Image Recognition of 85 Food Categories by Feature Fusion

Hajime Hoashi, Taichi Joutou and Keiji Yanai
Department of Computer Science, The University of Electro-Communications, Tokyo
1-5-1 Chofugaoka, Chofu-shi, Tokyo, 182-8585 Japan
{*hoashi-h, joutou-t, yanai*}*@mm.cs.uec.ac.jp*

***Abstract*—Recognition of food images is challenging due to their diversity and practical for health care on foods for people. In this paper, we propose an automatic food image recognition system for 85 food categories by fusing various kinds of image features including bag-of-features (BoF), color histogram, Gabor features and gradient histogram with Multiple Kernel Learning (MKL). In addition, we implemented a prototype system to recognize food images taken by cellular-phone cameras. In the experiment, we have achieved the 62.52% classification rate for 85 food categories.**

***Keywords*-food image recognition, feature fusion, multiple kernel learning**

I. Introduction

Since food image recognition helps record everyday meals easily, its realization is being awaited. However, since there are so many categories of everyday meals to be recognized, it was not easy to realize a food image recognition system with practicable performance. In fact, no practical systems for food image recognition exist at present.

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 the best classification performance at that time for the Caltech-101/256 databases which are de facto benchmark datasets for generic image recognition.

In [3], we proposed 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 helpful, since useful recognition features to recognize foods depend 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. In the experiment in [3], we achieved the 61.34% classification rate for 50 kinds of foods.

In this paper, we extended the system proposed in [3] in terms of image features and the number of food categories. As new image features, we added gradient histogram which can be regarded as simple version of Histogram of Oriented Gradient (HoG) [4], and we added 35 new food categories to the existing 50 categories. As a result, we obtained 62.52% classification rate for 85 food categories, which outperformed the result by the previous system for 50 food categories. This is because new features compensated decrease of classification rate due to increase of the number of categories.

The rest of this paper is organized as follows: Section 2 describes related work on object recognition including food image recognition. Section 3 explains the proposed method which is based on feature fusion of various kinds of image features with MKL. Section 4 describes the experimental results, and in Section 5 we conclude this paper.

II. Related Work

As food image recognition, D. Pishva et al.[5] proposed a bread recognition system which can handle 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. S. Yang et al.[6] proposed a food recognition system which was specialized for American fast-food such as hamburger, pizza and tacos. They defined eight basic food materials such as bread, beef and cheese, and recognized them and their relative position from images. Finally, they classified images into one of 61 categories using detected materials and their relations. However, the images used in their experiments were also taken with uniform backgrounds, which was far from practical settings. On the other hand, we treat 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 a combined kernel which is a weighted liner combination of several single kernels, while a normal SVM treats 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

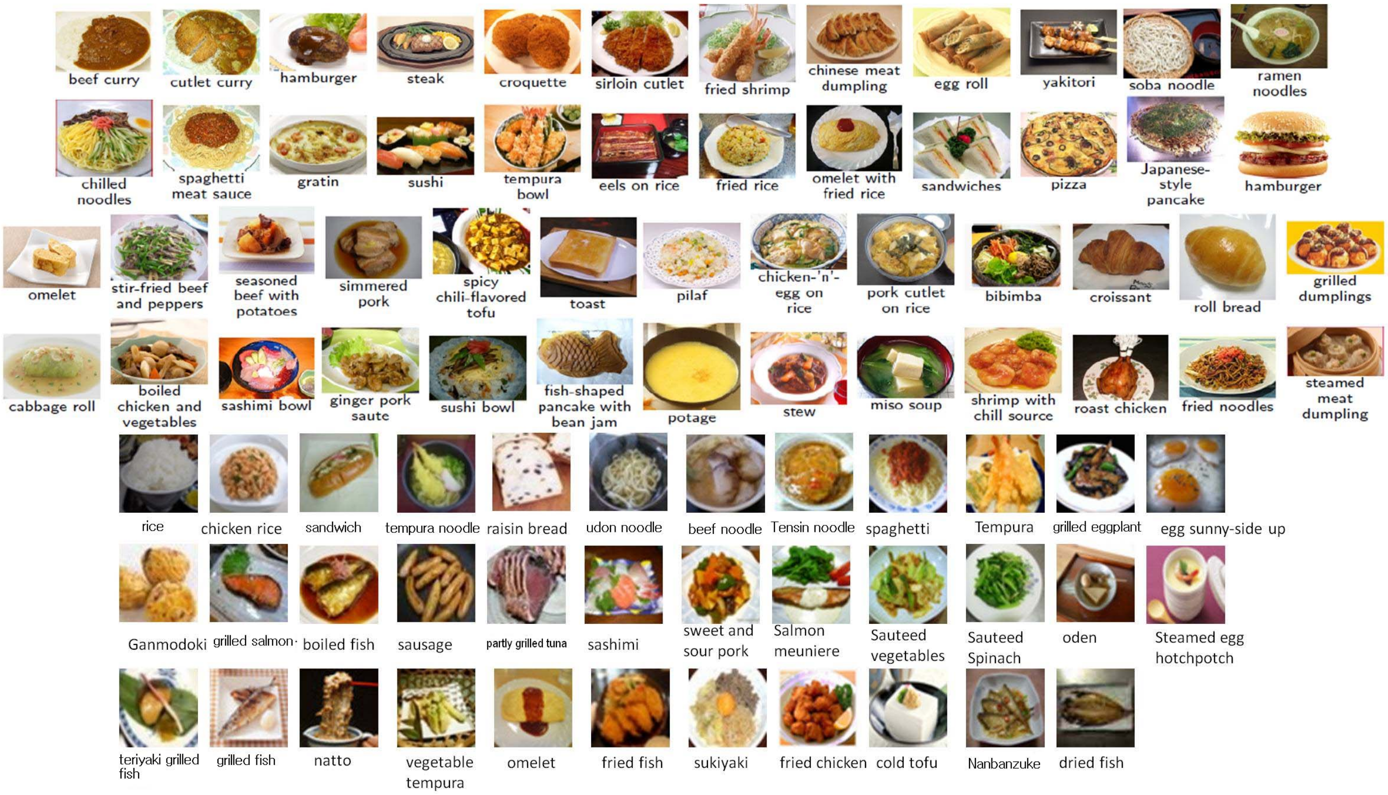

Figure 1. 85 kinds of food images which are recognition targets in the paper.

literature of machine learning [7], 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. [8] applied a MKL-based feature fusion into flower image classification. On the other hand, Kumar et al. [9] used MKL to estimate combination weights of the spatial pyramid kernels (SPK) [10] with a single kind of image features. Lampert et al. [11] estimated the degree of contextual relations between objects in the setting of multiple object recognition employing MKL. In this paper, we employ MKL-based feature fusion for image recognition of various kinds of foods.

III. Proposed Method

In this paper, we implement image recognition system which can handle many kinds of foods with high accuracy with MKL-based feature fusion method. In this paper, we prepare 85 kinds of food categories as shown in Figure 1, and classify an unknown food image into one of the predefined categories. As shown in Figure 1, 85 food dataset includes various kinds of food categories having various appearances.

In the training step, we extract various kinds of image features such as bag-of-features (BoF), color histogram, Gabor texture features and gradient histogram 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.

A. Image Features

In this paper, we use the following image features: bag-of-features (BoF), color, texture and gradient of images.

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 by grids, and visual descriptors are extracted by the Scale Invariant Feature Transform (SIFT) descriptor [12] 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 is 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 8 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 represents 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.

Gradient Histogram: Histogram of Oriented Gradients (HoG) was proposed by N. Dalal et al.[4]. It is similar SIFT in terms of how to describe local patterns which is based on gradient histogram. The difference between HoG and BoF is that BoF completely ignores location information of keypoints, while HoG keeps rough location information by building histograms for each dense grid and concatenating them as one feature vector. In short, HoG and BoF have different characteristics while both are composed of many local gradient histograms. In this paper, we used modified version of HoG. While the standard HoG divides a sub-window image into many small cells such as 6×12, the modified HoG in this paper divides a whole image into 1×1, 2×2, 4×4 or 8×8. In addition, we build two kinds of gradient histograms for each division. One takes into account “sign” of gradients and regards the direction spreading over 0 to 360 degrees, while the other regards the direction spreading over 0 to 180 degrees. Totally, we built 8 kinds of HoG vectors for an image.

B. Classification with Multiple Kernel Learning

In this paper, we carry out multi-class classification for 85 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 85 kinds of food detectors by regarding one category as a positive set and the other 84 categories as negative sets.

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 an adaptively-weighted combined kernel which fuses different kinds of image features. The combined kernel is as follows:

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

$$\text{with} \;\; \beta_j \geq 0, \;\; \sum_{j=1}^{K} \beta_j = 1. \qquad (1)$$

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.[13] 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 [13]. In the experiment, we use the MKL library included in the SHOGUN toolbox as the implementation of MKL.

IV. Experimental Results

In the experiments, we carried out 85-class image classification for 85 kinds of food images shown in Figure 1.

Before carrying out the experiments, we built a 85-category food image set by gathering food images from the Web and selecting 100 relevant images by hand for each category. Since Web images are taken by many people in various real situations, they can be considered to be challenging “real” food image data. 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), Gabor and gradient histogram as described in the previous section. 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. In addition, we prepared 8 kinds of gradient histograms as mentioned in the previous section. Totally, we extracted seventeen types of image feature vectors from one image. With MKL and without MKL, we integrated all of the seventeen features. In case of using no MKL to fuse features, we used uniform weights, which mean that all the weights are set as $1/17$.

We employ a SVM for training and classification. As a kernel function of the SVM, we used the χ^2 kernel which

Table I
RESULTS FROM SINGLE FEATURES AND FUSION BY MKL

image features	classification rate
BoF (dog1000)	33.47%
BoF (dog2000)	33.42%
BoF (grid1000)	30.73%
BoF (grid2000)	32.21%
BoF (random1000)	29.61%
BoF (random2000)	30.36%
Color	27.08%
Gabor (3×3)	23.60%
Gabor (4×4)	25.35%
Gradient (180, 1×1)	3.87%
Gradient (180, 2×2)	10.12%
Gradient (180, 4×4)	17.04%
Gradient (180, 8×8)	19.44%
Gradient (360, 1×1)	5.67%
Gradient (360, 2×2)	13.15%
Gradient (360, 4×4)	20.87%
Gradient (360, 8×8)	21.84%
SVM (uniform)	60.87%
MKL (mean-χ^2distance)	62.52%

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. [14] 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 I shows the classification results evaluated in the classification rate. While the best rate with a single feature were 33.47% by the BoF (dog1000), as the classification rate of feature fusion with the estimated weights by MKL, we obtained 62.52% for 85-class food image categorization, while 60.87% classification rate was obtained with uniform weights. This shows that MKL outperformed the uniform weights, although the difference was not so large. 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 increase to more than 80% as shown in Figure 2.

In addition, the classification rate obtained in [3] for 50-kind food classification was 61.34%, while we obtained 62.52% for 85-kind food classification. Although we constructed the 85-kind dataset by adding the 35 new classes to the dataset used in [3], the result for the 85-kind dataset in this paper outperformed the result of [3] for the 50-kind dataset. This is partly because the newly added features boosted the classification performance.

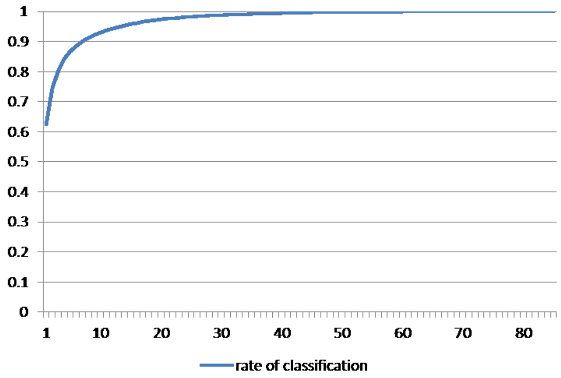

Figure 2. The classification rate when regarding that an image is correctly classified if the n top category candidates for the image contains the true category. The vertical axis and the horizontal axis represent the precision rate and the number of categories, respectively.

Table II
THE BEST FIVE AND WORST FIVE CATEGORIES IN TERMS OF THE RECALL RATE OF THE RESULTS BY MKL.

top 5	category	recall	worst 5	category	recall
1	soba noodle	95%	1	Ganmodoki	17%
2	eels on rice	94%	2	sandwich	24%
3	sauteed spinach	93%	3	toast	30%
4	miso soup	92%	4	grilled eggplant	30%
5	rice	90%	5	simmered pork	31%

Table II shows the best five and the worst five food categories in terms of the recall rate of the results obtained by MKL, and Figure 3 and Figure 4 shows food images of the best five categories and the worst five categories in terms of the recall rate, respectively. The variation of the appearances 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 “Ganmodoki” is less than 20%. This indicates that recognition accuracy 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 as future work.

Figure 5 shows the confusion matrix of 85-class food classification. Basically, the diagonal elements which correspond to the “correctly classified images” gathered most of the images. The categories into which incorrectly-classified images are classified were distributed over many categories except for some exceptions such as the confusion between “beef curry” and “curry” or between “pilaf” and “fried rice”.

Figure 6 shows the weights estimated by MKL for

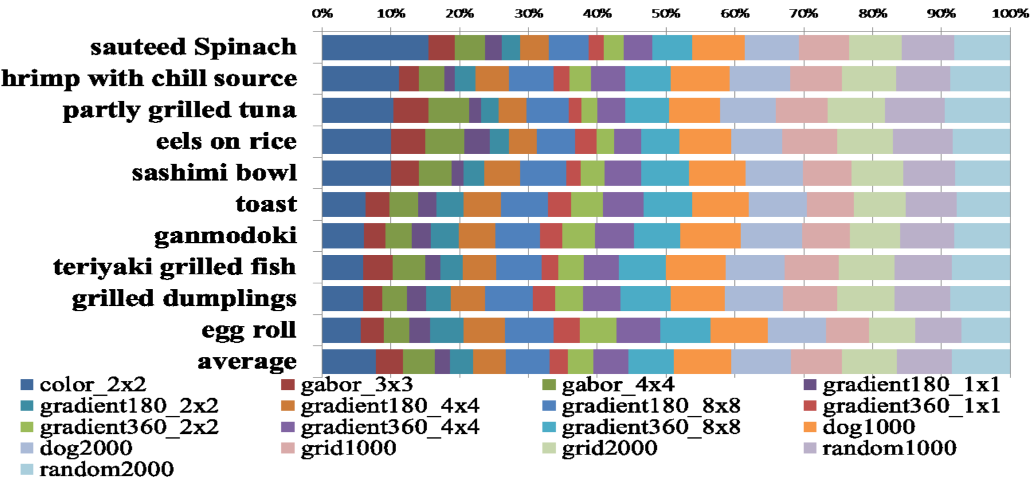

Figure 6. Estimated weights to combine features.

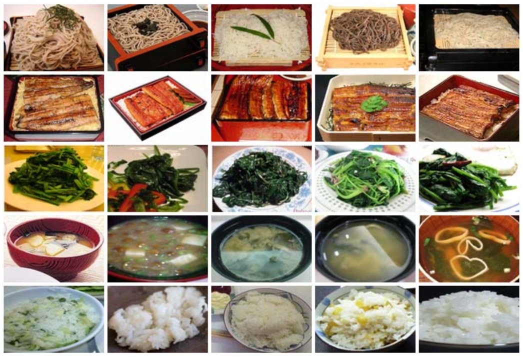

Figure 3. Some images of the best five categories of the results by MKL. From the top row, “soba noodle”, “eels on rice”, “sauteed spinach”, “miso soup”, and “rice” are shown in each row.

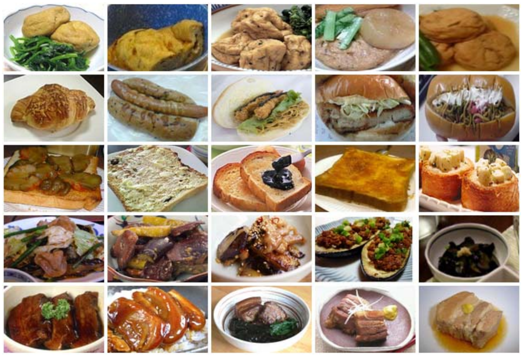

Figure 4. Some images of the worst five categories of the results by MKL. From the top row, “Ganmodoki”, “sandwich”, “toast”, “grilled eggplant”, and “simmered pork” are shown in each row.

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 8% and 8%, respectively. As a result, BoF occupied 49% weights out of 100%, while the total weight on Gradient was 35%. 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 than smaller ones regardless of sampling strategies. Among “Gradient” features, gradient180_8x8 was assigned with the largest weight. This shows the finer grid was better and the “sign” of gradients was not so important.

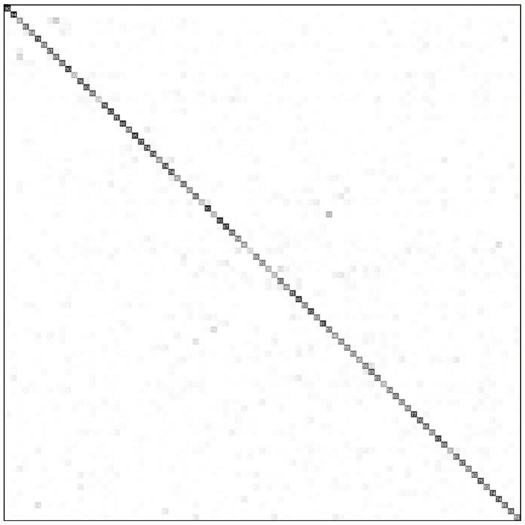

Figure 5. Confusion matrix. Grayscale indicates the ratio of classified images. The boxes on the diagonal line correspond to the correctly classified images, while the boxes out of the diagonal line mean confused classification results.

Figure 7 and Figure 8 shows some images of the most five categories and the least five categories in terms of the weight of “color”, respectively. The images shown in Figure 7 have typical colors within the same categories, while the images shown in Figure 8 have no typical colors and various colors within the same categories.

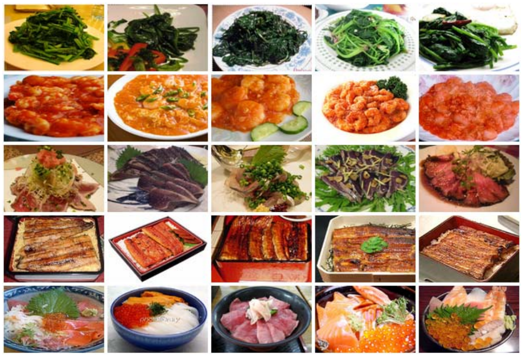

Figure 7. Some images of the most five categories in terms of "color" weights. From the top row, "sauteed Spinach", "shrimp with chill source", "partially-grilled tuna", "eels on rice", and "sashimi bowl" are shown in each row.

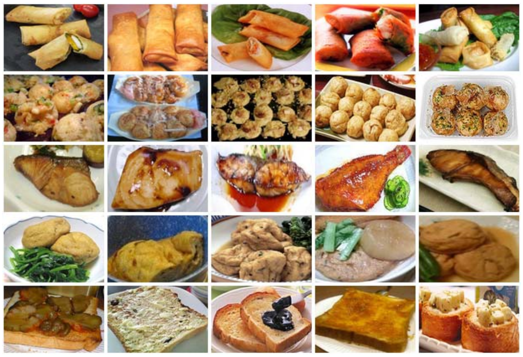

Figure 8. Some images of the least five categories in terms of "color" weights. From the top row, "egg roll", "grilled dumplings", "teriyaki grilled fish", "Ganmodoki", and "toast" are shown in each row.

A. 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 via e-mail. At present, the returned result includes the names of top ten categories with the values of their standard calories in the descending order of the output values of the 1-vs-rest classifiers.

We ran this system for the limited users for two years on trial. As a result, about 1000 food photos were uploaded, and the 85-kind food categories in our system converted 785 food images out of them. 356 images out of 785 images were correctly classified, which means the 45.3% classification rate. In case of relaxing evaluation within the top three and the top five categories, the 66.1% classification rate and the 69.4% classification rate were obtained, respectively.

In the experiment with cellular-phone photos, 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 an easy-to-be-recognized food photo.

V. Conclusions

In this paper, we extended the food recognition system proposed in [3] in terms of the number of categories and employed image features. By integrating seventeen kinds of image features with Multiple Kernel Learning, we obtained the 62.52% classification rate for 85-food-category classification evaluated by five-fold cross-validation. If we allow the system to return three candidate categories, the classification rate exceeded 80%. In addition, we implemented a prototype system to recognize food images taken by cellular-phone cameras, and we obtained 45.3% as the classification rate for 785 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 so as to cover most of the everyday foods of the average Japanese person. In the near future, we hope this system will be used as a food recognition engine which is a part of food health management service for cellular phone users.

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] T. Joutou and K. Yanai, "A food image recognition system with multiple kernel learning," in *Proc. of IEEE International Conference on Image Processing*, 2009.

[4] N. Dalal, B. Triggs, I. Rhone-Alps, and F. Montbonnot, "Histograms of oriented gradients for human detection," in *Proc. of IEEE Computer Vision and Pattern Recognition*, 2005, pp. 886–893.

[5] 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.

[6] S. Yang, M. Chen, D. Pomerleau, and R. Sukthankar, "Food recognition using statistics of pairwise local features," in *Proc. of IEEE Computer Vision and Pattern Recognition*, 2010.

[7] 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.

[8] 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.

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

[10] 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.

[11] 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.

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

[13] S. Sonnenburg, G. Rätsch, C. Schäfer, and B. Schölkopf, "Large Scale Multiple Kernel Learning," *The Journal of Machine Learning Research*, vol. 7, pp. 1531–1565, 2006.

[14] 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.