A practical algorithm for learning scene information from monocular video

Lin Zhu , Jie Zhou, Jingyan Song, Zhenlei Yan and Quanquan Gu

Automation Department of Tsinghua University, Beijing,P.R.China 100084

Zhul01@mails.tsinghua.edu.cn

Abstract: The estimate of the scene information, such as the region of ground/non-ground, the relative depth of the ground and the unevenness of ground, is important for applications such as video surveillance, mapbuilding and etc. Previous research in this field is based on specific assumptions which are difficult to satisfy in practical situations. In this paper a practical algorithm is proposed to estimate the scene information in monocular video. With the pedestrian detection results for a period of time, the Pedestrian-Scene Map (PS Map), consisting of the average width of a pedestrian and occurrence probability of a pedestrian at each position of the scene, is learned by integrating the pedestrian samples with different sizes at different positions of the scene. Furthermore, the relative depth of ground region, the ground/non-ground region and the unevenness of ground can be measured with PS Map. Experimental results illustrated the proposed method's effectiveness with stationary uncalibrated camera for unconstrained environment.

OCIS codes: (100.2000) Digital Image Processing; (110.6880)Three-dimensional image acquisition; (100.2960) Image analysis; (100.6890) Three-dimensional image processing; (150.6910) Three Dimensional Sensing; (150.6044) Smart cameras; (330.5000) Vision - patterns and; recognition (330.5020) Perception psychology.

References and links

1. D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," International Journal of Computer Vision **47**(1-3), 7–42 (2002).
2. D. Forsyth and J. Ponce, in *Computer Vision : A Modern Approach*, vol. Prentice Hall (2003).
3. R. Zhang, P. S. Tsai, J. E. Cryer, and M. Shah, "Shape from shading: A survey," IEEE Trans on Pattern Analysis and Machine Intelligence **21(8)**, 690–706 (1999).
4. A. Criminisi, I. Reid, and A. Zisserman, "Single view metrology," International Journal of Computer Vision **40**, 123–148 (2000).
5. D. Hoiem, A. Efros, and M. Hebert, "Geometric Context from a Single Image," Proceedings of the IEEE International Conference on Computer Vision **2**, 1284 –1291 (2005).
6. D. Hoiem, A. Efros, and M. Hebert, "Putting Objects in Perspective," Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on **2**, 2137 – 2144 (2006).
7. M. Greiffenhagen, V. Ramesh, D. Comaniciu, and H. Niemann, "Statistical modeling and performance characterization of a real-time dual camera surveillance system," Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition **2**, 335–342 (2000).
8. S. G. Jeong and et al, "Real-Time Lane Detection for Autonomous Vehicle," IEEE International Symposium on Industrial Electronics Proceedings (ISIE 2001) pp. 1466–1471 (2001).
9. N. Krahnstoever and P. R. S. Mendonca, "Bayesian autocalibration for surveillance," Proceedings of the IEEE International Conference on Computer Vision **2**, 1858–1865 (2005).
10. A. Saxena, S. H. Chung, and Y. N. Andrew, "3-D Depth Reconstruction from a Single Still Image," International Journal of Computer Vision 2007, http://ai.stanford.edu/ asaxena/learningdepth/ .

#89851 - $15.00 USD Received 19 Nov 2007; revised 2 Jan 2008; accepted 17 Jan 2008; published 18 Jan 2008

11. "Terminology relating to traveled Surface characteristics annual book of ASTM Standards," American society for testing and material(ASTM) . (1999)
12. "High Capacity Laser Profilograph," http://www.cedex.es/cec/documenti/survey.htm .
13. S. Se and M. Brady, "Vision-based Detection of Stair-cases," Proceedings of Fourth Asian Conference on Computer Vision ACCV pp. 535–540 (2000).
14. V. Nair and J. Clark, "An unsupervised, online learning framework for moving object detection," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition **2**, 317 – 324 (2004).
15. Z. Zhou and M. li, "Tri-training: exploiting unlabeled data using three classifiers," IEEE Transactions on Knowledge and Data Engineering **17(11)**, 1529–1541 (2005).
16. P. Viola and M. Jones, "Rapid Object Detection Using a Boosted Cascade of Simple Features," Proceedings of International Conference on Computer Vision and Pattern Recognition **1**, 511–518 (2001).
17. W. H. Ittelson, "Size as a cue to distance: static localization," American Journal of Psychology **64**, 54–67 (1951).
18. A. Yonas, L. Pettersen, and C. E. Granrud, "Infants' sensitivity to familiar size as information for distance," Child Development **53**, 1285–1290 (1982).

1. Introduction

Scene understanding is an important research problem in computer vision. The scene information, such as the regions of ground/non-ground, the relative depth of the ground and the unevenness of ground, has many applications in various fields, especially for video surveillance, object activity understanding and map-building.

Many research has been conducted in recent years. Most 3D measurement works are based on stereo vision [1] or multiple images with different viewpoints, such as structure from motion [2]. For many scenes for surveillance, cheap monocular stationary cameras are widely applied. So, estimating the scene information from monocular camera is very important.

There are some approaches that can perform depth reconstruction from single images with very specific assumptions. Ref [3] assumes that region's color or texture is uniform. Ref [4] assumes that vanishing lines and points are known, and Ref [5]and [6] assume that the ground is a plane with vertical walls. In some other works, camera viewpoint or calibration information is usually needed [7] [8] [9]. Szxena et al. [10] proposed to estimate 3D depth from a single image based on a large amount of image samples with depthmaps. The depthmaps are collected by a special 3-D laser scanner. All of the assumptions in the above research are not be easily satisfied in practical situations. Further more, these algorithms can not distinguish the ground/non-ground region and the relative evenness of ground, which are also very important for scene understanding and map-building in complex situations.

Unevenness of the ground is defined in accordance with ASTM E867 [11] as the perpendicular deviation of the surface from a horizontal reference plane. In the field of pavement measurement, to calculation of surface evenness indexes, equipments such as vehicle with several laser cameras are usually used [12]. Stairs and steps are omnipresent in man-made environment. In the field of autonomous vehicles and robots, the texture detection method [13] was used to estimate the stairs within a 5m range. This method was slow and far from real-time. If the distance is far away, the texture based method may not work.

In this paper, a practical framework is proposed to estimate the scene information in monocular surveillance videos by using pedestrian detection results without well-calibrated camera and plain ground assumptions. We only need a long duration of video (6 hours in this paper) captured by monocular stationary camera and small amount of pedestrian samples (800 in this paper) are manually marked. The pedestrian detector is trained with some conventional methods and the labeled samples can be replaced by the labeled pedestrian samples in advance (or from other applications). These can be easily satisfied in practical situations. We first use a semi-supervised learning method to learn the occurrence probability and reasonable size of pedestrians at every position of the scene, which we call the Pedestrian Scene Map (PS Map). Then, PS Map is used to exploit more scene information, such as the ground/non-ground re-

#89851 - $15.00 USD Received 19 Nov 2007; revised 2 Jan 2008; accepted 17 Jan 2008; published 18 Jan 2008

gion, the relative depth of ground and the unevenness of ground. The 'ground' mentioned in this paper indicates the ground where pedestrian samples passed by.

The remainder of this paper is organized as follows: In section 2, the definition of PS Map, perspective relationship of the PS Map and detector based framework (to learn the PS Map) are described respectively. Estimate of more scene information from the PS Map is explained in section 3. The result of our experimental study is presented in section 4 and we conclude in section 5.

2. Pedestrian Scene Map

2.1. *Definition of PS Map*

The video of the scene is captured by the monocular uncalibrated stationary camera. For any point (x_i, y_i) of the scene in video, the PS Map is defined as $M(x_i, y_i) = \{w_i(x_i, y_i), p(x_i, y_i)\}$, where $w_i(x_i, y_i)$ is the average width of the pedestrian bounding box with the point (x_i, y_i) as the middle point of bottom side, $p(x_i, y_i)$ is the occurrence probability of the pedestrian bounding box with the point (x_i, y_i) as the middle point of bottom side. With the ratio of the height to width of the pedestrian bounding box K, the height of the pedestrian bounding box $h_i(x_i, y_i) = K \times w_i(x_i, y_i)$.

2.2. *Perspective relationship of PS Map*

Before learning the PS Map, we try to find the true perspective relationship between the size of object and the position it rests on. Note that, we do not assume that all objects of interest rest on the plane ground. See Fig. 2 as an example.

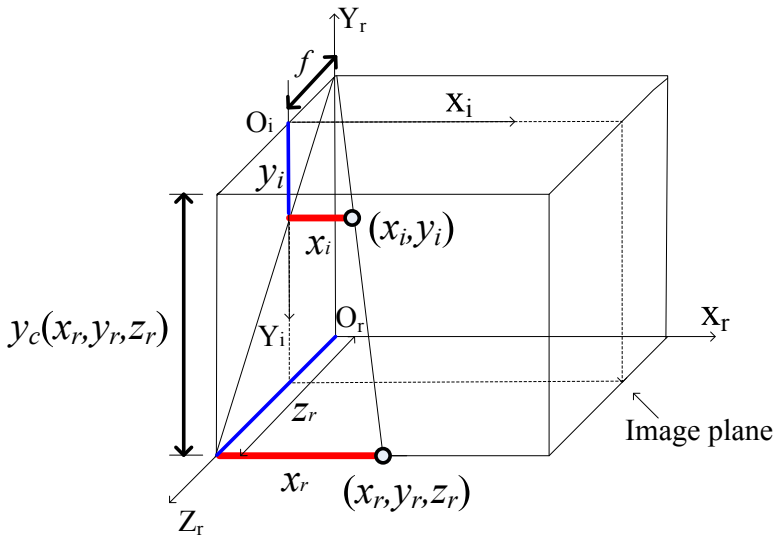

Fig. 1. The relationship between the world coordinate and the image coordinate.

In Fig. 1, (O_r, X_r, Y_r, Z_r) denotes the coordinate of real world, (O_i, X_i, Y_i) denotes the coordinate of image, and f denotes the camera focus. A camera with neither optical distortion nor roll and nor tilt is assumed. The point (x_r, y_r, z_r) on the ground surface of the scene are projected to point (x_i, y_i) in the image. $y_c(x_i, y_i)$ (or $y_c(x_r, y_r, z_r)$) denotes the camera height relative to the point of (x_r, y_r, z_r) on the ground.

#89851 - $15.00 USD Received 19 Nov 2007; revised 2 Jan 2008; accepted 17 Jan 2008; published 18 Jan 2008

As shown in Fig. 1, we can easily derive that

$$\frac{y_c(x_r, y_r, z_r)}{z_r} = \frac{y_i}{f}, \tag{1}$$

and

$$\frac{x_i}{x_r} = \frac{f}{z_r}. \tag{2}$$

Since $y_c(x_r, y_r, z_r)$ is related to (x_r, y_r, z_r), it cannot use the classic matrix notation for the stenope model.

From Eqn.2, we have

$$\Delta x_i = \frac{f}{z_r} \Delta x_r. \tag{3}$$

If the average width of pedestrians w_r ($w_r = \frac{h_r}{K}$) in the world coordinate is assumed a constant (where K is the ratio of the height to width of the pedestrian bounding box, h_r denotes the real average height of a pedestrian.), the corresponding width w_i in the image at different position (x_i, y_i) is proportional to z_r^{-1}, and z_r is the depth it rests on, i.e.,

$$w_i(x_i, y_i) = w_r f \frac{1}{z_r}. \tag{4}$$

With stationary camera (constant f), and approximate constant w_r, $w_i(x_i, y_i)$ embodies the depth information of point (x_r, y_r, z_r) on the ground surface (projected to point (x_i, y_i) in the image).

See Fig. 2 for example. There are three points (x_1, y_1),(x_2, y_2) and (x_3, y_3) shown in Fig. 2. Clearly, for point (x_2, y_2), it is on the wall. So the real occurrence probability of pedestrian $p(x_2, y_2)$ should be zero. For points (x_1, y_1) and (x_3, y_3), these points are on the ground of corridor. So the real occurrence probability of pedestrian $p(x_1, y_1) \neq 0$, $p(x_3, y_3) \neq 0$. Since the real depth of points (x_1, y_1) is smaller than that of point (x_3, y_3), the real width of pedestrian $w_i(x_1, y_1) > w_i(x_3, y_3)$.

For the point (x_i, y_i), PS Map only consider the cases that the complete pedestrian bounding box with the point (x_i, y_i) as the middle point of the bottom side. The case of partial occlusion as near the borders of the image will not be detected to estimate the PS Map.

Fig. 2. The pedestrian bounding boxes for three points at different positions in the scene.

#89851 - $15.00 USD Received 19 Nov 2007; revised 2 Jan 2008; accepted 17 Jan 2008; published 18 Jan 2008

2.3. Semi-supervised object detection

In our system, instead of estimating 3D geometry in a single image, the PS Map is learned by counting the pedestrian samples with the different size in the different position of scene for a period of time. Automatic pedestrian detection is needed for detecting the samples. Supervised training approach needs a large quantity of manually-labeled samples. In order to avoid manually labeling, some unsupervised approach makes use of the foreground blobs obtained by background subtraction as the positive training samples with the loss of accuracy [14]. Semi-supervised learning has recently become an active research area. It requires a little quantity of labeled data, and then uses unlabeled data to improve performance. Here, we investigate the use of semi-supervised learning approach, tri-training learning [15], on pedestrian detector training.

Tri-training was motivated from co-training. It generates three classifiers from the original labeled example set. Three classifiers are then refined using unlabeled examples in the tri-training process. An unlabeled example is labeled for a classifier if the other two classifiers agree on the labeling under certain conditions. This method can eliminate the requirement of co-training with sufficient and redundant views while effectively exploit unlabeled data to enhance the learning performance.

In this paper, a small quantity of manually labeled data (about 260 examples for each classifier, 800 examples totally) is used to train three initial pedestrian detectors (using the approach of Viola and Jones [16], 15 stages, detection rate is 99.5% and false positive is 0.3% for each stage). The original three classifiers can be trained off-line independently. So if the scene is changed or the camera is moved, these original classifiers need not be retrained. And for the scene to be estimated, no additional information from the human expert is needed.

For the tri-training process, by considering that the scene is monitoring by a stationary camera, foreground regions obtained by background subtraction are used for selecting training samples for tri-training learning. In this paper, unlabeled samples are selected from unlabeled video (within 6 hours, 20000 images). The flowchart of semi-supervised object detection is shown in Fig. 3. For details, please refer to Ref [15].

3. Estimate of scene information

The PS Map is very useful for pedestrian detection. For many tasks in video surveillance or robot navigation, sub-window features are used to recognize pedestrians, but may recognize false pedestrians at unreasonable positions in the scene, such as the top of the tree or the window of building. With the PS Map, the detection accuracy can be improved greatly by discarding false pedestrians with unreasonable sizes or unreasonable positions in the scene.

Additional information about the scene can also be estimated from the PS Map.

3.1. The relative depth of ground

The distribution of average width of a pedestrian reflects the relative depth of ground region in the scene. This is supported by the psychological experiment results [17] [18] and is easy to understand.

From the relationship between $w_i(x_i, y_i)$ and z_r in Eqn. 4, the depth of points (x_i, y_i) on the ground relative to the bottom of the image $D(x_i, y_i)$ can be obtained, which satisfies

$$D(x_i, y_i) \propto \frac{1}{w_i(x_i, y_i)}, \tag{5}$$

where $w_i(x_i, y_i)$ is the width of the pedestrian bounding box in the image with the point (x_i, y_i) as its middle point of the bottom side.

Further more, if the pedestrian size in the scene w_r and the focus f is known, the real depth of ground region can be obtained.

#89851 - $15.00 USD Received 19 Nov 2007; revised 2 Jan 2008; accepted 17 Jan 2008; published 18 Jan 2008

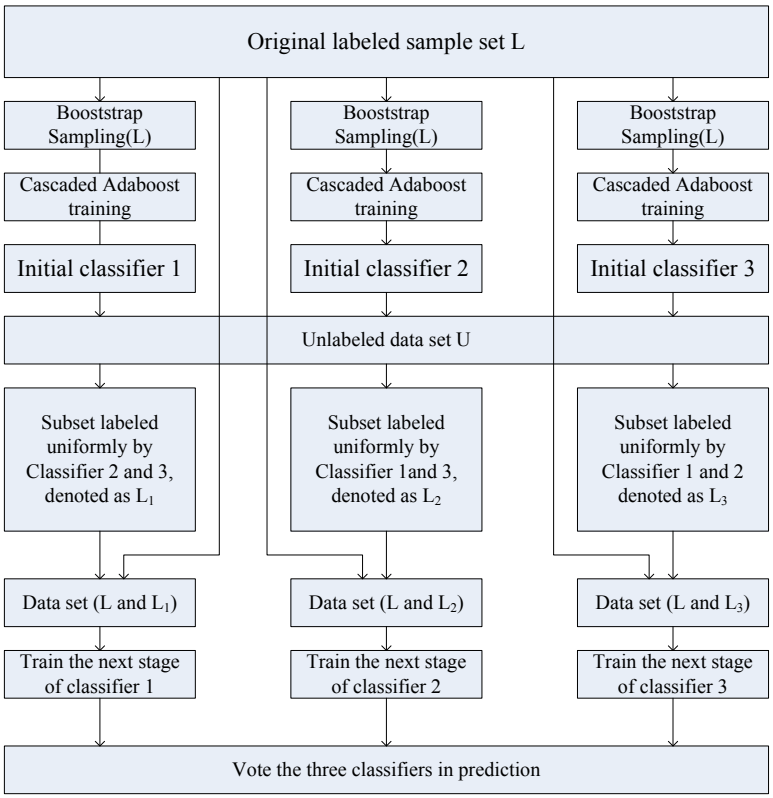

Fig. 3. The flow chart for Tri-training algorithm [15].)

3.2. The unevenness of ground

Let $s(x_r, y_r, z_r)$ denote the occurrence probability of a pedestrian with an upright pose at point (x_r, y_r, z_r) on the ground, which is assumed to homogeneous distribution without considering other factors that can influence where the pedestrian will stay such as heat, wind, et al. Pedestrians with other poses or partial occlusions can not be detected, and these can not affect the PS Map.

Let $e(.)$ denote the unevenness degree of the ground in the scene, which is defined as in Section 1. Intuitively, with more surface of ground within a unit region of horizontal plane, its unevenness degree is considered greater. So we use the unevenness degree of the ground in the scene to describe the surface area of ground in a unit region of horizontal plane. In the real unit area of ground, the number of pedestrians in one uneven place is greater than a flat place. This can help us to derive the unevenness degree of each place. From the definition of the unevenness, we know that $e(.)$ does not depend on z.

Let $\psi(.)$ denote the effect of perspective projection of the scene without regard to the unevenness of ground. It is used to describe the effect of perspective projection for flat and horizontal planes. In an image, the number of pedestrians in a far away place is larger than that in a closer place due to the perspective geometry.

The distribution of pedestrian occurrence probability $p(x_i, y_i)$ in the image is dependent on two aspects: one is the depth of position (perspective projection), and the other is the unevenness of ground surface. We have

$$p(x_i, y_i) = \psi\{e[s(x_r, y_r, z_r)]\}. \tag{6}$$

#89851 - $15.00 USD Received 19 Nov 2007; revised 2 Jan 2008; accepted 17 Jan 2008; published 18 Jan 2008

So the distribution of unevenness degree of ground in the scene can be measured by

$$e[s(x_r, y_r)] = \psi^{-1}[p(x_i, y_i)], \tag{7}$$

where the value of $e[s(x_r, y_r)]$ is larger if the ground at (x_r, y_r) is more uneven.

In order to find out the effect of perspective projection to the distribution of occurrence probability of pedestrian in the image, we try to find out the relationship between the area of one square on the ground in the real world and the corresponding area in the image, as shown in Fig. 4. Here, the ground is assumed flat and horizontal. The square on the ground in the real world has four points (denoted by 1, 2, A, B) with the coordinates of (z_{r1}, x_{r1}), (z_{r2}, x_{r2}), (z_{rA}, x_{rA}), (z_{rB}, x_{rB}); the corresponding trapezoid in the image has four corresponding points with the coordinates of (y_{i1}, x_{i1}), (y_{i2}, x_{i2}), (y_{iA}, x_{iA}), (y_{iB}, x_{iB}), and the line AB is the middle line of trapezoid.

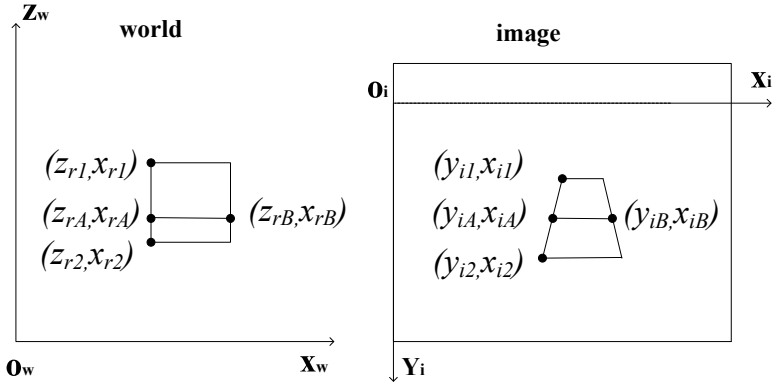

Fig. 4. A square on the ground in the real world and corresponding trapezoid in the image.

We can derive the relationship between the area of square in the world S_r and the area of corresponding trapezoid in the image S_i, as shown in Eqn. 8. The details are in the Appendix.

$$S_r \approx \frac{z_r^3}{f^2 y_c} S_i. \tag{8}$$

That is to say, due to the effect of perspective projection (depth z_r), the density of occurrence probability of a pedestrian in one point of the image(with depthz_r in the world) is approximately proportional to z_r^3, i.e.,

$$\psi(s(x_r, y_r)) \propto z_r^3, \tag{9}$$

where $s(x_r, y_r)$ is homogeneous distribution.

Then the unevenness of ground can be estimated by

$$e[s(x_r, y_r)] = \psi^{-1}[p(x_i, y_i)] \propto p(x_i, y_i) z_r^{-3}. \tag{10}$$

With the relationship between $w_i(x_i, y_i)$ and z_r in Eqn.4, the unevenness of ground can be measured using

$$e[s(x_r, y_r)] \propto p(x_i, y_i) w_i^3(x_i, y_i). \tag{11}$$

For point (x_i, y_i), the value of $e[s(x_r, y_r)]$ will be larger if the ground surface at point (x_i, y_i) is more uneven.

#89851 - $15.00 USD Received 19 Nov 2007; revised 2 Jan 2008; accepted 17 Jan 2008; published 18 Jan 2008

3.3. *The ground/non-ground region*

From the distribution of pedestrian occurrence probability, the regions of ground /non-ground can also be distinguished by

$$G(x_i, y_i) = \begin{cases} 255 & \text{if } p(x_i, y_i) \neq 0 \\ 0 & \text{if } p(x_i, y_i) = 0, \end{cases} \tag{12}$$

where (x_i, y_i) is regarded as the ground if $G(x_i, y_i) = 255$. Since the samples are not abundant in our paper, some ground locations have a limited amount of detected samples. So if pedestrian is once detected at one place, this place is regarded as ground.

4. Experimental results

To evaluate our approach, we have chosen a scene of video surveillance with a stationary IP camera. The scene is corridor with stairs (the ground is not flat) and the pedestrians have large scale changes in the scene, as shown in Fig. 5(a). Network and image compression latencies restrict the frame rate to about 1 Hz. This task is challenging for simple motion based detection methods. And the foreground can hardly be subtracted accurately due to large and abrupt lighting changes (e.g. when an office door opens).

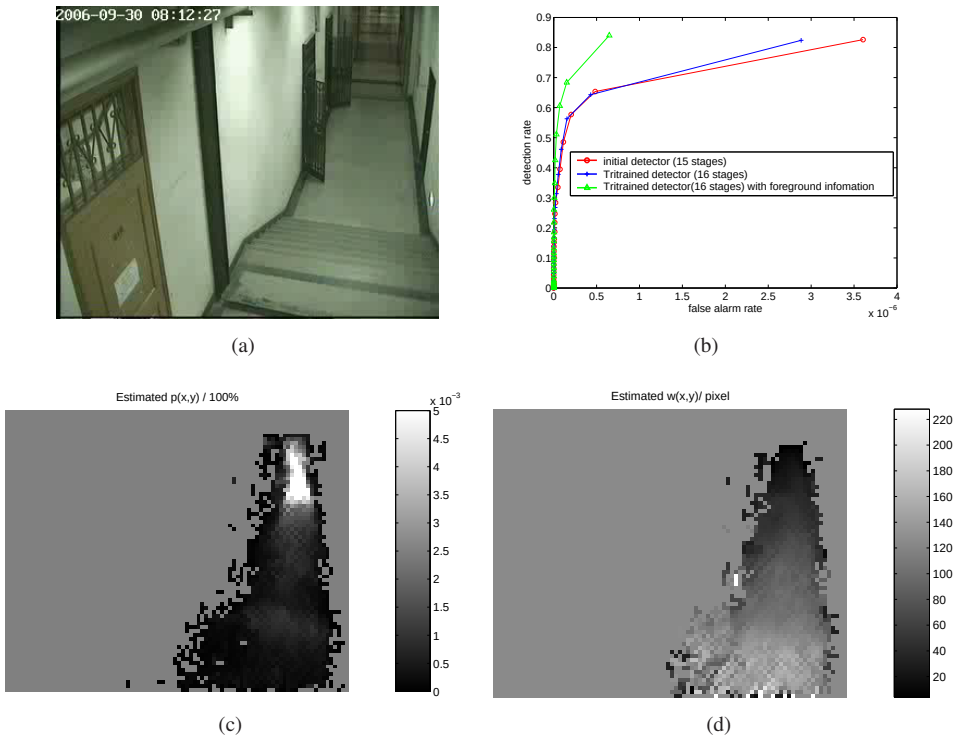

Fig. 5. (a): The corridor with stairs. (b): Receiver operating characteristic of the classifiers on the test set. (c): the estimated occurrence probability of pedestrians in the scene $p(x,y)$. (d): the estimated average width of pedestrians in the scene $w(x,y)$.

With tri-training based semi-supervised approach, a satisfying detection accuracy can be obtained with a small amount of manually labeled samples. The testing data set includes 584

#89851 - $15.00 USD Received 19 Nov 2007; revised 2 Jan 2008; accepted 17 Jan 2008; published 18 Jan 2008

images containing approximately 863 pedestrians. The count of negative patches is approximately 4.07 e +8. The false alarm rate is defined as the false alarm number divided by the total number of negative patches. The operating characteristics of the classifiers detected by the receiver using the test set are shown in Fig. 5(b). To decrease the false alarm rate, the positive patch labeled by tri-training detectors is used for scene learning only if the center of the patch is within the foreground regions. The final pedestrian detector's detection rate is 84% with false alarm rate $6.48e^{-7}$. With these steps, the false alarm rate decreases greatly, as shown in Fig. 5(b). The performance of this detector is enough to obtain the PS Map.

With the results of pedestrian detector, the PS Map can be obtained from the positive samples over time, as shown in Fig. 5 (c) and (d). In the Fig. 5(c), the brightness of each pixel embodies the value of occurrence probability of pedestrian in each position (4×4 pixel block) of the scene. The occurrence probability of pedestrian within the plane gray region is zero. In the Fig. 5(d), the brightness of each pixel embodies the average width of pedestrian bounding box with the points as the middle point of the bottom side (4×4 pixel block). The point in the plane gray region in the image represents that there is no pedestrian bounding box with the points as the middle point of the bottom side. With the increasing of the number of samples, the performance of estimated PS Map will be better.

In Fig. 6(a), the green region is the real ground region (manually labeled). According to Eqn.12, the region of ground/non-ground $G(x_i, y_i)$ is obtained, as shown in Fig. 6(b). The white region means the ground region, the black region means the non-ground region. After the estimated $G(x_i, y_i)$ (Fig. 6(b)) is smoothed by using erosion and dilation operators, the final estimation of ground region is shown in Fig. 6(c). This result shows that the ground region can be estimated by pedestrian samples over a period of time.

The real regions of stairs in the scene are manually labeled in red color, as shown in Fig. 6(d). With the estimated PS Map, according to Eqn.11, the unevenness of ground $E(x_i, y_i)$ can be measured, as shown in Fig. 6(e). The brighter point represents more uneven than the darker point. The regions with the point value over the threshold is shown in Fig. 6(f), which is just on the real positions of the stairs.

The real depth of ground region relative to the bottom side of scene is manually measured as shown in Fig. 6(g). According to Eqn.5, $D(x_i, y_i)$ the depth-relative to the bottom of the scene-of points on the ground can be estimated, as shown in Fig. 6(h). After the estimated $D(x_i, y_i)$ (Fig. 6(h)) is smoothed by using erosion and dilation operators, the final estimated depth of ground region is shown in Fig. 6(i).

The error of the estimated scene information are calculated using Eqns 13-15.

$$Error\ rate\ of\ estimated\ G(x,y) = \frac{pixel\ number\ of\ mislabeled\ in\ G(x,y)}{Total\ pixel\ number\ in\ G(x,y)}, \tag{13}$$

$$Error\ rate\ of\ estimated\ E(x,y) = \frac{Pixel\ number\ of\ mislabeled\ in\ E(x,y)}{Total\ pixel\ number\ in\ E(x,y)}, \tag{14}$$

and

$$Average\ error\ of\ estimated\ D(x,y) = \frac{\sqrt{\sum_{i=1}^{N}(d_{ei} - d_{ri})^2}}{N}, \tag{15}$$

where d_{ei} is the estimated depth value of i-th pixel, d_{ri} is the real depth value of i-th pixel, N is the total number of pixels that estimated and real depth value are both non-zero. The calculated errors of the estimated scene information corresponding to Fig. 6 are shown in Table 1.

From the experimental results, we can see that most regions of the ground are identified. The unidentified ground regions are very close to the doors, because few pedestrians would

#89851 - $15.00 USD Received 19 Nov 2007; revised 2 Jan 2008; accepted 17 Jan 2008; published 18 Jan 2008

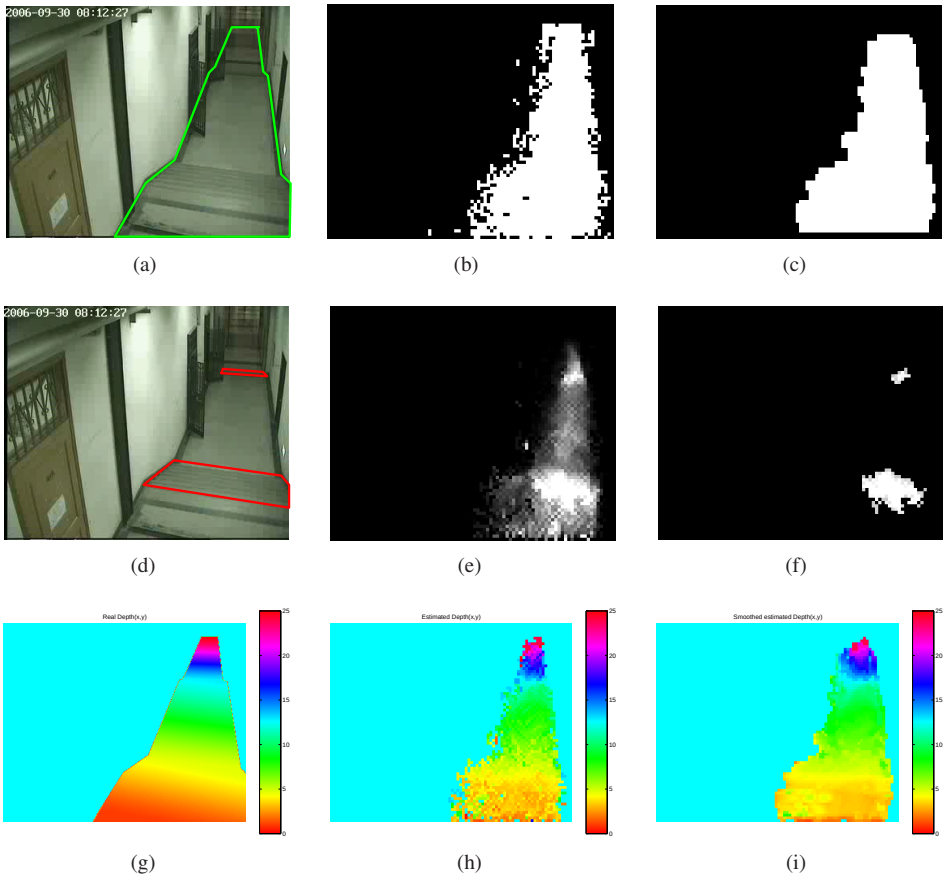

Fig. 6. Experimental results for the estimate of the corridor. (a): the corridor scene with the real ground region marked with green line. (b): coarse result of estimated ground/non-ground region. (c): final result of estimated ground/non-ground region. (d): the corridor scene with the real unevenness region marked with red line. (e): coarse result of estimated unevenness region. (f): final result of estimated unevenness region. (g): the real depth of ground relative to the bottom of the scene. (h): coarse result of estimated depth of ground relative to the bottom of the scene. (i): final result of estimated depth of ground relative to the bottom of the scene.

Table 1. Errors of the estimated scene information corresponding to Fig. 6.

Error rate of G(x,y)	Error rate of E(x,y)	Average error of D(x,y)
8.21%	5.29%	0.0884

stay in these regions. The region of uneven ground can also be identified rather well. If more pedestrian samples are used, the estimated uneven region will be much closer to the real uneven region. The estimated depth distribution of ground is very close to the real depth distribution of ground. Though pedestrians may vary in width at the same position, the width we used to estimate the depth information is the average value among the detected samples located in the same position. The research of pedestrian detection [16] showed that a satisfying detection results can be obtained by using a constant width with a standard detector. All pedestrians in the

#89851 - $15.00 USD Received 19 Nov 2007; revised 2 Jan 2008; accepted 17 Jan 2008; published 18 Jan 2008

image can be detected by scanning the detector over all positions and scales of the image. For unconstrained environments without scene geometry parameters or stereo vision information, the estimated results of using this method with the scene information are good enough to promote the development of many applications such as video surveillance, event understanding and map-building.

5. Conclusion

Scene information, such as the ground/non-ground region, the relative depth of the ground and the relative unevenness of the ground, is very important for many applications such as video surveillance, event understanding and map-building. To exploit this information from monocular surveillance video, we have proposed a practical framework to learn the scene information from the results of an object detector in video instead of from estimated 3D geometry parameters. The occurrence probability and reasonable size of pedestrians in each position of the scene are learned over time. With these results, more information about the scene can be exploited, such as the region of ground/non-ground, the relative depth of the ground and the unevenness of the ground. The proposed approach is not based on special assumptions and is more practical.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (No.60573062 and No.60673106) and the Specialized Research Fund for the Doctoral Program of Higher Education.

Appendix

The derivation of the relationship between the area of square in the world S_r and the area of corresponding trapezoid in the image S_i, is as follows.

From Eqn. 1 we can derive

$$\begin{cases} y_{i1} = \frac{f}{z_{r1}} y_c, \\ y_{i2} = \frac{f}{z_{r2}} y_c. \end{cases} \tag{16}$$

From Eqn. 2, we can derive

$$\begin{cases} x_{iA} = \frac{f}{z_{rA}} x_{rA}, \\ x_{iB} = \frac{f}{z_{rB}} x_{rB}, \\ z_{rA} = z_{rB}. \end{cases} \tag{17}$$

Since the line AB is the middle line, so the area of corresponding trapezoid in the image is

$$S_i = (x_{iB} - x_{iA})(y_{i2} - y_{i1}). \tag{18}$$

Substitute Eqn.16-17 into Eqn.18, and we get

$$S_i = \frac{f^2 y_c}{z_{rA} z_{r1} z_{r2}} S_r, \tag{19}$$

where S_r is the square area in the world.

Further more, since

$$y_{iA} - y_{i1} = y_{i2} - y_{iA}, \tag{20}$$

#89851 - $15.00 USD Received 19 Nov 2007; revised 2 Jan 2008; accepted 17 Jan 2008; published 18 Jan 2008

we can get

$$z_{rA} = \frac{2z_{r1}z_{r2}}{z_{r1}+z_{r2}}. \tag{21}$$

Substitute Eqn.21 into Eqn.19, and we get

$$S_i = \frac{f^2 y_c (z_{r1}+z_{r2})}{2z_{r1}^2 z_{r2}^2} S_r. \tag{22}$$

Since the side length of each square is very small (about 0.05 meter at the real depth of 6 meter, about 0.24 meter at the real depth of 30 meter), and the smallest ground depth of the real scene of Fig.6 is about 6 meters, it is reasonable to assume that $(z_{r2}-z_{r1}) \ll z_{r1}$, and $z_{r1} \approx z_{r2}$. Using z_r to replace z_{r1} and z_{r2}, we have the form of Eqn. 8 as

$$S_r \approx \frac{z_r^3}{f^2 y_c} S_i.$$

#89851 - $15.00 USD Received 19 Nov 2007; revised 2 Jan 2008; accepted 17 Jan 2008; published 18 Jan 2008