RECOGNITION OF MULTIPLE-FOOD IMAGES BY DETECTING CANDIDATE REGIONS

Yuji Matsuda, Hajime Hoashi and Keiji Yanai

Department of Informatics, The University of Electro-Communications, Tokyo
1-5-1 Chofugaoka, Chofu-shi, Tokyo, 182-8585 Japan
{matsuda-y,hoashi-h,yanai}@mm.cs.uec.ac.jp

ABSTRACT

In this paper, we propose a two-step method to recognize multiple-food images by detecting candidate regions with several methods and classifying them with various kinds of features. In the first step, we detect several candidate regions by fusing outputs of several region detectors including Felzenszwalb's deformable part model (DPM) [1], a circle detector and the JSEG region segmentation. In the second step, we apply a feature-fusion-based food recognition method for bounding boxes of the candidate regions with various kinds of visual features including bag-of-features of SIFT and CSIFT with spatial pyramid (SP-BoF), histogram of oriented gradient (HoG), and Gabor texture features.

In the experiments, we estimated ten food candidates for multiple-food images in the descending order of the confidence scores. As results, we have achieved the 55.8% classification rate, which improved the baseline result in case of using only DPM by 14.3 points, for a multiple-food image data set. This demonstrates that the proposed two-step method is effective for recognition of multiple-food images.

Index Terms— multiple-food image, region detection, window search, multiple kernel learning

1. INTRODUCTION

Recently, personal services to recode people's food habits using mobile phones have become popular. Users can become aware of own diet, and evaluate nutrition by recording their taken meals. When recording meals, inputing the names of food items by texts or selecting food items by hierarchical links is the common way. To record several items of foods in every meal in such way is a quite troublesome task. Therefore, it is desired to make recording of food items more easier and quickly. To this end, several methods to recognize food images have been proposed so far [2, 3, 4, 5].

However, all of these works assumed that one food image contained only one food item. They cannot handle an image which contains two or more food items such as a hamburger-and-french-fries image. In this paper, we propose a new method to recognize food images which contain two or more food items. In this paper, we call such images as "multiple-food images". The proposed method detects candidate regions with several methods including Felzenszwalb's deformable part model (DPM) [1], a circle detector and the JSEG region segmentation [6]. Then, we extract various kinds of image features from each candidate region. After applying the classification models trained by multiple kernel learning [7], we obtain the names of the top N food item candidates over the given image. The experimental results demonstrate that the proposed method is very effective for recognition of multiple-food images.

Note that the objective of our system is not associating extracted regions with names of food items directly, but listing all the names of the food items which are estimated to be shown in the given image, since our final objective is recording the lists of eaten items and calculating total amount of calorie of each meal automatically.

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. Section 4 describes the experimental results, and in Section 5 we conclude this paper.

2. RELATED WORK

2.1. Food Image Recognition

As food image recognition, S. Yang et al.[4] 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 in a food image. Finally, they classified images into one of 61 categories using detected materials and their relations. They achieve the 28.2% classification rate for Pittsburgh fast-food image dataset[8]. Zong et al.[5] also proposed a food recognition system employing SIFT detector and Local Binary Pattern (LBP). They achieved the better classification rate than S. Yang et al.'s results on the same fast food dataset. These two works focus on American fast foods, while we handle various 100 kinds of foods which are mainly common in Japan in this paper.

Joutou et al. proposed a food recognition system the target of which are 50 kinds of common food items in Japan [2]. They have proposed a method to recognize food images by fusing various kinds of image features include SIFT-based bag-of-features, Gabor, and color histograms using multiple kernel learning (MKL) [7], and have achieved the 61.34% classification rate. Hoashi et al. extended this system so as to recognize 85 kinds of food items [3].

However, all of above-mentioned works assumed that one food image contained only one food item, and the food item was shown as large as possible in an image. They cannot handle an image which contains two or more food items such

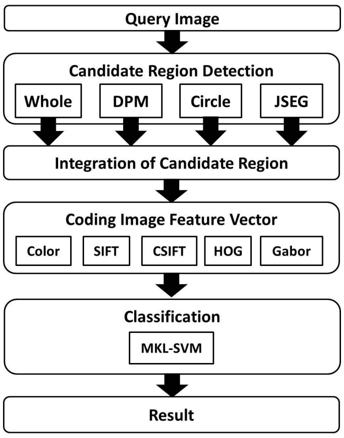

Query Image
Candidate Region Detection
Whole
DPM
Circle
JSEG
Integration of Candidate Region
Coding Image Feature Vector
Color
SIFT
CSIFT
HOG
Gabor
Classification
MKL-SVM
Result

Fig. 1. Recognition Flow

as an image that represents a meal set including a hamburger and french fries. In this paper, we propose a new method to recognize food images which contain two or more food items. To our best knowledge, this is the first work to recognize multiple-food items in one food image at the same time.

2.2. Object Detection

To detect objects in an image, a method based on sliding window is a standard method. Viola-Jones face detector is the most representative method among sliding-window-based methods [9]. Recently, as a method to detect generic objects the deformable part model (DPM) proposed by Felzenszwalb et al. [1] is quite popular because of its discriminative power and availability of source code. It is based on Histogram of Oriented Gradients (HoG) proposed by N. Dalal et al. [10]. In the DPM, several HoGs are used as part models, and their spatial relations are also modeled. As a training method, the latent SVM is used which is an extension of a standard SVM. The Felzenszwalb's DPM is widely-used as a standard method of generic object detection. Since the DPM is based on only HoG features and linear SVM, it is not sometimes effective to discriminate object categories which looks similar to each other.

To cope with this problem, A. Vedaldi et al. proposed an object detection method which combined linear-SVM-based window search and non-linear-SVM classifiers [11]. In the method, they detected several candidate regions using a jumping window search method roughly first which are slightly different from DPM, and next verified detected candidate regions with multiple kernel learning with non-linear kernel such as a chi-square-RBF kernel with high accuracy. Basically, we follow this idea to detect candidate regions of food items. In this paper, we use not only window search by DPM but circle detector and region segmentation algorithm to detect candidate regions over the given image.

3. PROPOSED METHOD

In this paper, we propose a food image recognition system which outputs the names of food items which are expected to be shown in a given food image. We show the overview of the processing flow of the proposed system in Figure 1. Given an input image, first, the system detects candidate regions of dishes. In this paper, we use four types of detectors including the deformable part model (DPM) [1], a circle detector, the JSEG region segmentation [6], and whole image. Next, we integrate bounding boxes of the candidate regions detected by the four methods. Then, we check the aspect ratio of width and height of the bounding boxes, and exclude irrelevant bounding boxes regarding their shapes from the candidate set. The system extracts various kinds of image features from the selected regions, and calculate SVM scores by multiple-kernel learning (MKL) [7] with non-linear kernels. Finally, we obtain the names of the top N food items over the given image regarding the descending order of the evaluation values.

3.1. Candidate Region Detection

The existing works on food image recognition assumed that one food image contained only one food item, and the food item was shown as large as possible in an image. On the other hand, we handle an image which contains two or more food items. To this end, at first, we estimate several candidate regions where food items are expected to exist before extracting image features. We use four kinds of candidate region detection methods including whole image, the deformable part model method (DPM) [1], a circle detector, and the JSEG region segmentation [6] as shown in Figure 1.

3.1.1. Whole Image

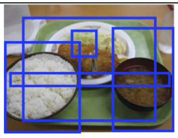
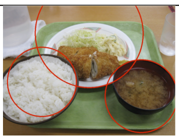
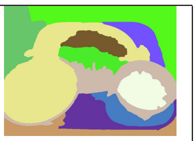
The simplest candidate region is a whole image. This candidate region is equivalent to the existing food recognition systems [2, 3, 4, 5] which assume that one image contains only one food item. This candidate is expected to work for an image containing one large dish, but does not suit for an image containing multiple small dishes.

3.1.2. Deformable Part Model (DPM)

As a main detector, we use the deformable part model (DPM) proposed by Felzenszwalb et al.[1]. The DPM is a two-layered hierarchical model, which consists of a global "root" filter and several part models. Each part model specifies a spatial model and a part filter. The spatial model defines a set of allowed placements for a part relative to a detection window, and a deformation cost for each placement. The score of a detection window is the score of the root filter on the window plus the sum over parts, of the maximum over placements of that part, of the part filter score on the resulting subwindow minus the deformation cost. Both root and part filters are scored by computing the dot product between a set of weights and HoG features [10] within a window.

To detect object regions, sliding window approach is adopted in the DPM. In addition, the DPM is defined at a fixed scale, and we detect objects by searching over an image pyramid. Therefore, to reduce computational cost, linear SVM are used in the DPM method. To compensate less discriminative power of linear SVM than non-linear kernel SVM, in the DPM, the latent SVM which is latent mixture of a standard SVM is used instead of a standard SVM. We trained the

Table 1. Candidate region detection method

Method	 **Whole**	 **DPM**	 **Circle**	 **JSEG**
# of candidates	1	100 (1 for each item)	4 (avg.)	14 (avg.)
Advantage	is suitable for larger food items	detect regions by HoG-based part model	detect dishes by contours of dishes.	detect dishes by segmentation.
Disadvantage	is unsuitable for small food items	is based on only gradient-based features.	dishes are not always circular.	segmentation sometimes fails.

DPMs with the latent SVM for each of 100 food categories in the experiments [1].

3.1.3. *Circle Detector*

A circle detector detects regions of dishes by extracting circular contours from an image. First, it converts a given image to a gray-scale image. Then, it extracts contours by the Canny Edge Detector. Finally, it detects circles by the Hough transform from extracted contours.

Note that we can detect not only circles but ellipses. However, in the preliminary experiments, ellipse detectors tend to detect too many ellipses. That is why we use a circle detector instead of more general ellipse detector.

3.1.4. *Region Segmentation*

Region segmentation is to divide an image into several pieces of regions. In this paper, we use the JSEG Algorithm proposed by Deng et al. [6] [2] as a region segmentation algorithm. JSEG divides an image by color space quantization and color class map. In JSEG, the number of segmented regions can be set as a parameter. In the experiment, we set the number of regions as 10.

We calculate circularity C of each region, and add only the regions where the circularity value exceeds the given threshold value into a region candidate set. The circularity C is calculated as follows:

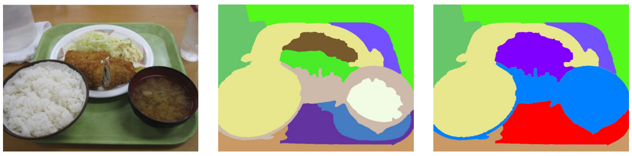
$$C = \frac{4\pi S}{L^2} \tag{1}$$

where S is the size of a region, L is the perimeter of a region. When C is closer to 1, the shape of the region is close to a circle. If C is much smaller than 1, the shape of the region is much far from a circle.

In addition, if circularity of the combined region of two adjacent regions is larger than original ones and the threshold, we add the combined region to a candidate set as shown in Figure 2.

3.2. Integration of Candidate Region

We simply aggregate all the candidate regions detected by four kinds of methods in a candidate set, and convert them into bounding boxes each of which circumscribes each detected region. In the second step, we extract various kinds of image features from the bounding boxes of all the detected regions, and apply non-linear kernel SVMs which are trained with MKL.



(a) input image (b) result of region segmentation (c) integration of region

Fig. 2. Candidate region detection by region segmentation

Moreover, we exclude apparently irrelevant regions from a candidate set to reduce classification cost and noisy candidates. In this paper, irrelevant regions are defined regarding their size and aspect ratio. We discard the bounding boxes the shorter side of which are less than 60 pixels and the bounding boxes the aspect ratios of which are more than twice standard deviations apart from the average aspect ratio of all the bounding boxes of food items in the training data.

3.3. Image Features

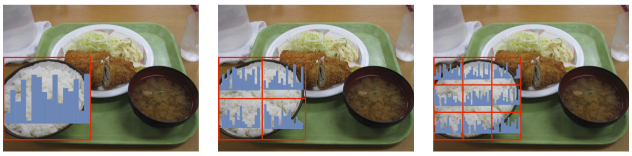
To recognize food images with visual features, just simply using the SIFT and color does not give good results. Therefore, in this paper, we integrate various kinds of image features in the same way as Joutou et al.'s work [2]. They used the multiple kernel learning (MKL) [7] to recognize whole images, while we use MKL to recognize each food items in each bounding box which is detected as a food candidate region. In this subsection, we describe the image features used in this paper including bag-of-features with spatial pyramid (SP-BoF), histogram of oriented gradient (HoG), Gabor texture features, and color histograms.

3.3.1. *Bag-of-features of SIFT and CSIFT*

The bag-of-features (BoF) representation [12] 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. In the scheme of BoF, first, a set of local image points is sampled and visual descriptors are extracted by the Scale Invariant Feature

[1] http://people.cs.uchicago.edu/˜pff/latent-release4/

[2] http://vision.ece.ucsb.edu/segmentation/jseg/software/



(a) level 1 (b) level 2 (c) level 3

Fig. 3. Three-level pyramids for BoF

Transform (SIFT) descriptor [13] on each point. In addition to SIFT, we also extract CSIFT [14] which is extracted SIFT from a RGB color space. CSIFT is proved to be robust against illumination changes [14]. In this paper, we use regular grid sampling with every 10 pixels for sampling local image points. Next, the resulting distribution of description vectors is then quantified by vector quantization against prespecified 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 the experiment, about several thousands of points depending on images are sampled by the grid sampling. We set the number of codewords as 1000.

3.3.2. Spatial Pyramid Representation

Since an image feature vector in the bag-of-features representation represented by a histogram of distribution of SIFT vectors, spatial information of SIFT vectors is discarded. Then, we use spatial pyramid representation [15] to take spatial information into account roughly. In spatial pyramid representation, object regions are divided by hierarchical grids. We extract a BoF vector from each of the grids and concatenate them into one long vector. In this paper, we use the three-level pyramid which consists of 1×1, 2×2 and 3×3 grids as shown in Figure 3. The dimension of the feature vector of spatial pyramid BoF (SP-BOF) is 1000 in the pyramid level 1, 4000 in the pyramid level 2, 9000 in the pyramid level 3, and 14000 totally. This SP-BoF feature is used not only for MKL-based region classification.

3.3.3. Histogram of Oriented Gradients

Histogram of Oriented Gradients (HoG) was proposed by N. Dalal et al. [10]. It is similar to 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 long feature vector. In short, HoG and BoF have different characteristics while both are composed of many local gradient histograms. In this paper, we divides a given region into 8×8 cells. Assuming that one block consists of 3×3 cells, one region corresponds to 6×6 blocks. Finally, we obtain a 2916-dim vector from each region.

3.3.4. Gabor texture feature

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, we divide a given region into 8×8 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 1536-dim vector for each region.

3.4. Classification for Candidate Region

After extraction of feature vectors from each candidate region, we calculate evaluation values of the candidate region regarding each of all the given categories using support vector machines (SVM) which are trained by multiple kernel learning (MKL).

As a kernel function of the SVM, in this paper, we use the $\chi^2 RBF$ kernel which were commonly used in object recognition tasks. $\chi^2 RBF$ kernel is defined as follows:

$$K(\mathbf{x}, \mathbf{y}) = \exp\Big(-\gamma \sum_i \frac{\|x_i - y_i\|^2}{x_i + y_i}\Big) \tag{2}$$

where γ is a kernel parameter. Zhang et al.[16] reported that the best results were obtained in case that they set the average of $\chi^2 RBF$ distance between all the training data to the parameter γ of the $\chi^2 RBF$ kernel. We followed this method to set γ.

In this paper, we use the multiple kernel learning (MKL) [7] 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 represented as follows:

$$K_{\text{combined}}(\mathbf{x}, \mathbf{x}') = \sum_{j=1}^{K} \beta_j K_j(\mathbf{x}, \mathbf{x}') \tag{3}$$

where $\beta_j \geq 0$ and $\sum_{j=1}^{K} \beta_j = 1$. β_j is weights to combine sub-kernels $\beta_j(\mathbf{x}, \mathbf{x}')$. MKL can estimate optimal weights from training data. In this paper, we train the support vector machine with one-vs-rest strategy using MKL in the same way as [2]. Since the number of the given categories was 100 in the experiment, we trained 100 models independently.

By applying trained models for each candidate regions regarding all the categories, we obtain evaluation values for each candidate region. We sort the evaluation values over all the candidate regions and all the categories in the descending order, and output the top N categories in terms of the evaluation values so that one food category is included in the output food name list only once.

Note that the objective of our system is not associating extracted regions with names of food items directly, but listing all the names of the food items which are estimated to be shown in the given image. Therefore, we do not output locations of bounding boxes of candidate regions.

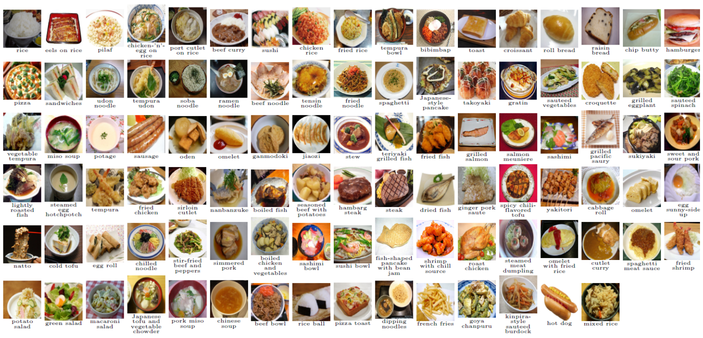


Fig. 4. 100 kinds of food used in the experiments. Please see this figure on a PDF viewer with magnification.

(a) Examples of multiple food-item images.

(b) Examples of single food-item images.

Fig. 5. Examples of multiple and single food-item images.

4. EXPERIMENTS

For experiments, we build a new food image dataset as shown in Figure 4 which includes 100 categories with bounding boxes on each food item. It contains about one hundred images for each category and 9060 images totally. It includes both of multiple and single food-item images as shown in Figure 5. In the experiments, we selected 500 multiple food-item images which contain 1200 food items, and 1200 single food-item images for testing, and we used the rest of all the images for training.

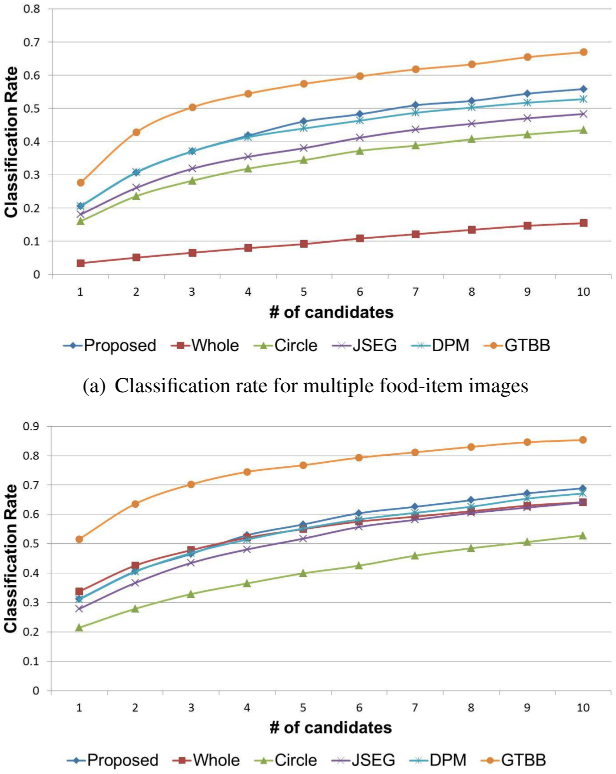
In the experiments, we trained two kinds of classifiers which are non-linear SVM to evaluate candidate regions, and linear-SVM to detect regions by the deformable part model (DPM). To train non-linear SVMs with MKL, we used all the image features extracted from the regions within the given bounding boxes as positive training samples, and image features extracted from background regions and images of other kinds of food items as negative training samples.

To evaluate the performance, we use a classification rate CR regarding food items, which is defined in the following equation:

$$\mathrm{CR} = \frac{\text{num. of correctly-detected food items in top N candidates}}{\text{num. of all the food items in all the test image}}$$

If the top N candidates include the names of the food items appearing in the given food image, we count them as the correctly-detected food items.

In the experiments, we compare the result by the proposed method with the results in case of using only single region candidate methods including four methods: whole image (Whole), deformable part model (DPM), circle detector (Circle) and region segmentation by JSEG (JSEG). "Whole" is equivalent to the existing methods as [2, 3], since they handled only single food-item images. Therefore, we regard "Whole" as a baseline method in this experiments. In addition, we compare the results with the result in case of using ground truth bounding boxes of the test images (GTBB), which means the case that region candidate detection is perfectly correct. This results can be regarded as being ideal results and the upper performance by introducing region detection.



(a) Classification rate for multiple food-item images

(b) Classification rate for single food-item images

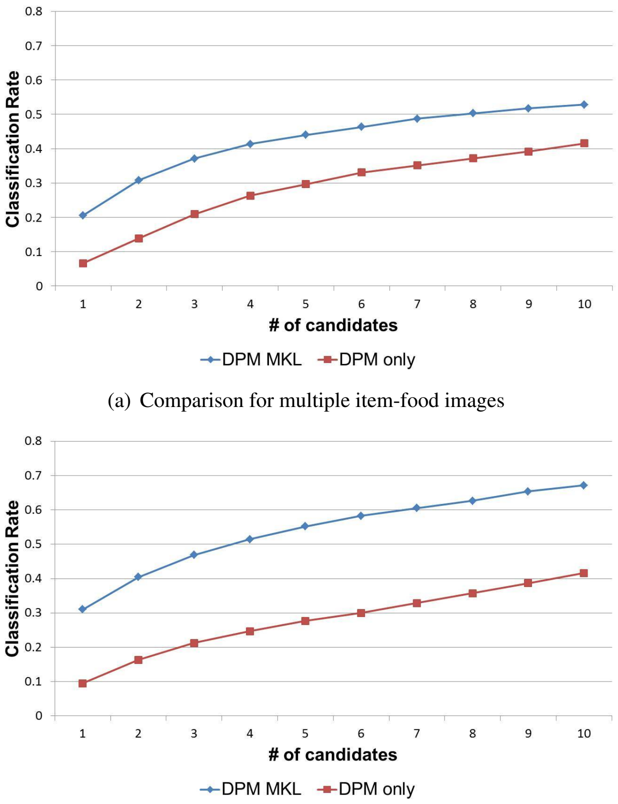
Fig. 6. Classification rates within the top N candidates

4.1. Experimental Results

Figure 6 (a) and (b) show the classification rates for multiple food-item images and for single food-item images, respectively, when varying the number of the top candidates N for evaluation. These figures contain the results by "Whole", "DPM", "Circle", "JSEG" and "GTBB" as well as the results by the proposed methods. Note that our main objective is to recognize food items in multiple food-item images, but we mention the results for single food-item images as well for reference.

When allowing top ten candidates, we achieved the 55.8% classification rate which were improved by 40.4 points compared to the baseline method for multiple food-item images, and the 68.9% classification rate which were improved by 4.7 points compared to the baseline method for single food-item images. The improvement for multiple food-item images was so much, which proves the effectiveness of the proposed method as a method to detect multiple food items.

Since a single food-item image contains only one food



(a) Comparison for multiple item-food images

(b) Comparison for single item-food images

Fig. 7. Comparison between DPM with MKL and only DPM

item, single food-item images tend to be covered with the foreground regions of food items and have relatively smaller background regions. Therefore, the difference between the proposed method and the baseline regarding classification rate is not so much. On the other hand, multiple food-item images tend to contain relatively larger background regions and several smaller foreground regions. Thus, the proposed method which recognizes not whole images but parts of images was required.

The results of "GTBB (ground-truth bounding boxes)" can be regarded as being ideal results and the upper performance by introducing region detection. The difference between the results by the proposed method and the "GTBB" results were about 10 points for both single and multiple food-item images, which is room to improve for future work in terms of candidate detection methods.

In the experiments, among single region detection methods, "DPM" was the best, since the DPM is one of the state-of-the-art object detectors. Then, we compare the results of "DPM MKL" which are obtained by applying both DPM and MKL, with the results of "only DPM" by using only DPM as shown in Figure 7. For multiple-food images, the classification rate of "DPM MKL" within top ten candidates was 52.8%, while "only DPM" was 41.5%. For single-food images, "DPM MKL" was 65.4%, while "only DPM" was 38.7%. These results shows that the classification rates were improved much by the second step where we apply non-linear kernel SVM by fusing various kinds of visual features.

5. CONCLUSIONS

In this paper, we proposed a two-step method to detect multiple food items from one food image. In the experiments, regarding the results of the top ten candidates, we have achieved 55.8% and 68.9% on the classification rate for multiple food-item images and single food-item images, respectively, which are improved by 40.4 points and 4.7 points compared to the results without region detection, and improved by 11.3 points and 26.7 points compared to the results by DPM alone. These results show the effectiveness of the proposed method.

For future work, we need to improve recognition accuracy by introducing co-occurrence probability between food items. In addition, to estimate calories of each food items shown in a food image accurately, we plan to study a method on estimate the amount of foods.

6. REFERENCES

[1] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010.

[2] T. Joutou and K. Yanai, "A food image recognition system with multiple kernel learning," in *Proc. of IEEE International Conference on Image Processing*, pp. 285–288, 2009.

[3] H. Hoashi, T. Joutou, and K. Yanai, "Image recognition of 85 food categories by feature fusion," in *Proc. of International Symposium on Multimedia*, pp. 296–301, 2010.

[4] 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.

[5] Z. Zong, D.T. Nguyen, P. Ogunbona, and W. Li, "On the combination of local texture and global structure for food classification," in *Proc. of International Symposium on Multimedia*, pp. 204–211, 2010.

[6] Y. Deng and B. S. Manjunath, "Unsupervised segmentation of color-texture regions in images and video," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 8, pp. 800–810, 2001.

[7] 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.

[8] M. Chen, K. Dhingra, W. Wu, L. Yang, R. Sukthankar, and J. Yang, "PFID: Pittsburgh fast-food image dataset," in *Proc. of IEEE International Conference on Image Processing*, pp. 289–292, 2009.

[9] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. of IEEE Computer Vision and Pattern Recognition*, pp. I:511–518, 2001.

[10] 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*, pp. 886–893, 2005.

[11] A. Vedaldi, V. Gulshan, M. Varma, and A. Zisserman, "Multiple kernels for object detection," in *Proc. of IEEE International Conference on Computer Vision*, 2009.

[12] 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*, pp. 59–74, 2004.

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

[14] A.E. Abdel-Hakim and A.A. Farag, "CSIFT: A SIFT descriptor with color invariant characteristics," in *Proc. of IEEE Computer Vision and Pattern Recognition*, vol. 2, pp. 1978–1983, 2006.

[15] 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*, pp. 2169–2178, 2006.

[16] 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.