CONTEXT BASED FOOD IMAGE ANALYSIS

Ye He, [1] *Chang Xu,* [1] *Nitin Khanna,* [2] *Carol J. Boushey,* [3] *and Edward J. Delp* [1]

[1] School of Electrical and Computer Engineering, Purdue University
[2] Department of Electronics and Communication Engineering, Graphic Era University, Dehradun, India
[3] Cancer Epidemiology Program, University of Hawaii Cancer Center

ABSTRACT

We are developing a dietary assessment system that records daily food intake through the use of food images. Recognizing food in an image is difficult due to large visual variance with respect to eating or preparation conditions. This task becomes even more challenging when different foods have similar visual appearance. In this paper we propose to incorporate two types of contextual dietary information, food co-occurrence patterns and personalized learning models, in food image analysis to reduce ambiguity in food visual appearance and improve food recognition accuracy. We evaluate our model on 1453 food images acquired by 45 participants in natural eating conditions. The result shows that incorporating contextual dietary information improves the food categorization accuracy by about 10%.

***Index Terms*—** Dietary Assessment, Contextual Information, Food Recognition, Image Segmentation

1. INTRODUCTION

Image classification is a challenging problem due to appearance variabilities caused for example by non-rigidity, background clutter, scale or lighting conditions. Designing good descriptors and classification methods for image classification has been extensively investigated in the literature[1, 2, 3, 4, 5]. The use of contextual information has gained more attention in psychology and computer vision with respect to its effects on visual search, localization and recognition performance [6, 7, 8]. Context in this setting refers to any information that is not directly produced by the visual appearance of an object in the image. Integrating contextual information with visual information in an object categorization framework is a challenging task. Several classifiers, such as boosting [9, 10] and Logistic Regression [11], have been developed that exploits contextual information in order to maximize the classification performance. Conditional random fields (CRF) [12, 13] have been used to increase object label agreement in object classification using contextual and visual features [14, 15].

We are developing an image analysis system to estimate the foods consumed at an eating occasion from food images acquired by mobile telephones[16]. The goal of our system is to locate and identify foods within an image. Traditional approaches to object categorization use visual features, such as color, texture and edge, as the main source of information for recognizing object classes in images. Visual features can capture variability in object classes up to a certain extent. Our previous work on food image analysis has shown that there are several issues that need to be addressed in an image-assisted dietary assessment system [16, 17]. These include the inability to distinguish visually similar food items, e.g. diet coke vs. coke, nonfat milk vs. 2% milk, solely based on their appearance in the image. Another issue is the selection of optimal training data for different classes. There is a large number of food combinations in the world and increasing the number of food training classes could cause a drastic increase in the food classification error.

We intend to improve our system by investigating methods for analyzing images and "learning" a user's diet that exploits contextual information which the user supplies to the system either explicitly or implicitly. Using contextual information in a mobile dietary system, henceforth called as "contextual dietary information", refers to the data that yields information about a users diet or can be used for diet planning. Contextual dietary information is incorporated as "side" information for the classifier. If an individual has a repeated dietary behavior (e.g. particular types of foods consumed or the number of foods consumed in a particular eating occasion), the classifier can dynamically select the food classes in the training dataset that match the eating pattern.

2. CONTEXTUAL DIETARY INFORMATION

Eating is more than just acquiring fuel for the body, it is a social experience and in many cases the how, where and with whom we eat can influence our diet. We want to exploit contextual dietary information that can be acquired when users have mobile telephones to take images of their eating occa-

This work was sponsored by grants from the National Institutes of Health under grants NIDDK 1R01DK073711-01A1 and NCI 1U01CA130784-01. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Institutes of Health. Address all correspondence to Edward J. Delp, ace@ecn.purdue.edu or see www.tadaproject.org.

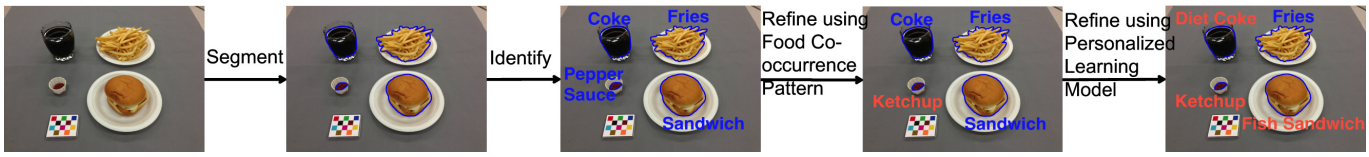

Fig. 1. Illustration of an idealized context based food segmentation and identification system. First, the input image is segmented. Then each segment is labeled by the identification system using only the visual features. Next, the food co-occurrence pattern is used to correct some of the labels. Finally, a personalized learning model is considered to provide further disambiguation based on the user's eating habit.

sions. This information can be either explicitly or implicitly acquired. In this paper we propose to incorporate two types of contextual dietary information, food co-occurrence patterns and personalized learning models. As an illustration of this idea, consider the example in Figure 1. In the scene of an eating occasion, a food image is first acquired by a user. The image is processed through a segmentation and food recognition system which generates an ordered list of possible food labels. For each segment, only the best match is shown in Figure 1. Without incorporating any contextual dietary information, "Coke", "Fries", "Pepper Sauce", and "Sandwich" would be the final labels; however, in context, these labels are not satisfactory. Considering the food co-occurrence pattern, "Pepper Sauce" is probably mis-labeled due to the similarity in visual appearance with "Ketchup". We could further improve the food recognition results by considering the personal eating habits of the user.

2.1. Food Segmentation and Identification

In our food segmentation and identification system, once a food image is acquired, image segmentation is used to locate object boundaries for food items in the image. To obtain a stable segmentation, we combine image segmentation and classification using a segmentation refinement step in which the classification feedback can be used to refine image segmentation until maximal classification confidence has been achieved [17, 18]. A flow chart of our food segmentation and identification system is shown in Figure 2.

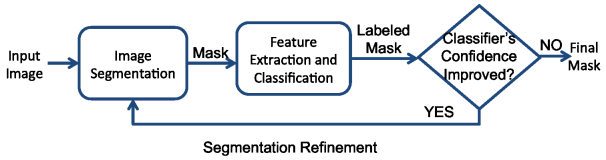

Fig. 2. Our image segmentation and food identification system.

In this work we adopt the framework of [19] to generate a shortlist of initial segmentations. Two regions A and B are segmented if the difference between the two regions is large relative to the internal difference within at least one of the two regions. The degree to which the difference between regions must be larger than minimum internal difference is controlled by a threshold k:

$$minInt(A,B) = min(Int(A) + \frac{k}{|A|}, Int(B) + \frac{k}{|B|}) \quad (1)$$

where $Int(A)$ is the internal difference of A; $|A|$ denotes the size of A. The input parameter k roughly controls the size of the regions in the resulting segmentation. Smaller values of k yield smaller regions and favor over-segmentation. In our implementation we initially set $k = 150$.

For each segmented region, we extract the color and texture features as described in [20, 21] and assign a category label to the segment based on a majority vote rule of the nearest neighbors. In our implementation, each segment is assigned to 4 food categories with the top 4 largest category labeling confidence score. The original image is sent back to the user to confirm the top food category, select from the next three, or designate a food category from a larger list of choices.

2.2. Food Co-Occurrence Patterns

A food co-occurrence pattern describes the likelihood of food combinations and is their mutual probability of existing together in a single eating occasion. In order to integrate a food co-occurrence pattern into our food segmentation and identification framework, machine learning techniques are used since they provide efficient and powerful probabilistic methods. Among various approaches, graphical models provide a simple way to visualize the structure of a probabilistic model. In this paper we use graphical models to incorporate contextual dietary features as a post-processing stage to promote agreement between the segment labels.

Using our image segmentation and food identification system, an input food image I is segmented into multiple segments $S_1, ..., S_N$ and each segment is assigned four food labels $f_1, ..., f_4$. A confidence score $p(f_k|S_n)$ measuring the probability that the food label f_k matches the segment S_n is estimated based on the distance of the visual features between the segmented region and the training data. We want to adjust the food labels to achieve maximal global contextual agreement with respect to the food co-occurrence pattern given the constraints of the segments' visual features. Since the number of food segments in an eating occasion is relatively

small, we use a fully connected undirected graph between all segments [22]. For example, the food image in Figure 1 is represented by the fully connected undirected graph in Figure 3. Each segment in the image is represented by a clique. For every subset of nodes (that is, the food labels for the segment) in Clique n, we have an associated probability function $p(f_k|S_n)$. We adjust the probability of each node by considering its association with all nodes in other cliques:

$$p'(f_k|S_n) = \frac{p(f_k|S_n)A(f,S)}{Z(\phi, S_1, ..., S_N)} \quad (2)$$

$$A(f,S) = \exp(\sum_{i=0}^{4} \sum_{j=1, j\neq n}^{N} \phi(f_{k,n}, f_{i,j}) p(f_i|S_j)) \quad (3)$$

where $f_{i,j}$ denotes node i in Clique j; $\phi(f_{k,n}, f_{i,j})$ is the co-occurrence probability between node $f_{k,n}$ and node $f_{i,j}$; $Z(\cdot)$ is the normalization constant obtained by summing the numerator over all nodes in the same clique.

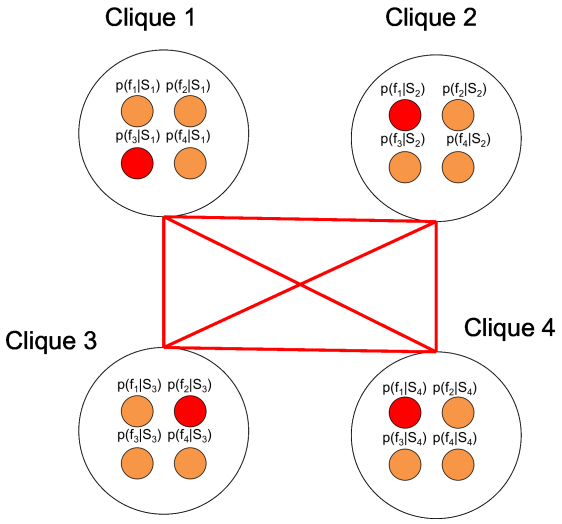

Fig. 3. An input food image is modeled by a fully connected undirected graph. A clique with four nodes represents a segment with four potential food labels.

To estimate the co-occurrence probability, we construct food co-occurrence matrices. These are symmetric, nonnegative matrices that contain the food co-occurrence probability among food labels in the training set of the database. Figure 4 illustrates the structure and content of a food co-occurrence matrix. The entry (i, j) in the matrix is the probability that food i appears in the same image as food j. The diagonal entry (i, i) corresponds to the probability that food label i appears in different locations of the same image.

2.3. Personalized Learning Models

As humans we all eat different foods, so incorporating an individual learning model into our system would allow the classifier to only select among the classes pertinent to an individual. Since foods eaten by an individual are fewer than foods eaten by an entire community or population, this reduces the class space and increases classification accuracy. According to data from nationally representative cross-sectional surveys collected between 1999-2008, <1,000 foods capture 99% of the foods consumed in the United States for individuals between 11 y and 65 y and the number of foods consumed by each person necessary for comprehensive dietary assessment is far less [23]. By building a personalized learning model the classifier can pre-select a subset of potential food classes that are commonly eaten by the individual.

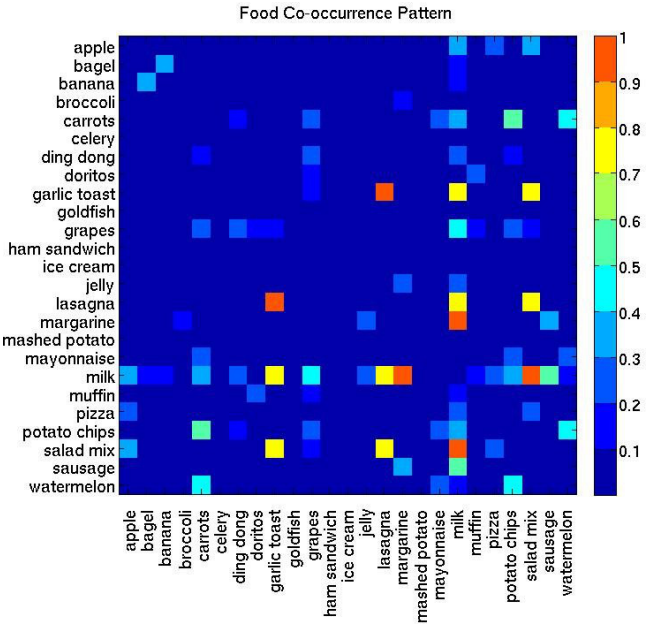

Fig. 4. An example of a food co-occurrence pattern.

Figure 5 shows the consumption frequency of a subset of foods eaten by participants in real life conditions. The experiment consists of 45 users. Each participant was asked to acquire a pair of before and after eating occasion images at each eating occasion for a 7-day period. In total 1453 before eating images were analyzed using the methods described in [17, 18] and 56 commonly eaten food items were classified. The largest number of images acquired by a participant was 88 images and the smallest was 23 images. The training data for these experiments were obtained beforehand by acquiring images in a series of food imaging sessions. We obtained training data for the foods that were to be commonly eaten by users in order to create a learning model for the classifier. From Figure 5, we can see the difference in individual eating habits. For example, user 11 drank milk quite frequently while user 3 rarely drank milk.

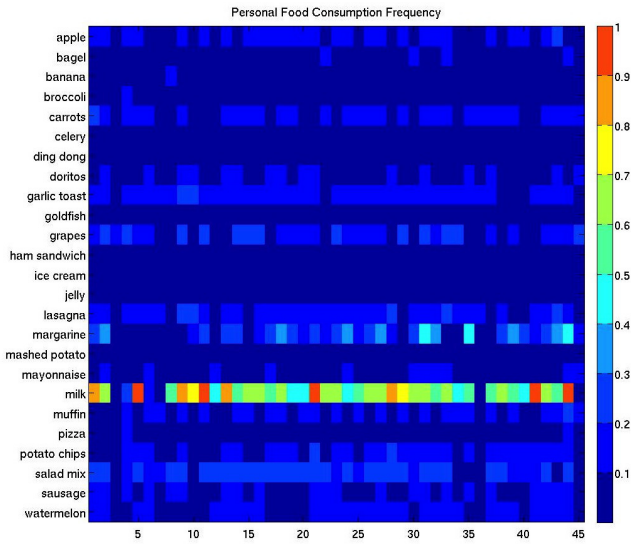

Fig. 5. An example of food consumption frequency by different users. Horizontal axis represents different users and vertical axis represents different food items.

3. EXPERIMENTAL RESULTS

Our proposed image segmentation and identification system with and without incorporating contextual dietary information were tested separately on 1453 food images with 56 unique food items (as shown in Figure 6). The food images were acquired by 45 users in natural eating conditions (also referred to "free-living" or "community-dwelling" such as home and on the go). The food identification accuracy is defined as:

$$accuracy = \frac{TP}{TP + FP + TN} \tag{4}$$

where TP indicates True Positives (correctly detected food segments); FP indicates False Positives (incorrectly detected food segments or misidentified foods); TN indicates True Negatives (food not detected). The identification accuracy of all the images using only visual features from the classifiers was 34%. It should be emphasized that in some of our other user studies where we controlled the foods the users consumed we achieved classification accuracy of more than 75%[21]. As indicated above the study used for this paper was a more "typical" use scenario. These results clearly indicate that food image analysis is a difficult problem and requires the use of more information than what appears in the image.

To improve the identification accuracy, we used the co-occurrence pattern to adjust the food labels to maximize their global contextual agreement. In our implementation, for each segmented area in an image the four most probable food labels and their corresponding probabilities were estimated. Then the probabilities were re-calculated using the food co-occurrence pattern according to Equation 2. The label with the largest probability was used as the final assignment for the segmented area. After the labeling process there were still several foods that we could not distinguish, e.g., diet coke vs. regular coke, nonfat milk vs. 2% milk. The personalized learning model was used to further refine the food labels. For example, we preferentially labeled a soda as diet coke after our personalized learning model learned this choice relative to an individual's diet. The food co-occurrence matrix and personalized learning database are updated as soon as new food images are received and this information will be used to identify future food images when they appear again in the user's diet.

Fig. 6. A list of all the food items used in the experiment.

After incorporating the contextual dietary information, the food identification accuracy increased to 44% for 56 food classes. This identification accuracy was calculated based on automatic segmentation results for food images acquired in natural eating conditions. We will further increase the identification accuracy by improving the image segmentation to localize food items. We feel that the 10% increase in accuracy indicates that our contextual models are promising and further investigation is warranted.

4. CONCLUSION AND FUTURE WORK

In this paper, we incorporated contextual dietary information in food image segmentation and identification as a post-processing step for aiding in food categorization/labeling. We used graphical models to incorporate the food co-occurrence pattern and personalized learning to help disambiguate food appearance. Experimental results showed that the food categorization accuracy was improved by 10%. To further improve the food categorization accuracy, we plan to explore other types of contextual information, such as geolocation and date and time of eating occasions.

5. REFERENCES

[1] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, October 2004.

[2] A. Bosch, A. Zisserman, and X. Muoz, "Image classifi-

cation using random forests and ferns," *Proceedings of the International Conference on Computer Vision*, Rio de Janeiro, Brazil, October 2007, pp. 1–8.

[3] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, New York, NY, June 2006, pp. 2169–2178.

[4] H. Zhang, A. C. Berg, M. Maire, and J. Malik, "Svm-knn: Discriminative nearest neighbor classification for visual category recognition," *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, New York, NY, June 2006, pp. 2126–2136.

[5] J. Zhang, M. Marszałek, 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, June 2007.

[6] I. Biederman, R. Mezzanotte, and J. Rabinowitz, "Scene perception: detecting and judging objects undergoing relational violations," *Cognitve Psychology*, vol. 14, no. 2, pp. 143 – 177, 1982.

[7] M. Bar and S. Ullman, "Spatial context in recognition," *Perception*, vol. 25, pp. 343 – 352, 1993.

[8] B. McFee, C. Galleguillos, and G. Lanckriet, "Contextual object localization with multiple kernel nearest-neighbor," *IEEE Transactions on Image Processing*, vol. 20, no. 2, pp. 570 – 585, 2011.

[9] M. Fink and P. Perona, "Mutual boosting for contextual inference," *Proceeding of Advances in Neural Information Processing Systems*, Vancouver, Canada, December 2003, pp. 1515–1522.

[10] L. Wolf and S. Bileschi, "A critical view of context," *International Journal of Computer Vision*, vol. 69, no. 2, pp. 251 – 261, August 2006.

[11] A. Torralba, K. P. Murphy, and W. T. Freeman, "Using the forest to see the trees: exploiting context for visual object detection and localization," *Communications of the ACM*, vol. 53, no. 3, pp. 107 – 114, March 2010.

[12] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," *Proceedings of the International Conference on Machine Learning*, Williamstown, WA, 2001, pp. 282 – 289.

[13] C. Sutton and A. McCallum, "An introduction to conditional random fields for relational learning," *Introduction to Statistical Relational Learning, MIT Press*, 2006.

[14] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie, "Objects in context," *Proceedings of the International Conference on Computer Vision*, Rio de Janeiro, Brazil, October 2007, pp. 1 – 8.

[15] C. Galleguillos and S. Belongie, "Context based object categorization: A critical survey," *Computer Vision and Image Understanding*, vol. 114, pp. 712 – 722, 2010.

[16] F. Zhu, M. Bosch, I. Woo, S. Kim, C. Boushey, D. Ebert, and E. Delp, "The use of mobile devices in aiding dietary assessment and evaluation," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 4, pp. 756 –766, August 2010.

[17] F. Zhu, M. Bosch, N. Khanna, C. Boushey, and E. Delp, "Multilevel segmentation for food classification in dietary assessment," *Proceedings of the 7th International Symposium on Image and Signal Processing and Analysis*, Dubrovnik, Croatia, September 2011, pp. 337–342.

[18] Y. He, N. Khanna, C. Boushey, and E. Delp, "Snakes assisted food image segmentation," *Proceedings of IEEE International Workshop on Multimedia Signal Processing*, Banff, Canada, September 2012, pp. 181–185.

[19] P. Felzenszwalb and D. Huttenlocher, "Image segmentation using local variation," *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Washington, DC, USA, June 1998, pp. 98 –104.

[20] B. Manjunath, J.-R. Ohm, V. Vasudevan, and A. Yamada, "Color and texture descriptors," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 11, no. 6, pp. 703 –715, June 2001.

[21] M. Bosch, F. Zhu, N. Khanna, C. Boushey, and E. Delp, "Combining global and local features for food identification and dietary assessment," *Proceedings of the International Conference on Image Processing*, Brussels, Belgium, September 2011, pp. 1789 – 1792.

[22] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, "An introduction to variational methods for graphical models," *Machine Learning*, vol. 37, no. 2, pp. 183–233, November 1999.

[23] H. A. Eicher-Miller and C. J. Boushey, "The most frequently reported foods and beverages differ by age among participants of nhanes 1999-2008," *The Journal of the Federation of American Societies for Experimental Biology (FASEB)*, vol. 26, p. 256.1, March 2012.