Preliminary Local Feature Selection by Support Vector Machine for Bag of Features

Tetsu Matsukawa† Koji Suzuki†
Takio Kurita ‡

†:University of Tsukuba
‡:National Institute of Advanced Industrial Science and Technology
{t.matsukawa, k-suzuki, takio-kurita}@aist.go.jp

Abstract This paper proposes a selection method of foreground local features for generic object recognition in "bag of features". Usually all local features detected from an given image are voted to a histogram of visual words in conventional bag-of-features method. But it may not be good choice because in the standard object recognition task, an image includes target regions and background regions. To distinguish the target from the background, a large number of visual-words are necessary because a variation of local features coming from the background regions is usually large. It is expected that the comparable classi cation performance will be achieved with a small number of visual words if such unimportant local features can be e ectively removed. Although it is di cult to correctly classify all local features into the target and the background, the number of visual-words can be reduced by simply neglecting many of the local features obtained from the background regions which are easily classi ed by Support Vector Machine (SVM). Experimental results showed the proposed method outperformed the conventional bag-of-features representation with a fewer number of visual-words by neglecting background features by the kernel SVM. The classi cation performance with linear SVM was also better than the conventional bag-of-feature when the number of visual words was small.

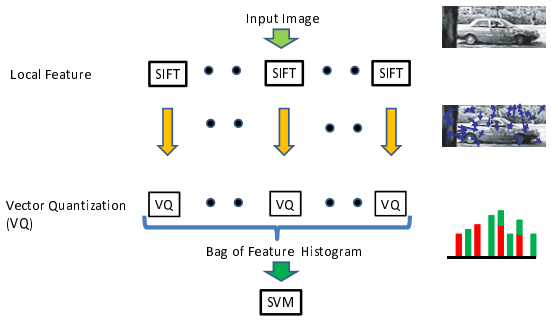
1 Introduction

Recently, bag-of-features method, which represents an image as an orderless collection of local features, has shown excellent results in generic object recognition problems[1,2,3,4,5,6]. Bag-of-features is an application of the text document classi cation method called "bag-of-words" to image classi cation problem by regarding local features as "visual words". Speci cally, bag-of-features method classi es local features into a large number of clusters using a vector quanti cation so that similar features assigned into the same cluster. Namely, each cluster can be regarded as a "visual word". After all local features of an image are mapped into the visual words, a histogram of the visual words is created in which each bin is corresponding to the visual word. The image category is recognized by Support Vector Machine (SVM)[7] using this histogram of visual words as the input feature vector.

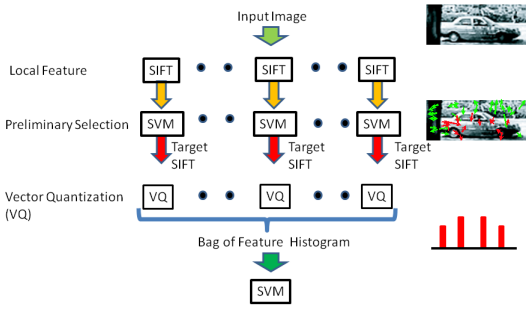
Usually all local features detected from an given image are voted to the histogram of visual words in conventional bag-of-features method. But it may not be good choice because in the standard object recognition task, the image includes target regions and background regions. A large number of visual words is necessary to e ectively distinguish the local features obtained from the target regions and the background regions because the variation of the local features obtained from the background regions is usually large.

In this paper, we propose a selection method of local features before calculating the histogram of visual words by removing the local features obtained from the background regions using SVM. Although it is di cult to correctly classify all local features into the the target's local features and the background's local features, the number of visual-words can be reduced with comparable classi cation performance by simply neglecting many of the local features which are easily classi ed as the background's features.

Previous Work Marszalek at el. proposed spatial weighting method to foreground regions by using the learned spatial relation of local features and target object mask [3]. They reported that the high



(a) Standard Bag of Features



(b) Proposed Method

Fig. 1: (a) Standard Bag of Features : histogram is constructed from all local features. (b) Proposed Method: histogram is constructed using only the local features classi ed to the target by SVM.

classi cation performances were achieved by reducing the in uences of the background. To learning the relation of the local features and the target object mask, they utilized the scale and the dominant gradient orientation of scale invariant detector. So, their method is applicable only for SIFT-like features. Shotton at el. proposed a method in which each local region is represented as semantic texton with probability of each category [4]. By assigning high weight for target category, they achieved high classi cation performance. Our proposed method is a simple extension of the standard bag-of-features and any local features can be used.

In section 2, we brie y review the standard bag-of-features method. Then the detail of the proposed method is described in section 3. The experimental result and conclusion are presented in section 4 and section 5, respectively.

2 Bag-of-Features

In this section, we brie y review the standard bag-of-features method (Fig.1(a)) [1][1]. The bag of features method is a classi cation method using orderless collection of quantized local features. The main step of bag-of-features are :

1. Detection and description of image patches
2. Assigning patch descriptors to a set of predetermined clusters (a vocabulary) with a vector quantization algorithm
3. Constructing a bag of features, which counts the number of patches assigned to each cluster
4. Applying a classi er by treating the bag of features as the features vector, and thus determine which category to assign to the image

In the step 1, we used Scale-Invariant Feature Transform (SIFT)[8] as local features. SIFT is a histogram of 8 gradients in 4x4 grid of spatial locations, giving a 128-dimension vector. We used the standard interest point detector proposed in [8][2]. In the step 2, we used k-means clustering to create visual vocabulary. In the step 4, we used linear SVM to classify the image category by using bag of features as the input features vector.

It is known that the bag of features method is robust for background clutter and produces good classi cation accuracy even without exploiting geometric information[3]. All visual-words calculated from an image are voted to a histogram in conventional bag-of-features method. But an image in the standard object recognition task usually includes the target regions and the background regions. Since the variation of the local features obtained from the background regions is usually large, the local features obtained from the background regions will a ect the recognition performance. Especially the classi cation performance will be decreased when the number of visual words is small, because the possibility to be assigned the local features from the target regions and the background regions into the same bin in the histogram of the visual words becomes higher. Namely a large number of visual words is necessary to e ectively distinguish the local features obtained from the target regions and the background regions. We expect that the comparable classi cation performance will be achieved with a small number of visual words if we can e ectively remove such unimportant local features obtained from the background regions.

[1]This method is called differently in many literatures such like "bag-of-visualwords", "bag-of-visterms" and "bag-of-keypoints". These terms indicate almost same method. The words "textons", "keypoints", "visual-words", "codebook" , "visal-terms" and "visterms" are also the same meanings.

The dense sampling[5] method which doesn't use interest point detector and effective sampling method[6] were also proposed .

[3]Recently, Spatial Pyramid Matching [2] were proposed to introduce coarse geometric information.

3 Proposed Method

Outline of the proposed method is shown in Fig. 1(b). Firstly, local features calculated from input image are classi ed to the background features and the target features by SVM. Using only the local features classi ed to the target, the histogram of visual words are created. The image category is recognized by using these histograms. In the following experiments, we applied the proposed method to the two-class classi cation task. The procedure for two-class classi cation tasks becomes as follows :

1. Detection and description of image patches
2. Classify local image patches to the foreground or the background by SVM.
3. Assigning the patch descriptors classi ed to the foreground to a set of predetermined clusters (a vocabulary) with a vector quantization algorithm
4. Constructing a bag of features, which counts the number of the foreground patches assigned to each cluster
5. Applying a classi er treating the bag of features as the features vector, and thus determine which category to assign to the image

In the step 2 and the step 5, we use SVM. The SVM determines the optimal hyperplane which maximizes the margin, where the margin is the distance between the hyperplane and the nearest samples. The decision function of linear SVM is de ned as

$$f(\mathbf{x}) = sgn(\sum_{i \in SV} \quad {}_i y_i \mathbf{x}_i \mathbf{x} \quad b), \tag{1}$$

where SV is a set of support vectors, y_i is the class label of $\mathbf{x}_i$ (+1 or 1), $_i$ is the learned weight of the training sample $\mathbf{x}_i$ and b is a learned weight of threshold parameter.

This assumes a linearly separable case. For a linearly non-separable case, the non-linear transform $\Phi(\mathbf{x})$ can be used. The training samples are mapped into the high dimensional space by $\Phi(\mathbf{x})$. By maximizing the margin in high dimensional space, nonlinear classi cation can be achieved. If the inner product $\Phi(\mathbf{x})^T \Phi(\mathbf{y})$ in the high dimensional space is computed by kernel $K(\mathbf{x}, \mathbf{y})$, then the training and the classi cation can be done without mapping into the high dimensional space. The decision function of kernel SVM is de ned by

$$f(\mathbf{x}) = sgn(\sum_{i \in SV} \quad {}_i y_i K(\mathbf{x}_i, \mathbf{x}) \quad b), \tag{2}$$

where, $K(\mathbf{x}_i, \mathbf{x})$ is the value of a kernel function for the training sample $\mathbf{x}_i$ and the test sample $\mathbf{x}$. In the following experiment, we used the Gaussian kernel de ned as

$$K(\mathbf{x}, \mathbf{y}) = exp(\quad \frac{|\ \mathbf{x} \quad \mathbf{y}\ |^2}{2}). \tag{3}$$

The regularization parameter C of SVM[7] and the kernel parameter were determined by using the grid search based on 5-fold cross validation.

4 Experiment

4.1 Experimental Settings

We evaluated the e ect of the proposed local features selection using UIUC Image Database for Car Detection[9]. The task in this experiment is to classify the input images into one of "car" and "not car" classes. Although this dataset has the training images and the test images, we used only the training images in this paper. The number of the training images is 1050 (550 car and 500 noncar). Examples of the training images are shown in Fig.2 (a). We created the mask images that indicate the car regions and the background regions. The examples of the mask images are shown in Fig.2 (b). In Fig.2 (b), the background regions are shown in black, the car regions are shown in red, and the occlusion regions are shown in green. In this paper, the occlusion regions are regard as the background regions. Using the mask images, we can automatically label the SIFT features detected from the image into the car regions or the background regions. The both of the linear and the kernel SVM were tried to remove the local features obtained from the background regions. LIBSVM is used for SVM implementation[10]. The SVM for SIFT selection was trained by using only SIFT features detected from the car images. Then the histograms were created from the car and non car images by applying the learned SVM for SIFT selection. SVM for classi cation was trained by these histograms. The classi cation performance of the proposed method was compared with the conventional bag-of-features method in the several conditions of the number of clusters. The classi cation accuracy was evaluated by 3-fold cross validation of 550 car and 500 non-car images.

4.2 E ect of Local Features Selection

In this section, we shows the e ects of the proposed local features selection. Fig.3 and 4 show examples of the selected SIFT features by using the trained kernel SVM. Fig.3 shows the results for a

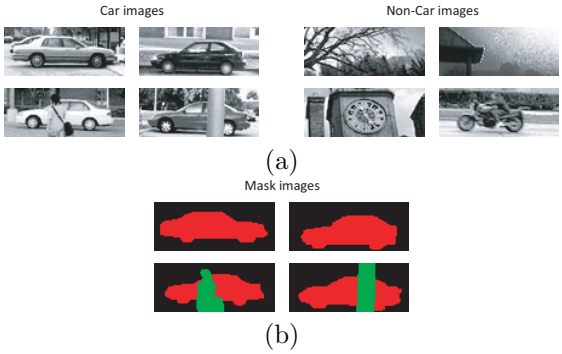


Fig. 2: Examples of the images in Dataset (a) The car and non car images (b) The mask images.

Table 1: Classi cation Performances of the SIFT Selection.

	Accuracy	$TP_{r\ te}$	$TN_{r\ tegrey}$
Kernel	80.79%	82.46%	78.76%
Linear	63.70%	75.85%	48.98%

car image. Although there are several misclassi - cation, the characteristic parts of car such as tires and windshield are correctly classi ed. Fig.4 shows the result for a non car image. It is noticed that some SIFT features are misclassi ed. To evaluate the performance of SIFT selection, three measures, Accuracy, True Positive Rate ($TP_{r\ te}$) and True Negative Rate ($TN_{r\ te}$) are used. These are de ned by,

$$Accuracy = 100 \quad \frac{TP+TN}{TP+FN+TN+FP}, \quad (4)$$

$$TP_{r\ te} = 100 \quad \frac{TP}{TP+FN}, \quad (5)$$

$$TN_{r\ te} = 100 \quad \frac{TN}{TN+FP}, \quad (6)$$

where TP means True Positive, TN means True Negative, FN means False Negative and FP means False Positive. These values are shown in Table.1. These are the average of 3-fold cross validation. The kernel SVM (Accuracy = 80.79%) were higher Accuracy than the linear SVM (Accuracy = 63.70 %). It is noticed that $TP_{r\ te}$ are higher than $TN_{r\ te}$ for both the kernel and the linear SVM, especially $TN_{r\ te}$ is very low in the linear SVM. This means that it is di cult to correctly distinguish the local features obtained from the background regions, especially for the linear SVM.

4.3 Results of Classification

The classi cation accuracies of the images are shown in Fig.5. The number of clusters are changed

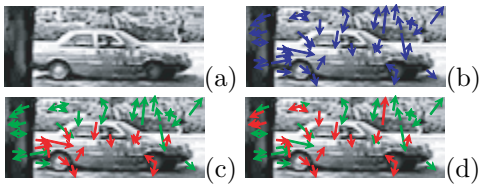


Fig. 3: SIFT selection results of a car image.(a): the original image, (b): all detected SIFT, (c): the correct classi cation, (d): the classi cation results by the kernel SVM (blue arrow: all detected SIFT, red arrow: target SIFT, green arrow: non-target SIFT).

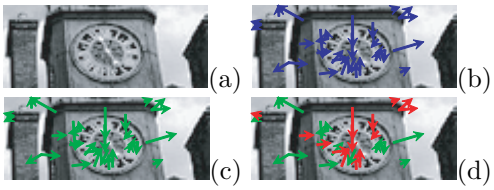


Fig. 4: SIFT selection results of a non car image.(a): the original image, (b): all detected SIFT, (c): the correct classi cation, (d): the classi cation results by kernel SVM (blue arrow: all detected SIFT, red arrow: target SIFT, green arrow: non-target SIFT).

from 50 to 500 with 50 intervals. The graph shows the average classi cation accuracies estimated by 3-fold cross validation and the standard errors. From Fig.5, it is noticed that the propose local features selection method using both the linear and the kernel SVM give higher accuracies when the number of clusters is small. When the number of clusters becomes larger than 150, the linear SVM based selection method gives lower accuracies than other methods. On the other hands, the kernel SVM based selection method is comparable to the conventional bag-of-features method even when the number of clusters is larger than 300.

Fig.6 shows the histograms of visual words computed from a car image. Similarly the histograms of visual words computed from a non car image are shown in Fig.7. In each histogram, red value indicates the SIFT obtained from the car regions and green indicates the SIFT obtained from the background regions. The bins of the histogram are sorted by the weights of the linear SVM for classi cation. This means that the left bins contribute more to support for "car" class and on the other hand, the right bins contribute more to support for "non car" class. From Fig.5, it is noticed that the accuracy of conventional method is lower than the proposed method when the number of clus-

ters is small. The reason of this is convinced from Fig.6 (a1). This is because the votes of SIFT features obtained form the car regions and the background regions are mixtured in the histogram when the number of clusters is small as shown in Fig.6 (a1). Thus the histograms of car (a1) and non car (c1) are hardly separable. On the other hand, the proposed method (kernel) could remove almost all votes from the background regions as shown in Fig.7 (a2). As the results, the histograms of car (a2) and non car (c2) could be easily separable. The linear SVM could not correctly remove the features obtained from the background regions as good as the case of the kernel SVM. Thus the votes from the background regions remains as shown in (a3). Also the linear SVM fails to classify the background features enough as shown in (c3) compared to the case of the kernel SVM in (c2).

When the number of clusters is large, the di erence of classi cation accuracies between the conventional bag-of-features method and the proposed methods (the kernel and the linear SVM) became smaller. This is because the bins from the foreground and the background becomes separable. This can be con rmed from the Fig.6 (b1), (b2), and (b3) and Fig.7 (d1), (d2), and (d3).

In summary, experimental results showed the proposed method outperformed the conventional bag-of-features representation when the number of clusters was small as shown in Fig.5. In the most cases, the proposed bag-of-features representation with 50 clusters achieved higher classi cation performance than the conventional bag-of-features method with 500 clusters. The classi cation performance with the linear-SVM based selection was also better than the conventional bag-of-feature when the number of clusters were lower than 100. In the local features selection performance, the kernel SVM achieved higher performance than the linear SVM. This means there is the positive correlation of the preliminary selection performance and the nal classi cation performance.

5 Conclusion

In this paper, we proposed a selection method of foreground local features for generic object recognition in "bag of features". Through experiments using UIUC Image Database for Car Detection, we have con rmed that the proposed method outperformed the conventional bag-of-features representation when the number of clusters is small. In the local features selection performance, the kernel SVM achieved higher performance than the linear SVM. For the nal classi cation performance were also better when the kernel SVM is used to select local features. This means that there is the positive correlation between the selection performance and the nal classi cation performance.

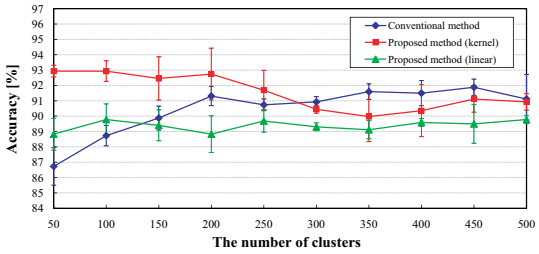


Fig. 5: Classi cation Accuracies of the images. Dot indicates the average classi cation accuracies and the vertical bar indicates the standard error computed by 3-fold cross validation.

For future works, we have to apply the proposed local features selection to other generic object recognition problems and to evaluate the e ectiveness of the proposed approach.

References

[1] G.Csurca, C.Dance, L.Fan and J.Willamowsli and C.Bray, Visual categorization with bags of keypoints, In ECCV International Workshop on Statistical Learning in Computer Vision, 2004.

[2] S.Lazebnik, C.Schmid, and J.Ponece, Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories, In CVPR, pp.2169-2178, 2006.

[3] M.Marszalek and C.Schmid, Spatial Weighting for Bag-of-Features, In CVPR, 2006.

[4] J.Shotton, M.Johnson, and R.Cipolla, Semantic Texton Forests for Image Categorization and Segmentation, In CVPR, 2008.

[5] F.Jurie and B.Triggs, Creating E cient Codebooks for Visual Recognition, In ICCV, Vol.1, pp.604-610, 2005.

[6] E.Nowak, F.Jurie and B.Triggs, Sampling Strategies for Bag-of-Features Image Classi cation, In ECCV, pp.490-503, 2006

[7] V.Vapnik, Statistical Learning Theory, John Wiley & Sones, 1998.

[8] D.Lowe, Distinctive image features form scale-invariant keypoints, IJCV, 60(2), pp.91-110, 2004.

[9] S.Agarwal, A.Awan and D.Roth, Learning to detect objects in images via a sparse, part-based representation. IEEE Trans. on PAMI, 26(11), pp.1475-1490, 2004.

[10] C-C. Chang and C-J. Lin, LIBSVM: a library for support vector machines, Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm, 2001.

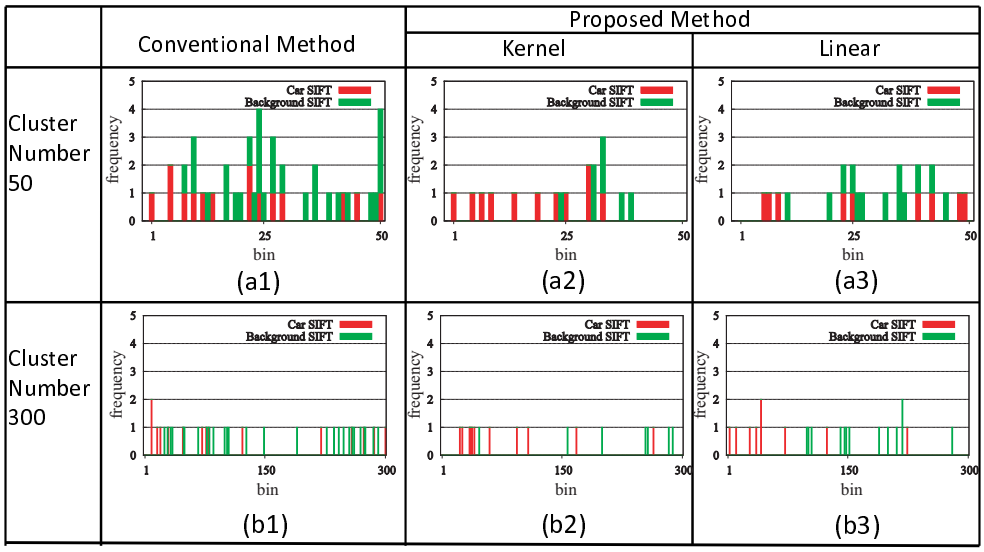


Fig. 6: Example of the histograms of visual words for a car image. (a1): the conventional method, the number of clusters is 50, (a2): kernel SVM , the number of clusters is 50, (a3): linear SVM, the number of clusters is 50, (b1): the conventional method, number of clusters is 300, (b2): kernel SVM, number of clusters is 300, (b3): linear SVM, the number of clusters is 300.

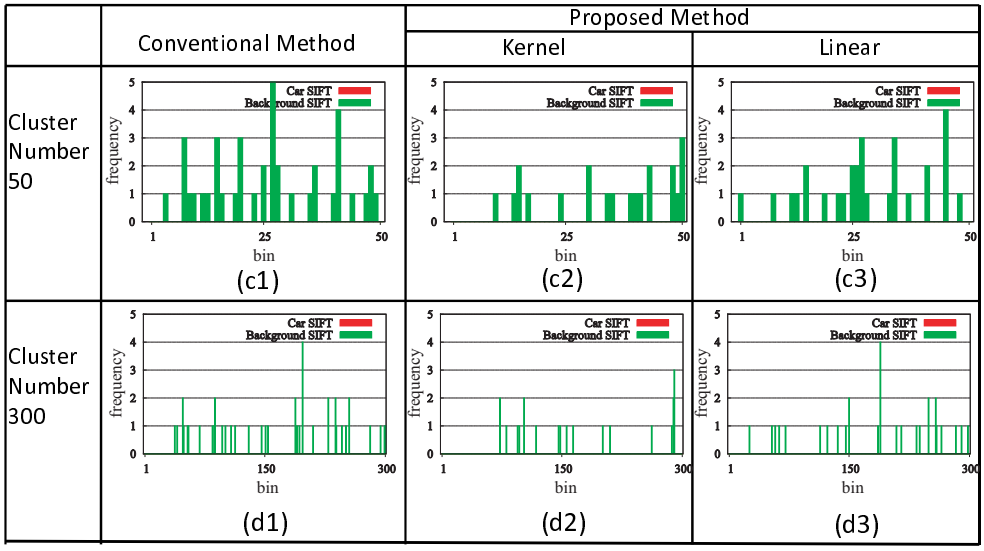


Fig. 7: Example of the histograms of visual words for a non car image. (c1): the conventional method, the number of clusters is 50, (c2): kernel SVM , the number of clusters is 50, (c3): linear SVM, the number of clusters is 50, (d1): the conventional method, the number of clusters is 300, (d2): kernel SVM, the number of clusters is 300, (d3): linear SVM, the number of clusters is 300.