GENERIC OBJECT RECOGNITION USING AUTOMATIC REGION EXTRACTION AND DIMENSIONAL FEATURE INTEGRATION UTILIZING MULTIPLE KERNEL LEARNING

Toru Nakashika, Akira Suga, Tetsuya Takiguchi, Yasuo Ariki

Kobe University, Graduate School of Engineering, Japan
{nakashika, akira1234}@me.cs.scitec.kobe-u.ac.jp, {takigu, ariki}@kobe-u.ac.jp

ABSTRACT

Recently, in generic object recognition research, a classification technique based on integration of image features is garnering much attention. However, with a classifying technique using feature integration, there are some features that may cause incorrect recognition of objects and a large amount of noise that causes a degradation in the recognition accuracy of image data. In this paper, we propose feature selection in an object area that is restricted by removing its background region, and multiple kernel learning (MKL) to weight each dimension, as well as the features themselves. This enables accurate and effective weighting since the weight is computed for each dimension using the selected feature. Experimental results indicate the validity of automatic feature selection. Classification performance is improved by using a background removing technique that utilizes saliency maps and graph cuts, and each dimensional weighting method using MKL.

Index Terms— Generic object recognition, Multi kernel learning, Feature integration, SIFT, HOG

1. INTRODUCTION

Recently, automatic classification, searching or tagging of images has become more important as digital cameras are now widely used all over the world, and the volume of image data on the web has become enormous. As a result, interest in object recognition by computer is increasing.

In object recognition research, the method using BoF (Bag-of-Features) is most widely used [1]. This is the technique derived from SIFT features [2] which describe local shapes of the target. In recent years, object recognition methods by MKL (Multiple Kernel Learning), which integrate varied image features, were proposed [3, 4]. We, human beings, recognize objects based upon various information, such as color, shape or texture. From this viewpoint, the feature-integration-based method has gained attention in the research of generic object recognition.

However, there is a problem with this method in that the feature weights could be learned incorrectly because the background in the image has a large amount of extraneous noisy features that may cause incorrect recognition of the target object. Therefore, we propose an object recognition method that combines automatic object region extraction and feature integration. Given an image set, a visually high attention region is extracted using a saliency map [6], and then the object region is cut out using a graph cuts algorithm [5] with seeds based on the saliency map. This process enables us to automatically delete the background region roughly and compute the adequate features of the target without prior knowledge. After that, 4 types of features (DoG-BoF, Grid-BoF, Color, Gabor) are obtained from each object-extracted image. Finally, using MKL-SVM, the image features are integrated, and at the same time an object classification is performed.

In the conventional method of object classification using feature integration by MKL, the weights of all the feature dimensions are learned equally in value [3]. For example, a simple color feature is composed of three feature dimensions, R, G and B, and their weights are learned to be equal. It is, however, true that both effective dimension and non-effective dimension for object recognition are included in a feature. Therefore, in this paper, we also propose a feature integration method in which the features are integrated with different weights to respective feature dimensions by expanding MKL in a way that prepares the kernels for each dimension.

We present the object extraction method using saliency maps and graph cuts in section 2. Then, the image features we use in this paper are introduced in section 3, and the feature integration method using MKL is described in detail in section 4. Finally, we demonstrate the effectiveness of our proposed method in section 5 and summarize our research and make conclusions in section 6.

2. OBJECT REGION EXTRACTION

In generic object recognition research, which attempts to recognize a target objects from its characteristic features in an image, the features included in the background region may negatively influence object recognition. To solve this problem, in this paper we introduce an effective method for object recognition that extracts an object region using a saliency map and then removes the background segment using graph cuts.

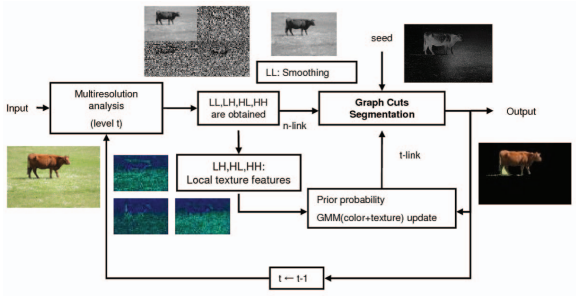

Fig. 1. Graph cuts using smoothing and local texture features.

Image segmentation is used to distinguish an object from background. In many proposed methods, the segmentation problem was solved as a minimization problem of energy; such as snakes [7], level set method [8] and graph cuts [5]. While in the two former approaches, a local minimum is searched from an energy function of the border, with the graph cuts algorithm, a global optimization can be calculated by minimizing an energy function including both region and border characteristics.

In this paper, we employ not the conventional graph cuts [5] but an expanded and more effective graph cuts approach that conducts image segmentation with an iterative smoothing using Multiresolution Analysis [9, 10]. Fig. 1 shows a flowchart of the segmentation.

The input image is decomposed into sub-bands (LL, LH, HL, HH) using Multiresolution Wavelet Analysis at level t. A smoothed image defined in a low-frequency range (LL) is used for the n-link, and local texture features defined in a high-frequency range (LH, HL, HH) are used for the t-link. The likelihood is derived from local texture features, as well as from color features. The prior probabilities are defined by a distance transform using the previous segmentation result, and the posterior probability obtained by multiplying one prior probability with the feature likelihood is set to the t-link edge as the edge cost. These processes are repeated until $t = 0$.

Letting the saliency map defined in the previous section be the seeds of the object for the graph cuts segmentation, the object region can be automatically selected with high accuracy. Even though it is difficult to perfectly remove the background region with this method, the background region can be cut off more or less accurately since most of the object has high saliency. Our goal is not to remove the background region completely but to be able to recognize the generic object. From this viewpoint, it is expected that the object recognition accuracy rate can be improved by eliminating the extra features included in the background region.

3. FEATURE DESCRIPTORS

3.1. Bag-of-Features (DoG & Grid)

First of all, SIFT features are extracted from training images. The features in an image are selected by using DoG (Difference-of-Gaussian) and Grid sampling [2]. DoG is a method that produces smoothing images by Gaussian filters with different scales, and determines the feature key points by detecting extreme values from the differential images. The BoF descriptor is, in this paper, obtained from an algorithm based on not only DoG but also Grid Sampling. In the Grid Sampling approach, multiple scales for feature key points are experimentally determined. The center of each circle indicates the position of key points, and the radius value of the circle denotes the scale size of the key points. SIFT can be obtained from orientation histograms containing samples from 4×4 sub-regions of the original neighborhood region around the key point. Each histogram covers 8 orientations; hence the SIFT descriptor has $4 \times 4 \times 8 = 128$ elements.

As a result, SIFT is an algorithm for detecting and extracting local feature descriptors that are reasonably invariant to changes in illumination, rotation, scaling, and small changes in viewpoint. For each obtained key point, a visual vocabulary is constructed through the clustering of key points using k-means algorithm [1]. Each key point cluster is regarded as a "visual word" in the vocabulary, and thus forms the Bag-of-Features (BoF) for describing visual content. In the BoF approach, the image is represented by a histogram with 1,000 bins of visual word frequency.

3.2. Color feature

Color information is not used in the SIFT features since it is extracted from gray-scale images. However, it is conceivable that color information is important for recognition of an object. So, in this paper, the color feature is represented as 3×3 color histograms including location information. The image is divided into 3×3 blocks, and a RGB color histogram with 64 bins is formed in each sub-region.

3.3. Gabor feature

A Gabor feature descriptor represents texture information with spatial localization, orientation and frequency. The features are obtained from convolving the image with Gabor filters of various frequencies and orientations.

Letting the image be divided into 3×3 sub-regions just like the color feature, a Gabor pattern in each subregion is extracted by the filters. The Gabor feature consists of 3×3 Gabor histograms, each of which contains the average values of the Gabor pattern with different conditions of the filter. We set the filters with $K = 6$ orientations and $S = 4$ frequencies.

The Gabor feature has the advantage of being robust in regard to varying illumination. Although it seems similar to SIFT features in terms of the local feature descriptor, we are able to prepare arbitrary patterns of scale and orientation for the Gabor feature.

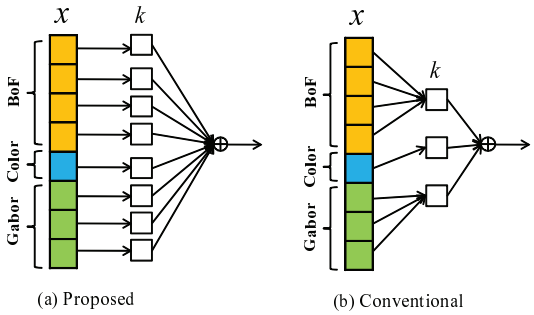

Fig. 2. In proposed method, each feature of each descriptor is integrated (a). Meanwhile, only features in descriptor level are integrated by MKL (b). x, k are a feature vector and a kernel vector of MKL, respectively.

4. DIMENSIONAL FEATURE INTEGRATION BY MKL

Multiple Kernel Learning (MKL) is a method for learning proper weights to the corresponding kernels, in using multiple classifiers with a kernel such as SVM (Support Vector Machine). A method for object recognition using MKL was proposed by Varma et al. [3]. In that paper, MKL was used as an integration method for image features by calculating appropriate weights corresponding to each feature. In this paper, object recognition is carried out using MKL for the feature integration in a similar way to Varma et al.'s method [3]. In Varma et al.'s approach, the MKL integrates each feature in feature descriptor level, whereas, in our method, all of the dimensions within each feature are integrated (Fig. 2).

We discuss the method of classifying objects by MKL and SVM (MKL-SVM) in more detail below. First a combined kernel $k(x, x')$ is defined as a linear combination of base kernels $k_l(x, x')$. The combined kernel is represented as the following:

$$k(x, x') = \sum_{l=1}^{K} \beta_l k_l(x, x')$$
$$\text{with } \beta_l \geq 0, \ \sum_{l=1}^{K} \beta_l = 1 \qquad (1)$$

where, x, x' are input super-vectors, which integrate the image features mentioned in Section 3. Each base kernel $k_l(x, x')$ corresponds to each dimension of the features in this paper. Therefore, the number of kernels K implies the number of dimensions used for object classification. The base kernel has its own weight β_l, and the weights can be optimized in the MKL framework. An efficient algorithm to solve MKL was proposed by Bach et al. as a corresponding dual problem [11].

As a result, the MKL-SVM procedure carries out the calculation of appropriate weights corresponding to each dimensional feature, as well as the learning of SVM used for the object classification.

5. EXPERIMENT RESULTS

Experiments start with the evaluation of the segmentation, which automatically cuts off the object region from an image using a saliency map and graph cuts. After that, we conducted object classification experiments using MKL-SVM with dimensional features. We also evaluated the validation of our proposed method by comparing object classifications with and without an object region extraction.

5.1. Experiments of Segmentation

We used Grab Cuts Database[1] for the segmentation experiment. The dataset includes 50 images with people, animals, cars or flowers, and each image has one object. Mask images corresponding to original images are also contained in the database.

Segmentation results are validated using an error rate, which is based on Over Segmentation (a detection rate of background in object region; Over) and Under Segmentation (a detection rate of object in background; Under) by use of mask images. The error rate is represented by combining Over and Under errors.

Table 1 shows the experimental results of the segmentation, changing a level t of Multiresolution Wavelet Analysis. As can be seen, automatic segmentation using saliency maps and graph cuts can be conducted merely at the error rate of around 7% in Over and Under by increasing the number of segmentation iterations. Segmentation with a larger number level t produces more effective results than normal graph cuts ($t = 1$). It is considered that this is because local errors are gradually corrected by changing level t. In the Multiresolution Wavelet Analysis approach, larger-level global segmentation shifts to smaller level local segmentation.

Even though the automatic segmentation approach can only extract the object region roughly, we consider that it is meaningful to partially extract the object region for object classification.

Table 1. Results of a segmentation experiment (%).

	t=1	t=2	t=3	t=4	t=5	t=6
Over	10.95	9.16	5.97	6.80	6.11	6.52
Under	18.03	9.56	7.25	7.26	7.26	7.27
Error	28.98	18.72	13.2	14.1	13.37	13.79

5.2. Classification Experiments

The Caltech-101 Database[2] was used for object recognition experiments, including images from 101 categories; such as

[1] http://research.microsoft.com/en-us/um/cambridge/projects/visionimagevideoediting/segmentation/grabcut/

[2] http://www.vision.caltech.edu/Image_Datasets/Caltech101/

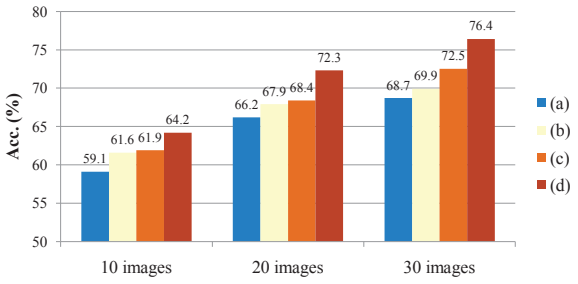

Fig. 3. Average classification rate with proposed method (%).

faces, cars, chairs, cameras, etc. We conducted the experiments with a different number of sets (10, 20 and 30 images for each category), and evaluated our method using cross-validation. We empirically employed χ^2 kernels for the kernel function of MKL-SVM. Preparing two-class MKL-SVMs for each category, we calculated the score values (distance to the hyperplane) of each MKL-SVM with different test images. The category with the highest score was chosen as the category the image belongs to. In order to verify the effectiveness of our method, we carried out 4 comparative experiments as follows:

(a) Without segmentation & feature-based MKL

(b) Without segmentation & dimension-based MKL

(c) With segmentation & feature-based MKL

(d) With segmentation & dimension-based MKL

Finally, we compared (a)∼(d) in terms of the classification accuracy rate.

The experiment results are summarized in Fig. 3, with (a)∼(d) in Fig. 3 corresponding to (a)∼(d) described, respectively. Through the results, it was confirmed that a dimensional weighting method (b) and an object region extraction method (c) both improved their classification accuracy, compared to a conventional method (a). Furthermore, the method applying both dimensional weighting and the object region extraction (d) is much more effective than any of the other methods (a)∼(c). It is believed that this is due to a synergistic effect between the removal of noisy background region and the detailed weighting of the extracted features.

6. CONCLUSION

In this paper, we proposed a method to classify objects by a combination of object region extraction using saliency maps and graph cuts and dimensional feature integration using Multiple Kernel Learning (MKL).

In a conventional method of feature integration, all dimensions in a feature are weighted with an equal value. Thus, the conventional method does not take the effectiveness or ineffectiveness of each dimension of the features into consideration. The feature integration method with noise in a background region also caused some errors in weighting.

In our proposed method, the features in an object region were segmented out by roughly removing a background region using saliency maps and graph cuts. Moreover, the detailed feature integration was realized by preparing kernels corresponding to each dimension of the features.

Experimental results showed the effectiveness of our proposed method. We confirmed that both the technique of object region extraction and the technique of dimensional feature weighting improved the accuracy of object classification.

7. REFERENCES

[1] G. Csurka, "Visual categorization with bags of keypoints," Proc. of ECCV Workshop on Statistical Learning in Computer Vision, 1–22, 2004.

[2] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," Journal of Computer Vision, pp.91–110, 2004.

[3] M. Varma, D. Ray, "Learning the discriminative power-invariance trade-off," In Proceedings of the IEEE International Conference on Computer Vision, Rio de Janeiro, Brazil, October. 2007.

[4] A. Vedaldi, V. Gulshan, M. Varma , A. Zisserman, "Multiple kernels for object detection," In Proceedings of the International Conference on Computer Vision, Kyoto, Japan, September. 2009.

[5] Y. Boykov, G. Funka-Lea, "Graph cuts and efficient N-D image segmentation," International Journal of Computer Vision, 70(2), pp. 109–131, 2006.

[6] R. Achanta, "Salient region detection and segmentation," 6th International Conference on Computer Vision Systems, pp.66–75, 2008.

[7] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: active contour models," International Journal of Computer Vision, vol. 1, no. 4, pp. 321–331, 1988.

[8] James A. Sethian, "Level set methods and fast marching methods: evolving interfaces in computational geometry fluid mechanics," Computer Vision, and Materials Science. Cambridge University Press, 1999.

[9] T.Nagahashi, H.Fujiyoshi, T.Kanade, "Image segmentaion using iterated graph cuts based on multi-scale smoothing," ACCV 2007, Part II, LNCS 4844, pp. 806-816, 2007.

[10] Keita Fukuda, Tetsuya Takiguchi, Yasuo Ariki, "Graph cuts by using local texture features of wavelet coefficient for image segmentation," ICME 2008 (International Conference on Multimedia and Expo), WE-PM2-L3.2, pp. 881-884, 2008.

[11] F. R. Bach, G. R. G. Lanckriet, M. Jordan, "Multiple kernel learning, conic duality, and the SMO algorithm," Proc. of International Conference on Machine Learning, 2004.

[12] S. Sonnenburg, G. R ¨ atsch, C. Sch ¨ afer, B. Sch ¨ olkopf, "Large scale multiple kernel learning," Journal of Machine Learning Research, vol. 7, pp. 1531–1565, 2006.