2007 IEEE International Conference on
Robotics and Automation
Roma, Italy, 10-14 April 2007

Visual Categorization Robust to Large Intra-Class Variations using Entropy-guided Codebook

Sungho Kim and In So Kweon
Korea Advanced Institute of Science and Technology
373-1 Guseong-dong Yuseong-gu Daejeon, Korea
{sunghokim, iskweon}@kaist.ac.kr

Chil-Woo Lee
Chonnam National University
300 Yongbong-dong Buk-gu Gwangju, Korea
leecw@chonnam.ac.kr

Abstract— **Categorizing visual elements is fundamentally important for autonomous mobile robots to get intelligence such as new object acquisition and topological place classification. The main problem of visual categorization is how to reduce the large intra-class variations, especially surface markings of man-made objects. In this paper, we present a robust method by introducing intermediate blurring and entropy-guided codebook selection in a bag-of-words framework. Intermediate blurring can filter out the high frequency of surface markings and provide dominant shape information. Entropy of a hypothesized codebook can provide the necessary measure for the semantic parts among training exemplars. From the first step, a generative optimal codebook for each category is learned using the MDL (minimum description length) principle guided by entropy information. From the second step, a final set of codebook is learned using the discriminative method guided by the inter-category entropy of the codebook. We select the necessary parameters through various evaluations and validate the effect of the surface marking reduction method using a Caltech-101 DB, which has large intra-class variations. Finally, we briefly introduce the impact of the method to the object categorization and segmentation problem.**

I. Introduction

Currently, many researchers have tried to develop human-like visual perception capabilities such as self-localization and object recognition for the intelligent mobile robots. Let's imagine that we have bought a new service robot and put it our home. The robot should adapt to the strange environment automatically. It will wander the house and categorize each room as a kitchen, bath room, or living room. Additionally, it will categorize novel objects such as the doors, sofas, dining tables, chairs or refrigerators. As we can see in this scenario, the two basic functions of an intelligent mobile robot are categorizing places and objects for automatic high-level learning about new environments. In the current state-of-the-art, topological localization remains at the level of image identification or matching to the same environment [9], [13]. Object identification (recognition) of the same objects is almost matured due to the introduction of local invariant features such as SIFT and its generalized version, G-RIF [7], [14]. Currently, the categorization of general objects or scenes is an active research area in computer vision society [6], [15].

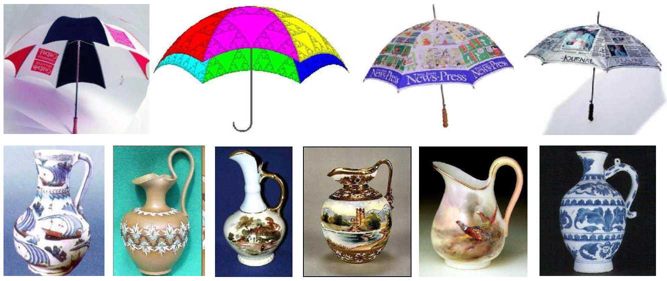


Fig. 1: Various types of surface markings for the man-made objects such as umbrellas and ewers.

However, vision-based categorization of visual elements is a very challenging problem due to large intra-class variations. Among many sources of them, such as geometric shape variations and photometric color variations, surface markings are dominant in man-made objects as shown in Fig. 1. Note the large variations of the surface markings at the interior regions of the objects. The effect of surface marking is much larger in man-made objects than in animals or plants due to creative design for beauty. These markings degrade the generalization capability of any categorization methods.

Until now, most researchers have focused on how to minimize the intra-class variations caused by the object shape. We can categorize the current object representation schemes according to the relation of the geometric strength and intra-class variation. As the strength of a geometric relation is weaker, the amount of intra-class variation is smaller. On the other hand, the discrimination power is reduced due to the weak spatial relation. Since the PCA (principle component analysis) can represent whole objects directly, it is very weak to the geometric variations [12]. The constellation model of visual parts can handle geometric variations more flexibly [6], [16]. Flexible shape samples using geometric blur can represent large variations of shapes [2]. Bag of words, derived from document indexing, is a very robust method to visual variation because it considers no geometrical relations [3]. Texton, which is a more generalized version of bag of words, can categorize textured regions such as forest, sky, and sea [19]. A compromise of both extremes is the implicit shape model, which assigns pose information for each codebook [11].

In this paper, our basic object representation is the bag of words approach to take advantage of its simplicity and

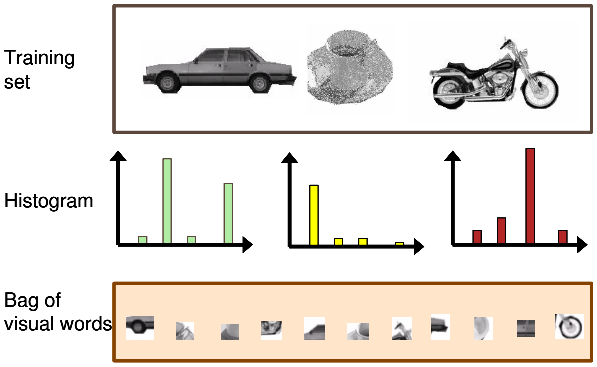


Fig. 2: The bag of visual words-based object representation scheme.

robustness to large geometric shape variations. However, we focus on how to reduce surface markings during visual word or codebook generation. First, we apply intermediate blurring to extract important object shape information. It is motivated from cognitive experiments showing that human visual systems can categorize blurry objects very quickly [1]. This means that low spatial frequency information is important to the visual categorization. Second, we utilize the information theory for the codebook selection. Entropy of a hypothesized codebook among training instances should be high for surface marking reduction, and entropy among different categories should be low for discrimination.

II. Object representation by bag of visual words

The term visual words originated from linguistics [5]. A paragraph consists of a set of words. From the distribution of words, we can determine a topic of the paragraph. Likewise, we can think of a scene or an object as composition of visual words, as shown in Fig. 2. Recently, the bag of visual words approach has shown very promising results on visual categorization problems [3], [4], [15], [19]. Although it is a very simple representation, it can handle large geometric variations because it discards geometric relationships among features or parts. So this approach is not suitable to the detection but suitable to the labeling of novel objects. The basic steps for the bag of visual words approach are visual word (codebook) generation, histogram building, and classifier learning. The critical issue of the visual word-based classification is how to learn the optimal set of visual words, or codebook. Csurka et al. and other researchers selected the optimal set of visual words by the well-known k-means clustering [3]. The size of k is empirically selected by cross validation of the training set. Winn et al. proposed a pair-wise feature clustering method that maximizes inter-class variation and minimizes intra-class variation [19]. Previous approaches do not consider surface marking problems explicitly for the optimal codebook generation.

III. Entropy-based robust object categorization

A. Overview of the proposed categorization method

The visual categorization system is composed of feature extraction, codebook generation, and classification, as shown in Fig. 3 illustrated with rectangle, triangle, and circle examples. First, we extract dense features after intermediate blurring. Then an intra-class codebook is learned using the proposed model selection method of entropy-guided MDL (minimum description length). These intra-class codebooks are further learned in a discriminative way using entropy-guided codebook selection. Then each training instance is represented by a histogram using the optimal set of codebook learning. Finally, classification is conducted using either NNC (nearest neighbor classifier) or SVM (support vector machine) by varying distance metrics. The most important blocks for surface marking reduction are intermediate blurring and entropy-guided codebook selection. Details of the system are explained in the following sub-sections.

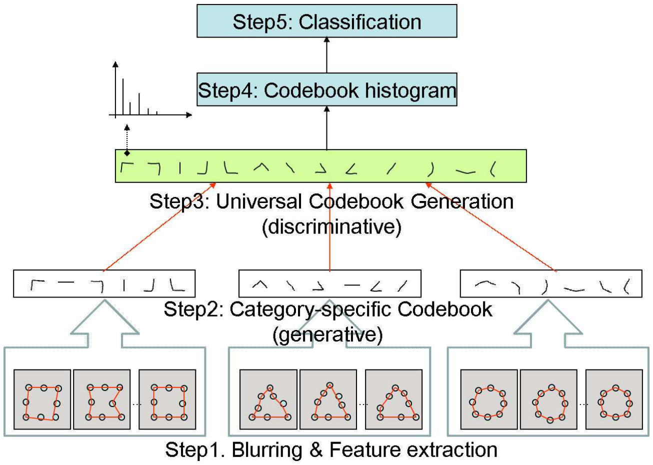

Step5: Classification
Step4: Codebook histogram
Step3: Universal Codebook Generation (discriminative)
Step2: Category-specific Codebook (generative)
Step1. Blurring & Feature extraction

Fig. 3: Composition of visual categorization system robust to surface marking of man-made objects.

B. Step 1: Feature detection by intermediate blurring and scale invariant dense feature

The first issue in the bag of visual words approach is how to extract local features. Direct application of sparse scale invariant features such as SIFT [14] and G-RIF [7] to Caltech-101 DB (available at http://www.vision.caltech.edu/htmlfiles/archive.html) shows very disappointing results: a 26.8% correct classification rate using 15 images for training and 15 images for testing (using Berg's evaluation method [2]). So, we need to find an optimal set of feature parameters, such as the level of blur, and location of sampling points. As a baseline method, we use the k-means clustering (k=650) and NNC classifier with Euclidean distance. We select optimal feature parameters that show best categorization performance with the baseline method. We choose 10 categories such as airplane, car side, cup, helicopter, motorbikes, binocular, camera, cell phone, umbrella, and watch from Caltech-101 DB. Note that those objects have strong surface markings. Randomly selected 15 images per category are used for training and rest 15 images per category are used for testing.

First, we evaluate the effect of blurring by changing the smoothing level in the original G-RIF. Fig. 4 shows the

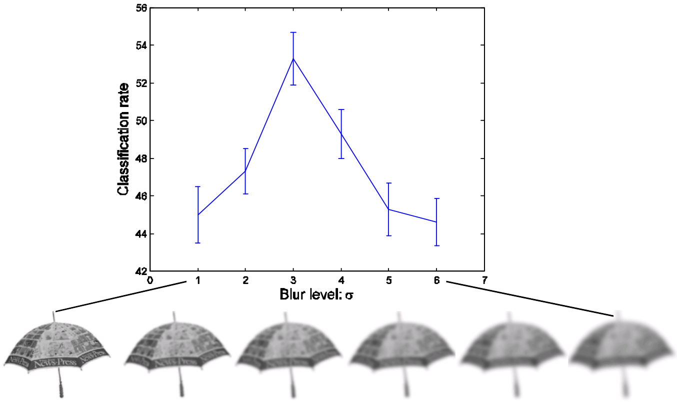


Fig. 4: Evaluation of blurring level in terms of categorization rate.

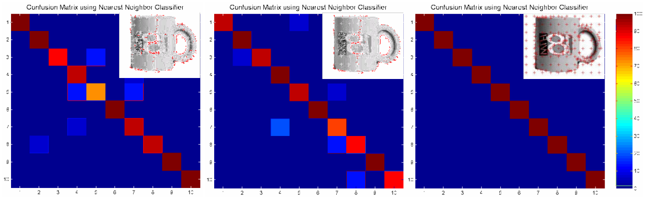


Fig. 5: Evaluation (confusion matrix) of sampling type: (left) edge sampling, (middle) grid sampling, (right) edge-grid sampling. Edge-grid sampling shows better performance.

evaluation results with the corresponding blurred objects. According to the maximum value, we set the blurring level as $\sigma = 3$.

Finally, we also evaluated the edge sample and dense grid sampling types with the selected blurring level. The evaluation was conducted using the same framework. The edge-grid sampling shows upgraded performance as shown in Fig. 5. So, we used edge-grid sampling with the selected feature parameters. Additionally, random samples instead of grid samples show similar performance.

C. Step 2: Intra-class codebook generation with MDL (generative)

In Step 2, we have to minimize intra-class variations. The main cause of large intra-class variation is surface markings, which have various patterns for object instances. As shown in Fig. 6, the surface markings can be removed by finding repeatable parts or high-entropy parts.

Based on this relation of entropy and surface markings, we can conduct model selection using MDL more efficiently. The MDL criteria can provide an optimal codebook in terms of fitting distortion and model complexity, as shown by the following equation [17]. The key point for surface marking reduction is to remove codebook candidates that have low entropy as shown in Fig. 7. An initial codebook is generated using two steps of agglomerative clustering (bottom-up) and k-means clustering (top-down) [8]. The detailed algorithm for intra-class codebook selection is shown in Algorithm 1.

$$\hat{\Lambda}(\mathbf{X},\theta) = \arg\min\left\{-\log L(\mathbf{X},\theta) + \frac{K(V+1)}{2}\log N\right\}$$

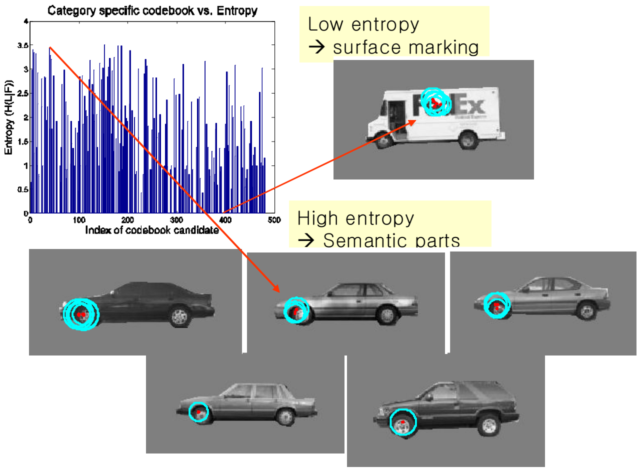


Fig. 6: Observation for repeatable parts (high entropy) and surface marking parts (low entropy).

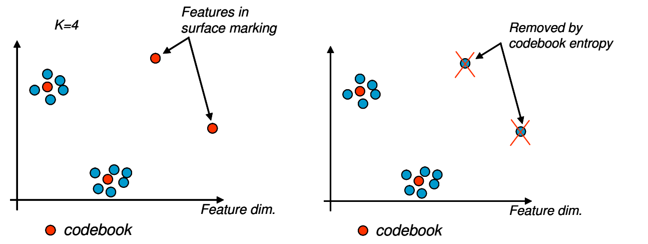
Algorithm 1 Class-specific codebook generation

```
Given: category-specific local features
Goal: make codebook
Step 1. Extract edge-grid features for each
category.
Step 2. Make initial codebook using
appearance-based clustering [8].
Step 3. Starting from this initial codebook.
  Evaluate MDL (Eq. 1)
  If MDL is minimum, stop.
  Else
    Remove one codebook that has lowest
entropy. Go to 1.
```

where L is likelihood of data fitting, $\mathbf{X}$ is training features, θ is parameters (mean and variance for codebook), K is the number of codebook, V is the number of parameters per codebook, and N is the number of features. Fig. 8 shows the MDL model selection curve and the properties of the selected codebook. Note that our codebook can find semantically meaningful parts for the training instances regardless of various surface markings.

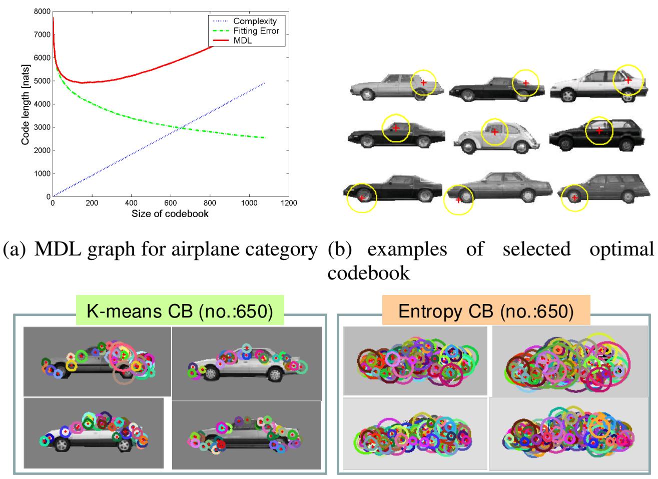
D. Step 3: Inter-class codebook generation (discriminative)

Given the category-specific codebooks learned in Step 1 and 2, we have to select a discriminative universal codebook for bag of visual words-based classification. We can obtain a discriminative codebook (F_{opt}) by maximizing the posterior of class labels given training examples and a hypothesized universal codebook. The key point in this approach is to select a removable codebook using the inter-category entropy of a codebook that has large entropy (ambiguous codebook). If we define $\{F\}$ as a hypothesized universal codebook, I_i^c as the i-th object instance belonging to category c, and l as the category label, then the posterior can be formulated as



(a) concept of k-means clustering (b) concept of entropy-guided codebook

Fig. 7: The mechanism of surface marking removal in entropy guided codebook compared to the conventional k-means clustering.



(a) MDL graph for airplane category (b) examples of selected optimal codebook

(c) Codebook (CB) label: k-means vs. proposed

Fig. 8: Entropy-guided MDL graph and its example parts corresponding to selected codebook. Note that similar parts are selected regardless of surface markings.

$$F_{opt} = \arg\max_{F}\{\prod_{c}\prod_{i\in c} p(l|I_i^c, \{F\}\}$$

$$= \arg\max_{F}\{log(\prod_{c}\prod_{i\in c} p(l|I_i^c, \{F\}))\}$$

$$since$$

$$p(l|I_i^c) = \frac{p(I_i^c|c,\{F\})p(c,\{F\})}{\sum_{c'} p(I_i^c|c',\{F\})p(c',\{F\})},$$

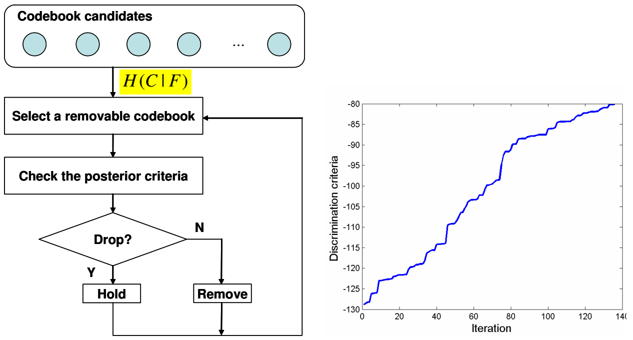
$$assume\ uniform\ p(c,\{F\})$$

$$F_{opt} = \arg\max_{F}\{\sum_{c}\sum_{i\in c}(\log p(I_i^c|c,\{F\}) - \log\sum_{c'} p(I_i^c|c',\{F\}))\}$$

$$where\ p(I_i^c|c,\{F\}) = p(H_i^c|H_M^c) \propto \exp\left(-KL(H_i^c, H_M^c)\right)$$

$$and\ KL(H_i^c, H_M^c) = \sum_{j=1}^{|F|}(H_i^c(j) - H_M^c(j))\log\frac{H_i^c(j)}{H_M^c(j)})$$

The posterior criterion in the 4th line of the equation is used to select the optimal set for a discriminative codebook. The distance measure (KL) can be anything introduced in the following sub-section. Fig. 9 shows the codebook search algorithm and its learning graph. First, we sort candidate codebooks based on the inter-category entropy. Candidate codebooks are just the collection of intra-class codebooks. Second, a high entropy codebook is removed from the candidate codebooks. Then we check the posterior using the hypothesized codebooks. If the posterior is dropped, the removable codebook is hold, otherwise it is removed. Such iteration continues until the posterior is converged. Fig. 10 shows the test results using only a set of the intra-class codebook ($|F| = 1062$) and the discriminatively learned universal codebook ($|F| = 926$) for 10 object categories. Note the upgraded categorization performance.



(a) optimization procedure (b) discriminative learning graph

Fig. 9: Inter-category entropy-guided universal codebook selection method.

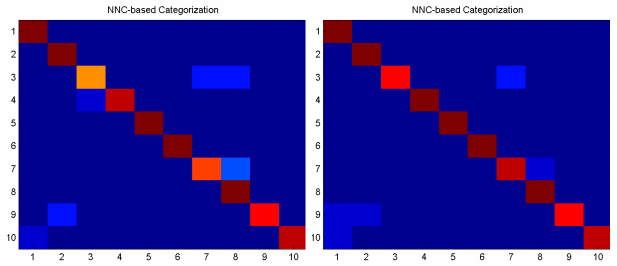


Fig. 10: Confusion matrix using non-discriminative codebook and discriminative codebook. (left) Before discriminative learning, (right) after discriminative learning.

E. Step 5: Distance metrics and classification

After histogram building from the discriminative codebook for all the training instances, we have to learn classifiers with certain distance metrics. We can summarize these as follows.

Distance metrics:$D(H_t, H_m)$

- Euclidean dist.: $D(H_t, H_m) = \sum_i (H_t(i) - H_m(i))^2$
- KL-divergence:
$D(H_t, H_m) = \sum_i (H_t(i) - H_m(i)) \cdot \log(H_t(i)/H_m(i))$
- Intersection:$D(H_t, H_m) = \sum_i \min(H_t(i), H_m(i))$
- χ^2 distance:$D(H_t, H_m) = \sum_i \frac{(H_t(i) - H_m(i))^2}{H_t(i) + H_m(i)}$

Classification

- NNC is the simplest classifier because it requires no specific learning method. Each training histogram is regarded as a single prototype. So, for an unknown test histogram, a category label is assigned with the nearest prototype in the model database.

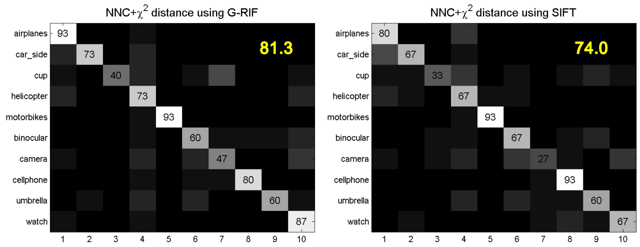


Fig. 11: Comparison of G-RIF and SIFT for 10 Caltech DB.

- Support vector machine (SVM) [18] can learn classification boundaries from training samples. It has been the most powerful classifier until now. Recently, a kernel-based SVM was introduced that can learn non-linear classification boundaries for complex data. We use the OvR (one vs. Rest) method for the multi class SVM.

In the extended Gaussian kernel, we can use the distance metrics described above. In the experiment section, we will compare these classification methods using codebooks that are robust to surface markings and discriminative.

IV. Performance evaluation

We evaluated our categorization system using a Caltech-101 DB. It consists of 48 man-made objects and 53 animals and plants. As an initial experiments, we selected 10 categories such as airplanes, cameras, cars, cell phones, cups, helicopters, motorbikes, scissors, umbrellas, and watch which have large intra-class variations due to surface markings. We randomly selected 15 examples for each category and tested 15 unlearned cluttered examples.

G-RIF vs. SIFT: As a starting point to multi-scale feature selection, first we compare our G-RIF to SIFT for the 10 category classification problem. For fair comparison, we make codebook using the conventional k-means clustering with k=650. We use NNC classifier with χ^2 distance. 15 images per category are used for training, rest 15 images are used to test. Fig. 11 shows the confusion matrices. The average categorization rate of SIFT is 74% and that of G-RIF is 81.3%. Since G-RIF is generalized version of SIFT in terms of information quantity, bag of visual words using G-RIF shows better performance than the same method using SIFT.

Entropy-guided codebook vs. k-means clustering: Next, we check the effect of codebook selection (our entropy-guided codebook, k-means clustering) to the categorization performance. We use the same G-RIF feature, number of codebook, and classifier except different codebook generation method. As shown in Fig. 12, bag of visual words using entropy-guided codebook shows better performance the the same method using k-means clustering. This fact means that our codebook can handle surface markings more robustly.

Optimal distance measure and classifier: Until now, we select G-RIF and entropy-guided codebook. Then what is optimal distance and classifier for the codebook histogram? We compared 4 distance measures of Euclidean, KL-divergence, histogram intersection, and χ^2 distance. As classifiers, we compare the well-known NNC and SVM especially OvR (one versus rest) for multi-category classification. Fig. 13 shows the confusion matrices for the combinations of distances and classifiers. Table I summarizes the overall performance. Note that the histogram intersection or χ^2 distance with SVM is the optimal choice for bag of visual words method.

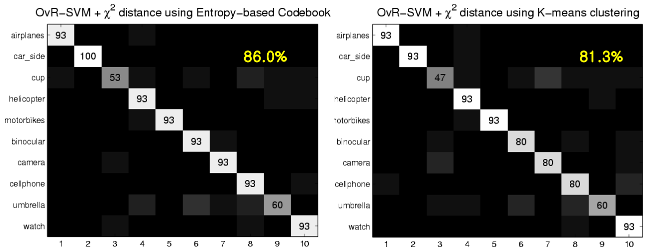


Fig. 12: Comparison of our entropy-guided codebook and k-means clustering.

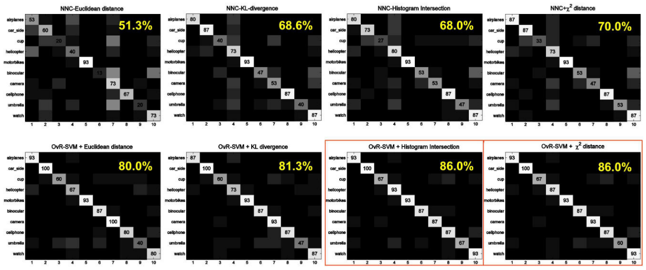


Fig. 13: Comparison of distance measures and classifiers.

Sparse sampling vs dense sampling in multi-scale: We have determined the optimal codebook selection method and classifier for bag of visual words-based categorization. In previous section, we uses dense sampling with fixed scale. Now, we compare the sparse sampling and dense sampling in multi-scale space. We set all the related parameters to the same value except sampling method in scale space. As shown in Fig. 14, the dense sampling in scale-space shows much better performance (87.3%).

Based on this finding, we extended the experiment to the whole database. We selected the SVM classifier with χ^2 distance. The DC (discriminative codebook) was learned from each category-specific GC (generative codebook). The average classification of our system was 48.58% for a cluttered test set. The current state-of-the-art for the same database using the conventional bag of visual words (single level, L=0, 15 training) shows 41.2% [10]. Most incorrect classifications are for animals and plants which have large

TABLE I: Categorization rate [%] according to the combinations of distance measures and classifiers

dist. measure	Euclidean dist.	KL-div.	hist. inters.	χ^2
NNC	51.3	68.6	68.0	70.0
SVM	80.0	81.3	86.0	86.0

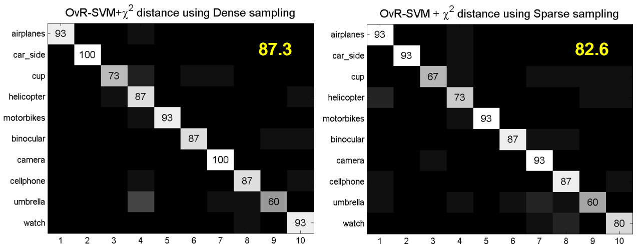


Fig. 14: Comparison of dense sampling and sparse sampling.

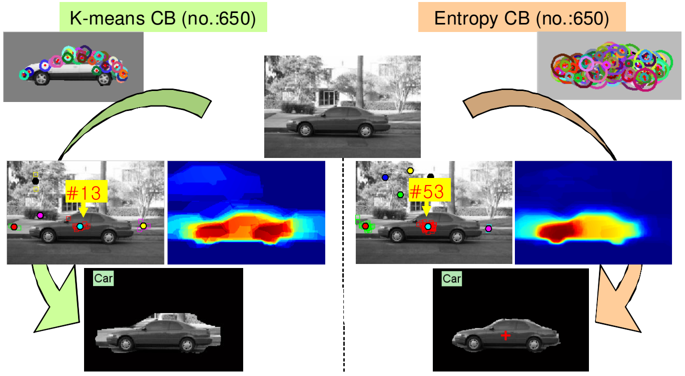


Fig. 15: Impact of the proposed codebook selection method to the object categorization and segmentation problem.

intra-class variations due to different shape.

V. Impact to the Categorization and Segmentation

The surface marking reduction strategy using the intermediate blurring and entropy-guided codebook is effective to the categorization of man-made objects. Furthermore, the codebook selection mechanism is useful to the simultaneous categorization and segmentation problem of man-made objects in cluttered environment. Our on-going research is to utilize the part-whole relation for that problem. As shown in Fig. 15, the entropy-guided codebook (CB) is useful to estimate the part-whole relation. Contrary to the k-means codebook, our entropy-based codebook shows more robust clustering of object center, more clean figure-ground segregation in cluttered environment.

VI. Conclusion

In this paper, we presented an object categorization method focusing on surface markings in the bag of visual words framework. We can minimize the effect of surface markings based on the entropy of the codebooks. High entropy in the intra-class codebook can remove surface marking parts (low entropy) in stage 1 learning. Additionally, a discriminative codebook is also selected from the category-specific codebook guided by the entropy of the inter-class codebook. The high entropy codebook is removed first because it gives ambiguous class labels. Finally, we evaluated those codebooks using NNC and SVM classifiers with different distance metrics. With the optimal set of features, codebooks, and classifiers, we can get upgraded performance in the bag of visual words framework. This work for codebook selection and classification can be applied to other complex categorization and segmentation problems.

VII. Acknowledgments

This work was supported in part by the Korea Science and Engineering Foundation (KOSEF) grant funded by the Korea government(MOST)(No. M1-0302-00-0064), by the MIC for the project, "Development of Cooperative Network-based Humanoids' Technology" of Korea.

References

[1] M. Bar. Visual objects in context. *Nature Reviews: Neuroscience*, 5:617–629, August 2004.

[2] A.C. Berg, T.L. Berg, and J. Malik. Shape matching and object recognition using low distortion correspondences. In *IEEE Computer Vision and Pattern Recognition or CVPR*, pages 26–33, 2005.

[3] G. Csurka, C.R. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *ECCV Workshop on Statistical Learning in Computer Vision*, 2004.

[4] G. Dorkó and C. Schmid. Selection of scale-invariant parts for object class recognition. In *International Conference on Computer Vision*, pages 634–640, 2003.

[5] S. Dumais, J. Platt, D. Heckerman, and M. Sahami. Inductive learning algorithms and representations for text categorization. In *Proceedings of the seventh international conference on Information and knowledge management (CIKM'98)*, pages 148–155, 1998.

[6] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *IEEE Computer Vision and Pattern Recognition or CVPR*, pages 264–271, 2003.

[7] S. Kim and I.-S. Kweon. Biologically motivated perceptual feature: Generalized robust invariant feature. In *Asian Conference on Computer Vision*, pages 305–314, 2006.

[8] S. Kim and I.-S Kweon. Simultaneous classification and visualword selection using entropy-based minimum description length. In *International Conference on Pattern Recognition*, pages 650–653, 2006.

[9] J. Kosecka and F. Li. Vision based topological markov localization. In *Proc. of IEEE International Conference on Robotics and Automation (ICRA)*, 2005.

[10] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *IEEE Computer Vision and Pattern Recognition or CVPR*, pages 2169–2178, 2006.

[11] B. Leibe, A. Leonardis, and B. Schiele. Combined object categorization and segmentation with an implicit shape model. In *Workshop on Stat. Learn. in Comp. Vis.*, 2004.

[12] B. Leibe and B. Schiele. Analyzing appearance and contour based methods for object categorization. In *IEEE Computer Vision and Pattern Recognition or CVPR*, volume 2, pages 409–415, 2003.

[13] Z. Lin, S. Kim, and I.S. Kweon. Recognition-based indoor topological navigation using robust invariant features. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'05)*, 2005.

[14] D.G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[15] K. Mikolajczyk, B. Leibe, and B. Schiele. Multiple object class detection with a generative model. In *IEEE Computer Vision and Pattern Recognition or CVPR*, pages 26–36, 2006.

[16] P. Moreels, M. Maire, and P. Perona. Recognition by probabilistic hypothesis construction. In *European Conference on Computer Vision*, pages 55–68, 2004.

[17] A. Vailaya, M. A. T. Figueiredo, A. K. Jain, and H.-J. Zhang. Image classification for context-based indexing. *IEEE Trans. Image Processing*, 10(1):117–130, 2001.

[18] V.N. Vapnik. *The nature of statistical learning theory*. Springer-Verlag New York, Inc., New York, NY, USA, 1995.

[19] J. Winn, A. Criminisi, and T. Minka. Object categorization by learned universal visual dictionary. In *International Conference on Computer Vision*, pages 1800–1807, 2005.