on Advanced Video and Signal-Based Surveillance (AVSS), 2012, pp. 31–36. doi: http://dx.doi.org/10.1109/AVSS.2012.73.
[12] D. Delannay, N. Danhier, C. De Vleeschouwer, Detection and recognition of sports(wo)men from multiple views, in: Third ACM/IEEE International Conference on Distributed Smart Cameras, ICDSC 2009, 2009, pp. 1–7. doi:http://dx.doi.org/10.1109/ICDSC.2009.5289407.
[13] S.M. Khan, M. Shah, Tracking multiple occluding people by localizing on multiple scene planes, IEEE Trans. Pattern Anal. Mach. Intell. 31 (3) (2009) 505–519.
[14] R. Eshel, Y. Moses, Homography based multiple camera detection and tracking of people in a dense crowd, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2008, IEEE, 2008, pp. 1–8.
[15] R. Eshel, Y. Moses, Tracking in a dense crowd using multiple cameras, Int. J. Comput. Vis. 88 (1) (2010) 129–143.
[16] W. Ge, R. Collins, Marked point processes for crowd counting, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009, 2009, pp. 2913–2920. doi:http://dx.doi.org/10.1109/CVPR.2009.5206621.
[17] A. Utasi, C. Benedek, A 3-d marked point process model for multi-view people detection, in: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011, pp. 3385–3392. doi:http://dx.doi.org/10.1109/CVPR.2011.5995699.
[18] D. Arsic, E. Hristov, N. Lehment, B. Hornler, B. Schuller, G. Rigoll, Applying multi layer homography for multi camera person tracking, in: Second ACM/IEEE International Conference on Distributed Smart Cameras, ICDSC 2008, 2008, pp. 1–9. doi:http://dx.doi.org/10.1109/ICDSC.2008.4635731.
[19] B. Babenko, P. Dollár, Z. Tu, S. Belongie, et al., Simultaneous learning and alignment: multi-instance and multi-pose learning, in: Workshop on Faces in'Real-Life'Images: Detection, Alignment, and Recognition, 2008.
[20] Z. Zivkovic, Improved adaptive Gaussian mixture model for background subtraction, in: Proceedings of the 17th International Conference on Pattern Recognition, ICPR 2004, vol. 2, IEEE, 2004, pp. 28–31.
[21] A. Ellis, A. Shahrokni, J.M. Ferryman, Pets2009 and winter-pets 2009 results: a combined evaluation, in: 2009 Twelfth IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS-Winter), IEEE, 2009, pp. 1–8.
[22] R. Kasturi, D. Goldgof, P. Soundararajan, V. Manohar, J. Garofolo, R. Bowers, M. Boonstra, V. Korzhova, J. Zhang, Framework for performance evaluation of face, text, and vehicle detection and tracking in video: data, metrics, and protocol, IEEE Trans. Pattern Anal. Mach. Intell. 31 (2) (2009) 319–336.

Peixi Peng received the B.S. degree from Xian Jiaotong University in 2011, Xian, China. At present Peng is pursuing the Ph.D. degree in the School of Electronics Engineering and Computer Science at Peking University, China. His research interests include surveillance video analysis, multimedia learning.

Yonghong Tian is currently a professor with the National Engineering Laboratory for Video Technology, School of Electronics Engineering and Computer Science, Peking University, Beijing, China. He received the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, China, in 2005, and was also a visiting scientist at Department of Computer Science/Engineering, University of Minnesota from November 2009 to July 2010. His research interests include machine learning, computer vision, video analysis and coding, and multimedia big data. He is the author or coauthor of over 110 technical articles in refereed journals and Conferences. Dr. Tian is currently an Associate Editor of IEEE Transactions on Multimedia, and a Young Associate Editor of the FRONTIERS OF COMPUTER SCIENCE, a member of the IEEE TCMC-TCSEM Joint Executive Committee in Asia (JECA). He was the recipient of the Second Prize of National Science and Technology Progress Awards in 2010; the best performer in the TRECVID content-based copy detection (CCD) task (2010–2011); the top performer in the TRECVID retrospective surveillance event detection (SED) task (2009–2012); the winner of the WikipediaMM task in ImageCLEF 2008. He is a senior member of IEEE, a member of ACM.

Yaowei Wang received the M.Sc. degree from the Department of Computer Science, Heibei University of Technology, Tianjin, China, in 2000, and the Ph.D. degree from the Graduate School of the Chinese Academy of Sciences, Beijing, China, in 2005. He is currently a Lecturer in the School of Information and Electronics, Beijing Institute of Technology, Beijing, China. He is the author or coauthor of more than 30 technical articles in refereed journals and conferences. His current research interests include multimedia analysis and surveillance video analysis.

Jia Li received the B.S. degree in electronic engineering from Tsinghua University, Beijing, China, in 2005, and the Ph.D. degree in computer science from the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, in 2011. He is currently an Associate Professor with State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, China. He is the author or coauthor of more than 30 technical articles in refereed journals and conferences. His current research interests include visual attention/saliency modeling, multimedia analysis, and online video advertising.

Tiejun Huang is a professor of the School of Electronics Engineering and Computer Science, the director of the Institute for Digital Media, Peking University. Professor Huang received the Ph.D. degree on Pattern Recognition and Intelligent System from Huazhong (Central China) University of Science & Technology in 1998, master and bachelor degree on computer science from Wuhan University of Technology in 1995 and 1992. His research area includes video coding, image understanding, digital right management (DRM) and digital library. He published more than 130 peer-reviewed papers and three books as author or co-author. He is the member of the Board of Director for Digital Media Project, the Advisory Board of IEEE Computing Now and the Board of Chinese Institute of Electronics.