Generic Object Recognition using CRF by Incorporating BoF as Global Features

Takeshi Okumura*, Tetsuya Takiguchi**, Yasuo Ariki**

* *Graduate School of Engineering, Kobe University, Japan*
** *Organization of Advanced Science and Technology, Kobe University, Japan*
E-mail: okumura@me.cs.scitec.kobe-u.ac.jp, takigu@kobe-u.ac.jp, ariki@kobe-u.ac.jp

Abstract

Generic object recognition by a computer is strongly required in various fields like robot vision and image retrieval in recent years. Conventional methods use Conditional Random Field (CRF) that recognizes the class of each region using the features extracted from the local regions and the class co-occurrence between the adjoining regions. However, there is a problem that CRF tends to fall into the local optimal recognition result because it uses only local features and the relation. To solve this problem, we propose a method that recognizes generic objects by incorporating Bag of Features (BoF) as the global feature into CRF. As a result of the experiment to the image dataset of 21 classes, the proposal method has improved the recognition rate by 6.5%.

1. Introduction

Generic object recognition means that a computer recognizes objects in images of a real world as the general name. This is one of the most challenging task in the computer vision. However, from the viewpoint of realizing the human vision by the computer, it is expected to be applied to the robot vision. Moreover, due to the popularization of digital cameras and the development of high-capacity hard disk drives in the recent years, it is getting difficult to classify and to retrieve enormous videos and images manually. Then, the computer is required to automatically classify and to retrieve videos and images. Especially the generic object recognition become more and more important.

There are two kinds of conventional approaches of the generic object recognition. One is the approach of recognizing the class of the image. This approach often uses Bag of Features (BoF) [1] that characterizes an image by a set of local features. This global feature is used in Support Vector Machine and probabilistic Latent Analysis, and the class of the image is recognized.

The other is the approach of recognizing the class of each pixel in the image. Low-level features, such as the color feature and texture feature, are extracted from the local region, and the class of the local region is recognized based on the features. One method of this approach [2] uses Gaussian Mixture Model, but since this method recognizes the class of the local regions independently, it is difficult to recognize the class of the regions from which the only ambiguous features are extracted. Also, there is a problem that the recognition result tends to become inconsistent as a whole.

To enable more consistent recognition, based on the idea that the relation of the co-occurrence exits among the objects in the image, the methods [4] that use Conditional Random Field [3], which is a graphical model, attract increasing attention. These methods recognize the class of each local region based on not only the features of the region but also the class co-occurrence between adjacent regions. The recognition result for the regions, from which only the ambiguous features are extracted, can be improved by considering the relation to adjacent regions. The class co-occurrence is a kind of contextual information. For example, class “cow” and class “grass” tend to coexist, but class “cow” and class “car” don't tend to coexist.

The former approach implicitly assumes that only one object exits in an image. This approach is not suitable for the general images where multi-class objects coexist. Therefore, we adopt the latter approach and use CRF for the recognition.

But many conventional methods have a common problem that the recognition result tends to fall in the local optimum. We think this is because the recognition is based on only the local features and relation. To solve this problem, we propose a method to incorporate BoF, which is robust to appearance change and occlusion, into CRF as the global feature.

This paper is organized as fellows. In Section 2, the proposed method is described that uses CRF incorporating BoF as global features. In Section 3, the performance of the proposed method is evaluated for 21 class image dataset. Section 4 is for paper summarization and discuss about the future work.

2. Proposed Method

The flow of the proposed method is shown in **Figures 1**.

ISSN 1975-4736 ©MITA 2009

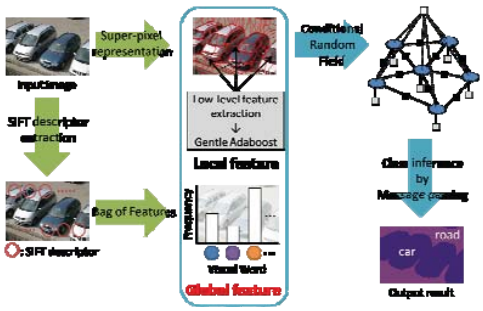

Figure 1: The flow of the proposed method

First, an input image is segmented into small regions called super-pixel by the super-pixel representation using Normalized Cuts [5]. The low-level features, such as the color and texture feature, are extracted from each super-pixel. Based on them, Gentle Adaboost computes the scores of all the classes to each super-pixel. These scores are local features used in Conditional Random Field [3].

Also, SIFT descriptors [6] are extracted from an input image by Grid Sampling, and they are vector-quantized by Bag of Features (BoF) [1]. This is the global feature used in CRF.

In the graph structure of CRF, each super-pixels are represented as an node, and all adjacent nodes are connected by an edge. Based on both the local feature and the global feature, the class distribution of each node is computed by CRF. Finally, after the final class distribution of each node is recomputed by message passing, taking class co-occurrence into consideration, the recognition of the class at each pixel is achieved.

Next, we describe each method that is used in our proposed method.

2.1 Image Segmentation and Local Feature

For the image segmentation, we use the super-pixel representation by using Normalized Cuts [5]. Normalized Cuts is one of the spectral clustering methods. The evaluation function is defined as the similarity of samples in the same cluster is high and the similarity of samples between the different clusters is low. By solving the minimization problem of this function as the eigenvalue problem, the cluster division is performed. The super-pixel representation means that the images are over-segmented by changing the parameter of Normalized Cuts. These regions are called super-pixels, and have the homogeneity of color and texture. Also, these are less redundant than pixels.

We extract the following low-level features from each super-pixel.

- Components of RGB, HSV, Lab, and YCbCr
- Filter response of Gabor filter and LoG
- Coordinates of centroid of super-pixel
- Area of super-pixel

For color and texture features, after feature extraction from each pixel, the statistics such as average, standard deviation, skewness and kurtosis are computed in each super-pixel.

These four low-level features characterize each super-pixel, and based on them, the scores for all the classes are computed by Gentle Adaboost. Gentle Adaboost is derived from Adaboost, a kind of Boosting that determines the output by the weighted voting of a lot of weak classifiers. Since Gentle Adaboost is the binary discriminant classifier, we prepare multiple classifiers whose number corresponds to the number of classes to be recognized. Sum of each classifier's score is normalized to 1 by using softmax function. This is the local feature for recognition.

2.2 Global Feature

We characterize the image by using Bag of Features (BoF) [1]. BoF originally uses SIFT (Scale-Invariant Feature Transform) feature [6]. The invariance for scale is obtained in the auto detection of the feature points and their scale by Difference-of-Gaussian (DoG).

However, in this detection method, since the feature points are detected at the large change of luminance gradient, there is a problem of causing bias in the characterization of the image. To solve this problem, we adopt Grid Sampling with multi-scale for detection. That is, the feature points are detected at regular intervals, and have multi-scale. These intervals and scales are determined empirically. The bias of characterization can be prevented by this detection method.

After the detection of the feature points, the representative orientation of each feature point is computed, and based on it, the luminance gradient histogram is obtained. Since it is always computed as the basis for the representative orientation, rotation invariance is achieved. This is called SIFT descriptor, and is the feature with 128 dimensions.

Moreover, since SIFT descriptor is extracted in gray scale space, color information is not taken into consideration. To incorporate color information into SIFT descriptor, after we convert the color space of the image into HSV color space, SIFT descriptor with color information is obtained by applying the same process to hue, saturation and value. This is the feature with 384 dimensions, and we call the former Gray SIFT descriptor, and the latter Color SIFT descriptor.

Next, as shown in **Figures 2**, after SIFT descriptors are extracted from all the training images, they are divided into W clusters by k-means. The centroid vector of each

cluster is called Visual Word, and the number of words W is determined empirically. In this way, the image is represented by the histogram of Visual Word frequency. Sum of all bins is normalized to 1.

The characterization of the image by BoF is robust to occlusion because it is expressed as aggregation of local feature. Also, it is robust to the change of appearance because of rotation invariance of SIFT descriptor.

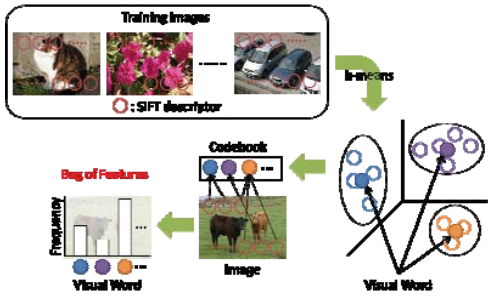

Figure 2: Bag of Features

2.3 Recognition by Conditional Random Field

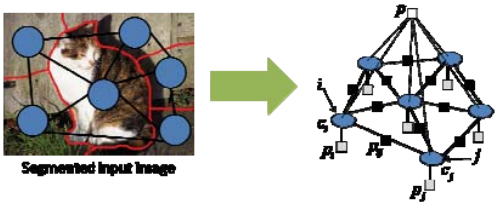

Figure 3: Graph representation of the image by CRF

Conditional Random Field (CRF) [3] is the graphical and discriminative model proposed in the domain of linguistic processing originally. This is used for estimating the class of the data with structure based on the observed feature. When this model is applied to the image, each super-pixel is represented as a node, and all adjacent nodes are connected by an edge. Therefore, the image is represented as the graph structure of CRF as shown in **Figures 3**.

Let $i \in S$ denote each super-pixel in an image S, N_i be the set of the super-pixels adjacent to the super-pixel i, $\mathbf{L} = \{\mathbf{l}_i\}_{i \in S}$ describe local feature (the score of Gentle Adaboost) extracted from each super-pixel, $\mathbf{g}$ represent global feature (Bag of Features) extracted from the image, and $\mathbf{c} = \{c_i\}_{i \in S}$ show the estimated class in each super-pixel. Then, the model formula of CRF is written as the following conditional distribution $P(\mathbf{c} \mid \mathbf{L}, \mathbf{g}; \boldsymbol{\theta})$. Its graph representation is shown in **Figures 3**.

$$P(\mathbf{c} \mid \mathbf{L}, \mathbf{g}; \boldsymbol{\theta}) = \frac{1}{Z} \exp\left[\sum_{i \in S} \{ p_i(c_i \mid \mathbf{l}_i; \boldsymbol{\alpha}) + p(c_i \mid \mathbf{g}; \boldsymbol{\beta}) \} + \sum_{i \in S} \sum_{j \in N_i} p_{ij}(c_i, c_j; \boldsymbol{\gamma}) \right] \quad (1)$$

where C is the number of the classes to be recognized, W is the number of Visual Word, Z is called partition function for regularization. $\boldsymbol{\theta} = \{\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}\}$ is the model parameter of CRF, and we decide them based on the following Maximum A Posteriori estimation by using all the training images with ground truth.

$$\theta^* = \arg\max_{\boldsymbol{\theta}} \left\{ \sum_{t=1}^{T} \log P(c^t \mid \mathbf{L}^t, \mathbf{g}^t; \boldsymbol{\theta}) - \frac{R}{2} \|\boldsymbol{\theta}\|^2 \right\} \quad (2)$$

where T is the number of the training images, R is the parameter for preventing over-fitting. $\boldsymbol{\theta}^*$ is computed analytically by L-BFGS method. However, since partition function Z is intractable because of the loop structure of the graph, it is approximated by PseudoLikelihood.

$p_i(c_i \mid \mathbf{l}_i; \boldsymbol{\alpha})$ and $p(c_i \mid \mathbf{g}; \boldsymbol{\beta})$ are the class distribution based on local feature and global feature respectively in each node. $p_{ij}(c_i, c_j; \boldsymbol{\gamma})$ is the class co-occurrence between the adjacent nodes. We define the class distribution $p(c_i \mid \mathbf{g}; \boldsymbol{\beta})$, derived from the global feature, as common to all the nodes in an image since we intend to obtain the class of each node in an image globally, and this class distribution works like a bias to each image. This formulation can prevent the recognition result from falling into the local optimum.

For the final class inference, we need to find the class of each node that maximizes the conditional distribution shown in Eq. (1). Since there is $p_{ij}(c_i, c_j; \boldsymbol{\gamma})$, we have to consider the class co-occurrence between the adjacent nodes in inference.

For the purpose, we use Maximizer of Posterior Marginal estimation.

$$c_i^* = \arg\max_{c_i} \sum_{\mathbf{c} \setminus c_i} P(\mathbf{c} \mid \mathbf{L}, \mathbf{g}; \boldsymbol{\theta}) \quad (3)$$

where c_i^* is the class maximizing the posterior marginal distribution. Since strict estimation is very difficult because of the loop structure, we approximately estimate it by loopy max-product algorithm that is a kind of loopy belief propagation.

3. Experimental Evaluation

3.1 Overview of Experiment

ISSN 1975-4736 ©MITA 2009

We used Microsoft Research Cambridge 21 Dataset for experiment. It includes 591 images with 21 classes. Each image is assigned ground truth at the pixel level, but black regions means "void", and these were not used for both learning and test. The size of the images is almost 320 × 240 pixels.

Recognition rate was computed as the class average accuracy. We randomly split the dataset into two subsets, and the half was used for learning, and the other was used for test. Also, we set the number of super-pixels in an image to about 200, the interval of Grid Sampling to 10 pixels, the scale of SIFT descriptor to {4, 8, 12, 16}, the number of Visual Word to {100, 200, ⋯, 900, 1000}.

We investigated the change of accuracy with or without the proposed method. In addition, as shown in **Table 1**, we carried out experiments on three different patterns of BoF.

3.2 Result and Discussion

Table 1: Recognition result

	Accuracy[%]
Grid Sampling + Color SIFT	64.6 (300word)
Grid Sampling + Gray SIFT	65.5 (500word)
DoG + Gray SIFT	62.3 (600word)
No Global Feature	59.0

From **Table 1**, we can confirm that the proposed method improves the accuracy by 6.5%, and Grid Sampling is more effective than DoG. The accuracy is high when the color information is not used, but we think this is because the variance of color in classes is very large, and characterization is difficult.

Some examples of the recognition result are shown in **Figures 4**.

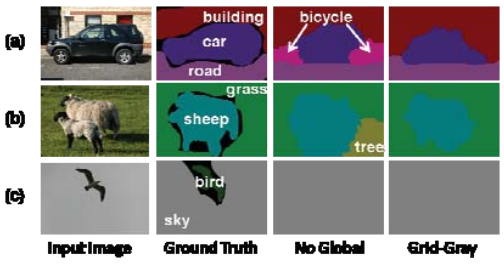

Figure 4: Examples of recognition result

(a) and (b) show the effectiveness of the proposed method. In the case (c), the proposed method cannot improve the accuracy since the object is very small, and feature extraction is very hard. This problem will be resolved by incorporating the concept of the object detectors.

In the future work, it is necessary to find more useful context information except the class co-occurrence since it is difficult to recognize generic objects only by low-level features due to high intra-class variance.

4. Conclusion

In this paper, we proposed the new method to recognize generic objects in the CRF framework by incorporating the local feature, global feature and the class co-occurrence. The local feature is the score of Gentle Adaboost based on the low-level features extracted from super-pixel. The global feature is Bag of Features (BoF) based on SIFT descriptor extracted by Grid Sampling. Recognition result is successfully prevented from falling into local optimum by the proposed method owing to the global feature BoF. As a result, recognition accuracy is improved by 6.5% compared to the method without using the global feature. In the future work, we will find more useful context information and more robust feature.

5. References

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

[2] K. Barnard, and D. Forsyth, "Learning the Semantics of Words and Pictures," *Proc. IEEE International Conference on Computer Vision, pp.408--415, 2001*

[3] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," *Proc. International Conference on Machine Learning, 2001*

[4] X. He, R. S. Zemel, and M. A. Carreira-Perpina, "Multiscale conditional random fields for image labeling," *Proc. IEEE Computer Vision and Pattern Recognition, pp.695-702, 2004*

[5] J. Shi, J. Malik, "Normalized cuts and image segmentation," *Proc. IEEE Computer Vision and Pattern Recognition, pp731-737, 1997*

[6] D. G. Lowe, "Object recognition from local scale-invariant features," *Proc. IEEE International Conference on Computer Vision, pp.1150-1157, 1999*