Multi-object Detection Based on Binocular Stereo Vision

Zhannan He[1], Qiang Ren[1], Tao Yang[1(✉)], Jing Li[2], and Yanning Zhang[1]

[1] School of Computer Science, Northwestern Polytechnical University, Xian, China
tyang@nwpu.edu.cn

[2] School of Telecommunications Engineering, Xidian University, Xian, China

Abstract. This paper proposes a new multi-object detection system based on binocular stereo vision. Firstly, we calibrate the two cameras to get intrinsic and extrinsic parameters and transformation matrix of the two cameras. Secondly, stereo rectify and stereo match is done to get a disparity map with image pairs acquired by binocular camera synchronously. Thus 3d coordinate of the objects is obtained. We then projects these 3D points to the ground to generate a top view projection image. Lastly, we propose distance and color based Mean shift cluster approach to classify the projected points, after which the number and position of objects can be determined. Binocular stereo vision based methods can overcome the problems of object occlusion, illumination variation, and shadow interference. Experiments in both indoor and corridor scenes show the advantages of the proposed system.

1 Introduction

Video surveillance is widely used in our life. It is very important in the area of public safety, traffic control, and intelligent human-machine interaction etc. How to detect multi objects accurately is one of the major concerned problems to the researchers. Monocular, binocular and multiple cameras are all used to detect objects. Existing object detection systems based on monocular vision [1–4] usually have problems in such conditions: (1) multiple objects with severe occlusion; (2) illumination variation; (3) the shadow interference. As for multi-camera based detection system [5,6], it can avoid occlusion because of multi-view and depth information. However, multi-camera based system needs additional processing, extra memory requirement, superfluous energy consumption, higher installation cost, and complex handling and implementation. For the above problems, binocular stereo vision based surveillance is a compromise between the above two systems. The binocular can solve occlusion problem and has a small computational cost and is easy to implement.

There is some work focus on stereo rectify [7,8] and stereo match [9]. Several approaches based on stereo vision have been proposed to solve object detection problems [10–16]. The work of Muñoz-Salinas et al. [10] combines information from multiple stereo cameras to get three different plan-view maps to detect objects. In [11], Cai et al. presents a new stereo vision-based model for

Z. Zhang and K. Huang (Eds.): IVS 2016, CCIS 664, pp. 114–121, 2016.
DOI: 10.1007/978-981-10-3476-3_14

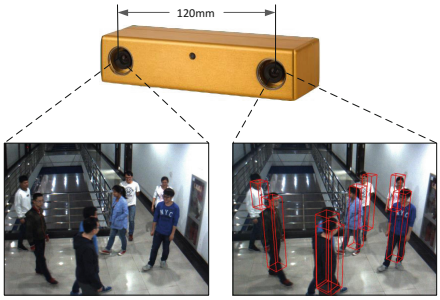

Fig. 1. The binocular stereo vision system. The left and right image are acquired by binocular stereo camera synchronously. The final result is Marked in right image with a stereo bounding box.

multi-object detection and tracking in surveillance system by projecting a sparse set of feature points to the ground plane. In [12], Schindler et al. study a system for detection and 3D localisation of people in street scene recorded from a mobile stereo rig. Colantonio et al. [13] uses a thermo-camera and two stereo visible-cameras synchronized to acquire multi-source information: three-dimensional data about target geometry, and its thermal information is combined to do object detection.

Our multi-object detection system based on binocular stereo vision is show in Fig. 1. It performs well when there exists severe occlusion, illumination variation and shadow interference. 3D coordinate of objects in the scene is obtained by binocular stereo vision, and then project it to the ground to get a top view projection image. In the projection image, the points on different objects are separated from each other so that object occlusion can be eliminated. In order to get the 3D coordinate of the object, the calibration of binocular stereo camera is needed. The framework of the proposed system is shown in Fig. 2.

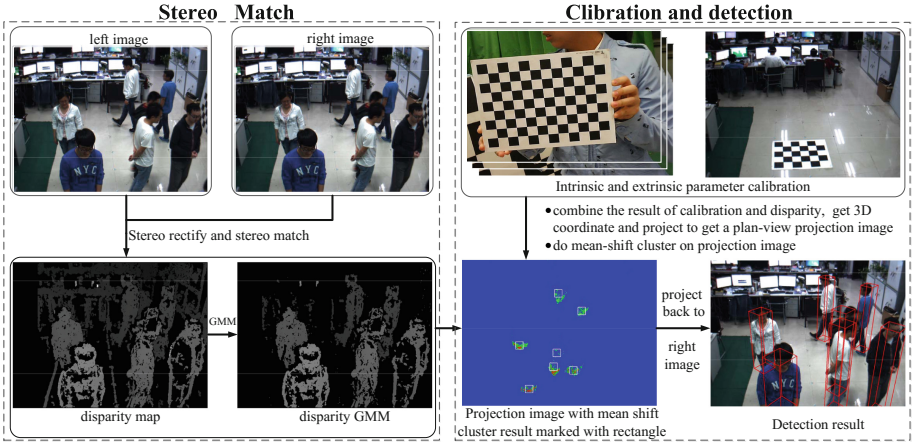

Fig. 2. The framework of the proposed binocular stereo vision system. Our work is divided into three main blocks: stereo match, calibration and detection.

The remainder of the paper is organised as follows: in Sect. 2 the calibration of binocular stereo camera is introduced. Section 3 describes the mean shift cluster method for object detection. Section 4 shows the result of out experiment. Conclusions and future works are contained in Sect. 5.

2 Binocular Stereo Vision System Calibration

In this part, we mainly talk about the camera calibration. Firstly, each camera is calibrated by Zhang's [17] planar calibration method and then rectified by Bouguet algorithm. Next, we match the same point in left and right images and get disparity map by BM algorithm. Through stereo calibration, we can get rotation and translation matrix which transforms the left camera coordinate system to the right camera coordinate system. Combining the disparity map between left and right images and calibration result, we can get the 3D coordinate of the object in the scene. The transformation of coordinate is introduced in detail below.

2.1 Camera Coordinate System to World Coordinate System

World coordinate is used to describe the position of camera, the rotation and translation matrixes between camera coordinate and world coordinate shown the transformation relationship between them. Assume that the point P's coordinate in the world coordinate system is (X_w, Y_w, Z_w), and its coordinate in the camera coordinate system is (X_c, Y_c, Z_c). According to geometric model of camera, select the right camera as the reference camera, as its center is the original point of the camera coordinate system. Hence:

$$\begin{bmatrix} X_c \\ Y_c \\ Z_c \\ 1 \end{bmatrix} = \begin{bmatrix} R & T \\ \mathbf{0} & 1 \end{bmatrix} \begin{bmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{bmatrix} = M_1 \begin{bmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{bmatrix} \tag{1}$$

Here we designate the original point of the world coordinate on the ground. R and T are the rotation and translation matrixes between camera coordinate system and world coordinate system.

2.2 World Coordinate System to Image Coordinate System

By the last step, we can get world coordinate of a point on the image. Assume that point P is the center of object on top view projection image, we need to reproject the point to the original right image to show the detection result, hence:

$$Z_c \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} f_u & 0 & u_0 & 0 \\ 0 & f_v & v_0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} R & T \\ \mathbf{0} & 1 \end{bmatrix} \begin{bmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{bmatrix} = M_2 M_1 \begin{bmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{bmatrix} \tag{2}$$

(u,v) is the coordinate of point P on right image. f_u and f_v are vertical and horizon focal length, M_2 is camera's intrinsic parameter matrix, M_1 is camera's extrinsic parameter matrix, which can be calibrated by Zhang's planar calibration method.

3 Object Detection - Mean Shift Cluster

After the binocular stereo camera calibration, we can get a projection image, in which each pixel represents one point on the ground and its value shows the quantity of points projected to the pixel. With the projection image, we can solve the object detection problem by clustering, as one cluster represents one object. We use a new distance and color based Mean shift cluster algorithm. Mean shift cluster [18,19] is a powerful non-parametric technique that does not require prior knowledge of the number of clusters and does not constrain the shape of the clusters.

The main idea behind Mean shift is to treat the points in the d-dimensional feature space as an empirical probability density function where dense regions in the feature space correspond to the local maxima or modes of the underlying distribution. It is a iterative algorithm. We use gradient ascent procedure on the local estimated density until convergence to solve the problem.

The mean shift procedure consists the following three steps:

1. choose an initial point as cluster center.
2. compute the mean shift vector from other points to this center.
3. move the center along the mean shift vector and get a new center, repeat step 1 until reach the termination condition.

When the two object are much close to each other, it is difficult to separate them by the distance based mean shift cluster.

Here we use both distance and color information of the projection image to do mean shift cluster. In this case, the kernel is $K_{h_s,h_c}(x)$ in Eq. 3.

$$K_{h_s,h_c}(x) = K(\left\| \frac{x^s - x_i^s}{h_s} \right\|) K(\left\| \frac{x^r - x_i^r}{h_r} \right\|) \tag{3}$$

To sum up, the steps are that the points are constantly moving along the direction of the probability density gradient. Mean Shift cluster can find the location of highest density by means of gradient descent.

We get n clusters centers by mean shift cluster method, each cluster represents one object. Then generate a bounding box, whose center is the cluster center. The object projection to the ground is inside the box. Combined with the height information of the object, we reproject these points back to the reference image. The detection result is shown in Fig. 3.

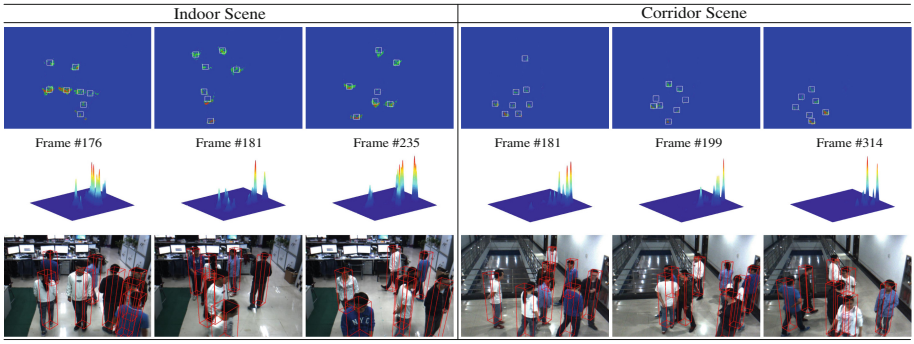

Fig. 3. The detection result by mean shift clustering. Top two rows are the 2D and 3D projection image, and the mean shift cluster results are marked with a white rectangle bounding box. Bottom row shows the detection result in original right image.

4 Experiment

In this section we show the performance of the proposed binocular stereo vision based multi-object detection system. We choose the point grey Bumblebee2 stereo vision camera (BB2-08S2C-60) and its baseline is 120 mm. The camera is placed in the ceiling with a certain angle. The area covered by the camera is 3 m × 4 m. The camera is synchronised and set to acquire images at 60 fps with a resolution of 1024 × 768 pixels. Our system has no specific requirements for the surveillance objects, and moving people is used as an example of multi-object detection in both indoor and corridor scenes, eight person are asked to walk casually.

Figure 4 shows a comparison between our system and four typical background model based methods, including AdaptiveBL [20], DP MeanBGS, Multi-LayerBG (MLBG) [21], Mixture of Gaussian V1BGS (MGV1BG) [20]. The detection results are marked with red stereo bounding boxes. It shows that our method can detect the occluded objects while other methods can't. Different objects are separate when the occlusion happens on the projection image.

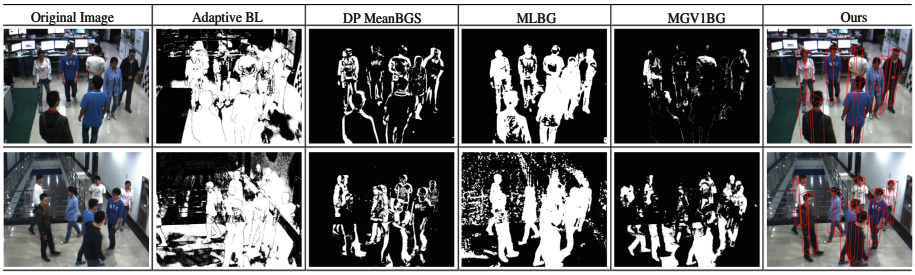

Fig. 4. Comparison to four background model based object detection methods. (Color figure online)

However, background model based methods can't separate them, as occluded objects and other objects are fused together and form a connected domain on the image.

4.1 Indoor Scene Experiment

The indoor scene always placed many things, like desks, cupboards etc., and presents a complex background on the image. Our system is a good solution to occlusion problem and complex background, as the object detection is solved by distance and color combined mean shift cluster on the top view projection image. The left part of Fig. 3 shows the detection result in indoor scenes. We can see that even though some objects are severely occluded, however, they can be detected properly. The detection results are marked with a red stereo bounding box.

4.2 Corridor Scene Experiment

As everyone knows, the corridor don't have sufficient light and the image quality is not good, especially at night. When people walking in the scene, severe occlusion, illumination and shadow change may happen. The state of art method based on monocular does not work well on this data set, but our method shows great performance.

We tested with 166 frames of indoor scene, 704 total number of real objects and 166 frames of corridor scene, 872 total number of real objects. In indoor scene experiment, only 21 real objects are not detected and there are 4 false alarms. In corridor scene experiment, our method also performs well, 25 real objects are not detected and there are 27 false alarms. The evaluation illustrates that our method shows high detection rate and low false detection rate.

5 Conclusions

In this paper, we present a multi-object detection system based on binocular stereo vision. 3D coordinate of object can be obtained by the binocular stereo vision based method. Then we can get a top view projection image. Mean shift cluster is used to determine the number and position of each object in the scene. Experiment in both indoor and corridor scenes shows that our method performs well to solve the problem of occlusion, illumination change and shadow interference. In future, we will make efforts to do track on this system.

Acknowledgements. This work is supported by the National Natural Science Foundation of China (No. 61672429, No. 61502364, No. 61272288, No. 61231016), ShenZhen Science and Technology Foundation (JCYJ20160229172932237), Northwestern Polytechnical University (NPU) New AoXiang Star (No. G2015KY0301), Fundamental Research Funds for the Central Universities (No. 3102015AX007), NPU New People and Direction (No. 13GH014604).

References

1. Szegedy, C., Toshev, A., Erhan, D.: Deep neural networks for object detection. In: Advances in Neural Information Processing Systems, pp. 2553–2561 (2013)
2. Tang, S., Andriluka, M., Schiele, B.: Detection and tracking of occluded people. Int. J. Comput. Vis. **110**(1), 58–69 (2014)
3. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 580–587 (2014)
4. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems, pp. 91–99 (2015)
5. Zhang, S., Wang, C., Chan, S.-C., Wei, X., Ho, C.-H.: New object detection, tracking, and recognition approaches for video surveillance over camera network. Sens. J. IEEE **15**(5), 2679–2691 (2015)
6. Raman, R., Sa, P.K., Majhi, B.: Occlusion prediction algorithms for multi-camera network. In: 2012 Sixth International Conference on Distributed Smart Cameras (ICDSC), pp. 1–6. IEEE (2012)
7. Kowalczuk, J., Psota, E.T., Perez, L.C.: Real-time stereo matching on CUDA using an iterative refinement method for adaptive support-weight correspondences. IEEE Trans. Circuits Syst. Video Technol. **23**(1), 94–104 (2013)
8. Nguyen, D.M., Hanca, J., Lu, S.-P., Munteanu, A.: Robust stereo matching using census cost, discontinuity-preserving disparity computation and view-consistent refinement. In: 2015 International Conference on 3D Imaging (IC3D), pp. 1–8. IEEE (2015)
9. Park, J., Choi, J., Seo, B.-K., Park, J.-I.: Fast stereo image rectification using mobile GPU. In: The Third International Conference on Digital Information Processing and Communications, pp. 485–488. The Society of Digital Information and Wireless Communication (2013)
10. Muñoz-Salinas, R., Medina-Carnicer, R., Madrid-Cuevas, F.J., Carmona-Poyato, A.: People detection and tracking with multiple stereo cameras using particle filters. J. Vis. Commun. Image Represent. **20**(5), 339–350 (2009)
11. Cai, L., He, L., Yiren, X., Zhao, Y., Yang, X.: Multi-object detection and tracking by stereo vision. Pattern Recogn. **43**(12), 4028–4041 (2010)
12. Schindler, K., Ess, A., Leibe, B., Van Gool, L.: Automatic detection and tracking of pedestrians from a moving stereo rig. ISPRS J. Photogramm. Remote Sens. **65**(6), 523–537 (2010)
13. Colantonio, S., Benvenuti, M., Di Bono, M.G., Pieri, G., Salvetti, O.: Object tracking in a stereo and infrared vision system. Infrared Phys. Technol. **49**(3), 266–271 (2007)
14. Kelly, P.: Pedestrian detection and tracking using stereo vision techniques. Ph.D. thesis, Dublin City University (2007)
15. Jafari, O.H., Mitzel, D., Leibe, B.: Real-time RGB-D based people detection and tracking for mobile robots and head-worn cameras. In: 2014 IEEE International Conference on Robotics and Automation (ICRA), pp. 5636–5643. IEEE (2014)
16. Hegger, F., Hochgeschwender, N., Kraetzschmar, G.K., Ploeger, P.G.: People detection in $3d$ Point clouds using local surface normals. In: Chen, X., Stone, P., Sucar, L.E., Zant, T. (eds.) RoboCup 2012. LNCS (LNAI), vol. 7500, pp. 154–165. Springer, Heidelberg (2013). doi:10.1007/978-3-642-39250-4_15

17. Zhang, Z.: A flexible new technique for camera calibration. IEEE Trans. Pattern Anal. Mach. Intell. **22**(11), 1330–1334 (2000)
18. Tao, W., Jin, H., Zhang, Y.: Color image segmentation based on mean shift and normalized cuts. IEEE Trans. Syst. Man Cybern. Part B: Cybern. **37**(5), 1382–1389 (2007)
19. Cheng, Y.: Mean shift, mode seeking, and clustering. IEEE Trans. Pattern Anal. Mach. Intell. **17**(8), 790–799 (1995)
20. KaewTraKulPong, P., Bowden, R.: An improved adaptive background mixture model for real-time tracking with shadow detection. In: Remagnino, P., Jones, G.A., Paragios, N., Regazzoni, C.S. (eds.) Video-Based Surveillance Systems, pp. 135–144. Springer, Heidelberg (2002)
21. Yao, J., Odobez, J.-M.: Multi-layer background subtraction based on color and texture. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2007, pp. 1–8. IEEE (2007)