A PROBABILISTIC MODEL FOR FOOD IMAGE RECOGNITION IN RESTAURANTS

Luis Herranz, Ruihan Xu, Shuqiang Jiang

Key Lab of Intell. Info. Process, Chinese Academy of Sciences, Beijing 100190, China

ABSTRACT

A large amount of food photos are taken in restaurants for diverse reasons. This dish recognition problem is very challenging, due to different cuisines, cooking styles and the intrinsic difficulty of modeling food from its visual appearance. Contextual knowledge is 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 geolocation 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 geolocations. We apply that model in three different tasks: dish recognition, restaurant recognition and geolocation refinement. Experiments on a dataset including 187 restaurants and 701 dishes show that combining multiple evidences (visual, geolocation, and external knowledge) can boost the performance in all tasks.

Index Terms— food recognition, geolocation, mobile

1. INTRODUCTION

Food images are present in many multimedia applications, including food logs[1], dietary assessment systems[2] and food-related social networks where users can share their recipes and culinary experiences. This has motivated an increasing interest in automatic food recognition. Early works were able to classify among a few dozen types of food[3, 4, 5]. Recently, Kawana and Yanai[6] proposed a mobile food recognition system that can recognize 256 food categories. However, large-scale food recognition, covering multiple cuisines and fine-grained classification, is still a very challenging problem.

In order to address complex recognition problems, humans incorporate prior and contextual knowledge. Intelligent systems can also leverage external knowledge to simplify the problem. The most representative example is mobile recognition of landmarks[7, 8] based on geolocation and image retrieval techniques to find photos of the same landmark from geotagged photo databases, and use them to annotate the test image. Geolocation 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[9] or using vocabulary trees[8, 7]. As landmarks are rigid and geometrically almost invariant, retrieving similar images and performing geometric verification often finds the corresponding landmark[8, 7]. Classifiers can also be used instead of retrieval techniques. In this case geolocation helps to restrict the classification to the landmarks in the geographic neighborhood (i.e. shortlists the candidate classes).

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 shared in social networks, used to find information about the dish, or keep a personal record of favorite dish. The user is often not familiar with the particular dish or even the restaurant (e.g. 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 location) and exploit geolocation to simplify the problem and improve the performance.

We adopt a probabilistic approach, because allows us to design flexible models for each of the components of the problem, and often leads to improved performance. Thus, we propose a probabilistic model that connects locations, restaurants, dishes and visual features. By combining visual and geolocation 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.

This work was supported in part by National Basic Research Program of China (973 Program): 2012CB316400, in part by the National Natural Science Foundation of China: 61322212 and 61450110446, in part by National Hi-Tech Development Program (863 Program) of China: 2014AA015202, in part by the Key Technologies R&D Program of China: 2012BAH18B02, and in part by the CAS President's International Fellowship Initiative: 2011Y1GB05. This work is also funded by Lenovo Outstanding Young Scientists Program (LOYS).

We would also like to thank José Miguel Hernández-Lobato and Daniel Hernández-Lobato for their suggestions.

2. DISH RECOGNITION IN RESTAURANTS

The objective of food or dish recognition is to identify the class s of an input image, represented by some visual descriptor $\mathbf{x}$. This is achieved using certain visual classifier $p(s|\mathbf{x})$. We consider the particular problem of dish recognition *in restaurants*, assuming that the user is located in a restaurant. Thus, in addition to the visual classifier, we have access to the menu of the restaurant and to the geographic location of both the restaurants and the user.

2.1. Framework

The input to the recognition system is a pair $(\boldsymbol{\mu}_q, \mathbf{x})$, where $\boldsymbol{\mu}_q$ are the local coordinates (estimated by the location services of the device) and the visual descriptor $\mathbf{x}$. When a new image is captured, we assume that the mobile phone has an estimation of its current location $\Psi_q = (\lambda_q, \phi_q)$, with λ_q and ϕ_q denoting latitude and longitude.

Similarly, the main properties of a restaurant k are its menu M_k (i.e. the dish categories found in that particular restaurant) and its geographic location $\boldsymbol{\Psi_k} = (\lambda_k, \phi_k)$. For simplicity we project the location onto a local coordinate system (with origin at the average coordinates of the dataset), and use 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.

2.2. Shortlist approach

First, we adapt the shortlist approach, commonly used in landmark recognition[10]. This approach is based on the reasonable assumption that the user is likely to be in one of the landmark or buildings within a small area centered at $\boldsymbol{\mu}_q$.

Similarly, we assume that the user is located in one of the restaurants within a geographical neighborhood, so only the dishes in the menus of those restaurants are likely to be the true dish. Thus, 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

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

The set of candidate restaurants H_q is obtained as

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

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

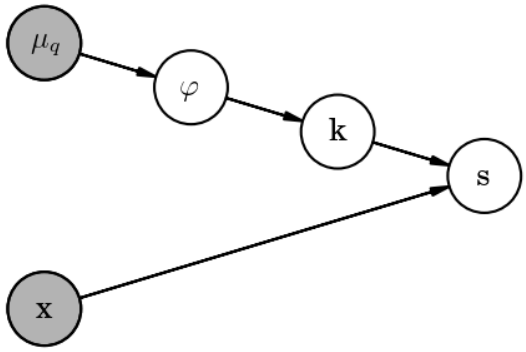

Fig. 1. The proposed probabilistic model, where the estimated location $\boldsymbol{\mu}_q$ and the visual feature $\mathbf{x}$ are observed variables, and the actual location $\boldsymbol{\varphi}$, the restaurant k and the dish s are hidden variables.

3. PROBABILISTIC MODEL

3.1. Model

We model the problem as the generative process represented in Fig. 1, in which the device provides the estimated location $\boldsymbol{\mu}_q$ and the visual feature $\mathbf{x}$. We introduce explicitly the dependency between the restaurant and the dish (via the menu), the visual feature and the dish (via the 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.

From the graphical model, we can obtain the joint distribution $p\left(s, k, \boldsymbol{\varphi}|\boldsymbol{\mu}_q, \mathbf{x}\right)$ given the coordinates $\boldsymbol{\mu}_q$ and the visual feature $\mathbf{x}$ as

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

We can identify three submodels: 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 $\boldsymbol{\varphi}$

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

The predicted dish can be obtained by solving

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

3.2. Revisiting the shortlist approach

We first briefly review the shortlist approach from the perspective of the model described in Fig. 1 by comparing (1) and (3). The neighborhood model is simply a circle of radius ϵ centered at $\boldsymbol{\mu_q}$

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

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

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

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

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

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

$$p_{SL}\left(s|\boldsymbol{\mu}_{\mathbf{q}}, \mathbf{x}\right) \propto p\left(s|\mathbf{x}\right)\left[s \in \bigcup_{k \in H_q} M_k\right] \quad (9)$$

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

3.3. An alternative model

Variations of the shortlist approach have been widely used combined with retrieval or classification techniques for landmark recognition. However, both the neighborhood and the restaurant location models have obvious limitations.

The hard-threshold neighborhood model consider 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. Thus, instead of (6), we use a Gaussian model for the neighborhood

$$p_G\left(\boldsymbol{\varphi}|\boldsymbol{\mu}_{\boldsymbol{q}}\right) = \mathcal{N}\left(\boldsymbol{\varphi}|\boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q\right) \quad (10)$$

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

Representing a restaurant with a point is not realistic, as they cover certain 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 a Gaussian model

$$p_G\left(k|\boldsymbol{\varphi}\right) = \mathcal{N}\left(\boldsymbol{\varphi}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right) \quad (11)$$

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

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 $p\left(s|k\right) = \frac{\left[s \in M_k\right]}{|M_k|}$. The resulting restaurant-dependent visual model is

$$p_R\left(s|k, \mathbf{x}\right) = p\left(s|\mathbf{x}\right)\frac{\left[s \in M_k\right]}{|M_k|} \quad (12)$$

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

$$p\left(s|\boldsymbol{\mu}_{\mathbf{q}}, \mathbf{x}\right) \propto p\left(s|\mathbf{x}\right)\sum_{k=1}^{K}\frac{\left[s \in M_k\right]}{|M_k|}\mathcal{N}\left(\boldsymbol{\mu}_k|\boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q + \boldsymbol{\Sigma}_k\right) \quad (13)$$

4. OTHER APPLICATIONS

So far, we have performed inference over the joint distribution $p\left(s, k, \boldsymbol{\varphi}|\boldsymbol{\mu}_q, \mathbf{x}\right)$ to predict 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 .3.3.

4.1. Restaurant recognition

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

$$p\left(k|\boldsymbol{\mu}_q, \mathbf{x}\right) = \sum_{s=1}^{D} p\left(s|k, \mathbf{x}\right)\int_{\boldsymbol{\varphi}} p\left(\boldsymbol{\varphi}|\boldsymbol{\mu}_q\right) p\left(k|\boldsymbol{\varphi}\right)\mathrm{d}\boldsymbol{\varphi} \quad (14)$$

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

$$p\left(k|\boldsymbol{\mu}_q, \mathbf{x}\right) \propto \mathcal{N}\left(\boldsymbol{\mu}_k|\boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q + \boldsymbol{\Sigma}_k\right)\frac{\sum_{s \in M_k} p\left(s|\mathbf{x}\right)}{|M_k|} \quad (15)$$

The predicted restaurant is obtained as

$$k^* = \underset{k \in \{1, \ldots K\}}{\arg\max}\, p\left(k|\boldsymbol{\mu}_q, \mathbf{x}\right) \quad (16)$$

4.2. Location refinement

As a byproduct, the proposed model can leverage restaurant location data and visual evidence to improve the initial estimation of the location. This is particularly useful in indoor environments with restaurants such as shopping malls, where some location signals (e.g. GPS) may not be available.

Marginalizing (3) over s and k we obtain

$$p\left(\boldsymbol{\varphi}|\boldsymbol{\mu}_q, \mathbf{x}\right) = p\left(\boldsymbol{\varphi}|\boldsymbol{\mu}_q\right)\sum_{k=1}^{K} p\left(k|\boldsymbol{\varphi}\right)\sum_{s=1}^{S} p\left(s|k, \mathbf{x}\right) \quad (17)$$

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

$$p\left(\boldsymbol{\varphi}|\boldsymbol{\mu}_q, \mathbf{x}\right) \propto \sum_{k=1}^{K}\omega_k\mathcal{N}\left(\boldsymbol{\varphi}|\boldsymbol{\theta}_k, \boldsymbol{\Lambda}_k\right) \quad (18)$$

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 $\boldsymbol{\Lambda}_k$, $\boldsymbol{\theta}_k$ and ω_k using (19), (20) and (21)
end for
Initialize $\boldsymbol{\varphi} = \boldsymbol{\mu}_q$
repeat
  for $k = 1 : K$ **do**
    Compute $\gamma_k(\boldsymbol{\varphi})$ using (23)
  end for
  Update estimated location $\boldsymbol{\varphi}$ using (22)
until *converged*
return $\boldsymbol{\varphi}$

with

$$\boldsymbol{\Lambda}_k = \left(\boldsymbol{\Sigma}_q^{-1} + \boldsymbol{\Sigma}_k^{-1}\right)^{-1} \tag{19}$$

$$\boldsymbol{\theta}_k = \boldsymbol{\Lambda}_k \left(\boldsymbol{\Sigma}_q^{-1}\boldsymbol{\mu}_q + \boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k\right) \tag{20}$$

$$\omega_k = \frac{\sum_{s \in M_k} p(s|\mathbf{x})}{|M_k|} \tag{21}$$

In (18) 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 the restaurant, the location $\boldsymbol{\varphi}$ is a continuous variable. To find the location that maximizes (18) 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

$$\boldsymbol{\varphi} = \frac{1}{\sum_{j=1}^{K} \gamma_j(\boldsymbol{\varphi})} \sum_{k=1}^{K} \gamma_k(\boldsymbol{\varphi}) \boldsymbol{\theta}_k \tag{22}$$

where we define

$$\gamma_k(\boldsymbol{\varphi}) = \frac{\omega_k \mathcal{N}(\boldsymbol{\varphi}|\boldsymbol{\theta}_k, \boldsymbol{\Lambda}_k)}{\sum_{j=1}^{K} \omega_j \mathcal{N}(\boldsymbol{\varphi}|\boldsymbol{\theta}_j, \boldsymbol{\Lambda}_j)} \tag{23}$$

Unfortunately, (22) is not a closed-form expression due to the dependency of $\gamma_k(\boldsymbol{\varphi})$ on $\boldsymbol{\varphi}$. However we can alternatively compute and update their estimations (see Algorithm 1).

5. EXPERIMENTS

5.1. Dataset and settings

Current food benchmarks do not include restaurant nor geographic location. For that reason we collected our own dataset to evaluate the proposed method. Restaurant information was collected from a restaurant review website[1], where users post their own photos of specific dishes taken in an specific restaurant. We crawled a large city and selected all the restaurants with at least three different dishes (i.e. menu) and at least 15 images per dish. We use 10 images for training and the rest as test images. The geographic location of the restaurant is also collected from the website. Our dataset contains a total of 187 restaurants and 701 unique dish categories (considering the same dish in different restaurants as different categories the dataset would contain 1173 dish categories). We tried SVMs over two types of visual features: bag-of-words with LLC[11] and deep features with DeCAF[12].

[1] http://www.dianping.com

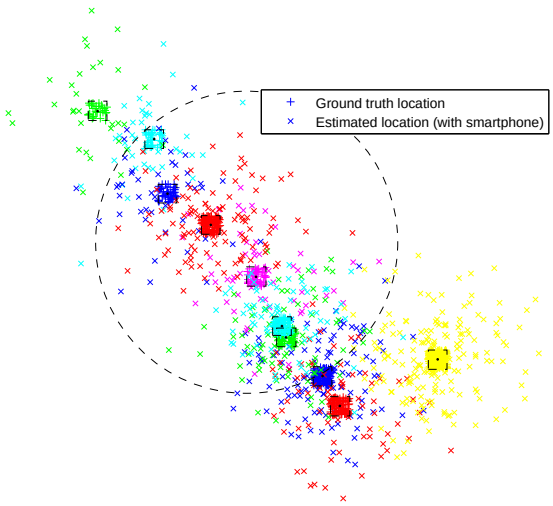

Fig. 2. Example of dense neighborhood with several restaurants and the simulated test locations ($\sigma_{LOC} = 40$ meters). A neighborhood of radius 200 meters is shown for reference.

We simulated the location of the test images assuming a simple location model for the restaurant (see Fig. 2 for an example). The layout and dimensions of the restaurants are not available, so we modeled restaurants as squares of 25x25 meters centered in the coordinates collected from the restaurant website, and the location of the user can be any place in that square (randomly sampled from a uniform distribution). Note that this model is very different from the model used in the proposed method (Gaussian). In this way, we simulated the location for the test queries and then we add Gaussian noise (zero mean and $\sigma_{LOC} = 40$ meters) to simulate the error due to the smartphone's location service error.

We evaluated the different methods for different values of the radius ϵ (we used [10:20:90, 100:100:1000]). Note that both ϵ is a parameter of the shortlist method, which cannot be compared directly with the parameter σ_q of the probabilistic method. After first inspecting the trends in the dish and restaurant recognition accuracy, for better comparison we align them using $\epsilon = 3\sigma_q$. 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 $5\sigma_q$.

Table 1. Dish recognition accuracy.

Feature	Radius ϵ (SL) $3\sigma_q$ (PR)	Accuracy (%) All (≥ 0 restaurants) VS	SL	PR	Dense (≥ 5 restaurants) VS	SL	PR
LLC	50	21.72	30.80	52.56	N/A	N/A	N/A
	200		53.30	**54.54**	23.15	45.15	48.64
	500		49.31	52.33	23.27	43.47	48.42
	1000		44.90	48.99	23.14	42.23	45.66
	Best		53.30	**54.54**		45.94	**48.79**
	(ϵ, $3\sigma_q$)		(200)	(200)		(100)	(400)
DeCAF	50	48.35	42.23	76.30	N/A	N/A	N/A
	200		76.61	**77.58**	48.64	68.04	71.78
	500		74.05	76.54	49.58	68.46	72.58
	1000		71.32	74.47	50.04	68.81	72.51
	Best		76.61	**77.62**		70.73	**72.88**
	(ϵ, $3\sigma_q$)		(200)	(100)		(100)	(400)

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

Table 2. Restaurant recognition accuracy.

Feature	Radius ϵ (SL) $3\sigma_q$ (PR)	Accuracy (%) All (≥ 0 restaurants) LC	SL	PR	Dense (≥ 5 restaurants) LC	SL	PR
LLC	50	82.69	51.42	88.56	N/A	N/A	N/A
	200		87.85	**92.69**	53.77	68.55	81.36
	500		79.43	90.31	69.78	64.75	82.87
	1000		70.38	85.44	75.06	63.14	80.33
	Best		88.79	**92.69**		79.10	**83.87**
	(ϵ, $3\sigma_q$)		(100)	(200)		(70)	(400)
DeCAF	50	82.69	52.37	94.35	N/A	N/A	N/A
	200		94.37	**96.14**	53.77	83.26	89.16
	500		90.23	93.73	69.78	82.55	87.96
	1000		85.63	90.40	75.06	81.71	87.19
	Best		94.37	**96.14**		89.55	**92.54**
	(ϵ, $3\sigma_q$)		(200)	(200)		(70)	(70)

LC: nearest (only location), **SL**: shortlist, **PR**: probabilistic.

Due to the scarcity of user-contributed data for most restaurants in the restaurant website, only a fraction of the restaurants meets the demanding requirements (at least 3 dishes and at least 15 images per dish). As a result the dataset is relatively sparse in geographic location and the most common case is finding only one restaurant in the neighborhood. However, the benefit of the proposed method is more evident in more complex cases where the density of restaurants is high (and consequently the number of candidate categories is significantly higher). For this reasons 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. 2 has a relatively high density of restaurants).

5.2. Dish recognition

We compare the average accuracy of the *shortlist* approach (Section 2.2) and the *probabilistic* method (Section 3.3) for different sizes of the neighborhood model (see Table 1). We also include the *visual* classifier (without considering location information) as baseline.

As expected, visual classification over DeCAF achieves a remarkable accuracy of 48.35%. Including information about the location increases the performance around 30% for both types of visual features, which makes the system much more competitive. Both *shortlist* and *probabilistic* achieve a similar best accuracy, with the later being slightly better. However, *shortlist* is much more dependent on the specific choice of the neighborhood size ϵ, while the accuracy of *probabilistic* depends less on σ_q, and in general benefits from larger neighborhoods. If we focus on the more interesting case of dense areas, the problem is considerably harder and as a result the accuracy drops in both *shortlist* and *probabilistic*.

In practice the error in the estimation depends on many factors (e.g. indoor/outdor, availability of positioning signals, building density). We evaluated the performance varying the amount of location error σ_{LOC}. Fig. 3 compares both methods for $\epsilon = 200$ meters and $\epsilon = 1000$ meters. We observe that a smaller neighborhood provides higher accuracy but a rapid degradation of the performance if the location error is larger than expected. A larger neighborhood shows more robustness to variations in σ_{LOC} at the cost of some drop in accuracy. In both cases, *probabilistic* performs better than *shortlist*.

5.3. Restaurant recognition

We evaluate now the accuracy for restaurant recognition using the proposed probabilistic model (Section 4.1). 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 2.

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.

5.4. 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 3). We compare the restaurant location estimated using

Table 3. Location refinement error.

Feature	Radius ϵ (SL) $3\sigma_q$ (PR)	Average error (meters) All (≥ 0 restaurants) LC	SL	PR	Dense (≥ 5 restaurants) LC	SL	PR
LLC	50	49.97	34.63	25.26	N/A	N/A	N/A
	200		12.28	8.40	48.84	32.80	19.78
	500		46.61	6.91	50.04	76.96	13.08
	1000		133.38	**6.70**	50.28	159.31	**9.45**
	Best		8.50	**6.70**		15.56	**9.45**
	(ϵ, $3\sigma_q$)		(100)	(1000)		(70)	(1000)
DeCAF	50	49.97	34.43	24.11	N/A	N/A	N/A
	200		5.37	6.40	48.84	17.02	14.29
	500		21.24	4.74	50.04	35.91	9.24
	1000		62.70	**4.50**	50.28	77.10	**6.24**
	Best		5.37	**4.50**		8.97	**6.24**
	(ϵ, $3\sigma_q$)		(200)	(1000)		(90)	(1000)

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

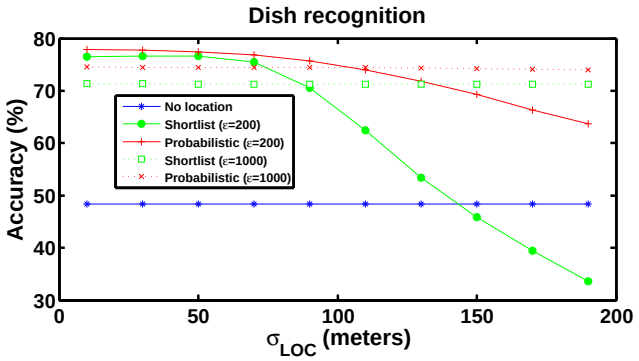

Fig. 3. Accuracy for different values of σ_{LOC}.

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, achieved by *probabilistic*. This 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. 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. Again, *probabilistic* can handle better these cases, achieving a remarkable error of only 6.24 meters in dense areas compared to 8.97 meters with *shortlist* in the best case.

6. CONCLUSIONS

Focusing on a dining out scenario, we describe an integrated approach to recognize the dish and restaurant and refine the estimation of the location by taking advantage of visual information, geo-context, and prior knowledge about the restaurants. Formulating the problem in a probabilistic framework allows us to perform inference over different hidden variables leading to different recognition tasks. We compare the proposed methods with the shortlist approach, widely used for landmark recognition, adapted to our restaurant scenario. The proposed method outperforms the shortlist approach and other baselines in all the tasks. We showed that not only knowing the geo-context can help dish recognition, but also having visual evidence about the dish can help to improve the estimation of the location via knowledge about the restaurant.

7. REFERENCES

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

[2] F. Kong and J. Tan, "Dietcam: Regular shape food recognition with a camera phone," in *BSN*, 2011, pp. 127–132.

[3] M. Chen, K. Dhingra, W. Wu, L. Yang, and R. Sukthankar, "Pfid: Pittsburgh fast-food image dataset," in *ICIP*, 2009, pp. 289–292.

[4] H. Hoashi, T. Joutou, and K. Yanai, "Image recognition of 85 food categories by feature fusion," in *ISM*, 2010, pp. 296–301.

[5] 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.

[6] Y. Kawano and K. Yanai, "Automatic expansion of a food image dataset leveraging existing categories with domain adaptation," in *Proc. of ECCV Workshop on Transferring and Adapting Source Knowledge in Computer Vision (TASK-CV)*, 2014.

[7] 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, March 2011.

[8] Z. Li and K.-H. Yap, "Content and context boosting for mobile landmark recognition," *IEEE Signal Processing Letters*, vol. 19, no. 8, pp. 459–462, Aug 2012.

[9] T. Li, T. Mei, I.-S. Kweon, and X.-S. Hua, "Contextual bag-of-words for visual categorization," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 21, no. 4, pp. 381–392, April 2011.

[10] K.-H. Yap, T. Chen, Z. Li, and K. Wu, "A comparative study of mobile-based landmark recognition techniques," *IEEE Intelligent Systems*, vol. 25, no. 1, pp. 48–57, Jan 2010.

[11] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *CVPR*, 2010.

[12] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, "DeCAF: A deep convolutional activation feature for generic visual recognition," in *ICML*, 2014.

Supplementary material for *A probabilistic model for food recognition in restaurants*

A. DETAILED DERIVATIONS

We provide details about the derivation of some equations.

A.1. Shortlist approach

Using the neighborhood model (6), the restaurant location model (7) and the visual model (8) in (4) we obtain

$$\begin{aligned} p\left(s|\boldsymbol{\mu}_{\mathbf{q}},\mathbf{x}\right) &\propto p\left(s|\mathbf{x}\right)\times\sum_{k=1}^{K}\left[s\in M_k\right] \\ &\times\int_{\boldsymbol{\varphi}}\left[\left\|\boldsymbol{\varphi}-\boldsymbol{\mu}_q\right\|\leq\epsilon\right]\delta\left(\left\|\boldsymbol{\varphi}-\boldsymbol{\mu}_k\right\|\right)\mathrm{d}\boldsymbol{\varphi} \\ &= p\left(s|\mathbf{x}\right)\sum_{k=1}^{K}\left[s\in M_k\right]\int_{\boldsymbol{\varphi}\in B_q}\delta\left(\left\|\boldsymbol{\varphi}-\boldsymbol{\mu}_k\right\|\right)\mathrm{d}\boldsymbol{\varphi} \\ &= p\left(s|\mathbf{x}\right)\left[s\in\bigcup_{k\in H_q}M_k\right] \end{aligned} \tag{24}$$

where $B_q=\left\{\boldsymbol{\varphi}|\left\|\boldsymbol{\varphi}-\boldsymbol{\mu}_q\right\|\leq\epsilon\right\}$ is the ϵ-circular neighborhood of the query, and $H_q=\left\{\boldsymbol{\varphi}|\left\|\boldsymbol{\varphi}-\boldsymbol{\mu}_q\right\|\leq\epsilon\right\}$.

A.2. An alternative model

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

$$\begin{aligned} p\left(s|\boldsymbol{\mu}_{\mathbf{q}},\mathbf{x}\right) \propto& p\left(s|\mathbf{x}\right)\times\sum_{k=1}^{K}\frac{\left[s\in M_k\right]}{|M_k|} \\ &\times\int_{\boldsymbol{\varphi}}\mathcal{N}\left(\boldsymbol{\varphi}|\boldsymbol{\mu}_q,\boldsymbol{\Sigma}_q\right)\times\mathcal{N}\left(\boldsymbol{\varphi}|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k\right)\mathrm{d}\boldsymbol{\varphi} \end{aligned} \tag{25}$$

Using the following relation for the product of two multivariate Gaussians

$$\begin{aligned} &\mathcal{N}\left(\boldsymbol{\varphi}|\boldsymbol{\mu}_q,\boldsymbol{\Sigma}_q\right)\mathcal{N}\left(\boldsymbol{\varphi}|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k\right) \\ =&\mathcal{N}\left(\boldsymbol{\mu}_k|\boldsymbol{\mu}_q,\boldsymbol{\Sigma}_q+\boldsymbol{\Sigma}_k\right)\mathcal{N}\left(\boldsymbol{\varphi}|\boldsymbol{\theta},\boldsymbol{\Lambda}\right) \\ &\boldsymbol{\Lambda}=\left(\boldsymbol{\Sigma}_q^{-1}+\boldsymbol{\Sigma}_k^{-1}\right)^{-1} \\ &\boldsymbol{\theta}=\boldsymbol{\Lambda}\left(\boldsymbol{\Sigma}_q^{-1}\boldsymbol{\mu}_q+\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k\right) \end{aligned} \tag{26}$$

in (25) we further obtain

$$\begin{aligned} p\left(s|\boldsymbol{\mu}_{\mathbf{q}},\mathbf{x}\right) \propto& p\left(s|\mathbf{x}\right)\times\sum_{k=1}^{K}\frac{\left[s\in M_k\right]}{|M_k|} \\ &\times\int_{\boldsymbol{\varphi}}\mathcal{N}\left(\boldsymbol{\mu}_k|\boldsymbol{\mu}_q,\boldsymbol{\Sigma}_q+\boldsymbol{\Sigma}_k\right)\times\mathcal{N}\left(\boldsymbol{\varphi}|\boldsymbol{\theta},\boldsymbol{\Lambda}\right)\mathrm{d}\boldsymbol{\varphi} \\ \propto& \ p\left(s|\mathbf{x}\right)\sum_{k=1}^{K}\frac{\left[s\in M_k\right]}{|M_k|}\mathcal{N}\left(\boldsymbol{\mu}_k|\boldsymbol{\mu}_q,\boldsymbol{\Sigma}_q+\boldsymbol{\Sigma}_k\right) \end{aligned} \tag{27}$$

A.3. Restaurant recognition

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

$$\begin{aligned} p\left(k|\boldsymbol{\mu}_q,\mathbf{x}\right) &= \sum_{s=1}^{D}\int_{\boldsymbol{\varphi}}p\left(s,k,\boldsymbol{\varphi}|\boldsymbol{\mu}_q,\mathbf{x}\right)\mathrm{d}\boldsymbol{\varphi} \\ &= \sum_{s=1}^{D}p\left(s|k,\mathbf{x}\right)\int_{\boldsymbol{\varphi}}p\left(\boldsymbol{\varphi}|\boldsymbol{\mu}_q\right)p\left(k|\boldsymbol{\varphi}\right)\mathrm{d}\boldsymbol{\varphi} \end{aligned} \tag{28}$$

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

$$\begin{aligned} p\left(k|\boldsymbol{\mu}_q,\mathbf{x}\right) &\propto \sum_{s=1}^{D}p\left(s|\mathbf{x}\right)\times\frac{\left[s\in M_k\right]}{|M_k|} \\ &\times\int_{\boldsymbol{\varphi}}\mathcal{N}\left(\boldsymbol{\varphi}|\boldsymbol{\mu}_q,\boldsymbol{\Sigma}_q\right)\times\mathcal{N}\left(\boldsymbol{\varphi}|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k\right)\mathrm{d}\boldsymbol{\varphi} \\ &\propto \mathcal{N}\left(\boldsymbol{\mu}_k|\boldsymbol{\mu}_q,\boldsymbol{\Sigma}_q+\boldsymbol{\Sigma}_k\right)\frac{\sum_{s\in M_k}p\left(s|\mathbf{x}\right)}{|M_k|} \end{aligned} \tag{29}$$

The predicted restaurant is obtained as

$$k^*=\underset{k\in\{1,\ldots K\}}{\arg\max}\,p\left(k|\boldsymbol{\mu}_q,\mathbf{x}\right) \tag{30}$$

A.4. Location refinement

Marginalizing (3) over s and k we obtain

$$\begin{aligned} p\left(\boldsymbol{\varphi}|\boldsymbol{\mu}_q,\mathbf{x}\right) &= \sum_{k=1}^{K}\sum_{s=1}^{D}p\left(s,k,\boldsymbol{\varphi}|\boldsymbol{\mu}_q,\mathbf{x}\right) \\ &= p\left(\boldsymbol{\varphi}|\boldsymbol{\mu}_q\right)\sum_{k=1}^{K}p\left(k|\boldsymbol{\varphi}\right)\sum_{s=1}^{S}p\left(s|k,\mathbf{x}\right) \end{aligned} \tag{31}$$

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

$$p\left(\boldsymbol{\varphi}|\boldsymbol{\mu}_q,\mathbf{x}\right)\propto\sum_{k=1}^{K}\omega_k\mathcal{N}\left(\boldsymbol{\varphi}|\boldsymbol{\theta}_k,\boldsymbol{\Lambda}_k\right) \tag{32}$$

with

$$\boldsymbol{\Lambda}_k=\left(\boldsymbol{\Sigma}_q^{-1}+\boldsymbol{\Sigma}_k^{-1}\right)^{-1} \tag{33}$$

$$\boldsymbol{\theta}_k=\boldsymbol{\Lambda}_k\left(\boldsymbol{\Sigma}_q^{-1}\boldsymbol{\mu}_q+\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k\right) \tag{34}$$

$$\omega_k=\frac{\sum_{s\in M_k}p\left(s|\mathbf{x}\right)}{|M_k|} \tag{35}$$

Fig. 4. Geographic distribution of the restaurants in the dataset.

To find the location that maximizes (18) 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

$$\begin{aligned} 0 &= -\sum_{k=1}^{K} \frac{\omega_k \mathcal{N}\left(\boldsymbol{\varphi}|\boldsymbol{\theta}_k, \boldsymbol{\Lambda}_k\right)}{\sum_{j=1}^{K} \omega_j \mathcal{N}\left(\boldsymbol{\varphi}|\boldsymbol{\theta}_j, \boldsymbol{\Lambda}_j\right)} \boldsymbol{\Lambda}_k \left(\boldsymbol{\varphi} - \boldsymbol{\theta}_k\right) \\ &= -\sum_{k=1}^{K} \gamma_k\left(\boldsymbol{\varphi}\right) \boldsymbol{\Lambda}_k \left(\boldsymbol{\varphi} - \boldsymbol{\theta}_k\right) \end{aligned} \tag{36}$$

where we define

$$\gamma_k\left(\boldsymbol{\varphi}\right) = \frac{\omega_k \mathcal{N}\left(\boldsymbol{\varphi}|\boldsymbol{\theta}_k, \boldsymbol{\Lambda}_k\right)}{\sum_{j=1}^{K} \omega_j \mathcal{N}\left(\boldsymbol{\varphi}|\boldsymbol{\theta}_j, \boldsymbol{\Lambda}_j\right)} \tag{37}$$

As $\boldsymbol{\Lambda}_k$ is not singular, we can multiply (36) by $\boldsymbol{\Lambda}_k^{-1}$ and rearrange to obtain

$$\boldsymbol{\varphi} = \frac{1}{\sum_{j=1}^{K} \gamma_j\left(\boldsymbol{\varphi}\right)} \sum_{k=1}^{K} \gamma_k\left(\boldsymbol{\varphi}\right) \boldsymbol{\theta}_k \tag{38}$$

B. DATASET

Fig. 4 shows the geographic distribution of restaurants in the dataset. Due to the availability of data in the review website, and the minimum requirements (at least 3 dishes and 15 images per dish), there are some regions with isolated restaurants, and some regions with higher density.

C. EXPERIMENTS

Figs. 5, 6 and 7 show more detailed experimental results for dish recognition, restaurant recognition and location refinement, respectively.

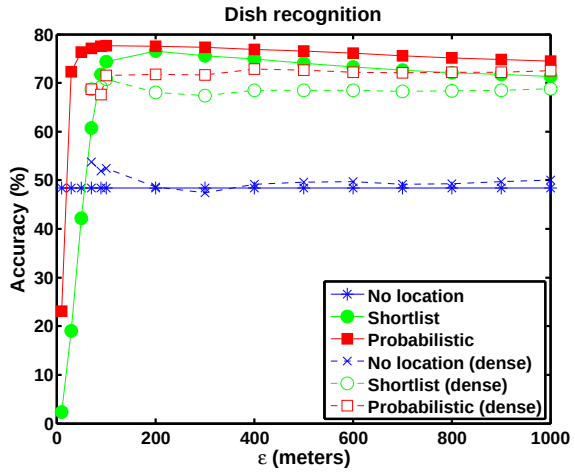

Fig. 5. Dish recognition accuracy.

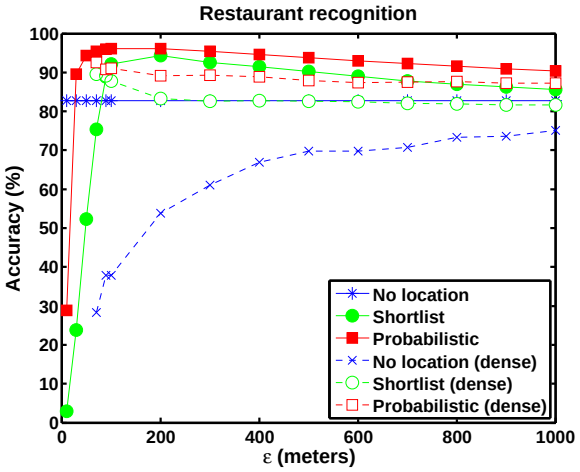

Fig. 6. Restaurant recognition accuracy.

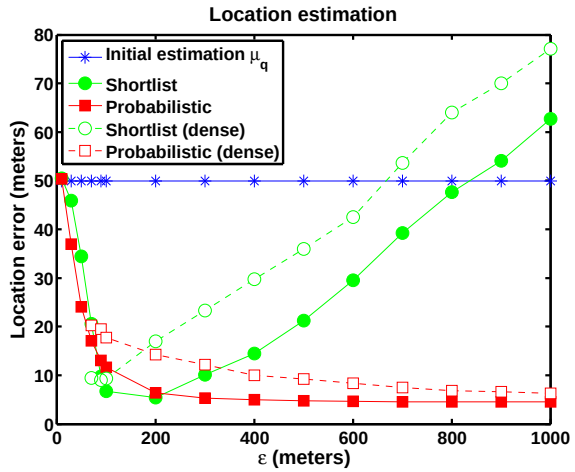

Fig. 7. Location refinement error.