IET Research Journals

Submission Template for IET Research Journal Papers

IET Journals
The Institution of Engineering and Technology

Moving Shadow Detection via Binocular Vision and Colour Clustering

ISSN 1751-8644
doi: 0000000000
www.ietdl.org

Lei Lu[1], Ming Xu[2,3*], Jeremy S. Smith[3], Yuyao Yan[2]

[1] School of Information and Communications Engineering, Faculty of Electronic and Information Engineering, Xi'an Jiaotong Unversity, Xi'an, Shaanxi 710049, P. R. China
[2] Department of Electrical and Electronic Engineering, Xi'an Jiaotong-Liverpool University, Suzhou, Jiangsu 215123, P. R. China
[3] Department of Electrical Engineering and Electronics, University of Liverpool, L69 3BX, Liverpool, UK
* E-mail: ming.xu@xjtlu.edu.cn

Abstract: A pedestrian segmentation algorithm in the presence of cast shadows is presented in this paper. The novelty of this algorithm lies in the fusion of multiview and multiplane homographic projections of foregrounds and the use of the fused data to guide colour clustering. This brings about an advantage over the existing binocular algorithms in that it can remove cast shadows while keeping pedestrians' body parts which occlude shadows. The phantom detections, which are inherent with the binocular method, are also investigated. Experimental results with real-world videos have demonstrated the efficiency of this algorithm.

1 Introduction

Intelligent video surveillance is receiving more and more attention from the computer vision community and industry. It aims to automatically detect, track, recognize moving targets (e.g. pedestrians and vehicles) and identify abnormal events. It has been widely used for the security of public spaces, traffic monitoring, battlefield surveillance, etc. The recent progresses in artificial intelligence and GPUs foster the growth in intelligent video surveillance.

The success of an intelligent video surveillance system greatly depends on the robustness of its detection algorithm for moving targets. On the other hand, the background subtraction method, which is the most favourable one for foreground detection, is sensitive to the cast shadows of moving targets. As moving cast shadows change the local scene appearance in the same way as moving objects, they are often misclassified as foregrounds. These misclassified cast shadows can distort the shape, size and colour distribution of the relevant foreground regions and connect foreground regions which are not adjacent to each other. This may mislead the tracking algorithm [1], because the shape, size and colour of a foreground region usually constitutes the measurements for a tracker and each moving target in a multi-target foreground region has independent dynamics as well as different paths. In these cases, the individual targets cannot be readily separated and the further processing becomes unreliable. Therefore, moving cast shadows are one of the major challenges in intelligent video surveillance.

In this paper, we present a novel algorithm for the detection of moving cast shadows by using multiview geometric projection and colour segmentation. In this algorithm, the foregrounds extracted from individual camera views are projected to a virtual top view according to the homographies for the ground plane and a parallel plane at the average height of pedestrians' waists. The intersections of multiview foreground projections by using the ground-plane homographies correspond to the locations of moving cast shadows on which the pedestrians are standing. On the other hand, the intersections of multiview foreground projections by using the waist-plane homographies report the locations and widths of the pedestrians. In the second stage of this algorithm, pairs of such foreground intersection regions are warped back to the original camera views. K-means colour clustering is carried out on each shadow region to identify pedestrians' body parts from the shadow region. The clustering is initialized on the basis of the pair of warped back intersection regions, which makes the clustering converge quickly.

The contributions of the proposed algorithm are fourfold. Firstly, unlike the previous shadow detection algorithms using binocular vision, this algorithm can remove moving cast shadows while keeping pedestrians' body parts (e.g. feet and legs) which occlude shadows. Secondly, it is the first work in using multi-plane foreground projections to guide and accelerate the colour clustering in the shadow region. Thirdly, it is innovative to investigate the phantom detections which are inherent with shadow detection using binocular vision. Finally, a scale-space Hough transform is proposed to estimate the vertical vanishing point from the head-foot observations of pedestrians, which adds robustness to the estimation of the multiplane homography when camera calibration data is not available.

The remaining part of this paper is organized as follows: In Section 2, related work is discussed. Section 3 details the multi-plane homography estimation by using either calibrated or uncalibrated cameras. In Section 4, the extraction of pairs of foreground interaction regions is introduced. Section 5 describes the strategy to filter out phantom detections in the foreground intersection regions. Section 6 details the colour clustering to segment shadow regions. Experimental results are shown in Section 7. The conclusions are presented in Section 8.

2 Related Work

Shadows can be divided into cast shadows and self shadows, or moving shadows and static shadows. We will focus on the detection of moving cast shadows in video sequences. The reason is that, by using the background subtraction method, static cast shadows in videos can be correctly detected as backgrounds and moving self shadows can be correctly detected as foregrounds. In addition, we need to differentiate our target from the research works on shadow detection in still images in which supervised shadow learning can be carried out in an off-line manner. There exists a large volume of literature on moving shadow detection. Prati et al. [2] classified these methods into three categories according to the types of features used, that is, spectral, spatial and temporal features. Sanin et al. [3] further divided spectral features into intensity, chromaticity and physical properties, and divided spatial features into geometry and textures. A more recent survey can be found in [4], which further subdivides geometry features in terms of shapes and light directions, and separates edges from textures.

The *chromaticity* approach assumes that in outdoor environments the sunlight and ambient illumination is white so that cast shadows

© The Institution of Engineering and Technology 2015

reduce luminance values while maintaining the chromaticity values of background pixels. These methods often choose a colour space, which separates chromaticity from intensity, such as normalized rgb space in [5][6] and HSV space in [7][8]. This approach has a low computational cost. However, the chromaticity or hue component is poorly defined and very noisy in dark shadows. In addition, a part of foreground regions may meet this assumption and thus are lost in detection. The *physical* approach [9][10][11][12][13][14] realizes that there are two light sources (the sun and sky) in outdoor environments and cast shadows are bluish, due to the scattered light by the sky, rather than proportional to the background in RGB components. The methods in this approach model both light sources and learn the appearance of shadow pixels to better predict the colour change of shadow regions. They are more accurate than the chromaticity approach but still suffer from the colour similarity between moving objects and shadows. The *geometry* approach [15][16] utilizes the prior knowledge in light source orientation and specific object types to separate shadows from objects. Hsieh et al. [15] assumes that pedestrians are standing upright and have a different orientation from their shadows. The vertical histogram projection of each foreground region was used to split a coarse shadow region from the foreground region. Then the location, size and intensity of the coarse shadow region are modelled by a mixture of Gaussians, which is later used to refine the shadow detection. The geometry approach cannot deal with multiple light sources and objects with shadows in the same orientation as the objects. The *texture* approach [17][18][19][20][21][14] assumes that background regions under shadows maintain their texture. Therefore, coarse shadow regions are detected first by using spectral features and they are further classified into shadows or foregrounds by using texture correlation with the backgrounds. The disadvantage of this approach is in its slow speed. These approaches to moving shadow removal are summarized in Table 1.

Table 1 Methodology in moving shadow removal.

Methods	Pros	Cons
Chromaticity	simple, fast	white light sources, poor on dark shadows
Physical	more accurate	objects and backgrounds with similar chromaticity
Geometry	simple	known light source direction, upright people
Texture	insensitive to illumination	slow correlation
Binocular	any object type	shadows on flat planes and in overlapping field of view

With the deployment of multi-camera video surveillance systems, one good solution for moving shadow removal is to utilize the information redundancy in multiple camera views by exploiting multiview homographies [22][23][24][25][26]. Although more synchronized cameras are required, it improves the robustness of the detection owing to information fusion. Onoguchi [22] proposed an algorithm using two cameras and assuming that moving objects are standing on the ground plane. Then one camera view is warped to the other by a homographic transformation based on the ground plane. The intensity at each location of the second view is compared with that of the warped image from the first camera view. If they are highly correlated, then that pixel is classified as background or a cast shadow. The disadvantage of this algorithm is that, if a part of the cast shadow is occluded by a foreground object in one camera view but visible in the other view, it is classified as foregrounds in both camera views because the corresponding pixels in both camera views are not similar in their colours. In [23], Lanza et al. extracted the change mask image in each of the multiple camera views. These change mask images are projected to a virtual top view by homographic transformations. It was found that pedestrians and their cast shadows are always located in the intersections of these projected change masks from the multiple views. Then the intersection regions are warped back to and subtracted from the single-view change masks. This method often gives rise to phantom detections due to the intersections of the projected foregrounds of non-corresponding objects, e.g. between two pedestrians or between a pedestrian and the cast shadow of another pedestrian. These two methods remove not only the shadows but also the pedestrians' body parts. In [24], Jeong and Jaynes used Onoguchi's method to detect coarse shadow regions in two camera views and applied a Gaussian mixture model to learn the colours in such shadow regions at initial frames. The colour model was then used to refine coarse shadow regions. This paper differs from their work in that it does not rely on the assumptions of the same colour sensitivity in multiple cameras, the ground plane being a uniform colour and a dominant shadow area in each coarse shadow region. In addition, phantom detections are also considered in this paper. In [25] foreground regions are projected from multiple camera views to a stack of parallel planes and the across-plane intersections are used to compensate the lost feet in [23]. However, it is sensitive to pedestrians' poses such as striding people. In this paper, the foreground intersections on the waist plane are used to guide the colour clustering in the foreground intersection regions on the ground plane. There is no across-plane foreground intersection.

3 Homography Estimation

Planar homography is defined by a 3×3 transformation matrix between a pair of captured images of the same plane from two camera views. Let $\mathbf{u}$ and $\mathbf{u}'$ be the image coordinates of a point on such a plane in the two views. They are associated by the homography matrix $\mathbf{H}$ as follows:

$$\widetilde{\mathbf{u}}' \cong \mathbf{H}\widetilde{\mathbf{u}} \tag{1}$$

where $\cong$ denotes the equivalence defined up to scale and the vectors with a tilde represent their homogeneous coordinates.

3.1 Homography Estimation with Calibrated Cameras

As the homography transformation $\mathbf{H}$ is a special variation of the projective transformation, a 3×4 projection matrix $\mathbf{M} = [\mathbf{m}_1, \mathbf{m}_2, \mathbf{m}_3, \mathbf{m}_4]$, which is built by using the intrinsic and extrinsic parameters of each camera, can be used to determine the homography matrix for a specific plane.

The homography, from the top view to camera view c, for the ground plane is [27]:

$$\mathbf{H}_0^{t,c} = (\mathbf{H}_0^{c,t})^{-1} = [\mathbf{m}_1, \mathbf{m}_2, \mathbf{m}_4]\ . \tag{2}$$

The homography, from the top view to camera view c, for the plane parallel to the ground plane and at a height of h is as follows [27], where $[\mathbf{0}]$ is a 3×2 zero matrix:

$$\mathbf{H}_h^{t,c} = [\mathbf{m}_1, \mathbf{m}_2, h\mathbf{m}_3 + \mathbf{m}_4] = \mathbf{H}_0^{t,c} + [\mathbf{0}|h\mathbf{m}_3]\ . \tag{3}$$

3.2 Homography Estimation with Uncalibrated Cameras

When camera calibration data is not available, a three-step process can be used to estimate the homographies for the parallel planes at different heights:

1. Estimate the ground-plane homography $\mathbf{H}_g^{i,j}$ from at least four pairs of corresponding landmark points in two camera views.
2. Estimate the vertical vanishing point $\mathbf{v}$.
3. Calculate the homography $\mathbf{H}_h^{i,j}$ for a plane parallel to and at a height h above the ground plane, by using (**??**)[28], where $\gamma \propto h$:

$$\mathbf{H}_h^{i,j} = (\mathbf{H}_g^{i,j} + [\mathbf{0}_{3\times 2}]\gamma\mathbf{v})\left(\mathbf{I}_{3\times 3} - \frac{1}{1+\gamma}[\mathbf{0}_{3\times 2}]\gamma\mathbf{v}\right)\ . \tag{4}$$

The vanishing point is where parallel lines converge, when they are projected from 3D space to an image plane under a perspective projection. The vertical vanishing point can be estimated from

© The Institution of Engineering and Technology 2015

static scene structures such as the vertical lines of buildings [29][30]. In the scenarios which lack these vertical lines, the principal axes of pedestrians can be used to estimate the vertical vanishing point [31]. Due to the measurement errors and outliers in line segment extraction, the most favoured approaches for vanishing point detection are based on the Hough transform [29][32][33], RANSAC or RANSAC-like algorithms [30][34] or the clustering of the intersection points of line segment pairs [35]. To cope with the measurement errors, Shufelt [29] cast a swath of uniformly distributed votes into the Hough accumulator space; Szeliski [32] assigned more weights to long, non-collinear line segments; Xu et al [30] modelled the two end points of each line segment with Gaussians and measured the consistency between that line segment and each vanishing point candidate.

In this paper, the vertical vanishing point was estimated, from the head and foot positions of pedestrians, by using a scale-space Hough transform. A graphical interface was used to browse the video sequence of each camera view and collect the image coordinates $\mathbf{u}_f$ for the feet and $\mathbf{u}_h$ for the top of the head of each selected pedestrian standing upright. The outcome of this process is a set of foot-and-head landmark pairs $\{(\mathbf{u}^i_f, \mathbf{u}^i_h)\}|^m_{i=1}$ for a specific camera view. Although an automatic tool to collect such image coordinates may be developed by extracting the principal axes of the observed pedestrians, it is not trivial to reliably identify outliers such as shadows, vehicles, grouped pedestrians, cyclists, people with prams or luggage etc.

As the top of the head of each pedestrian can be easily identified and the majority of the observation errors come from the feet of striding pedestrians, it is assumed that: (1) the top of the head of each pedestrian is accurately observed; (2) the line L_i which connects $\mathbf{u}^i_h$ and $\mathbf{u}^i_f$ is most likely the direction of the i-th pedestrian's torso; (3) while $\mathbf{u}^i_f$ is most likely the foot position of the pedestrian, the other positions along the two sides of $\mathbf{u}^i_f$ may be the ground-truth foot location with decreased likelihoods. This is illustrated in Fig. 1(a), where the potential foot points are located along an circular arc with the circle centre at $\mathbf{u}^i_h$ and with the arc centre at $\mathbf{u}^i_f$. The closer a foot candidate is to $\mathbf{u}^i_f$, the more likely it is the foot point. θ_0 represents the largest span for this fan-shaped region and is determined by the averaged width to height ratio of the pedestrians.

To determine the projection from the torso line of each pedestrian into the Hough accumulator, a projection template in the vertical direction is built as a fan-shaped Gaussian, as shown in Fig. 1(b). The angle between any pixel $\mathbf{u}'$ in this template and the u'-axis is calculated as:

$$\theta = \arccos\left(\frac{\mathbf{u}'_0 \cdot \mathbf{u}'}{\|\mathbf{u}'_0\|\|\mathbf{u}'\|}\right) . \qquad (5)$$

where $\mathbf{u}'_0$ is a unit vector in the u'-axis. The projection template is defined as:

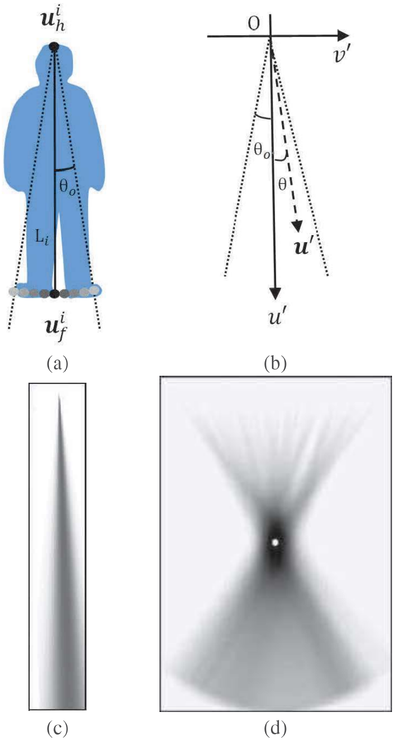
$$\mathbf{A}_t(\mathbf{u}') = \begin{cases} exp(-\theta^2/(2\sigma^2_\theta)), & if \quad \theta < \theta_0 \\ 0, & otherwise \end{cases} \qquad (6)$$

where $\sigma_\theta = \theta_0/3$.

The projection $\mathbf{A}_i(\mathbf{u})$ from the torso line of the i-th pedestrian into the Hough accumulator (see Fig. 1(c)) is obtained by applying an image rotation and translation operation on $\mathbf{A}_t(\mathbf{u}')$:

$$\begin{aligned} \phi_i &= \arccos\left(\frac{\mathbf{u}_0 \cdot (\mathbf{u}^i_f - \mathbf{u}^i_h)}{\|\mathbf{u}_0\|\|\mathbf{u}^i_f - \mathbf{u}^i_h\|}\right) && (7) \\ \mathbf{u} &= \begin{pmatrix} \cos\phi_i & \sin\phi_i \\ -\sin\phi_i & \cos\phi_i \end{pmatrix} \mathbf{u}' + \mathbf{u}^i_h \ . && (8) \end{aligned}$$

where $\mathbf{u}_0$ is a unit vector in the u-axis of the original camera view, which is facing downwards; ϕ_i is the angle between the torso line L_i and the u-axis, and its sign is determined by whether $\mathbf{u}^i_f$ is to the left or right of $\mathbf{u}^i_h$.



(a) (b)

(c) (d)

Fig. 1: (a) The head point and potential foot points of a pedestrian, (b) the image coordinates for a projection template, (c) the projection from a pedestrian into the Hough accumulator, and (d) the Hough accumulator and the vanishing point.

The projections overlaid in the Hough accumulator and the vertical vanishing point, as shown in Fig. 1(d), are:

$$\begin{aligned} \mathbf{A}(\mathbf{u}) &= \sum_{i=1}^{m} \mathbf{A}_i(\mathbf{u}) && (9) \\ \mathbf{v} &= \operatorname*{argmax}_{\mathbf{u}} \mathbf{A}(\mathbf{u}) \ . && (10) \end{aligned}$$

4 Coarse Shadow Detection

The homography which reveals the position transformation for coplanar objects between different camera views is utilized to acquire coarse shadow regions.

4.1 Single-View Foregrounds

The foreground regions, in each camera view, are detected through a background subtraction approach. The intensity of each pixel throughout a video is modeled by a mixture of Gaussians [36]. The most stable Gaussians correspond to the adaptive background. If the value of a pixel in the new frame is significantly different from that of the background, that pixel becomes a foreground pixel. Connected component analysis is used to connect the foreground pixels into foreground regions, which is followed by a morphological closing operation to bridge split body parts and a size filter to remove noise. The cast shadows of pedestrians are included in such a foreground change mask M as they make significant changes to the background appearance.

© The Institution of Engineering and Technology 2015

4.2 Ground-Plane Foreground Intersections

The foreground change mask of each camera view (say camera view c), M^c, is projected onto the top view t with the homography for the ground plane, which is formulated as:

$$M_g^{t,c} = \mathbf{H}_g^{c,t}(M^c) \tag{11}$$

The matrix $\mathbf{H}_g^{c,t}$ denotes the ground-plane homography from camera view c to the top view. The intersections of the projected change masks from multiple camera views, M_g^t, are as follows:

$$M_g^t = \bigcap_c M_g^{t,c} \tag{12}$$

which are also the intersection patches of moving objects with the ground plane. The intersection patches include pedestrians' feet and cast shadows, since both touch the ground.

The intersection patches for the ground plane are then warped back to the individual camera views, according to the ground-plane homography:

$$M_g^c = (\mathbf{H}_g^{c,t})^{-1}(M_g^t) \tag{13}$$

The warped back region M_g^c is usually located on the ground. To cope with the inaccuracy in the homography estimation and foreground detection, the warped back patches are morphologically dilated by a square structure element B before being intersected with the original change mask M^c. The size of B is proportional to the tolerance of the inaccuracy.

$$M_g^c = M^c \cap (M_g^c \oplus B) \tag{14}$$

The rectified ground region M_g^c then fits well with the corresponding foreground mask. Since the ground region M_g^c contains both pedestrians' feet and shadows, it can be thought of as a coarse shadow region. Some previous work used $M^c - M_g^c$ for pedestrian detection and thus lost the pedestrians' body parts such as feet and legs [23].

4.3 Waist-Plane Foreground Intersections

The foreground change mask of each camera view, M^c, is projected onto the top view t with the homography for a plane parallel to the ground and at the average waist height. The intersection of the multi-view foreground projections, according to the waist-plane, is a patch representing the waist section in that plane:

$$M_w^t = \bigcap_c \mathbf{H}_w^{c,t}(M^c) \tag{15}$$

where $\mathbf{H}_w^{c,t}$ denotes the waist-plane homography from camera view c to the top view. When it is warped back to the individual camera views according to the ground-plane homography,

$$M_w^c = (\mathbf{H}_g^{c,t})^{-1}(M_w^t) \tag{16}$$

the warped back patch is like the projection of a pedestrian's torso on the ground and is usually close to the pedestrian's feet. Therefore, it is referred to as a bottom region. The position and size of the bottom region can be used as a reference in the subsequent colour clustering method on the ground region.

5 Phantom Pruning in Foreground Intersections

Phantoms may appear in the warped back region M_g^c, when there are more than one pedestrian adjacent to each other in a camera view. This is due to the intersection of foreground projections of different pedestrians in the top view [37]. For the ground-plane based projections, real ground regions are usually located at the bottom of each foreground region in a single view, while the phantoms are always hidden behind the pedestrians in the top view and above the ground region in the single view. On the other hand, a pedestrian hidden behind others may have a foreground intersection region above the ground region in the single view.

Algorithm 1 Validation of ground and bottom regions

```
Input: Foreground maps $M^c$, $c \in [1, C]$;
Input: Ground region maps $M_g^t$ and $M_g^c$, $c \in [1, C]$;
Input: Bottom region maps $M_w^t$ and $M_w^c$, $c \in [1, C]$;
Output: Validated ground and bottom regions $R_g^c$, $R_w^c$, $c \in [1, C]$;
% step 1: validate ground regions
for each intersection region in $M_g^t$ do
    if its counterpart in $M_g^c$ is at the bottom of a region in $M^c$ then
        It joins $R_g^c$ for all camera views
    end if
end for
% step 2: validate bottom regions
for each intersection region in $M_w^t$ do
    if its counterpart in $M_w^c$ is outside any region of $R_g^c$ then
        Remove it in all camera views
    end if
end for
for each intersection region in $M_w^t$ do
    if its counterpart in $M_w^c$ is the only one in a region of $R_g^c$ then
        It joins $R_w^c$ for all camera views
    else if its counterpart in $M_w^c$ is the lowest in a region of $R_g^c$ then
        It joins $R_w^c$ for all camera views
    end if
end for
return [$R_g^c$, $R_w^c$], $c \in [1, C]$
```

For the waist-plane based projections M_w^c, phantoms may appear above or below the ground region. The phantoms above the ground region in a single view are those foreground intersections behind the pedestrians in the top view. Those below the ground region are the intersections in front of pedestrians. In the assumption that each pedestrian is not simultaneously hidden behind others in all camera views, Algorithm 1 is developed to identify validated ground regions and bottom regions.

6 Colour Clustering on Ground Regions

The ground region M_g contains both the feet and shadows. To distinguish between these two regions, the colour intensity values of both regions in each ground region are modeled by two Gaussians and then a rectified K-means algorithm is applied to identify these two regions in each ground region.

6.1 Initialization of K-Means Algorithm

The K-means algorithm [38] is an iterative process for cluster analysis. On the basis of several initial means, the algorithm proceeds by alternating between an assignment step and an update step. The assignment step is performed by assigning each observation to a cluster so that the least within-cluster sum of squared distances is achieved. Then the update step is conducted through calculating the new means as the centroids of the observations in the new clusters. In practice, an initial mean must be assigned to each cluster. The ground region R_g^c and bottom region R_w^c for each foreground region are used in the selection of the initial means.

A staged process is followed to remove the cast shadows. A fixed number of N_{sample} points are randomly sampled from the ground region R_g^c. A distance list is acquired by calculating the L2 distances between each of these points and the centre of R_w^c. This list is sorted in ascending order of the distances. Then, the top

$N_{sample}[area(R_w^c)/area(R_g^c)]$ points with the nearest distance to the centre of R_w^c are averaged in colour intensity and the mean is used as the seed point for the foot region. The other sample points with farther distances to the centre of R_w^c are used to estimate the seed point for the shadow region.

6.2 Implementation of Modified K-Means Algorithm

Let Φ denote the label of the class which the pixels in the ground region belong to, i.e., $\Phi = \{c_{shadow}, c_{feet}\}$. Then label decision for the pixels in the ground region could be modeled with a Bayesian maximum posterior optimization, i.e.,

$$c_{optimal} = \underset{c \in \Phi}{\operatorname{argmax}}\, p(c|\mathbf{I}) \ . \tag{17}$$

where $\mathbf{I}$ is the colour intensity observation and $c_{optimal}$ is the optimal label to interpret $\mathbf{I}$. The term $p(c|\mathbf{I})$ could be further decomposed, with the Bayesian rule, into:

$$p(c|\mathbf{I}) = \frac{p(\mathbf{I}|c)p(c)}{p(\mathbf{I})} \propto p(\mathbf{I}|c)p(c) \ . \tag{18}$$

The likelihood $p(\mathbf{I}|c_i)$ for class c_i can be formulated as:

$$p(\mathbf{I}|c_i) = (2\pi)^{-\frac{N}{2}} |\boldsymbol{\Sigma}_i|^{-\frac{1}{2}} exp\{-\frac{1}{2}(\mathbf{I} - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1}(\mathbf{I} - \boldsymbol{\mu}_i)\} \tag{19}$$

where $\boldsymbol{\mu}_i$ is the mean colour of each cluster and N is the dimension of the colour space ($N = 3$). The prior probability $p(c_i)$ can be approximated with the area ratio in the ground region:

$$p(c_i) = \begin{cases} area(R_w^c)/area(R_g^c), & if \quad c_i = c_{feet} \\ 1 - area(R_w^c)/area(R_g^c), & if \quad c_i = c_{shadow} \end{cases} \tag{20}$$

where R_w^c is the bottom region approximating the foot region. Based on the modeling process above, the modified K-means algorithm can be processed in four steps:

1. Initialization: k initial means (k=2) are generated.
2. Assignment: Assign each pixel to its corresponding cluster based on (17) and (18).
3. Updating: calculate the new means to be the centroids of the observations in the new clusters.
4. Iteration: Alternate between steps 2 and 3.

Based on the modified K-means algorithm, the shadow region and foot region can be identified effectively. Fig. 2 shows the colour histogram and the clustering result of the K-means algorithm for a ground region at frame 329 (Fig. 5) of the EPFL Campus dataset [39]. As the foot region is darker and much smaller than the shadow, its colour histogram is concentrated on the left and is of lower magnitudes.

7 Experimental Results

The proposed scale-space Hough transform for vanishing point detection and the shadow removal algorithm have been tested using real-world videos and/or Monte Carlo simulation tests. Both qualitative and quantitative performance evaluations were carried out.

7.1 Performance Evaluation of Vanishing Point Estimation

A number of experiments were carried out on the vanishing point estimation by using both real-world surveillance videos, such as the EPFL Campus [39] and PETS'2001 datasets [40], and a few Monte Carlo simulation tests. Fig. 3 shows the estimation results of the vanishing point and multi-plane homographies by using camera view 2 of the PETS'2001 dataset. Fig. 3(a) is the 3D visualization of the Hough accumulator, where the top of the ridge in dark red corresponds to the vanishing point. Fig. 3(b) shows the sampled torso lines in the original camera view (at the top) and projected into the Hough accumulator space. As pedestrians are short targets and may change poses during walking, these torso lines are very noisy and contain some outliers.

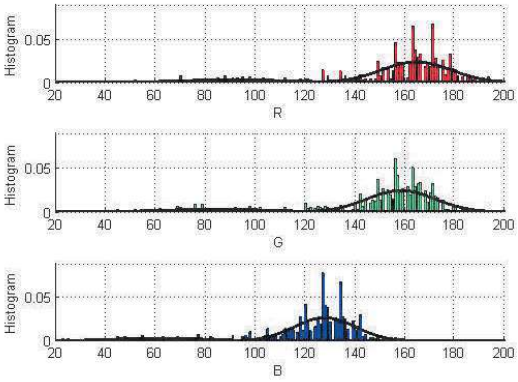


Fig. 2: The colour histogram and the result of the K-means clustering for a ground region. The subfigures from top to bottom are the R, G and B channels, respectively

Figs. 3(c) and (d) are used to verify the estimation of the vanishing point and homographies. Fig. 3(c) shows the framelets of a small number of pedestrians overlaid on the background image at their original locations. There is no building line segment available in this camera view. Fig. 3(d) is a top view from Google Maps for the same site and is used as a reference image. The ground-plane homography between view 2 and the top view was calculated from the landmarks. The feet of the pedestrians in Fig. 3(c) were manually localized and labelled with crosses. They are projected into the top view according to the ground-plane homography. The projections in the top view, which correspond to the locations of these pedestrians, are then back projected to Fig. 3(c) according to the head-plane homography and are labelled with circles. If a pedestrian is standing upright and of average height, the back-projection corresponds to the head position. In Fig. 3(c), the circles are indeed at the head positions of the pedestrians. Each circle is also on the line connecting the pedestrian's feet with the vanishing point.

Monte Carlo simulation tests were carried out to evaluate the quantitative performance of the vanishing point estimation. A greyscale image of 500 × 1000 pixels was used for the Hough accumulator space. Along the top border of this image are 100 points evenly distributed to simulate the head points of pedestrians, which are denoted as $P = \{p_1, p_2, ..., p_{100}\}$. In the middle of this image is the ground-truth location of the vanishing point v. Therefore, $\overline{p_i v}$ simulates the ground-truth torso line of a pedestrian. This is illustrated in the left of Fig. 4(a). Then a disturbance, which was sampled from the Gaussian $(1/(\sqrt{2\pi}\sigma_\theta))exp(-\theta^2/(2\sigma_\theta^2))$, was added into the angle of each torso line (see the middle of Fig. 4(a)), where $\sigma_\theta = \phi/3$ and $\phi = \arctan(1/9) \approx 6.34^\circ$ is a constant. The Hough accumulator generated using the largest span $\theta_0 = \phi$ is shown in the right of Fig. 4(a).

In the proposed algorithm for vanishing point estimation, θ_0 is the only parameter. To investigate its impact on the performance of the vanishing point estimation, a range of values for θ_0 were tested. For each selected value, 100 Monte Carlo simulations were carried out and the distances between the estimated vanishing points and the ground truth were compared. Fig. 4(b) shows the Hough accumulators when 2ϕ, 4ϕ, or 6ϕ was used as the largest span θ_0, in which the green dot represents the estimated vanishing point and the cross is the ground truth. The localization errors of the vanishing points using different θ_0 values are shown in Table 2. When the span of

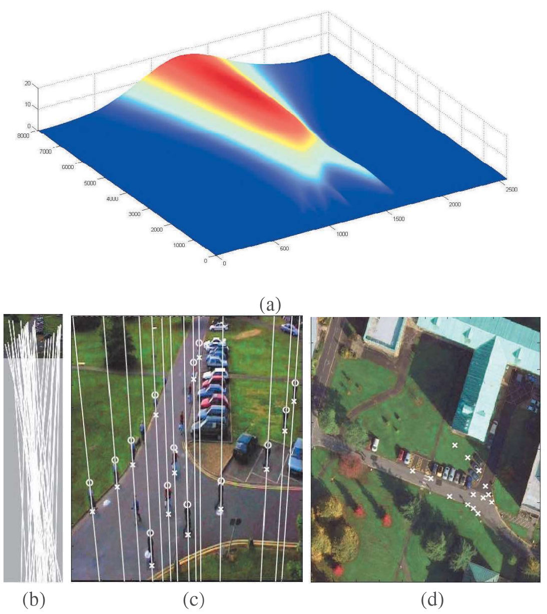


Fig. 3: The vanishing point estimation in the PETS'2001 dataset: (a) 3D visualization of the Hough acumulator, (b) the torso lines, (c) verification of the estimated vanishing point and homographies, and (d) the corresponding pedestrian locations in a top view.

the projection template varies from ϕ to 9ϕ, the localization errors stay at low levels and are not sensitive to the choice of the θ_0 value. When the span is less than ϕ, the localization errors increase rapidly, because this approximates the Hough algorithm without considering the observation uncertainty.

Table 2 Localization errors of the vanishing point under different θ_0 values. The unit is one pixel.

Span	0.3ϕ	0.5ϕ	0.7ϕ	ϕ	1.5ϕ	2ϕ	2.5ϕ
Mean	25.6	17.8	12.8	9.4	9.0	8.7	8.2
STD	14.8	12.8	8.9	6.1	6.6	5.8	5.9
Span	3ϕ	4ϕ	5ϕ	6ϕ	7ϕ	8ϕ	9ϕ
Mean	8.9	8.6	8.1	8.9	7.9	7.4	8.2
STD	6.2	5.5	5.5	6.0	5.1	5.2	5.6

The proposed algorithm was also compared with the RANSAC algorithm for vanishing point estimation. 200 torso lines were used in the comparison, in which different outlier numbers 0, 10, 20 and 40 were tested and the other torso lines were randomly disturbed with $\sigma_\theta = \phi/3$. For each selected outlier rate, 100 Monte Carlo simulations were carried out and the distances between the estimated vanishing points and the ground truth were compared. The localization errors of the proposed and the RANSAC algorithms are shown in Table 3. The proposed scale-space Hough algorithm outperforms the RANSAC algorithm with a two-thirds mean and a halved variance (square of standard deviation STD) for localization errors.

7.2 Performance Evaluation of Moving Shadow Detection

The proposed algorithm for moving shadow detection was evaluated by using the EPFL Campus video sequences [39]. This video was selected since it is the only public video dataset which has been used for multiview shadow detection [23]. It was captured by three cameras at a frame rate of 25 fps and with a resolution of 360×288 pixels. The ground-plane homography matrix from each camera view to a virtual top view is provided but camera calibration data is not available. In our experiments, only two camera views (view 0 and view 2) were used; the waist-plane homography, from each camera view to the top view, was estimated from the vertical vanishing point and the given ground-plane homography. The vertical vanishing point was estimated by using the scale-space Hough transform of the observed torso lines.

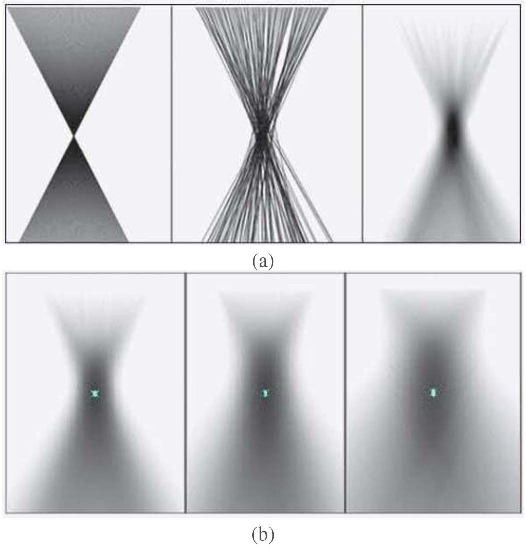


Fig. 4: Simulation tests on vanishing point estimation: (a) the ground-truth vanishing point and torso lines (left), the torso lines with disturbance (middle), and the Hough accumulator with the largest span $\theta_0 = \phi$ (right); (b) the Hough accumulators with $\theta_0 = 2\phi$ (left), $\theta_0 = 4\phi$ (middle) and $\theta_0 = 6\phi$ (right).

Table 3 Localization error comparisons of the vanishing point using the proposed and RANSAC algorithms. The unit is one pixel.

Outlier Rate (%)	0	5	10	20
Mean (proposed)	9.4	10.5	10.9	11.8
Mean (RANSAC)	15.7	16.9	15.3	17.5
STD (proposed)	6.1	6.5	6.8	7.7
STD (RANSAC)	8.7	9.3	9.0	10.7

Fig. 5 shows the results of the shadow removal algorithm at frame 329, which contains a single pedestrian. The original images, the single-view change masks M^c and the final foregrounds in both camera views are shown in the 1st row of Figs. 5. It is observed that the cast shadows are removed but the feet and legs remain. The single-view change masks M^c are projected and intersect in the top view with the ground-plane homography and with the waist-plane homography, respectively, as shown in the 2nd row of Fig. 5. There is no phantom foreground intersection in the top view. Warping the intersection patches from the ground plane to both camera views leads to the black ground region that contains both the pedestrian's feet and their shadow. On the other hand, warping the intersection patches from the waist plane to both camera views leads to the black bottom region of the pedestrian.

Fig. 6 shows the results of the shadow removal algorithm at frame 3638, which contains two pedestrians. The original images, the single-view change masks and the final foregrounds are shown in the 1st row of Figs. 6. The cast shadows are removed but the feet and legs remain. The single-view change masks are projected and intersect in the top view with the ground-plane homography, as shown in the 2nd row of Fig. 6. The two intersection regions behind the pedestrians are phantoms due to different people. The intersection patches

© The Institution of Engineering and Technology 2015

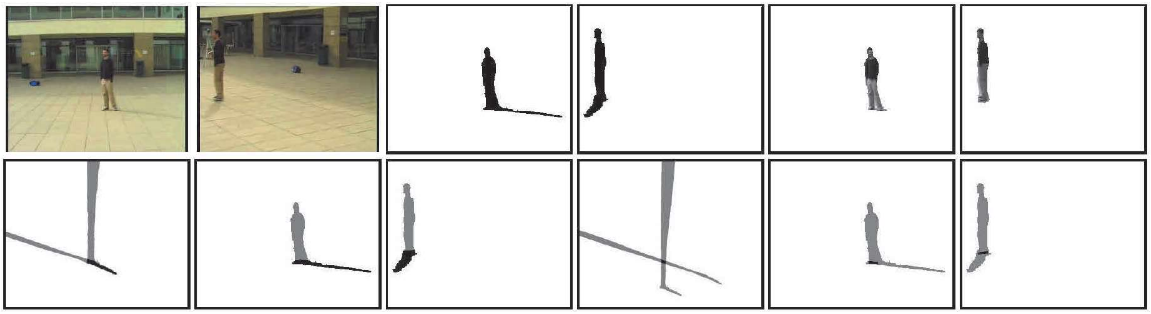


Fig. 5: Shadow removal for one pedestrian at frame 329 of the Campus dataset. 1st row: the two original camera views, the single-view change masks and the final foregrounds. 2nd row: (left) the ground-plane foreground intersections and the warped back intersections after morphological dilation, (right) the waist-plane foreground intersections and the warped back intersections after morphological dilation.

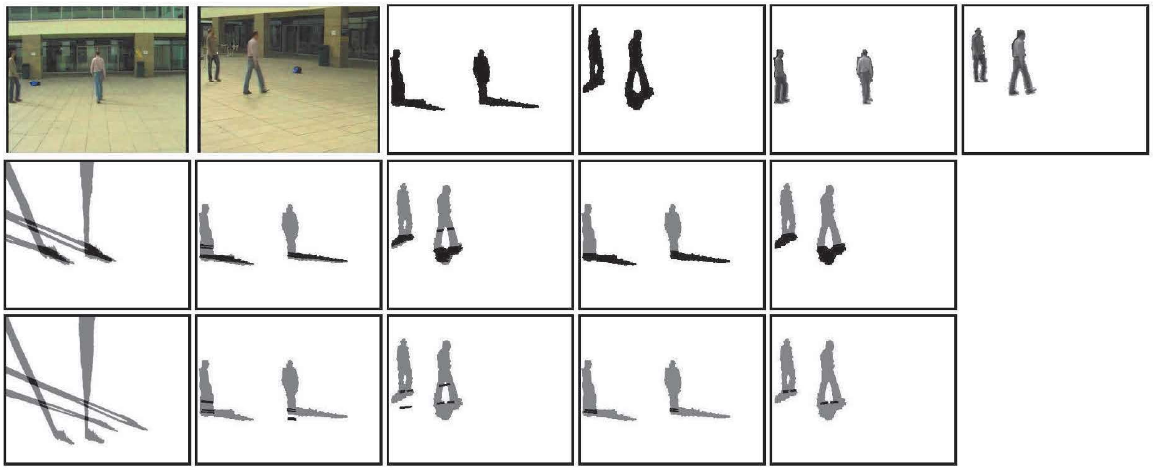


Fig. 6: Shadow removal for two pedestrians at frame 3638 of the Campus dataset. 1st row: the two original camera views, the single-view change masks and the final foregrounds. 2nd row: the ground-plane foreground intersections, and the warped back intersections before or after phantom removal and morphological dilation. 3rd row: the waist-plane foreground intersections, and the warped back intersections before and after phantom removal and morphological dilation.

are warped back to both camera views and shown in black. The black region at the bottom of each foreground region is a ground region, while those above the ground region are phantoms. The results after the phantom removal and morphological dilation are shown in the right of the 2nd row. The top-view foreground intersections using the waist-plane homography is shown in the left of the 3rd row of Fig. 6. The two intersection regions in front of or behind the pedestrians are phantoms. The intersection patches are warped back to both camera views. The black regions within or partly overlapping the ground region are bottom regions, while the others are identified as phantoms. The results after the phantom removal are shown in the right of the 3rd row.

To evaluate the quantitative performance of the shadow removal algorithm, the detection errors of both pedestrians and shadows were investigated. As the detected foreground regions for pedestrians and cast shadows were obviously affected by the foreground detection algorithm and the morphological closing operation, manual segmentation of the foreground regions as ground truths was not justified. Instead, extraction of the ground-truth foregrounds for each pedestrian and the corresponding shadow was based on the detected foreground regions. Then the ground-truth region of the pedestrian was manually segmented from the detected foreground region and the remaining foreground region was thought of as the ground-truth region for the pedestrian's shadow.

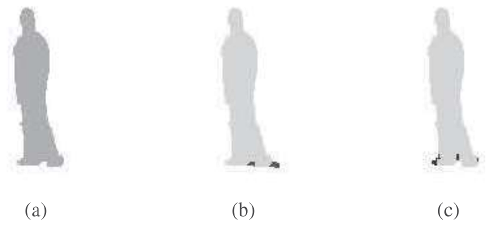


Fig. 7: The detection errors of a pedestrian: (a) the ground-truth foreground region, (b) the ground-truth foreground region (grey) and false positives (black), and (c) the detected ground-truth foreground region (grey) and false negatives (black).

The results of the proposed algorithm were compared with the ground-truth regions for both the pedestrian and the corresponding shadow, as shown in Fig. 7. The pixels, which belong to the ground-truth region of the pedestrian but are detected as shadows, were thought as false negatives for the pedestrian and false positives for the shadow. The pixels, which belong to the ground-truth region

© The Institution of Engineering and Technology 2015

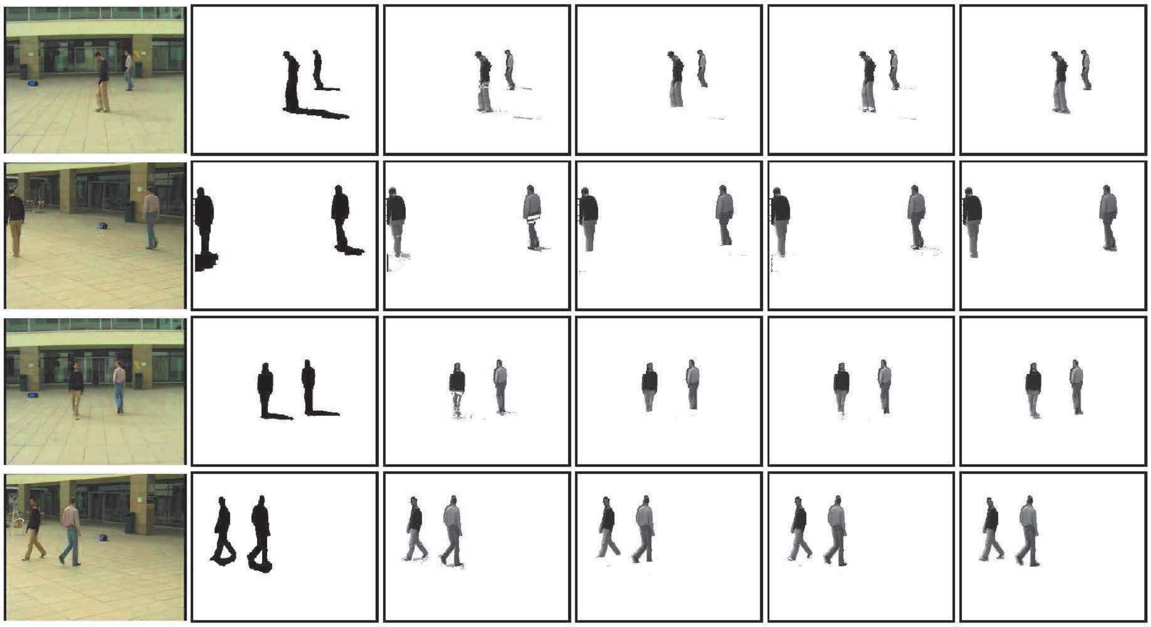


Fig. 8: A qualitative comparison of shadow removal algorithms in both camera views of frames 4253 and 5087 in the EPFL Campus dataset. 1st column: the original images, 2nd column: foreground regions, 3rd column: Onoguchi algorithm, 4th column: Lanza A (general purpose) algorithm, 5th column: Lanza B (shadow focused) algorithm, and 6th column: the proposed algorithm.

of the shadow but are detected as its corresponding pedestrian, were thought as false positives for the pedestrian and false negatives for the shadow. Fig. 7 illustrates the ground truth and detection errors of a pedestrian at frame 329. Fig. 7(a) is the ground-truth region of the pedestrian. Fig. 7(b) compares the ground-truth and the false positives of the pedestrian. Fig. 7(c) compares the detected ground-truth region and the false negatives of the pedestrian.

Table 4 shows the detection error rates of pedestrians and shadows using the proposed algorithm and benchmark binocular methods for moving shadow removal. The numbers of the detection errors were normalised by the ground-truth pixel number for either the pedestrian or the shadow. As the pedestrians tend to be larger than their cast shadows in this video, the false positive rate (FPR=FP/GT) and false negative rate (FNR=FN/GT) for shadows are higher than those for pedestrians. The average FPR and FNR for pedestrians in the proposed algorithm are 2.49% and 0.60%, respectively. Those for their shadows are 2.51% and 9.07%, respectively. The results were further compared with those of the Onoguchi algorithm [22], Lanza A (general purpose) method and Lanza B (shadow focused) method [23]. Since Onoguchi's algorithm tends to lose a part of the foregrounds for pedestrians due to intensity correlation with the pixelwise projection of the other view, it has a higher pedestrian FNR and shadow FPR; Since the Lanza A method considers ground regions as cast shadows and underestimates the pedestrian regions, its shadow FNR and pedestrian FPR are almost zero, but the pedestrian FNR and shadow FPR are very high; The Lanza B method tends to extract incomplete feet and legs of pedestrians and thus gives rise to higher pedestrian FNR and shadow FPR. To give a fair comparison, the total error rate (TER) was used, where TER=FPR+FNR. In Table 4, the total error rates in both the pedestrian and shadow detections using the proposed algorithm are the lowest in these four binocular methods for moving shadow removal. A qualitative comparison of these four algorithms is shown in Fig. 8. To give a fair comparison, the implementations of these four algorithms were based on the same set of foreground detection results as shown in column 2. The proposed algorithm demonstrates the best quality in terms of pedestrian completeness and shadow removal.

Table 4 A comparison on the detection error rates of the proposed and traditional algorithms.

	Pedestrians (%)			Shadows (%)		
	FPR	FNR	TER	FPR	FNR	TER
Onoguchi [19]	3.59	6.74	10.33	25.30	14.57	39.87
Lanza A [20]	0.02	8.85	8.87	33.32	0.08	33.40
Lanza B [20]	1.11	3.62	4.73	13.61	4.12	17.73
Proposed	2.49	0.60	3.09	2.51	9.07	11.58

The proposed algorithm was implemented in C/C++. Its speed was tested by using a PC with an Intel Core i7-9700K CPU (8 cores) running at 3.6GHz and a 16GB RAM. The results are shown in Table 5. The execution time for running the proposed algorithm consists of four steps: foreground detection, foreground projection, phantom pruning and colour clustering. Foreground detection was carried out by using Gaussian mixture model. Foreground regions were projected to the top view by using a contour-based real-time implementation [41]. The average time in processing one frame is 26.5 ms, which corresponds to a frame rate of 37.7 fps.

Table 5 Execution time for running the proposed algorithm.

No.	Steps	Time (ms)	Percentage (%)
1	Foreground Detection	3.1	11.7
2	Foreground Projection	9.7	36.6
3	Phantom Pruning	0.1	0.4
4	Colour Clustering	13.6	51.3
	Total	26.5	100.0

8 Conclusions and Discussion

We have proposed a moving object segmentation algorithm by using multiple cameras, which is robust in the presence of cast shadows. The novelty of the work lies in the foreground fusion by using multiview and multiplane homography mapping and a novel colour segmentation technique guided by the outcome of the data fusion. This algorithm is effective in the scenarios where the shadows are cast on a flat road plane and within the overlapping fields of view of two cameras.

It is worth noting that the proposed algorithm is related to deep learning based object detection [42] in that both can detect pedestrians in videos. However, significant differences exist: (1) the former aims to detect general moving objects, while the latter is focused on specific classes of objects that have been trained; (2) the former is based on widely used background subtraction inherent to moving shadows, while the latter is insensitive to shadows; (3) the former is very efficient, while the latter needs dedicated GPUs for training with large-scale datasets and is much slower even at the detection stage; (4) a major challenge of the latter is occlusion, which can be coped with by using multiple cameras as in the former.

9 Acknowledgments

This work was supported by the National Natural Science Foundation of China (NSFC) under Grant 60975082 and Xi'an Jiaotong-Liverpool University under Grant RDF-17-01-33.

10 References

1 Xu, M., Ellis, T., Godsill, S. J., Jones, G. A.: 'Visual tracking of partially observable targets with suboptimal filtering', *IET Computer Vision*, 2011, 5, (1), pp. 1–13

2 Prati, A., Mikic, I., Trivedi, M., Cucchiara, R.: 'Detecting moving shadows: algorithms and evaluation', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2003, 25, (7), pp. 918–923

3 Sanin, A., Sanderson, C., Lovell, B. C..: 'Shadow detection: a survey and comparative evaluation of recent methods', *Pattern Recognition*, 2012, 45, (4), pp. 1684–1695

4 Russell, M.,Zou, J. J., Fang, G.: 'An evaluation of moving shadow detection techniques', *Computational Visual Media*, 2016, 2, (3), pp. 195–217

5 Elgammal, A., Harwood, D., Davis, L.: 'Non-parametric model for background subtraction'. Proc. European Conf. on Computer Vision, 2006, pp. 751–767

6 Xu, M., Ellis, T.: 'Illumination-invariant motion detection using colour mixture models'. Proc. British Machine Vision Conf., 2001, pp. 163–172

7 Cucchiara, R., Grana, C., Piccardi, M., Prati, A.: 'Detecting moving objects, ghosts, and shadows in video streams', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2003, 25, (10), pp. 1337–1342

8 Nagarathinam, K., Kathavarayan, R. S.: 'Moving shadow detection based on stationary wavelet transform and zernike moments', *IET Computer Vision*, 2018, 12, (6), pp. 787–795

9 Nadimi, S., Bhanu, B.: 'Physical models for moving shadow and object detection in video', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2004, 26, (8), pp. 1079–1087

10 Martel-Brisson, N., Zaccarin, A.: 'Learning and removing cast shadows through a multidistribution approach', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2007, 29, (7), pp. 1133–1146

11 Liu, Z., Huang, K., Tan, T.: 'Cast shadow removal in a hierarchical manner using MRF', *IEEE Trans. Circuits Syst. Video Techn.*, 2012, 22, (1), pp. 56–66

12 Wang, B., Chen, C. P., Li, Y., Zhao, Y.: 'Hard shadows removal using an approximate illumination invariant'. Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, 2018, pp. 1628–1632

13 Shi, H., Liu, C.: 'A new cast shadow detection method for traffic surveillance video analysis using color and statistical modeling', *Image and Vision Computing*, 2020, 94, 103863

14 Yi, Y., Dai, J., Wang, C., et al: 'An Effective Framework Using Spatial Correlation and Extreme Learning Machine for Moving Cast Shadow Detection', *Applied Sciences*, 2019, 9, 5024

15 Hsieh, J. W., Hu, W. F., Chang, C. J., Chen, Y. S.: 'Shadow elimination for effective moving object detection by gaussian shadow modeling', *Image and Vision Computing*, 2003, 21, (6), pp. 505–516

16 Yan, Y., Xu, M., Smith, J. S.: 'Generalized vertical projection histograms using multi-plane homology', *IET Electronics Letters*, 2019, 55, (10), pp. 593–595

17 Leone, A., Distante, C.: 'Shadow detection for moving objects based on texture analysis', *Pattern Recognition*, 2007, 40, (4), pp. 1222–1233

18 St-Charles, P. L., Bilodeau, G. A., Bergevin, R.: 'SuBSENSE: A universal change detection method with local adaptive sensitivity', *IEEE Trans. on Image Processing*, 2015, 24, (1), pp. 359–373

19 Gomes, V., Barcellos, P., Scharcanski, J.: 'Stochastic shadow detection using a hypergraph partitioning approach', *Pattern Recognition*, 2017, 63, pp. 30–44

20 Russell, M., Zou, J., Fang, G., Cai, W.: 'Feature-based image patch classification for moving shadow detection', *IEEE Trans. Circuits Syst. Video Techn.*, 2019, 29, (9), pp. 2652–2666

21 Zhang, H., Qu, S., Li, H., Luo, J., Xu, W.: 'A Moving shadow elimination method based on fusion of multi-feature', *IEEE Access*, 2020, 8, pp. 63971–63982

22 Onoguchi, K.: 'Shadow elimination method for moving object detection'. Proc. Int. Conf. on Pattern Recognition, 1998, vol. 1, pp. 583–587

23 Lanza, A., Stefano, L. D., Berclaz, J., Fleuret, F., Fua, P.: 'Robust multi-view change detection'. Proc. British Machine Vision Conf., 2007

24 Jeong, K., Jaynes, C.: 'Moving shadow detection using a combined geometric and color classification approach'. Proc. IEEE Workshop on Motion and Video Computing, 2005, pp. 36–43

25 Iwama, H., Makihara, Y., Yagi, Y.: 'Foreground and shadow segmentation based on a homography-correspondence pair'. Proc. Asian Conf. on Computer Vision, 2010, pp. 702–715

26 Xu, M., Lu, L., Jia, T., Ren, J., Smith, J.: 'Cast shadow removal in motion detection by exploiting multiview geometry'. Proc. IEEE Int. Conf. on Syst., Man and Cybern., 2012, pp. 762–766

27 Ren, J., Xu, M., Smith, J. S., Cheng, S.: 'Multi-view and multi-plane data fusion for effective pedestrian detection in intelligent visual surveillance', *Multidimensional Systems and Signal Processing*, 2016, 27, (4), pp. 1007–1029

28 Khan, S. M., Shah, M.: 'Tracking multiple occluding people by localizing on multiple scene planes', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2009, 31, (3), pp. 505–519

29 Shufelt, J. A.: 'Performance evaluation and analysis of vanishing point detection techniques', *IEEE Trans. Pattern Anal. Mach. Intell.*, 1999, 21, (3), pp. 282–288

30 Xu, Y., Oh, S., Hoogs, A.: 'A minimum error vanishing point detection approach for uncalibrated monocular images of man-made environments'. Proc. IEEE Conf. on Computer Vision and Pattern Recognition, 2013, pp. 1376–1383

31 Lv, F., Zhao, T., Nevatia, R.: 'Camera calibration from video of a walking human', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2006, 28, (9), pp. 1513–1518

32 Szeliski, R.: 'Computer Vision: Algorithms and Applications' (Springer, 2010)

33 Zhang, Y., Su, Y., Yang, J., Ponce, J., Kong, H.: 'When dijkstra meets vanishing point: A stereo vision approach for road detection', *IEEE Trans. on Image Processing*, 2018, 27, (5), pp. 2176–2188

34 Zhou, Z., Farhat, F., Wang, J.: 'Detecting dominant vanishing points in natural scenes with application to composition-sensitive image retrieval', *IEEE Trans. on Multimedia*, 2016, 19, (12), pp. 2651–2665

35 Ding, W., Li, Y., Liu, H.: 'Efficient vanishing point detection method in unstructured road environments based on dark channel prior', *IET Computer Vision*, 2016, 10, (8), pp. 852–860

36 Stauffer, C., Grimson, W. E. L.: 'Adaptive background mixture models for realtime tracking'. Proc. IEEE Conf. on Computer Vision and Pattern Recognition, vol. 2, 1999, pp. 2246–2252

37 Ren, J., Xu, M., Smith, J., Zhao, H., Zhang, R.: 'Multi-view visual surveillance and phantom removal for effective pedestrian detection', *Multimed Tools Appl.*, 2018, 77, (14), pp. 18801–18826

38 MacKay, D.: 'Information Theory, Inference and Learning Algorithms' (Cambridge Univ. Press, 2003)

39 'EPFL dataset', http://cvlab.epfl.ch/data/pom

40 'PETS2001 dataset', http://www.cvg.rdg.ac.uk/datasets/index.html

41 Xu, M., Ren J., Chen D., Smith J. S. and Wang, G.: 'Real-time detection via homography mapping of foreground polygons from multiple cameras', *IEEE Int. Conf. on Image Processing*, 2011, pp. 3593-3596

42 Liu, L., Ouyang, W., Wang, X., et al: 'Deep learning for generic object detection: A survey', *arXiv:1809.02165v1*, 2018

© The Institution of Engineering and Technology 2015