Research Paper

Interest Point Detection Based on Stochastically Derived Stability

Ukrit Watchareeruetai,[†1,†2] Akisato Kimura,[†1]
Robert Cheng Bao,[†1,†3] Takahito Kawanishi[†1]
and Kunio Kashino[†1]

We propose a novel framework called *StochasticSIFT* for detecting interest points (IPs) in video sequences. The proposed framework incorporates a stochastic model considering the temporal dynamics of videos into the SIFT detector to improve robustness against fluctuations inherent to video signals. Instead of detecting IPs and then removing unstable or inconsistent IP candidates, we introduce *IP stability* derived from a stochastic model of inherent fluctuations to detect more stable IPs. The experimental results show that the proposed IP detector outperforms the SIFT detector in terms of repeatability and matching rates.

1. Introduction

Recently, local features based on interest point (IP) detection have been successfully used in many computer vision tasks, such as image indexing [1),2)], stereo matching [3)] and object recognition [4),5)]. The advantages of using IPs are that they are robust against partial occlusions and do not require a segmentation process. Many researchers have proposed methods for extracting IPs from a still image, e.g., Lindeberg [6)], Mikolajczyk and Schmid [7)] and Lowe [8)]. Among current IP detectors, the Scale-invariant Feature Transform (SIFT) detector [8)] is the most appealing. The SIFT detector was designed to be robust against changes in scale and rotation, and partially tolerant to changes in illumination conditions and viewpoints.

We are interested in recognizing objects in video sequences, namely detecting an object appearing in a video sequence when given a query image of that object. The SIFT detector is effective in detecting IPs in many cases; however, it was designed for still images, not video sequences. That means the SIFT detector localizes IPs frame by frame and does not use any motion information or temporal smoothing, which might be useful for IP detection.

Various methods for detecting IPs in a video sequence have been proposed [9),10)]. These methods explore singular points (or discriminative points) in the spatio-temporal domain; consequently, they would be useful for certain tasks such as event detection or behavior/gait recognition. In contrast, this article focuses on detecting IPs that are discriminative in the spatial domain, but smooth in the temporal domain, which would be useful for object recognition and video retrieval tasks. For this purpose, various post-processing techniques have been proposed. The most widely used approach employs geometrical consistency among detected IPs in the spatial domain based on RANdom SAmpling Consensus (RANSAC) [11)] or Least Median of Squares (LMeds) [12)]. Tracking-based approaches are also widely employed [13),14)] for IP detection in a video sequence. However, the use of geometrical consistency alone may be insufficient to appropriately extract IPs from a video sequence, and tracking-based methods are not robust against occlusions, sudden illumination changes and noise.

To this end, we propose a new framework called *StochasticSIFT* for detecting IPs from video signals. The proposed framework incorporates a stochastic model that simulates temporal dynamics inherently included in video signals into the SIFT detector. This approach enables us to simultaneously deal with both inherent fluctuations and the temporal smoothness of video signals in natural ways. Instead of independently detecting IPs from each frame followed by a process for removing unstable or inconsistent IPs, we perform IP detection after evaluating the *stability* of IPs derived from the stochastic model, resulting in IP detection that would be robust against noise or illumination changes. In contrast with approaches that localize IPs in the spatio-temporal domain as described in[9),10)], our approach detects points that are discriminative in the spatial domain but smooth in the temporal domain. Therefore, the proposed method is more suitable for object recognition tasks.

Another feature of the proposed framework is that it enables us to incorporate

[†1] NTT Communication Science Laboratories, NTT Corporation
[†2] International College, King Mongkut's Institute of Technology Ladkrabang
[†3] University of British Columbia

some constraints *before* performing IP detection. Most previous IP detectors for video signal applications firstly detect IP candidates, and then remove unstable candidates by using point tracking and/or geometrical constraints. However, this kind of approach sometimes fails to track or select appropriate IPs because temporal and/or geometrical constraints can be applied *only* to the detected IPs. On the other hand, our approach can naturally introduce such constraints *before* detecting IP candidates, resulting in point detection that would be robust against certain kinds of noise or distortion.

The rest of this article is organized as follows. Section 2 briefly describes the SIFT method. Section 3 explains the proposed method, i.e., StochasticSIFT. Section 4 presents our datasets, experimental setup and results. Section 5 provides our conclusion.

2. Scale-invariant Feature Transform

SIFT[8] can be divided into two main steps, i.e., (1) IP localization and (2) descriptor construction. In the first step, IPs are localized in both scale and spatial domains. In the second step, rotation-invariant descriptors are constructed and added to the detected IPs.

2.1 Interest Point Localization

The IP localization process in SIFT is illustrated in **Fig. 1**. This process can be briefly explained as follows:

(1) *Creating pyramids of difference of Gaussian (DoG) images:* Firstly, an input image is repeatedly convolved by a Gaussian filter to create a set of scale-space images, called the first *octave*. Each filtered image is characterized by the *scale*, i.e., the number of Gaussian filtering. Then, two filtered images with adjacent scales are subtracted to obtain a DoG image. The next octave is then created by down-sampling a filtered image (the third image in the figure) of the current octave, and repeating the same process several times.

(2) *Finding local extrema in both spatial and scale domains:* Every pixel value in all the DoG images is compared with eight neighbors of the same scale, nine neighbors of the upper scale, and nine neighbors of the lower scale, making a total of 26 neighbors. The position and scale of the local extrema are reserved as an IP candidate.

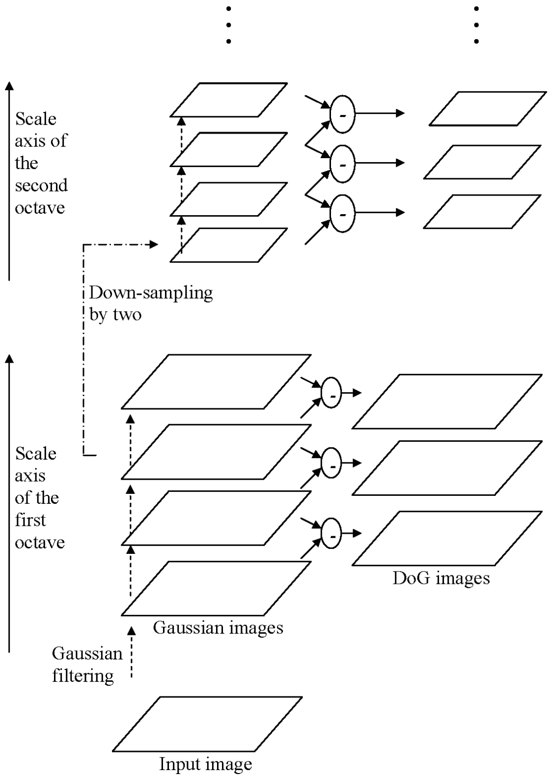

Fig. 1 Creating difference of Gaussian (DoG) pyramid.

(3) *Performing post-processing to remove inaccurate IPs:* In this step, the contrast of the IP candidates is calculated. If the contrast of a candidate is less than a threshold value $Th_{contrast}$, it will be removed from the candidate set. In this work, the threshold value $Th_{contrast}$ is set at 0.04. Then, the remaining IPs are checked to determine whether they are edge-like IPs, i.e., IPs located on an edge. Edge-like IPs will be removed, and the IPs remaining in the candidate set are considered detected IPs. Post-processing is described in detail in Ref. 8).

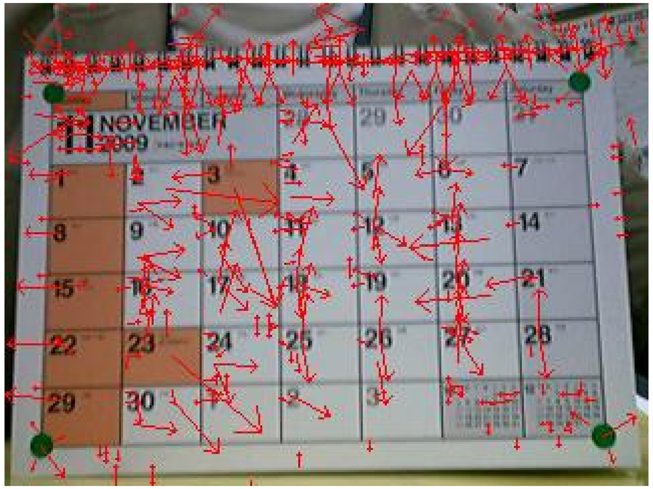

Fig. 2 Example of interest points, scales and their principal directions detected by SIFT.

2.2 Descriptor Construction

(1) *Assigning orientation:* The magnitude and orientation of gradients at each IP and its neighbors are calculated from a Gaussian filtered image with the same scale as that IP. Then, an orientation histogram (36 bins covering 360 degrees) is constructed. Here a Gaussian-weighted circular window is used to emphasize the features near the IP. The highest peak of the histogram is assigned as the principal orientation of the IP (**Fig. 2**).

(2) *Creating descriptors:* An IP descriptor is created from the orientation histogram. The descriptor coordinate is rotated relative to the principal IP orientation to make it invariant to rotation changes. Neighboring pixels are divided into 4 × 4 subregions to create a local histogram of eight orientation bins. These are used to construct a feature vector with a size of $4 \times 4 \times 8 = 128$. Then the feature vector is normalized to unit-length to make it partially tolerant to affine and illumination changes.

3. Proposed Method: StochasticSIFT

3.1 Overview

As described in Section 1, the SIFT detector is not very robust against sudden illumination changes and pulse noise, even if appropriate post-processing is executed. Considering mechanisms in which such kinds of clutter arise, they can be modeled as stochastic events. The proposed method, *StochasticSIFT*, introduces a stochastic framework including those mechanisms.

Figure 3 shows the framework of StochasticSIFT, which consists of five layers connected hierarchically. We briefly describe each layer in the following paragraphs.

$\bar{\boldsymbol{S}}(t) = \{\bar{\boldsymbol{s}}(\omega, t)\}_{\omega \in \Omega}$ is a set of DoG images calculated from the t-th video frame $\boldsymbol{i}(t)$, where ω represents a scale, Ω is a set of scales, and $\bar{\boldsymbol{s}}(\omega, t)$ is the DoG image at scale ω and time t. Details of how to extract DoG images are provided in Section 2.

$\boldsymbol{S}(t) = \{\boldsymbol{s}(\omega, t)\}_{\omega \in \Omega}$ is a stochastic representation of DoG images at the t-th frame, each of which $\boldsymbol{s}(\omega, t)$ is called a *stochastic DoG image* with scale ω and time t. We would like to note that the introduction of a stochastic representation of DoG images is the first contribution of our proposed method. Stochastic DoG images $\boldsymbol{S}(t)$ are calculated from the current DoG images $\bar{\boldsymbol{S}}(t)$ and previous stochastic DoG images $\boldsymbol{S}(t-1)$, where temporal dynamics induced by noise or illumination changes are introduced. More details will be described in Section 3.2.

$\boldsymbol{F}(t) = \{\boldsymbol{f}(\omega, t)\}_{\omega \in \Omega}$ is a set of *discriminant maps* at time t. Each pixel value $f(\omega, t, \boldsymbol{x})$ of a discriminant map $\boldsymbol{f}(\omega, t)$ represents the discriminative degree of this pixel $\boldsymbol{x}$ in the stochastic DoG image $\boldsymbol{s}(\omega, t)$. This value can be viewed as *IP stability*, which determines whether or not the pixel should be regarded as an IP. An evaluation of the IP stability based on the stochastic representation of DoG images is the second contribution of StochasticSIFT. A set of IPs, denoted by $\boldsymbol{X}(t)$, is extracted from the discriminant maps. We discuss how to select IPs in Section 3.3.

3.2 Stochastic DoG Images

The presence of noise, illumination changes and viewpoint changes means that the observed DoG image might be corrupted, and different from its true state,

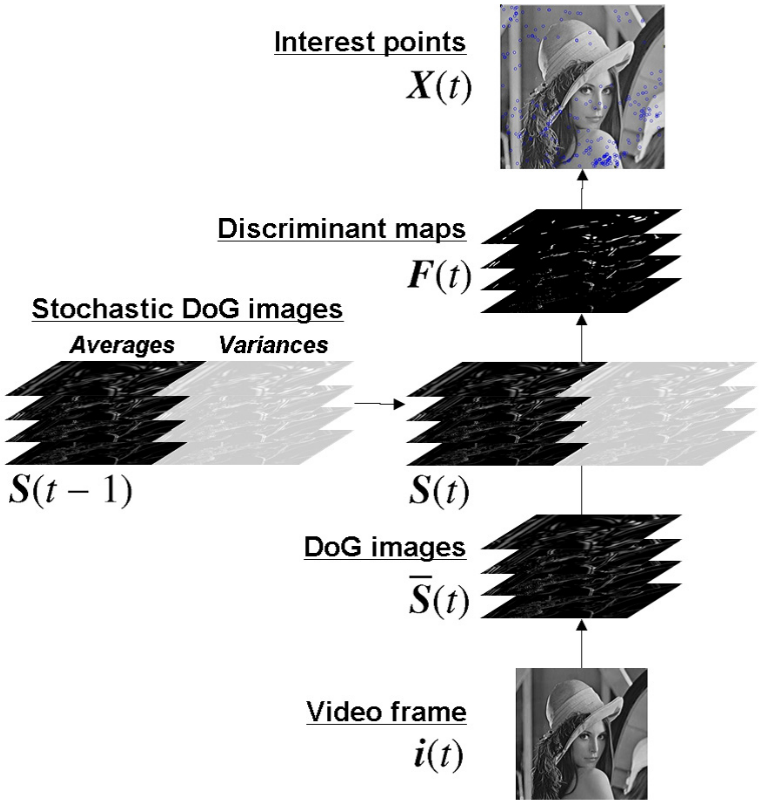

Fig. 3 Framework of StochasticSIFT.

resulting in missed or spurious IPs. To this end, we introduce a stochastic representation of DoG images, which we call stochastic DoG images, to enable us to handle such fluctuations.

A stochastic DoG image $\boldsymbol{s}(\omega, t)$ is defined as a set of Gaussian random variables, each of which corresponds to a pixel $\boldsymbol{x} = (x, y)$. We denote a random variable at pixel $\boldsymbol{x}$, time t and scale ω as $s(\boldsymbol{x}, \omega, t)$, assuming the following state space model[15]:

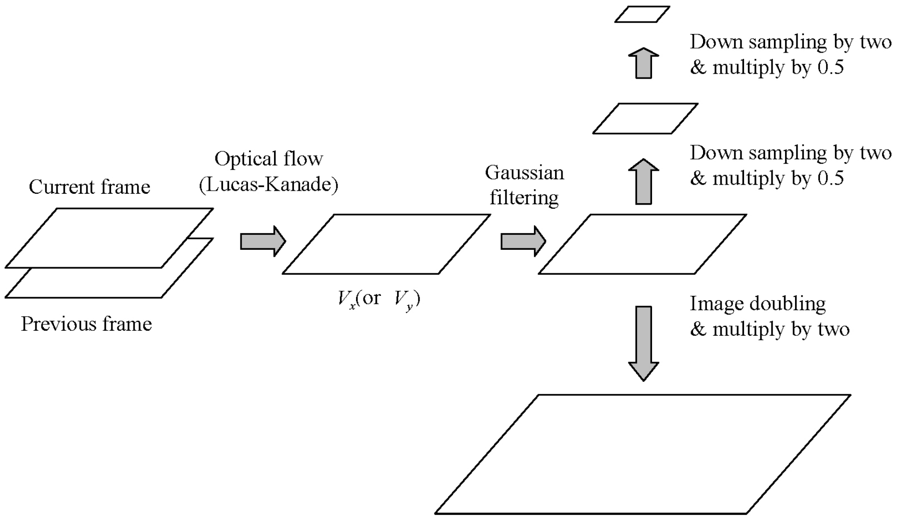

Fig. 4 Calculation of optical flow pyramid.

$$s(\boldsymbol{x}, \omega, t) = s(\widetilde{\boldsymbol{x}}, \omega, t-1) + \epsilon_1, \tag{1}$$

$$\overline{s}(\boldsymbol{x}, \omega, t) = s(\boldsymbol{x}, \omega, t) + \epsilon_2, \tag{2}$$

$$\widetilde{\boldsymbol{x}} = (\widetilde{x}, \widetilde{y}) = (x - \Delta x(\boldsymbol{x}), y - \Delta y(\boldsymbol{x})), \tag{3}$$

where ϵ_i $(i = 1, 2)$ is a Gaussian random variable with zero mean and variance σ_i^2, $\overline{s}(\boldsymbol{x}, \omega, t)$ is the pixel value of the DoG image $\overline{\boldsymbol{s}}(\omega, t)$ at pixel $\boldsymbol{x}$, and $\widetilde{\boldsymbol{x}}$ is the position after compensating for the optical flow $(\Delta x(\boldsymbol{x}), \Delta y(\boldsymbol{y}))$ (the movement during time $t-1$ and t) at pixel $\boldsymbol{x}$. Eq. (1) simulates the temporal dynamics of the DoG images, whereas Eq. (2) models observation noise.

We must compensate for the optical flow because objects in the video sequence do not necessarily remain in the same position. If the estimation model does not include any compensation for motion information, it will estimate the value of the stochastic feature from different locations, resulting in poor performance. Here we calculate optical flow based on the method proposed by Lucas and Kanade[13], and create a pyramid of optical flow images as shown in **Fig. 4**. In particular, optical flow components $\Delta \boldsymbol{X}(t)$ and $\Delta \boldsymbol{Y}(t)$ are calculated from the current and previous video frames ($\boldsymbol{i}(t)$ and $\boldsymbol{i}(t-1)$, respectively). Gaussian filtering is used to smooth the optical flow. Then, each component is down-sampled by two and each pixel value multiplied by 0.5 to obtain the next higher level flow component (for the lower level, if any, we perform image doubling and multiply each pixel value by two instead). Note that the number of levels in the pyramid equals the number of octaves of the pyramid of DoG images, and the optical flow of each level is used for all scales in the corresponding octave.

We can recursively estimate the mean $\widehat{s}(\boldsymbol{x},\omega,t)$ and variance $\sigma_s^2(\boldsymbol{x},\omega,t)$ of a random variable $s(\boldsymbol{x},\omega,t)$ as follows:

$$\begin{aligned}\hat{s}(\boldsymbol{x},\omega,t) =& \frac{\sigma_2^2}{\sigma_1^2+\sigma_2^2+\sigma_s^2(\widetilde{\boldsymbol{x}},\omega,t-1)}\hat{s}(\widetilde{\boldsymbol{x}},\omega,t-1)\\ &+ \frac{\sigma_1^2+\sigma_s^2(\widetilde{\boldsymbol{x}},\omega,t-1)}{\sigma_1^2+\sigma_2^2+\sigma_s^2(\widetilde{\boldsymbol{x}},\omega,t-1)}\bar{s}(\boldsymbol{x},\omega,t), \end{aligned} \tag{4}$$

$$\sigma_s^2(\boldsymbol{x},\omega,t) = \frac{\sigma_2^2\cdot(\sigma_1^2+\sigma_s^2(\widetilde{\boldsymbol{x}},\omega,t-1))}{\sigma_1^2+\sigma_2^2+\sigma_s^2(\widetilde{\boldsymbol{x}},\omega,t-1)}, \tag{5}$$

We introduce an adaptation of the variance parameter σ_1^2 of the observed noise to estimate the nature of the input videos. σ_1^2 is updated based on the following equation.

$$\sigma_1^2(\boldsymbol{x},\omega,t) = \frac{(t-1)}{t}\sigma_1^2(\boldsymbol{x},\omega,t-1)+\frac{1}{t}(\bar{s}(\boldsymbol{x},\omega,t)-\hat{s}(\boldsymbol{x},\omega,t|t))^2 \tag{6}$$

Note that the variance parameter σ_2^2 is a constant. The initial values of σ_1^2 and σ_2^2 are $= 3.331\times10^{-2}$ and 3.825×10^{-2} [⋆1], respectively.

Sometimes, it is possible that we may obtain unreasonable optical flows, which would lead to an inaccurate estimation of stochastic DoG images. In such cases, the estimation step should be re-initialized to avoid any cumulative errors caused by such unreasonable optical flows. Here, we introduce the following two criteria to re-initialize the estimation when at least one criterion is satisfied:

(1) *Unreasonably large optical flow:* $|\Delta x|>\theta$ or $|\Delta y|>\theta$, where θ is a threshold value. Here we set the threshold $\theta = H/8$, where H is the image height. This means the threshold value is different in each octave.

(2) *Impossible optical flow:* $\widetilde{\boldsymbol{x}}$ is outside of image.

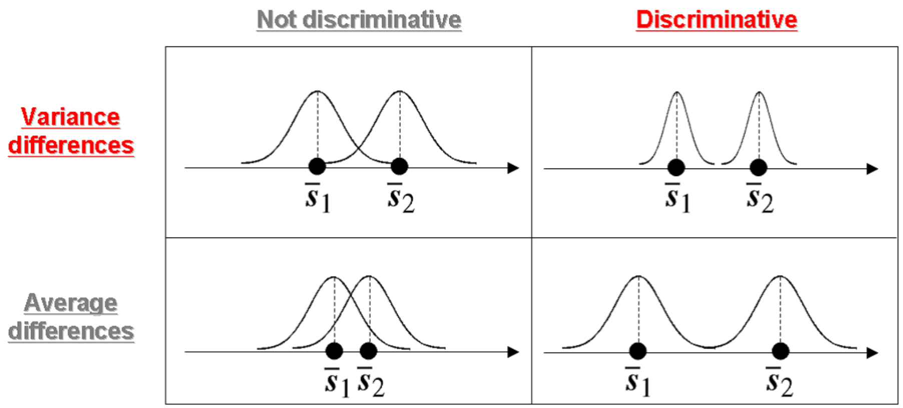

Fig. 5 Intuitive description of "discriminability".

3.3 Discriminant Maps

As described in Section 3.1, a discrimination map represents the discriminative degree at each pixel of stochastic DoG images against those at neighboring pixels.

Figure 5 intuitively depicts the importance of "discriminability" when detecting IPs. The traditional SIFT detector looks for local extrema in DoG images as IP candidates. However, when the pixel values of a local extremum and its neighbors are very close to each other (Fig. 5 bottom left), the extremum may disappear by just a little amount of fluctuations of pixel values. Also, the extremum may be lost when there is a large amount of fluctuation (Fig. 5 top left). The above discussion implies that the "discriminability" of local extrema in DoG images is directly related to the "stability" of IPs.

From this viewpoint, we newly introduce a measure that evaluates IP stability based on a stochastic representation of DoG images. If the density of a pixel is sufficiently discriminative and its average is a local extremum, we regard the pixel as a *stable* IP candidate. As a measure for evaluating stability, we utilize the sum of two-class Fisher linear discriminants [16]:

$$f(\boldsymbol{x},t,\omega) = \sum_{(\widetilde{\boldsymbol{x}},\widetilde{\omega})\in N(\boldsymbol{x},\omega)} \frac{|\widehat{s}(\boldsymbol{x},t,\omega)-\widehat{s}(\widetilde{\boldsymbol{x}},t,\widetilde{\omega})|}{\sqrt{\sigma_s^2(\boldsymbol{x},t,\omega)+\sigma_s^2(\widetilde{\boldsymbol{x}},t,\widetilde{\omega})}}, \tag{7}$$

where $N(\boldsymbol{x},\omega)$ stands for a set of neighboring pixels and scales of $(\boldsymbol{x},\omega)$. In

[⋆1] They are heuristically selected and the same values are used for all tests.

Eq. (7), we compute two quantities: 1) the absolute difference of the mean between an extremum and one of its neighbors, and 2) the square root of the sum of the variance of the extremum and that neighbor. We then calculate the ratio between these two quantities. This ratio would reflect the "discriminability" between the extremum and that neighbor (as illustrated in Fig. 5). Consequently, we use the sum of this ratio over all neighbors to determine the "discriminability" of the considered extremum.

3.4 IP Selection

Then, we determine whether a pixel should be regarded as an IP based on the following processes:

(1) Find the points that are local extrema in both the DoG images and the average of the stochastic DoG images, and include these points in the set of IP candidates.

(2) Remove the candidates with low stability scores calculated by Eq. (7).

(3) Remove the remaining candidates with low intensities in the average of the stochastic DoG images.

(4) Remove the remaining candidates that seem to be at the edges.

Note that the last two processes are similar to the post-processing techniques of the SIFT detector (see Section 2).

4. Evaluations

4.1 Datasets

To evaluate the performance of the proposed IP detector, we performed experiments with five datasets: Proximity card, Yurica, Calendar, Bus card, and Perfume box datasets. Each dataset consists of one query image and one video sequence, and the name of each dataset comes from the planar object presented in the dataset (see **Fig. 6**). In the evaluation, we localize IPs in the query image and in each frame of the video sequence, and then perform IP matching based on their descriptors. The size of the query images is 320 × 240 pixels. Each video sequence consists of 100 frames that are the same size as the query images. We used only planar objects because we needed a ground truth of point correspondence to perform numerical evaluations. Note that our approach can also function properly for non-planar objects.

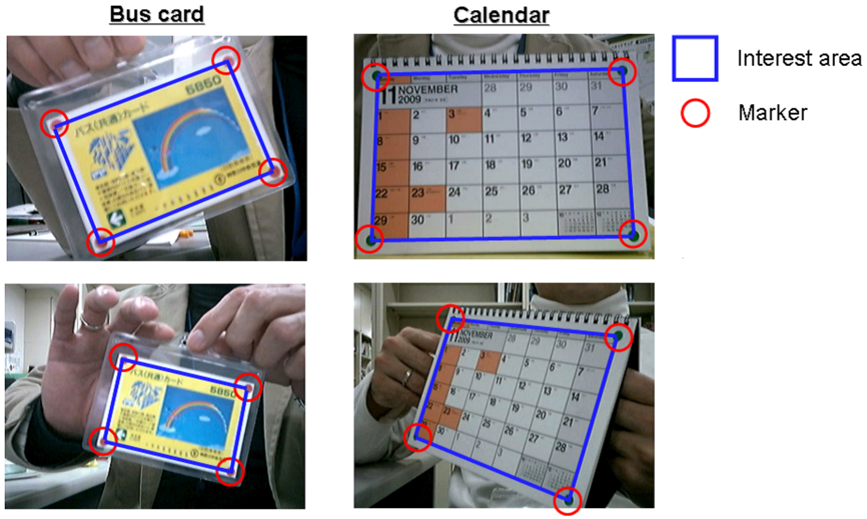

Fig. 6 Examples of datasets used in experiments.

As shown in Fig. 6, four markers were stuck each object to define an interest area (rectangular area with four markers as its corners) and extract a homography matrix of every pair of a query image and a video frame to obtain the ground truth. We considered only IPs located in the interest areas, and discarded the rest. Note that the complete interest area of objects can be seen in each frame without any occlusion; however, there are some changes in positions, rotations, scales, and/or viewpoints, as shown in Fig. 6.

We added artificial Gaussian noise with different variances (25, 50, 75, and 100[⋆1]) to every video sequence to observe the robustness of the proposed IP detectors against noise. Note that no extra noise is added to the query image.

4.2 Evaluation Measures

We adopt two measures for evaluating IP detectors. One is known as the *repeatability rate*[17], which evaluates only positional matches. In particular, once an IP of a query image is mapped into a video frame, based on the homography

[⋆1] The variance of the noise intensity added to each pixel.

matrix, we say that the mapped IP *positionally matches* an IP in the video frame if the distance[⋆1] between the mapped IP from a query and the actual IP of a video frame is less than the threshold $\epsilon = 1.5$. The repeatability rate is defined as

$$R = \frac{1}{T}\sum_{t=1}^{T} \frac{pm_t}{\min(n_0, n_t)}, \tag{8}$$

where pm_t is the number positional matches at the t-th frame, n_0 and n_t respectively are the number of IPs in the query image and the t-th frame, and T is the total number of frames (in the experiments, $T = 100$).

The other evaluation measure is called the *matching rate*, which considers both positional and descriptor matches. In particular, we first match the descriptor of an IP in the query image with IPs in the video frame, and find the best and second best matches. Then we use the following criterion to check whether we should accept this descriptor match.

$$d_0 < TH * d_1, \tag{9}$$

where d_0 and d_1 are the distances between the descriptors of the best and second best matches, respectively, and TH is a constant (here $TH = 0.49$). If the above criterion is satisfied, we say that their descriptors are matched. After that, we check their positions to determine whether they are positionally matched. If they are also positionally matched, we say that they are correctly matched; otherwise, they are falsely matched. The matching rate is defined as follows:

$$MR = \frac{\sum_{t=1}^{T} cm_t}{\sum_{t=1}^{T}(cm_t + fm_t)}, \tag{10}$$

where cm_t and fm_t respectively are the numbers of correct and false matches at the t-th frame.

4.3 Results

We compared our proposed IP detector, i.e., StochasticSIFT, with the SIFT detector implemented by Hess[⋆2]. The common parameters of SIFT and StochasticSIFT were set as follows: The number of octaves was six, the number of scales in each octave was five, the variance of Gaussian filter was 1.6.

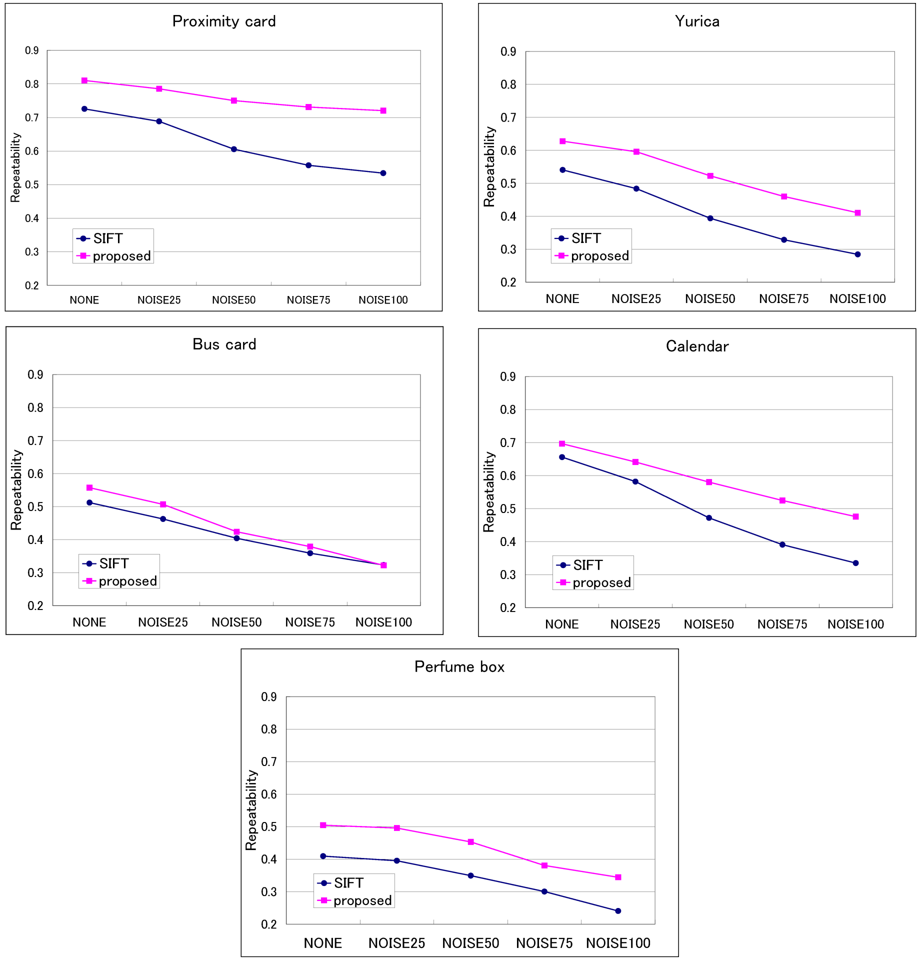

Fig. 7 Results by repeatability rates.

The evaluation results with the repeatability rate are shown in **Fig. 7**. The results indicate that StochasticSIFT provides better repeatability rates than SIFT for all the five datasets. Although the repeatability rates decrease according to increases in Gaussian noise, StochasticSIFT outperformed the original SIFT for

[⋆1] The distance is measured in the original scale image.
[⋆2] http://web.engr.oregonstate.edu/˜hess/

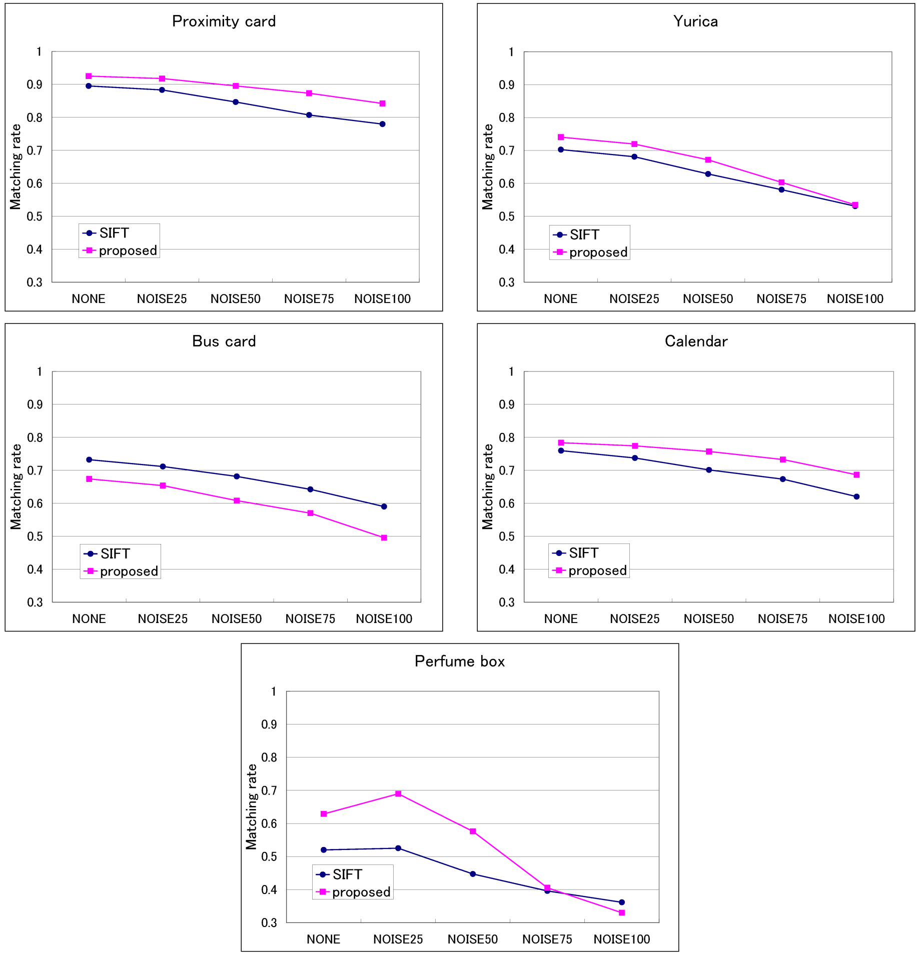

Fig. 8 Results by matching rates.

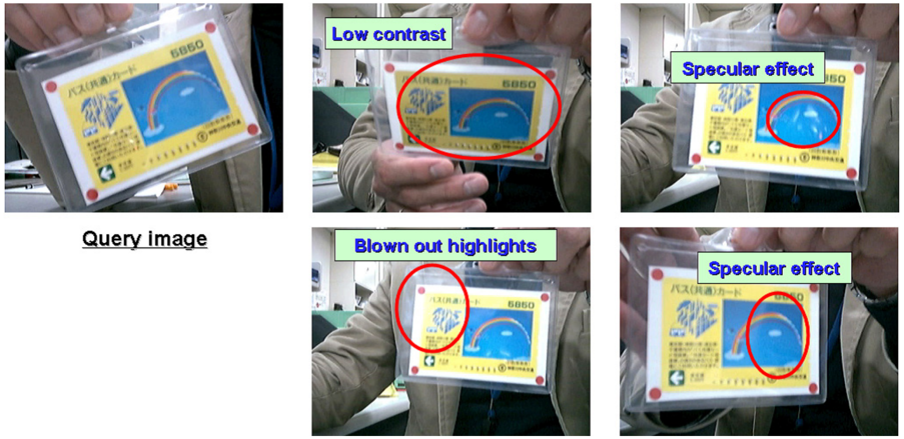

Fig. 9 Several fluctuations found in the "bus card" dataset that the proposed method could not handle.

Table 1 Comparison of the average numbers of detected IPs for some datasets.

	SIFT					
Noise	Proximity		Yurica		Bus	
level	query	video	query	video	query	video
NONE	103	111.2	199	150.3	276	200.2
25	103	113.6	199	148.4	276	188.9
50	103	108.6	199	137.9	276	162.9
75	103	94.1	199	123.4	276	131.8
100	103	74.4	199	106.4	276	109.0
	StochasticSIFT					
Noise	Proximity		Yurica		Bus	
level	query	video	query	video	query	video
NONE	67	62.2	104	78.3	128	89.7
25	67	60.8	104	75.9	128	84.3
50	67	57.9	104	68.5	128	73.2
75	67	50.6	104	60.3	128	58.6
100	67	40.1	104	49.3	128	44.3

any noise levels. And for some datasets, the differences in the repeatability rates became larger as the noise increased.

The evaluation results with the matching rate are shown in **Fig. 8**. The matching method we used was the same for both of the IP detectors, and is described in Ref. 18). The results indicate that StochasticSIFT provided better matching rates for all but one dataset (Bus card). **Figure 9** shows several fluctuations found in the "Bus card" dataset that the proposed method could not mitigate. The proposed method achieved worse matching rates than SIFT with this dataset

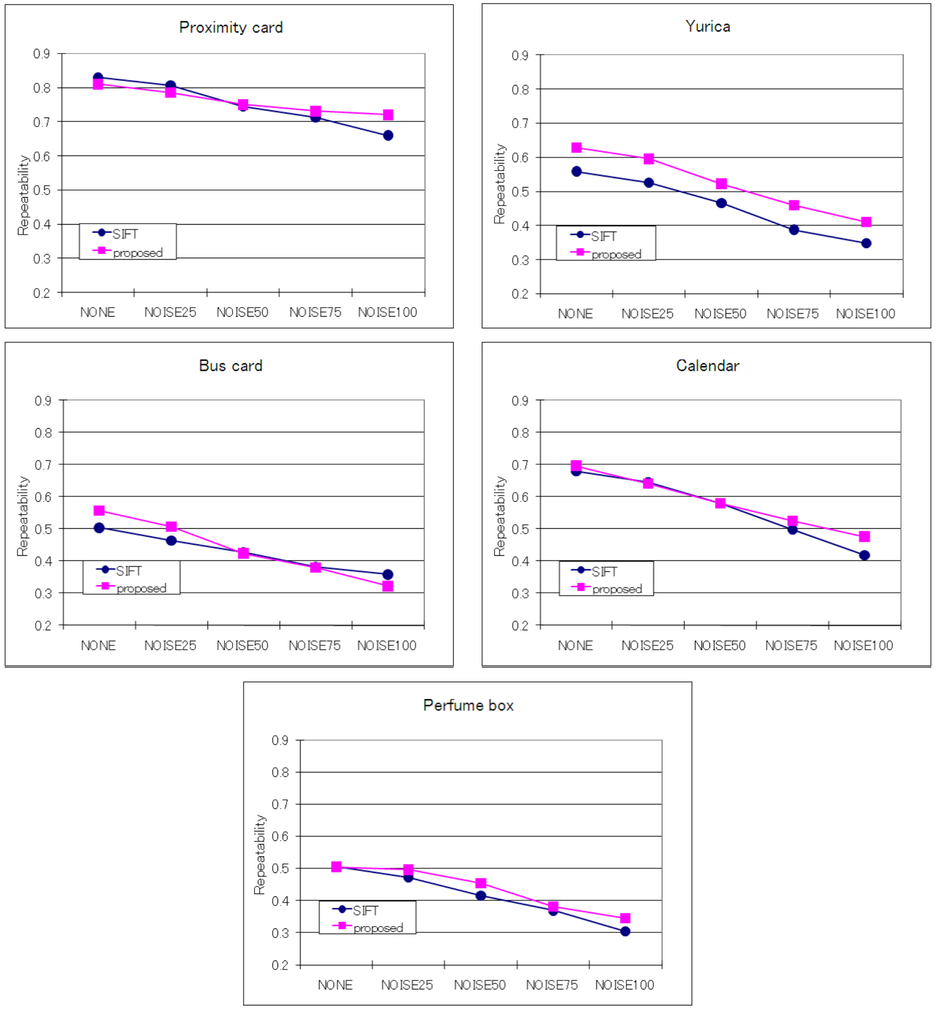

Fig. 10 Results by repeatability rates. The threshold used for removing low-contrast IPs with the SIFT detector was adjusted so that the number of detected IPs was almost the same as that with StochasticSIFT.

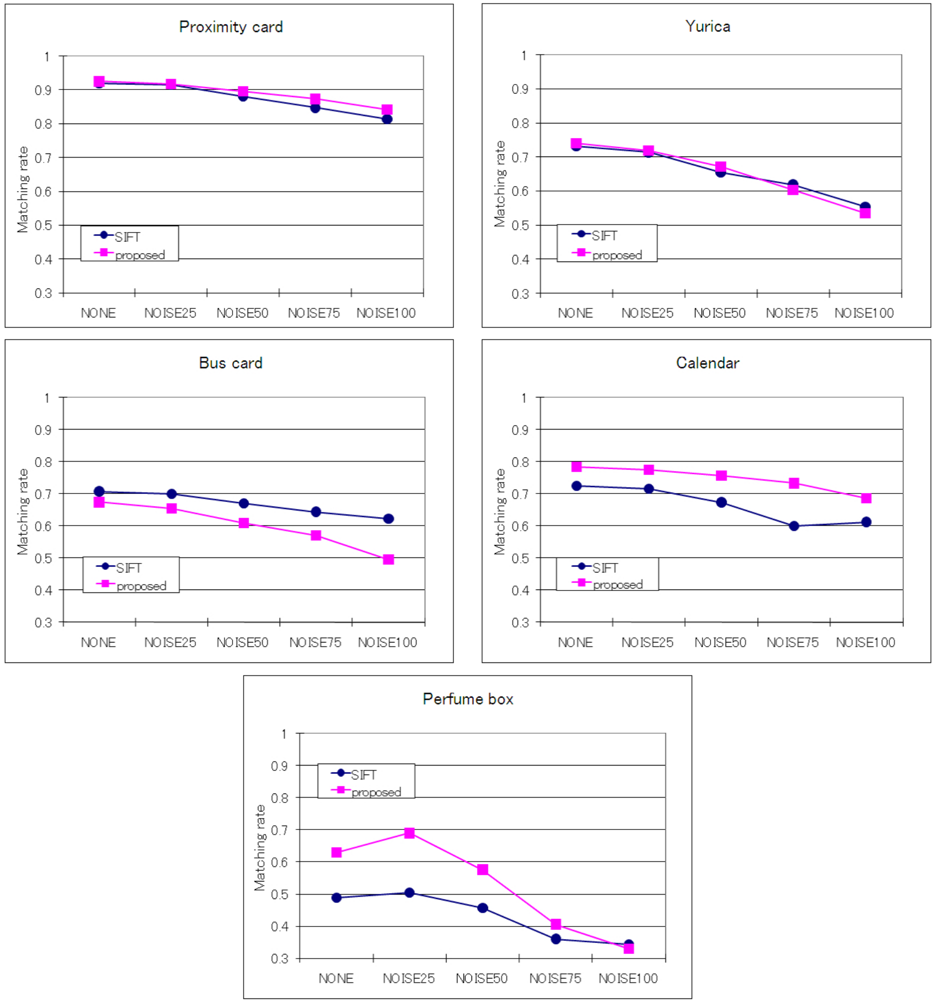

Fig. 11 Results by matching rates. The threshold used for removing low-contrast IPs with the SIFT detector was adjusted so that the number of detected IPs was almost the same as that with StochasticSIFT.

Table 2 Comparison of average computation cost (msec. per frame).

SIFT	StochasticSIFT	StochasticSIFT / SIFT
284.09	651.51	2.29

because it includes a lot of fluctuations throughout the frame, and therefore unreasonable optical flows often occurred.

Table 1 provides additional information about the number of detected IPs. Generally, StochasticSIFT detects fewer IPs than the original SIFT detector. This is because StochasticSIFT takes account of the temporal dynamics needed to remove spurious IPs that might be caused by noise.

This information might raise the following questions. Does the number of IPs affect the repeatability and matching rates? What happens if we compare the two methods when they give the same number of IPs? To answer these questions, we conducted another experiment by adjusting the threshold $Th_{contrast}$[⋆1] of the SIFT detector so that SIFT will give (almost) the same number of IPs (detected from the query image) as StochasticSIFT for each dataset. We then compare the repeatability and matching rates of the two methods. The results are shown in **Fig. 10** and **Fig. 11**. The results show that the repeatability rate of SIFT is greatly improved. In this case, the repeatability rates of the two methods are comparable, except as regards the Yurica dataset where StochasticSIFT is clearly better. However, the matching rate obtained with SIFT is not greatly improved and is even worse in some cases. The result indicates that IPs detected by StochasticSIFT are more stable than those detected with SIFT even if the two methods generate the same numbers of IPs.

We have measured the computation costs of the proposed method and the SIFT detector (not include the matching process). The costs were measured on a Core2Duo 2.2 GHz PC, and the average values are shown in **Table 2**. The computation time with the proposed method was around 2–3 times (average 2.29 times) of the SIFT detector, depending on the number of IPs in each dataset. One of the most intense processes is the calculation of optical flow. The results suggest that an efficient way to estimate optical flow is needed to improve the calculation cost of the proposed method.

5. Concluding Remarks

We proposed a new stochastic framework for interest point detection that we called StochasticSIFT, which extends the SIFT detector. It incorporates a stochastic representation of DoG images, which takes into account the temporal dynamics inherent to video signals into the SIFT detector. Experimental results suggest that the proposed method has certain advantages over the SIFT detector in terms of both repeatability and matching rate. We have not yet used any geometrical constraints to remove unreliable IPs, which might help improve the performance of StochasticSIFT.

Acknowledgments The first and third authors contributed to this work during their internship at NTT Communication Science Laboratories. The authors thank Dr. Yoshinobu Tonomura (currently Ryukoku Univ.), Dr. Naonori Ueda, Dr. Hiroshi Sawada, Dr. Kenji Nakazawa (currently NTT Advance Technologies Inc.) and Dr. Junji Yamato of NTT Communication Science Laboratories, and Prof. Noboru Ohnishi of Nagoya University for their help to the internship.

References

1) Kise, K., Noguchi, K. and Iwamura, M.: Memory Efficient Recognition of Specific Objects with Local Features, *Proc. International Conference on Pattern Recognition* (*ICPR*), pp.1–4 (2008).

2) Li, F.-F. and Perona, P.: A Bayesian Hierarchical Model for Learning Natural Scene Categories, *Proc. Conference Computer Vision and Pattern Recognition* (*CVPR*), Vol.2, pp.524–531, Washington, DC, USA, IEEE Computer Society (2005).

3) Sato, T. and Yokoya, N.: New multi-baseline stereo by counting interest points, *Proc. Canadian Conference on Computer and Robot Vision* (*CRV*), pp.96–103 (2005).

4) Sivic, J., Russell, B., Efros, A., Zisserman, A. and Freeman, W.: Discovering objects and their location in images, *Proc. International Conference on Computer Vision* (*ICCV*), Vol.1, pp.370–377 (2005).

5) Lazebnik, S., Schmid, C. and Ponce, J.: Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories, *Proc. Conference on Computer Vision and Pattern Recognition* (*CVPR*), Vol.2, pp.2169–2178 (2006).

[⋆1] This threshold is used for removing low contrast IPs. It is a major factor affecting the stability of IPs detected with the SIFT detector (see Section 2.1).

6) Lindenberg, T.: Feature detection with automatic scale selection, *International Journal of Computer Vision*, Vol.30, No.2, pp.79–116 (1998).

7) Mikolajczyk, K. and Schmid, C.: Scale and Affine invariant point detectors, *International Journal of Computer Vision*, Vol.60, No.1, pp.63–86 (2004).

8) Lowe, D.: Distinctive image features from scale-invariant keypoints, *International Journal of Computer Vision*, Vol.60, No.2, pp.91–110 (2004).

9) Laptev, I. and Lindeberg, T.: Space-time interest points, *Proc. International Conference on Computer Vision* (*ICCV*), pp.432–439 (2003).

10) Wong, S. and Cipolla, R.: Extracting spaciotemporal interest points using global information, *Proc. International Conference on Computer Vision* (*ICCV*), pp.1–8 (2007).

11) Fischler, M.A. and Bolles, R.C.: Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography, *Communications of the ACM*, Vol.24, No.6, pp.381–395 (1981).

12) Rousseeuw, P.J. and Leroy, A.W.: *Robust Rgression and Outlier Detection*, Wiley, New York, USA (1987).

13) Lucas, B. and Kanade, T.: An iterative image registration technique with an application to stereo vision, *Proc. International Joint Conference on Artificial Intelligence*, pp.674–679 (1982).

14) Shi, J. and Tomasi, C.: Good features to track, *Proc. Conference on Computer Vision and Pattern Recognition* (*CVPR*), pp.593–600 (1994).

15) Ristic, B., Arulampalam, S. and Gordon, N.: *Beyond the Kalman filter: Particle filters for tracking applications*, Artech House Publishers, Boston (2004).

16) Fisher, R.A.: The use of multiple measurements in taxonomic problems, *Annals Eugen.*, Vol.7, pp.179–188 (1936).

17) Schmid, C., Mohr, R. and Bauckhage, C.: Evaluation of Interest Point Detectors, *International Journal of Computer Vision*, Vol.37, No.2, pp.151–172 (2000).

18) Beis, J.S. and Lowe, D.G.: Shape Indexing Using Approximate Nearest-Neighbour Search in High-Dimensional Spaces, *Proc. 1997 Conference on Computer Vision and Pattern Recognition* (*CVPR '97*), pp.1000–1006, Washington, DC, USA, IEEE Computer Society (1997).

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

(Communicated by *Vincent Lepetit*)

Ukrit Watchareeruetai received his B.Eng. in Electrical Engineering from Kasetsart University in 2002 and Master of Information Science and Doctor of Engineering degrees from the Graduate School of Information Science, Nagoya University in 2007 and 2010, respectively. From 2002 to 2004, he worked as a research assistant at Kasetsart Signal and Image Processing Laboratory. During his internship in 2009, he was working at NTT Communication Science Laboratories. Currently, he is a lecturer at the Department of Engineering and Technology, International College, King Mongkut's Institute of Technology Ladkrabang (KMITL). He received the IPSJ Digital Courier Funai Young Researcher Encouragement Award in 2011. His research interests include computer vision, pattern recognition and evolutionary computation. He is a member of IEEE, IPSJ, and ECTI.

Akisato Kimura received his B.E., M.E. and D.E. degrees in Communications and Integrated Systems from Tokyo Institute of Technology, Japan in 1998, 2000 and 2007, respectively. Since 2000, he has been with NTT Communication Science Laboratories, NTT Corporation, where he is currently a senior research scientist in Innovative Communication Laboratory. He has been engaged in content-based multimedia content identification, computational models of human visual attention, automatic image/video annotation, and social media mining. His research interests include pattern recognition, computer vision, image/video processing, human visual perception, statistical signal processing, machine learning and information theory. He is a senior member of IEICE and IEEE.

Robert Cheng Bao was born in 1985. He received his B.A.Sc. degree from the University of British Columbia, Canada in 2011, with Major in Engineering Physics Electrical Option, Minor in Commerce, and Minor in Economics. He worked with Ventyx Corp., Canada in 2007, NTT Corp., Japan in 2008, and Microsoft Corp., U.S.A. in 2009, respectively. During his internship with NTT Corp., he was engaged in research involving interest point detection with stochastic features.

Takahito Kawanishi was born in 1973. He received his B.E. degree from Kyoto University in 1996, his M.E. and Ph.D. degrees from Nara Institute of Science and Technology in 1998 and 2006, respectively. He has been working in NTT Corp. since 1998 and now is a senior research scientist of the Communication Science Laboratories of NTT. He has been engaged in the research area of media content identification, monitoring and search. He received the FIT 2003 Young Researcher Award in 2003. He is a member of IPSJ and IEICE.

Kunio Kashino is a Leader of Media Recognition Research Group at NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation. He is also a visiting professor at National Institute of Informatics, Japan. He has been working on audio and video analysis, search, retrieval, and recognition algorithms and their implementation. He received his Ph.D. degree from the University of Tokyo in 1995. He is a member of IPSJ, IEICE, JSAI, ASJ, ACM, and a Senior Member of the IEEE.