IPSJ Transactions on Computer Vision and Applications Vol.5 129–133 (July 2013)

[DOI: 10.2197/ipsjtcva.5.129]

Express Paper

Object Detection Based on Spatio-temporal Light Field Sensing

Atsushi Shimada[1,a)] Hajime Nagahara[1,b)] Rin-ichiro Taniguchi[1,c)]

Received: March 11, 2013, Accepted: April 24, 2013, Released: July 29, 2013

Abstract: This paper discusses about object detection based on spatio-temporal light field sensing. Our proposed method generates an arbitrary in-focus plane in the surveillance scene, and the background region can be filtered out by out-focusing. A new feature representation, called Local Ray Pattern (LRP), is introduced to evaluate the spatial consistency of light rays. The combination of LRP and GMM-based background modeling realizes object detection on the in-focus plane. Experimental results demonstrate the effectiveness and applicability for video surveillance.

Keywords: light field sensing, Local Ray Pattern, background modeling, object detection

1. Introduction

Object detection based on background modeling has often been used for visual surveillance applications. It detects a change of image signal by making a statistical model of observed pixel values. A change should be caused by detection targets such as walking people, moving cars and so on, however, background changes including waving trees, cast shadows, etc. also become a factor for false positive detection. Although traditional studies have been discussed about an effective background modeling [1], [3], there is no radical solution to overcome the problem.

Our study proposes a new sensing strategy, which enables a system to exclude background region from a space of interest at the stage of imaging [*1]. In the following of this paper, we call "space of interest" an in-focus area, where target detection is performed. Against the in-focus area, the background region is called out-focus area. Our sensing strategy can make an in-focus area with an arbitrary shape like **Fig. 1**. Therefore, the background region (a tree in Fig. 1) will be captured with a blur. To realize such an imaging strategy, we introduce a light field camera to capture light rays from a scene, and apply a technique of digital refocusing for the generation of in-focus/out-focus areas. Object detection is also performed by processing the light rays from the viewpoint of spatio-temporal light field consistency.

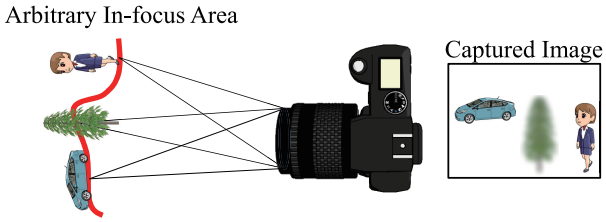

Fig. 1 The proposed in-focus area. An area in which target objects will appear is only focused. Other areas are captured with a blur.

[1] Faculty of Information Science and Electrical Engineering, Kyushu University, Fukuoka 819–0395, Japan
[a)] atsushi@limu.ait.kyushu-u.ac.jp
[b)] nagahara@ait.kyushu-u.ac.jp
[c)] rin@ait.kyushu-u.ac.jp

[*1] We intentionally use the words of "space of interest" instead of "region of interest (ROI)" since ROI has a 2D-like meaning.

2. Related Work

A light field camera was originally proposed for image-based rendering for use in the graphics community, and has been used for a variety of different visualization applications, such as computer imaging through a virtual aperture, 3D graphics, and digital refocusing [6]. In recent years, the light field camera has been applied to solve a difficult computer vision and pattern recognition problem, such as transparent object recognition [5].

A single-view camera has mainly been used for visual surveillance applications. Background modeling based object detection is one of the fundamental techniques [1], [3]. Object detection is performed on a 2D image plane, therefore, traditional approaches often suffer from background changes. Our proposed method utilizes a light field camera, and represents a detection field as a 3D space, which enables to filter out the background region from the detection field.

A 3D intrusion detection system with multiple cameras is proposed [4]. An arbitrary 3D volumetric restricted area is generated by multiple cameras surrounding the target area. Intrusion detection is achieved by computing intersections of an object and a sensitive plane, which is the boundary of the restricted area. Our proposed method also generates such a restricted area for object detection, but the camera arrangement and the detection algorithm is a lot different from the related work. The proposed method introduces a digital refocusing technique to generate a restricted area, and object detection is achieved by light ray processing.

3. Overview of the Proposed Idea

Firstly, we set up an in-focus area where object detection is performed. As shown in Fig. 1, the in-focus area does not have to be a 2D plane. For example, it is possible to intentionally set

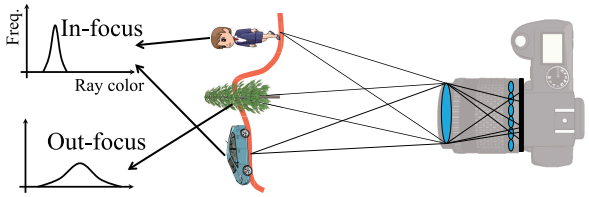

Fig. 2 Distribution of light ray colors. The distribution will have a small variance if the light rays come from an in-focus area, otherwise the variance will become larger.

up some "out-focus area," in which the background changes often occur (refer to Section 6 for an actual situation). To configure in-focus/out-focus areas, we use a digital refocusing technique with a light field camera.

Secondly, we evaluate two factors; spatial consistency and temporal consistency of observed light rays. If we assume Lambertian reflectance, the same colored rays are recorded by a light field camera. In other words, the color distribution of light rays has a small variance if they come from the same point in the in-focus area as shown in **Fig. 2**. We use this characteristic for evaluation of the spatial consistency. On the other hand, the temporal consistency of the light rays is modeled by the Gaussian Mixture background model, which is often utilized for change detection.

Finally, two evaluation results (i.e., spatial consistency and temporal consistency) are integrated to determine foreground masks which denote the object detection result. A Markov random field based approach is introduced to assign foreground/background labels for all the light rays.

4. In-focus Area Configuration

We use 4D-ray representation of a light field image $L(s,t,u,v)$ determined by the intersection of a camera plane (s,t) and a slant of ray (u,v). We use a commercial light field camera, ProFUSION 25, which has 25 VGA resolution (640×480 pixels) cameras and can simultaneously capture images from 25 viewpoints. Each image is independently recorded as a 2D image $p(x,y)$ from the respective camera located at a 2D coordinate (s,t). We obtain a light field image $L(s,t,u,v)$ by projecting these images to the parallel image coordinates prescribed by the slant (u,v). The projection from the ray images $p(x,y)$ to the slant of the light field $p_{s,t}(u,v)$ can be calculated from:

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = {}^{d}H_{s,t} \cdot K_{s,t}^{-1} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{1}$$

where $K_{s,t}$ is a matrix dependent on intrinsic parameters of each camera at (s,t). ${}^{d}H_{s,t}$ is a matrix describing the homography between two views:

$${}^{d}H_{s,t} = R_{0,0} \cdot R_{s,t}^{-1} - \frac{T_{s,t}n^{T}}{d} \tag{2}$$

where $R_{s,t}$ and $T_{s,t}$ are the rotation matrix and translation vector of the camera located at (s,t). We assume that the normal vector n is parallel to the optical axis. The notation d denotes the distance from the camera. Actually, the d is parameterized by (x,y) as $d(x,y)$ so that the homography can be independently configured on each pixel (x,y). In the following experiments, we set up the homography of each pixel semi-automatically [*2].

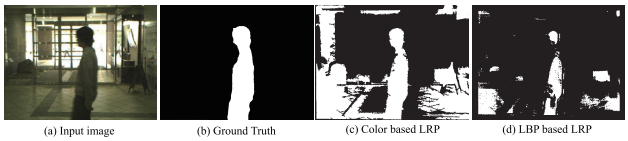

Fig. 3 In-focus/out-focus assignment based on LRP. (a) Input image, (b) Ground truth, (c) Color based LRP, (d) LBP based LRP.

5. Object Detection Based on Local Ray Pattern

5.1 Local Ray Pattern (LRP)

In this section, a new representation of lay feature is defined. A light field camera captures light rays from many different directions of a sensing space. If an object has a Lambertian surface, the same colored rays from the object are imaged on the array of cameras. The relationship of the color rays of the same coordinate (u,v) is evaluated between the basis camera $(0,0)$ and the other cameras (s,t) as follows.

$$LRP_{u,v} = \sum_{s,t} |f_{s,t}(u,v) - f_{0,0}(u,v)| \tag{3}$$

where the $f_{s,t}(u,v)$ denotes a feature of light ray (u,v) captured by the camera (s,t). The above formula calculates the similarity of the light rays. We call this light ray relationship "Local Ray Pattern (LRP)." Note that the correspondence between the light rays is well calibrated according to the strategy mentioned in Section 4. Therefore, the value of $LRP_{u,v}$ ideally becomes zero if the light rays come from the same point on the in-focus area.

5.2 Light Ray Feature

One of the simple ways to represent a light ray feature is to use the color information of light ray of (u,v). **Figure 3** (c) shows an example of in-focus/out-focus assignment based on thresholding the $LRP_{u,v}$. The pixel whose $LRP_{u,v}$ is smaller than the threshold is painted in white color. Compared with Fig. 3 (b), which represents the ideal output of the scene, we can see that the object (human) region can be detected. Meanwhile, some pixels on the out-focus area are also detected. The simple feature such as color information sometimes make $LRP_{u,v}$ smaller even if the light rays do not come from the same point. For example, a uniformly colored plane such as a wall, a floor and so on would be misdetected as shown in Fig. 3 (c).

The proposed method utilizes the light ray texture, whose idea is inspired by the Local Binary Pattern (LBP) [7]. The original LBP operator labels the pixels of an image block by thresholding the neighborhood of each pixel with the center value and considering the result as a binary number of LBP code (see **Fig. 4**). The proposed method takes the similar calculation to acquire $LBP^{P,R}_{s,t,u,v}$ by

$$LBP^{P,R}_{s,t,u,v} = \sum_{p=0}^{P-1} f(g_p - g_c)2^p \tag{4}$$

[*2] Homography is calculated for all the range of d in advance, then we manually set up the distance from the camera to generate the in-focus area.

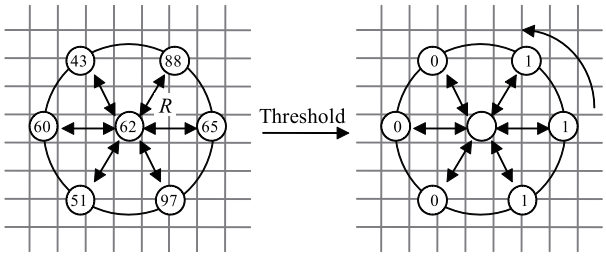

Fig. 4 An example of Local Binary Pattern.

$$f(x) = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0 \end{cases} \tag{5}$$

where g_c and g_p correspond to the brightness of the center light ray (u, v) and its surrounding P light rays which are captured in the same camera (s, t) [*3]. This kind of rich information is helpful to reduce the false positive detection (compare Fig. 3 (c) and (d)).

5.3 Light Field Background Modeling

To detect changes on the in-focus area, a background modeling based approach is introduced. A light field camera captures multiple light rays from the same point in the scene, and the calibration among them is well done to focus on the sensing area only as mentioned above. Therefore, a background model can be created for each light ray in principle at the expense of computational time. The proposed method implicitly utilizes all the rays to reduce the computational cost for model update, etc. A synthetic aperture image is generated from all the light rays, and the background modeling is performed by using the pixel value of X^t of the synthetic aperture image.

$$P(X^t) = \sum_{k=1}^{K} w_k^t \eta(X^t | \mu_k^t, \mathbf{\Sigma}_k^t), \tag{6}$$

where K is the number of distributions. The variables w_k^t, μ_k^t and $\mathbf{\Sigma}_k^t$ are an estimate of the weight, mean value and covariance matrix of the k-th Gaussian in the mixture for frame t, respectively, while η is the Gaussian probability density function. Each parameter is updated to adapt to an observed pixel value frame by frame. According to the change in pixel value, the number of distributions changes dynamically. For further details of the algorithm, refer to Ref. [8].

5.4 Object Detection in In-focus Area

The proposed object detection integrates two evaluation results acquired by LRP and background modeling. Light rays from the object region should have a smaller value of $LRP_{u,v}$ and a smaller probability of $P(X_{u,v})$ (a larger probability of $1 - P(X_{u,v})$). To evaluate these conditions, an energy function is defined according to a Markov random field and a light ray on the basis camera $(0, 0)$ [*4] is given a proper label (foreground or background) by minimizing the energy function. The energy function is defined as follows. To simplify the notation, the subscripts u, v are replaced by i, and the notation of time t is ignored.

[*3] Note that LBP is calculated in each camera (s, t) independently. Meanwhile, LRP is calculated by integrating the LBPs among the cameras.

[*4] The following formula omits s, t to simplify the notation.

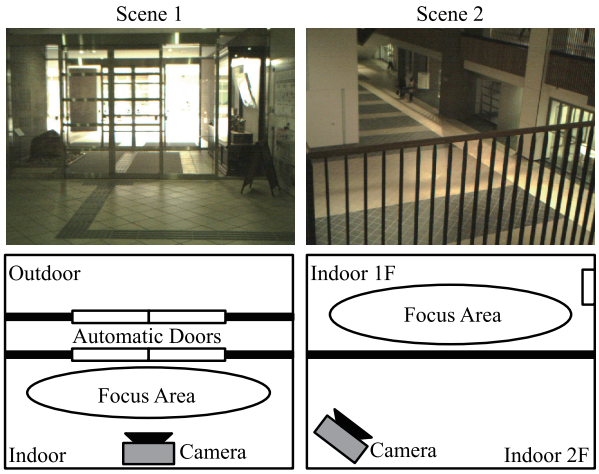

Fig. 5 Captured scenes and their overhead views.

$$E(L) = \lambda \sum_{(i) \in \mathcal{V}} G(l_i | LRP_i, X_i) + \sum_{(i,j) \in \mathcal{E}} H(l_i, l_j | X_i, X_j) \tag{7}$$

where $L = (l_1, \ldots, l_{|\mathcal{V}|})$ is a binary label vector, $l_i = 0$ and $l_i = 1$ denote the background label and the foreground label respectively. $\mathcal{V}$ and $\mathcal{E}$ represent a set of all rays and a set of four adjacent rays respectively. The penalty term $G(l_i | LRP_i, X_i)$ is actually calculated as

$$G(l_i = 0 | LRP_i, X_i) = f(g_i - Th) \tag{8}$$

$$G(l_i = 1 | LRP_i, X_i) = 1 - G(l_i = 0 | LRP_i, X_i) \tag{9}$$

where Th is a predefined threshold, and

$$g_i = \frac{1 - P(X_i)}{LRP_i}. \tag{10}$$

The $f(\cdot)$ is a step function represented by Eq. (5).

On the other hand, the smoothing term $H(l_i, l_j | X_i, X_j)$ is calculated by

$$H(l_i, l_j | X_i, X_j) = \frac{1}{\ln(|X_i - X_j| + 1 + \epsilon)}. \tag{11}$$

The smoothing term evaluates the similarity of adjacent light rays to assign the same label if they have a similar feature. The energy is minimized using a graph-cut algorithm [2].

6. Experimental Results

6.1 Outline

We captured two scenes by ourselves since there is no open dataset captured by a light field camera. The scenes were captured by ProFUSION 25 camera at 5 fps. The description of each scene is as follows (also see **Fig. 5** which shows an example shot and an overhead view of the scene.).

Scene 1 "Entrance of building" includes people moving inside/outside of an automatic door. The target is a human who comes into the building. A human walking outside, a moving door, etc. are not the detection target.

Scene 2 "Lobby of building" captures the scene from the upper floor of the lobby. Moving people appear not only in the lobby but also in the near distance from the camera. The target is a human who is walking on the lobby. A human walking on the upper floor is not the detection target.

The in-focus area of each scene was manually set up according to the method in Section 4.

Object detection results of the proposed method were compared with the following two methods.

LRP Foreground/Background is judged by thresholding $LRP_{u,v}$.

BM (Background Model) Foreground/Background is judged by background modeling with a synthetic aperture image.

6.2 Results and Discussion

Figures 6 and **7** show foreground masks for each scene and **Table 1** gives the evaluation results based on precision, recall and F-measure manner. A value given in bold type is the best F-measure among the three methods.

With regard to the Scene 1, the proposed method could detect target regions accurately, which was supported by the fact of a high precision, recall and F-measure as shown in Table 1. LRP also detected the target region as well as the proposed method, meanwhile the wall and floor region were mistakenly detected because of the same reason mentioned in Section 5.2. Uniformed texture regions were detected by LRP even using texture information. Besides the silhouette of the target was shrunk since the texture of the target contour involved not only the target region but also the background region. Such low consistency made $LRP_{u,v}$ larger, it resulted in the false negative problem. These problems were solved in the proposed method by introducing light field background modeling and MRF based smoothing. The results of BM were much worse. The biggest factor of low accuracy was detection of the automatic door. Most of background modeling approaches have suffered from this problem.

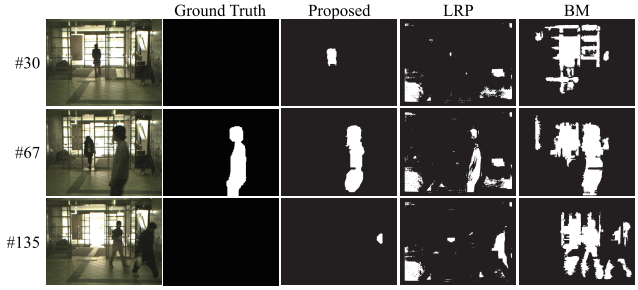

Fig. 6 Object detection result in Scene 1.

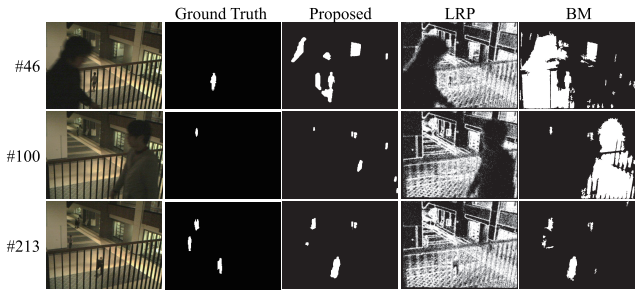

Fig. 7 Object detection result in Scene 2.

Table 1 Object detection accuracy.

		Scene 1	Scene 2
Proposed	Precision	0.83	0.17
	Recall	0.83	0.78
	F-Measure	**0.83**	**0.28**
LRP	Precision	0.23	0.01
	Recall	0.40	0.54
	F-Measure	0.29	0.03
BM	Precision	0.26	0.03
	Recall	0.91	0.87
	F-Measure	0.41	0.05

In Scene 2, a person moving in the near distance from the camera was out of target (see the column of ground truth in Fig. 7). The BM could not regard such a near distance person as out of target since it just handled temporal changes in the field of view. Evaluation of the consistency of light rays from the in-focus area solved the problem in the methods with LRP. The proposed method furthermore improved the accuracy with the combination of LRP and BM. The score of precision in Scene 2 was much worse than in Scene 1 in Table 1. The size of non-target was much bigger than the target in Scene 2. If each method mistakenly detected the non-target region (i.e., a person walking in the near distance), the precision was drastically reduced eventhough the target region was correctly detected.

Finally, the current algorithm works at about 3 fps with Intel Core i7 2.1 GHz processor. There is room for improvement of the computational strategy, e.g., calculating LRP and LBP by introducing a speed-up technique.

7. Conclusion

This paper discussed object detection based on statio-temporal light field sensing. Through the experiments, we confirmed that the proposed method had a new possibility of using light field sensing for visual surveillance. It easily solves the problems most traditional methods have suffered from, such as false positive detection of automatic doors.

In future work, further scenes need to be used in evaluating the proposed method. Besides, some supporting system to make in-focus area more easily should be considered since the current system takes much time and energy for the in-focus area setting.

Acknowledgments This work was partially supported by JSPS KAKENHI Grant-in-Aid for Challenging Exploratory Research No.25540072 and Strategic Information and Communications R&D Promotion Program (SCOPE) No.121810005.

References

[1] Bouwmans, T.: Recent Advanced Statistical Background Modeling for Foreground Detection: A Systematic Survey, *Recent Patents on Computer Science*, Vol.4, No.3, pp.147–176 (2011).

[2] Boykov, Y. and Kolmogorov, V.: An experimental comparison of min-cut/max-flow algorithms for energy minimization in computer vision, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.26, No.9, pp.1124–1137 (2004).

[3] Hassanpour, H., Sedighi, M. and Manashty, A.: Video Frame's Background Modeling: Reviewing the Techniques, *Journal of Signal and Information*, Vol.2, No.2, pp.72–78 (2011).

[4] Kawabata, S., Hiura, S. and Sato, K.: 3D Intrusion Detection System with Uncalibrated Multiple Cameras, *8th Asian Conference on Computer Vision* (*ACCV2007*), Yagi, Y., Kang, S., Kweon, I. and Zha, H. (Eds.), Lecture Notes in Computer Science, Vol.4843, pp.149–158, Springer, Berlin, Heidelberg (2007).

[5] Maeno, K., Nagahara, H., Shimada, A. and Taniguchi, R.: A Background Invariant Feature for Transparent Object Recognition, *Proc. 8th Joint Workshop on Machine Perception and Robotics* (2013).

[6] Ng, R., Levoy, M., Brédif, M., Duval, G., Horowitz, M. and Hanrahan, P.: Light Field Photography with a Hand-Held Plenoptic Camera, Technical report (2005).

[7] Ojala, T., Pietikainen, M. and Harwood, D.: A Comparative Study of

Texture Measures with Classification Based on Feature Distributions, *Pattern Recogn.*, Vol.29, No.1, pp.51–59 (1996).

[8] Shimada, A., Arita, D. and Taniguchi, R.: Dynamic Control of Adaptive Mixture-of-Gaussians Background Model, *CD-ROM Proc. IEEE International Conference on Advanced Video and Signal Based Surveillance* (2006).

(Communicated by *Tomokazu Sato*)