Multi-Object Segmentation in a Projection Plane Using Subtraction Stereo

Toru Ubukata
Chuo University / CREST, JST
ubukata@sensor.mech.chuo-u.ac.jp

Kenji Terabayashi
Chuo University / CREST, JST
terabayashi@mech.chuo-u.ac.jp

Alessandro Moro
University of Trieste / CREST, JST
alessandoro.moro@stud.units.it

Kazunori Umeda
Chuo University / CREST, JST
umeda@mech.chuo-u.ac.jp

Abstract

We propose a method for multi-object segmentation in a projection plane. Our algorithm requires a stereo camera system called Subtraction Stereo, which extracts foreground information with a fixed stereo camera. The main contribution of this paper is how the image sequences that include partial occlusion of the foreground objects can be accurately segmented using mean shift clustering in real-time processing. The proposed method is suitable for inside a medium-sized environment, such as a room. Finally, we try to segment the sequences that include occlusion and show the accuracy of the proposed method.

1. Introduction

The goal of our work is to detect and track objects automatically for a video surveillance system. The detection of moving objects from a video stream is a basic and fundamental problem of tracking and traffic control. Object tracking is an important function of surveillance for situational recognition and the flow of pedestrians from image sequences. The studies of such systems have become increasingly popular [1, 2, 3, 4].

Foreground objects, such as pedestrians, are often occluded by each other when using a video surveillance system. Many studies deal with such occlusion problems, and there are several methods. The first is a feature-based object classification, such as the HOG feature [2, 5]. The second is optimization, such as the Markov random field [3]. These methods have been used, particularly, in recent years. Each requires a considerable amount of computation time. Instead of these methods, we have decided to use 3-D information [1, 6] obtained with a stereo camera because the surveillance system requires real-time processing for emergency situations.

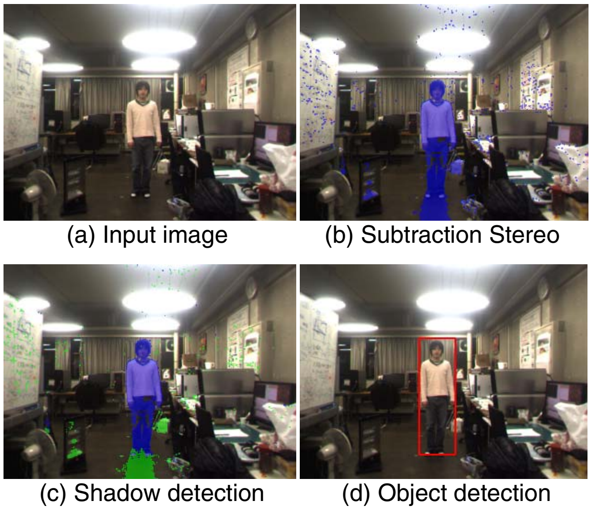
To manage the occlusion problem, we improved the plan-view analysis [1]. This method cannot segment multiple objects when the distance between two or more objects is too small (Fig. 3(a)). This problem occurs because the system cannot segment into the *projected blobs* (see Sec. 3.1) accurately. In order to segment such multiple objects, we propose using the mean shift [4], often used as a tracking method, as a clustering method on the histogram which shows the density of the projected points. In this way, we can use the 3-D information effectively.

This paper is organized as follows. In Section 2, we explain the acquisition of foreground information, which consists of Subtraction Stereo and shadow detection. In Section 3, we present a multi-object segmentation method using mean shift clustering. In Section 4, we show the experimental result for the proposed method. Section 5 is the conclusion and an explanation of future work.

2. Foreground detection

2.1. Subtraction Stereo

We use Subtraction Stereo [7] with an algorithm based on a background subtraction method and stereo matching. Subtraction Stereo extracts foreground objects in a sequence with the background subtraction method first, and stereo matching is then applied to the extracted regions. Therefore, the processing regions for stereo matching are restricted to the foreground, and we can obtain the foreground information (Fig. 1(b)) with less computation time.



(a) Input image (b) Subtraction Stereo

(c) Shadow detection (d) Object detection

Figure 1. Flow of the object detection: The blue region is in the foreground, and the green region is the detected shadow. The red bounding box shows the final detection.

2.2. Shadow detection

Shadow detection is used to refine the foreground. The image obtained using Subtraction Stereo includes noise affected by the shadow. This noise seriously affects the projection plane.

We have improved the shadow detection described in [8]. When we define $I(x,y)$ as the intensity of the pixel located in the 2-D image position (x,y) and $I'(x,y)$ as the intensity of the background pixel, the equation for the evaluation of shadow is described as

$$\theta_{(t+1,x,y)} = \begin{cases} \alpha\Psi_{(x,y)} + \beta\Lambda_{(x,y)} + (1-\alpha-\beta)\theta_{(t,x,y)}, & \\ \qquad if\ \frac{I_{(x,y)}}{\eta} < I'_{(x,y)} & \\ \infty, otherwise & \end{cases} \qquad (1)$$

where θ corresponds to a shadow value. This value will be applied a threshold to determine if a pixel is a shadow. A small shadow value corresponds to a shadow point. The functions Ψ and Λ show color constancy between pixels and within pixel. α, β, and η are constant weights of textures, colors, and intensity, which are determined empirically in our experiments. The details of this method are explained in [9]. The result of the shadow detection is shown in Fig. 1(c).

3. Multi-object segmentation

There is a problem with foreground detection in which the foreground objects occlude each other (Fig. 2(a)). When the camera position and angle are known, the camera's 3-D coordinate system can be projected to a world 3-D coordinate system. Thus, to solve the occlusion problem, foreground pixels are projected into a certain plane to usefully segment multiple objects. We call this plane the projection plane. To accurately segment multiple objects, we ultimately use mean shift clustering.

3.1. Projection plane

Foreground pixels are segmented into blobs based on 8-neighborhood connections in the foreground. After removing small blobs that are less than a threshold, the blobs are defined as $\{B_i | i = 0, \ldots, n\}$, where n is the total number of blobs.

Each pixel involved in the foreground has 3-D information obtained with a stereo camera. These pixels are projected to the world coordinate X-Y plane (Fig. 2(c)), which is selected to be useful to segment in the occluded sequence. This plane is called projection plane. In this paper, the ground plane is given as the projection plane. The projected point which was located at (x,y), is defined as $p(\mathrm{x},\mathrm{y})$, which denotes the position of the projection plane.

To deal with the projection plane easily, we create a cell defined as 5 × 5 cm in this plane. At the same time, we create a 2-D histogram (Fig. 2(d)) for each cell by counting the projected points in the cell. This histogram is defined as

$$H(c) = \left\{\sum N_{(x,y,c)} \middle| \forall(x,y) \in B_i\right\}$$
$$where\ \ N_{(x,y,c)} = \left\{\begin{array}{c} 1, if\ p(x,y) \subseteq c \\ 0, otherwise \end{array}\right\} \qquad (2)$$

where c is the region of a cell in the projection plane.

After creating the histogram, we segment the cells whose frequency is more than a threshold into the blobs that are based on 8-neighborhood connections in the projection plane. These blobs are called *projected blobs* (Fig. 2(d)). If a *projected blob* is so small that it is considered a noise, it is removed. The *projected blobs* are defined as $\{PB_j | (j = 0, \ldots, m)\}$, where m is the total number of *projected blobs*. The result of the *projected blobs* reflects 2-D image (Fig. 2(b)) because the system has a relationship between the cell position and the 2-D image position.

If it is difficult to segment multiple objects by creating *projected blobs* (Fig. 3(a, c)), the system processes the mean shift clustering described in the next section.

3.2. Mean shift clustering

The mean shift algorithm is a nonparametric

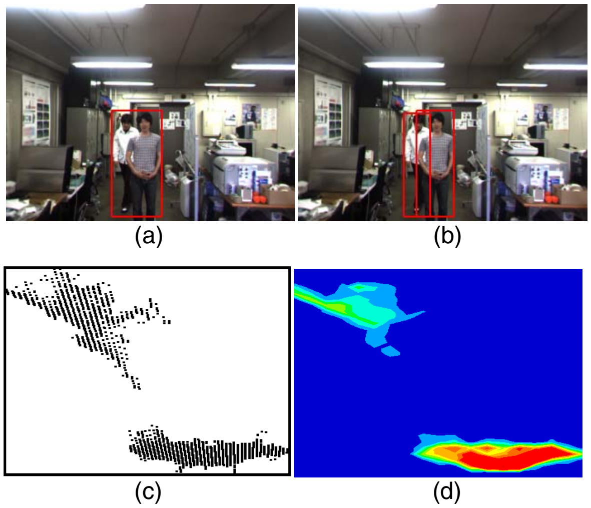


Figure 2. Comparison with and without a projection plane: (a) Error of object detection; (b) Result of using the projection plane; (c) Projected points; (d) *Projected blobs*. We show the frequency of the histogram in blue (low) to red (high) in (d). Two main blobs are evident.

clustering technique that does not require prior knowledge of the number of clusters and does not constrain the shape of the clusters. We use the mean shift clustering in the projection plane. However, mean shift clustering does not work on the projected points directly. Because the system can compute in less time, it works on the histogram created previously. Given the position vector P_c, which points to the cell position in PB_j , the mean shift vector $m(v)$ always points toward the direction of the maximum increase in the density. The mean shift vector $m(v)$is defined as

$$m(v) = \frac{\sum_{c \in rectangle} P_c \, H(c)}{\sum_{c \in rectangle} H(c)} - v \qquad (3)$$

where v is the gravity point and $H(c)$ is the frequency of the histogram in Eq. (2). $H(c)$ is used as the weight because it represents the density of the projected points. To achieve real-time processing, we define the *rectangle* simply as a kernel, which is considered a variance of the projected points of only one human at a certain distance. The center of this *rectangle* is located at v and the initial value is set evenly. The mean shift procedure is obtained by iterating the following procedures: (1) computation of the mean shift vector $m(v)$ using only the cells that are included in the *rectangle*; and (2) translation of the center of the *rectangle* $v^{t+1} = v^t + m(v^t)$. This iterative calculation is guaranteed to converge to a point in which the gradient of the density function is zero, and stops when $m(v)$ becomes sufficiently small.

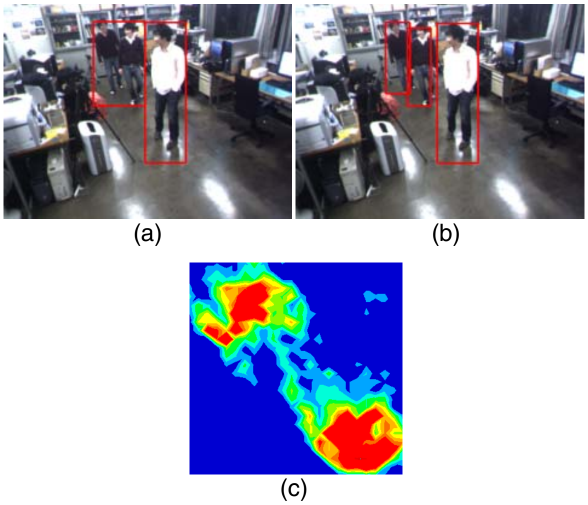


Figure 3. Comparison with and without mean shift clustering: (a) Plan-view analysis [1]; (b) Processing result using the mean shift clustering; (c) *Projected blobs* which are obtained in the left bounding box in (a). Only one main blob is visible; however, there should be two people in (b).

After the mean shift process on each PB_j , we integrate the shifted *rectangles* which converged on the same cell or nearby cell by the mean shift. Multiple objects can then be segmented by the clustering along the *rectangle*. The result of the clustering is shown in Fig. 3(b).

4. Experimental result

In this section, we show the experimental result to evaluate the accuracy of the multi-object segmentation method. We used a Point Gray Research Bumblebee2 camera with 640 × 480 pixel resolution. These videos are obtained at a rate of 30 frames per second. The results were obtained using an Intel Core2 Extreme CPU, 2.93 GHz with 3 GB ram.

For the evaluation, we analyzed 1,000 frames of two sequences (Fig. 4): in the room and in front of the elevator. These figures were selected from 1,000 frames used for the evaluation. Table 1 shows the results of our analysis. We compare our method with another one that was proposed by [1]. As is evident in Table 1, our method can detect the objects with higher accuracy than that of [1] (e.g., Fig. 3(a, b)).

Figure 5 shows examples of the error scenes. The system cannot detect an object that is fully occluded, such as that in Fig. 5(a). The system can detect a partial occluded object that is equivalent to at least one half of a human body. Figure 5(b) is the mean shift error that converges on a point of local maximum.

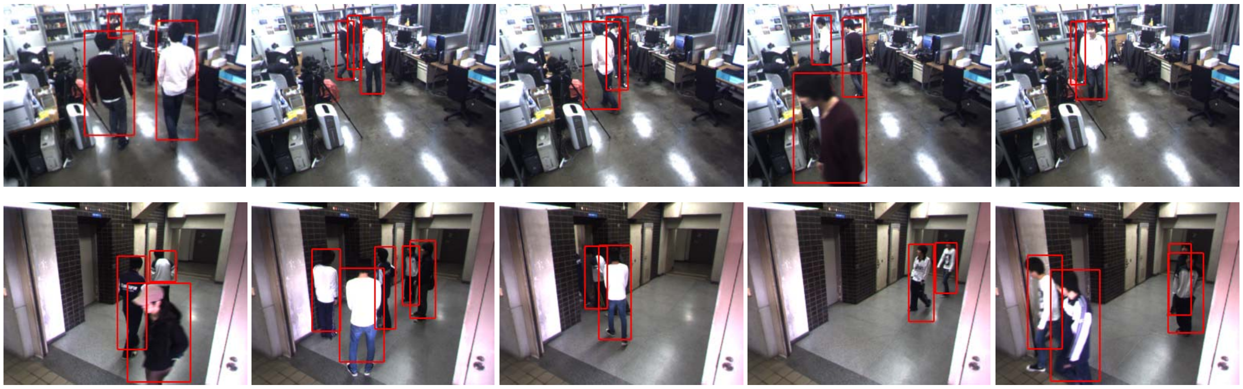
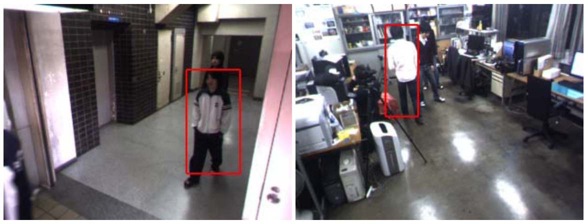


Figure 4. Experimental results: in the room (above); in front of the elevator (below)



(a) (b)

Figure 5. Examples of error scenes

The proposed method is performed in an indoor medium-sized environment because 3-D information depends on the distance from the camera. The reliability of 3-D information diminishes in proportion to the distance from the camera.

The frame rate of our new method is normally between 13 and 19 frames per second and is sufficient to integrate our tracking method [10].

5. Conclusion

In this paper, we have presented a new method of multi-object segmentation. The key factors in this paper are: (1) shadow detection removes the shadow region; (2) the projection plane which used the stereo information on the world coordinate X-Y plane can refine the foreground information; (3) mean shift clustering can segment multiple objects when the distance between the objects is insufficient. As our experiments show, the proposed method achieves multi-object segmentation with high accuracy and real-time processing.

In future work, to adjust to greater distances in outdoor sequences, we will introduce a feature-based object classification into real-time processing using Subtraction Stereo. Therefore, we can construct a more robust human tracking system for surveillance.

Table 1. Evaluation result

Sequence	T. Pos.	F. Neg.	F. Pos.
Room	88.9 %	11.1 %	0.2 %
Room [1]	80.7 %	19.3 %	4.8 %
Elevator	88.4 %	11.6 %	1.5 %
Elevator [1]	81.3 %	18.7 %	2.4 %

References

[1] S. Bahadori, et al., "Real-time people localization and tracking through fixed stereo vision," Applied Intelligence, Vol.26, No.2, pp.83-97, 2007.

[2] M. D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, L. V. Gool, "Robust Tracking-by-Detection using a Detector Confidence Particle Filter," In *Proc. ICCV*, pp. 1515–1512, 2009.

[3] W. Chaohui, et al., "Segmentation, Ordering and Multi-Object Tracking using Graphical Models," In *Proc. ICCV*, pp. 747–754, 2009.

[4] D. Comaniciu, V. Ramesh, P. Meer, "Real-time tracking of non-rigid objects using mean shift," In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Hilton Head Island, SC, 2000, pp.142–149.

[5] N. Dalal, B. Triggs, "Histograms of oriented gradients for human detection," In *Proc. CVPR*, CA,USA, 2005, pp. 886 - 893.

[6] Y. Ma, S. Worrall, A. M. Kondoz, "Depth Assisted Visual Tracking," In *Proc. WIAMIS*, pp. 157–160, 2009.

[7] K. Umeda, et al., "Subtraction Stereo -A Stereo Camera System That Focuses on Moving Regions -," In *Proc. of SPIE-IS&T Electronic Imaging*, Vol.7239 Three-Dimensional Imaging Metrology, 723908, 2009.

[8] M.-T. Yang, K.-H. Lo, C.-C. Chiang, W.-K. Tai, "Moving cast shadow detection by exploiting multiple cues," Image Processing, IET, Vol. 2, pp. 95-104, 2008.

[9] A. Moro, et al., "Auto-adaptive threshold and shadow detection approaches for pedestrians detection," In *Proc. AWSVCI,* pp. 9-12, 2009.

[10] Y. Hoshikawa, et al., "Human Tracking Using Subtraction Stereo and Color Information," In *Proc. AWSVCI,* pp. 5-8, 2009.