Modeling Restaurant Context for Food Recognition

Luis Herranz, Shuqiang Jiang, *Senior Member, IEEE*, and Ruihan Xu

***Abstract*—Food photos are widely used in food logs for diet monitoring and in social networks to share social and gastronomic experiences. A large number of these images are taken in restaurants. Dish recognition in general is very challenging, due to different cuisines, cooking styles, and the intrinsic difficulty of modeling food from its visual appearance. However, contextual knowledge can be crucial to improve recognition in such scenario. In particular, geocontext has been widely exploited for outdoor landmark recognition. Similarly, we exploit knowledge about menus and location of restaurants and test images. We first adapt a framework based on discarding unlikely categories located far from the test image. Then, we reformulate the problem using a probabilistic model connecting dishes, restaurants, and locations. We apply that model in three different tasks: dish recognition, restaurant recognition, and location refinement. Experiments on six datasets show that by integrating multiple evidences (visual, location, and external knowledge) our system can boost the performance in all tasks.**

***Index Terms*—Food recognition, image recognition, location, mobile applications, probabilistic modeling.**

I. Introduction

EATING is an essential activity, with food being connected to countless aspects and events in our life. With the development of recent technologies, such as smartphones and computer vision, food-related applications have flourished. Health monitoring is an important research area. Examples of health-related applications are food logs [1]–[3], calorie intake estimation [4]–[6] and nutrition analysis [7]–[9]. Dietary self-monitoring has been proved effective for changing eating habits, helping people to lose weight [10], [11]. Another popular area is cooking-related activities. Examples are cooking video indexing [12] and authoring [13], cooking activity recognition [14], [15], menu planning [16]–[18], recipe recommendation [19], [20], enhanced recipes [21], [22] and cooking support and assistance [23], [24]. Finally, other works focus on food images taken in a social context, providing automatic annotation [25]–[27] and retrieval of similar images. Social networks are also useful to predict consumption patterns [28]–[30] and food analysis [30], [31].

In order to realize these applications effectively, recognizing food directly in images [25], [32], [33] is highly desirable. However, unrestricted food recognition is still extremely challenging even for humans, especially relying only in the visual information. Actually, when addressing complex recognition problems, humans incorporate prior and contextual knowledge [34]. Similarly, intelligent systems can also leverage external knowledge to simplify the problem.

In this paper, we focus on the specific but popular scenario of dining out in restaurants and taking photos of food (i.e. dishes). Those photos can be kept in personal food logs, used to retrieve nutritional information, recipes or any other information of interest, or shared in social networks as new experiences. The user is often not familiar with the particular dish or even the restaurant (e.g. while traveling in a foreign country) so automatic recognition is convenient. In that scenario, two important tags are the name of the dish and the restaurant. Unconstrained dish recognition in such scenario is extremely complex due to the large number of classes and great variation due to different cooking and presentation styles across restaurants. For that reason we leverage external information (menu and restaurant information) and exploit geographic location to simplify the problem and improve the performance.

We adopt a probabilistic approach, allowing us to design flexible models for each of the components of the problem, and often leading to improved performance. Thus, we propose a probabilistic model that connects locations, restaurants, dishes and visual features. By combining visual and location signals, and knowledge about the restaurants, we can significantly improve the performance of automatic annotation of dish and restaurant names. Additionally, we can refine the estimated location, which is particularly useful in indoor environments where the estimation is more difficult.

The rest of the paper is organized as follows. Section II reviews the related work. Section III introduces the problem of dish recognition in restaurants. The proposed model is described in Sections IV and V. Experiments and conclusions are presented in Sections VI and VII.

II. Related Work

In our particular scenario (i.e. dish recognition in restaurants) we can identify two relevant groups of related works: food recognition and context-based image recognition.

Manuscript received March 1, 2016; revised June 15, 2016 and July 20, 2016; accepted September 9, 2016. Date of publication October 3, 2016; date of current version January 17, 2017. This work was supported in part by the National Basic Research 973 Program of China under Grant 2012CB316400, in part by the National Natural Science Foundation of China under Grant 61550110505, Grant 61532018, and Grant 61322212, in part by the National High Technology Research and Development 863 Program of China under Grant 2014AA015202, in part by the Beijing Municipal Commission of Science and Technology under Grant D161100001816001, and in part by the Lenovo Outstanding Young Scientists Program. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Benoit Huet.

The authors are with the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China (e-mail: luis.herranz@vipl.ict.ac.cn; shuqiang.jiang@vipl.ict.ac.cn; ruihan.xu@vipl.ict.ac.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2016.2614861

Early works in food recognition were able to classify among a few dozen types of food [33], [35], [36]. Kawana and Yanai [37] proposed a mobile food recognition system that can recognize 256 food categories. Convolutional neural networks (CNNs) [38], [39] have been applied successfully to food recognition [32], [40], [41]. However, large-scale food recognition, covering multiple cuisines and fine-grained classification, is still a very challenging problem.

When humans face complex recognition problems, they often exploit contextual information, which is often more important than the content itself [34]. Similarly, modern devices can exploit different sources of knowledge (e.g. websites, databases) and contextual information (e.g. GPS, accelerometer). The most representative example is mobile recognition of landmarks [42], [43] based on location and image retrieval techniques to find photos of the same landmark from geotagged photo databases, and use them to annotate the test image. location can effectively bound the search to only a subset of images. Typically, local features such as SIFT are extracted, and encoded with a bag-of-words representation [44] or using vocabulary trees [42], [43]. As landmarks are rigid and geometrically almost invariant, retrieving similar images and performing geometric verification often finds the right landmark. Classifiers can also be used instead of retrieval techniques. In this case location helps to restrict the classification to the landmarks in the geographic neighborhood (i.e. shortlists the candidate classes).

Recently, three works [1], [45], [46] almost simultaneously proposed restaurant-oriented food recognition, where the restaurant context (menu and images) and location are leveraged to improve food recognition. They basically reduce the candidate categories to those in the menus of the neighboring restaurants (i.e. *shortlist* the candidates). In particular, Bettadapura *et al.* [45] focus on automatic food logging, aided by online restaurant information. Menu-Match [1] also retrieves nutritional information. They evaluate their methods with relatively small datasets (4350 images from 10 restaurants, and 645 images from 3 restaurants, respectively). Xu *et al.* [46] focus on a larger scale scenario, with the data obtained by crawling online restaurant databases (collecting menu, location and user-contributed images of dishes). The resulting Dishes dataset contains around 115 K images collected from 646 restaurants in 6 cities. Their study focuses on classification under geolocalized conditions, showing that geolocalized training can improve classification performance and efficiency. In contrast, here we focus on better modeling contextual data and the relation with the other components, rather than on the visual classifier itself.

III. Dish Recognition in Restaurants

A. Dish Recognition in Restaurants Problem

Traditional food or dish recognition tries to identify the class s of an input image from its visual descriptor $\mathbf{x}$, using certain visual classifier $p(s|\mathbf{x})$. We focus on the more constrained problem of dish recognition *in restaurants*, assuming that the user is located within a restaurant. Thus, in addition to the visual model, the system has access to additional contextual information, in particular the menu of the restaurant and the geographic location of both restaurants and users.

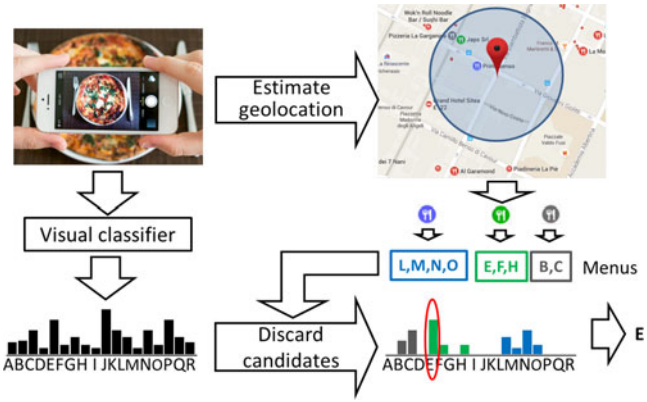

Fig. 1. Overview of the baseline framework for dish recognition in restaurants (i.e., *shortlist* method).

The recognition system takes the pair $(\boldsymbol{\mu}_q, \mathbf{x})$ as input, where $\boldsymbol{\mu}_q$ are the local coordinates and the visual descriptor $\mathbf{x}$. When a new image is captured, we assume that the mobile device has estimated its current location $\boldsymbol{\Psi}_q = (\lambda_q, \phi_q)$ via its location services, where λ_q and ϕ_q denote latitude and longitude.

For a given restaurant k, the information the system exploits is its menu M_k (i.e. the list of dish categories served in restaurant k) and its geographic location $\boldsymbol{\Psi}_k = (\lambda_k, \phi_k)$. For simplicity we use the local coordinates $\boldsymbol{\mu_k} = (u_k, v_k)$. The restaurant database contains K restaurants with a combined total of $D = \left|\bigcup_{k=1}^{K} M_k\right|$ dishes. The menu is represented as $M_k = \{s_1, \ldots, s_{D_k}\}$, where $s_i \in \{1, \ldots D\}$ is the i-th dish in the restaurant menu M_k, with D_k different dishes.

B. Approach 1: Shortlist

A simple yet effective way to include geolocalized knowledge is by discarding unlikely candidates and thus reducing the complexity of the problem. This approach is commonly used in landmark recognition, often referred to as the *shortlist* approach [47]. This approach uses location to discard all the landmarks or buildings outside an area centered at $\boldsymbol{\mu}_q$, and then search for similar images within the remaining ones. Since the remaining images belong to a fraction of the candidate classes, the problem is easier, and can both save computation cost and increase the accuracy.

This method can be easily adapted to our scenario [1], [45], [46], in which the user takes a photo of the dish and the smartphone estimates the location via the operating system's location services. Since the photo is taken in one of the restaurants within the geographical neighborhood, only the dishes in the menus of those restaurants are likely to be the actual dish in the photo, so the remaining classes can be ignored in the result of the visual classifier (see Fig. 1). Given the coordinates $\boldsymbol{\mu}_q$ and the visual feature $\mathbf{x}$, predicting the dish is equivalent to finding the dish with maximum probability among the candidates

$$s^* = \underset{s \in \bigcup_{k \in H_q} M_k}{\arg\max}\ p(s|\mathbf{x}) \tag{1}$$

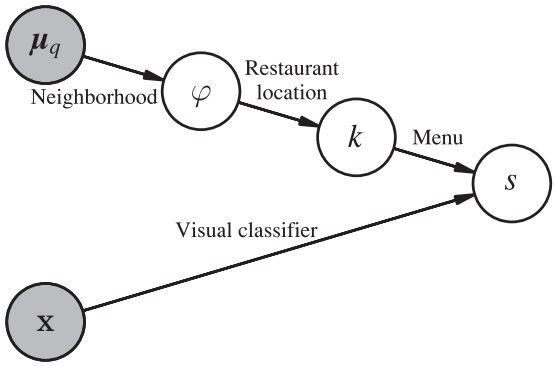

Fig. 2. Proposed probabilistic model. The estimated location $\boldsymbol{\mu}_q$ and the visual feature $\mathbf{x}$ are observed variables, and the actual location $\boldsymbol{\varphi}$, the restaurant k, and the dish s are latent variables.

where H_q is the set of candidate restaurants obtained as

$$H_q = H\left(\boldsymbol{\mu}_q, \epsilon\right) = \left\{k \mid \left\|\boldsymbol{\mu}_k - \boldsymbol{\mu}_q\right\| \leq \epsilon \quad \forall k = 1, \ldots, K\right\} \tag{2}$$

where ϵ is the maximum distance from the candidate restaurants to the test image.

IV. Probabilistic Framework

A. Model

While the underlying idea of the shortlist approach is very intuitive, previous works [1], [45], [46] have implemented it in a simple way and based on simple rules to connect each module. In contrast, we adopt a probabilistic perspective, modeling the system as the generative process of Fig. 2. In this way we can use probabilistic models to connect the different components, rather than heuristic rules.

In our model, the device provides the estimated location $\boldsymbol{\mu}_q$ and the visual feature $\mathbf{x}$, which are the observed variables. The actual location $\boldsymbol{\varphi}$, the restaurant k and the dish s are latent variables. We introduce explicitly the dependency between the restaurant and the dish (via the menu), the visual feature and the dish (via the visual classifier) and the restaurant and the location of the user. We explicitly introduce a new variable $\boldsymbol{\varphi}$ denoting the (true) location of the user, which is different from the observed location $\boldsymbol{\mu}_q$ estimated by the location services of the device.

Given the previous observed and latent variables, and the graphical model, the joint distribution $p\left(s, k, \boldsymbol{\varphi} | \boldsymbol{\mu}_q, \mathbf{x}\right)$ can be factorized as

$$p\left(s, k, \boldsymbol{\varphi} | \boldsymbol{\mu}_q, \mathbf{x}\right) = p\left(\boldsymbol{\varphi} | \boldsymbol{\mu}_q\right) p\left(k | \boldsymbol{\varphi}\right) p\left(s | k, \mathbf{x}\right). \tag{3}$$

In this factorization we can identify three factors: the *neighborhood model* $p\left(\boldsymbol{\varphi} | \boldsymbol{\mu}_q\right)$, the *restaurant location model* $p\left(k | \boldsymbol{\varphi}\right)$ and the *(restaurant-conditioned) visual model* $p\left(s | k, \mathbf{x}\right)$, which accounts for the explicit dependency on the menu of k.

To predict the dish, we marginalize (3) over k and $\boldsymbol{\varphi}$

$$p\left(s | \boldsymbol{\mu}_q, \mathbf{x}\right) = \sum_{k=1}^{K} p\left(s | k, \mathbf{x}\right) \int_{\boldsymbol{\varphi}} p\left(\boldsymbol{\varphi} | \boldsymbol{\mu}_q\right) p\left(k | \boldsymbol{\varphi}\right) \mathrm{d}\boldsymbol{\varphi}. \tag{4}$$

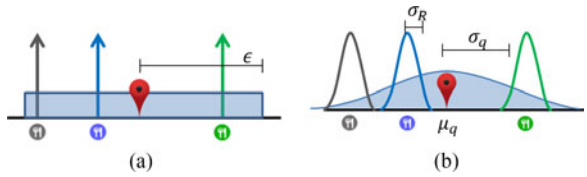

Fig. 3. Modeling neighborhoods and restaurant locations: (a) shortlist approach (piecewise disc and delta, respectively), and (b) alternative models using Gaussian distributions.

The predicted dish can be then obtained by solving

$$s^* = \underset{s \in \{1, \ldots D\}}{\arg\max} p\left(s | \boldsymbol{\mu}_q, \mathbf{x}\right). \tag{5}$$

B. Revisiting the Shortlist Approach

Now we can revisit the shortlist approach of Fig. 2 from this probabilistic perspective. Comparing (1) and (3), we can easily identify the neighborhood model as a circle of radius ϵ centered at $\boldsymbol{\mu}_q$

$$p_{\mathrm{SL}}\left(\boldsymbol{\varphi} | \boldsymbol{\mu}_q\right) = \left[\left\|\boldsymbol{\varphi} - \boldsymbol{\mu}_q\right\| \leq \epsilon\right]. \tag{6}$$

Restaurants are represented as points. Thus, the corresponding restaurant location can be modeled with the delta function as

$$p_{\mathrm{SL}}\left(k | \boldsymbol{\varphi}\right) = \delta\left(\left\|\boldsymbol{\varphi} - \boldsymbol{\mu}_k\right\|\right). \tag{7}$$

For each restaurant, only the dishes in its menu are candidate categories, and thus have non-zero probability. We can include this fact in the visual model as

$$p_{\mathrm{SL}}\left(s | k, \mathbf{x}\right) \propto p\left(s | \mathbf{x}\right) \left[s \in M_k\right] \tag{8}$$

where $[P]$ is 1 if the statement P is true, and 0 otherwise. Note that (8) can be normalized to recover the full probability.

Using (6), (7) and (8) in (4) we obtain

$$\begin{aligned} p_{\mathrm{SL}}\left(s | \boldsymbol{\mu}_{\mathbf{q}}, \mathbf{x}\right) &\propto p\left(s | \mathbf{x}\right) \times \sum_{k=1}^{K} \left[s \in M_k\right] \\ &\quad \times \int_{\boldsymbol{\varphi}} \left[\left\|\boldsymbol{\varphi} - \boldsymbol{\mu}_q\right\| \leq \epsilon\right] \delta\left(\left\|\boldsymbol{\varphi} - \boldsymbol{\mu}_k\right\|\right) \mathrm{d}\boldsymbol{\varphi} \\ &= p\left(s | \mathbf{x}\right) \sum_{k=1}^{K} \left[s \in M_k\right] \int_{\boldsymbol{\varphi} \in H_q} \delta\left(\left\|\boldsymbol{\varphi} - \boldsymbol{\mu}_k\right\|\right) \mathrm{d}\boldsymbol{\varphi} \\ &= p\left(s | \mathbf{x}\right) \left[s \in \bigcup_{k \in H_q} M_k\right] \end{aligned} \tag{9}$$

where $H_q = \left\{\boldsymbol{\varphi} \mid \left\|\boldsymbol{\varphi} - \boldsymbol{\mu}_q\right\| \leq \epsilon\right\}$ is the ϵ-circular geographical neighborhood of the test image. Note that solving (5) for (9) is equivalent to solving (1).

C. Alternative Neighborhood and Restaurant Location Models

Fig. 3(a) illustrates the neighborhood and restaurant models described in the previous section. We can see that both models have obvious limitations. The hard-threshold neighborhood

model considers all the candidate classes equally probable, no matter the restaurant is in the border of the neighborhood or very close to the estimated location. A model with soft decay would be more realistic (see Fig. 3). Thus, instead of (6), we use a Gaussian model for the neighborhood

$$p_{\mathrm{G}}\left(\boldsymbol{\varphi}|\boldsymbol{\mu_q}\right) = \mathcal{N}\left(\boldsymbol{\varphi}|\boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q\right) \tag{10}$$

with $\boldsymbol{\Sigma}_q = \sigma_q^2\mathbf{I}$.

Similarly, representing a restaurant with a point is not realistic, as they cover certain spatial area. If we had full access to the dimensions and layout of each restaurant we could use it as $p\left(k|\boldsymbol{\varphi}\right)$. Unfortunately, we do not have that information, so for convenience we simply use another Gaussian model

$$p_{\mathrm{G}}\left(k|\boldsymbol{\varphi}\right) = \mathcal{N}\left(\boldsymbol{\varphi}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right) \tag{11}$$

with $\boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}_{\mathrm{R}} = \sigma_{\mathrm{R}}^2\mathbf{I}$, where we assume the same model for all the restaurants. Note that (11) collapses to the model of (7) when $\sigma_{\mathrm{R}} = 0$.

Using a probabilistic interpretation, we can consider the menu as a prior over the global visual classifier model $p\left(s|\mathbf{x}\right)$, with the menu modeled as $p\left(s|k\right) = \frac{[s\in M_k]}{|M_k|}$. The resulting restaurant-dependent visual model is

$$p_{\mathrm{R}}\left(s|k,\mathbf{x}\right) = p\left(s|\mathbf{x}\right)\frac{[s\in M_k]}{|M_k|}. \tag{12}$$

Using the new models from (10), (11) and (12) in (4) we obtain the new marginal probability

$$\begin{aligned} p\left(s|\boldsymbol{\mu_q},\mathbf{x}\right) \;\propto\; & p\left(s|\mathbf{x}\right)\times\sum_{k=1}^{K}\frac{[s\in M_k]}{|M_k|} \\ & \times\int_{\boldsymbol{\varphi}}\mathcal{N}\left(\boldsymbol{\varphi}|\boldsymbol{\mu}_q,\boldsymbol{\Sigma}_q\right)\times\mathcal{N}\left(\boldsymbol{\varphi}|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k\right)\mathrm{d}\boldsymbol{\varphi}. \end{aligned} \tag{13}$$

Using the following relation for the product of two multivariate Gaussians:

$$\begin{aligned} &\mathcal{N}\left(\boldsymbol{\varphi}|\boldsymbol{\mu}_q,\boldsymbol{\Sigma}_q\right)\mathcal{N}\left(\boldsymbol{\varphi}|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k\right) \\ &\quad = \mathcal{N}\left(\boldsymbol{\mu}_k|\boldsymbol{\mu}_q,\boldsymbol{\Sigma}_q+\boldsymbol{\Sigma}_k\right)\mathcal{N}\left(\boldsymbol{\varphi}|\boldsymbol{\theta},\boldsymbol{\Lambda}\right) \\ &\boldsymbol{\Lambda} = \left(\boldsymbol{\Sigma}_q^{-1}+\boldsymbol{\Sigma}_k^{-1}\right)^{-1} \\ &\boldsymbol{\theta} = \boldsymbol{\Lambda}\left(\boldsymbol{\Sigma}_q^{-1}\boldsymbol{\mu}_q+\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k\right) \end{aligned} \tag{14}$$

in (14) we further obtain

$$\begin{aligned} p\left(s|\boldsymbol{\mu_q},\mathbf{x}\right) \;\propto\; & p\left(s|\mathbf{x}\right)\times\sum_{k=1}^{K}\frac{[s\in M_k]}{|M_k|} \\ & \times\int_{\boldsymbol{\varphi}}\mathcal{N}\left(\boldsymbol{\mu}_k|\boldsymbol{\mu}_q,\boldsymbol{\Sigma}_q+\boldsymbol{\Sigma}_k\right)\times\mathcal{N}\left(\boldsymbol{\varphi}|\boldsymbol{\theta},\boldsymbol{\Lambda}\right)\mathrm{d}\boldsymbol{\varphi} \\ \propto\; & p\left(s|\mathbf{x}\right)\sum_{k=1}^{K}\frac{[s\in M_k]}{|M_k|}\mathcal{N}\left(\boldsymbol{\mu}_k|\boldsymbol{\mu}_q,\boldsymbol{\Sigma}_q+\boldsymbol{\Sigma}_k\right). \end{aligned} \tag{15}$$

TABLE I
DATA USED IN THE EXPERIMENTS

City	#restaurants	#dishes		#images	
Total	Per restaurant	Total	Per dish		
Beijing	187	1173	6.27	45541	38.82
Shanghai	198	1253	6.33	37590	30.00
Tianjin	78	435	5.58	10811	24.85
Nanjing	64	328	5.13	7895	24.07
Hangzhou	62	371	5.98	9124	24.59
Guangzhou	57	272	4.77	6543	24.06

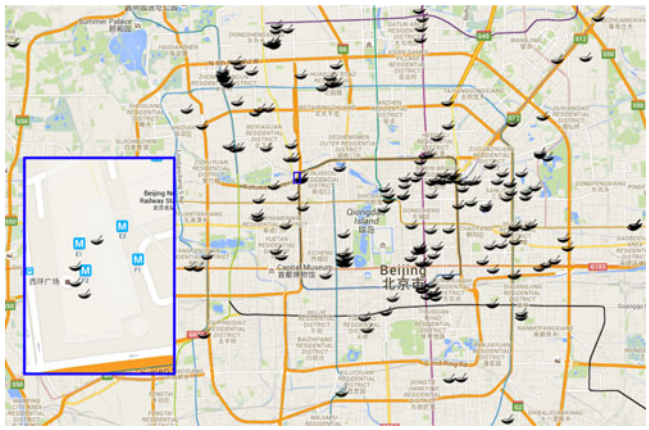

Fig. 4. Geographic distribution of the restaurants in the Beijing dataset. The window shows an example of dense area.

V. SOLVING OTHER TASKS

In this probabilistic framework, the joint distribution $p\left(s,k,\boldsymbol{\varphi}|\boldsymbol{\mu}_q,\mathbf{x}\right)$ can be used to perform inference over any latent variables. So far, we focused on predicting the dish. However, by marginalizing over other variables we can also infer the restaurant and even the location. For these problems we focus on the alternative model described in Section IV-C.

A. Restaurant Recognition

Marginalizing (3) over s and $\boldsymbol{\varphi}$ we obtain

$$\begin{aligned} p\left(k|\boldsymbol{\mu}_q,\mathbf{x}\right) \;&=\; \sum_{s=1}^{D}\int_{\boldsymbol{\varphi}}p\left(s,k,\boldsymbol{\varphi}|\boldsymbol{\mu}_q,\mathbf{x}\right)\mathrm{d}\boldsymbol{\varphi} \\ &=\; \sum_{s=1}^{D}p\left(s|k,\mathbf{x}\right)\int_{\boldsymbol{\varphi}}p\left(\boldsymbol{\varphi}|\boldsymbol{\mu}_q\right)p\left(k|\boldsymbol{\varphi}\right)\mathrm{d}\boldsymbol{\varphi} \end{aligned} \tag{16}$$

and using (6), (7) and (12) we obtain

$$\begin{aligned} p\left(k|\boldsymbol{\mu}_q,\mathbf{x}\right) \;\propto\; & \sum_{s=1}^{D}p\left(s|\mathbf{x}\right)\times\frac{[s\in M_k]}{|M_k|} \\ & \times\int_{\boldsymbol{\varphi}}\mathcal{N}\left(\boldsymbol{\varphi}|\boldsymbol{\mu}_q,\boldsymbol{\Sigma}_q\right)\times\mathcal{N}\left(\boldsymbol{\varphi}|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k\right)\mathrm{d}\boldsymbol{\varphi} \\ \propto\; & \mathcal{N}\left(\boldsymbol{\mu}_k|\boldsymbol{\mu}_q,\boldsymbol{\Sigma}_q+\boldsymbol{\Sigma}_k\right)\frac{\sum_{s\in M_k}p\left(s|\mathbf{x}\right)}{|M_k|}. \end{aligned} \tag{17}$$

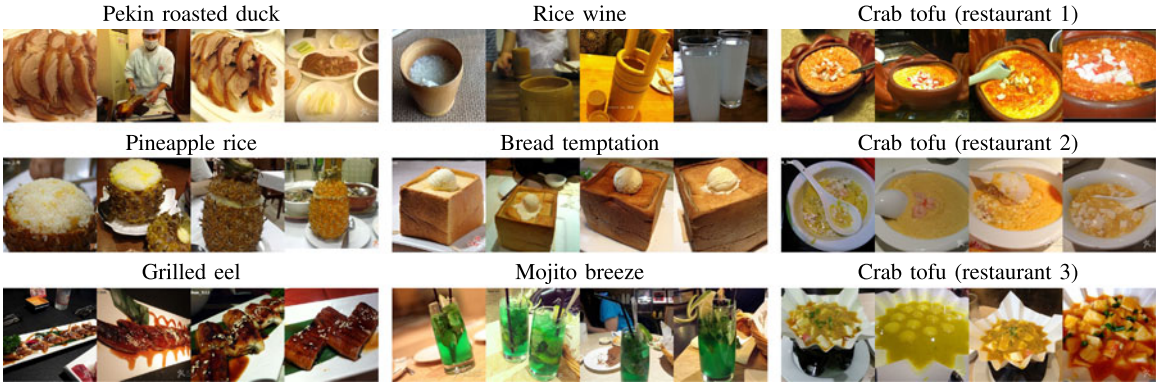

Fig. 5. Examples of dish names and photos available in Dishes. Users tend to share exotic dishes and dishes with attractive presentations.

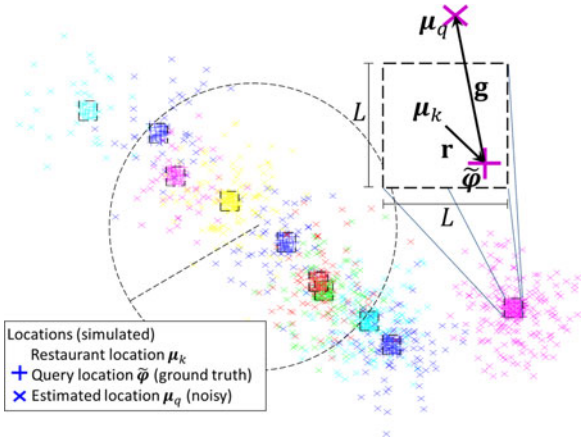

Fig. 6. Dense neighborhood with several restaurants and the resulting simulated locations (with $\sigma_{LOC} = 40$ meters). The top right corner shows how the estimated locations are simulated: from the location of the restaurant $\boldsymbol{\mu}_k$ we randomly sample a location $\tilde{\boldsymbol{\varphi}}$ within squared-shaped restaurants (considered ground truth location) and then add noise to simulate the location $\boldsymbol{\mu}_q$ (used as the noisy estimation obtained in the device). A circular neighborhood of radius 200 meters is shown for reference. Different colors represent different restaurants. Better view in electronic version.

The predicted restaurant is obtained as

$$k^* = \underset{k\in\{1,\dots K\}}{\arg\max} p\left(k|\boldsymbol{\mu}_q, \mathbf{x}\right). \tag{18}$$

B. Location Refinement

Typically, location services in mobile devices only leverage radio signals, such as those from GPS or mobile stations. As a byproduct of the probabilistic approach, we can also integrate the knowledge about restaurants and the visual evidence to improve the initial estimation of the location. This is particularly useful in indoor environments with many restaurants such as shopping malls, where some location signals (e.g. GPS) may not be available or not reliable.

Marginalizing (3) over s and k we obtain

$$p\left(\boldsymbol{\varphi}|\boldsymbol{\mu}_q, \mathbf{x}\right) = p\left(\boldsymbol{\varphi}|\boldsymbol{\mu}_q\right) \sum_{k=1}^{K} p\left(k|\boldsymbol{\varphi}\right) \sum_{s=1}^{S} p\left(s|k, \mathbf{x}\right) \tag{19}$$

Algorithm 1: Location estimation algorithm.

Input: Initial location $\boldsymbol{\mu}_q$ and visual feature $\mathbf{x}$
Output: Location $\boldsymbol{\varphi}$

```
for k = 1 : K do
    Compute Λ_k, θ_k and ω_k using (21), (22) and (23)
end for
Initialize φ = μ_q
repeat
    for k = 1 : K do
        Compute γ_k (φ) using (25)
    end for
    Update estimated location φ using (24)
until converged
return φ
```

and using (10), (11) and (12) we obtain

$$p\left(\boldsymbol{\varphi}|\boldsymbol{\mu}_q, \mathbf{x}\right) \propto \sum_{k=1}^{K} \omega_k \mathcal{N}\left(\boldsymbol{\varphi}|\boldsymbol{\theta}_k, \boldsymbol{\Lambda}_k\right) \tag{20}$$

$$\boldsymbol{\Lambda}_k = \left(\boldsymbol{\Sigma}_q^{-1} + \boldsymbol{\Sigma}_k^{-1}\right)^{-1} \tag{21}$$

$$\boldsymbol{\theta}_k = \boldsymbol{\Lambda}_k \left(\boldsymbol{\Sigma}_q^{-1}\boldsymbol{\mu}_q + \boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k\right) \tag{22}$$

$$\omega_k = \frac{\sum_{s\in M_k} p\left(s|\mathbf{x}\right)}{|M_k|}. \tag{23}$$

In (20) we see that $p\left(\boldsymbol{\varphi}|\boldsymbol{\mu}_q, \mathbf{x}\right)$ is modeled as a mixture of Gaussians. The mean $\boldsymbol{\phi}_k$ and covariance $\boldsymbol{\Lambda}_k$ of the component k depend both on the initial estimation of the location and the restaurant model. The weight ω_k accounts for the evidence that the visual feature $\mathbf{x}$ comes from the restaurant k.

In contrast to the dish and restaurant, the location $\boldsymbol{\varphi}$ is a continuous variable. To find the location that maximizes (20) we use a maximum likelihood approach. Setting $\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\varphi}} \ln p\left(\boldsymbol{\varphi}|\boldsymbol{\mu}_q, \mathbf{x}\right) = 0$ we obtain

$$\boldsymbol{\varphi} = \frac{1}{\sum_{j=1}^{K} \gamma_j\left(\boldsymbol{\varphi}\right)} \sum_{k=1}^{K} \gamma_k\left(\boldsymbol{\varphi}\right) \boldsymbol{\theta}_k \tag{24}$$

where we define

$$\gamma_k\left(\boldsymbol{\varphi}\right) = \frac{\omega_k \mathcal{N}\left(\boldsymbol{\varphi}|\boldsymbol{\theta}_k, \boldsymbol{\Lambda}_k\right)}{\sum_{j=1}^{K} \omega_j \mathcal{N}\left(\boldsymbol{\varphi}|\boldsymbol{\theta}_j, \boldsymbol{\Lambda}_j\right)}. \tag{25}$$

TABLE II
DISH RECOGNITION ACCURACY

Dataset	Radius	All (≥ 0 restaurants)								Dense (≥ 5 restaurants)		
	ϵ (SL)	Accuracy (%)				Average class accuracy (%)				Accuracy (%)		
	$3\sigma_q$ (PR)	VS	CX	SL	PR	VS	CX	SL	PR	CX	SL	PR
Beijing	50		11.21	42.45	77.5		12.96	40.17	74.34	N/A	N/A	74.14
	200		11.59	77.93	**78.24**		13.85	74.31	**74.76**	7.36	76.47	**78.64**
	500	54.75	10.92	76.22	76.70	50.31	12.44	72.14	73.19	9.47	72.99	76.52
	1000		11.09	73.48	74.76		12.97	68.86	70.82	9.62	72.42	74.79
	2000		11.37	69.85	71.70		13.54	64.71	67.33	11.12	69.06	71.63
	Best		11.67	77.93	**78.58**		14.20	74.31	**75.28**		-	
	$(\epsilon, 3\sigma_q)$		(700)	(200)	(100)		(700)	(200)	(100)			
Shanghai	50		11.28	41.32	75.49		11.03	39.63	72.28	2.06	69.94	70.38
	200		11.12	74.99	**75.97**		10.72	71.70	**72.54**	4.63	71.01	**74.74**
	500	54.04	11.45	72.46	74.09	51.67	11.24	69.11	70.67	8.99	70.29	73.50
	1000		11.36	69.60	72.03		10.96	66.41	68.49	10.27	67.82	71.66
	2000		11.33	65.26	69.04		11.21	61.90	65.24	10.82	64.50	69.04
	Best		11.91	74.99	**76.37**		11.56	71.70	**73.03**		-	
	$(\epsilon, 3\sigma_q)$		(90)	(200)	(100)		(90)	(200)	(100)			
Tianjin	50		12.55	41.97	78.10		14.34	41.71	77.81	2.08	59.72	70.98
	200		12.35	77.77	**78.35**		13.89	77.23	**78.02**	5.56	72.23	**77.65**
	500	61.45	12.52	75.11	76.74	59.31	14.00	74.35	76.23	7.81	72.55	75.75
	1000		13.08	72.81	74.62		14.16	71.75	73.67	10.79	71.74	74.38
	2000		12.83	69.52	72.05		13.89	67.95	70.71	12.39	68.74	71.98
	Best		13.31	77.77	**78.78**		14.50	77.23	**78.45**		-	
	$(\epsilon, 3\sigma_q)$		(100)	(200)	(90)		(100)	(200)	(90)			
Nanjing	50		15.49	41.95	76.84		16.42	42.10	77.03	N/A	N/A	68.83
	200		15.90	**75.74**	74.92		16.39	76.80	**77.48**	5.66	**75.84**	74.92
	500	60.15	16.25	74.13	75.67	60.37	17.36	73.86	75.84	12.86	70.70	74.49
	1000		15.54	69.25	72.94		16.32	69.06	72.73	13.99	66.58	71.75
	2000		16.23	66.44	69.21		17.64	66.22	69.04	15.14	64.34	68.72
	Best		16.34	76.77	**77.44**		17.82	76.80	**77.59**		-	
	$(\epsilon, 3\sigma_{\mathbf{q}})$		(400)	(200)	(100)		(400)	(200)	(100)			
Hangzhou	50		14.19	46.21	82.71		16.18	44.89	79.38	N/A	N/A	74.88
	200		14.37	83.47	**84.78**		16.29	79.67	**81.32**	6.12	79.27	83.75
	500	69.56	14.19	81.47	83.39	63.85	16.15	77.21	79.74	10.83	79.18	**83.98**
	1000		14.98	78.83	81.55		17.21	74.00	77.47	13.80	75.84	81.43
	2000		13.54	75.23	78.65		14.75	70.52	73.42	12.95	75.13	78.65
	Best		14.98	83.47	**84.85**		17.21	79.67	**81.52**			
	$(\epsilon, 3\sigma_q)$		(1000)	(200)	(100)		(1000)	(200)	(100)			
Guangzhou	50		16.30	40.72	74.84		16.45	39.59	73.42	6.06	79.80	72.68
	200		16.77	75.46	**75.52**		16.77	74.11	**74.21**	7.93	71.90	**74.35**
	500	62.88	16.06	73.53	74.52	61.56	16.15	72.12	73.25	11.19	71.31	73.76
	1000		16.69	70.86	72.85		16.62	69.22	71.47	15.51	69.33	72.76
	2000		16.66	68.38	70.15		16.70	66.71	68.41	16.20	68.09	69.98
	Best		17.21	75.46	**75.67**		17.79	74.11	**74.37**			
	$(\epsilon, 3\sigma_q)$		(70)	(200)	(100)		(70)	(200)	(100)			

VS: visual (no location), **CX**: only context (no visual), **SL**: shortlist, **PR**: probabilistic.

Unfortunately, (24) is not a closed-form expression due to the dependency of $\gamma_k(\boldsymbol{\varphi})$ on $\boldsymbol{\varphi}$. However, we can alternative estimate $\boldsymbol{\Lambda}_k$, $\boldsymbol{\theta}_k$ and ω_k for fixed $\gamma_k(\boldsymbol{\varphi})$, and then estimate $\gamma_k(\boldsymbol{\varphi})$ with the updated $\boldsymbol{\Lambda}_k$, $\boldsymbol{\theta}_k$ and ω_k (see Algorithm 1).

VI. EXPERIMENTAL RESULTS

A. Experimental Setup

Dataset: Most food benchmarks do not include restaurant [25], [35]–[37], [48] or geographic location [1]. Dishes[1][46] is a restaurant-oriented food recognition dataset that includes menus, restaurant locations and dish images, crawled from www.dianping.com for six Chinese cities. The selected restaurants have at least three different dishes in the menu, and at least 15 images per dish. Following Xu *et al.* [46], we use 10 images for training and the rest as test images. We separate the data in the different cities and studied them independently (more details about the datasets are shown in Table I). Fig. 4 shows the geographic distribution of restaurants in the Beijing dataset.

Overview of the content: Dishes differs from other food datasets in the type of content. To understand the type of content we must pay attention to how the data was collected. While

[1][Online]. Available: http://vipl.ict.ac.cn/isia/datasets_dish/index.html

most food datasets [1], [35], [48], [49] define a set of classes and then collect data (either by manually taking photos or querying a search engine), Dishes was collected in a restaurant basis, without targeting specific classes. Only restaurants with more than 3 dishes in the menu and dishes with more than 15 images are kept. Since the data is based on user contributions, having enough images just depends on the interest of users in taking photos and sharing them. Thus, the content rarely includes uninteresting and everyday dishes (e.g. *beef noodles*), because they are rarely shared, and thus not likely to be found in the dataset. In contrast, popular and exotic dishes, or nicely presented dishes are more commonly shared and consequently likely to appear in Dishes (see Fig. 5).

Thus, while a majority of restaurants and dishes are Chinese, there is also a significant diversity of other cuisines (e.g. Western, Japanese, Korean). Fig. 5 shows examples of dishes from the dataset. Most of them have attractive presentations, which often differ from restaurant to restaurant.

Simulating test locations: Images in Dishes are crowdsourced from web data and lack accurate location data, other than the location of the corresponding restaurant. This indirect information is too coarse and not suitable to evaluate properly the proposed methods. Thus, following Xu *et al.* [46], we simulated the location of the test images assuming a simple query location model.[2] This model includes two factors: the location of the user within the restaurant $\mathbf{r}$ and the location error $\mathbf{g}$

$$\tilde{\boldsymbol{\varphi}} = \boldsymbol{\mu}_k + \mathbf{r} \tag{26}$$

$$\boldsymbol{\mu}_q = \tilde{\boldsymbol{\varphi}} + \mathbf{g} \tag{27}$$

where $\tilde{\boldsymbol{\varphi}}$ is the location within the restaurant (used as ground truth location) and $\boldsymbol{\mu}_q$ is the actual location (used as estimated location, i.e. provided by the location services of the device). Both are obtained from the location $\boldsymbol{\mu}_k$ of the corresponding restaurant k, combined with the relative location within the restaurant $\mathbf{r}$ and the error $\mathbf{g}$ of the location service of the device (e.g. GPS). Since we do not have the layout of the restaurant, we assume a square of $L \times L$ ($L = 25$ meters in our experiments), which is modeled by two unidimensional uniform distributions

$$\mathbf{r} = \frac{L}{2}\left(\mathcal{U}\left(-1,1\right), \mathcal{U}\left(-1,1\right)\right) \tag{28}$$

and the error $\mathbf{g}$ is assumed Gaussian $\boldsymbol{\mu}_q$

$$\mathbf{g} = \mathcal{N}\left(0, \sigma^2_{\text{LOC}}\mathbf{I}\right) \tag{29}$$

with $\sigma_{\text{LOC}} = 40$ meters in our experiments. This model is illustrated in Fig. 6, along with examples of simulated test queries and their corresponding ground truth locations and restaurants

Neighborhood parameters: The most important parameter in the model is the size of the neighborhood, with the radius specified either by ϵ or σ_q. Note that ϵ is a parameter of the shortlist method, which cannot be compared directly with the parameter σ_q of the probabilistic method. For better comparison, we inspected the recognition accuracy curves for dish recognition and

[2]This model is just for simulation purposes. Not to be confused with the models in Sections IV and V.

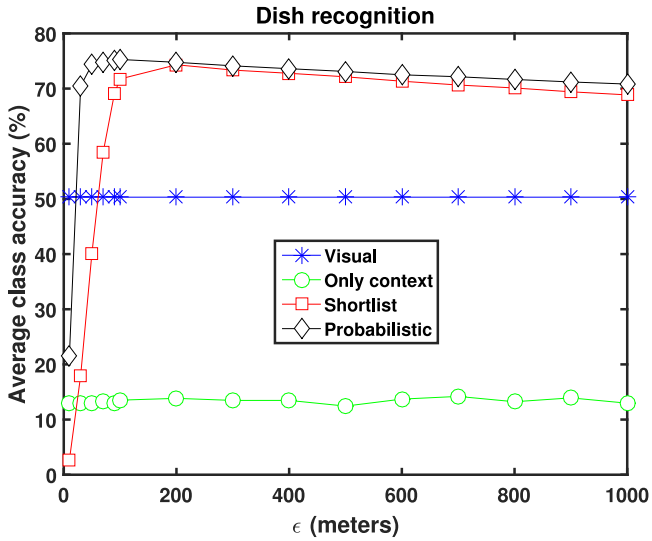

Fig. 7. Dish recognition accuracy (Beijing dataset).

restaurant recognition accuracies, and we found that $\epsilon = 3\sigma_q$ gives a reasonable alignment. For the probabilistic model, the support of a Gaussian function is infinite, but in practice we set the probability to zero for restaurants whose distance to the location of the test image is larger than $10\sigma_q$. In our experiments we evaluated ϵ in a range from 10 to 2000 meters.

Visual features and classifier: For the visual classifier we use a deep network (AlexNet architecture trained on ILSVRC2012 [38]), implemented with Caffe [50]. We extract the activation of the layer *fc7*, and then train a regularized logistic regressor for the particular dataset using Liblinear [51].

Training data: A problem with Dishes is that it imposes very strict constrains on the required data in order to train visual classifiers (at least 3 dishes per restaurant and at least 15 images per dish). This results in only a fraction of the restaurants meeting these demanding requirements, and the dataset is very sparse in geographic location. The consequence is that often there is only one restaurant in the neighborhood. In order to evaluate more realistic and challenging settings, we also report the performance in cases with high density of restaurants (e.g. shopping malls, food streets), defined as those test queries whose $\epsilon-$ neighborhood has at least 5 restaurants (the example in Fig. 6 has a relatively high density of restaurants). Note that in this case the test set depends on ϵ, so the accuracies for different values of ϵ are not directly comparable, since they do not include the same queries. For small neighborhoods, the number of test queries may be too low to be representative.

Tasks and methods: We evaluate the three tasks described earlier, i.e. dish recognition, restaurant recognition and location refinement. We consider the following methods:

1) *Visual* (**VS**): only considers visual information, ignores contextual (i.e. location nor restaurant information).
2) *Contextual* (**CX**): only considers contextual information, ignores visual.
3) *Location* (**LC**): the input location information (i.e. $\boldsymbol{\mu}_q$ in our simulations).
4) *Shortlist* (**SL**): baseline described in Section III-B.

TABLE III
RESTAURANT RECOGNITION ACCURACY

Dataset	Radius ϵ (SL) $3\sigma_q$ (PR)	All (≥ 0 restaurants) Accuracy (%) CX	SL	PR	Average class accuracy (%) CX	SL	PR	Dense (≥ 5 restaurants) Accuracy (%) SL	PR
Beijing	50		52.17	95.27		51.93	96,79	N/A	87.79
	200		95.19	**95.97**		95.63	**97.14**	86.94	**93.06**
	500	84.30	92.25	93.00	88.75	91.56	94.55	85.40	92.17
	1000		87.55	88.81		86.18	90.82	84.58	88.25
	2000		82.25	82.93		79.49	85.44	80.64	82.89
	Best $(\epsilon, 3\sigma_q)$		95.19 (200)	**96.56** (100)		95.63 (200)	**97.60** (100)		
Shanghai	50		51.53	93.68		51.27	94.06	81.33	83.84
	200		92.13	**94.01**		91.74	**94.30**	83.04	**92.48**
	500	71.42	87.56	90.67	72.22	87.56	90.67	83.39	89.94
	1000		82.81	85.82		82.81	85.82	80.36	85.41
	2000		76.48	80.22		76.48	80.22	75.55	80.22
	Best $(\epsilon, 3\sigma_q)$		92.13 (200)	94.72 (100)		91.74 (200)	94.98 (100)		
Tianjin	50		51.37	94.77		51.19	95.95	80.56	87.73
	200		94.23	95.19		95.34	**96.50**	88.04	**93.32**
	500	68.58	90.19	**92.37**	75.44	90.82	94.19	85.89	91.20
	1000		86.78	88.35		86.62	90.62	84.02	88.19
	2000		81.83	83.67		80.99	85.88	80.71	83.58
	Best $(\epsilon, 3\sigma_q)$		94.23 (200)	**95.68** (100)		95.34 (200)	**96.75** (100)	-	
Nanjing	50		53.11	96.99		52.93	96.96	N/A	88.83
	200		96.47	**97.46**		96.23	**97.64**	91.26	**96.87**
	500	82.21	91.53	94.30	83.49	90.50	95.00	88.22	93.73
	1000		84.07	89.01		83.51	90.39	81.52	88.02
	2000		79.91	81.84		79.03	84.45	77.90	81.22
	Best $(\epsilon, 3\sigma_q)$		96.47 (200)	**97.77** (100)		96.23 (200)	**97.82** (100)		
Hangzhou	50		53.44	95.07		53.87	95.61	N/A	85.10
	200		95.36	**97.19**		95.19	**97.25**	88.51	**95.71**
	500	82.66	92.37	94.70	85.85	91.42	95.28	87.69	93.84
	1000		88.64	91.24		86.78	92.04	83.81	91.00
	2000		83.80	85.50		81.30	87.06	83.28	85.50
	Best $(\epsilon, 3\sigma_q)$		95.36 (200)	**97.45** (100)		95.19 (200)	**97.57** (100)	-	
Guangzhou	50		51.50	95.29		51.00	95.67	92.93	91.07
	200		95.81	**96.00**		95.35	**96.59**	91.93	**94.23**
	500	77.43	92.00	94.04	78.94	91.79	94.80	87.79	93.60
	1000		87.58	90.35		87.50	91.91	85.67	90.14
	2000		83.31	85.59		83.27	87.75	82.82	85.22
	Best $(\epsilon, 3\sigma_q)$		95.81 (200)	**96.42** (100)		95.35 (200)	**96.93** (100)		

CX: only context (no visual), **SL**: shortlist, **PR**: probabilistic.

5) *Probabilistic* (**PR**): proposed methods described in Sections IV-C and V.

Evaluation metrics. In most of the experiments we report the recognition accuracy. However, since the dataset is imbalanced with different classes having different number of test images, we also report the average per class accuracy as a complementary quality index. For location refinement we report the distance to the (simulated) ground truth location $\|\boldsymbol{\varphi} - \tilde{\boldsymbol{\varphi}}\|$.

B. Dish Recognition

Table II and Fig. 7 compare the average accuracy of the *shortlist* and *probabilistic* methods. In order to evaluate the contribution of the content and the context, we also include two baselines: the *visual* classifier (without considering location nor prior knowledge) and a purely *contextual* classifier, which ignores visual information. The latter only considers the dishes of the restaurants the neighborhood, and then chooses randomly one of the candidate dishes, since they are equally probable (this is the best we can do without visual information).

As expected, visual information is very important, and *visual* already achieves remarkable accuracies between 54–70%. The context by itself is less reliable, but *contexutal* can reach up to 17%. Combining both type of information increases notably the performance (by 13–25%), which makes the system much

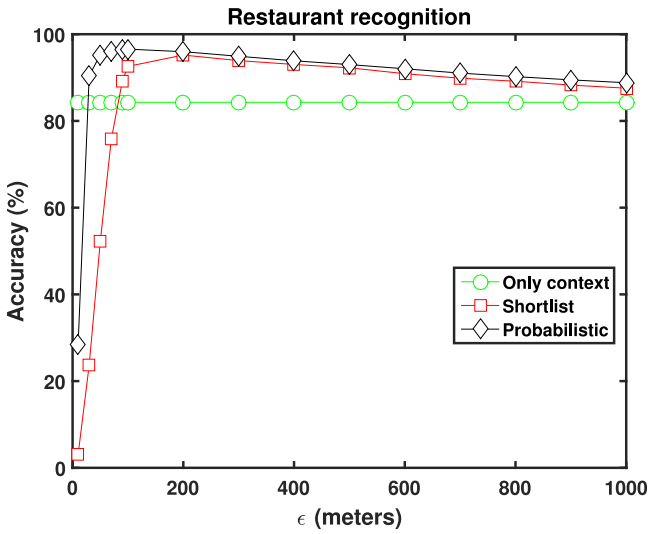

Fig. 8. Restaurant recognition accuracy (Beijing dataset).

more competitive. Both *shortlist*and *probabilistic* achieve a similar best accuracy, with the later being slightly better. However, *shortlist* is very sensitive to the specific choice of the neighborhood size ϵ, while *probabilistic* is much more robust and the accuracy depends less on σ_q, in general benefiting from larger neighborhoods.

C. Restaurant Recognition

For this second task we evaluate the accuracy for restaurant recognition using the proposed probabilistic model (Section V-A). Using only location information, we include the nearest restaurant to the estimated *location* $\boldsymbol{\mu}_q$ as a baseline. We also include another baseline based on selecting the coordinates of the restaurant with the dish detected by *shortlist* (if several restaurants have that dish, we select the nearest to $\boldsymbol{\mu}_q$). The results are shown in Table III and and Fig. 8.

Due to the sparsity in the location and the large number of cases with only one restaurant in the neighborhood, a purely *location*-based approach has already good performance. In this case, visual classification is not so reliable unless the accuracy is very high. Otherwise a wrong prediction would often lead to a wrong restaurant, and a drop in restaurant recognition accuracy. Thus, the performance here is also very dependent on the particular choice of ϵ. Finally, *probabilistic* is more robust to the choice of σ_q and significantly outperforms the other two methods by effectively combining both location and visual information, with a remarkable accuracy of 91.06% in dense areas.

D. Location Refinement

Finally, we evaluate the potential of the proposed model to refine the estimated location by incorporating visual evidence about the dish and prior information about the restaurants. As we simulated the location of test images, we can measure the error in the estimated location using different methods (see Table IV). We compare the restaurant location estimated using the iterative method of Algorithm 1 (*probabilistic*), and compared with the initial estimation $\boldsymbol{\mu}_q$ and the coordinates of the restaurant predicted by *shortlist*, as in the previous section.

By incorporating visual evidence and prior knowledge about the location of the restaurant, the error in the estimation can be reduced dramatically, from 50 to less than 5 meters. The *probabilistic* method generally improves for larger ϵ, while *shortlist* is very sensitive to the performance of the visual classifier, and consequently to the value of ϵ (see Fig. 9). When the visual accuracy drops, either due to a more complex problem in denser areas or to a not suitable value of ϵ, the error increases dramatically.

TABLE IV
LOCATION REFINEMENT ERROR

Dataset	Radius ϵ (SL) / $3\sigma_q$ (PR)	Average error (meters) All (≥ 0 restaurants) LC	CX	SL	PR
Beijing	50			34.86	34.28
	200			4..75	6.51
	500	50.14	8.53	15.93	4.78
	1000			58.01	4.53
	2000			170.23	**4.47**
	Best $(\epsilon, 3\sigma_q)$			4.75 (200)	**4.47** (2000)
Shanghai	50			34.58	24.09
	200			6.65	6.06
	500	50.27	9.58	25.17	4.33
	1000			68.51	4.07
	2000			227.28	**4.01**
	Best $(\epsilon, 3\sigma_q)$			6.65 (200)	**4.01** (2000)
Tianjin	50			34.83	23.78
	200			4.75	5.70
	500			21.47	3.98
	1000			51.70	3.72
	2000			157.23	**3.66**
	Best $(\epsilon, 3\sigma_q)$			4.75 (200)	**3.66** (2000)
Nanjing	50	50.21	10.24	34.16	23.65
	200			3.25	4.89
	500			23.54	3.03
	1000			95.39	2.74
	2000			180.91	**2.69**
	Best $(\epsilon, 3\sigma_q)$			3.25 (200)	**2.69** (2000)
Hangzhou	50	50.21	6.85	34.60	24.65
	200			4.80	7.40
	500	50.15	10.11	17.40	5.74
	1000			51.48	5.49
	2000			143.10	**5.43**
	Best $(\epsilon, 3\sigma_q)$			**4.80** (200)	5.43 (2000)
Guangzhou	50			34.25	23.34
	200			3.48	4.64
	500	50.16	8.11	17.26	2.89
	1000			57.92	2.62
	2000			137.09	**2.58**
	Best $(\epsilon, 3\sigma_q)$			3.48 (200)	**2.58** (2000)

LC: initial location, i.e. $\boldsymbol{\mu}_q$, **SL**: shortlist, **PR**: probabilistic.

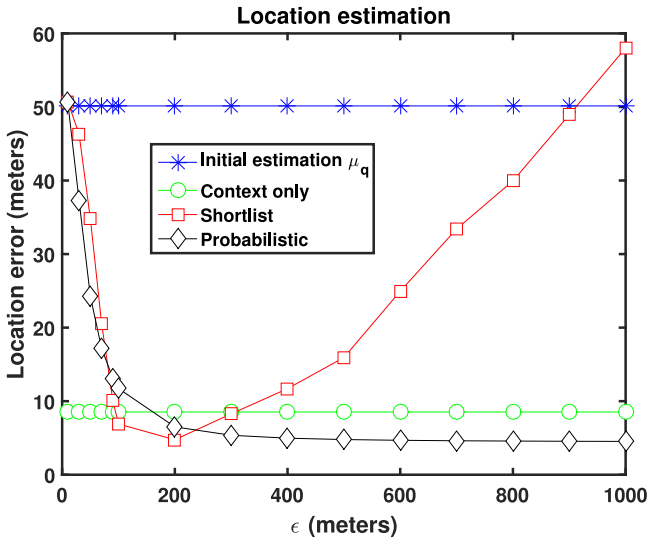

Fig. 9. Location refinement error (Beijing dataset).

VII. Discussion and Conclusion

Even for humans, dish recognition in the wild is extremely challenging, and in general cannot be solved only from the visual image without some prior and contextual knowledge. Integrating different visual and contextual cues is a natural process in humans to formulate an educated guess. Similarly, in this paper we describe an approach to perform different recognition tasks related with the dining out in restaurants scenario, by taking advantage of visual information, geographical context, and prior knowledge about the restaurants. We formulate the problem in a probabilistic framework, which allows us to perform inference over different hidden variables leading to different recognition tasks. Compared with a more simple model (the shortlist approach), the proposed probabilistic approach combines better and more robustly the different cues achieving better performance. Often, integrating heterogeneous and apparently unrelated cues is the key to solve complex problems. For example, we showed experimentally that taking a look to your meal may be helpful to better estimate where you are, provided you are familiar with the restaurants in the area.

The restaurant scenario poses many challenges in practice that can be addressed in future works. Current datasets and recognition methods still have some limitations. While the Dishes dataset is an important step towards evaluate food recognition in a realistic restaurant context, we currently face two limitations. First, photos still lack real location and other useful contextual information, which are desirable for more realistic experiments. One possible direction is designing more accurate models for neighborhoods and restaurants. New types of information can be also incorporated in the framework (e.g. time). In addition, the proposed approach requires training discriminative classifiers, which limits its applicability to those restaurants with enough training images (10 in our experiments). Future works can address this limitation and propose solutions that can deal with fewer training samples or leveraging other type of information could increase the coverage in practice.

References

[1] O. Beijbom, N. Joshi, D. Morris, S. Saponas, and S. Khullar, "Menu-match: Restaurant-specific food logging from images," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, Jan. 2015, pp. 844–851.

[2] K. Aizawa and M. Ogawa, "Foodlog: Multimedia tool for healthcare applications," *IEEE Multimedia*, vol. 22, no. 2, pp. 4–8, Apr. 2015.

[3] K. Kitamura, C. de Silva, T. Yamasaki, and K. Aizawa, "Image processing based approach to food balance analysis for personal food logging," in *Proc. IEEE Int. Conf. Multimedia Expo.*, Jul. 2010, pp. 625–630.

[4] T. Miyazaki, G. C. de Silva, and K. Aizawa, "Image-based calorie content estimation for dietary assessment," in *Proc. IEEE Int. Symp. Multimedia*, Dec. 2011, pp. 363–368.

[5] P. Pouladzadeh, S. Shirmohammadi, and R. Al-Maghrabi, "Measuring calorie and nutrition from food image," *IEEE Trans. Instrum. Meas.*, vol. 63, no. 8, pp. 1947–1956, Aug. 2014.

[6] W. Wu and L. Yang, "Fast food recognition from videos of eating for calorie estimation," in *Proc. IEEE Int. Conf. Multimedia Expo.*, Jun.-Jul. 2009, pp. 1210–1213.

[7] F. Kong and J. Tan, "Dietcam: Regular shape food recognition with a camera phone," in *Proc. Int. Conf. Body Sensor Netw.*, 2011, pp. 127–132.

[8] K. Aizawa, Y. Maruyama, H. Li, and C. Morikawa, "Food balance estimation by using personal dietary tendencies in a multimedia food log," *IEEE Trans. Multimedia*, vol. 15, no. 8, pp. 2176–2185, Dec. 2013.

[9] A. Mariappan *et al.*, "Personal dietary assessment using mobile devices," in *Proc. SPIE Electron. Imag.*, 2009, Art. no. 72460Z.

[10] D. L. Helsel, J. M. Jakicic, and A. D. Otto, "Comparison of techniques for self-monitoring eating and exercise behaviors on weight loss in a correspondence-based intervention," *J. Amer. Dietetic Assoc.*, vol. 107, no. 10, pp. 1807–1810, 2007. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0002822307014800

[11] L. E. Burke, J. Wang, and M. A. Sevick, "Self-monitoring in weight loss: A systematic review of the literature," *J. Amer. Dietetic Assoc.*, vol. 111, no. 1, pp. 92–102, 2011. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0002822310016445

[12] K. J. Oh, M. D. Hong, S. Y. Sim, and G. S. Jo, "Automatic indexing of cooking video by using caption-recipe alignment," in *Proc. Int. Conf. Behavior Econ. Social Comput.*, Oct. 2014, pp. 1–6.

[13] Y. Hayashi, K. Doman, I. Ide, D. Deguchi, and H. Murase, "Automatic authoring of a domestic cooking video based on the description of cooking instructions," in *Proc. Workshop Multimedia Cooking Eating Activities*, 2013, pp. 21–26. [Online]. Available: http://doi.acm.org/10.1145/2506023.2506028

[14] Y. Ji, Y. Ko, A. Shimada, H. Nagahara, and R.-I. Taniguchi, "Cooking gesture recognition using local feature and depth image," in *Proc. Workshop Multimedia Cooking Eating Activities*, 2012, pp. 37–42. [Online]. Available: http://doi.acm.org/10.1145/2390776.2390785

[15] A. Iscen and P. Duygulu, "Knives are picked before slices are cut: Recognition through activity sequence analysis," in *Proc. Workshop Multimedia Cooking Eating Activities*, 2013, pp. 3–8. [Online]. Available: http://doi.acm.org/10.1145/2506023.2506025

[16] D. Elsweiler and M. Harvey, "Towards automatic meal plan recommendations for balanced nutrition," in *Proc. ACM Conf. Recommender Syst.*, 2015, pp. 313–316. [Online]. Available: http://doi.acm.org/10.1145/2792838.2799665

[17] F.-F. Kuo, C.-T. Li, M.-K. Shan, and S.-Y. Lee, "Intelligent menu planning: Recommending set of recipes by ingredients," in *Proc. Workshop Multimedia Cooking Eating Activities*, 2012, pp. 1–6. [Online]. Available: http://doi.acm.org/10.1145/2390776.2390778

[18] B. K. Seljak, "Dietary menu planning using an evolutionary method," in *Proc. Int. Conf. Intell. Eng. Syst.*, 2006, pp. 108–113.

[19] Y. Kawano, T. Sato, T. Maruyama, and K. Yanai, "[Demo paper] Mirurecipe: A mobile cooking recipe recommendation system with food ingredient recognition," in *Proc. IEEE Int. Conf. Multimedia Expo. Workshops*, Jul. 2013, pp. 1–2.

[20] A. Yajima and I. Kobayashi, ""Easy" cooking recipe recommendation considering user's conditions," in *Proc. Int. Joint Conf. Web Intell. Intell. Agent Technol.*, Sep. 2009, vol. 3, pp. 13–16.

[21] L. Buykx and H. Petrie, "What cooks needs from multimedia and textually enhanced recipes," in *Proc. IEEE Int. Symp. Multimedia*, Dec. 2011, pp. 387–392.

[22] K. Doman, C. Y. Kuai, T. Takahashi, I. Ide, and H. Murase, "Smart video-cooking: A multimedia cooking recipe browsing application on portable devices," in *Proc. ACM Int. Conf. Multimedia*, 2012, pp. 1267–1268. [Online]. Available: http://doi.acm.org/10.1145/2393347.2396435

[23] R. Hamada, J. Okabe, I. Ide, S. Satoh, and S. Sakai, "Cooking navi: Assistant for daily cooking in kitchen," in *Proc. ACM Int. Conf. Multimedia*, 2005, pp. 371–374. [Online]. Available: http://doi.acm.org/10.1145/1101149.1101228
[24] I. Ide, Y. Shidochi, Y. Nakamura, D. Deguchi, and T. Takahashi, "Multimedia supplementation to a cooking recipe text for facilitating its understanding to inexperienced users," in *Proc. IEEE Int. Symp. Multimedia*, Dec. 2010, pp. 242–247.
[25] Y. Kawano and K. Yanai, "Foodcam: A real-time mobile food recognition system employing fisher vector," in *Proc. Int. Conf. Multimedia Model.*, 2014, pp. 369–373.
[26] R. Kusumoto, X. H. Han, and Y. W. Chen, "Sparse model in hierarchic spatial structure for food image recognition," in *Proc. Int. Conf. Biomed. Eng. Informat.*, Dec. 2013, pp. 851–855.
[27] Y. Matsuda and K. Yanai, "Multiple-food recognition considering co-occurrence employing manifold ranking," in *Proc. Int. Converence Pattern Recog.*, 2012, pp. 2017–2020.
[28] S. Abbar, Y. Mejova, and I. Weber, "You tweet what you eat: Studying food consumption through twitter," in *Proc. ACM Conf. Human Factors Comput. Syst.*, 2015, pp. 3197–3206. [Online]. Available: http://doi.acm.org/10.1145/2702123.2702153
[29] M. De Choudhury, S. Sharma, and E. Kiciman, "Characterizing dietary choices, nutrition, and language in food deserts via social media," in *Proc. ACM Conf. Comput.-Supported Cooperative Work Social Comput.*, 2016, pp. 1157–1170. [Online]. Available: http://doi.acm.org/10.1145/2818048.2819956
[30] R. West, R. W. White, and E. Horvitz, "From cookies to cooks: Insights on dietary patterns via analysis of web usage logs," in *Proc. Int. Conf. World Wide Web*, 2013, pp. 1399–1410. [Online]. Available: http://doi.acm.org/10.1145/2488388.2488510
[31] J. Noronha, E. Hysen, H. Zhang, and K. Z. Gajos, "Platemate: Crowdsourcing nutritional analysis from food photographs," in *Proc. ACM Symp. User Interface Softw. Technol.*, 2011, pp. 1–12. [Online]. Available: http://doi.acm.org/10.1145/2047196.2047198
[32] Y. Kawano and K. Yanai, "Food image recognition with deep convolutional features," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput. Adjunct Publication*, 2014, pp. 589–593. [Online]. Available: http://doi.acm.org/10.1145/2638728.2641339
[33] D. T. Nguyen, Z. Zong, P. O. Ogunbona, Y. Probst, and W. Li, "Food image classification using local appearance and global structural information," *Neurocomputing*, vol. 140, pp. 242–251, 2014.
[34] R. Jain and P. Sinha, "Content without context is meaningless," in *Proc. ACM Int. Conf. Multimedia*, 2010, pp. 1259–1268.
[35] M. Chen, K. Dhingra, W. Wu, L. Yang, and R. Sukthankar, "PFID: Pittsburgh fast-food image dataset," in *Proc. IEEE Int. Conf. Image Process.*, Nov. 2009, pp. 289–292.
[36] H. Hoashi, T. Joutou, and K. Yanai, "Image recognition of 85 food categories by feature fusion," in *Proc. IEEE Int. Symp. Multimedia*, Dec. 2010, pp. 296–301.
[37] Y. Kawano and K. Yanai, "Automatic expansion of a food image dataset leveraging existing categories with domain adaptation," in *Proc. ECCV Workshop Transferring Adapting Source Knowl. Comput. Vis.*, 2014, pp. 3–17.
[38] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1106–1114.
[39] J. Donahue *et al.*, "DeCAF: A deep convolutional activation feature for generic visual recognition," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 647–655.
[40] K. Yanai and Y. Kawano, "Food image recognition using deep convolutional network with pre-training and fine-tuning," in *Proc. IEEE Int. Conf. Multimedia Expo. Workshops*, Jun. 2015, pp. 1–6.
[41] S. Christodoulidis, M. Anthimopoulos, and S. Mougiakakou, "Food recognition for dietary assessment using deep convolutional neural networks," in *Proc. Int. Conf. Image Anal. Process. Workshops*, 2015, pp. 458–465.
[42] B. Girod, V. Chandrasekhar, R. Grzeszczuk, and Y. Reznik, "Mobile visual search: Architectures, technologies, and the emerging MPEG standard," *IEEE Multimedia*, vol. 18, no. 3, pp. 86–94, Mar. 2011.
[43] Z. Li and K.-H. Yap, "Content and context boosting for mobile landmark recognition," *IEEE Signal Process. Lett.*, vol. 19, no. 8, pp. 459–462, Aug. 2012.
[44] T. Li, T. Mei, I.-S. Kweon, and X.-S. Hua, "Contextual bag-of-words for visual categorization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 4, pp. 381–392, Apr. 2011.
[45] V. Bettadapura, E. Thomaz, A. Parnami, G. D. Abowd, and I. Essa, "Leveraging context to support automated food recognition in restaurants," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, Jan. 2015, pp. 580–587.
[46] R. Xu, L. Herranz, S. Jiang, S. Wang, and X. Song, "Geolocalized modeling for dish recognition," *IEEE Trans. Multimedia*, vol. 17, no. 8, pp. 1187–1199, Aug. 2015.
[47] K.-H. Yap, T. Chen, Z. Li, and K. Wu, "A comparative study of mobile-based landmark recognition techniques," *IEEE Intell. Syst.*, vol. 25, no. 1, pp. 48–57, Jan. 2010.
[48] L. Bossard, M. Guillaumin, and L. Gool, "Food-101. mining discriminative components with random forests," in *Proc. Eur. Conf. Comput. Vis.*, Jan. 2014, pp. 446–461.
[49] Y. He, C. Xu, N. Khanna, C. J. Boushey, and E. J. Delp, "Analysis of food images: Features and classification," in *Proc. IEEE Int. Conf. Image Process.*, Oct. 2014, pp. 2744–2748.
[50] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. ACM Int. Conf. Multimedia*, 2014, pp. 675–678.
[51] R. Fan, K. Chang, C. Hsieh, X. Wang, and C. Lin, "Liblinear: A library for large linear classification," *J. Mach. Learn. Res.*, vol. 9, pp. 1871–1874, 2008.

Luis Herranz received the Ph.D. degree in computer science and telecommunication from the Universidad Autónoma de Madrid, Spain, in 2010.

From 2003 to 2010, he was with the Escuela Politécnica Superior, Universidad Autónoma de Madrid as a Researcher and Teaching Assistant. From 2010 to 2011, he was with the Mitsubishi Electric R&D Centre Europe, U.K. He is currently a Postdoctoral Research Fellow with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China. His research interests include image understanding, computer vision, machine learning, and multimedia indexing and retrieval.

Shuqiang Jiang (S'04–M'06–SM'08) is a Professor with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China. He is also with the Key Laboratory of Intelligent Information Processing, Chinese Academy of Sciences. His research interests include multimedia processing and semantic understanding, pattern recognition, and computer vision. He has authored or coauthored more than 100 papers on the related research topics.

Prof. Jiang is a Senior Member of CCF and a Member of ACM. He is an Associate Editor of the *IEEE MultiMedia Magazine* and *Multimedia Tools and Applications*. He was supported by the New-Star program of Science and Technology of Beijing Metropolis in 2008, the NSFC Excellent Young Scientists Fund in 2013, and the Young Top-Notch Talent of Ten Thousand Talent Program in 2014. He is the Vice Chair of ACM SIGMM China chapter, and the General Secretary of IEEE CASS Beijing Chapter. He is the General Chair of ICIMCS 2015, Program Chair of ICIMCS2010, Special Session Chair of PCM2008 and ICIMCS2012, Area Chair of PCIVT2011, Publicity Chair of PCM2011, Web Chair of ISCAS2013, and Proceedings Chair of MMSP2011. He has also served as a TPC Member for more than 20 conferences, including ACM Multimedia, CVPR, ICCV, ICME, ICIP, and PCM. He was the recipient of the Lu Jiaxi Young Talent Award from the Chinese Academy of Sciences in 2012, and the CCF Award of Science and Technology in 2012.

Ruihan Xu received the B.S. degree in automation from Xidian University, Xi' an, China, in 2012, and the M.Eng. degree in computer science from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2015.

His current research interests include multimedia content analysis, computer vision, and pattern recognition.