# Automatic Expansion of a Food Image Dataset Leveraging Existing Categories with Domain Adaptation

Yoshiyuki Kawano Keiji Yanai

Department of Informatics, The University of Electro-Communications
1-5-1 Chofugaoka, Chofu-shi, Tokyo 182-8585 JAPAN
{kawano-y,yanai}@mm.inf.uec.ac.jp

**Abstract.** In this paper, we propose a novel effective framework to expand an existing image dataset automatically leveraging existing categories and crowdsourcing. Especially, in this paper, we focus on expansion on food image data set. The number of food categories is uncountable, since foods are different from a place to a place. If we have a Japanese food dataset, it does not help build a French food recognition system directly. That is why food data sets for different food cultures have been built independently category so far. Then, in this paper, we propose to leverage existing knowledge on food of other cultures by a generic "foodness" classifier and domain adaptation. This can enable us not only to built other-cultured food datasets based on an original food image dataset automatically, but also to save as much crowd-sourcing costs as possible. In the experiments, we show the effectiveness of the proposed method over the baselines.

**Keywords:** dataset expansion, food image, foodness, domain adaptation, crowdsourcing, adaptive SVM

## 1 Introduction

Recently, needs for food image recognition become larger, since food habit recording services for smartphones are spreading widely for everyday health care. For food habit recording, conventional ways such as inputing food names by texts or selecting food items from menus are very tedious, which sometimes prevent users from using such systems regularly. Then, several works on food recognition have been proposed so far [1–5] to make it easy to use food habit recording. In these works, the number of food categories is 100 at most, which is not enough for practical use. In fact, all of the foods we eat in our everyday life cannot be covered with only one hundred food categories, and the number of foods which can be recognized should be increased much more.

On the other hand, in these years, large-scale image classification is paid attention, and many methods for that have been proposed recently [6–9]. Due to these works, the number of categories to be recognized have been increased up to 1000. For example, in ImageNet Large Scale Visual Recognition Challenge

(ILSVRC), the number of categories to be classified is 1000. The data set for ImageNet Challenge is a subset of ImageNet [10], which is known as the largest visual database where the number of categories are more than 20,000. Large-scale image data sets such as ImageNet cannot be created by researchers by themselves. Most of them use crowd-sourcing Web services such as Amazon Mechanical Turk to build them semi-automatically.

In this paper, we propose a novel framework to expand an existing image dataset automatically leveraging existing categories. Especially, in this paper, we focus on expansion on food image data set.

While ImageNet covers comprehensive concepts, our target is restricted to foods. In ImageNet, annotation of each concept is gathered independently. On the other hand, since foods look more similar to each other, visual knowledge on foods of a certain country is expected to help collect annotations of food photos of the other countries. Then, in this paper, we propose a novel effective framework which utilizes knowledge on food of other countries by domain adaptation.

Basically, we gather food image candidates on novel food categories from the Web, and select good photos and add bounding boxes by using crowd-sourcing. In general, raw Web images include many noise images which are irrelevant to a given keyword. Especially, in this work, non-food images can be regarded as noise images. To exclude them from the gather images, we filter and re-rank Web images related to a given food category by using visual knowledge extracted from the existing food dataset.

Firstly, we built a generic "foodness" classifier from a Japanese food data set, UEC-Food100 [4]. We cluster all the food categories in the exist food image set into several food groups the member of which are similar to each other in terms of image feature vectors, and we train SVMs regarding each food group independently. Then, we evaluate unknown images using the trained SVMs on the food groups, and regards the maximum value of the output values of all the SVM as the "foodness" value of the given image. We can decide if a given image of a unknown category is a food photo or not based on the "foodness" value. In addition, because we select the maximum value from all the output valued of food groups, we estimate the most related food group to a given photo.

After "foodness" filtering, we obtain a food photo set. However, it might include food photos irrelevant to the given food keyword. Secondly, we select and re-rank more relevant images from the images judged as food photos by using transfer learning with visually similar categories in the source food photo data set. As a method of transfer learning, we use Adaptive SVM (A-SVM) [11] which can learn a discriminative hyper-plane in the target domain taking into account source-domain training data. In this work, the labeled data of the source categories which are visually similar to the target food photos are used as source-domain training data. As an initial target-domain training data, we use upper-ranked photos by a unsupervised image ranking method, VisualRank (VR) [12]. Then, we select food candidate images to be submitted for the crowd-sourcing by applying a trained A-SVM. By the experiments, the precision of the food

candidate photos by A-SVM has been proved to outperformed the results by only VisualRank and by normal standard SVM.

The contributions of this paper are as follows:

(1) Propose a novel framework to extend an existing image dataset with a generic "foodness" classifier and domain transfer learning.

(2) Three-step crowd-sourcing: selecting representative sample images, excluding noise photos, and drawing bounding boxes.

(3) Evaluate and compare accuracy of built food datasets and costs regarding the proposed method and two baselines.

(4) Apply the proposed framework in a large scale, and build a new 100-category food dataset based on the existing 100-category food dataset automatically.

## 2 Related Works

In the above-mentioned work, the target foods are limited to the foods which are common in a certain country. For example, US food [1, 3, 13], Chinese food [2] and Japanese food [4, 14]. From this observation, it is assumed that these food datasets were built to implementing food recognition systems the target of which are only the foods in the specific countries.

In addition, in the above-mentioned works, the number of target food categories is limited to 100 at most. From a practical point of view, 100 food categories is not enough for recognizing everyday foods for generic people. In fact, the number of foods we eat in our everyday life is much more than one hundred, and the number of foods which can be recognized should be increased much more.

Then, in this work, to make it easy to add the number of food categories and to implement food image recognition systems for other country foods or all the country foods, we propose a method to use an existing food dataset to build additional or another food dataset automatically by applying transfer learning.

On the Web, there are various kinds and huge amounts of images. It is very easy to collect images associated with a given keyword using Web API such as Bing Image Search API, Flickr API and Twitter API. However, raw Web images contain many noise images which are irrelevant to the given keyword. Therefore, many works on re-rank Web images regarding the given keyword have been proposed since ten years ago [15, 16]. Most of these works employed object recognition methods to select relevant images to given keywords from "raw" images collected from the Web using Web image search engines.

After spreading Amazon Mechanical Turk (AMT) which is the world-largest crowd-sourcing Web platform, it is commonly used for a task to select relevant images. AMT enables us to build a very huge-scale image dataset such as ImageNet [10], to build a middle- or large-scale dataset with bounding boxes [17], and to add attributes to a large-scale dataset [18].

In some works, AMT was incorporated into object recognition procedures, which was called "humans in the loop". Vijayanarasimhan et al. [17] proposed to combine active learning of object detectors and AMT crowd-sourcing tasks to draw bounding boxes as a loop procedure to raise accuracy of object detection gradually. On the other hand, Branson et al. [19] proposed complementary use of AMT with object classifiers by giving AMT workers simple easy questions to tackle difficult fine-grained object classification.

In addition, thanks to crowd-sourcing, many kinds of image datasets have released such as "bird" [20], "aircraft" [21], and "flower" [22]. They are intended to be built for fine-grained visual categorization research.

In this work, we use AMT as a crowd-sourcing service to select relevant images and add bounding boxes to selected food images. The objective is similar to [17]. However, while Vijayanarasimhan et al. [17] collected relevant images and their bounding boxes on each category independently, we collect images using knowledge of the known categories in the existing database with a "foodness" classifier and transfer learning.

In addition, as a pre-step of image selection, we prepare a task to ask the best representative photos regarding the given category. Some small number of representative photos are used to be shown workers as example photos to raise the accuracy of the image selection step.

## 3 Proposed Method

In this paper, we propose a novel framework to expand an existing image dataset automatically. The proposed framework consists of two stages: (1) the image selection stage, and (2) the crowd-sourcing stage.

In the image selection stage, we collect images from the Web with the given category names, and filter out noise images using a "foodness" classifier and adaptive SVM [11], both of which we train using knowledge of the existing food image database.

Then, in the crowd-sourcing stage, we crowdsource three kinds of tasks. First one is selecting representative images for the given new food category, the second one is discriminating relevant images from noise ones, and the third one is drawing bounding boxes on each of the selected images.

The processing flow of the proposed framework is shown in Fig.1. Each of the processing steps is explained as follows:

(1) Collect target food images associated with the given new food category from the Web.

(2) Evaluate "foodness" on each of the collected images, and select only high "foodness" images.

(3) Rank the selected food images with VisualRank, and train adaptive SVMs(A-SVM) [11] with upper ranked images as pseudo positive samples.

(4) Evaluate collected images again by A-SVM.

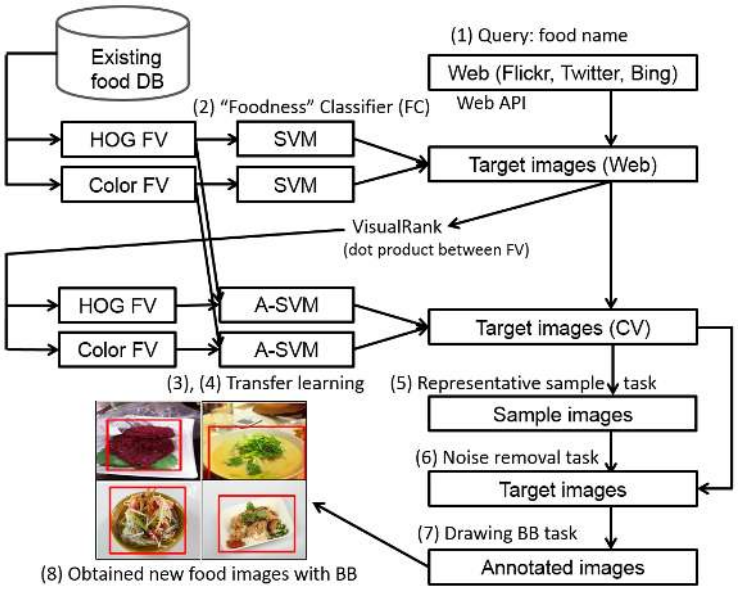


**Fig. 1.** Processing flow of the proposed framework.

(5) Crowdsource a task to select representative samples from the top 30 images in terms of A-SVM scores

(6) Crowdsource a task to discriminate relevant images from noise images for the images ranked higher by A-SVM

(7) Crowdsource a task to draw bounding boxes on the selected images.

(8) Add the annotated food images to a food image dataset.

### 3.1 Foodness Classifier

We construct a "Foodness" Classifier (FC) for discriminating and evaluating food images. FC evaluates if the given image is a food photo or not. We use FC to remove noise images from the images gathered from the Web.

We construct a FC from the existing multi-class food image dataset. Regarding feature extraction and coding, we adopt the same way as our mobile food recognition system [14]. Firstly, we train linear SVMs [23] in the one-vs-rest strategy for each category of the existing multi-class food image dataset. As image features, we adopt HOG patches [24] and color patches. Regarding foods, rotation and scale invariance is not so important. We regard fast extraction as more important, since we originally use these features for mobile food recognition. Both descriptors are coded by Fisher Vector (FV) [9, 25], and they are integrated in the late fusion manner. We perform multi-class image classification

in the cross-validation using the trained liner SVMs, and we build a confusion matrix according to the classification results.

Secondly, we make some category groups based on confusion matrix of multi-class classification results. This is inspired by Bergamo et al.'s work [26]. They grouped a large number of categories into superordinate groups the member categories of which are confusing to each other recursively. In the same way, we perform confusion-matrix-based clustering for all the food categories. We intend to obtain superordinate categories such as meat, sandwiches, noodle and salad automatically.

To build a "foodness" classifier (FC), we train a linear SVM of each of the superordinate categories. The objective of FC is judging if a food photo candidate which never corresponds to any food category in the existing dataset is a food photo or not. Therefore, abstracted superordinate categories are desirable to be trained, rather than training of all the food categories directly. The output value of FC is the maximum value of SVM output of all the superordinate food groups.

When training SVMs, we used all the images of the categories under the superordinate category as positive samples. For negative samples, we built a negative food image set in advance by gathering images using the Web image search engines with query keywords which are expected to related to noise images such as "street stall", "kitchen", "dinner party" and "restaurant" and excluding food photos by hand. All the images are represented by Fisher Vector of HOG patches and color patches. SVMs are trained in the late fusion manner with uniform weights. In the experiments, we will show the effectiveness of FC for evaluating "foodness" of food images of novel unknown categories based on visual knowledge of known food categories in the existing database.

### 3.2 Re-ranking with Domain Transfer

After "foodness" filtering, most of the remaining images are food images. However, they might includes other kinds of foods than the given food category. Since the objective of the proposed framework is collecting food images of novel unknown categories. To filtering out other food images than the target food category, we adopt discriminative approach with pseudo-positive samples in the similar way as Schroff et al. [16], since we have no labeled samples on the given novel category initially.

To get pseudo-positive samples, we rank the remaining images by the VisualRank [12] method. In addition, we use adaptive SVM [11] to leverage visual knowledge of the existing food image database. Because VisualRank is a unsupervised method to rank images which have many visually similar images in the upper ranking, it is useful to select relevant images from a noisy image dataset. However, it has drawback to narrow diversity of images. To compensate it, we use a domain adaptation method to leverage the existing food image database for classifying novel unknown food images.

**VisualRank** To select pseudo-positive images, we apply VisualRank [12] to the top $N$ images in term of "foodness" scores. We set $N$ as 300 in the experiments.

For computing VisualRank scores, we obtained similarity matrix $S$ as being dot product of Fisher Vectors (concatenated vectors of HOG FV and Color FV) instead of the number of matched local features. It has been proved that dot-product of FV can be regarded as a good similarity measure, and dot-product of L2-normalized vectors is equivalent to the cosine similarity [9, 27]. In addition, according to the following equation, we normalize each element $s$ of similarity matrix $S$ so that $s_{normalized}$ varies within only $[0, 1]$ range, because the value of elements of FV can be negative.

$$s_{\text{normalized}} = \frac{s - s_{min}}{s_{max} - s_{min}} \tag{1}$$

where $s_{max}$ and $s_{min}$ are maximum and minimum values among all the elements of similarity matrix $S$. Next, we column-normalized $S$ for computation of VisualRank. Regarding a bias vector $v$, we assign only the top $m$ images with uniform weights in the same way as [12]. We set $m$ as 100 in the experiments. Note that, we use the same Gaussian Mixture Model(GMM)s for FV coding estimated in the the previous step. This mean that the descriptors of unknown category are not modeled independently.

We calculate VisualRank score $r$ in the following equation:

$$r = \alpha * Sr + (1 - \alpha) * v, \tag{2}$$

where $\alpha$ is a damping factor where we set it to 0.85 according to [12].

**Selection of source domain samples and target domain samples** To train and apply adaptive SVM [11], we need to prepare source-domain labeled samples as well as target-domain labeled samples. Because for both domains we need to prepare positive and negative samples, totally we prepare four kinds of samples for training of A-SVM.

As target-domain positive samples, we use the top $M$ images in terms of the VisualRank scores, while as target-domain negative samples, we use the images with lower "foodness" scores in the initial image set gathered from the Web. In the experiments, we selected at most 300 images the "foodness" score of which were less than -0.6. Because the objective of this A-SVM-based re-ranking step is excluding noise images from the initial image set, not classifying generic images into one of food categories, we use negative samples which are peculiar to the given food category.

As source-domain positive samples, we use all the samples in the most related food group to the given new food category. As mentioned in the previous subsection, a "foodness" classifier can estimate the most related food group as well as a "foodness" score. We select the most frequent food group among the top 100 "foodness" images in the initial image set as the most related food group to the given new food category. As source-domain negative sample, we use the same negative food image set used in the previous step of a "foodness" classifier.

In the next step, we select positive samples to exclude noise images, and select effective negative samples for training. Regarding source-domain samples,

in general, the distribution of source-domain samples are wider than one of target-domain samples. Regarding target-domain samples, they are unreliable and tend to include outliers, since target-domain samples are selected automatically by a "foodness" classifier and VisualRank. Then we select the samples which are closed to target-domain samples as source-domain samples, and the samples which are closed to other target-domain samples as target-domain samples according to the following heuristics:

- Select the target-domain positive samples each of which has more than 3 positive samples among the nearest 5 samples over the space of all the target-domain (positive and negative) samples.
- Select the target-domain negative samples each of which has 5 negative samples among the nearest 5 samples over the target-domain space.
- Select the source-domain positive samples which are included in the union set of 7 closest source-domain all (positive and negative) samples to each of the selected target-domain positive samples.
- Select the source-domain negative samples which are included in the union set of 7 closest source-domain all samples to each of the selected target-domain positive samples.

This process is called "Sample Selection (SS)" in the section on experiments.

After source/target-domain positive/negative samples are selected finally, we train the adaptive SVM, and apply the trained model to re-rank the images in the image set after filtering by the "foodness" classifier. We use only higher-ranked images for crowd-sourcing tasks.

### 3.3 Crowd-sourcing

The final objective is obtaining a novel food image dataset with bounding boxes. In the previous steps, we applied "foodness" filtering and adaptive SVM re-ranking. However, the obtained food imageset is not perfect, and has no bounding box information. As the final steps, we crowdsource the following three kinds of tasks: (1) selecting representative sample images, (2) removing irrelevant images, and (3) drawing bounding boxes. As a crowd-sourcing service, we use Amazon Mechanical Turk (AMT).

**Representative Image Selection Task** We assumes that AMT workers does not have knowledge about various kinds of foods. Therefore, it cannot be expected to obtain highly accurate results without any preparations. Then, we prepare a task to select representative sample images as a pre-process step.

In this task, we ask AMT workers to select less than 10 representative images to the given food category from the top 30 image of A-SVM output scores, after studying about the given food category by visiting Wikipedia, Google Web search and Google Web image search with the name of the given food category as a query word. We design the task page so that AMT workers cannot submit the results without clicking the Web links to Wikipedia and Google sites. After

collecting results from 5 workers, we select the top 5 or 7 images as representative samples based on the number of votes by the workers. In the experiments, we set one HIT (Human Intelligence Task, which is a task unit in AMT.) of this task as 0.06$.

**Noise Removal Task** In this task, we ask AMT workers to annotate if the shown images are relevant to the given food category or not. In the task page, we show the representation images selected in the previous task. We believe this will be helpful for works who have never seen the target food. In one HIT, we use randomly-selected 25 images in the higher rank of the A-SVM scores. To prevent irresponsible worker, if there are more than four unchecked images, the result cannot be submitted. The results will be combined based on the majority voting. In the experiments, we requested each HIT for 5 workers. We set 1 HIT as 0.03$.

**Drawing Bounding Box Task** As the final task, we ask AMT workers to draw bounding boxes on the selected food images until the previous step. One HIT contains ten image annotation. In this step, worker can still mark irrelevant images as "noise" in the same way as the noise removal task, if they discover. After obtaining the results, we combine them by averaging the position of bounding boxes excluding images with no bounding boxes and too small bounding boxes. We add the finally selected images with bounding box to a new food image database as a ground-truth data. In the experiments, we requested each HIT for 4 workers. We set one HIT as 0.05$.

## 4 Experiments

In this section, we perform the following three experiments to evaluate the effectiveness of the proposed method.

- Performance comparison on food image filtering by a "foodness" classifier and adaptive SVM to leverage knowledge of the existing food dataset
- Evaluation of the final results after crowd-sourcing and analysis of crowd-sourcing cost.

Before evaluation, we describe a dataset, feature representation and initial food image collection from the Web. As an existing food dataset, we use "UEC-Food100" dataset [4] which consists 14361 food photos. Its number of food categories is 100, most of which are Japanese food categories. When building a "foodness" classifier (FC), we clustered 100 food categories into 13 food groups based on confusion matrix as shown in Tab.1. Note that the type of food groups in the table are named by hand for explanation.

As feature representation, we used 32-dim HOG local patches (8 orientations, 2x2) and 24-dim color local patches (mean and variance of RGB, 2x2) both of which are densely sampled from an image at difference 2 scales. After applying

**Table 1.** 13 food groups and their member foods.

| type of food group | food categories |
|---|---|
| noodles | udon nooles, dipping noodles, ramen |
| yellow color | omlet, potage, steamed egg hotchpotch |
| soup | miso soup, pork miso soup, Japaneses tofu and vegetable chowder |
| fried | takoyaki, Japaneses-style pancake, fried noodle |
| deep fried | croquette, sirloin cutlet, fried chicken |
| salad | green salad, macaroni salad, macaroni salad |
| bread | sandwiches, raisin bread, roll bread |
| seafood | sashimi, sashimi bowl, sushi |
| rice | rice, pilaf, fried rice |
| fish | grilled salmon, grilled pacific saury, dried fish |
| boiled and seasoned | seasoned beef with potatoes, simmered ganmodoki, seasoned beef with potatoes |
| sauteed | sauteed vegetables, go-ya chanpuru, kinpira-style sauteed burdock |
| sauce | stew, curry, stir-fried shrimp in chili sauce |

PCA, local descriptors are coded into Fisher Vector with GMM codebook (k=64) and a level-1 spatial pyramid (SPM) [28]. The GMM was estimated from the existing food dataset in advance.

Regarding initial food image collection from Web, we collected food images via Flickr API, Twitter API and Bing Image Search API based on query words associated with the given food category. We collected more than 600 images for each category. As query words, we used the words of both local language and English. We excluded duplicated URLs using a URL hash table after putting together all the image URLs gathered from three different APIs.

### 4.1 Evaluation on Image Filtering Results

For evaluation, we collected 35 categories of food image sets including 5 country foods with 7 categories for each country. All the 35 categories do not overlap with the categories in "UEC-FOOD100". We evaluated the precision of the top 300 food images (Precision@300) for each category. Note that we regarded badly-conditioned food images as being irrelevant. For example, an image with very small food region and an image in which only small portion of original food region is visible are not relevant.

We compare Precision@300 after filtering by the following six methods: (1) VisualRank with Fisher Vector, (2) "foodness" classifier (FC), (3) normal SVM using only target-domain training samples without "Sample Selection (SS)" after FC filtering (4) normal SVM using only target-domain training samples with SS after FC filtering (5) adaptive SVM using both source/target-domain training samples without SS after FC filtering, and (6) adaptive SVM using both source/target-domain training samples with SS after FC filtering. The last method (FC + A-SVM(SS)) corresponds to the proposed methods. Note that "Sample Selection (SS)" means the step to select of training samples for A-SVM

**Table 2.** Precision@300 of the food images ranked by six methods. The bottom method† is the proposed methods.

| no | Method | American | Japanese | Chinese | Thai | Indonesian | **Average** |
|---|---|---|---|---|---|---|---|
| (1) | VisualRank(VR) | 58.47 | 54.95 | 60.66 | 62.19 | 58.71 | 59.00 |
| (2) | "Foodness" (FC) | 78.00 | 75.33 | 77.61 | 82.85 | 78.61 | 78.48 |
| (3) | FC + SVM | 84.52 | 82.90 | 84.80 | 88.80 | 81.95 | 84.60 |
| (4) | FC + SVM(SS) | 85.57 | 83.38 | 85.09 | 89.23 | 82.23 | 85.10 |
| (5) | FC + A-SVM | 86.95 | 85.71 | 86.19 | 89.66 | 82.71 | 86.24 |
| (6) | FC + A-SVM(SS)† | **89.61** | **87.76** | **87.76** | **91.38** | **84.09** | **88.12** |

**Table 3.** Precision@200 of "pseudo-positive samples" which is provided to SVM or A-SVM as positive target samples.

| no | Method | American | Japanese | Chinese | Thai | Indonesian | **Average** |
|---|---|---|---|---|---|---|---|
| (A) | only FC | 79.21 | 77.00 | 80.21 | 83.78 | 83.00 | 80.64 |
| (B) | FC+VR | 85.00 | 83.78 | 85.78 | 89.14 | 86.21 | 85.98 |

or SVM (see Sec. 3.2), and VisualRank is still used for positive sample selection in (4)(5)(6)(7).

Tab.2 shows average Precision@300 of the results after filtering by each of the seven methods over 5 country foods and all 35 kinds of foods. Overall, the proposed methods outperformed other six baseline methods for all the regional foods.

Compared between VR and other supervised methods, the precision value by unsupervised VisualRank is not so good as the results by supervised discriminative classifiers such as FC and FC+A-SVM. In fact, FC improved Precision by about 20.0 points compared to VisualRank. This indicates that using existing categories helps improve filtering accuracy much, although they are different from the newly collected categories.

To use supervised methods such as SVM and A-SVM, we selected pseudo-positive samples from the top 300 images ranked by FC with Visual Rank(VR), and we used the top 200 images ranked by VR as pseudo-positive samples in the after steps. Tab. 3 shows the precision at the top 200 images before and after applying VR. Compared with two results, Precision@200 was improved by 5.34 points, which shows the effectiveness of applying VR after FC.

In case of FC+SVM, we used only target-domain training samples where positive samples are selected by VisualRank from the unlabeled samples, and manually-constructed common negative samples are used as negative samples. Although FC+SVM employs supervised SVM, the step itself is unsupervised because positive samples are "pseudo-positive" samples collected automatically. Even without supervision, FC+SVM improved by about 6 points compared to FC in terms of Precision@300. After adding training sample selection (SS) for SVM, the Precision was slightly improved.

FC+A-SVM and FC+A-SVM(SS) introduced a transfer learning method, adaptive SVM, which takes into account source-domain training samples as well. From their results, introducing domain transfer helps improve accuracy of image

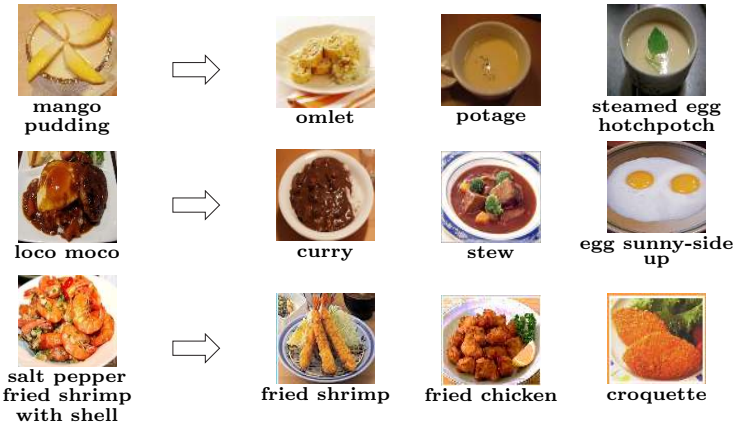


**Fig. 2.** The target-domain food images in the left column, and three source-domain samples in the most related food groups.

filtering, and the proposed method (FC+A-SVM(SS)) has achieved the best result, which proves the effectiveness of the proposed method.

Fig.2 shows the top 3 food categories in terms of Precision@300 among all the 35 categories, "mango pudding", "loco moco" and "fried shrimp with shell" in the left column, and three source-domain samples in the corresponding food groups. The images in each food group are used as source-domain positive samples when training adaptive SVM. The target-domain images looks similar to source-domain images in terms of color, shape or ingredients. All the foods in the first row in the figure are light-yellow, the foods in the second row have brown-colored source, and the foods in the bottom row have fried ingredients. From these results, the new category images can be classified with visually similar images of the existing categories in the most related food group by using transfer learning. This is a part of the contributions of this work.

### 4.2 Evaluation accuracy and costs of crowdsourcing

We evaluate the effectiveness of showing representative samples to workers, accuracy of obtained image sets and crowdsourcing costs.

**Workers' Evaluation on Representative Samples** We prepared a task to select representative sample images as a pre-process step. Selected representation images were shown in the page of noise removal task and drawing bounding box task in order to teach workers what relevant food photos look like. To evaluate its effectiveness, we asked workers in each HIT if sample images shown in the HIT page are useful, so so or useless. As a result, 3495 and 5359 answers are obtained in noise removal task and annotation bounding box task. Tab.4 shows the ratio of each answer, which shows the effectiveness of showing representative samples in both noise removal task and annotation bounding box task.

**Table 4.** Evaluation by workers on representative samples images (%)

| | useful | so so | useless |
|---|---|---|---|
| noise removal task | 89.59% | 7.90% | 2.52 % |
| drawing bounding box task | 91.68% | 7.02% | 1.31 % |

**Table 5.** Precision of food images on dataset by difference 3 methods.

| | precision | gain |
|---|---|---|
| FC + BB task | 91.10 | - |
| FC + A-SVM + BB task | 94.19 | +3.09 |
| FC + A-SVM + NR task + BB task | 97.83 | +3.64 |

**Evaluation of accuracy and costs** To evaluate accuracy and costs including crowdsourcing, we constructed three kinds of datasets by the following different combination of filtering steps: (1) FC + drawing bounding box task (BB task), (2) FC + A-SVM(SS) + BB task, and (3) FC + A-SVM(SS) + noise removal task (NR task) + BB task. In case of (1) and (2), workers have to mark irrelevant images as "noise" in addition to drawing bounding boxes to relevant images in the BB task, because the noise removal task is not included. The combination (1) is the simplest, and in (2) adaptive SVM was added. The last combination where noise removal task is prepared as an independent task is equivalent to the proposed framework. Note that all the combination includes representative sample selection task.

Tab.5 shows the precision of food images on the constructed dataset after crowsourcing by each combination. The precision by "FC+BB task" was 91.1%, while the precision by "FC+A-SVM(SS)+BB task" was 94.19%. Introducing A-SVM(SS) improved 3.09%, while it improved about 10% regarding the precision of filtered image sets before crowdsoucing. Although both combinations employs human annotation via crowdsourcing, the difference in precision appeared after crowdsourcing. This is estimated to come from the accuracy of dataset to supply workers. From this observation, to get more accurate results from crowdsourcing, more accurate data should be provided to crowdsourcing workers.

Compared between "FC+A-SVM(SS)+BB task" and "FC+A-SVM(SS)+NR task+BB task", separating noise removal task from drawing bounding box improved the precision, although provided datasets are the same. This indicates that crowdsourcing tasks (HITs) should be include only one kinds of jobs. Of course, increase of the number of crowdsourcing steps means increase of economical costs. We compare costs among the three cases in the next.

Tab.6 shows the recovery ratio and cost for the three combinations. The recovery ratio means the ratio of the number of the images which were finally annotated with correct bounding boxes over the number of provided images to workers for crowdsourced annotation. If the recovery ratio is low, many irrelevant samples are provided to workers, which means economical costs increase. "Costs"

**Table 6.** Recovery ratio(%) and costs($) to get annotated 100 images.

| | noise removal | | bounding box | | total |
|---|---|---|---|---|---|
| | recovery ratio | cost | recovery ratio | cost | total |
| FC + BB task | - | - | 64.2 | 3.11 | 3.11 |
| FC + A-SVM + BB task | - | - | 74.7 | 2.68 | 2.68 |
| FC + A-SVM + NR task + BB task | 80.9 | 0.74 | 86.7 | 2.31 | 3.16 |

shown in the table means the money ($) paid for AMT to get 100 annotated images. To avoid wasting money, the recovery rate should be high, hopefully close to 100%.

"FC+BB task" was apparently a bad strategy, because the total cost is high and the accuracy of the obtained results shown in Tab.5 is worst. "FC+A-SVM+BB" performed the best among the three strategies in terms of cost. Adding the noise removal task, the cost increased, because the number of crowdsourcing steps also increased. However, the precision of the final obtained results was the best as shown in Tab.5. That shows that there is a trade-off between cost and accuracy. It depends on the policy when building a dataset. Of course, Tab.6 shows just one case. If a unit price for HIT of each task is changed, the result of cost analysis will be changed. Regarding cost, more accurate data should be provided to crowdsourcing workers to raise the recovery rate. To do that, introducing "foodness" classifier and adaptive SVM is very effective.

## 5 Conclusions

In this paper, we proposed a novel framework to expand an existing image dataset automatically employing generic classifiers and domain adaptation to leverage visual knowledge in the existing dataset. Especially, in this paper, we focused on expansion on food image data set. In the experiments, we showed the effectiveness of the proposed method over baselines in terms of the proposed image filtering methods and the proposed procedure for crowdsourcing.

For future work, we will make further analysis on the difference between a hand-collected food image dataset and an automatically collected dataset by the proposed framework. In addition, we plan to extend the framework to other categories than foods such as clothes and animal.

We will release a new large-scale food photo dataset collected by the proposed framework by the time of ECCV 2014. It will include more than 256 kinds of foods from various countries such as French, Italian, US, Chinese, Thai, Vietnamese, Japanese and Indonesia.

## References


1. Yang, S., Chen, M., Pomerleau, D., Sukthankar, R.: Food recognition using statistics of pairwise local features. In: CVPR. (2010)
2. Chen, M., Yang, Y., Ho, C., Wang, S., Liu, S., Chang, E., Yeh, C., Ouhyoung, M.: Automatic chinese food identification and quantity estimation. In: SIGGRAPH Asia 2012 Technical Briefs. (2012)
3. Bosch, M., Zhu, F., Khanna, N., Boushey, C.J., Delp, E.J.: Combining global and local features for food identification in dietary assessment. In: ICIP. (2011)
4. Matsuda, Y., Yanai, K.: Multiple-food recognition considering co-occurrence employing manifold ranking. In: ICPR. (2012)
5. Kawano, Y., Yanai, K.: Real-time mobile food recognition system. In: Proc. of IEEE CVPR International Workshop on Mobile Vision (IWMV). (2013)
6. Vedaldi, A., Zisserman, A.: Efficient additive kernels via explicit feature maps. IEEE Trans. on PAMI **34**(3) (2012) 480–492
7. Wang, J., Yang, J., Yu, K., Lv, F., Huang, T., Gong, Y.: Locality-constrained linear coding for image classification. In: CVPR. (2010) 3360–3367
8. Zhou, X., Yu, K., Zhang, T., Huang, T.S.: Image classification using super-vector coding of local image descriptors. In: ECCV. (2010) 141–154
9. Perronnin, F., Sánchez, J., Mensink, T.: Improving the fisher kernel for large-scale image classification. In: ECCV. (2010)
10. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A Large-Scale Hierarchical Image Database. In: CVPR. (2009)
11. Yang, J., Yan, R.and Hauptmann, A.G.: Cross-domain video concept detection using adaptive svms. In: ACM MM. (2007)
12. Jing, Y., Baluja, S.: Visualrank: Applying pagerank to large-scale image search. IEEE Trans. on PAMI (2008)
13. Chen, M., Dhingra, K., Wu, W., Yang, L., Sukthankar, R., Yang, J.: PFID: Pittsburgh fast-food image dataset. In: ICIP. (2009) 289–292
14. Kawano, Y., Yanai, K.: Rapid mobile food recognition using fisher vector. In: ACPR. (2013)
15. Yanai, K., Barnard, K.: Probabilistic Web image gathering. In: ACM SIGMM WS Multimedia Information Retrieval. (2005) 57–64
16. Schroff, F., Criminisi, A., Zisserman, A.: Harvesting image databases from the web. In: ICCV. (2007)
17. Vijayanarasimhan, S., Grauman, K.: Large-scale live active learning: Training object detectors with crawled data and crowds. In: CVPR. (2011) 1449–1456
18. Patterson, G., Hays, J.: Sun attribute database: Discovering, annotating, and recognizing scene attributes. In: CVPR. (2012) 2751–2758
19. Branson, S., Wah, C., Schroff, F., Babenko, B., Welinder, P., Perona, P., Belongie, S.: Visual recognition with humans in the loop. In: ECCV
20. Welinder, P., Branson, S., Mita, T., Wah, C., Schroff, F., Belongie, S., Perona, P.: Caltech-ucsd birds 200. Technical report, California Institute of Technology (2010)
21. Maji, S., Kannala, J., Rahtu, E., Blaschko, M., Vedaldi, A.: Fine-grained visual classification of aircraft. Technical report, arXiv (2013)
22. : Oxford flower 102 `http://www.robots.ox.ac.uk/~vgg/data/flowers/`.
23. Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J.: LIBLINEAR: A library for large linear classification. The Journal of Machine Learning Research **9** (2008) 1871–1874

24. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR. (2005)
25. Perronnin, F., Dance, C.: Fisher kernels on visual vocabularies for image categorization. In: CVPR. (2007)
26. Bergamo, A., Torresani, L.: Meta-class features for large-scale object categorization on a budget. In: CVPR. (2012)
27. Perronnin, F., Liu, Y., Sánchez, J., Poirier, H.: Large-scale image retrieval with compressed fisher vectors. In: CVPR. (2010)
28. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: CVPR. Volume 2., IEEE (2006) 2169–2178