Generic Object Recognition by Tree Conditional Random Field based on Hierarchical Segmentation

Takeshi OKUMURA† Tetsuya TAKIGUCHI‡ Yasuo ARIKI‡
† *Graduate School of Engineering, Kobe University*
‡ *Organization of Advanced Science and Technology, Kobe University*
okumura@me.cs.scitec.kobe-u.ac.jp {*ariki, takigu*}*@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 the discriminative ability of the features extracted from local regions is insufficient, and these methods is not robust to the scale variance. To solve this problem, we propose a method that integrates the recognition results in multi-scales by tree conditional random field based on hierarchical segmentation. As a result of the image dataset of 7 classes, the proposed method has improved the recognition rate by 2.2%.

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 (SVM) and probabilistic Latent Analysis (pLSA), 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 (GMM), 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. Furthermore, 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 [3][4] that use Conditional Random Field (CRF) [5], 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

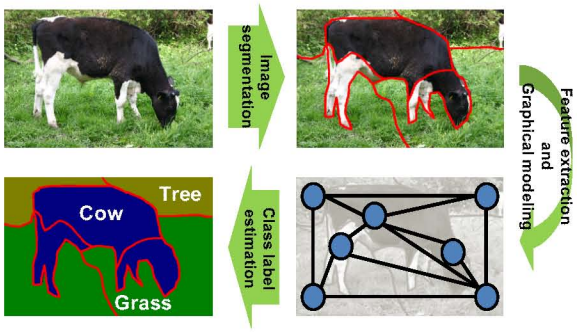

Figure 1. Recognition of the class of each pixel in the image by CRF

1051-4651/10 $26.00 © 2010 IEEE
DOI 10.1109/ICPR.2010.741

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 (see Fig.1). 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.

But many conventional methods have a problem that the features extracted from the local regions are not discriminative, and these are not robust to the scale variance in objects. We think this is because the scale of the image segmentation as a pre-processing is a single scale. To solve this problem, we propose a method that integrates the recognition results in multi-scales by tree conditional random field based on hierarchical segmentation.

This paper is organized as follows. In Section 2, the proposed method is described. In Section 3, the performance of the proposed method is evaluated for 7 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 Fig.2. First, an input image is applied to Segmentation by

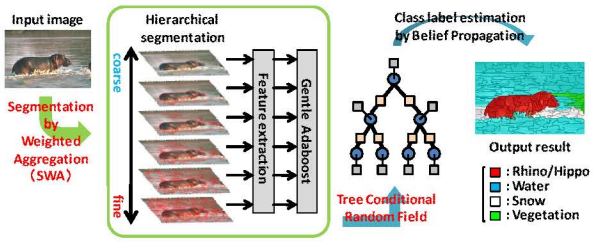

Figure 2. The flow of the proposed method

Weighted Aggregation (SWA) [6]. SWA is a hierarchical segmentation method that the lower the layer is, the more finely the image is segmented, and the higher the layer is, the more coarsely the image is segmented. The image in the top layer is not segmented. A segment in a layer corresponds to multi-segments in the lower layer.

Then, the low-level features, such as the color feature and texture feature, are extracted from each segment in each layer except for the highest layer. Only in the highest layer, Bag of Features (BoF) [1], that is suitable for characterizing the whole image, is extracted. Based on them, Gentle Adaboost computes all the class reliability to all segments. These are local features used in the proposed method.

Finally, according to the relation between segments in hierarchical segmentation, Tree Conditional Random Field (TCRF) is constructed. The sum of the class reliability of each segments and class co-occurrence between segments is defined as energy function, and the class assignment maximizing this is estimated by Belief Propagation (BP) [7].

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

2.1 Hierarchical Segmentation by SWA

For hierarchical segmentation, we use Segmentation by Weighted Aggregation (SWA) [6]. The image is regarded as a weighted graph. Nodes correspond to pixels, and edges correspond to connecting neighboring nodes. The evaluation function $\Gamma(\mathbf{u})$ is defined as

$$\Gamma(\mathbf{u}) = \frac{\sum_{i>j} w_{ij}(u_i - u_j)^2}{\sum_{i>j} w_{ij} u_i u_j} = \frac{\mathbf{u}^T L \mathbf{u}}{\frac{1}{2}\mathbf{u}^T W \mathbf{u}} \quad (1)$$

where $\mathbf{u} = \{u_1, u_2, \cdots, u_n\}$ is a state vector, with a state variable $u_i = 1$ if a pixel i belongs to a segment S and $u_i = 0$ if a pixel i belongs to a segment S. Also, L is the laplacian matrix of the graph, W is the weight matrix, the numerator of Eq.1 denotes the cutting cost function, and the denominator denotes the size of a segment S. The purpose of this function is that the internal weight of a segment becomes low and the size of all segments becomes as uniform as possible.

For the optimal segmentation, the minimization problem of this function is solved as the eigen problem $L\mathbf{u} = \lambda W \mathbf{u}$ with minimal positive eigenvalue λ. Algebraic MultiGrid (AMG) procedure solves this problem by the approximate recursive coarsening $\mathbf{u} \approx P\mathbf{U}$ with the sparse interpolation matrix P and the coarsening state vector $\mathbf{U}$. Thus, the image is represented by the pyramid structure. The proposed method uses each layer of this hierarchy and the relation between the layers .

2.2 Features

After hierarchical segmentation, we extract the following low-level features from each segment in each layer except for the top layer.

- 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 segment.

Also, only in the top layer, we extract Bag of Features (BoF) [1] from the image. BoF is the appearance-based method that the local features such as SIFT (Scale-Invariant Feature Transform) [8] is extracted from the image, and they are divided into W clusters by k-means. The centroid vector of each cluster is called Visual Words, and the number of words W is determined empirically. In this way, the image is represented by the histogram of Visual Words frequency. The characterization of the image by BoF is robust to occlusion because it is expressed as aggregation of local features, and it is robust to the change of appearance because of vector quantization by k-means.

These features characterize each segment in each layer, and based on them, the class reliability for all 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 and features should be trained in each layer, we prepare multiple classifiers whose number corresponds to the number of classes to be recognized times the number of layers.

2.3 Recognition by TCRF

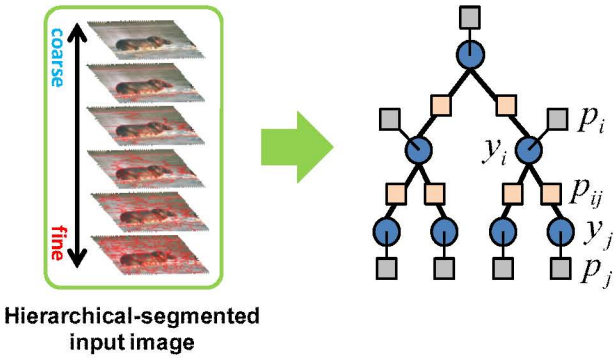

Figure 3. Graph representation of the image by TCRF

Conditional Random Field (CRF) [5] 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 hierarchical-segmented image, each segment is represented as a node, and all segments that have relation between layers are connected by an edge. Therefore, the image is represented as the graph structure as shown in Fig.3, and we call this model Tree Conditional Random Field (TCRF) since the graph structure is tree structure that has no loop structure.

Let $i \in N$ denote each segment in a hierarchical-segmented image, $\pi(i)$ be the set of the child nodes for the parent node i, $\mathbf{X} = \{\mathbf{x}_i\}_{i \in N}$ describe the class reliability in each segment by Gentle Adaboost, and $\mathbf{y} = \{y_i\}_{i \in N}$ show the estimated class in each node. Then, the model formula of TCRF is written as the following conditional distribution $P(\mathbf{y}|\mathbf{X};\boldsymbol{\theta})$.

$$P(\mathbf{y}|\mathbf{X};\boldsymbol{\theta}) = \frac{1}{Z}\exp\left\{\sum_{i \in N} p_i(y_i|\mathbf{x}_i;\boldsymbol{\alpha}) + \sum_{i \in N}\sum_{j \in \pi(i)} p_{ij}(y_i, y_j;\boldsymbol{\beta})\right\} \quad (2)$$

where Z is called partition for regularization. $\boldsymbol{\theta} = \{\boldsymbol{\alpha}, \boldsymbol{\beta}\}$ is the model parameter of TCRF, and we decide them based on the following Maximum A Posteriori (MAP) estimation by using all the training images with ground truth.

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}}\left\{\sum_{t=1}^{T} \log P(\mathbf{y}^t|\mathbf{X}^t;\boldsymbol{\theta}) - \frac{R}{2}||\boldsymbol{\theta}||^2\right\} \quad (3)$$

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 [9].

$p_i(y_i|\mathbf{x}_i;\boldsymbol{\alpha})$ is the class reliability distribution in each node based on the output of Gentle Adaboost. $p_{ij}(y_i, y_j;\boldsymbol{\beta})$ is the class co-occurrence between the adjacent nodes. For the final class estimation, we need to find the class of each node that maximizes the conditional distribution shown in Eq.2.

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

$$y_i^* = \arg\max_{y_i} \sum_{\mathbf{y} \setminus y_i} P(\mathbf{y}|\mathbf{X};\boldsymbol{\theta}) \quad (4)$$

where y_i^* is the class maximizing the posterior marginal distribution. Since the graph structure is tree structure, the global optimal estimation can be done by Belief Propagation [7].

By decreasing the segmentation error, we regard the estimation result in the bottom layer as the final estimation result. Since this estimation considers all estimation results in all layers of the hierarchy and is global optimal, the proposed method is robust to the scale variance of objects.

3. Experimental Evaluation

3.1 Overview of Experiment

We used Corel 7 Dataset for experiment. It includes 100 images with 7 classes. Each image is assigned ground truth at the pixel level. The size of image is 180×120 pixels.

Recognition rate was computed as the class average accuracy. We did training and test by leave-one-out method. Also, in hierarchical segmentation, we regarded the layer whose number of segments is about 200 as the bottom layer. This is called super-pixel representation.We set the number of layers to 6 layers, the number of Visual Words to 500 words.

We investigated the change of accuracy with or without the proposed method.

3.2 Result and Discussion

Table 1. Recognition Result

	No Hierarchization (NH)	proposal
Accuracy	70.2%	72.4%

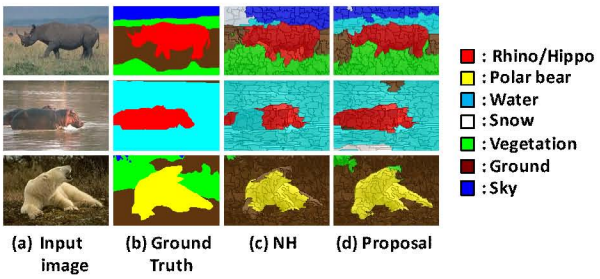

Figure 4. Example of recognition result

From Table.1, we can confirm that the proposed method improves the accuracy by 2.2%.

For discussion, some examples of the recognition result is shown in Fig.4 Compared with the conventional method (c), especially, the proposed method (d) corrects the false recognition near the border between different classes. This is because the proposed method can consider multi-scale by constructing TCRF based on hierarchical segmentaion.

4. Conclusion

In this paper, we proposed the new method to recognize generic objects in the CRF framework by incorporating the hierarchical structure based on Segmentation by Weight Aggregation. We called this tree structured model Tree Conditional Random Field. Because of considering multi-scale by hierarchization, the class estimation became robust to the scale variance. As a result, recognition accuracy is improved by 2.2%. In the future, we will find more useful context information except for class co-occurrence and new feature such as 3D information.

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. Shotton, J. Winn, C. Rother, and A. Criminisi, "Textonboost: joint appearance, shape and context modeling for multi-class object recognition and segmentation, " Proc. IEEE European Conference on Computer Vision, pp.1-15, 2006.

[4] S. Gould, J. Rodgers, D. Cohen, G. Elidan and D. Koller, "Multi-Class segmentation with relative location prior, " International Journal of Computer Vision, pp.300-316, 2008.

[5] 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.

[6] E. Sharon, A. Brandt, and R. Basri, "Fast multiscale image segmentation, " Proc. IEEE Computer Vision and Pattern Recognition, pp70-77, 2000.

[7] C. M. Bishop, "Pattern Recognition and Machine Learning, " Springer, Chapter.8, 2006.

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

[9] J. Nocedal, "Updating Quasi-Newton Matrices With Limited Storage, " Mathematics of Computation, pp.773-782, 1980.