A FOOD IMAGE RECOGNITION SYSTEM WITH MULTIPLE KERNEL LEARNING

Taichi Joutou and Keiji Yanai

Department of Computer Science, The University of Electro-Communications
1-5-1 Chofugaoka, Chofu-shi, Tokyo 182-8585 JAPAN

ABSTRACT

Since health care on foods is drawing people's attention recently, a system that can record everyday meals easily is being awaited. In this paper, we propose an automatic food image recognition system for recording people's eating habits. In the proposed system, we use the Multiple Kernel Learning (MKL) method to integrate several kinds of image features such as color, texture and SIFT adaptively. MKL enables to estimate optimal weights to combine image features for each category. In addition, we implemented a prototype system to recognize food images taken by cellular-phone cameras. In the experiment, we have achieved the 61.34% classification rate for 50 kinds of foods. To the best of our knowledge, this is the first report of a food image classification system which can be applied for practical use.

***Index Terms*—** food image, multiple kernel learning, generic object recognition

1. INTRODUCTION

Since health care on foods is drawing people's attention recently, a food image system that can record everyday meals easily is being awaited. However, since there are so many categories of everyday meals to be recognized, it was impossible to realize a food image recognition system with practicable performance before. In fact, no practical systems for food image recognition have been proposed so far.

In these five years, researches on generic object recognition have progressed greatly due to developments of new feature representations and machine learning methods. Especially, the bag-of-features (BoF) representation [1] and kernel methods with a support vector machine (SVM) have made great breakthroughs. To improve image classification performance, recently integration of various image features such as color and texture in addition to BoF is being paid attention to. Varma et al.[2] proposed employing a multiple kernel learning (MKL) method to integrate various kinds of image features. They achieved 89.56% and 60.55% as the best classification rate for the Caltech-101/256 database which is one of de facto benchmark datasets for generic image recognition.

In this paper, we propose introducing a multiple kernel learning (MKL) into food image recognition. MKL enables to integrate various kinds of image features such as color, texture and BoF adaptively. This property of MKL is desirable, since useful recognition cues to recognize foods varies depending on foods. For example, while color seems to be useful to recognize "potage", texture is likely to be more useful to recognize "hamburger". By employing the MKL, we can estimate optimal mixing weights of image features for each category. Moreover, we implement a prototype system to recognize food images taken by cellular-phone cameras with the proposed method. In the experiment, we have achieved the 61.34% classification rate for 50 kinds of foods shown in Figure 1. If we accept the third candidate categories at most in terms of the output values of the 1-vs-rest classifiers, the classification rate reaches 80.05%.

2. RELATED WORK

As food image recognition, D. Pishva et al.[3] proposed a bread recognition system which can treat with 73 kinds of hand-made bread with the 95% classification rate. However, images in their dataset are taken by a special fixed camera setting in order to let the center of bread fit to the center of an image, and they used uniform background to separate bread regions from backgrounds easily. On the other hand, we treat with food images taken by many people in various settings. In fact, in the experiment, we use food images gathered from the Web.

To tackle such difficult problem, we use a Multiple Kernel Learning (MKL) to integrate various kinds of image features. MKL is a kind of extensions of a support vector machine (SVM). MKL treats with a combined kernel which is a weighted liner combination of several single kernels, while a normal SVM treats with only a single kernel. MKL can estimates the weights for a linear combination of kernels as well as SVM parameters simultaneously in the train step. The training method of a SVM employing MKL is sometimes called as MKL-SVM. Since MKL-SVM is a relatively new method which was proposed in 2004 in the literature of machine learning [4], there are only few works which applied MKL into image recognition.

Since by assigning each image feature to one kernel MKL can estimate the weights to combine various kinds of image feature kernels into one combined kernel, we can use MKL as a feature fusion method. As mentioned before, Varma et al.[2] proposed using MKL to fuse various kinds of image features and made experiments with Caltech-101/256. Similarly, Nilsback et al.[5] applied a MKL-based feature fusion into flower image classification. On the other hand, Kumar et al.[6] used MKL to estimate combination weights of the spatial pyramid kernels (SPK) [7] with a single kind of image features. Lampert et al.[8] estimated the degree of contextual relations between objects in the setting of multiple object

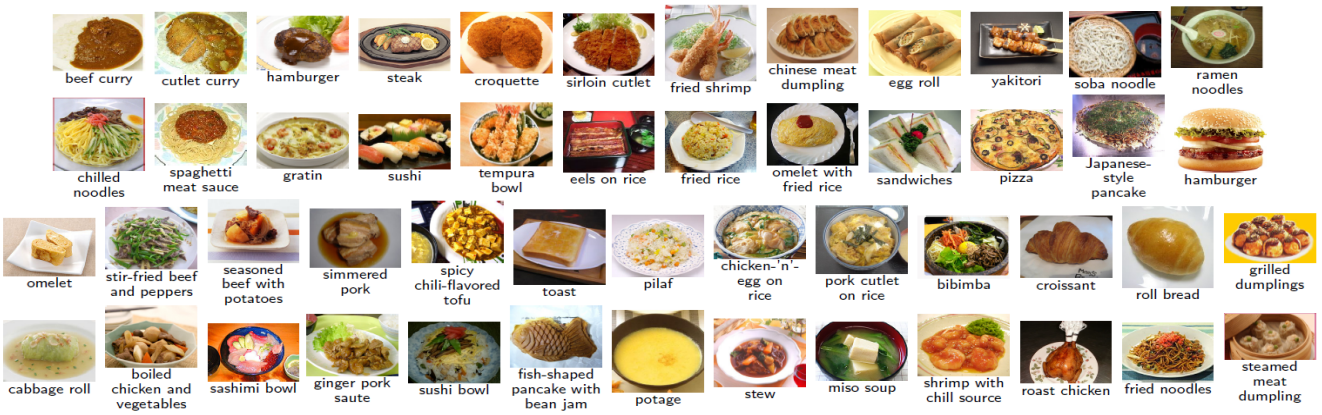

Fig. 1. 50 kinds of food images which are recognition targets in the paper.

recognition employing MKL. In this paper we propose food image recognition employing the MKL-based feature fusion method.

3. PROPOSED METHOD

In this paper, we realize image recognition of many kinds of foods with high accuracy by introducing a MKL-based feature fusion method into food image recognition. In our recognition, we prepare 50 kinds of food categories as shown in Figure 1, and classify an unknown food image into one of the categories. There has been no systems which can handle such many kinds of food categories so far.

In the training step, we extract various kinds of image features such as bag-of-features (BoF), color histogram and Gabor texture features from the training images, and we train a MKL-SVM with extracted features.

In the classification step, we extract image features from a given image in the same way as the training step, and classify it into one of the given food categories with the trained MKL-SVM.

3.1. Image Features

In this paper, we use the following image features: bag-of-features, color and texture.

Bag-of-Features: The bag-of-features representation [1] attracts attention recently in the research community of object recognition, since it has been proved that it has excellent ability to represent image concepts in the context of visual object categorization / recognition in spite of its simplicity. The basic idea of the bag-of-features representation is that a set of local image points is sampled by an interest point detector, randomly, or grid-based, and visual descriptors are extracted by the Scale Invariant Feature Transform (SIFT) descriptor [9] on each point. The resulting distribution of description vectors is then quantified by vector quantization against pre-specified codewords, and the quantified distribution vector is used as a characterization of the image. The codewords are generated by the k-means clustering method based on the distribution of SIFT vectors extracted from all the training images in advance. That is, an image are represented by a set of "visual words", which is the same way that a text document consists of words. In this paper, we use all of the following three kinds of strategies to sample: Difference of Gaussian (DoG), random sampling and regular grid sampling with every 10 pixels. In the experiment, about 500-1000 points depending on images are sampled by the DoG keypoint detector, and we sample 3000 points by random sampling. We set the number of codewords as 1000 and 2000.

Color Histogram: A color histogram is a very common image representation. We divide an image into 2×2 blocks, and extract a 64-bin RGB color histogram from each block with dividing the space into $4 \times 4 \times 4$ bins. Totally, we extract a 256-dim color feature vector from each image.

Gabor Texture Features: A Gabor texture feature represent texture patterns of local regions with several scales and orientations. In this paper, we use 24 Gabor filters with four kinds of scales and six kinds of orientations. Before applying the Gabor filters to an image, we divide an image into 3×3 or 4×4 blocks. We apply the 24 Gabor filters to each block, then average filter responses within the block, and obtain a 24-dim Gabor feature vector for each block. Finally we simply concatenate all the extracted 24-dim vectors into one 216-dim or 384-dim vector for each image.

3.2. Classification with Multiple Kernel Learning

In this paper, we carry out multi-class classification for 50 categories of food images. As a classifier we use a support vector machine (SVM), and we adopt the one-vs-rest strategy for multi-class classification. In the experiment, we build 50 kinds of food detectors by regarding one category as a positive set and the other 49 categories as negative sets.

A normal SVM can treat with only a single kind of image feature. Then, in this paper, we use the multiple kernel learning (MKL) to integrate various kinds of image features. With MKL, we can train a SVM with a adaptively-weighted combined kernel which fuses different kinds of image features. The combined kernel is as follows:

$$K_{comb}(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^{K} \beta_j K_j(\mathbf{x}, \mathbf{y})$$

$$\text{with} \quad \beta_j \geq 0, \quad \sum_{j=1}^{K} \beta_j = 1. \tag{1}$$

where β_j is weights to combine sub-kernels $K_j(\mathbf{x}, \mathbf{y})$. MKL can estimate optimal weights from training data.

By preparing one sub-kernel for each image features and estimating weights by the MKL method, we can obtain an optimal combined kernel. We can train a SVM with the estimated optimal combined kernel from different kinds of image features efficiently.

Sonnenburg et al.[10] proposed an efficient algorithm of MKL to estimate optimal weights and SVM parameters simultaneously by iterating training steps of a normal SVM. This implementation is available as the SHOGUN machine learning toolbox at the Web site of the first author of [10]. In the experiment, we use the MKL library included in the SHOGUN toolbox as the implementation of MKL.

4. EXPERIMENTAL RESULTS

In the experiments, we carried out image classification for 50 kinds of food images shown in Figure 1 to evaluate the proposed method.

First of all, we build a 50-category food image set by gathering food images from the Web and selecting 100 relevant images by hand for each category, since images on the Web are taken by many people in various real situations, which is completely different from an artificial setting for experiments. Basically we selected images containing foods which are ready to eat as shown in Figure 1. For some images, we clipped out the regions where the target food was located. Because originally our targets are common foods in Japan, some Japanese unique foods are included in the dataset, which might be unfamiliar with other people than Japanese.

The image features used in the experiments were color, bag-of-features (BoF) and Gabor. As color features, we used a 256-dim color histogram. To extract BoFs, we tried three kinds of point-sampling methods (DoG, random, and grid) and two kinds of codebook size (1000 and 2000). Totally, we prepared six kinds of the BoF vectors. As Gabor texture features, we prepared 216-dim and 384-dim of Gabor feature vectors which are extracted from 3×3 and 4×4 blocks, respectively. Totally, we extracted nine types of image feature vectors from one image. With MKL, we integrated all of the nine features.

We employ a SVM for training and classification. As a kernel function of the SVM, we used the χ^2 kernel which were commonly used in object recognition tasks:

$$K_f(\mathbf{x}, \mathbf{y}) = \sum_{f=1}^{K} \beta_f \exp\Big(-\gamma_f \chi_f^2(\mathbf{x}_f, \mathbf{y}_f)\Big)$$

$$\text{where } \chi^2(\mathbf{x}, \mathbf{y}) = \sum \frac{(x_i - y_i)^2}{x_i + y_i}$$

where γ_f is a kernel parameter. Zhang et al. [11] reported that the best results were obtained in case that they set the average of χ^2 distance between all the training data to the parameter γ of the χ^2 kernel. We followed this method to set γ.

For evaluation, we adopted 5-fold cross validation and used the classification rate which corresponds to the average value of diagonal elements of the confusion matrix. To compare between categories, we used the recall rate which is calculated as (the number of correctly classified images)/(the number of all the image in the category).

Table 1. Results from single features and fusion by MKL

image features	classification rate
color	38.18%
BoF (dog1000)	26.52%
BoF (dog2000)	27.48%
BoF (grid1000)	26.10%
BoF (grid2000)	27.68%
BoF (random1000)	28.42%
BoF (random2000)	29.70%
Gabor3x3	31.28%
Gabor4x4	34.64%
MKL (after fusion)	61.34%

Table 2. The best five and worst five categories in the recall rate of the results by MKL.

top 5	category	recall	worst 5	category	recall
1	miso soup	97%	1	simmered pork	18%
2	soba noodle	94%	2	ginger pork saute	28%
2	eels on rice	94%	3	toast	31%
4	potage	91%	4	pilaf	39%
5	omelet with fried rice	87%	4	egg roll	39%

Table 1 shows the classification results evaluated by the classification rate. While the best rate with a single feature were 38.18% by the color histogram, as the classification rate with MKL-based feature fusion we obtained 61.34% for 50-class food image categorization. If we accept three candidate categories at most in the descending order of the output values of the 1-vs-rest classifiers, the classification rate increases to 80.05%.

Table 3 shows the best five and the worst five food categories in terms of the recall rate of the results obtained by MKL, and Figure 2 shows food images of the best five categories in the recall rate. Variation of the food images belonging to the best five categories was small. Four kinds of food images out of the best five exceeded 90%, while "simmered pork" is less than 20%. This indicates that recognition accuray varies depending on food categories greatly. One of the reasons is that some of categories are taxonomically very close and their food images are very similar to each other. For example, images of "beef curry" and ones of "cutlet curry" are very similar, since both of them are variations of curry. Although selecting categories to be classified is not an easy task in fact, we need to examine if all the categories used in the experiments are appropriate or not carefully.

Figure 3 shows the weights estimated by MKL for the 1-vs-rest classifiers of ten categories, and the average weights. BoF (DoG2000) was assigned the largest weight, and BoF (random2000) became the second in terms of the average weight. The weight of color and Gabor were about only 10% and 7%, respectively. As a result, BoF occupied 93% weights out of 100%. This means that BoF is the most importance feature for food image classification, and DoG and random sampling are more effective than grid sampling to build BoF vectors. In terms of codebook size, 2000 is more useful than 1000, which shows larger codebooks is better

Fig. 2. Food images of the best five categories of the results by MKL.

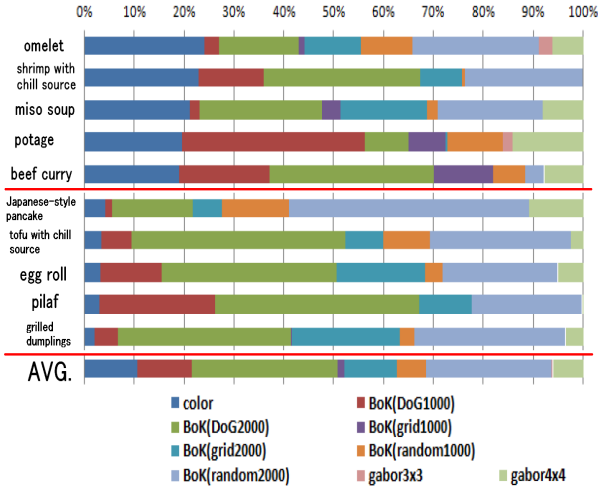

Fig. 3. Estimated weights to combine features.

than smaller ones regardless of sampling strategies.

4.1. Evaluation with a Prototype System

We implemented a prototype system to recognize food images taken by cellular-phone cameras. We can upload food images taken just before eating to the system from anywhere, and obtain a recognition result by a cellular-phone e-mail. Currently, the returned result includes only the names of top ten categories in the descending order of the output values of the 1-vs-rest classifiers. As future work, we plan to return the amount of calories and some advices on the meal the user is about to eat.

We have made this system available for the limited users for ten months on trial. As a result, 166 food photos were uploaded, and 62 images out of them were correctly classified, which means the 37.35% classificatino rate. In case of relaxing evaluation within the top three, the 55.43% classification rate was obtained which exceeded 50%.

In the experiment with the prototype system, since we did not instruct users how to take a food photo in advance, some uploaded food images were taken in the bad condition such that foods were shown in the photo as a very small region or images taken in the dark room were too dark to recognize. Therefore, the accuracy for the prototype system might be improved by instructing users how to take a easy-to-be-recognized food photo.

5. CONCLUSIONS

In this paper, we propose a food image recognition system employing the MKL-based feature fusion method. By estimating the optimal weight to combine different image features with MKL, we have achieved the 61.34% classification rate for 50 kinds of foods with the cross-validation-based evaluation. If we allow the system to return three candidate categories at most, the classification rate reached 80.05%. In addition, we implemented a prototype system to recognize food images taken by cellular-phone cameras, and we obtained 37.55% as the classification rate for 166 food images which were actually uploaded by the trial users.

As future work, we plan to extend the food image database by adding more categories, and to add more image features. To the best of our knowledge, this is the first report of a food image classification system which can be applied for practical use.

6. REFERENCES

[1] G. Csurka, C. Bray, C. Dance, and L. Fan, "Visual categorization with bags of keypoints," in *Proc. of ECCV Workshop on Statistical Learning in Computer Vision*, 2004, pp. 59–74.

[2] M. Varma and D. Ray, "Learning the discriminative power-invariance trade-off," in *Proc. of IEEE International Conference on Computer Vision*, 2007, pp. 1150–1157.

[3] D. Pishva, A. Kawai, K. Hirakawa, K. Yamamori, and T. Shiino, "Bread Recognition Using Color Distribution Analysis," *IEICE Trans. on Information and Systems*, vol. 84, no. 12, pp. 1651–1659, 2001.

[4] G. R. G. Lanckriet, N. Cristianini, P. Bartlett, L. E. Ghaoui, and M. I. Jordan, "Learning the kernel matrix with semidefinite programming," *Journal of Machine Learning Research*, vol. 5, pp. 27–72, 2004.

[5] M. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in *Proc. of Proceedings of the Indian Conference on Computer Vision, Graphics and Image Processing*, 2008.

[6] A. Kumar and C. Sminchisescu, "Support kernel machines for object recognition," in *Proc. of IEEE International Conference on Computer Vision*, 2007.

[7] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. of IEEE Computer Vision and Pattern Recognition*, 2006, pp. 2169–2178.

[8] Christoph H. Lampert and Matthew B. Blaschko, "A multiple kernel learning approach to joint multi-class object detection," in *Proc. of the German Association for Pattern Recognition Conference*, 2008.

[9] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[10] S. Sonnenburg, G. Rätsch, C. Schäfer, and B. Schölkopf, "Large scale multiple kernel learning," *Journal of Machine Learning Research*, vol. 7, pp. 1531–1565, 2006.

[11] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid, "Local Features and Kernels for Classification of Texture and Object Categories: A Comprehensive Study," *International Journal of Computer Vision*, vol. 73, no. 2, pp. 213–238, 2007.