Combination Way of Local Properties, Classifiers and Saliency in Bag-of- Keypoints Approach for Generic Object Recognition

Shogo Nakamoto and Takashi Toriu[†],

Graduate school of Engineering, Osaka City University, Osaka, 558-8585 Japan

Summary
Csurka et. al. proposed a bag-of-keypoints approach which represents an image by a histogram of the number of occurrences of local properties at keypoints. In this approach, Scale Invariant Feature Transform (SIFT) descriptors are utilized for generic object recognition. As an alternative to SIFT, a method based on Speeded Up Robust Features (SURF) are reported to give better performance at greater speeds than SIFT. In this paper, we investigate combination way of SIFT or SURF and current various classifies such as Naïve Bayes, SVM and so on. We also propose a scheme in which a saliency map is utilized for removing irrelevant keypoints. We demonstrate that removing keypoints based on saliency improves classification rate in some situation.

Key words:
Generic object recognition, SIFT, SURF, Saliency map, bag-of-keypoints.

1. Introduction

Generic Object Recognition is the task of classifying an individual object in a real scene into a generic category. This task is one of the most important and the most difficult research topics in computer vision. Csurka et. al. proposed a bag-of-keypoints approach to generic object recognition [1]. In this approach, a feature vector is extracted from the image by constructing a histogram of the number of occurrences of descriptors associated with SIFT (Scale Invariant Feature Transform), and based on this feature vector, the image is categorized using one of classifiers such as Support Vector Machine (SVM), Naïve Bayes and so on. Milajczyk and Schmid [2] compared several descriptors in the scheme of Bag-of-Keypoints and showed that the descriptors in SIFT has highest performance among GLOH (Gradient Location and Orientation Histogram) [2], PCA-SIFT [3] etc.

In the meanwhile, human does not process whole area of an input visual image uniformly, but he usually focuses his visual attention on a limited area. Allocating computational resources intensively to the attended area would enable not only rapid but also accurate reaction. Itti, and Koch [4] proposed a model that calculates a saliency map, which represented degree of attractiveness of attention at each point of the image. Toriu et. al. [5] proposed another method to obtain saliency map, which had a learning step based on PCA and could take account of effect of visual experiences.

In this paper, first, we investigate combination way of two kinds local properties of SIFT and SURF utilized in the scheme of bag-of-keypoints and currently used various classifies such as Naïve Bayes, SVM and so on. As a result, we show that SVM with a non-linear kernel has highest performance and SURF descriptor has higher performance than SIFT descriptor. Secondly, we investigate what happens when saliency is used to reduce keypoints so as to enhance feature in object region and show that removing keypoints based on saliency improves classification rate in some situation.
In the next section, we outline the scheme of bag-of-key points and several current classifiers. In section 3, we summarize a method to obtain saliency map based on two-step PCA, and propose a scheme in which a saliency map is utilized for removing irrelevant keypoints. In section4, we conduct two experiments. In the first experiment, we compare combination ways of SIFT or SURF and current various classifies. In the second experiment, we investigate effectiveness of the saliency map when it is used for removing irrelevant keypoints. We discuss the results of the experiments in section 5, and conclude in section 6.

2. Bag-of-keypoints, Local properties and Classifiers

2.1 Bag-of-keypoints

In the scheme of bag-of-keypoints, the image is represented as a collection of local properties at keypoints. In the learning stage, a code book is constructed by clustering the local properties at keypoints for many learning images by using the k-clustering method. The code corresponding to each cluster is treated as a visual word. By mapping the local properties to the visual word, it is possible to construct a histogram by counting the frequency of the visual words. The histogram is a feature of that image and it is called the bag-of-keypoints. Figure 1 shows a flow of constructing bag-of keypoints. When a set of images are represented by their bag-of-keypoints,

Manuscript received May 5, 2011
Manuscript revised May 20, 2011

any classifier can be applied to these feature vectors.

2.2 Local properties

In this paper, we compare two methods of obtaining local properties, SIFT [6] and SURF [7]. In both methods, first, keypoints are detected. In the method of SIFT, keypoints are obtained using DoG (Difference of Gaussian), and the local properties construct a 128 dimensional vector. In the method of SURF, keypoints are obtained by Hesse matrix and, the local properties construct a 256 dimensional vector.

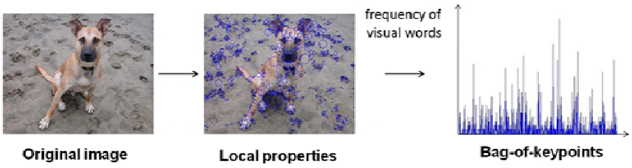


Fig. 1 Flow of bag-of-keypoints construction.

2.3 Classifier

We compare four classifiers, Naïve Bayse [8], linear SVM [9], SVM with Gaussian kernel and SVM with heavy trailed rbf kernel.

In the classification by Naïve Bases, an image I_i is classified to the class C_j such that the conditional probability $P(C_j \mid I_i)$ is maximum, where P(C_j|I_i) is evaluated as

$$P(C_j \mid I_i) \propto P(C_j)P(I_i \mid C_j)\,. \tag{1}$$

Let v_t be a visual word, V be a set of all visual words, and $N(t,i)$ be a number of visual word v_t . Then,

$$P(I_i \mid C_j) = \prod_{t=1}^{|V|} P(v_t \mid C_j)^{N(t,i)}\,, \tag{2}$$

where

$$P(v_t \mid C_j) = \frac{1+\sum_{\{I_i \in C_j\}} N(t,i)}{|V| + \sum_{s=1}^{|V|}\sum_{\{I_i \in C_j\}} N(s,i)}\,. \tag{3}$$

In equation (3), Laplace smoothing is performed by adding 1 to frequency of visual words to avoid the zero frequency problem.

In addition to the linear kernel, we employ following two kernels: Gaussian kernel

$$\kappa(\mathbf{x},\mathbf{y}) = \exp(-\frac{(\mathbf{x}-\mathbf{y})^2}{2\sigma^2})\,, \tag{4}$$

and heavy tailed rbf (HTRBF) kernel

$$\kappa(\mathbf{x},\mathbf{y}) = \exp(-\rho\textstyle\sum_i | x_i{}^a - y_i{}^a |^b)\,. \tag{5}$$

3. Saliency map

The degree to which visual attention easily attracted has been quantified as saliency. Itti and Koch [4] proposed a model to obtain a saliency map from local image features such as brightness, color, and orientation. The saliency map depends only on the input image and is independent of visual experience.

Toriu et. al proposed a model that obtains the saliency map based on not only the input image but also the results of learning [5]. Figure 2 shows the outline of the model.

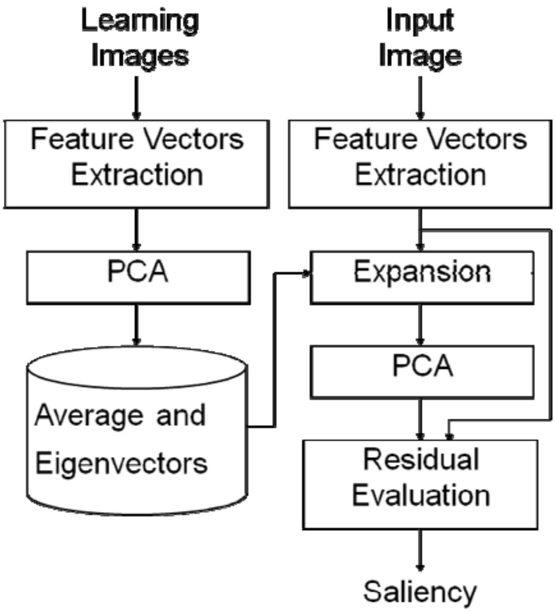


Fig. 2. Outline of the method to detect saliency based on two-step Principal Component Analysis.

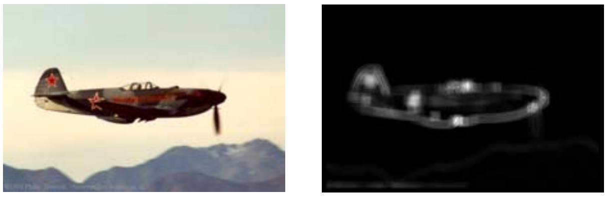
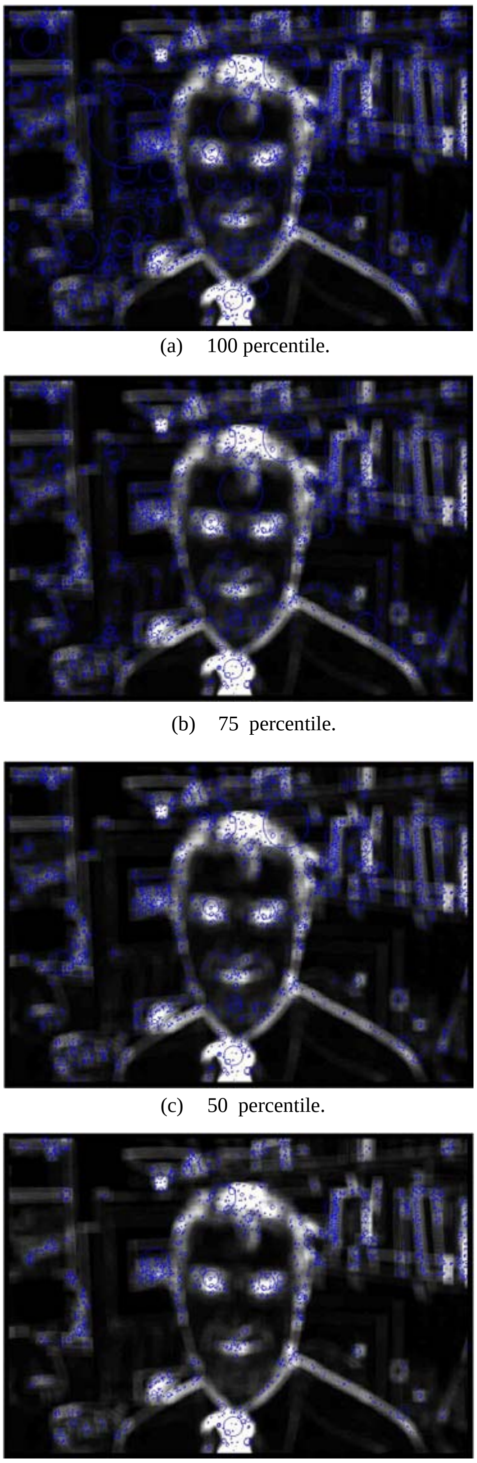


Fig. 3 An example of a saliency map.



(a) 100 percentile.

(b) 75 percentile.

(c) 50 percentile.

(d) 25 percentile.

Fig.4. Examples of detecting keypoints in the high saliency area.

In the learning step, first, feature vectors are extracted at all positions of numerous learning images. Then Principal Component Analysis (PCA) is applied to the feature vectors. By the PCA the average of the feature vectors and a number of eigenvectors are obtained as principal basis vectors. In the searching step, the feature vector is extracted at each position of the input image in the same way as in the learning step. Then each feature vector is expanded on the basis memorized in the dictionary. Then the weights are combined into a vector. The set of the vectors extracted from the input image in this way is analyzed by another PCA. The residual when the feature vector is approximated by the principal basis vectors in the second Principal Component Analysis is output as the saliency.

Figure 3 shows an example of a saliency map. The left hand side image is the original image and the saliency map of this image is shown on the right hand side as a grey scale image, where high intensity means high saliency. Saliency map represents where attention should be focused. If saliency is high at a certain area the keypoints in this area are expected to be of more importance. The area where saliency is high can be extracted by applying thresholding operation to the image of the saliency map. The threshold is determined, for example, by percentile thresholding Once the high saliency area is extracted, keypoints are detected in this area. Figure 4 shows examples of results of detecting SIFT keypoints in the high saliency area extracted by percentile thresholding.

4. Experiments

We conducted two experiments. In the first experiment, we compared several combination ways of local properties and classifies. In the second experiment, we investigated effectiveness of using saliency map to restrict the area where keypoints are detected. Table 1 shows the combination ways in the experiment 1, and Table 2 shows the combination ways in experiment 2. In the second experiment, we used SVM (HTRBF) as a classifier, which has the best performance in the first experiment.

Table 1. Experiment 1.

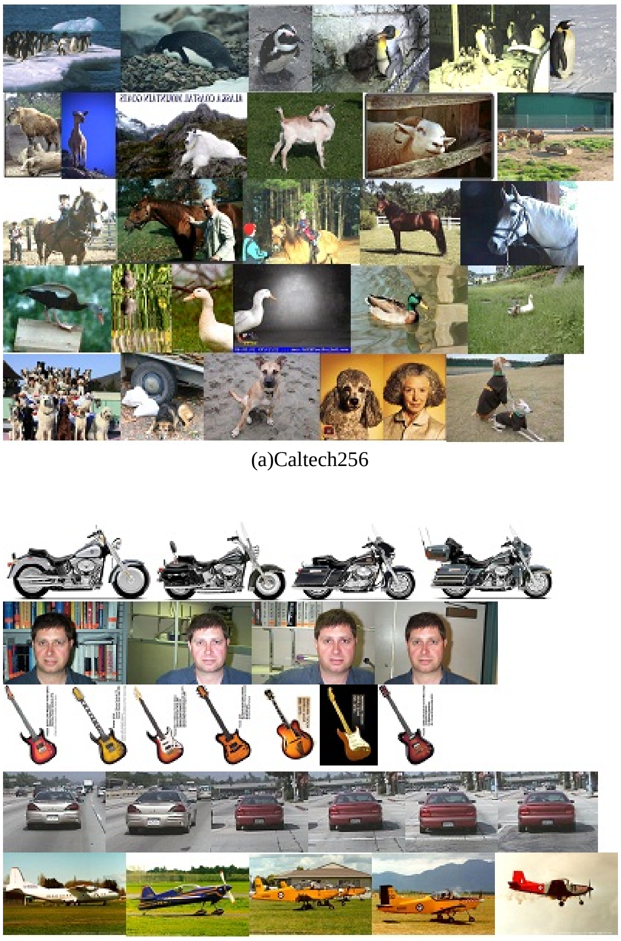
Local properties	Classifier

SIFT	Naïve Bayes
	SVM(linear)
	SVM(Gaussian)
	SVM(HTRBF)

SURF	Naïve Bayes
	SVM(linear)
	SVM(Gaussian)
	SVM(HTRBF)

Table 2. Experiment 2.

Local properties	Classifier
SIFT	100 percentile 75 percentile 50 percentile 25 percentile
SURF	100 percentile 75 percentile 50 percentile 25 percentile



(a)Caltech256

(b)Oxfor

Fig. 5 Examples of images of the data sets used in the experiment.

We used two image data sets, one is Caltech256 data set [10], and the other is Oxford data set [11]. We selected 5 categories in each data set; “dog”, “duck”, “goat”, “horse”, and “penguin” in the Caltech256 data set, and “airplanes_side”, “cars_brad”, “faces”, “guitars”, and “motorbikes side” in the Oxford data set. Figure 5 shows examples of images in each data set. We can see that inner class variation of the Caltech256 is large, while that of the Oxford data set is not so large.

4.1 Results of experiment 1

4.1 Results of experiment 1

Table 3 shows the numbers of images used in the first experiment. We used 50 images in each category for learning and 30 images in each category for testing in the case of the Caltech256 data set. As for the Oxford data set, we used 100 images in each category for learning and 100 images in each category for testing.

Table 3. The numbers of images used in the experiment 1.

	learning	testing	total
Caltech256	250	150	400
Oxford	500	500	1000

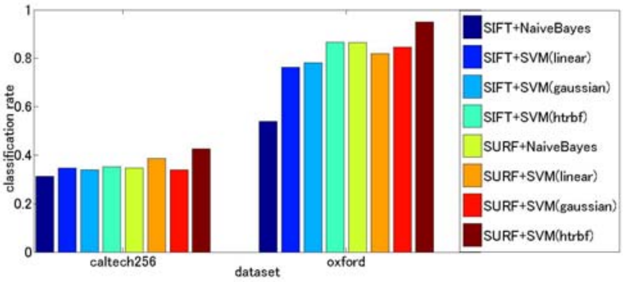


Fig. 6 Classification rates in the experiment 1.

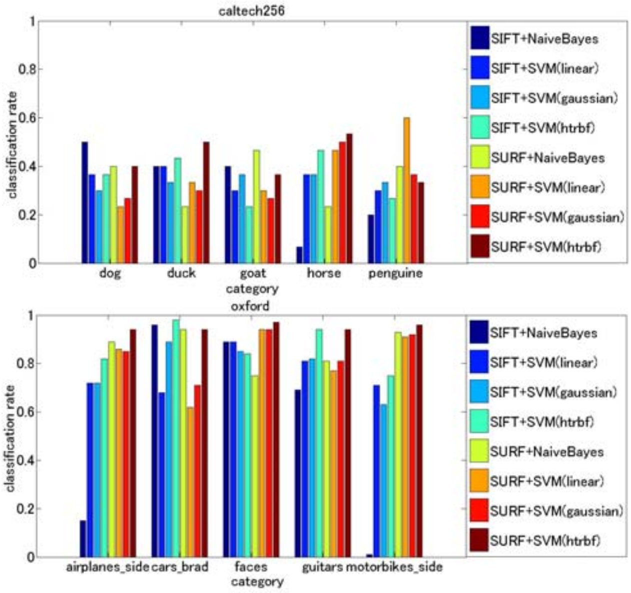


Fig. 7 Classification rates for each category in the experiment 1.

Figure 6 shows the results of the experiment. It shows the classification rate for each combination of local properties and classifiers. The classification rate is defined as the ratio of the number of correctly classified images to the number of all test images. We can see that SVM has higher performance than Naïve Bases, and as for SVM, non-linear kernel is superior to linear one. We can also see that SURF descriptor has higher performance than SIFT descriptor. The combination of SURF descriptor and SVM with HTRBF kernel has the highest performance.

Figure 7 shows the classification rates for each category. The classification rate is defined as the ratio of the number of correctly classified images to the number of all images included in that category. We can see that SVM with HTRBF kernel has the highest performance for any category, and that SURF descriptor has higher performance than SIFT descriptor irrespectively of the category.

Table 4. The number of images used in the experiment 2.

	learning	Testing	total
Caltech256	250	150	400
Oxford	400	250	650

4.2 Results of experiment 2

Table 4 shows the numbers of images used in the second experiment. The same number of images is used in the case of the Caltech256 data set. As for the Oxford data set, we used 100 images in each category for learning and 50 images in each category for testing.

Figure 8 shows the results of the second experiment. In the case of Oxford data set, which has comparatively low inner class variation, the classification is highest for the combination of the SURF descriptor and zero percentile thresholding. It means that removing keypoints using the saliency map is not effective in this case. In the case of the SIFT descriptor, 25 percentile thresholding for the saliency map has the highest classification rate but the difference is not so significant. When we use Caltech256 data set, which has comparatively large inner class variation, the effect of removing keypoints according to saliency is more distinctive. 50 percentile thresholding is best in the case of the SIFT descriptor and 25

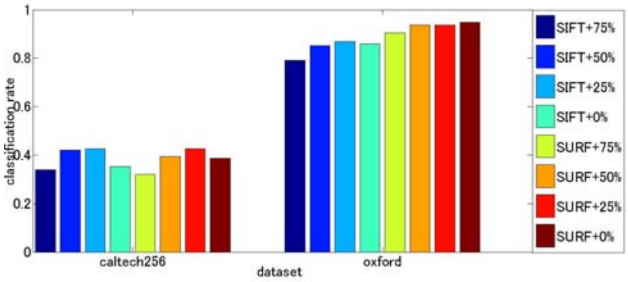


Fig. 8 Classification rates in the experiment 2.

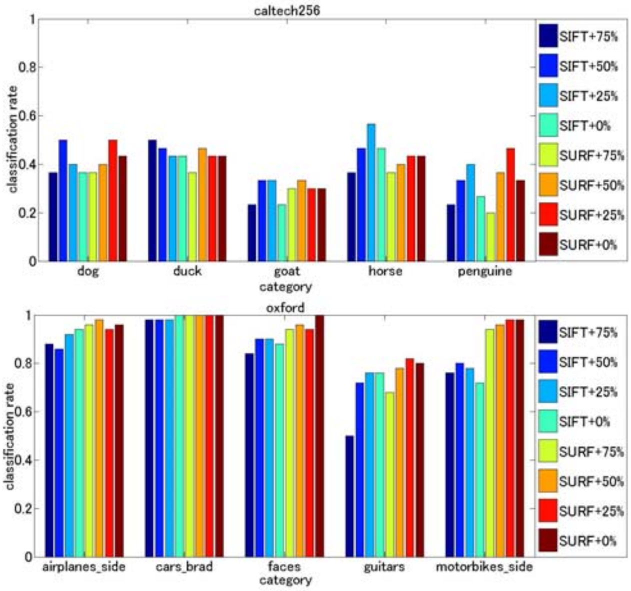
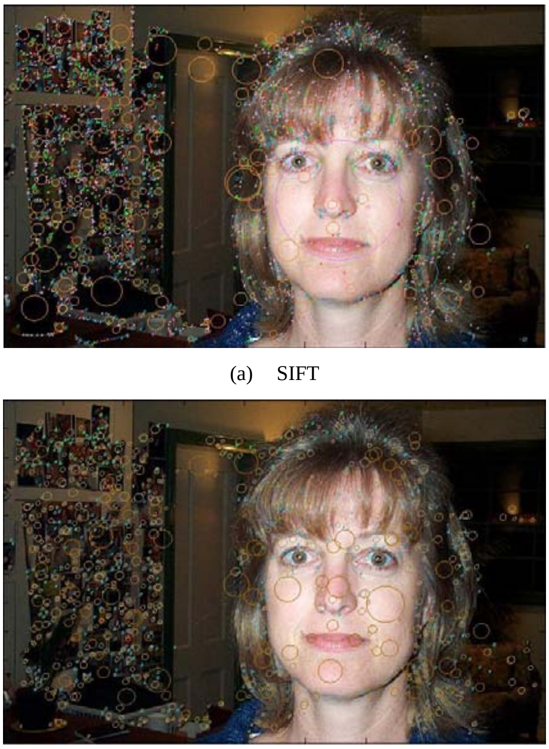


Fig. 9 Classification rates for each category in the experiment 2.

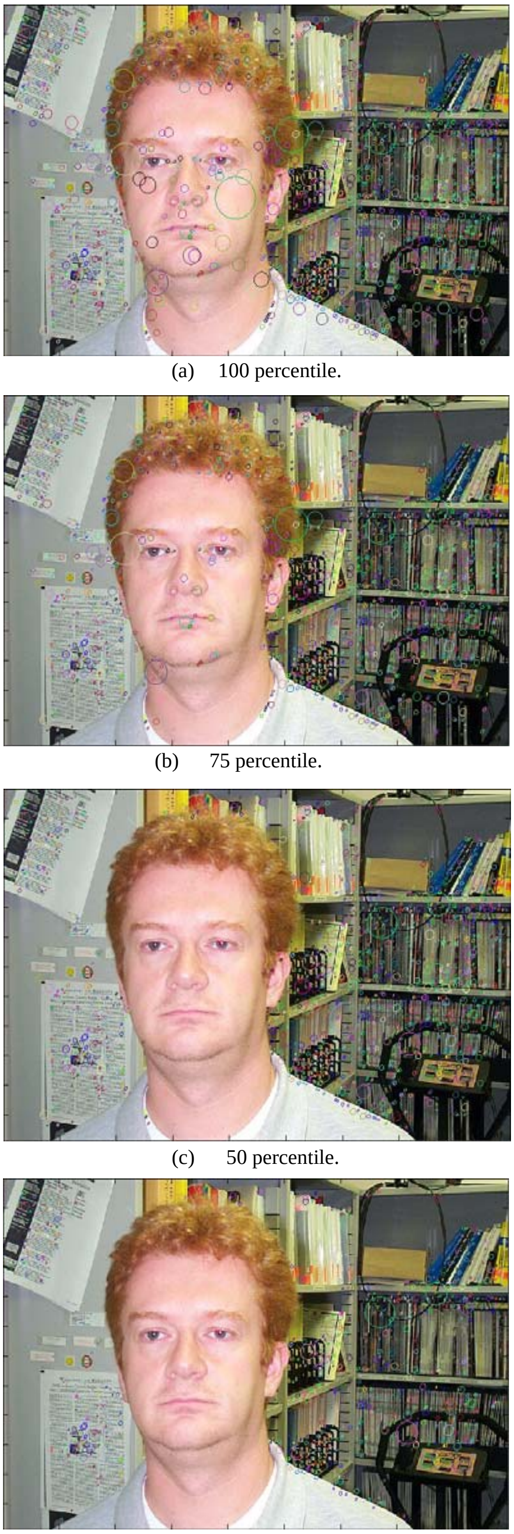


(a) SIFT

(b) SUR

Fig. 10 Comparison between SIFT and SURF.

percentile thresholding is best in the case of the SURF descriptor. It should be noted that when the saliency map is not used the SURF descriptor is superior to the SIFT descriptor but when the saliency map is used to remove irrelevant keypoints the SIFT descriptor with 25 percentile thresholding has the highest classification rate. Figure 9 shows the classification rate for each category. In the case of Oxford data set, when the saliency map is used for removing irrelevant keypoints, the classification rate often decreases. On the contrary, in the case of Caltech256 data set, the classification rate increases when the saliency map is used for removing irrelevant keypoints.



(a) 100 percentile.

(b) 75 percentile.

(c) 50 percentile.

(d) 25 percentile.

Fig. 11SURF keypoints distribution for an image in Oxford data set.

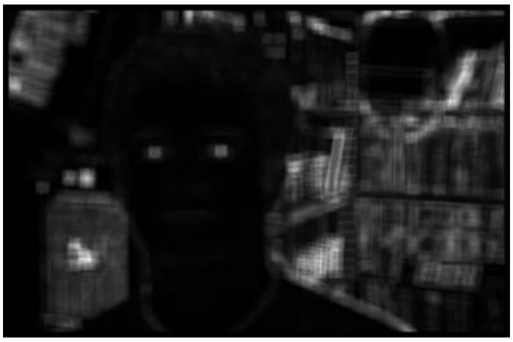
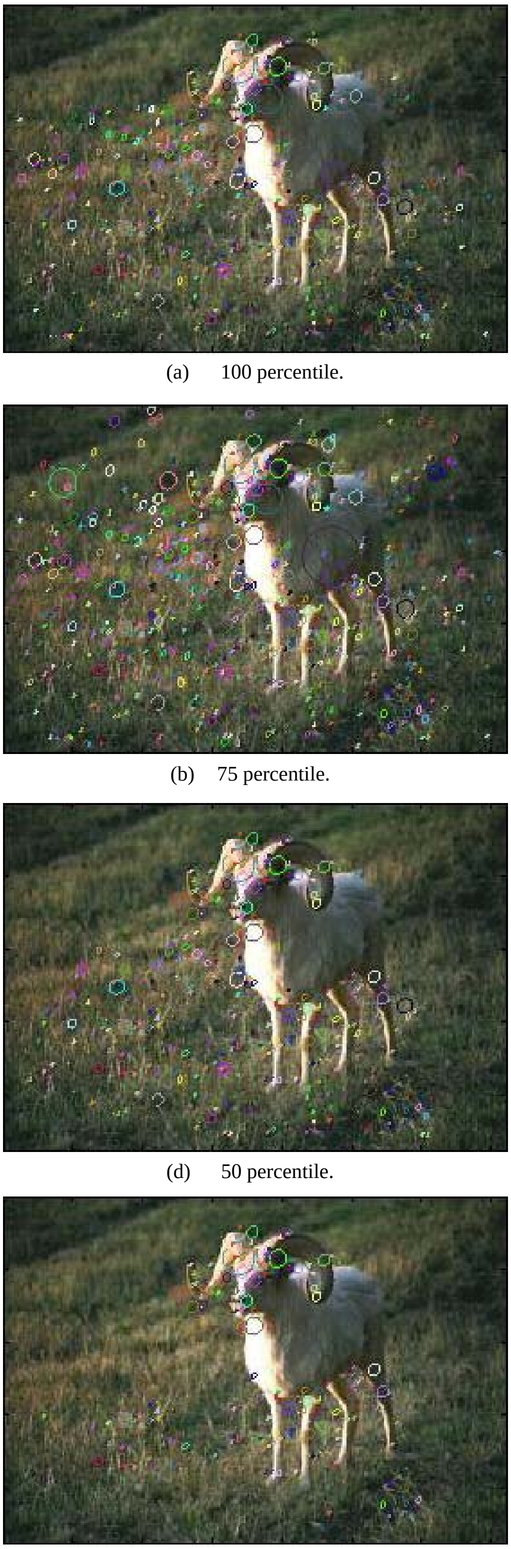


Fig. 12The saliency map corresponding to the images in figure 11.

5. Discussion

In the first experiment, we saw that SURF was superior to SIFT, while there had been reported that SIFT was superior to SURF in terms of matching accuracy of keypoints when images were geometrically transformed by view point change [12]. We consider the reason why this difference comes in the following. Figure 10 (a) shows SIFT keypoints distribution and Figure 10 (b) shows SURF keypoints distribution. In these images, the circles show keypoints, and the radius of each circle shows the scale. Comparing these two images, we can see that the number of keypoints is different with each other. The number of SIFT keypoints is larger than SURF keypoints in the background. This is considered to be the reason why SURF is superior to SIFT in the first experiments.

In the second experiment, we investigated the effect of the saliency map which was used for removing irrelevant keypoints. As for Oxford data set, the classification rate became low when keypoints were removed using the saliency map. Figure 11 shows the SURF keypoints distribution when the high saliency area is extracted by the percentile thresholding method, and figure 12 shows the saliency map. The saliency is high in the background rather than the face region. Accordingly, keypoints are often removed in the object region rather than in the background. This is considered to be the reason why the classification became low when keypoints were removed using saliency map. On the other hand, in the case of Caltech256 data set, the classification rate became high when keypoints were removed using the saliency map. Figure 13 shows the SIFT keypoints distribution when the high saliency area is extracted by the percentile thresholding method, and figure 14 shows the saliency map. In this case, saliency is relatively high in the region of the object, and as a result, keypoints in the region of background are efficiently removed.



(a) 100 percentile.

(b) 75 percentile.

(d) 50 percentile.

(e) 25 percentile.

Fig. 13 SIFT keypoints distribution for an image in Caltech256 data set.

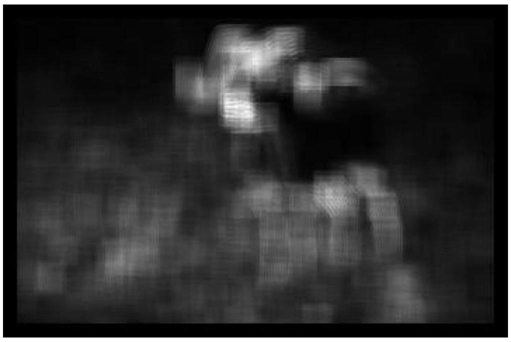


Fig. 14 The saliency map corresponding to the images in figure 13.

5. Conclusion

In this paper, we investigated combination way of local properties utilized in the scheme of bag-of-keypoints and currently used various classifies such as Naïve Bayes, SVM and so on. We found that SVM has higher performance than Naïve Bases, and SVM with non-linear kernel is superior to that with linear kernel. We also found that SURF descriptor has higher performance than SIFT descriptor. The combination of SURF descriptor and SVM with HTRBF kernel has the highest performance.

Secondly, we investigated the effectiveness of the saliency map to reduce keypoints so as to enhance feature in object region. We found that the classification rate becomes higher when the saliency map is used for Caltech256 data set, but the effect is not seen in the case for Oxford data set. The reason might be that background sometimes has more saliency than the object in Oxford data set. More elaborate investigation is needed to evaluate the effectiveness of using the saliency map in the scheme of bag-of-keypoints. It is also necessary to compare the saliency map used in this paper and the original saliency map proposed by Itti et. al. [4].

Acknowledgments

This work is partially supported by SCOPE: Strategic Information and Communications R&D Promotion Program (102307010).

References

[1] G. Csurka, C. Bray, C. Dance, and L. Fan, “Visual categorization with bags of keypoints,” Proc. of ECCV Workshop on Statistical Learning in Computer Vision, pp. 1-22, 2004.

[2] K. Mikolajczyk, C. Schmid, “A performance evaluation of local descriptors,” IEEE Trans. Pattern Analysis and Machine Inteligence, vol. 27, no. 10, pp. 1615-1630, 2005.

[3] Y. Ke, R. Sukthankar, “PCA-SIFT:A more distinctive representation for local image descriptors,” Proc. of IEEE Conference on Computer Vision and Pattern Recognition, pp. 511-517, 2004.

[4] L.Itti, C.Koch and E.Niebur, “ A Model of Saliency-based Visual Attention for Rapid Scene Analysis, ”IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 20, No. 11, pp. 1254-1259,1998.

[5] T. Toriu and S. Nakajima, "A method of calculating saliency of images and of optimizing efficient distribution of image windows,” International Journal of Innovative Computing, Information and Control, Vol.3, No.6, pp. 1407-1419, 2007.

[6] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” Proc. of International Journal of Computer Vision, vol. 60, no. 2, pp. 91-110, 2004.

[7] H. Bay, T. Tuytelaars, and L. J. Van Gool, “SURF: Speeded Up Robust Features,” Proc. European Conference on Computer Vision, pp. 404-417, 2006.

[8] D. D. Lewls, “Naive Bayes at forty: The independence assumption in information retrieval,” Proc. of European Conference on Machine Learning, pp. 4-15, 1998.

[9] V. Vapnik, Statistical Learning Theory, John Wiley Sons, New York, NY, 1998.

[10] http://www.vision.caltech.edu/Image Datasets/Caltech256/

[11] http://www.robots.ox.ac.uk/˜vgg/data/datacats.html

[12] J. Bauer, N. S¨underhauf, and P. Protzel, “Comparing Several Implementations of Two Recently Published Feature Detectors,” In Proc. of the International Conference on Intelligent and Autonomous Systems, IAV, Toulouse, France, 2007.

Shogo Nakamoto received the B.E. in Information Engineering from Osaka City University, Osaka, Japan in 2011. He is now in his Master first year in Information Engineering, Graduate School of Engineering, Osaka City University. His research interests include image processing.

Takashi Toriu received the B.Sc. in 1975, M.Sc. and Ph.D. degree in physics from Kyoto University, Kyoto, Japan, in 1977 and 1980, respectively. He was a researcher in Fujitsu Laboratories Ltd. from 1982 to 2002, and now he is a Professor of Osaka City University. His research interests are in the areas of image processing, computer vision, and especially in modeling of human visual attention. He is a member of IEEE, IEICE, IPSJ, ITE and IEEJ.