Multi-View Visual Surveillance and Phantom Removal for Effective Pedestrian Detection

Jie Ren[1], Ming Xu[2, 3], Jeremy S. Smith[3], Huimin Zhao[4*], Rui Zhang[5]

[1] College of Electronics and Information, Xi'an Polytechnic University, Xi'an, China

[2]Department of Electrical and Electronic Engineering, Xi'an Jiaotong-Liverpool University, Suzhou, China

[3] Department of Electrical Engineering and Electronics, University of Liverpool, Liverpool, U.K.

[4] School of Computer Science, Guangdong Polytechnic Normal University, Guangzhou, China

[5] Department of Mathematical Science, Xi'an Jiaotong-Liverpool University, Suzhou, China

*Corresponding author.

Abstract: To increase the robustness of detection in intelligent video surveillance systems, homography has been widely used to fuse foreground regions projected from multiple camera views to a reference view. However, the intersections of non-corresponding foreground regions can cause phantoms. This paper proposes an algorithm based on geometry and colour cues to cope with this problem, in which the homography between different camera views and the Mahalanobis distance between the colour distributions of every two associated foreground regions are considered. The integration of these two matching algorithms improves the robustness of the pedestrian and phantom classification. Experiments on real-world video sequences have shown the robustness of this algorithm.

Keywords: motion detection, video surveillance, homography.

1. INTRODUCTION

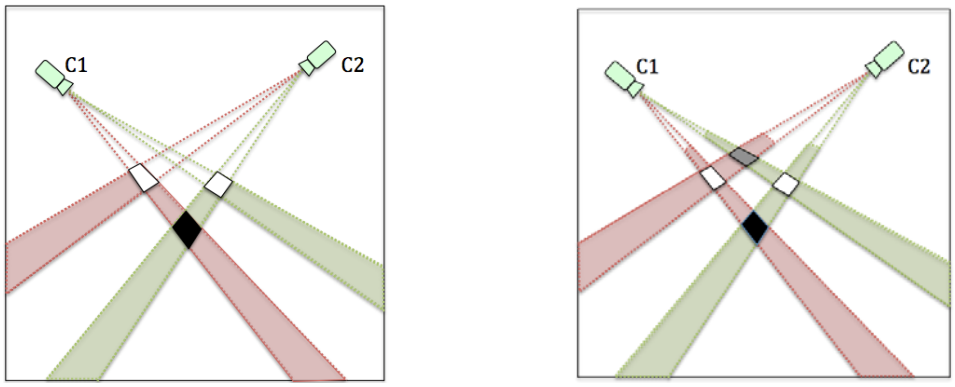
Intelligent visual surveillance is an active research area in artificial intelligence and computer vision. The aim of an intelligent visual surveillance system is to detect, track, classify objects and recognize events automatically. Moving object detection is an essential process before tracking and event recognition in video surveillance can take place. Using multiple cameras is a reasonable solution to occlusions, because when an object is occluded in one camera view, it may be visible in other camera views. Furthermore, multiple camera views can extend the overall field of view.

Since many cameras are used, an active research topic is how to utilize the information from the multiple cameras. According to the level of information fusion, the multi-camera approaches can be divided into three categories. The first category is the low-level information fusion which

starts tracking with a single camera view and switches to another camera when the system predicts that the current camera will no longer have a good view. In the intermediate level of information fusion, measurements, extracted features or tracked targets are first detected in the individual camera views and then integrated to obtain the global estimation. The third category no longer extracts features or tracks targets but provides foreground bitmap information from the individual camera views. The foreground information from all camera views is fused in the fusion centre and then detection and tracking based on the fused information are undertaken. The third category has emerged in recent years and belongs to the category of high-level information fusion. This approach can help object detection when the scene is crowded.

To associate camera views and to fuse information from all the camera views, one useful assumption is that in all camera views the objects of interest are on a common plane. This assumption is valid for most scenarios in intelligent visual surveillance systems. Then, homography, a geometric transformation which shows a pixelwise mapping between two views according to a common plane, can be used as an efficient method to associate multiple camera views. Using a homography transformation, foreground regions detected from each of the multiple camera views can be projected to a reference view according to the homography for a specific plane. The intersection regions of the foreground projections indicate the locations of moving objects on that plane. This method achieved good results in detection and is robust in coping with occlusion. In Khan and Shah's work [1], the foreground likelihood image, which is extracted from each of the multiple camera views, is warped to a reference camera view according to the ground-plane homography and overlaid with those from other camera views. A threshold is applied in the reference view to determine the locations of people on the ground plane. Then, the homographies for a set of parallel planes at different heights are employed to increase the robustness of the detection. This work achieves good results in moderately crowded scenes, because regions at the locations of true objects reinforce each other whereas the false locations are scattered around.

One problem with the homography approach is that the intersections of non-corresponding foreground regions can cause false-positive detections known as phantoms. Fig. 1 is a schematic diagram to illustrate how non-corresponding foreground regions intersect and give rise to false positives. The warped foreground region in the top view is observed as the intersection of the groundplane and the cones swept out by the silhouette of the underlying object. When the foreground regions for the same object are warped from multiple views to the top view, they will intersect at a location where the object touches the ground. However, if the warped foreground regions from different objects intersect in the top view, the intersection region will lead to a phantom detection. In Fig. 1 (a), the foreground regions of two objects are projected from two camera views into the top view. The foreground projections intersect in three regions on the ground plane. The white intersection regions are the locations of the two objects, whilst the black region may be a phantom. When homography mapping is based on a plane parallel to but higher than the ground plane, the projected foreground regions will move towards the cameras and additional phantoms may be generated. In Fig. 1 (b), the projected foregrounds from the two cameras intersect in four regions on plane *p* which is parallel to and off the ground plane. The grey region is an addition phantom.



(a) (b)

Fig. 1 A schematic diagram of phantom occurrence using (a) ground-plane homography and (b) a plane higher than the ground plane.

The research described in this paper is an extension of Khan and Shah's work, which focuses on multi-camera object detection using multi-layer homography mapping while solving false-positive detections. The main contributions of this paper are as follows:

1) To identify false-positive detections caused by the foreground intersections of non-corresponding objects in the top view, geometrical information of the foreground regions is utilised. A height matching algorithm is proposed to match each intersection region in the top view with its associated foreground regions in individual camera views and to identify whether the intersection region is due to the same object.

2) An appearance model and a colour matching algorithm are proposed, in which the colour similarity of the two foreground regions associated with each intersection region is calculated. In addition, the height matching algorithm and the colour matching approach are combined to further improve the robustness in the classification of the foreground intersection regions.

The remainder of this paper is organized as follows: In Section 2 the related work in this area is reviewed. The overall framework is presented in Section 3. In Section 4 foreground segmentation in individual camera views is introduced. Section 5describes the techniques used to estimate the homographies and how to apply the homographic transformations to the foreground regions in the individual camera views. Section 6 presents an integrated method in which the height matching and the colour matching are applied successively to identify whether each foreground intersection region in the top view is due to the same object or not. The experimental results are demonstrated in Section 7, followed by conclusions.

2. Related Work

There are a number of algorithms which aim to remove phantoms in foreground projection intersections. One solution is to avoid the generation of phantoms. Adding more cameras can provide a wider field of view and reduce the probability that an object is occluded in all views. Although additional cameras can reduce the sizes and number of phantoms, it is limited by the

cost of the additional cameras [9]. Stering et al. [10] applied the idea of generalized Hough voting in the homography projection. Hough voting relates all foreground probabilities to a position on the ground plane and restrains the shadow generation. However, the authors stated that they cannot handle the case when objects are invisible in all camera views.

Since phantoms are often gradually created and merge back into real objects, distinguishing and detecting them on the basis of positions is difficult [11]. Therefore, another approach often removes phantoms in the tracking process. In [12], Yang et al. pointed out that phantoms appear from nowhere and checked their temporal coherence to test if a foreground intersection region existed in the previous frame. Khan and Shah [1] also filtered out the phantoms according to the temporal coherence. In Liem and Gavrila's work, they assumed that phantoms are often unsteadily detected and checked the temporal coherence during a 'hidden' time rather than between every two consecutive frames. If such a candidate cannot survive over the hidden time in tracking, then it is classified as a phantom. They also proposed that a new object can only appear from the boundary of the overlapping field of views (FOVs); objects which are first detected in the middle of the overlapping FOVs are phantoms [11, 13].

The geometric approach is built on the comparison of features between phantoms and real objects. This approach can be further divided into two sub-classes: 3D space and 2D image methods, according to the types of geometric constraints that are used. The features applied in the geometric approach include heights and sizes. In the 3D space method, the comparison is in 3D space or in a virtual top view. Tong et al. [14] utilized foreground projection on multiple planes at different heights to removed phantoms. In [12], Yang et al. pointed out that the size of a phantom is often smaller than the minimum object size in the top view. However, this assumption is related to the height and viewing angle of the camera, and it does not work when a phantom region is covered by real objects in all camera views. Eshel and Moses [9] used the height information and assumed that the cameras are looking downwards. They found that if the viewing rays from two cameras intersect behind a true object, the phantoms are lower than the true object, while taller phantoms occur when the rays intersect in front of true objects. By limiting the heights of real objects within an appropriate range, they could remove some phantoms.

Some methods use the 2D information to identify phantoms. Arics and Hristov[6] warped the intersection regions from the top view back into each camera view and checked if they are totally covered by foreground regions. If the warped back regions are totally covered in all views, they are considered as phantoms. Peng et al. [15] learned an occlusion relationship by using a Bayesian network in each camera view and then removed phantoms according to a multi-view Bayesian network. In [2], a filtering algorithm is applied to remove covered pixels by checking whether a pixel on the virtual ground plane is occluded in all views. In Eshel and Moses's work [9], they applied the pixelwise intensity correlation between aligned frames in a reference view to remove phantoms. In [16], the foreground masks from all camera views are projected to a centroid plane to generate an occupancy likelihood map. The occupancy likelihood map is transformed to occupancy likelihood rays in the polar coordinate representation in each camera view, in which the origin of the polar coordinate is at the camera centre. The distance between the intersection region and the origin and the angle that each intersection region covered illustrate the depth information and the size of that intersection region. Then, the depth

information and covered angles are used to identify the occlusion relationship and remove phantoms.

3. PROPOSED FRAMEWORK

In this section, the framework of the proposed approach is outlined. Fig. 2 shows a flowchart of the proposed phantom removal algorithm based on both geometrical information and colour cues. Firstly, the foreground regions detected in each camera view are warped into a virtual top view according to the homography mapping for a plane at some height. The intersection regions indicate all the possible regions that contain real objects or phantoms. By assuming the pedestrians are standing upright, the intersection regions can be thought as the positions where pedestrians touch that plane.

As the matching is carried out in each camera view, the intersection regions in the top view are warped back to the individual camera views according to the ground-plane homography. For each foreground region in a camera view, the warped back patches corresponding to the same foreground region are grouped into a patch set for that foreground region. Height matching and colour matching are applied successively to identify whether each warped back patch in the patch set can match that foreground region.

The height matching is based on the position analysis between each foreground region and the warped back patches corresponding to that foreground region. The position analysis is derived from the observation that if an intersection region of the foreground projections from different camera views contains a real object, the warped back patch of that intersection region by using the ground-plane homography will be located at or above the bottom of that foreground region. If more than one warped back patches are matched to the same foreground region in the height matching, they are further classified in the colour matching.

The colour matching is based on the Mahalanobis distance between the colours of a pair of foreground regions each from a different camera view and intersecting in the top view. After the position analysis is applied to the patch classification in each camera view, the classification results from both camera views are integrated to classify the foreground intersection regions in the top view.

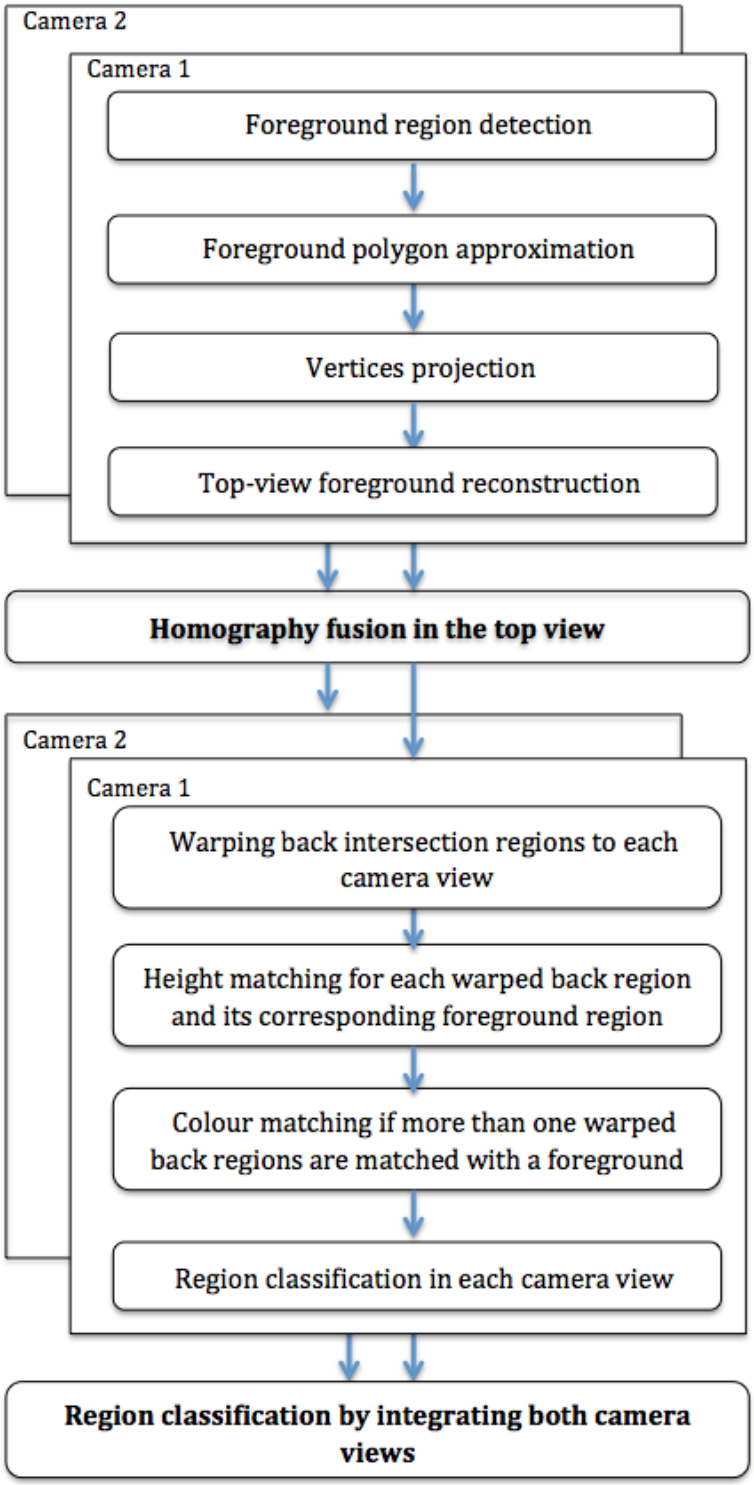


Fig. 2 Flowchart of the proposed phantom removal algorithm based on heights and colours.

4.FOREGROUND POLYGONS

The object detection starts with a single-camera foreground detection, in which a Gaussian mixture model and background subtraction are used to detect the foreground pixels in the

individual camera views. Then, the detected foreground pixels in each frame are grouped into foreground regions by applying connected component analysis, morphological operations and a size filter. Once the foreground regions have been identified in a camera view, the foreground regions need to be projected to a reference view. As a pixelwise homographic transformation is time consuming, each foreground region is approximated by the polygon of the foreground region's contour. The Douglas-Peucker (DP) method [17] has been used for the polygon approximation.

4.1 Foreground Segmentation in a Single View

As an essential process in visual surveillance systems, foreground segmentation aims to separate moving objects from a background image in each frame. The Mixture of Gaussians (MoG) model is a widely used method to cope with switching background elements (e.g., waving trees) [18]. Stauffer and Grimson[19] used a mixture of Gaussian distributions to model switching, multiple backgrounds. The sum of the probability density functions weighted by the corresponding priors represents the probability that a pixel is observed at a particular intensity or colour. KaewTraKulPong and Bowden[20] proposed an improved Mixture of Gaussians model which reduces the learning time and can remove moving shadows from foreground regions.

The colour value of each pixel is modelled by a mixture of K Gaussian distributions which are used to represent the variations of the background. Let $\mathbf{p}_t = [R_t G_t B_t]^T$ be the value of a pixel at time t, the probability of that pixel taking this value is:

$$P(\mathbf{p}_t) = \sum_{j=1}^{K} \frac{w_j}{(2\pi)^{\frac{d}{2}} \left|\boldsymbol{\Sigma}_j\right|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{p}_t-\boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}_j^{-1}(\mathbf{p}_t-\boldsymbol{\mu}_j)} \quad (1)$$

where d is the dimension of the colour value (currently d = 3), w_j is the weight, $\boldsymbol{\mu}_j$ is the temporal mean and $\sum_j$ is the covariance matrix for the j-th distribution. Let σ_j^2 be the trace of $\sum_j$. These K distributions are ordered according to w_j/σ_j^2, which means a distribution occurring frequently with low variation has a high rank. The first B ranked distributions, whose sum of weights is over a threshold T, are thought of as background models. After a new frame $\mathbf{I}_t$ arrives at time t, each pixel $\mathbf{I}_t(r,c)$ is compared with its background models. If it is more than 2.5 times the standard deviation away from all the B distributions, it is regarded as a foreground pixel.

$$F_t = \left\{(r,c): \left\|\mathbf{I}_t(r,c) - \boldsymbol{\mu}_{t-1,j}(r,c)\right\| > 2.5\sigma_{t-1,j}(r,c)\right\} \quad j \in [1,B] \quad (2)$$

If the pixel value is matched with one of the B background distributions, then the matched background distribution k is updated by incorporating the observed pixel value. The weights, means and standard deviations of the other K-1 distributions remain the same. The weight of each distribution is normalized by the sum of the new K weights. If the pixel value fails to match any of the K background distributions, it will be used to build a new distribution to replace the distribution which has the least weight in the K distributions. After the foreground pixels in each camera view are detected, these pixels are grouped into foreground regions by applying connected component analysis, morphological operations and a size filter.

4.2 Foreground Polygons

Once the foreground regions have been identified in a camera view, each foreground region is projected to a reference view according to the homography for a certain plane. Instead of applying a pixelwise homography mapping, the algorithm focuses on the vertices of each foreground polygon. Then, the image-level projection is replaced by the projection of a small number of the vertices of each foreground polygon.

Each foreground region in the foreground image can be represented by the contour of that foreground region. Let F_i^a be the *i*-th foreground region detected in camera view a. The contour of F_i^a is represented by an ordered set of *N* points $C_i^a = \{p_1, p_2, \ldots, p_N\}$ on the contour curve. The algorithm proposed by Suzuki and Abe [21] is used to extract the contour of each foreground region. To make the representation of the contour point set C_i^a more compact, the original contour is approximated by a polygon; that is, to find a subset of these contour points that can best represent the contour. The Douglas-Peucker (DP) algorithm [17] is used for the polygon approximation.

5. HOMOGRAPHY MAPPING

5.1 Homography Estimation

Planar homography is a special relationship, defined by a 3×3 transformation matrix **H** between a pair of captured images of the same plane with different cameras:

$$\mathbf{H} = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix} \quad (3)$$

Let (x, y) and (x', y') be a pair of corresponding points on that plane in the two images. $\mathbf{x} = [x, y, 1]^T$ and $\mathbf{x}' = [x', y', 1]^T$ are the homogeneous coordinates of those two points. They are associated by the homography matrix **H**:

$$\mathbf{x}' \cong \mathbf{H}\,\mathbf{x} \quad (4)$$

where $\cong$ denotes that the homography is given up to an unknown scalar.

The homography matrix **H** with eight unknowns can be recovered from at least four pairs of corresponding points in the two camera views. The more pairs of the corresponding points, the better the estimation of **H** obtained. In addition, the estimated homography matrix performs better if these points are homogenously distributed.

5.2 Fusion of Foreground Polygons

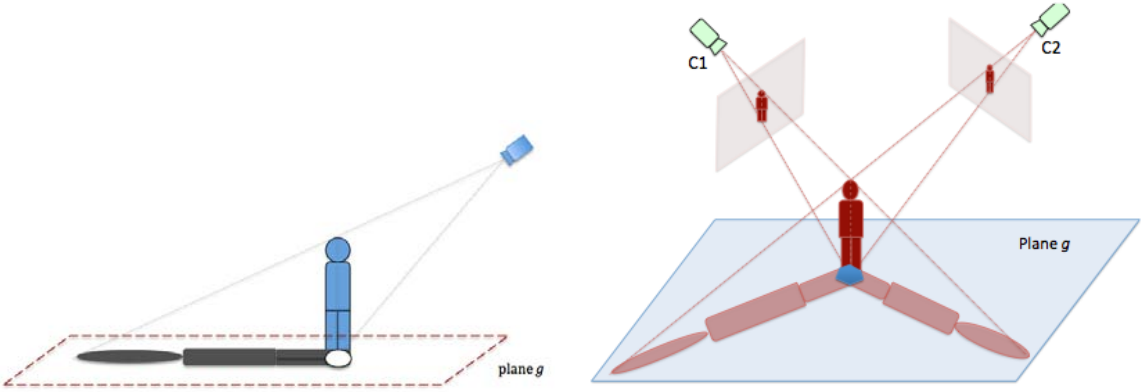
To improve the computational efficiency of the homography projection, the vertices of the foreground polygons in each camera view are projected into a virtual top view according to the homography for a certain plane. The ground-plane homography $H_g^{a,t}$ is used to project the

vertices V_i^a of the *i*-th foreground polygon from camera view *a* to the top view *t*. Let $V_{i,g}^{a,t}$ be the set of projected vertices in the top view *t*, which can be described as:

$$V_{i,g}^{a,t} = H_g^{a,t}(V_i^a) \quad (5)$$

Since the vertices in V_i^a are arranged in order, connecting each projected vertex with its neighbour sequentially can generated a new contour in the top view *t*. This new contour approximates the contour of the projected foreground region $F_{i,g}^{a,t}$ which is the projection of F_i^a from camera view *a* to the top view *t* according to the ground-plane homography. Then $F_{i,g}^{a,t}$ is rebuilt by filling the internal area of the projected foreground polygon with a fixed value using the ray-casting algorithm [22].

When each projected foreground polygon is reconstructed, the results approximate the bitmap projection of the foreground image F^a from camera view *a* to the top view *t*, which is denoted as $F_g^{a,t}$.The warped foreground region of an object in the top view is observed as the intersection of the ground-plane and the cones swept out by the silhouette of that object. Fig. 3 (a) illustrates the homography projection based on a single camera view according to the ground plane *g*. If the camera is considered as a light source, the grey region which is the projected foreground region is like the shadow of the blue object on plane *g*.



(a) (b)

Fig. 3 A schematic diagram of the homography projection according to the ground plane: (a) from a single camera view and (b) from two camera views (the overlaid foreground projections and the intersection region).

The foreground projection is extended from a single camera to multiple cameras. Let lower-case *c* be the index of the cameras. The fusion of the projected foregrounds in the top view is carried out by overlaying the projected foreground images from the multiple camera views:

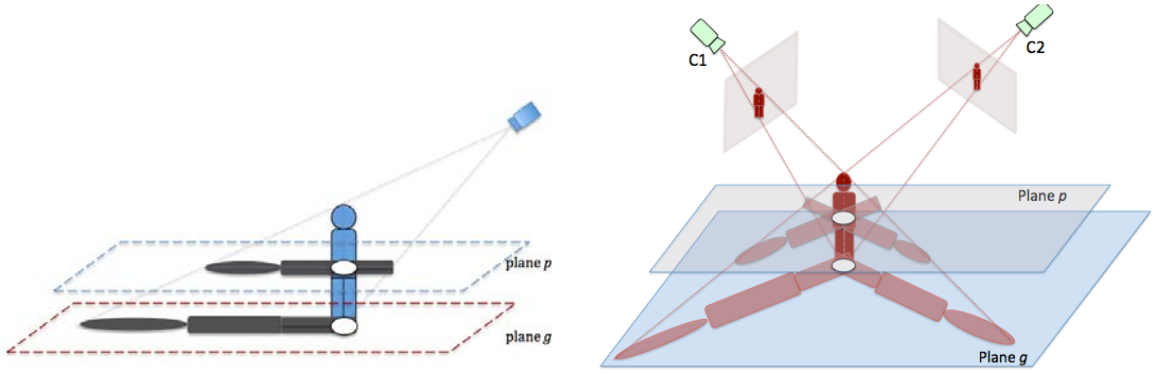
$$F_g^t = \sum_c F_g^{c,t} \quad (6)$$

The projected foreground regions from different camera views may intersect in the top view. The intersection regions correspond to enhanced regions in the overlaid foreground projection image F_g^t and indicate the locations of moving objects on the ground plane. The intersection regions are denoted by:

$$P_g^t = \bigcap_c F_g^{c,t} \quad (7)$$

When the foreground regions for the same object are warped from multiple views to the top view, they will intersect at a location where the object touches the ground. Fig. 3(b) shows a schematic diagram of the overlaid foreground projections and the intersection region. Although the intersection region corresponds to the location of the object on plane *g*, the size and the shape of that region is not exactly the cross section of the object.

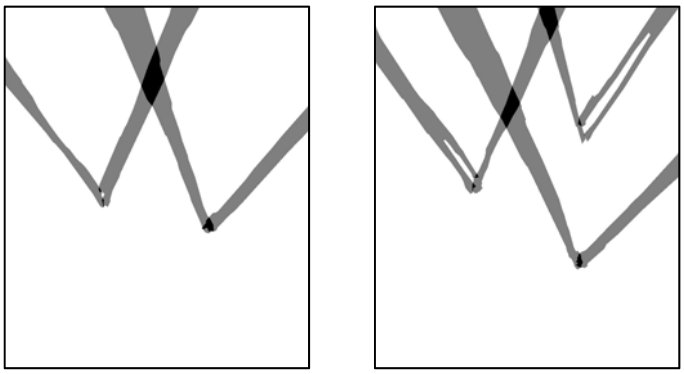
To improve the robustness of the object detection, the foreground projection and the fusion in the top view can be extended from the ground plane to a set of parallel planes. Fig. 4 shows a schematic diagram of the homography projection according to a plane parallel to the ground plane. Plane *p* is an imaginary plane parallel to the ground plane *g* and at the height of a person's waist. In Fig. 4 (a), the projected foreground region in plane *p* moves to the camera when the height of plane *p* increases. In Fig. 4 (b), the projected foregrounds from the two camera views to plane *p* intersect at the waist of the person.



(a) (b)

Fig. 4 A schematic diagram of the homography projection according to a plane parallel to the ground plane: (a) from a single camera view and (b) from two camera views.

Although the ground plane is the most commonly used plane in homography mapping, the foreground projections of the same object, each from one of multiple camera views, may have missed intersections in the reference view. This may happen in at least three scenarios. Firstly, pedestrians' feet are quite small objects and are frequently missed in detection, when a pedestrian is striding and hence has his two legs separated. Furthermore, their feet are not necessarily touching the ground while they are walking. Finally, homography estimation errors are another reason for missed intersections. These are illustrated in Fig. 5. Fig. 5 (a) shows an example of missed intersections due to inaccurate foreground detection when the homography mapping based on the ground plane is applied. Fig. 5 (b) is another example for missed intersections when one foot of a pedestrian is not touching the ground.



(a) (b)

Fig. 5 Examples of missed intersections by using ground-plane homography mapping.

6. PHANTOM REMOVAL

When the foreground images in the individual camera views are projected into the top view according to the homography for the ground plane or a plane parallel to the ground plane and at some height, the foreground regions from the different camera views may intersect in the top view, in which the intersections indicate the regions which may contain objects. If the intersecting foreground regions from the different camera views correspond to the same object, the intersection region reports the location where the object touches the plane used in the homography projection. If the intersection regions are caused by non-corresponding foreground regions from different camera views, they are false positive detections or phantoms. This is an important problem in multi-camera object detection using foreground homography mapping.

The objective of the research described in this paper is to identify the false-positive detections. A height matching algorithm is proposed to identify the false-positive detections, which is based on the geometry between the individual camera views. However, when two or more objects are close to each other in one camera view, the warped back patches of these objects may be close to the feet of the same pedestrian in another camera view. This brings difficulties to the Nearest-Neighbourhood based height matching, when there exist homography estimation errors and foreground detection errors. On the other hand, colour is a strong cue to distinguish between objects. The colours of foreground regions in the individual camera views can be used to identify whether each intersection region in the top view is due to the same object or not. In this paper, the colours of each foreground region are used to build an appearance model and a colour matching algorithm based on the Mahalanobis distance is applied to calculate the similarity of two associated foreground regions in their colour.

6.1 Patch Sets

Given the foreground region F_i^a from camera view a and F_j^b from camera view b, $F_{i,p}^{a,t}$ and $F_{j,p}^{b,t}$ are the projected foreground regions from the two camera views to the top view according

to the homographies $H_p^{a,t}$ and $H_p^{b,t}$ for the waist plane. Then the foreground projections are overlaid in the top view. If the two projected foreground regions intersect in the top view, they and their original foreground regions in the individual camera views are defined as a pair of projected foreground regions and a foreground region pair respectively. The intersection region of the projected foreground regions $F_{i,p}^{a,t}$ and $F_{j,p}^{b,t}$ are denoted as:

$$P_{i,j,p}^{t} = F_{i,p}^{a,t} \cap F_{j,p}^{b,t} = \left(H_p^{a,t}(F_i^a)\right) \cap \left(H_p^{b,t}(F_j^b)\right) \quad (8)$$

If the intersection region $P_{i,j,p}^{t}$ is formed by an object, it indicates the location where the object is intersected by plane p. When plane p is at different heights and parallel to the ground plane, the intersection region $P_{i,j,p}^{t}$ varies in its size and shape, which approximates the widths of the corresponding body parts at different heights. Assuming that pedestrians are standing upright, the ground plane and D virtual planes at different heights are considered. Let h be the height of plane p with a height range of $[0, 2]$ metres. $\{P_{i,j,p}^{t}\}_{p \in [0,D]}$ represents a set of foreground intersection regions at different heights but at the same location in the top view. When $P_{i,j,p}^{t}$ with different h values are projected onto the ground plane, they are at the same position in the ground plane. Therefore, $P_{i,j,p}^{t}$ can be observed at the location where the object touches the ground. Then, for the intersection region $P_{i,j,p}^{t}$, the index p of the plane can be removed.

Since the phantom classification is based on each camera view, each intersection region in the top view is warped back to the individual camera views first. Given an intersection region $P_{i,j}^{t}$ in the top view, the image patch in camera view a, which is warped back from the top view using the ground-plane homography, is as follows:

$$P_{i,j}^{a} = \left(H_g^{a,t}\right)^{-1}\left(P_{i,j}^{t}\right) \quad (9)$$

For each foreground region in camera view a, the image patches which are warped back on that foreground region are grouped into a patch set of that foreground region. For example, if the i-th foreground region in camera view a is F_i^a and the J foreground regions in camera view b are $\{F_j^b\}_{j \in [1,J]}$, there will be up to J intersection regions $\{P_{i,j}^{t}\}_{j \in [1,J]}$ in the top view, which are associated with F_i^a. When these intersections $\{P_{i,j}^{t}\}_{j \in [1,J]}$ are warped back into camera view a, the image patches $\{P_{i,j}^{a}\}_{j \in [1,J]}$ is defined as the patch set corresponding to the foreground region F_i^a in camera view a.

6.2 Height Matching in a Single View

In the height matching algorithm, geometrical relationships are utilized to identify the top-view intersection regions that are due to corresponding pedestrians in the individual camera views. The foreground correspondence is determined by comparing the bottom of a foreground region and the warped back patches associated with that foreground region in an individual camera view.

6.2.1 Normalized Distances

The normalized distance is the distance between the centroid of a warped back patch and the bottom of that patch's corresponding foreground region in a camera view. Given a foreground region F_i^a and a warped back patch $P_{i,j}^a$ whose corresponding foreground region in camera view a is F_i^a, the distance between the centroid of $P_{i,j}^a$ and the bottom of F_i^a is denoted as $h_{i,j}^a$. To remove the perspective effects, $h_{i,j}^a$ is normalized by h_i^a, which is the height of F_i^a:

$$d_{i,j}^a = \frac{h_{i,j}^a}{h_i^a} \tag{10}$$

The normalized distance $d_{i,j}^a$ indicates the likelihood that $P_{i,j}^a$ is located around the foot area of F_i^a and that $P_{i,j}^a$ contains an object. Fig. 6 shows a schematic diagram of how to calculate the normalized distance in camera view a. $h_{i,j}^a$ can be either positive or negative. When $P_{i,j}^a$ is located below the bottom of F_i^a, $h_{i,j}^a$ has a negative value, otherwise it has a zero or positive value. Therefore, the range of the normalized distance $d_{i,j}^a$ is from a negative value to 1.

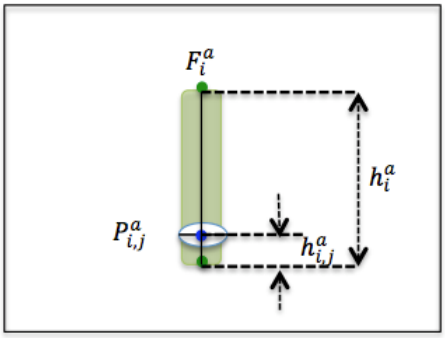


Fig. 6 A schematic diagram of height matching in a camera view.

6.2.2 Height Matching of a Patch Set

Given a patch set $\{P_{i,j}^a\}_{j\in[1,J]}$ for the i-th foreground region in camera view a, the normalized distance of each patch in $\{P_{i,j}^a\}_{j\in[1,J]}$ is calculated and the normalized distance set is denoted as $\{d_{i,j}^a\}_{j\in[1,J]}$. The patches which have normalized distances within a threshold and the number of such the patches are:

$$J_i^a = \{j: |d_{i,j}^a| \le Th_d, \quad j \in [1,J]\} \tag{11}$$

$$n_i^a = \#J_i^a \quad (12)$$

Ideally, only one patch in $\{P_{i,j}^a\}_{j\in[1,J]}$ should be located around the foot area of F_i^a and be recognized as the correct match of F_i^a. However, when two or more objects are very close to each other in one camera view, the warped back patches of these objects may be close to the feet of the same foreground region simultaneously. According to the value of n_i^a, the height matching can be divided into three pathways. If $n_i^a = 1$, there is only one matched patch in $\{P_{i,j}^a\}_{j\in[1,J]}$; the normalized distance $d_{i,j}^a$ of that matched patch $P_{i,j}^a$ is selected as the matched height of foreground region F_i^a and $d_i^a = d_{i,j}^a$. The matched height will be used to decide upper patches and lower patches afterwards. If $n_i^a = 0$, the matched height of foreground region F_i^a is set to zero and $d_i^a = 0$. If $n_i^a > 1$, the n_i^a patches will be further classified in the colour matching.

6.3 Colour Matching

Since colour is a strong cue to differentiate objects, the colours of the foreground regions in individual camera views are utilized to identify whether two foreground projections from different camera views are due to the same object. The first step of colour matching is to generate the appearance model of each foreground region. Then, the colour similarity of two foreground regions, which intersect in the top view, is measured according to the Mahalanobis distance of their appearance models.

6.3.1 Appearance models

The appearance model of each torso region is built by using the colours of all the pixels in the torso region. The torso region is defined as the part of a foreground region, which is within a specified range of heights. The torso region is used because it has a large area and can provide stable colour cues and discriminative features to distinguish between pedestrians. To handle the multiple colours in the torso region, the appearance model is developed by using the Gaussian mixture model. The K-means algorithm[23] and the Expectation-Maximization (EM) algorithm[24] are widely used to find the parameters of the probability density functions in a Gaussian mixture model. Since the clustering of the pixels in a torso region is based on the hue component, the cyclic property of the hue is considered. Given a foreground region F_i^a, the colour appearance of the torso region A_i^a is modeled by K Gaussian distributions: $\mathcal{N}(\pi_{i,n}^a, \mu_{i,n}^a, \Sigma_{i,n}^a)$, $n \in [1, K]$, where $\pi_{i,n}^a$, $\mu_{i,n}^a$ and $\Sigma_{i,n}^a$ are the weight, mean and covariance of the *n-th* Gaussian distribution. The K Gaussians are ordered according to the magnitudes of the weights and $\pi_{i,1}^a$ is the greatest weight.

6.3.2 Appearance matching with a single patch

After the appearance model of each torso region is constructed, the Mahalanobis distance is calculated to measure the colour similarity between every two foreground regions which have intersecting projections in the top view. Given another foreground region F_j^b, its torso region

A_j^b can also be modeled by K Gaussian distributions: $\mathcal{N}(\pi_{j,m}^b, \mu_{j,m}^b, \Sigma_{j,m}^b)$, $m \in [1, K]$, where $\pi_{j,m}^b$, $\mu_{j,m}^b$ and $\Sigma_{j,m}^b$ are the weight, mean and covariance of the *m-th* Gaussian distribution respectively. A cross matching method can be used to calculate the Mahalanobis distance between A_i^a and A_j^b. Since the Gaussian distributions in each GMM are ranked in a descending order, the first distribution is always the dominant distribution in the GMM. Ideally, only the dominant distribution from each GMM should be involved in the colour matching. However, it is often necessary to consider some non-dominant distributions in the colour matching, when the underlying torso region lacks a dominant colour or its dominant colour is partly occluded by a non-dominant colour in another camera view. An example for the former scenario is a textured T-shirt. An example of the latter scenario is a red T-shirt in one camera view, which is partly occluded by an arm in the other view. The distributions used in the cross matching are decided by the weights of the individual distributions which must be above a threshold. For the torso region A_i^a, the Gaussian distributions involved in the cross matching, which are called significant distributions, are represented by a set and the number of such a set is:

$$N_i^a = arg \max_{n \in [1,K]} \{\pi_{i,n}^a : \pi_{i,n}^a \geq T_g\} \quad (13)$$

The colour matching between the torso region A_i^a in camera view a and the torso region A_j^b in camera view b is carried out in three steps. In the first step, the Mahalanobis distances between the dominant distribution $\mathcal{N}(\pi_{i,1}^a, \mu_{i,1}^a, \Sigma_{i,1}^a)$ of A_i^a and all the significant distributions of A_j^b, $\mathcal{N}(\pi_{j,m}^b, \mu_{j,m}^b, \Sigma_{j,m}^b)$, $m \in [1, N_j^b]$, are calculated:

$$c_{i,j,m}^a = (\mu_{i,1}^a - \mu_{j,m}^b)^{\mathrm{T}} (\Sigma_{i,1}^a + \Sigma_{j,m}^b)^{-1} (\mu_{i,1}^a - \mu_{j,m}^b) \quad (14)$$

In the second step, the Mahalanobis distances between the dominant distribution of A_j^b and all the significant distributions of A_i^a are calculated. The result is denoted as $c_{i,j,n}^b$, $n \in [1, N_i^a]$. Then the Mahalanobis distances between A_i^a and A_j^b is a combination of $\mathrm{c}_{i,j}^a = \{c_{i,j,m}^a\}_{m \in [1,N_j^b]}$ and $\mathrm{c}_{i,j}^b = \{c_{i,j,n}^b\}_{n \in [1,N_i^a]}$:

$$\mathrm{S}_{i,j}^{a,b} = \mathrm{c}_{i,j}^a \cup \mathrm{c}_{i,j}^b \quad (15)$$

where $S_{i,j}^{a,b}$ can be rewritten as $S_{i,j}^{a,b} = \{S_{i,j,k}^{a,b}\}_{k \in [1,L]}$; the number of the Mahalanobis distances L is $(N_i^a + N_j^b - 1)$.Then, the minimum value in $\{S_{i,j,k}^{a,b}\}_{k \in [1,L]}$ is thought of as the colour distance between the pair of colour appearance models:

$$c_{i,j}^{a,b} = \min_{k \in [1,L]} (S_{i,j,k}^{a,b}) \quad (16)$$

6.3.3Appearance matching with multiple patches

The Mahalanobis distance between the appearance models of a pair of foreground regions reflects the likelihood that these two foreground regions, each in a different camera view, are coming from the same object. For a set of warped back patches $\{P_{i,j}^a\}_{j \in J_i^a}$ in camera view a, the Mahalanobis distances of these patches are $\{c_{i,j}^{a,b}\}_{j \in J_i^a}$. The patch that has the least Mahalanobis distance is identified as the matched patch.

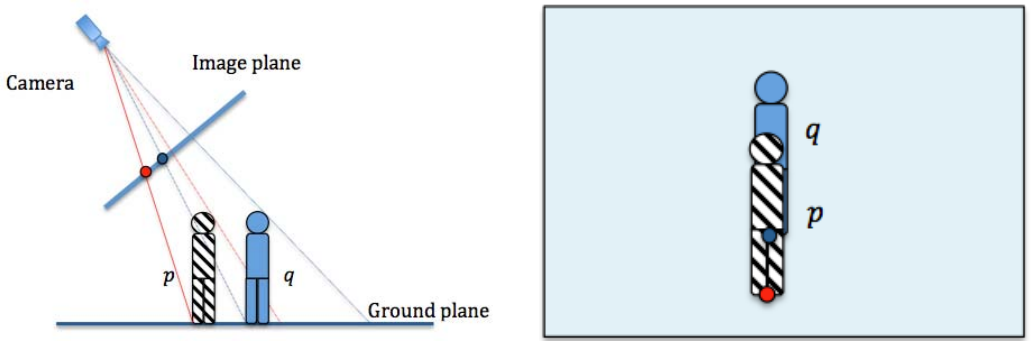
$$j_{min} = \arg\min_{j \in J_i^a} \{c_{i,j}^{a,b}\} \quad (17)$$

Then, its normalized distance $d_{i,j_{min}}^a$ is used as the matched height of F_i^a.

6.4 Region Classification

6.4.1 Position Analysis

Position analysis is applied in the patch classification in the individual camera views. It is based on the normalized distances of the patches. In the position analysis, the camera is assumed to be viewing downward. Therefore the vanishing point is in the direction of positive infinity in the image coordinates. This assumption is satisfied in most visual surveillance systems. According to the projective geometry, if an object is closer to the camera in the top view, that object will move downward in the direction of positive infinity in that camera view. Fig. 7 shows a schematic diagram of the position analysis in a camera view. There are two objects p and q on the ground plane in Fig. 7 (a), in which object p is closer to the camera than object q. They are in the same ray passing through the camera centre. Therefore, object p is located below object q and may partly occlude object q in the camera view (Fig. 7 (b)).



(a) (b)

Fig. 7 A schematic diagram of position analysis in a camera view.

6.4.2 Patch classification in a single view

During the height matching in a camera view, the warped back patch which matches the foreground region for the patch set is identified and the matched height is determined. The normalized distances of the other patches in the same patch set are compared with the matched height to decide whether the other patches are above or below the matched patch. The patches in that patch set can be divided into three categories: the object patch (Op), upper patches (Up) and lower patches (Lp). The object patch 'Op' corresponds to the foot location of an object which is visible in that camera view. If the normalized distance of a patch is greater than the matched height, then that patch is identified as an upper patch (Up), which corresponds to an intersection region behind that for an object. If the normalized distance of a patch is less than the matched height, then that patch is identified as a lower patch (Lp), which corresponds to a foreground intersection region in front of that for an object.

6.4.3 Region classification in both views

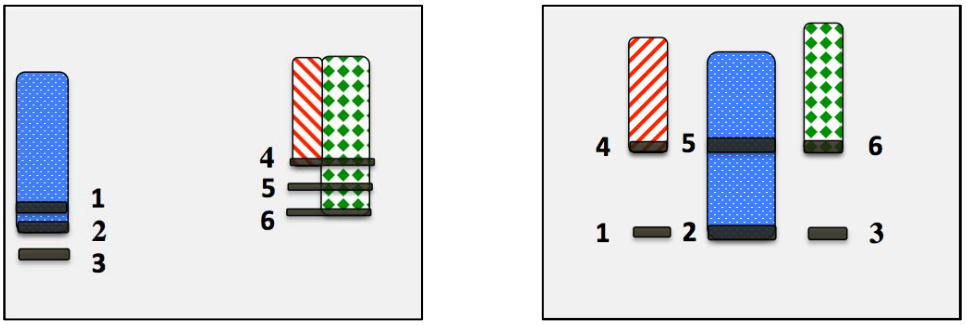
After the warped back regions are classified in a single camera view, the classification results from both views are incorporated to classify the intersection regions in the top view. The intersection regions in the top view are classified into four categories: object regions (Ob), occluded regions (Oc), covered regions (Cv) and phantoms (Ph). Table 1 summarizes the classification of the intersection regions from the two camera views.

Table 1 Classification of the intersection regions from two camera views.

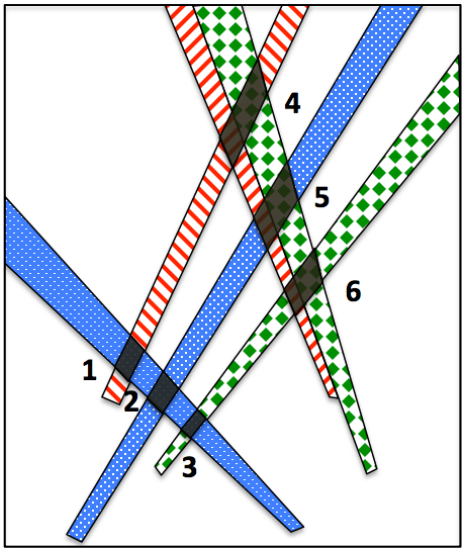
Camera View *a* / Camera View *b*	Op	Up	Lp
Op	Ob	Oc	Ph
Up	Oc	Cv	Ph
Lp	Ph	Ph	Ph

In Table 1, if the warped back patches of an intersection region are identified as object patches in both camera views, that intersection region contains an object (Ob) and is visible in both camera views. If an intersection region is identified as an object patch in one camera view and an upper patch in the other view, the corresponding object is visible in the first camera view and occluded in the second camera view. As a result, that intersection region is classified as an occluded region (Oc). When an intersection region is identified as an upper patch in both camera views, it is labelled as a covered region (Cv), because it may either be a phantom or contain a real object behind other objects. If the warped back patch for an intersection region is identified as a lower patch in either camera view, it is determined as a phantom (Ph).

Fig. 8 shows an example of the position analysis in the top view. In Fig. 8 (a) and (b), there are three objects which are illustrated in red stripes, green squares and blue dots in each camera view. The objects with the same colour and pattern in the two camera views are the same object. Since the green square object occludes the red striped object in Fig. 8 (a), they are grouped into a single foreground region. The foreground regions are projected from the two camera views to the top view according to the homography for a plane parallel to the ground plane and at some height. The two foreground regions from camera view *a* and three foreground regions from camera view *b* intersect in 6 regions in the top view. Fig. 8 (c) shows the overlaid foreground projections and darker intersection regions which are labelled with 1 to 6 in the top view. The ground-truth object locations are intersection regions 2, 4 and 6. The warped back patches of these intersection regions from the top view to the individual camera views according to the ground-plane homography are the black patches in Fig. 8 (a) and (b). Each warped back patch is given the same label as the corresponding intersection region in the top view.



(a) (b)



(c)

Fig. 8 A schematic diagram of the position analysis in two camera views, (a) warped back patches in camera view *a*, (b) warped back patches in camera view *b*, and (c) overlaid foreground projections in the top view.

Intersection region 1, which corresponds to an upper patch in Fig. 8 (a) and a lower patch in Fig. 8 (b), is a phantom region. Intersection region 3 is a phantom, as its warped back patches are lower patches in both camera views. The warped back patches of intersection regions 2 and 6 are located at the foot area of the corresponding foreground objects in the two camera views. Those intersection regions are object regions and indicate the locations of the blue dot object and the green square object in the top view. Intersection region 4 is an occluded region, because its warped back patch is an object patch in camera view *b* but is an upper patch in camera view *a*, indicating that intersection region contains an object but is occluded by another object in camera view a. Intersection region 5 is an upper patch in both camera views and corresponds to a covered region. It is occluded by the green squared object at intersection region 6 in camera view *a* and the blue dot object at intersection region 2 in camera view *b*.

The details of the phantom pruning algorithm are described as **Algorithm 1**.

Algorithm 1: Phantom Pruning

1:	**for** each camera view **do**
2:	each intersection region $P_{i,j}^{t}$ of foreground projections in the top view are warped back to the camera view by using the homography for the ground plane;
3:	**for** each foreground region in the camera view (using a as the index of the camera) **do**
4:	the patch set $\{P_{i,j}^{a}\}_{j\in[1,J]}$ of F_i^a is generated;

5:	**for** each of the warped back patche in $\{P_{i,j}^a\}_{j\in[1,J]}$ **do**
6:	calculate the normalized distance $d_{i,j}^a$;
7:	$J_i^a = \{j: \lvert d_{i,j}^a \rvert \leq Th, j \in [1,J]\}$ $n_i^a = \#J_i^a$
8:	**if** $n_i^a = 0$ **then**
9:	the matched height $d_i^a = 0$;
10:	**else if** $n_i^a = 1$**then**
11:	$d_i^a = d_{i,j}^a$;
12:	**else begin**
13:	**for** each patch $P_{i,j}^a$ that satisfy $j \in J_i^a$ **do**
14:	calculate the Mahalanobis distance $c_{i,j}^{a,b}$ between A_i^a and A_j^b;
15:	**end for**
16:	Patch $P_{i,j_{min}}^a$, where $j_{min} = arg\min_{j\in J_i^a}\{c_{i,j}^{a,b}\}$, is identified as the object patch for F_i^a; $d_i^a = d_{i,j_{min}}^a$;
17	**end**
18:	**end if**
19:	**If** $d_{i,j}^a == d_i^a$, **then** $P_{i,j}^a$ is labeled as Object Patch (Op);
20:	**else if** $d_{i,j}^a > d_i^a$, **then** $P_{i,j}^a$ is labeled as Upper Patch (Up);
21:	**else** $P_{i,j}^a$ is labeled as Lower Patch (Lp);
22:	**end if**
23:	**end for**
24:	**end for**
25:	Classification of the foreground intersection regions based on integration of the results from all camera views.
26:	**end for**

7 EXPERIMENTAL RESULTS

The phantom removal algorithm has been tested over a range of video sequences. The dataset was captured in the author's campus, where the cameras were placed close to pedestrians. People walked around within a $4.0m \times 2.4m$ region to ensure some degree of occlusion. There

are 2790 frames captured in each camera view with a resolution of 640 × 480 pixels and a frame rate 15 fps. The test of the phantom removal algorithm was evaluated over 142 frames, each of which was periodically sampled from 2155 frames of the testing video (the first 660 frames contain no pedestrians or only one pedestrian).

In these experiments, the homography mapping was based on a plane parallel to the ground plane and at a height of one metre, which is at the waist level of the pedestrians of average height. Each foreground polygon in a camera view was warped to the top view according to the homography for this plane. A threshold was applied to the overlaid foreground projections in the top view. The approximated polygon of each intersection region in the top view was warped back into the individual camera views according to the ground-plane homography. The distance between the warped back patch and the bottom of its associated foreground region is calculated in each camera view. The location of each warped back patch in a single camera view is represented by its centroid. If the normalised distance was less than 0.1, the warped back patch was thought of as being located near the foot area of its associated foreground region.

The height matching algorithm and the colour matching algorithm are combined to improve the robustness of classification. The colour matching method uses the colours of foreground regions in the individual camera views to identify whether each intersection region in the top view is due to the same object or not. Fig. 9 shows the procedure of the phantom removal algorithm using the height matching and colour matching at frame 1200. Fig. 9 (a)-(d) are the original images and the results of foreground detection in the two camera views. In each camera view, there are three pedestrians which are labelled with 1 to 3 in camera view *a* and are labelled with a to c in camera view *b*. Fig. 9 (e) shows the overlaid foreground projections from the two camera views to the top view with the homography for a plane at a height of one metre. Their intersection regions in the top view are shown in Fig. 9 (f). Fig. 9 (g) and (h) are the warped back patches overlaid in the original camera views. Each intersection region and its corresponding warped back patches are given a similar label to indicate the corresponding foreground regions in both camera views. The torso regions in the two camera views are illustrated in Fig. 9 (i) and (j).

(a) (b) (c) (d)

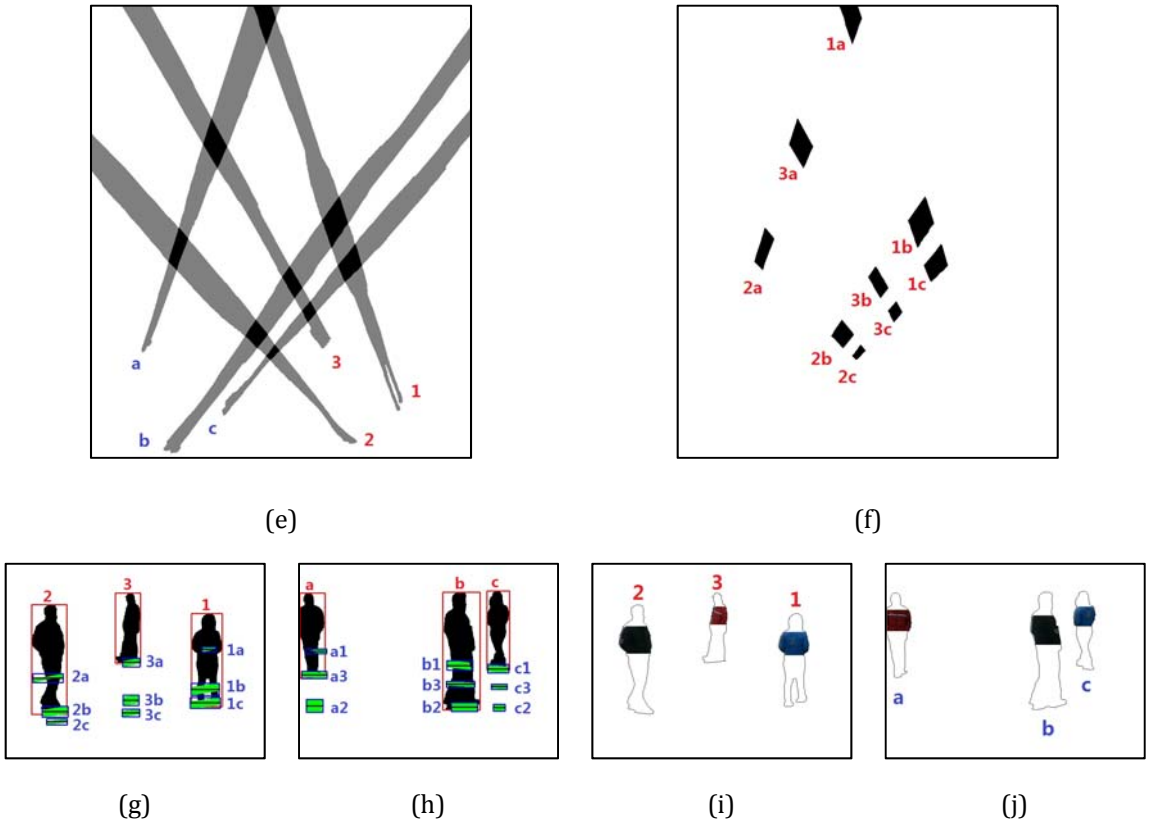


Fig. 9 The process of phantom removal using the height matching and colour matching at frame 1200, (a)(b) the original images in two camera views, (c)(d) the foreground regions, (e) the overlaid foreground projections in the top view, (f) the intersection regions in the top view, (g)(h) the warped back patches in the two camera views, and (i)(j) the torso regions in the two camera views.

Table 2 and Table 3 show the results of the height matching and colour matching for the warped back patches in the two camera views. For foreground region 1 in camera view *a*, it is related to three warped back patches labelled as 1a, 1b and 1c. Patch 1c, which has the minimal normalized distance less than the threshold 0.1, is identified as an object patch. Patches 1a and 1b which have normalized distances larger than that for patch 1c are recognized as upper patches. For foreground region 3 in camera view *a*, patch 3a is identified as an object patch. Patches 3b and 3c are identified as lower patches because their normalize distances are less than that of patch 3a. Since warped back patches 2b and 2c have normalized distances less than the threshold 0.1, colour matching is applied to further identify which may contain a real object. Then, patch 2b which has a lower colour distance is selected as the object patch of foreground region 2 in camera view *a*. The other patches in the two camera views can be classified using the height matching only.

Table 2 Height matching and colour matching at frame 1200 in camera view *a*.

Foreground Region in Camera View *a*	Foreground Region in Camera View *b*	Normalized Distance	Colour Distance	Classification Result

1	a	0.613	3540780.00	Up
	b	0.187	11784.40	Up
	c	**0.043**	16.32	Op
2	a	0.332	460419.00	Up
	b	**0.026**	**22.41**	Op
	c	**-0.067**	4742.66	Lp
3	a	**0.012**	179.88	Op
	b	-0.523	499446.00	Lp
	c	-0.701	1.371490.00	Lp

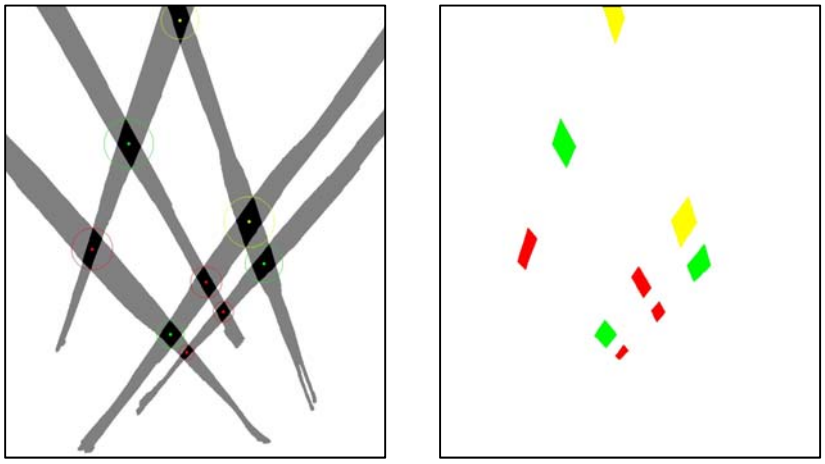
Table 3 Height matching and colour matching at frame 1200 in camera view *b*.

Foreground Region in Camera View *b*	Foreground Region in Camera View *a*	Normalized Distance	Colour Distance	Classification Result
a	1	0.310	3540780.00	Up
	2	-0.329	460419.00	Lp
	3	**0.038**	179.88	Op
b	1	0.378	11784.40	Up
	2	**0.024**	22.41	Op
	3	0.220	499446.00	Up
c	1	**0.015**	16.32	Op
	2	-0.477	4742.66	Lp
	3	-0.210	1371490.00	Lp

The classification results in the two camera views are combined to make a final decision according to Table 1. Table 4 shows the classification results of the intersection regions. To visualize the classification results, in **Error! Reference source not found.**, each intersection region in the top view is filled with a different colour, in which red indicates phantom regions, green indicates object regions, yellow is for covered regions that are occluded or invisible in both camera views.

Table 4 Classification results for the foreground intersections at frame 1200 using both height matching and colour matching.

Region	1a	1b	1c	2a	2b	2c	3a	3b	3c
Label	Cv	Cv	Ob	Ph	Ob	Ph	Ob	Ph	Ph



(a) (b)

Fig. 10 Classification results of the intersection regions at frame 1200 using both height matching and colour matching, (a) in the overlaid foreground projection image and (b) in the foreground intersection image.

The phantom removal algorithm which uses height matching and colour matching has been tested over the 142 sampled frames. Table 5 and Table 6 show the performance evaluation of the phantom removal algorithm. The classification results are compared with ground truth data. The 786 intersection regions from 142 frames are classified into four categories: object regions, phantom regions, covered regions and occluded regions.

Table 5 Performance evaluation of the classification using height matching and colour matching.

		Classification Results with Height and Colour Matching				Number of the Ground Truth
		Object Regions	Phantom Regions	Covered Regions	Occluded Regions	
Ground Truth	Object Regions	307	0	10	2	319
	Phantom Regions	0	309	5	0	314
	Covered Regions	0	0	112	0	112
	Occluded Regions	0	0	0	41	41

Total number of Classification	307	309	127	43	786

In Table 5, the confusion matrix of the classification results is given, along with the ground truths. For each category, let GT and CR be the ground-truth numbers and actual classification numbers of that category. The false negatives (missed detections), FN, are the intersection regions which belong to that category but are misclassified as the other category. The false positives (FP) or false alarms are the intersection regions which belong to the other category but are misclassified as that category. The false negative rate (R_{FN}) is obtained as the ratio between the number of the false negatives and the number of ground truths. The false positives rate (R_{FP}) is the ratio between the number of the false positives and the ground-truth number.

$$R_{FN} = FN/GT \qquad (12)$$

$$R_{FP} = FP/GT$$

The false negative rate and the false positive rate of each category were calculated and the results are shown in Table 6. The ground-truth number of object regions was 319, where 307 were correctly identified. 10 object regions were misclassified as covered regions because pedestrians in these object regions are invisible in both camera views. 2 object regions were misclassified as occluded regions. Since no region was misclassified as an object region, the false negative rate is 3.76% and the false positive rate is 0.00%.

Table 6 The classification errors with the height matching and colour matching.

	False Negative Rate R_{FN} (%)	False Positive Rate R_{FP} (%)
Object Regions	3.76	0.00
Phantom Regions	1.59	0.00
Covered Regions	0.00	13.39
Occluded Regions	0.00	4.88

8 CONCLUSIONS

In this paper, an approach based on geometrical information and colour cues has been proposed to identify phantoms in multi-view pedestrian detection. The former is a height matching algorithm based on the geometry between the camera views. The latter is a colour matching algorithm based on the Mahalanobis distance of the colour distributions of every two associated foreground regions. Since the height matching is uncertain in the scenarios with adjacent pedestrians, the two algorithms are combined to improve the robustness of the foreground intersection classification. The robustness of the proposed algorithm is demonstrated in real-world image sequences.

The limitation of this algorithm is that the foreground segmentation error is assumed to be relatively low. When the foreground segmentation error is high, a higher threshold should be

applied in height matching, which may increase the rate of misclassification. As such, future investigations should be focused to tackle this new challenge with techniques such as denoising and enhancement [25-27], feature mining [28-29], deep learning [30, 32] and model-based tracking even with sub-pixel accuracy [31, 33, 34].

Acknowledgements

This work was supported by the National Natural Science Foundation of China (NSFC) under Grants 60975082/61672008 and the Scientific Research Program funded by Shaanxi Provincial Education Department, P. R. China, under Grant 15JK1310.

REFERENCES

[1] S. M. Khan and M. Shah, "Tracking multiple occluding people by localizing on multiple scene planes," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 31, pp. 505-519, 2009.

[2] T. T. Santos and C. H. Morimoto, "Multiple camera people detection and tracking using support integration," *Pattern Recognition Letters,* vol. 32, pp. 47-55, 2011.

[3] K. Kim and L. S. Davis, "Multi-camera tracking and segmentation of occluded people on ground plane using search-guided particle filtering," in *Proceedings of the European Conference on Computer Vision*, 2006, pp. 98-109.

[4] W. Hu, M. Hu, X. Zhou, T. Tan, J. Lou, and S. Maybank, "Principal axis-based correspondence between multiple cameras for people tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 28, pp. 663-671, 2006.

[5] W. Du and J. Piater, "Multi-camera people tracking by collaborative particle filters and principal axis-based integration," in *Proceedings of the Asian Conference of Computer Vision*, 2007, pp. 365-374.

[6] D. Arsic, E. Hristov, N. Lehment, B. Hornler, B. Schuller, and G. Rigoll, "Applying multi layer homography for multi camera person tracking," in *Proceedings of the ACM/IEEE International Conference on Distributed Smart Cameras*, 2008, pp. 1-9.

[7] A. Alahi, L. Jacques, Y. Boursier, and P. Vandergheynst, "Sparsity driven people localization with a heterogeneous network of cameras," *Journal of Mathematical Imaging and Vision,* vol. 41, pp. 39-58, 2011.

[8] J. Berclaz, F. Fleuret, and P. Fua, "Principled Detection-by-Classification from Multiple Views," in *Proceedings of the Conference on Computer Vision Theory and Applications*, 2008, pp. 375-382.

[9] R. Eshel and Y. Moses, "Tracking in a dense crowd using multiple cameras," *International journal of computer vision,* vol. 88, pp. 129-143, 2010.

[10] S. Sternig, T. Mauthner, A. Irschara, P. M. Roth, and H. Bischof, "Multi-camera multi-object tracking by robust hough-based homography projections," in *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, 2011, pp. 1689-1696.

[11] M. Liem and D. M. Gavrila, "Multi-person tracking with overlapping cameras in complex, dynamic environments," in *Proceedings of the British Machine Vision Conference*, 2009, 38(3): 199-218.

[12] D. B. Yang, H. H. González-Baños, and L. J. Guibas, "Counting people in crowds with a real-time network of simple image sensors," in *Proceedings of the International Conference on Computer Vision*, 2003, pp. 122-129.

[13] M. Liem and D. M. Gavrila, "Multi-person localization and track assignment in overlapping camera views," in *Pattern Recognition*, ed: Springer, 2011, pp. 173-183.

[14] X. Tong, T. Yang, R. Xi, D. Shao, and X. Zhang, "A Novel Multi-planar Homography Constraint Algorithm for Robust Multi-people Location with Severe Occlusion," in *Proceedings of theInternational Conference on Image and Graphics*, 2009, pp. 349-354.

[15] P. Peng, Y. Tian, Y. Wang, and T. Huang, "Robust multiple cameras pedestrian detection with multi-view bayesian network," Pattern Recognition, 48(5):1760–1772, 2015.

[16] W. Ge and R. T. Collins, "Crowd detection with a multiview sampler," in *Proceedings of the European Conference on Computer Vision*, 2010, pp. 324-337.

[17] D. H. Douglas and T. K. Peucker, "Algorithms for the reduction of the number of points required to represent a digitized line or its caricature," *The International Journal for Geographic Information and Geovisualization,* vol. 10, pp. 112-122, 1973.

[18] T. B. Moeslund, A. Hilton, and V. Krüger, "A survey of advances in vision-based human motion capture and analysis," *Computer vision and image understanding,* vol. 104, pp. 90-126, 2006.

[19] C. Stauffer and W. E. L. Grimson, "Adaptive background mixture models for real-time tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1999, 2.

[20] P. KaewTraKulPong and R. Bowden, "An improved adaptive background mixture model for real-time tracking with shadow detection," in *Video-Based Surveillance Systems*, ed: Springer, 2002, pp. 135-144.

[21] S. Suzuki, "Topological structural analysis of digitized binary images by border following," *Computer Vision, Graphics, and Image Processing,* vol. 30, pp. 32-46, 1985.

[22] I. E. Sutherland, R. F. Sproull, and R. A. Schumacker, "A characterization of ten hidden-surface algorithms," *ACM Computing Surveys (CSUR),* vol. 6, pp. 1-55, 1974.

[23] S.Lloyd, "Least squares quantization in PCM," *IEEE Transactions on Information Theory,* vol. 28, pp. 129-137, 1982.

[24] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B (Methodological),* pp. 1-38, 1977.

[25] J. Jiang et al, "Live: An integrated production and feedback system for intelligent and interactive tv broadcasting", *IEEE Trans. Broadcasting*, vol. 57, no. 3, pp. 646-661, 2011.

[26] J. Ren et al, "Fusion of intensity and inter-component chromatic difference for effective and robust colour edge detection", *IET Image Processing*, vol. 4, no. 4, pp. 294-301, 2010.

[27] J. Zabalza et al, "Novel two-dimensional singular spectrum analysis for effective feature extraction and data classification in hyperspectral imaging", *IEEE Trans. Geoscience and Remote Sensing*, vol. 53, no. 8, pp. 4418-4433, 2015.

[28] J. Ren et al, "Hierarchical modeling and adaptive clustering for real-time summarization of rush videos", *IEEE Trans. Multimedia*, vol. 11, no. 5, pp. 906-917, 2009.

[29] J. Ren et al, "Real-time modeling of 3-d soccer ball trajectories from multiple fixed cameras," IEEE Trans. CSVT. vol. 18, no. 3, pp. 350-362, 2008.

[30]. J Han et al, "Background prior-based salient object detection via deep reconstruction residual," IEEE Trans. CSVT, vol. 25, no. 8, pp. 1309-1321, 2015.

[31] J. Ren et al, "High-accuracy sub-pixel motion estimation from noisy images in Fourier domain," IEEE Trans. Image Processing, vol. 19, no. 5, pp. 1379-1384, 2010.

[32] J. Zabalza, "Novel segmented stacked autoencoder for effective dimensionality reduction and feature extraction in hyperspectral imaging," Neurocomputing, vol. 185, pp. 1-10, 2016.

[33]. J Ren et al, "Tracking the soccer ball using multiple fixed cameras," Computer Vision and Image Understanding, vol. 113, no. 5, pp. 633-642, 2009.

[34] J. Ren et al, "Efficient detection of temporally impulsive dirt impairments in archived films", *Signal Processing*, vol. 87, no. 3, pp. 541-551, 2007.