Estimating Nutritional Value From Food Images Based on Semantic Segmentation

Kyoko Sudo
NTT Media Intelligence Laboratories
1-1 Hikarinooka,Yokosuka-shi, Kanagawa 239-0847 JAPAN
sudo.kyoko@lab.ntt.co.jp

Jun Shimamura
NTT Media Intelligence Laboratories
1-1 Hikarinooka,Yokosuka-shi, Kanagawa 239-0847 JAPAN
shimamura.jun@lab.ntt.co.jp

Kazuhiko Murasaki
NTT Media Intelligence Laboratories
1-1 Hikarinooka,Yokosuka-shi, Kanagawa 239-0847 JAPAN
murasaki.kazuhiko@lab.ntt.co.jp

Yukinobu Taniguchi
NTT Media Intelligence Laboratories
1-1 Hikarinooka,Yokosuka-shi, Kanagawa 239-0847 JAPAN
taniguchi.yukinobu@lab.ntt.co.jp

UbiComp '14, September 13 - 17 2014, Seattle, WA, USA

http://dx.doi.org/10.1145/2638728.2641336

Abstract
Estimating the nutritional value of food based on image recognition is important to health support services employing mobile devices. The estimation accuracy can be improved by recognizing regions of food objects and ingredients contained in those regions. In this paper, we propose a method that estimates nutritional information based on segmentation and labeling of food regions of an image by adopting a semantic segmentation method, in which we consider recipes as corresponding sets of food images and ingredient labels. Any food object or ingredient in a test food image can be annotated as long as the ingredient is contained in a training food image, even if the menu containing the food image appears for the first time. Experimental results show that better estimation is achieved through regression analysis using ingredient labels associated with the segmented regions than when using the local feature of pixels as the predictor variable.

Author Keywords
recipe; food image; ingredients; image segmentation

ACM Classification Keywords
I.5.4. Pattern recognition: Applications

Introduction

In recent years, there have been various Internet based health care services, and they provide applications that aid in controlling calorie consumption or providing nutritional information. Maintaining a record of meals is effective in maintaining good health, and there have been studies on an application interfaces for this purpose [1]. It was shown that a system with an image based interface is more effective in maintaining a log than a text based interface [2]. An image assisted system can also provide services such as searching for recipes and obtaining information regarding nutrition as depicted in Fig. 1. Moreover, capturing images and sharing data on a network helps to support users and provide motivation to keep records. There have been studies on recognizing food images targeting these services including crowd-sourcing [3] and machine learning [4].

There are two main approaches for estimating the nutritional value from food images. One is to estimate the category of food and output information associated with the category. The other is to estimate the nutritional value directly from image features using regression analysis. Most conventional methods follow the first approach. In the first approach, a database search is performed based on the food name, and information is presented related to the food. However, in order to manage many types of food, the food names in the database should be finely categorized. Some databases of Internet applications cover all the menus from several large food chains. Here also, the interface is important to the search and selection of items from those large databases. However, it is still difficult to cover home-style cooked food, dishes served in restaurants that have original combinations of

Figure 1. System provides information on food and searches for recipes based on image recognition.

ingredients, or dishes that are decorated in the serving style of that particular restaurant. We also may not know what food name is the most appropriate for a specific dish. The second approach, based on regression analysis of image features such as a color histogram and local features is effective because generally nutrition is correlated with color. However, since this type of approach cannot provide other information than just numerical nutritional information, it is difficult to construct an interface that provides results to users in a satisfying manner. For these reasons, a method that can estimate the nutritional value accurately, and obtain a middle state of image recognition is desirable.

In this paper we present an algorithm that estimates the nutritional value through semantic segmentation. We extract a label histogram, which expresses the frequency of occurrence of text tags of ingredients, based on the results of partitioning of an image into regions that are associated with ingredients. Then we estimate the nutritional value using regression analysis based on the label histogram. In the experiments, we

compare the results of the regression analysis using image features to those using a label histogram.

Related work and proposed approach

One approach to construct a system to search for nutritional information from a photo query is visual segmentation and ingredient recognition. By extracting visually areas of segmented ingredients and recognizing the types of ingredients, we can estimate the nutritional information even when the serving style or the food itself is original. The method proposed by Eskin *et al.* estimates the kinds of food based on color information by dividing food areas into small categories such as meat, vegetables and fruit[5]. The method proposed by He *et al.* estimates the weight of food using geometric information of a shape template, as well as the kinds of food based on color information[6]. In an ingredient recognition approach, taking advantage of the co-occurrence of multiple ingredients is also efficient. The method proposed by Yang *et al.* recognizes the kind of fast food using the co-occurrence of multiple kinds of food items[7]. Their database has various types of fast food, but there are not many variations in the ingredients. So, it is difficult to adapt these methods to recognize foods that have many kinds of ingredients and various combinations of them.

We propose a method based on semantic segmentation, in which the segmentation results are obtained with labels indicating to what kind of food that the segmented area is associated. We adopt an annotation method that uses spatial co-occurrence and co-occurrence of text tags. Based on the annotation results, we obtain the frequency of occurrence of the text tags each of which is related to a segmented region. Then we estimate certain kind of nutritional value such as the number of calories based on

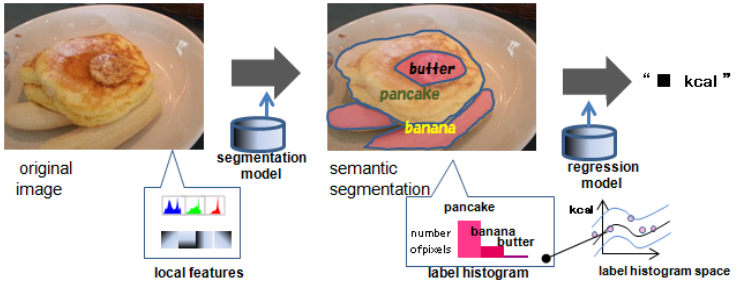

Figure 2. Proposed approach for calorie estimation based on semantic segmentation of images and regression analysis.

value such as the number of calories based on regression analysis with the frequency of occurrence of text tags as the predictor variable (Figure 2).

The proposed algorithm has two main steps, each of which has a statistical model. The first step is the semantic segmentation step, in which the food image is segmented and labeled. The output of this step is a label histogram, that shows the frequency of occurrence of the ingredient labels. This process has a semantic segmentation model using which visual features are related to ingredient labels. The second step is to estimate the nutritional value from the label frequency histogram using a label-nutrition regression model. We use regression analysis of the label frequency histogram and the numerical data of the nutritional information, which is the output of this step, to train the model.

Semantic segmentation of food image

Examples of groups of food	Examples of ingredient names
Fish	Salmon, tuna, sole, crab
Meat	Beef, chicken, pork, bacon
Egg	Egg, whole egg,yolk, egg white
Colored vegetables	Spinach, carrot, paprika(bell pepper)

Table 1. Example of food groups and the ingredient names which belong to each group.

Semantic segmentation is an image annotation technique that is mostly applied to scene understanding. Labels are predicted for each pixel of an image so that the image is divided into object regions. Each region consists of pixels associated with the same label. We adapt the semantic segmentation to a food image to obtain the regions of the food object with ingredient labels.

- Semi-supervised semantic segmentation

First, we extract SIFT and color (Lab) features from pixels randomly selected in the image for all training samples. They are then coded using k-means clustering. The visual features are obtained as Bag-of-words for each super-pixel. A super-pixel is a small region obtained by partitioning an image based on edge and local features where there are no boundaries between different objects. We divide an image into super-pixels based on Liu's method [8].

To train the semantic segmentation model, we have a copious amount of recipe data with corresponding sets of food images and ingredient labels. Since the ingredient labels are not associated with the pixel or the object region of the image but to the whole image, the data are considered as a weakly labeled training set. In this paper, we model the training set as a Generalized Multi-Image Model (GMIM) [9]. We train this model to yield the same label for pixels with similar appearance in different images as well as pixels spatially close to an image based on an incremental optimization technique. We give each super-pixel a visual feature code as an initial label and train iteratively.

- Food groups and label histogram

There are many ingredients in a recipe database, and many ingredients have multiple popular names. Since semantic segmentation uses co-occurrence of labels not only in one image but also between multiple images, it is important to select words and quantize text code to train the model efficiently. We generate classes of food groups. Examples of the classes are given in Table 1. All words of ingredients belong to a food group. We set not only the "food" group, i.e., the groups for the edible region, but also a "background" group. This makes it possible to partition an image into food regions and non-food regions without manual segmentation. A label histogram of an image is obtained by counting food labels associated to super-pixels by semantic segmentation. The dimension of label histograms is the number of food groups. In this work, we categorize ingredients into 29 food groups, so the number of dimensions of label histograms is set to 29.

Regression from label histogram

The results of semantic segmentation show that each pixel has an ingredient label. We estimate the nutritional value from the label frequency histogram based on a support vector regression (SVR) model.

SVR is an algorithm that adapts a Support Vector Machine to estimate the regression parameters by mapping a feature into a higher-dimensional space using a kernel function. Although label histograms are sparse and outliers are contained in them due to rareness of the menu itself or failure of the segmentation process, we expect that the SVR model will become robust through training.

Experiment

In the database used in the experiments, we have 2500 recipes from a cooking recipe site. Of those, 1250 recipes are used for training and the rest are used for testing. Each data set comprises a food image, an ingredient list, and cooking directions. We also obtained nutritional data for each recipe to train the estimation model. The nutritional data includes items such as the number of calories, vitamins, and sodium levels. In the experiments, we use calories, fat, protein, calcium, sodium, and iron. We compared the results of SVR nutritional estimation using the label histogram feature as a prediction variable to those of SVR using the color histogram feature as a prediction variable. The RBF kernel is used for SVR. The number of variables in the color histogram is 75.

An example of the segmentation results of an image is shown in Fig. 3. The color indicates the food label. The labels for meat, fish, and vegetables correspond fairly closely with the super-pixels. Also, the background region is well estimated. Although there are more than a few labeling errors, we use the label histogram as the feature even though it allows some labeling errors. Examples are shown in Fig. 4. The results of the top three images show that foods that are shaped or decorated with fruit or vegetables yield better results when using our method than using simple color histogram. This is because the colors of the fruits or vegetables have a greater influence on the color features than high calories but brightly colored ingredients mainly contain carbohydrates or proteins. In the lower left image, the areas of cakes are associated with the label “sugar” and not with “egg” or “flour,” so the calorie estimate is very high. This is because a super-pixel currently has only a single label. There is the possibility that different patterns of labels are contained in the same super-pixel. Improvements that would allow super-pixels to have multiple labels are for future work. In the lower middle image, the food labels are added to even background areas that

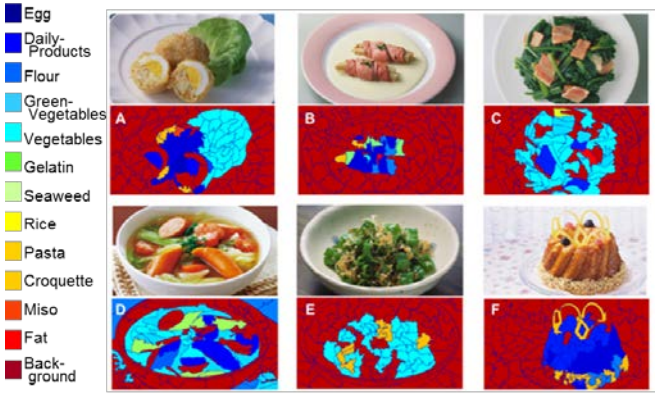

Figure 3. Examples of semantic segmentation results.

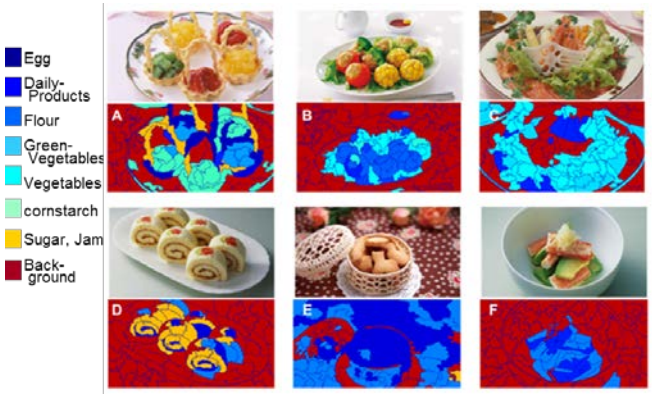

Figure 4. Examples of semantic segmentation results. The pairs in the top two rows are examples where the label histogram yields better results than the color histogram.

have many textures, which results in a large error in estimation. The lower right image is a small bowl of Japanese food, which represents a difficult case to estimate nutritional information because there are few instances that are similar in the recipe data.

The nutritional estimation results based on regression analysis are given in Table 2. In this experiment, regression analysis using a label histogram obtained by segmentation based on image features for predictor variables yields better results than directly using image features as predictor variables in terms of the error rate of calories, fat, and protein. The results are almost the same or less compared to the color histogram for calcium, sodium, and iron, which are strongly related to color.

	Average error of nutrition prediction	
Nutrition	Color +SVR	Label +SVR
Calorie	33.6%	31.8%
Fat	38.6%	37.6%
Protain	38.8%	37.9%
Calcium	40.6%	40.2%
Sodium	37.9%	37.7%
Iron	37.2%	38.7%

Table 2. Results of average error of nutrition prediction based on regression using color histogram or ingredient labels as predictor variants.

Conclusion

We proposed a method that estimates the nutritional information based on an image using semantic segmentation. By estimating the regions and ingredient labels, the proposed method enables us to present nutritional information to users in a more satisfactory manner. Moreover, it is possible to estimate the nutritional information of unknown recipes even when the combination of ingredients is original or new, as long as the models for this method have learned recipes that include the same ingredients.

References

[1] A. Andrew, G. Borriello, J. Fogarty. Simplifying Mobile Phone Food Diaries. In *Proc. Int. Conf. on Pervasive Health*, 2013, 260-263.

[2] K. Aizawa, K. Maeda, M. Ogawa, Y. Sato, M. Kasamatsu, K. Waki and H. Takimoto, Comparative Study of the Routine Daily Usability of FoodLog: A smartphone-based food recording tool associated by image retrieval. *Journal of Diabetes Science and Technology,* 2014.

[3] J. Noronha, E. Hysen, H. Zhang and K. Z. Gajos, PlateMate: Crowdsourcing nutrition analysis from food photographs. In *Proc. the 24th annual ACM symposium on User interface software and technology (UIST), 2011.*

[4] Y. Kawano and K. Yanai, Real-Time Mobile Food Recognition System, In *Proc. CVPR2013 Mobile Vision Workshop,*1-7*.*

[5] Y. Eskin and A. Mihailidis, An intelligent nutritional assessment system. In *Proc. AAAI Fall Symposium2012.*

[6] Y. He, C. Xu, N. Khanna, C.J.Boushey and E.J. Delp, Food image analysis: Segmentation, identification and weight estimation. In Proc. *ICME2013*, 1-6.

[7] S. Yang, M. Chen, D. Pomerleau and R. Sukthankar, Food recognition using statistics of pairwise local features. In *Proc. CVPR 2010,* 2249-2256*.*

[8] M.Y. Liu, O. Tuzel, S. Ramalingam and R. Chellappa, Entropy Rate Superpixel Segmentation. In *Proc. CVPR 2011,* 2097-2104.

[9] A. Vezhnevets, V. Ferrari and J. M. Buhmann, Weakly Supervised Structured Output Learning for Semantic Segmentation. In *Proc. CVPR 2012,* 845-852.