The Use of Temporal Information in Food Image Analysis

Yu Wang[1(✉)], Ye He[2], Fengqing Zhu[1], Carol Boushey[3], and Edward Delp[1]

[1] School of Electrical and Computer Engineering, Purdue University, West Lafayette, Indiana, USA
wang1317@purdue.edu
[2] Google, Mountain View, California, USA
[3] Cancer Epidemiology Program, University of Hawaii Cancer Center, Honolulu, Hawaii, USA

Abstract. We have developed a dietary assessment system that uses food images captured by a mobile device. Food identification is a crucial component of our system. Achieving a high classification rates is challenging due to the large number of food categories and variability in food appearance. In this paper, we propose to improve food classification by incorporating temporal information. We employ recursive Bayesian estimation to incrementally learn from a person's eating history. We show an improvement of food classification accuracy by 11% can be achieved.

1 Introduction

Mobile devices will transform the healthcare industry by increasing accessibility to quality care and wellness management. Dietary intake provides valuable insights for fitness monitoring as well as mounting intervention programs for chronic diseases. Accurate methods to assess food and nutrient intake are essential [7,15]. We have developed a dietary assessment system, known as the mobile Food Record (mFR) [1,18], to automatically estimate food type, volume, nutrients, and energy from a food image captured by a mobile device [18,19]. The mFR system consists of: a web-based user interface, a mobile application and a backend system including a computational server and an associated database system [2,18].

To achieve high classification accuracy in food images is challenging due to lighting and pose variations, background noise and occlusion. One type of food can have various serving styles (different portion sizes, distinct appearance). As a result, the use of contextual information may reduce the complexity of food image analysis. "Context" refers to any prior knowledge that is not derived from the image pixel values [4]. The use of contextual information has gained attention in psychology and computer vision with respect to its effects on visual search,

E. Delp—This work was sponsored by grant from the National Institutes of Health under grant NIEH/NIH 2R01ES012459-06. Address all correspondence to E. J. Delp: ace@ecn.purdue.edu or see www.tadaproject.org.

V. Murino et al. (Eds.): ICIAP 2015 Workshops, LNCS 9281, pp. 317–325, 2015.
DOI: 10.1007/978-3-319-23222-5_39

localization and recognition [4,14,16]. There has been work in using contextual information in food image analysis. Matsuda et al [13] proposed to use manifold ranking method to improve food classification rate using food co-occurrence statistics. Beijbom et al [5] made use of geographic location as context and focused on identifying foods in restaurants. In previous work [9] we incorporated two types of contextual knowledge, food co-occurrence patterns and an individual's food consumption frequency for a week.

In this paper, we propose to incorporate temporal information to learn a person's dietary pattern based on a recursive Bayesian model to improve food classification accuracy. The learning process is achieved by incorporating user feedback in the food classification. The user feedback consists of confirmed, modified, or added food labels based on the food image analysis.

2 Food Image Analysis

2.1 Image Acquisition and User Feedback

In our mFR system, a mobile application is used to capture a pair of before and after meal images at each eating occasion [18]. The images are sent to the server for automatic image analysis. Results are sent back to the user for confirmation and review using the mobile application [2]. In this paper, we used food images collected from one of our dietary assessment studies, where 45 participants were asked to acquire a pair of before and after meal images at each eating occasion for roughly 7 days. A total of 1453 food images were analyzed classifying 56 commonly eaten food items using the methods described in [10,11,19]. Figure 1 shows the food consumption pattern of a subset of foods from our data. Each square in Fig.1 indicates the consumption frequency of a particular food, λ_i, for a participant, S_j, where the consumption frequency is defined as followed,

$$Freq(\lambda_i^{S_j}) = \frac{N_i}{N_{\text{total}}} \quad . \tag{1}$$

N_i is the number of times that S_j has consumed λ_i and N_{total} is the total number of food items that S_j has consumed. The food consumption pattern for S_j is $[Freq(\lambda_1^{S_j}), \ldots, Freq(\lambda_n^{S_j})]$, where n is the total number of food categories.

To evaluate our learning model, we manually selected participants with similar food consumption patterns to build personalized eating datasets for a month. We measure the similarity using Euclidean distance between each food consumption pattern and used *K-means* to find clusters. For example, one of the personalized eating dataset contains food images from participant 14, 17, 20 and 32, which can be treated as images from a single participant for a month. As illustrated in Fig.1, participant 14, 17, 20 and 32 all show relatively high consumption frequency of milk, salad mix and lasagna. We constructed three separate datasets, each features a different food consumption style and contains approximately 120 images. We labeled them as *Dataset 1, 2 and 3* from *User 1, 2 and 3*.

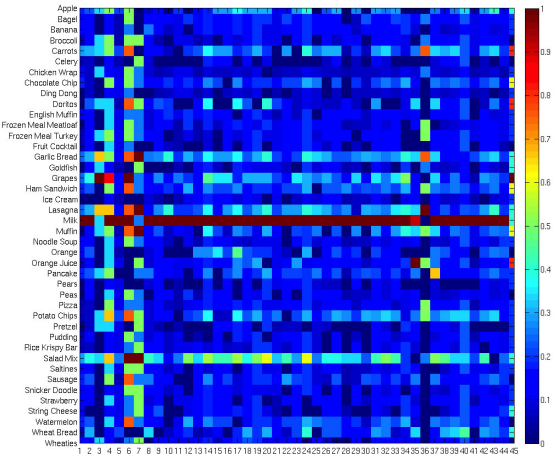


Fig. 1. A food consumption pattern. Horizontal axis represents the participants and vertical axis represents food items.

2.2 Image Segmentation and Food Identification

We used graph based segmentation using local variation [8,10]. The internal difference of a segmented region is defined to be the largest weight in the minimum spanning tree while the difference between two segmented regions is defined to be the minimum weight edge connecting the two regions [8]. The ratio of the region difference to the internal difference within at least one of the two regions determines whether two regions are segmented or not. The degree to which the difference between regions must be larger than minimum internal difference is controlled by a threshold k, where k roughly controls the size of the regions in the resulting segmentation. In our experiments, k was initially set to 150. For each segment, a set of color, texture and local region features are extracted and classified using k-Nearest Neighbor (KNN), vocabulary trees and Support Vector Machines (SVM) [12,19]. Finally we combine the decisions from all the feature channels using a majority vote rule.

3 Temporal Context

Temporal context in this paper refers to which days a person eats a particular food. An example might be that person A drinks 2% milk everyday while person B never has milk but eats Greek yogurt everyday. Such time-related eating habits can be of great help when designing a personalized training model because it allows the classifier to only select among the food classes relevant to an individual's dietary pattern [9]. According to representative cross-sectional surveys collected between 1999-2008, less than 1,000 foods capture 99% of the foods consumed in the United States for individuals between 11-65 years old and that the number of foods consumed by each person is far less [6].

3.1 Recursive Bayesian Model

In this paper, we use recursive Bayesian estimation to incrementally learn a person's dietary pattern [3,17]. We model whether a person, S_j eats a particular food, λ_i as an independent Bernoulli trial,

$$W = \begin{cases} 1, X \\ 0, 1 - X \end{cases} .$$

where $W = 1, X$ represents S_j eats λ_i with a possibility, X, and X is assumed to follow a Gaussian-like distribution with the support from 0 to 1.

We would like to estimate the probability, P_{λ_i}, that a person, S_j, will eat a particular food, λ on the next day given the past. Let $p_{\lambda_i}(x^k)$ be the probability density function (PDF) representing S_j eats λ_i on the k^{th} day, and z^k be the observation whether S_j eats λ_i on the k^{th} day. In our case, the observation, z^k is obtained from the user feedback in the mFR. The following equations describe the posteriori update step in the recursive Bayesian network,

$$p_{\lambda_i}(x^k|z^{1:k}) = \frac{p_{\lambda_i}(z^k|x^k)p_{\lambda_i}(x^k|z^{1:k-1})}{p_{\lambda_i}(z^k|z^{1:k-1})} = \frac{\text{likelihood} \times \text{prior}}{\text{normalization term}} . \quad (2)$$

Initially, $p_{\lambda_i}(x^1)$ is assumed to follow a Gaussian-like distribution centered at 0.5 with unit variance. The likelihood and prior PDFs are updated according to the user's feedback. If the user eats λ on the k^{th} day, $p_{\lambda_i}(z^k|x^k)$ becomes the Gaussian-like distribution centered at 1 with unit variance, otherwise the distribution centers at 0. $p_{\lambda_i}(x^k|z^{1:k})$ is used to predict $p_{\lambda_i}(x^{k+1}|z^{1:k})$ and the PDF is computed by multiplying the likelihood and prior followed by normalization between 0 and 1. On the $k+1^{\text{th}}$ day, the optimal estimate of P_{λ_i} is computed as $P_{\lambda_i} = \arg\max_x p_{\lambda_i}(x^k|z^{1:k})$.

For all the foods in the training dataset, we have a set of probabilities,

$$\Psi^{k+1} = \left[P_{\lambda_1}^{k+1}, \ldots, P_{\lambda_n}^{k+1}\right]^{\mathrm{T}} .$$

where n is the total number of food categories. We further define the context-based confidence scores (CCS) to be:

$$\Phi^{k+1} = \left[\phi_{\lambda_1}^{k+1}, \ldots, \phi_{\lambda_n}^{k+1}\right]^{\mathrm{T}} = \left[\omega P_{\lambda_1}^{k+1}, \ldots, \omega P_{\lambda_n}^{k+1}\right]^{\mathrm{T}} . \quad (3)$$

where ω controls the trust weight we assigned to the context-based decisions (more details in Sect.3.2).

3.2 Decision Fusion

So far, we have obtained the confidence scores from both image analysis based on multiple feature channels and temporal context. From the image analysis, a set of candidate classes was assigned to each segmented region, S_q, associated with the corresponding confidence scores for each food class:

$$\Lambda_{\text{cand}}^{\text{auto}} = \left[\lambda_1^{\text{auto}}, \ldots, \lambda_n^{\text{auto}}\right]^{\mathrm{T}} , \ \Phi_{\text{cand}}^{\text{auto}} = \left[\phi_1^{\text{auto}}, \ldots, \phi_n^{\text{auto}}\right]^{\mathrm{T}} .$$

where $\Lambda_{\text{cand}}^{\text{auto}}$ represents the candidate set, $\Phi_{\text{cand}}^{\text{auto}}$ indicates the corresponding confidence scores of n different food classes in the training dataset, i.e., $n = 56$ food classes in our dataset. From (3), we know the context-based decision for each food class on a certain day. To combine the above two source of decisions, we used a strategy of maximum confidence score. The final score is determined as:

$$\Phi_{\text{cand}}^{\text{final}} = \left[\phi_1^{\text{auto}}, \ldots, \phi_n^{\text{auto}}\right]^{\text{T}} + \left[\omega P_{\lambda_1} \ldots \omega P_{\lambda_n}\right]^{\text{T}} \quad . \tag{4}$$

ω, also in (3), is set to be $1/h$ of the maximum automatic analysis based confidence score: $\omega = \frac{1}{h} \max(\Phi_{\text{cand}}^{\text{auto}})$. In our experiments, we observed best results when h was set to 4-5.

4 Experimental Results

Three separate datasets (i.e. *Dataset 1, 2 and 3* described above) with a total of 358 food images were used with and without temporal context. Fifty-six unique food items were contained in the datasets. Each dataset features different food composition and consumption style. For example, milk, lasagna, mixed salad and garlic bread are the most frequently-consumed foods in *Dataset 1* while *Dataset 2* does not have any frequently-consumed foods except milk. *Dataset 3* represents a significant dietary pattern change within a month. The first three weeks in *Dataset 3* have similar food consumption style as *Dataset 1*. However, we selected data from participant 7 in Fig.1 for the last week, which has noticeably different eating pattern.

Figure 2 shows how the recursive Bayesian network updates the prediction probabilities for three example food items in *Dataset 1*. On Day 1, every food has the same prediction of 0.5. In the end, the predictions of milk, orange and pretzel converge to 0.99, 0.34 and 0.12 respectively. Milk was consumed almost every day in *Dataset 1*, so the blue curve in Fig.2 gradually increases to show improved confidence. Note the prediction for pretzel decreases and the red curve for orange oscillates around 0.3 because pretzel or orange were consumed less frequently.

Note we used the food label with the highest confidence score from classifier, and define the classification accuracy as, $\Theta = \frac{TP}{TP+FP+TN}$, where TP denotes True Positives (correctly detected food segments), FP denotes False Positives (incorrectly detected food segments or misidentified foods) and TN denotes True Negatives (food not detected).

As we show below the classification accuracy from the highest confidence score is in the range of 50-65%. In the typical operation of our mFR system we report the top 4 food labels and have a classification accuracy of 80-85% [11]. In this paper we want to emphasize how the contextual information improves the classifier performance.

Figure 3 demonstrates the food classification accuracy improvement. The blue lines in Fig.3(a), 3(b) and 3(c) indicate the average daily food classification accuracy with temporal context, Θ_{context}, while the red lines indicate the one

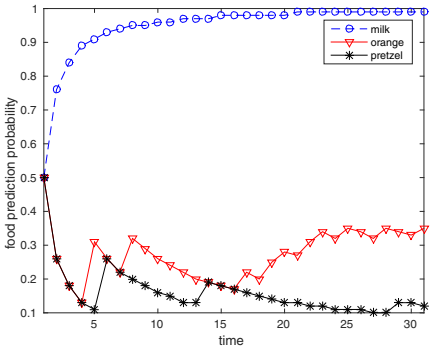
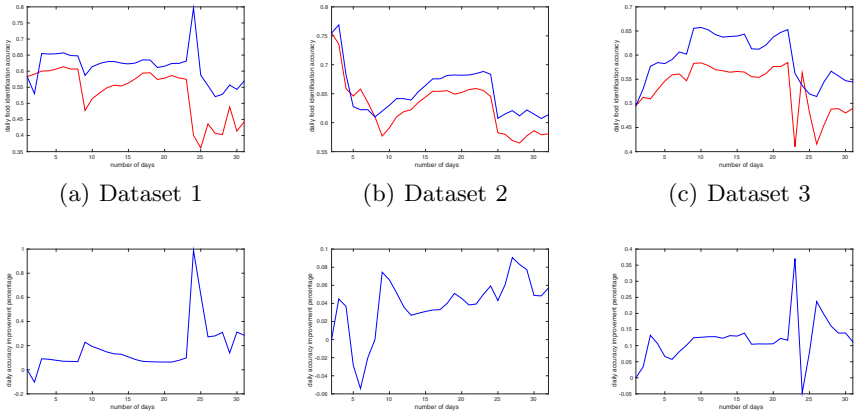


Fig. 2. Food occurrence prediction of three food items (Blue: Milk; Red: Orange; Black: Pretzel).

without, Θ_{auto}. The accuracy improvement is illustrated in Fig.3(b) is determined by $\frac{\Theta_{\text{context}} - \Theta_{\text{auto}}}{\Theta_{\text{auto}}}$. As shown in Fig.3(d), the accuracy improvement drops from Day 10 to Day 20 as the baseline (classification accuracy without context) increases from 47% to 57%. This implies that the proposed method is more effective when the automatic image analysis does not work well. If we set a threshold for the baseline, the average accuracy improvement when the baseline is above 0.55 is just 6.92% compared to 32.97% when the baseline is below 0.55. The 80% accuracy rate achieved with temporal context on Day 25 demonstrates the effectiveness of the proposed method when the automatic image analysis result is poor. In *Dataset 2* , the classification accuracy without context is always above 0.55 (see the red line in Fig.3(b)). The deep valley shown in Fig.3(e) implies the learning process in the first week. Nevertheless, Fig.3(d) and Fig.3(e) both illustrate an ascent trend of accuracy improvement.

We selected the images of the last 7 days to have a noticeably different food consumption pattern compared to the first 23 days in *Dataset 3*. We would like to verify the behavior of our training model under circumstance where a person may change their eating style. As expected, the blue line and the red line intersect in Fig.3(c) on Day 24. We witnessed a huge drop in Fig.3(f) followed by the re-learning state. The accuracy improvement is negative on Day 24, because the context-based prediction puts more confidence on the specific food, which *Dataset 3* no longer contains after changing one's eating habit, for example, milk is not consumed on Day 24. Due to the dietary change in *Dataset 3*, the increasing trend of classification accuracy is not as conspicuous as Fig.3(d) and Fig.3(e). Table 1 summarizes two statistics for the datasets, average daily classification accuracy (in %) and average daily accuracy improvement (in %). Due to our dataset selection, the classification accuracy using automatic image analysis alone in *Dataset 2* is significantly higher than other datasets. Thus, the accuracy improvement for *Dataset 2* is expected to be lower (3.85%). The fact that *Dataset 2* has less frequently-consumed foods also contributed to the lower



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

(d) Dataset 1 (e) Dataset 2 (f) Dataset 3

Fig. 3. Learning curves for one month. Daily classification rates with(blue) and without(red) temporal context are illustrated in (a),(b) and (c). Corresponding accuracy improvements are shown in (d),(e) and (f).

accuracy improvement. When a person has a more consistent eating pattern, such as *User 1*'s dataset, the classification accuracy gain using temporal contextual information is significantly higher (18.45%). On average, the proposed method of utilizing temporal context shows approximately 11%($\approx \frac{18.45+3.85+12.39}{3}$%) improvement (see Table 1).

Table 1. Food classification performance statistics

statistics	user ID	with context	without context
average daily classification accuracy(%)	user1	61.88	53.23
	user2	65.25	62.90
	user3	59.69	53.28
average daily accuracy improvement(%)	user1	18.45	
	user2	3.85	
	user3	12.39	

5 Conclusions

In this paper we investigated the use of temporal context to improve food classification accuracy. We used a recursive Bayesian network to achieve active learning. Experimental results showed the classification accuracy was improved by 11% on average. In the future, we plan to extend our learning model by combining various ranges of eating history and separating eating occasions based on the time of day.

References

1. The TADA project. http://tadaproject.org
2. Ahmad, Z., Khanna, N., Kerr, D., Boushey, C., Delp, E.: A mobile phone user interface for image-based dietary assessment. In: Proceedings of the IS&T/SPIE Conference on Mobile Devices and Multimedia 9030, 903007-1-9, San Francisco, CA (February 2014)
3. Arulampalam, S., Maskell, S., Gordon, N., Clapp, T.: A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking. IEEE Transactions on Signal Processing **50**(2), 174–188 (2002)
4. Bar, M.: Visual objects in context. Nature Reviews Neuroscience **5**(8), 617–629 (2004)
5. Beijbom, O., Joshi, N., Morris, D., Saponas, S., Khullar, S.: Menu-match: Restaurant-specific food logging from images. In: Winter Conference on Applications of Computer Vision, pp. 844–851, Waikoloa, HI (January 2015)
6. Eicher-Miller, H., Boushey, C.: 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 **26**, 256–261 (2012)
7. Fagot-Campagna, A., Saaddine, J., Flegal, K., Beckles, G.: Diabetes, impaired fasting glucose, and elevated hba1c in us adolescents: the third national health and nutrition examination survey. Diabetes Care **24**(5), 834–837 (2001)
8. Felzenszwalb, P., Huttenlocher, D.: Image segmentation using local variation. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, pp. 98–104, Santa Barbara, CA (June 1998)
9. He, Y., Xu, C., Khanna, N., Boushey, C., Delp, E.: Context based food image analysis. In: Proceedings of IEEE International Conference on Image Processing, pp. 2748–2752, Melbourne, Australia (September 2013)
10. He, Y., Xu, C., Khanna, N., Boushey, C., Delp, E.: Food image analysis: segmentation, identification and weight estimation. In: Proceeding of the IEEE International Conference on Multimedia and Expo, pp. 1–6, San Jose, CA (July 2013)
11. He, Y., Xu, C., Khanna, N., Boushey, C., Delp, E.: Analysis of food images: Features and classification. In: Proceedings of the IEEE International Conference on Image Processing, pp. 2744–2748, Paris, France (October 2014)
12. Manjunath, B., Ohm, J.R., Vasudevan, V., Yamada, A.: Color and texture descriptors. IEEE Transactions on Circuits and Systems for Video Technology **11**(6), 703–715 (2001)
13. Matsuda, Y., Yanai, K.: Multiple-food recognition considering co-occurrence employing manifold ranking. In: Proceedings of the IEEE International Conference on Pattern Recognition, pp. 2017–2020, Tsukuba, Japan (November 2012)
14. McFee, B., Galleguillos, C., Lanckriet, G.: Contextual object localization with multiple kernel nearest-neighbor. IEEE Transactions on Image Processing **20**(2), 570–585 (2011)
15. Ogden, C., Carroll, M., Curtin, L., Lamb, M., Flegal, K.: Prevalence of high body mass index in us children and adolescents, 2007–2008. Jama **303**(3), 242–249 (2010)
16. Rabinovich, A., Vedaldi, A., Galleguillos, C., Wiewiora, E., Belongie, S.: Objects in context. In: Proceedings of IEEE International Conference on Computer vision, pp. 1–8, Rio de Janeiro, Brazil (October 2007)

17. Sarkka, S.: Bayesian Filtering and Smoothing. Cambridge University Press, Cambridge (2013)
18. Zhu, F., Bosch, M., Woo, I., Kim, S., Boushey, C., Ebert, D., Delp, E.: The use of mobile devices in aiding dietary assessment and evaluation. IEEE Journal of Selected Topics in Signal Processing **4**(4), 756–766 (2010)
19. Zhu, F., Bosch, M., Khanna, N., Boushey, C., Delp, E.: Multiple hypotheses image segmentation and classification with application to dietary assessment. IEEE Journal of Biomedical and Health Informatics **19**(1), 377–388 (2015)