IPSJ Transactions on Computer Vision and Applications Vol. 3 198–210 (Dec. 2011)

Research Paper

Statistical Local Difference Pattern for Background Modeling

Satoshi Yoshinaga,[†1] Atsushi Shimada,[†1]
Hajime Nagahara[†1] and Rin-ichiro Taniguchi[†1]

Object detection is an important task for computer vision applications. Many researchers have proposed a number of methods to detect the objects through background modeling. To adapt to "illumination changes" in the background, local feature-based background models are proposed. They assume that local features are not affected by background changes. However, "motion changes", such as the movement of trees, affect the local features in the background significantly. Therefore, it is difficult for local feature-based models to handle motion changes in the background. To solve this problem, we propose a new background model in this paper by applying a statistical framework to a local feature-based approach. Our proposed method combines the concepts of statistical and local feature-based approaches into a single framework. In particular, we use illumination invariant local features and describe their distribution by Gaussian Mixture Models (GMMs). The local feature has the ability to tolerate the effects of "illumination changes", and the GMM can learn the variety of "motion changes". As a result, this method can handle both background changes. Some experimental results show that the proposed method can detect the foreground objects robustly against both illumination changes and motion changes in the background.

1. Introduction

A fundamental problem in computer vision is detecting a region or object of interest from an image sequence. Background subtraction, which removes a background image from the input image, is still widely used for detecting moving objects in practical applications. However, when it comes to outdoor surveillance, the cameras are often installed at a high place to provide a large field of view, and then their "long shot" scenes often include not only the objects but also background changes caused by illumination conditions or disturbances in these scenes. Background changes which occur in the outdoors can be classified into two types:

- **Illumination changes** – changes caused by lighting conditions such as the sun rising, setting, or being blocked by clouds;
- **Motion changes** – changes caused by the motion of, for example, tree branches, leaves, grass, waves on water or clouds.

To robustly detect the foreground objects, we should handle the background changes which occur in the outdoors. Many researchers have proposed background modeling approaches for dealing with these effects [1)–13)].

The intensity of illumination changes is often observed to be locally related to that of neighboring pixels, since illumination affects multiple pixels. Local feature-based approaches that use this characteristic have been suggested to cope with illumination changes. Early research proposed the use of edge features as a local feature for background modeling. Jabri et al. [2)] proposed using the edges of an image as well as pixel intensity for the background model. Manson et al. [3)] divided the first frame of a video sequence into blocks and calculating color edge histograms for each block. An edge feature is a derivative of image intensity, and hence is less affected by uniform illumination changes. Local Binary Patterns (LBP) [4),5)] is a well known local feature for background modeling. LBP is defined by the signed differences between a target pixel and neighboring pixels. LBP is also not affected by local intensity changes caused by illumination, since it contains a binary pattern describing lower or higher intensity relations between neighboring pixels. The distance of neighboring pixel is related to the scene context for a local feature-based method. Radial Reach Filter (RRF) [6)] extends LBP to adaptively determine the distance. These approaches assume that local features are not affected by the background changes. However, a surveillance scene also often includes motion changes, and they affect the local features in the background significantly. Therefore, it is difficult for local feature-based background models to handle motion changes in the background.

Statistical methods [7)–11)] have been used to cope with motion changes. In these approaches, background pixels are modeled by a distribution of the previously observed intensity values of each pixel. Background pixel values are usually observed with higher probabilities if we assume a foreground object is moving. When we use multiple distributions for the pixels, we can treat multi-modal

†1 Kyushu University

backgrounds caused by motion changes in a scene. A Gaussian mixture model is used for representing the multiple distributions in the literature [7),8)]. Non-parametric statistical methods [9)–11)] which use kernel density estimation have also been proposed. All of the current statistical approaches model the background pixel by pixel. Hence, there has been no research that uses statistical models for local features.

Some hybrid methods [12),13)], which use multiple different background models, have been also proposed. To avoid falsely classifying the object regions as background, Yoshimura et al. [12)] used a local feature-based background model in addition to the one focused on each pixel, and combined the results of them using a "logical OR" operation. However, their method tends to falsely detect the background regions as object regions. On the other hand, to cope with both illumination and motion changes in the background, Tanaka et al. [13)] used both local feature-based and statistical background models, and the results of them were combined using a "logical AND" operation. Then, their method divides the foreground regions, since only positive regions from both algorithms are accepted and all other regions are rejected. Therefore, these methods are a kind of tandem system, and a logical combination of the detection results does not lead to an improvement of the accuracy of the foreground detection.

In this paper, we propose a new background model suitable for outdoor surveillance[⋆1]. We combine the concepts of a local feature-based approach and a statistical approach into a single framework. This new framework for background modeling is the main contribution of this work, and it is completely different from previous hybrid methods [12),13)]. Our method uses illumination invariant local features, and describes their distribution by Gaussian Mixture Models (GMMs). The local feature has the ability to tolerate the effects of illumination changes, and the GMM can learn the variety of motion changes. Therefore, our proposed method can detect the foreground objects robustly against both illumination and motion changes. This is also our contribution, and we expect that our method can support a high recall ratio and high precision ratio at the same time.

⋆1 Our target scenes are mainly "long shot" scenes in the outdoors, and our proposed method is not intended for "close-up shot" scenes such that a foreground object is very large.

2. Statistical Local Difference Pattern

In the proposed model, we apply a Gaussian Mixture Model (GMM) to a local feature called the *Local Difference* (LD) to get a statistical local feature called the *Statistical Local Difference* (SLD). Finally, we define *Statistical Local Difference Pattern* (SLDP) for the background model by using several SLDs (see **Fig. 1**). In Section 2.1, we explain the concept and advantages of SLDP. The construction of LD is discussed in Section 2.2, and the representation of SLD using GMM in Section 2.3. Finally, we explain the construction and detection rules for SLDP in Section 2.4.

2.1 Concept of Statistical Local Difference Pattern

Previous statistical approaches [7)–11)] can handle multi-modal backgrounds but not illumination changes. Conversely, local feature-based approaches [4)–6)] can deal with illumination changes but not multi-modal backgrounds.

To solve these problems, we propose a new background model by applying a statistical framework to a local feature-based approach as shown in Fig. 1. **Figure 2** shows the advantages of using SLDP. In most cases where illumination changes, there are small changes in the difference between a target pixel and its neighboring pixel, since the values of pixels in a localized region increase or

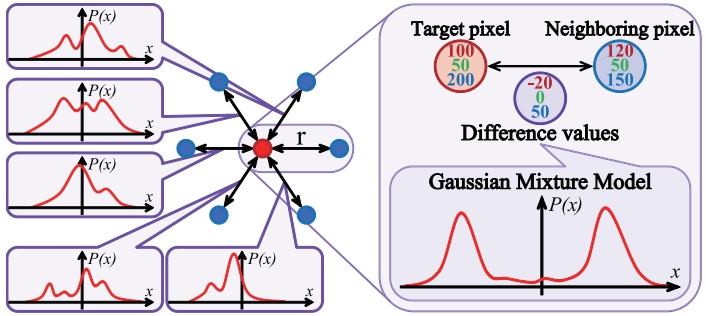
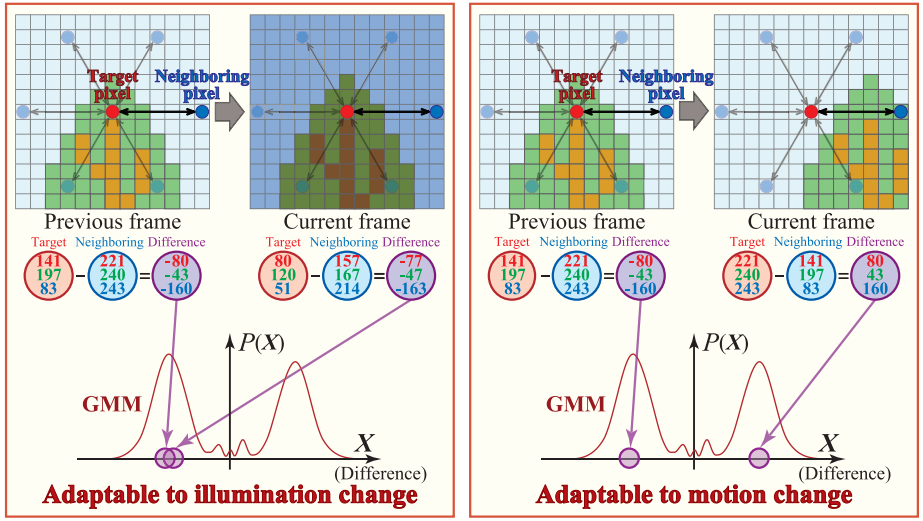


Fig. 1 Proposed background model based on *Statistical Local Difference Pattern*: Local Difference (LD) is a local feature, and is defined by the difference between a target pixel and a neighboring pixel. LD is modeled using a GMM to represent its distribution, making it a statistical local feature called the Statistical Local Difference (SLD). Our proposed model defines the Statistical Local Difference Pattern (SLDP) using several SLDs for the background model (this figure shows an example with six SLDs).



(a) Illumination change (b) Motion change of the background

Fig. 2 Adaptivities of the proposed model to background fluctuation: (a) shows the case of illumination changing suddenly (e.g., when sunlight is blocked by clouds). SLDP can adapt to illumination changes. This is because LD has the ability to tolerate the effects of illumination changes which affect the target pixel value in proportion with others. (b) shows the case of texture changing periodically (e.g., the effect of movement of tree or grass). GMMs can adapt to these kinds of motion changes, since they can learn the variety of background hypotheses.

decrease proportionally. Due to the invariance of the difference value with respect to illumination changes, SLDP has the ability to tolerate the changes as shown in Fig. 2 (a), since it uses the difference value as a local feature. Furthermore, our proposed method can also cope with motion changes, since SLDP can learn the variety of the changes as shown in Fig. 2 (b). This is because a GMM, which can handle a multi-modal background, is applied to LD which is an important component of SLDP. Thus, our background model can combine the concepts of both statistical and local feature-based approaches into a single framework.

2.2 Construction of Local Difference

A target pixel and a neighboring pixel in an observed image are described by the vectors $\boldsymbol{p}_c = (x_c, y_c)^T$ and $\boldsymbol{p}_j = (x_j, y_j)^T$ respectively. We can then represent a local feature $\boldsymbol{X}_j$, called the *Local Difference* (LD), by using the difference between the target and neighboring pixel:

$$\boldsymbol{X}_j = f(\boldsymbol{p}_c) - f(\boldsymbol{p}_j), \tag{1}$$

where $f(\boldsymbol{p})$ is the image intensity at pixel $\boldsymbol{p}$.

In cases where illumination changes occur, the changes in the LD are smaller than the pixel values, since the pixels in the localized region show a similar change. Therefore, the value of LD is more stable than each pixel value under the illumination changes.

2.3 Construction of Statistical Local Difference

We apply a Gaussian Mixture Model (GMM) to LD to represent probability density functions (PDF) for LD. This gives a statistical local feature called *Statistical Local Difference* (SLD). We define the SLD $P(\boldsymbol{X}_j^t)$ (PDF for LD) at time t by:

$$P(\boldsymbol{X}_j^t) = \sum_{k=1}^{K} w_{j,k}^t \eta(\boldsymbol{X}_j^t | \boldsymbol{\mu}_{j,k}^t, \boldsymbol{\Sigma}_{j,k}^t), \tag{2}$$

where $w_{j,k}^t$, $\boldsymbol{\mu}_{j,k}^t$ and $\boldsymbol{\Sigma}_{j,k}^t$ are the weight, the mean and the covariance matrix of the k-th Gaussian in the mixture at time t respectively, and η is the Gaussian probability density:

$$\eta(\boldsymbol{X}_j^t | \boldsymbol{\mu}_j^t, \boldsymbol{\Sigma}_j^t) = \frac{1}{(2\pi)^{\frac{d}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} (\boldsymbol{X}_j^t - \boldsymbol{\mu}_j^t)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{X}_j^t - \boldsymbol{\mu}_j^t) \right). \tag{3}$$

We construct the background model by updating the GMM (that is, the SLD). The updating method for the GMM is based on the statistical approach proposed by Shimada et al. [8]. This method allows automatic changes of K (the number of Gaussian distributions) in response to background changes. That is, K increases when the background has many hypotheses because of motion changes, for example. On the other hand, when pixel values are constant for a while, some Gaussian distributions are eliminated or integrated, and K consequently decreases.

2.4 Background Model Using a Statistical Local Difference Pattern

In our proposed method, each pixel has a pattern of SLD in the background model. We call this pattern of SLD the *Statistical Local Difference Pattern*

(SLDP), and SLDP $\boldsymbol{S}^t$ at time t is defined as follows:

$$\boldsymbol{S}^t = \{P(\boldsymbol{X}_1^t), \ldots, P(\boldsymbol{X}_j^t), \ldots, P(\boldsymbol{X}_N^t)\}, \tag{4}$$

where N represents the number of SLDs (Fig. 1 shows an example in which $N = 6$). The N SLDs $P(\boldsymbol{X}_j^t)$ $(j = 1, \ldots, N)$ are defined using a target pixel $\boldsymbol{p}_c = (x_c, y_c)^T$ and N neighboring pixels $\boldsymbol{p}_j = (x_j, y_j)^T$. When a directional vector $\boldsymbol{a}_j$ $(j = 1, \ldots, N)$, which describes the direction from the target pixel to each neighboring pixel, is defined as

$$\boldsymbol{a}_j = \left(\cos\frac{j-1}{N}2\pi, \sin\frac{j-1}{N}2\pi\right)^T, \tag{5}$$

then the neighboring pixel $\boldsymbol{p}_j$ is given by:

$$\boldsymbol{p}_j = \boldsymbol{p}_c + r\boldsymbol{a}_j. \tag{6}$$

In Eq. (6), r is a radial distance, and all of the neighboring pixels lie on a circle of radius r centered at a target pixel $\boldsymbol{p}_c$. We can also refer to N as the number of neighboring pixels.

Foreground detection using SLDP uses a voting method to judge whether a target pixel $\boldsymbol{p}_c$ belongs to the background or the foreground. When the pattern of N LDs is given as $\boldsymbol{D}^t = \{\boldsymbol{X}_1^t, \ldots, \boldsymbol{X}_j^t, \ldots, \boldsymbol{X}_N^t\}$, foreground detection based on SLDP is decided according to:

$$\Phi(\boldsymbol{p}_c) = \begin{cases} \text{background} & \text{if } \phi(\boldsymbol{D}^t|\boldsymbol{S}^t) \geq th, \\ \text{foreground} & \text{otherwise,} \end{cases} \tag{7}$$

where th is a threshold for determining whether a target pixel $\boldsymbol{p}_c$ belongs to the background or the foreground. In Eq. (7), $\phi(\boldsymbol{D}^t|\boldsymbol{S}^t)$ is a function which returns a value between 0 and 1, and is defined by

$$\phi(\boldsymbol{D}^t|\boldsymbol{S}^t) = \frac{1}{N}\sum_{j=1}^{N}\psi(\boldsymbol{X}_j^t), \tag{8}$$

where $\psi(\boldsymbol{X}_j^t)$ is a function which returns 0 or 1, depending on whether or not the LD $\boldsymbol{X}_j^t$ matches the SLD $P(\boldsymbol{X}_j^t)$ at time t. The LD is said to match the SLD if it falls within 2.5 standard deviations of the mean. For further details, we refer the reader to the literature [8].

3. Experimental Result

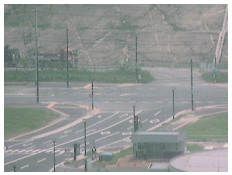
We conducted four types of experiments. First, we compared the overall foreground detection performance of our method with competing approaches. Second, we evaluated the validity of our method using Wallflower dataset [1]. Third, we investigated the effect of the parameters r and N on foreground detection. Finally, the robustness of the method for all types of background changes, illumination changes and motion changes, was examined.

Except for the validation using Wallflower dataset in Section 3.2, the datasets for the five outdoor scenes illustrated in **Fig. 3** were used. As we can see from Fig. 3, they are long shot scenes, and are the targets for our proposed background model. Scene1 and scene2 are taken from PETS (PETS2001)⋆1, while scene3, scene4 and scene5 are our original datasets which are available from our website⋆2. The PETS datasets involve not only pedestrian movement though the streets, but also illumination changes (sunlight blocked by clouds) and motion changes (tree swaying and cloud movement) in the background. Our original datasets include

(a) scene1 (PETS2001)

(b) scene2 (PETS2001)

(c) scene3



(d) scene4

(e) scene5

Fig. 3 The datasets for evaluation.

⋆1 Benchmark data of the International Workshop on Performance Evaluation of Tracking and Surveillance. Available from ftp://pets.rdg.ac.uk/PETS2001/

⋆2 Several kinds of test image are available from http://limu.ait.kyushu-u.ac.jp/dataset/

Table 1 Two kinds of performance evaluation results for foreground detection based on Recall, Precision and the F-measure: one is with respect to each dataset (scene) and the other evaluates whole datasets (scenes).

		PETS datasets		Our original datasets			Total
		scene1	scene2	scene3	scene4	scene5	
Proposed method	Recall	72.0	77.6	80.9	57.1	28.1	46.9
	Precision	88.9	62.4	80.5	92.9	79.3	80.6
	F-measure	79.6	69.2	80.7	70.8	41.5	59.3
Hybrid method [13]	Recall	38.6	51.1	68.9	42.2	22.5	34.2
	Precision	89.9	75.0	81.0	96.3	88.6	86.6
	F-measure	54.0	60.8	74.5	58.7	35.9	49.1
GMM method [8] (proposed without local feature-based approach)	Recall	72.6	68.6	88.1	54.4	27.8	45.1
	Precision	38.1	32.1	67.3	88.7	76.2	59.8
	F-measure	50.0	43.8	76.3	67.4	40.7	51.4
LM method (proposed without statistical approach)	Recall	15.2	17.5	23.6	21.2	15.0	17.2
	Precision	8.4	41.4	91.1	87.2	87.9	48.8
	F-measure	10.8	24.6	37.4	34.2	25.6	25.4

several different sizes of moving objects such as pedestrians, cars, buses, etc.

3.1 Comparison with Previous Methods

We compared the overall performance of foreground detection with three different approaches, the GMM method [8], the Local Magnitude (LM) method and the Hybrid method [13]. The GMM method [8] removes the local feature-based framework from our proposed method, and is consistent with a statistical approach using Gaussian Mixture Models. The LM method removes the statistical framework from our proposed method, and models local magnitude relations between a target pixel and its neighboring pixels. The Hybrid method [13] combines a statistical model and a local feature-based model. We used the GMM and LM methods to evaluate the effectiveness of the statistical and local feature-based approaches, respectively. The Hybrid method [13] was used to indicate that our new framework is better than hybrid approaches which used the ad hoc solutions by logical combination.

In these experiments, the radial distance is $r = 10$, the number of neighboring pixels is $N = 6$ and the detection threshold is $th = 5$. Although the details of GMM are not explained in Section 2.3, we also indicate the parameter settings in GMM for reproducibility: the learning rate is $\alpha = 0.05$, the initial weight is $W = 0.05$ and the threshold of choosing the background model $T = 0.7$. For details of GMM, we refer the reader to the literature [8]. The effects of varying the parameters r and N are investigated in Section 3.3.

Three measures, Recall, Precision ratio and the F-measure, were used for evaluation against manually-produced ground truth datasets[⋆1], and are calculated as follows:

$$\text{Recall}\,(\%) = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \times 100, \tag{9}$$

$$\text{Precision}\,(\%) = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \times 100, \tag{10}$$

$$\text{F-measure}\,(\%) = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \times 100, \tag{11}$$

where **True Positive**, **False Positive** and **False Negative** represent the number of pixels which are correctly classified as foreground, the number of pixels which are incorrectly classified as foreground and the number of pixels which are incorrectly classified as background, respectively. The recall ratio is the fraction of the foreground pixels detected correctly, and the precision ratio is the frac-

⋆1 A ground truth image denotes the foreground regions which should be detected by background subtraction. The ground truth datasets for several benchmark datasets, including those used in this paper, are published on http://limu.ait.kyushu-u.ac.jp/dataset/

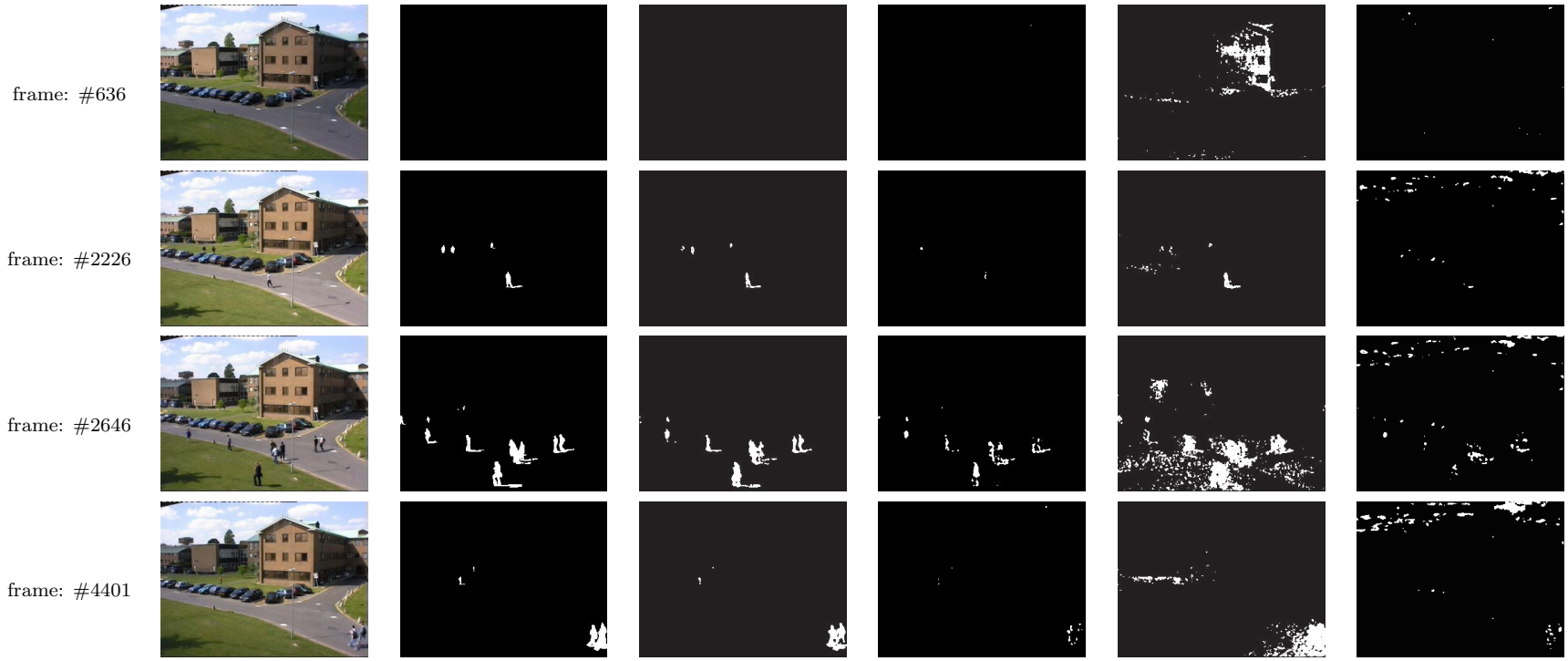


Fig. 4 The results of foreground detection for the proposed method, the Hybrid, GMM and LM methods.

tion of detected pixels which belong to the foreground. The F-measure is the harmonic mean of Precision and Recall. **Table 1** shows two kinds of performance evaluation results for foreground detection based on Recall, Precision and F-measure, one with respect to each scene (dataset) and the other evaluating whole scenes (datasets). To demonstrate the experimental results, we show the results of foreground detection for scene1 in **Fig. 4**.

The PETS datasets (scene1 and 2) include illumination changes and motion changes in the background region. Table 1 shows that our proposed method and the Hybrid method both achieve high precision ratios for the PETS datasets, since they can adapt to both illumination and motion changes. Therefore, little noise is detected in Fig. 4 (c) and (d), which further shows that these methods can adapt to both types of background change. On the other hand, Table 1 shows that the GMM and LM methods have low precision rates for the PETS datasets. In the case of GMM, it cannot adapt to illumination change and detects a number of noises (see Fig. 4 (e)-frame #636, #2646 and #4401). On the other hand, LM cannot adapt to motion changes, and Fig. 4 (f) shows that it detects cloud movement (note especially the area of sky in Fig. 4 (f)-frame #2226, #2646 and #4401).

In the cases of our original datasets (scene3, 4 and 5), neither illumination changes nor motion changes are severe. Therefore all of the methods achieve a

high precision ratio for these datasets in Table 1. In terms of the recall ratio, however, there are differences between the methods in Table 1. In the case of the LM method, it is robust against illumination changes but it has difficulty detecting entire foreground objects because the operator is too simple. Therefore, Table 1 shows that the LM method has the lowest recall ratio for all datasets. Table 1 also shows that the recall ratio for the Hybrid method is lower than for our proposed method and for the GMM method. This is because the Hybrid method combines the results of several different models using the "logical AND" operation, and false-negative pixels in either of the two models reduce the true-positive rate. This is confirmed in Fig. 4 (d)-frame #2226, #2646 and #4401, in which there are a number of false-negative pixels in the object regions. In the cases of our method and the GMM method, their models can be constructed using a single framework, and therefore they maintain a high recall ratio. As a result, Table 1 shows that our proposed method and the GMM method both achieve a high recall ratio over whole scenes (datasets). For these reasons, we conclude that our proposed method can combine the best aspects of both local feature-based and statistical approaches.

3.2 Validation Using Wallflower Dataset

In this section, to investigate what kind of scenes our proposed method can handle apart from our target scenes, we have used Wallflower dataset [1)⋆1]. This dataset contains not only long shot scenes but also close-up shot scenes which observe a large foreground, and includes the background changes which are not observed in the outdoor. Regarding the parameters, we employed the same ones as used in Section 3.1. We show the visual and numerical results in **Fig. 5** and **Table 2** respectively, in which the results of Wallflower are cited from its literature [1)]. In Table 2, the column of total errors indicates the summation of false-negative and false-positive pixels in each scene.

With respect to total errors, the performance of our method is lower than Wallflower. This is because our method detects many false-negative pixels in three close-up shot scenes: "Light Switch", "Camouflage" and "Foreground Aperture", as shown in Fig. 5 and Table 2. These close-up shot scenes are not the targets for our method, and it is difficult to robustly detect the foreground objects. One reason for this is that the SLDP does not model the background color but rather the difference between a target pixel and its neighboring pixels. In most cases of close-up shot scenes, the background has a uniform texture, and

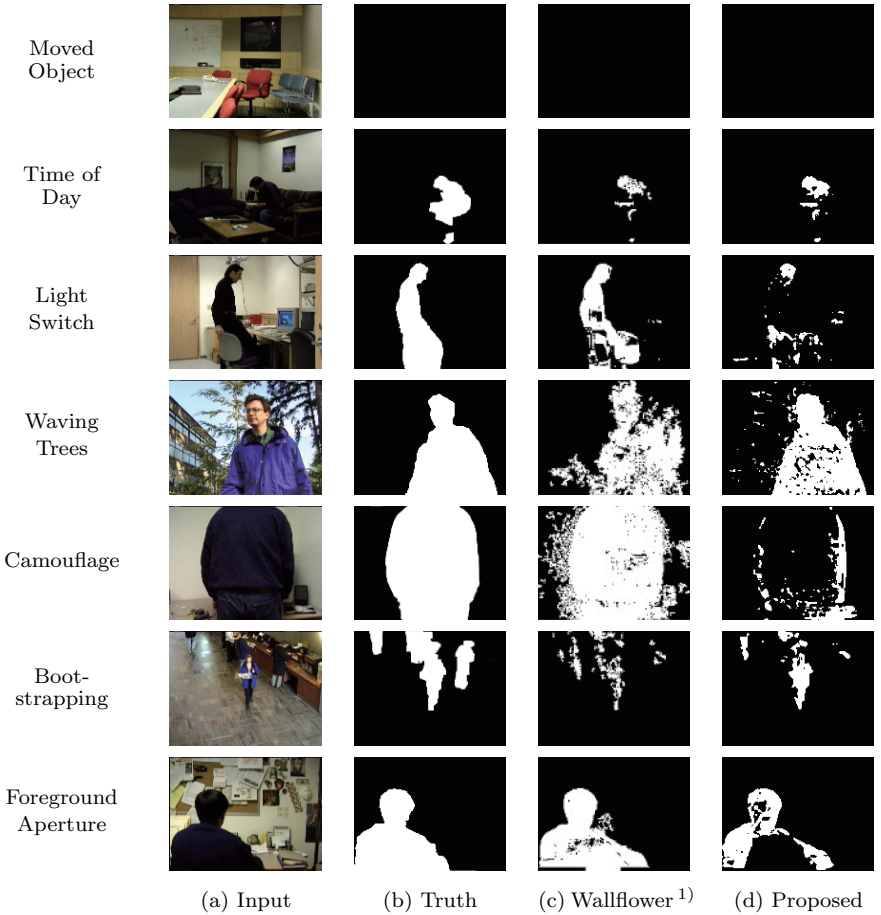


Fig. 5 The results of foreground detection for the proposed method and Wallflower [1)].

⋆1 Wallflower dataset contains images and their ground truth data for various background subtraction issues.

Table 2 Performance evaluation using Wallflower dataset.

Algorithm	Error Type	Moved object	Time of Day	Light Switch	Waving Trees	Camouflage	Bootstrap	Foreground aperture	Total Errors
Wallflower [1)]	False Negative	0	961	947	877	229	2025	320	
	False Positive	0	25	375	1999	2706	365	649	11478
Proposed method	False Negative	0	791	2369	600	8887	1439	2690	
	False Positive	0	44	280	788	387	132	553	18960

then the change in the SLDP is hardly-detectable when an object with a uniform texture appears. Therefore, our method mistakenly regards the foreground object as the background in the close-up shot scenes, and that is a limitation of our method. Another reason is that our method has no post-processing for complementing the object regions with a spatiality such as color similarity. In the case of Wallflower, as the post-processing phase, the method complements the object regions with color similarity and can achieve a reduction of false-negative pixels. If we adopt a post-processing such as Graph-Cut, etc., we expect that our proposed method can also reduce the number of false-negative pixels.

On the other hand, in scenes other than those listed above, our method can outperform Wallflower as we can see from Fig. 5 and Table 2. This is because these scenes except for "Waving Trees" are long shot scenes, and "Waving Trees" has a relatively complicated texture in its background. Then the SLDP can distinguish the object regions from the background without confusing the object with the background as discussed above. These scenes also involve illumination and motion changes in the background. The SLDP can adapt to both changes, and therefore our method detects few false-positive pixels. From the results of Section 3.1 and this section, we can confirm that our method can detect foreground objects accurately in the long shot scenes and the scenes which have a relatively complicated texture in their background.

3.3 Analysis of SLDP Parameters

Our proposed model is based on the SLDP, which has two important parameters. One is the number of neighboring pixels N and the other is the radial distance r. We examined the accuracy of foreground detection by changing these parameters as they are thought to affect the accuracy of our proposed method. In this section, Recall, Precision ratio and the F-measure were used to evaluate the results.

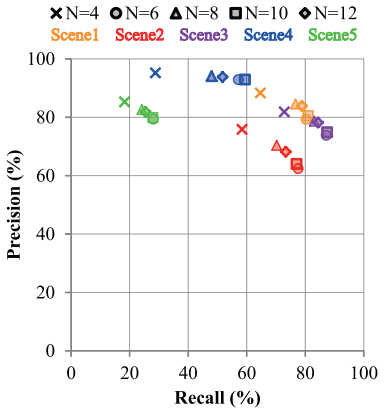


Fig. 6 Foreground detection accuracy in relation to the number of neighboring pixels N. The difference in symbol and color represents the difference in N and scene respectively.

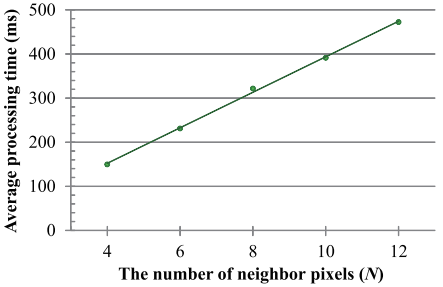


Fig. 7 Computational cost in relation to the number of neighboring pixels N.

3.3.1 Analysis of the Number of Neighboring pixels

N controls the amount of information maintained by each pixel, and it is considered to affect the accuracy of our proposed method. Therefore, we investigated the relationship between the accuracy of foreground detection and N, for $N = 4, 6, 8, 10, 12$. The results for the five outdoor scenes (in Fig. 3) are shown in **Fig. 6**. Then, we also investigated appropriate th (the detection threshold in Eq. (7)) for each N. Each scene indicates a similar tendency, and therefore we show the result for scene1 in **Fig. 8**. Figure 8 shows the highest accuracy at around $th = 0.8$ without reference to N, therefore we adopt $th = 0.8$ as the

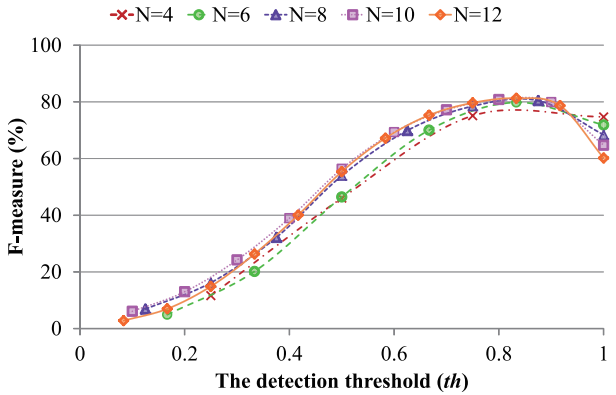


Fig. 8 Analysis of appropriate th (the detection threshold) for the number of neighboring pixels N using PETS2001 dataset (Scene1 in Fig. 3).

appropriate threshold in Eq. (7). Figure 6 indicates that the accuracy is not sensitive to N when $N \geq 6$, then we judged the appropriate N from the relationship between computational time[⋆1] and N as shown in **Fig. 7**. Figure 7 shows that the computational cost increases proportional to N. Therefore, we selected $N = 6$ and $th = 0.8$ as the optimal parameters in terms of the balancing accuracy and computational cost.

3.3.2 Effect of Radial Distance

The r controls the local feature and the localized regions focused by each pixel, and they are considered to affect the accuracy of our proposed method. Therefore, to investigate the effect of r on the detection accuracy, we selected typical objects of various sizes from the datasets illustrated in Fig. 3. Several examples which illustrate the sensitivity of r relative to object size, in which the object region is enlarged, are shown in **Fig. 9**. In Fig. 9, semitransparent blue regions and pixels represent the regions of interest and the pixels detected as the background respectively, and several good results are bounded by the red rectangles.

The red rectangles in Fig. 9 show that the suitable size of r becomes larger with

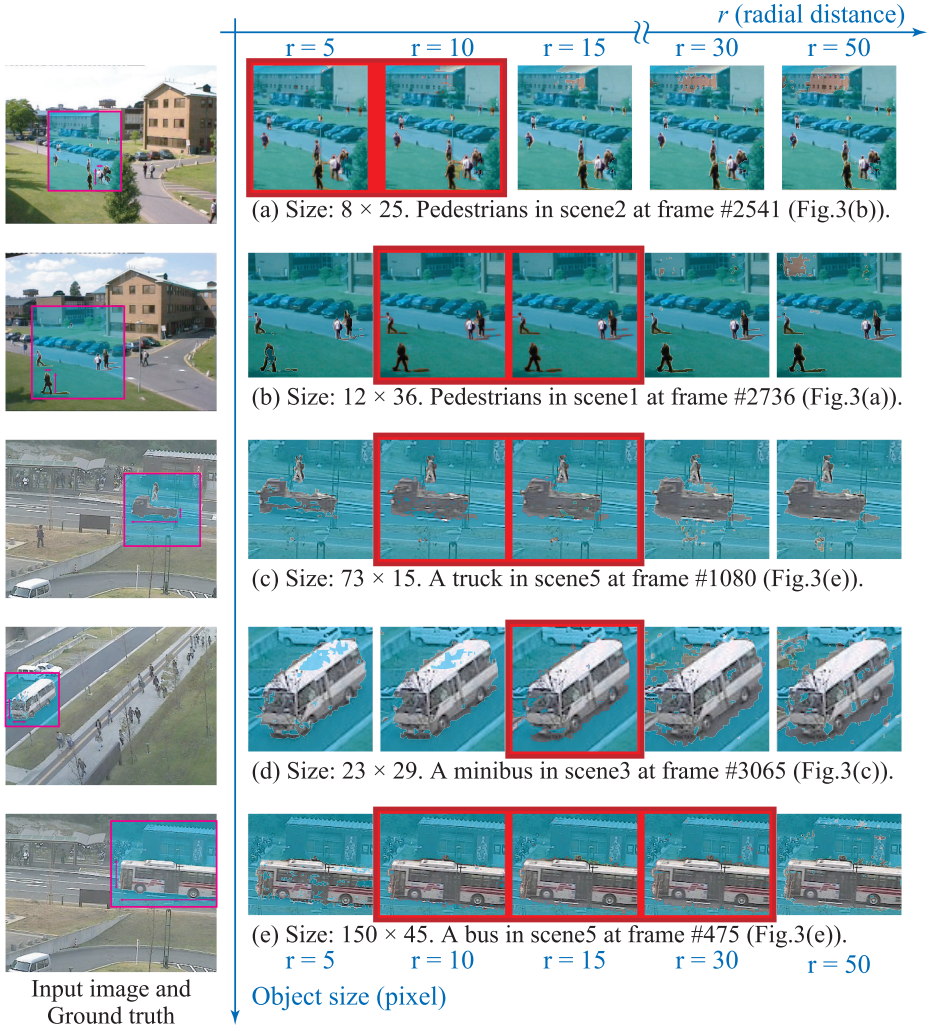


Fig. 9 Examples of variation in foreground detection results for different radial distances r. The semitransparent blue regions and pixels represent the regions of interest and the pixels detected as the background respectively. Several good detection results are bounded by the red rectangles.

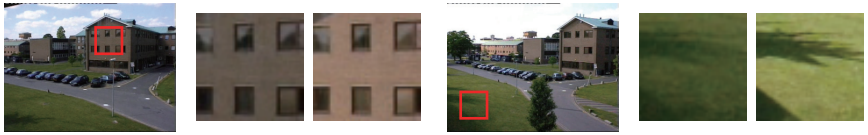
⋆1 We used a PC which has Core 2 Duo 2.8 GHz CPU and 4 GB memory, and the image size was 320 × 240 (pixel).

the increasing size of the object. Then, in Fig. 9 (b)–(d), there are many false-negative pixels when the value of r is smaller than half the short side of the object. The objects in these scenes have uniform textures and their background regions are also uniform. In such cases, there is little change in the SLDP before and after the object appears, since the SLDP does not model background color but rather the difference between a target pixel and its neighboring pixels. This is why a part of the object region was mistakenly regarded as background. Figure 9 also shows that the number of false-positive pixels increases as the size of r becomes larger, although large r can detect large objects. This is because r controls the localized regions focused by each pixel, and then the SLDP can adapt to the illumination changes which occur in the region. Therefore, there is a trade-off between the detection performance of r and the adaptivity to illumination changes. Because of this trade-off, the range of suitable size of r is limited, and its upper bound depends on the scene. However, in most cases, its lower bound depends mainly on the object size. In most applications (e.g., surveillance, security, etc.), we can predict the size of the objects, since the camera is stationary and observes similar objects in these applications. Then, we can easily estimate the lower bound of suitable size of r, and it is reasonable to choose r that is close to the lower bound. Hence, it does not lose a generality or effectiveness of the proposed method.

On the other hand, in certain cases where the object (such as Fig. 9 (e)) or the background has a relatively complicated texture, we can choose a smaller size of r than the size mentioned above, as we can see from Fig. 9 (e). Therefore, a task for future research is to automatically select the optimal size of r by using the background texture information, and it will lead to eliminating the trade-off of r.

3.4 Evaluation of the Adaptivity to Background Changes

In this section, we evaluate the adaptivity of our proposed model to illumination changes and motion changes in the background. We compared the performance of our proposed method with three different approaches. The GMM method [8] and LM method were used for evaluating the effectiveness of the statistical framework and the local feature-based framework as in Section 3.1. We also used the Hybrid method [13] to indicate that our new framework is better than hybrid methods which used the ad hoc solutions by logical combination. The evaluation frames



(a) Change observed on the wall (b) Change observed on the ground

Fig. 10 Typical region in which illumination change occurs.

used here involve the background region only. That is, all of the pixels should be regarded as background. Therefore, we introduced a new criterion, True-Negative-Rate, calculated by:

$$\text{True-Negative-Rate}\,(\%) = \frac{\text{True Negative}}{\text{True Negative} + \text{False Positive}} \times 100, \qquad (12)$$

where **True Negative** represents the number of pixels which are correctly detected as background.

3.4.1 Adaptivity to Illumination Change

We used outdoor scenes (92 frame images from scene1 and 132 frame images from scene2) in which the illumination conditions changed rapidly, and manually cropped two typical regions showing these rapid illumination changes for evaluation. **Figure 10** shows two sample frame images and the cropped regions whose size was 50×50 pixels. Examples of the results of foreground detection are shown in **Fig. 11**, while **Table 3** shows True-Negative-Rate for the illumination changes.

Figure 11 shows that the GMM method [8] detects a number of false-positive pixels, and Table 3 shows that the True-Negative-Rate of the GMM method is low. These results are typical evidence of the weakness of a statistical approach regarding illumination changes.

Meanwhile, we see that the methods which use a local feature-based framework (LM, Hybrid [13] and our proposed method) are robust against illumination changes, which is demonstrated by the few false-positive pixels in Fig. 11. Table 3 shows numerically that these three local feature-based methods can achieve a high True-Negative-Rate. This is because these methods can adapt to illumina-

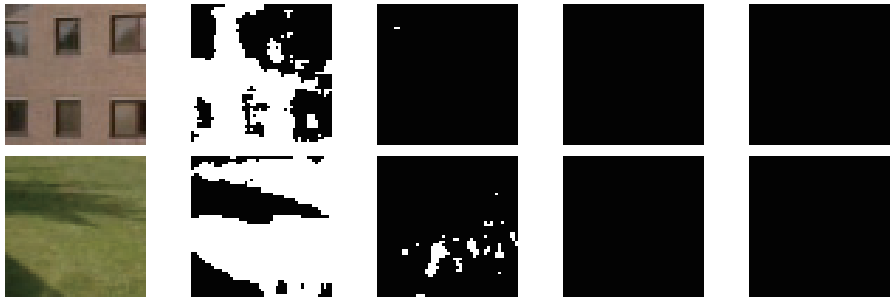


Fig. 11 Examples of the adaptivity to illumination change.

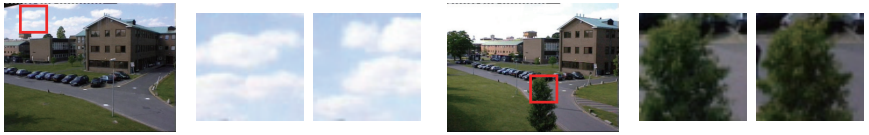
Table 3 True-Negative-Rate for illumination change.

Method	True-Negative-Rate (%)
GMM method [8] (proposed without local feature-based approach)	73.7
LM method (proposed without statistical approach)	97.6
Proposed method	99.8
Hybrid method [13]	99.9

tion changes by using the relations between the target pixel and its neighboring pixels. In the case of our proposed method, we assume that illumination changes affect localized regions proportionally, and our method can tolerate the effects of illumination changes which occur in the localized region. On the other hand, a limitation of our proposed method is that it cannot adapt to unusual illumination which causes peaked changes, such as specular reflection.

3.4.2 Adaptivity to Motion Change

The same two scenes were also used for the evaluation of motion changes in background regions. We manually cropped two typical regions displaying motion changes including cloud movement (see **Fig. 12** (a)) and trees swaying (see Fig. 12 (b)). The cropped image sequences consist of 3780 frame images in scene1 and 722 frame images in scene2. The size of regions was 50×50 pixels. Examples of the results of foreground detection are shown in **Fig. 13**, while **Table 4** shows



(a) Change caused by cloud movement (b) Change caused by movement of a tree

Fig. 12 Region in which motion change occurs and a example of changes.

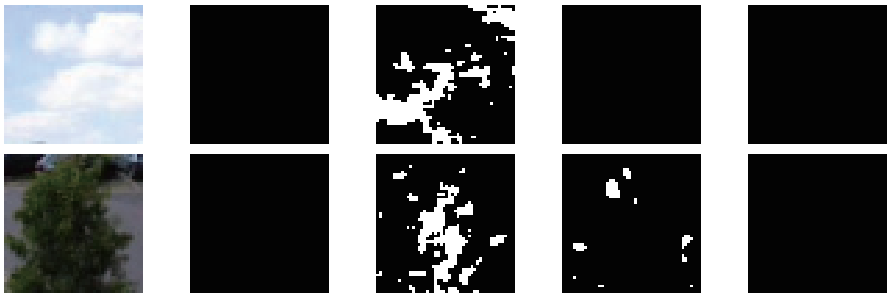


Fig. 13 Examples of the adaptivity to motion change.

Table 4 True-Negative-Rate for motion change.

Method	True-Negative-Rate (%)
GMM method [8] (proposed without local feature-based approach)	98.2
LM method (proposed without statistical approach)	80.3
Proposed method	99.3
Hybrid method [13]	98.3

the True-Negative-Rate.

We see that the LM method detects a number of false-positive pixels from Fig. 13, and its True-Negative-Rate is low from Table 4. These results are typical evidence of a weakness of the local feature-based approach when there are motion changes.

Meanwhile, as shown in Fig. 13, the methods using a statistical framework (GMM [8], Hybrid [13] and our proposed method) output a smaller number of false-

positive pixels. Table 4 shows numerically that these three statistical methods achieve a high True-Negative-Rate. This is because these methods can maintain multiple hypotheses of multi-modal backgrounds by their statistical models, however there is a limitation to the periodicity which can be coped with by them. For these methods to adapt to motion changes, the changes need to be modeled by the statistical models which forget the observation of the past over time. In the case of our method, we use GMM as the statistical model, and the periodicity is controlled by varying GMM parameters: α, W and T (the learning rate, the initial weight and the threshold of choosing the background model, respectively). For details of GMM, we refer the reader to the literature [8]. Although we can make minor adjustments to the periodicity, our method cannot cope with the case where the motion changes are periodically observed after a long interval.

4. Conclusion

In this paper, we have proposed a new background model based on the *Statistical Local Difference Pattern* (SLDP). Our main contribution is proposing a method that combines the concepts of a local feature-based approach and a statistical approach into a single framework. The result is that our proposed method adapts to both illumination changes and motion changes in the background. This is because the SLDP uses illumination-invariant local features which have the ability to tolerate the effects of illumination changes, and describes their distribution by GMMs which can learn the variety of motion changes. As the experimental results, we have confirmed that our proposed method can detect the foreground objects robustly against illumination changes and motion changes, especially in long shot scenes.

However, our proposed method also has two main constraints:

Computational time

Our proposed method does not work very fast, since each pixel has N GMMs in SLDP and needs to update them, where N is the number of the neighboring pixels. In the case where the image size was 320×240 pixels and $N = 6$, the computational time was about $230\,ms$ using a PC running a Core 2 Duo 2.8 GHz CPU with 4 GB memory. We think that this problem is not so critical for our method, since the problem will be able to be solved by the development of the computer. However, because of the demand for fast processing in an application such as security, reduction of the computational time is one of our future researches.

Dependence of radial distance r on the object size

Our proposed method also has a problem associated with the object size as discussed in Section 3.3.2. In particular, when the object is too big or the radial distance r is smaller than half the short side of the object, our method does not work well and many false-negative pixels are observed in the object region. On the other hand, when the background or the object has a complicated texture, we confirmed that r does not depend strongly on the size of the object. Therefore, future research will aim to eliminate the dependence of r on the object size by using the background texture information.

References

1) Toyama, K., Krumm, J., Brumitt, B. and Meyers, B.: Wallflower: Principle and Practice of Background Maintenance, *International Conference on Computer Vision*, pp.255–261 (1999).
2) Jabri, S., Duric, Z. and Wechsler, H.: Detection and location of people in video images using adaptive fusion of color and edge information, *15th International Conference on Pattern Recognition*, Vol.4, pp.627–630 (2000).
3) Mason, M. and Duric, Z.: Using histograms to detect and track objects in color video, *30th Applied Imagery Pattern Recognition Workshop*, pp.154–159 (2001).
4) Heikkila, M., Pietikainen, M. and Heikkila, J.: A texture-based method for detecting moving objects, *British Machine Vision Conference* (2004).
5) Heikkila M. and Pietikainen M.: A Texture-Based Method for Modeling the Background and Detecting Moving Objects, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.28, No.4, pp.657–662 (2006).
6) Satoh, Y., Kaneko, S., Niwa, Y. and Yamamoto, K.: Robust object detection using a Radial Reach Filter (RRF), *Systems and Computers in Japan*, Vol.35, No.10, pp.63–73 (2004).
7) Stauffer, C. and Grimson, W.E.L.: Adaptive background mixture models for real-time tracking, *IEEE International Conference on Computer Vision and Pattern Recognition* (*CVPR*), Vol.2, pp.246–252 (1999).
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).
9) Elgammal, A., Duraiswami, R., Harwood, D. and Davis, L.: Background and Foreground Modeling using Non-parametric Kernel Density Estimation for Visual

Surveillance, *Proc. IEEE*, Vol.90, pp.1151–1163 (2002).
10) Monari, E. and Pasqual, C.: Fusion of Background Estimation Approaches for Motion Detection in Non-static Backgrounds, *CD-ROM Proc. IEEE International Conference on Advanced Video and Signal Based Surveillance* (2007).
11) Tanaka, T., Shimada, A., Arita, D. and Taniguchi, R.: A Fast Algorithm for Adaptive Background Model Construction Using Parzen Density Estimation, *CD-ROM Proc. IEEE International Conference on Advanced Video and Signal Based Surveillance* (2007).
12) Yoshimura, H., Iwai, Y., and Yachida, M.: Object Detection with Adaptive Background Model and Margined Sign Cross Correlation, *International Conference on Pattern Recognition* (*ICPR 2006*) (2006).
13) Tanaka, T., Shimada, A., Taniguchi, R., Yamashita, T. and Arita, D.: Towards robust object detection: Integrated background modeling based on spatio-temporal features, *9th Asian Conference on Computer Vision* (Sep. 2009).

(Received November 10, 2010)
(Accepted October 17, 2011)
(Released December 28, 2011)

(Communicated by *Jun Takamatsu*)

Satoshi Yoshinaga received his B.E. and M.E. degrees from Kyushu University in 2009 and 2011. He is a Ph.D. student at Kyushu University. He is also a Research Fellow of the Japan Society for the Promotion of Science. He has been engaged in image processing.

Atsushi Shimada received his M.E. and D.E. degrees from Kyushu University in 2004 and 2007. Since 2007, he has been an assistant professor in Graduate School of Information Science and Electrical Engineering at Kyushu University. He has been engaged in image processing, pattern recognition and neural networks.

Hajime Nagahara received his B.E. and M.E. degrees in electrical and electronic engineering from Yamaguchi University in 1996 and 1998, respectively. He received his Ph.D. in system engineering from Osaka University in 2001. He was a Research Associate of Japan Society for the Promotion of Science in 2001–2003. He was a Research Associate of the Graduate School of Engineering Science, Osaka University, in 2003–2006. He was a Visiting Associate Professor at CREA University of Picardie Jules Verns, France in 2005. He was an Assistant Professor of Graduate School of Engineering Science in 2007–2010. He was a Visiting researcher at Columbia University, USA in 2007–2008. He has been an Associate Professor of Faculty of Information Science and Electrical Engineering, Kyushu University, since 2010. Computational photography, Image processing, Computer vision and Virtual reality are his research subjects. He achieved an ACM VRST2003 Honorable Mention Award in 2003.

Rin-ichiro Taniguchi received his B.E., M.E., and D. degrees from Kyushu University in 1978, 1980, and 1986. Since 1996, he has been a professor in the Graduate School of Information Science and Electrical Engineering at Kyushu University, where he directs several projects including multiview image analysis and software architecture for cooperative distributed vision systems. His current research interests include computer vision, image processing, and parallel and distributed computation of vision-related applications.