Semantic features for food image recognition with geo-constraints

Xinhang Song, Shuqiang Jiang, Ruihan Xu, Luis Herranz
Key Laboratory of Intelligent Information Processing
of Chinese Academy of Sciences (CAS),
Institute of Computer Technology, CAS,
Beijing, 100190, China.
{xinhang.song,shuqiang.jiang,ruihan.xu,luis.herranz}@vipl.ict.ac.cn

Abstract*—Food-related photos have become very popular, due to social networks, food recommendation and dietary assessment systems. Reliable annotation is essential in those systems, but user-contributed tags are often non-informative and inconsistent, and unconstrained automatic food recognition still has relatively low accuracy. Most works focus on exploiting only the visual content while ignoring the context. In this paper, we improve food image recognition using two novel components. First, different from the conventional approach representing image in a visual feature space, we represent images in a semantic space, where we model context information. Secondly, we leverage the geographic context of the user and information about restaurants to simplify the classification problem. Thus, we propose a food recognition framework based on semantic features and location-adaptive classification. We collected a restaurant-oriented food dataset with food images, dish tags and restaurant-level information, such as the menu and geographic location. Experiments on this dataset show that exploiting geolocation improves around 30% the recognition performance, and the semantic feature has a gain of 3%-10% to the other visual features.

Index Terms—**semantic featuures; location-adaptive; food image recognition;**

I. Introduction

Recently, we are witnessing a fast increase of the visual data both on internet platforms and mobile devices. Instead of only relying on textual information, nowadays users tend to use images to record their daily activities and interact with other people. Images are easy to capture, convenient to store and share, and more expressive to capture a particular moment. However, automatically and accurately recognizing images still remains a big challenge and hinders many potential applications. Although there exist many tags and surrounding contextual descriptions for internet and mobile photos, they are often noisy, and inconsistent, leading to unreliable tags. Although, automatic image recognition has been studied for a long time[11], [23], [16], image recognition performance is still far from satisfactory for many real applications. The main challenge lies in how to effectively obtain effective visual representations from images and transform them to semantic descriptions.

In most cases, image recognition systems do not exploit useful knowledge about the particular tasks, . For instance, the contextual information directly related with the images. Images captured from mobile devices are not only composed of pixels, but they also have external contextual properties such as geographic information, temporal information, etc., which can provide valuable information for image recognition. How to design task-oriented image recognition techniques by flexibly integrating contextual information into the recognition algorithms is an important problem. In this paper, we focus on a specific yet important task: food image recognition. Food is not only an important part in many social events, but also the central topic in daily social interactions. Food images are widely available on social platforms and mobile devices. According to the Huffington Post[1], food has become the most popular category, with 57% of Pinterest[2] users interacting with food-related photos. A recent market research report suggests that 49% of the consumers learn about food through social networks[21]. Automatic and accurate food image recognition should be useful for many applications such as dish photo retrieval and organization, food recommendation and dietary assessment. Food image recognition is a challenging problem, as the actual appearance of food varies greatly for different dishes. Photos of the same dish may vary greatly, due to different locations and restaurants.. Even for images of the same dish taken in the same restaurant, the visual appearance may also vary due to the potential point of view, background, and lighting conditions.

How to design an effective method to represent dish images is essential in food recognition. For instance, Joutou and Yanai [8] proposed an automatic food image recognition system based on multiple kernel learning (MKL) integrating features, like color, texture and SIFT. Zong *et al.*Kawano *et al.*[9] use Fisher vectors over HOG patches. Yang *et al.* [26] propose using pairwise local features to capture important shape characteristics and spatial relationships between food ingredients. Thus, in general, the techniques used for image representation are based on visual features. The widely used frameworks are based on the bag-of-words (BOW) model. The first step is extracting local visual features, such as SIFT[14] or HOG[6], then encoding them into a global representation using the BoW model[20], or its sparse variations[25], [23]. However, one limitation of visual features is that they do not explore the semantic context[18]. Some works[17], [19],

[1] http://www.huffingtonpost.com
[2] http://www.pinterest.com

978-1-4799-4274-9/14 $31.00 © 2014 IEEE
DOI 10.1109/ICDMW.2014.144

IEEE computer society

[18], [12], [13], [22], [10] design features in a semantic space. Object bank[12] proposes a semantic representation that encodes the response of a number of pretrained object classifiers at different spatial locations. Classemes[22] are intermediate semantic representations based on a set of 2659 basis classes. In [18], Dirichlet mixture models (DMM) are used to model co-occurrence patterns. Similarly, the semantic manifold (SM)[10] is a discriminative alternative which uses a support vector machine (SVM) over SMNs combined with a suitable kernel for the multinomial simplex: the negative geodesic kernel (NGD)[27], implicitly addressing the co-occurrence modeling.

In most cases, dish images are taken when users are dining out in a restaurant. In this scenario, external geographical information directly associated with the image context can be obtained. Built-in GPS receivers in mobile cameras can be used for outdoor positioning[15], while network/WiFi-based location techniques can obtain geographical locations even in indoor environments [1], [5]. In general, we can assume that an approximate location of the dish photo can be guaranteed. As illustrated in Fig. 1, from the estimated geographical location, dish photos can be restricted in a specified area, so the candidate restaurants where the photos come from can be estimated. The goal of dish photo recognition is to estimate the most probable dish labels, which we assume are contained in the menus of the candidate restaurants. So the candidate dish label set for image recognition is also limited to a smaller group. The red arrows in Fig. 1 show the three aspects of the geographical context associated with the photos. The first one is the direct geographical location, the other two are indirect information related with candidate restaurants and dish labels. We need to effectively use this information to boost the performance of dish recognition. Geolocation has been widely studied to restrict the candidate images or categories for image retrieval and recognition, yet most of earlier works focus on landmark applications[24], [28], [4]. Landmarks can be consider rigid and intrinsically invariant. Photos of a particular landmark contain consistent visual patterns despite different lighting conditions, capturing positions, and background possibilities, so partial duplicate and BOW retrieval techniques based on local features can be directly applied for this scenario. Moreover, landmark images have obvious geographical properties they can be effectively integrated into the retrieval framework. Based on the retrieval results, KNN-based classification can be used for landmark recognition[13], [7]. However, dish recognition is a different problem. Dishes are non-rigid objects, but highly deformable. Visual appearances of dish photos are different no matter whether they belong to the same dish category or the same dish instances. On the other hand, in most cases, the available training data for the dish recognition is limited, especially for a dish in a particular geographical area. Based on the above analysis, investigations on dish image representation and classification techniques by fully considering the properties of dish images and effectively utilizing geographical information are desired.

In this paper, we propose a framework to represent dish images using semantic features for location-adaptive classification. The proposed semantic features are represented by the dish posterior probability distributions. We firstly learn these probability distribution using GMM. Then, using the semantic features, discriminative classifiers with a suitable feature embedding for the semantic space are trained for all the dish categories in the database. For a query image including its geographical location, candidate restaurants can be obtained by searching nearest restaurants, then only those available in the candidate restaurants are evaluated to make the prediction. Our technique is suitable for close-up dish photos as it occupies most of the dish photos. Based on the proposed semantic features for location-adaptive classification, the contribution of this paper is twofold:

- We use semantic features for dish recognition, improving the performance compared to of visual features.
- We propose a recognition framework which exploits geolocation in the semantic space, adapted to the problem of dish recognition in restaurants.

II. Semantic feature

A. Model and learning

The probability distribution of a vocabulary of dishes is estimated from a set of local visual features defined in some visual feature space X. Each image from the dataset is represented as a bag of local visual descriptors $I = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, $\mathbf{x}_n \in X$, densely sampled in a grid with N local patches. Note that visual features are localized in patches, but there are no corresponding patch labels. To address this problem, dish categories are also used as patch theme vocabulary, so we will refer to them as dish categories (or semantic concepts, in general) to both image categories and patch themes.

Images are modeled using a generative process, in which a dish category w is first sampled, and then N (patch) feature vectors are generated from $P_{\mathbf{X}|W}(\mathbf{x}_n|w)$. Given a new image, the category can be predicted using the Bayes rule

$$P_{W|\mathbf{X}}(w|I) = \frac{\prod_{n=1}^{N} P_{\mathbf{X}|W}(\mathbf{x}_n|w)\, P_W(w)}{\prod_{n=1}^{N} P_{\mathbf{X}}(\mathbf{x}_n)} \propto \frac{\prod_{n=1}^{N} P_{\mathbf{X}|W}(\mathbf{x}_n|w)}{\prod_{n=1}^{N} P_{\mathbf{X}}(\mathbf{x}_n)} \quad (1)$$

which assumes a uniform prior for $P_W(w)$.

As patch labels are not available, theme conditional distributions $P_{\mathbf{X}|W}(\mathbf{x}_n|w)$ are learned using weak supervision via image labels. In this setting, we have a single label available for each image, indicating that the theme (i.e. labeled dish category) is present in the bag of patches, but it does not imply that other themes are absent. Each theme conditional distribution $P_{X|W}(x|w)$ is modeled as a Gaussian mixtures model (GMM), learned over a set of training images[3].

B. Inference

Once we have learned theme GMMs, we can estimate the distribution of concepts for a new image I using the posterior

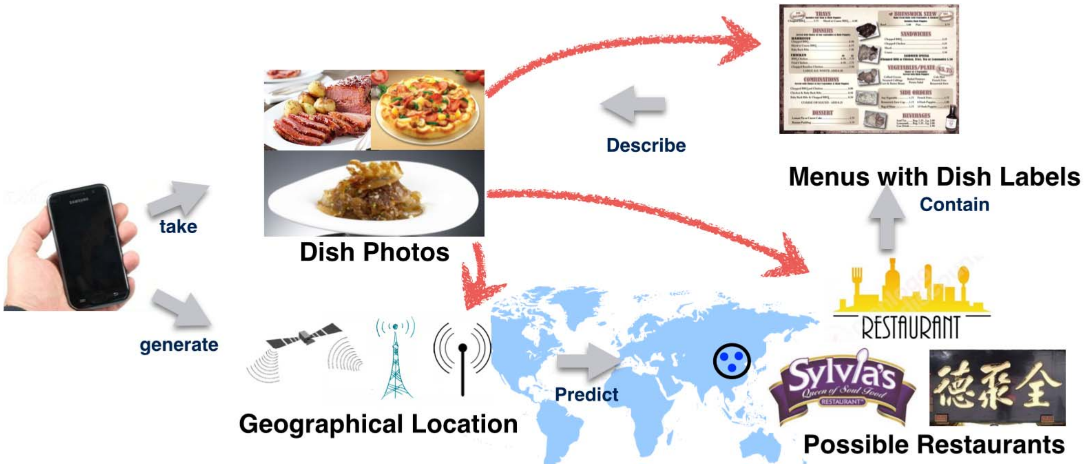

Fig. 1. The association of dish photos with geographical location and extended geo-information

probability $P_{W|\mathbf{X}}(w|I)$. For a vocabulary with M scene categories, the vector of posterior probabilities $\mathbf{s} = (s_1, ..., s_M)^T$ with $s_w = P_{W|\mathbf{X}}(w|I)$, is referred to as the *semantic multinomial* (SMN) of the image I[17]. The SMN is a probability vector of concepts that lies on the simplex Δ^{M-1} (referred to also as semantic space or semantic simplex) and provides a compact yet rich semantic description of the image. The whole process can be seen as a mapping $f : X^N \mapsto \Delta^{M-1}$ from the set of local visual descriptors in the image to the semantic space. Finally, given a new image, a label can be predicted from a SMN by simply selecting the concept with maximum probability (we will refer to this decision method as Bayes classification).

Note that this mapping is defined for images, but can also be used over patches to extract local SMNs as $f_{patch} : X \mapsto \Delta^{M-1}$. This alternative view allows us to infer the image SMN from patch SMNs. Thus, we can define a patch-to-image operation. In particular, for (1) the corresponding operation is just a product of the semantic multinomials

$$s_w = \Omega_w^{\text{prod}}(I) = \prod_{n=1}^{N} s_{nw} \tag{2}$$

where

$$s_{nw} = P_{W|\mathbf{X}}(w|\mathbf{x}_n) = \frac{P_{\mathbf{X}|W}(\mathbf{x}_n|w)\, P_W(w)}{P_{\mathbf{X}}(\mathbf{x}_n)} \tag{3}$$

is the w-th element of the patch SMN $\mathbf{s}_n$.

C. Multi-feature combination

Instead of a single type of visual feature, we now consider a set of complementary ones V (in our experiments $V = \{\text{gradient, shape, color}\}$). Each feature $v \in V$ generates a set of local visual descriptors $I^v = \{\mathbf{x}_1^v, \ldots, \mathbf{x}_N^v\}$, $\mathbf{x}_n^v \in \mathbf{X}^v$, and $I = \{I^1, \ldots, I^{|v|}\}$ represents all the features in the image. Now we assume that we learn feature-specific theme models $P_{\mathbf{X}^v|W}(\mathbf{x}_n^v|w^v)$, learned independently in the same way as in the single feature case. Similarly to (3), we can define the feature-specific patch SMN as

$$s_{nw}^v = P_{W^v|\mathbf{X}^v}(w^v|\mathbf{x}_n^v) = \frac{P_{\mathbf{X}^v|W^v}(\mathbf{x}_n^v|w^v)\, P_{W^v}(w^v)}{P_{\mathbf{X}^v}(\mathbf{x}_n^v)} \tag{4}$$

Using the Bayes rule we can extend (1) as

$$s_w = \frac{\prod_{v\in V} P_{W^v|W}(w^v|w)\, P_W(w) \prod_{n=1}^{N} P_{\mathbf{X}^v|W^v}(\mathbf{x}_n^v|w^v)}{\prod_{n=1}^{N} P_{\mathbf{X}^v}(\mathbf{x}_n^v)} \tag{5}$$

where $P_{W^v|W}(w^v|w)$ is uniform between each type of feature. Using (4) we can further rearrange (5) into

$$s_w = \prod_{n=1}^{N} \prod_{v\in V} P_{W|W^v}(w|w^v)\, s_{nw}^v \tag{6}$$

III. Semantic feature embedding

In this section, we discuss about the distance measurement and distance embedding for the feature in the semantic space. As the SMN $\mathbf{s} = (\mathbf{s_1}, ..., \mathbf{s_M})^\mathbf{T}$ is defined in the semantic simplex Δ^{M-1}, to exploit the geometry of the simplex, a suitable distance is the geodesic distance $d_{GD}(\mathbf{s}, \mathbf{s}') = 2\arccos\left(\left\langle\sqrt{\mathbf{s}}, \sqrt{\mathbf{s}'}\right\rangle\right)$where $\sqrt{\mathbf{s}}$ denotes element wise square root. A negative geodesic distance (NGD) kernel can be defined from this distance as $k_{NGD}(\mathbf{s}, \mathbf{s}') = -d_{GD}(\mathbf{s}, \mathbf{s}')$[27], which is used in combination with SVM. However, using NGD kernel requires kernel SVM, which has more the computational complexity than using linear SVM, especially on the large-scale dataset. To adapt NGD kernel to the linear SVM, one way is embedding this kernel into the feature. In [10], the authors propose an approximate embedding for the NGD kernel, and use it with the linear SVM. This approximate embedding can decrease the computational complexity, but also lose the accurracy for the recognition.

Now we propose an exact feature embedding method using a dictionary. We assume a mapping $\phi(\mathrm{s})$ to a different space (i.e. kernel space), with the kernel dictionary U, it can be mapped as:

$$\bar{\alpha} = \arg\min_{\alpha} \left\| \phi(s) - U\alpha \right\|^2 + \lambda \left\| \alpha \right\|_1 \tag{7}$$

where α is the codes, $U = [\phi(z_1), \ldots, \phi(\mathbf{z}_C)]$ is the kernel dictionary, formed by the dictionary in $Z = [z_1, z_2 ... z_C]$ mapped to the kernel space, C is the size of the dictionary. Instead of learning the dictionary in the kernel space, we simply use K-means in the semantic space to learn Z (in fact, as we will see, we do not need to explicitly compute U). In order to use the geodesic distance between two filtered SMNs s and s$'$, we formulate the whole filtering and embedding as a new kernel, obtained from the projected vectors on the NGD kernel space as

$$\begin{aligned} k_{KCNF}(\mathrm{s}, \mathrm{s}') = k_{NGD}(\bar{\mathrm{s}}, \bar{\mathrm{s}'}) = \phi_{NGD}(\bar{\mathrm{s}})^{\intercal} \phi_{NGD}(\bar{\mathrm{s}'}) \\ = (U\bar{\alpha})^{\intercal} (U\bar{\alpha'}) \end{aligned} \tag{8}$$

where $\bar{\alpha}$ and $\bar{\alpha'}$ are the codes for $\bar{\mathrm{s}}$ and $\bar{\mathrm{s}}'$. The corresponding analytic solution is $\bar{\alpha} = (U^{\intercal}U)^{-1} U^{\intercal} \phi_{NGD}(\bar{\mathrm{s}})$ (note that we do not have this solution for $\lambda \neq 0$). By applying some appropriate matrix transformations we can rearrange (8) as

$$\begin{aligned} k_{KCNF}(\mathrm{s}, \mathrm{s}') = (U^{\intercal} \phi_{NGD}(\mathrm{s}))^{\intercal} (U^{\intercal}U)^{-1} (U^{\intercal} \phi_{NGD}(\mathrm{s}')) \\ = K_{zs}(s)^{\intercal} K_{zz}^{-1} K_{zs}(s') \end{aligned} \tag{9}$$

where $K_{zs}(\mathrm{s}) = U^{\intercal} \phi_{NGD}(\mathrm{s})$ and $K_{zz} = U^{\intercal}U$. Note that no explicit mapping ϕ_{NGD} is necessary to compute any of those matrices. As K_{zz} is positive definite, we can find a decomposition $G^{\intercal}G = K_{zz}^{-1}$ (e.g. using the Cholesky decomposition), which leads to

$$\begin{aligned} k_{KCNF}(\mathrm{s}, \mathrm{s}') &= (GK_{zs}(\mathrm{s}))^{\intercal} (GK_{zs}(\mathrm{s}')) \\ &= \phi_{KCNF}(\mathrm{s})^{\intercal} \phi_{KCNF}(\mathrm{s}') \end{aligned}$$

where we have an explicit mapping

$$\phi_{KCNF}(\mathrm{s}) = GK_{zs}(\mathrm{s}) \tag{10}$$

that only depends on a set of words z_i and the original (unfiltered) SMN s and s$'$. With (10) we can obtain embeddings of semantic features that can be used for large scale classification.

IV. Geo-constrained classification

A. Geographically nearest classes

To use the geographical information associated to dish photos, we first need to obtain the geographical neighborhood relationship of dish images taken in restaurants. Since the main properties of a restaurant are its geographic location and its menu (i.e. the dish categories found in that particular restaurant), a restaurant can be represented as a pair $R = (\Psi, menu)$ where $\Psi = (\lambda, \phi)$ is the geographic location, λ and ϕ denote latitude and longitude. The restaurant database contains J restaurants with a combined total of D dishes. For a given restaurant $j \in \{1, \ldots, J\}$, the menu can be represented as $menu_k = \{r_1, \ldots, r_{D_j}\}$, where $r_i \in \{1, \ldots D\}$ is the i-th dish in the restaurant menu $menu_j$, with D_j different dishes.

For a query image, we assume that the mobile phone will obtain its current location $\Psi_q = (\lambda_q, \phi_q)$, which can be typically estimated by triangulation or multilateration using multiple sources, such as GPS satellites, cell phone towers or wireless networks. In our work we also assume that the error in the estimation is relatively isotropic and thus we select as candidates the restaurants found within a circular area of radius ϵ. The distance between the current location Ψ_q and the location Ψ_j of the restaurant j can be approximated by the spherical law of cosines

$$d(\Psi_q, \Psi_j) = rad \times (\sin\phi_q \sin\phi_j + \cos\phi_q \cos\phi_j \cos(\lambda_q - \lambda_j)) \tag{11}$$

where rad is the radius of the Earth. Finally, the set of its nearest restaurants is obtained as

$$H_q = H(\Psi_q, \epsilon) = \{ j \mid d(\Psi_q, \Psi_j) \leq \epsilon, \forall j = 1, \ldots, J \} \tag{12}$$

B. Classification model with geo-constraints

With the geographically nearest restaurants obtained, we can select these restaurants as candidate restaurants, and the unlikely dish categories which are not within the union of the menus of these restaurants can be discarded. We first train classifiers for all the dish categories in the database and then evaluate only those available in the candidate restaurants. We use linear SVMs due to the high dimensionality of the aggregated semantic feature. We denote a (binary) classifier $f : T \mapsto \mathbb{R}$ that maps a feature vector $t \in T$ to a certain score. A typical (one-against-all) SVM classifier combines multiple binary classifiers (one per category) and predicts the category with the maximum score. Note that the negative samples are selected from all the restaurants.

We assume that we have trained a pool of binary dish classifiers $\{f_1, f_2, \ldots, f_D\}$, where f_i denotes the global classifier for the dish category i. Once we have a set of candidate restaurants H_q, the set of possible dish categories can be obtained from their menus as $M(H_q) = \bigcup_{j \in H_q} M_j$. Based on the geographical context related to the photo, we effectively limit the potential dish categories. Thus we refer to this as a geographical constraint or *geo-constraint*. Then, for a feature vector x and a candidate restaurant set H, the classifier predicts a category based on the candidate classifier with maximum score

$$y^* = \mathop{\arg\max}_{i \in menu(H_q)} f_i(t) \tag{13}$$

V. Experiments

A. Dish dataset

We collected data about restaurants in Beijing from an online restaurant review site[3]. Each restaurant in our dataset includes its location (coordinates of latitude and longitude), the list of dishes (i.e. menu) and photos of each dish. We discarded restaurants with less than 3 dishes in the menu and fewer than

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

TABLE I
OVERALL STATISTICS OF THE DATASET.

	Min	Average	Max	Total
Restaurants	-	-	-	187
Dishes/restaurant	3	6.27	25	1173
Images/dish	15	38.82	438	45541

TABLE II
COMPARISON OF DIFFERENT FEATURES FOR DISH RECOGNITION.

Method	No Location	With Location
BoW[11]	7.3	41.7
Kdes[2]	16.8	51.1
LLC[23]	22.4	56.3
Fisher Vector[16]	26.4	59.0
Proposed	**29.2**	**61.2**

15 images per dish. Table.I shows the overall statistics of the datasets. The dataset contains 187 restaurants with a combined 1173 dish categories (701 unique dish categories).

B. Comparison results

We learn semantic features using three types of the kernel descriptors[2], and combine them in the semantic follow the sectionII-C. We train 512 Gaussian mixture models for each dish category to learn the SMN. For the feature embedding, we train a 1024 size dictionary using the learned SMNs. We use the same size dictionary for the other visual features.

We compare the proposed semantic feature with other visual features. We consider four types of visual features: BoW, kernel descriptors (Kdes), locality-constrained linear coding (LLC) and Fisher Vector. The evaluation includes global models with and without considering geo-constraints. The results are shown in Table II. Comparing with the visual feature with best performance (Fisher Vector), the proposed method has a gain of 3%. And it also can be seen that the location-adaptive model improves the performance significantly, with a gain about 30%. Note that, we learned the proposed semantic features using the same Kdes visual descriptors. Comparing with Kdes, the proposed method has a gain of 12.4% without location-adaptive model, and a gain of 10.1% with location-adaptive model.

VI. CONCLUSION

This paper investigates food image recognition using two novel components:: food images are represented using semantic features and the classification is adapted to the geographic location via geo-constraints. We collected a restaurant-oriented food dataset with food images, dish tags and restaurant-level information, such as the menu and geographic location. We evaluated the proposed method comparing with conventional approaches based on visual features that ignore the geographical context. The comparison results show that semantic features have better performance with a gain around 3% over the best visual feature. Moreover, the location-adaptive model can dramatically improve the performance with a gain about 30%.

ACKNOWLEDGMENT

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 61350110237, in part by National Hi-Tech Development Program (863 Program) of China: 2014AA015202, and in part by the Chinese Academy of Sciences Fellowships for Young International Scientists: 2011Y1GB05. This work is also funded by Lenovo Outstanding Young Scientists Program (LOYS).

REFERENCES

[1] A. Bensky. *Wireless positioning technologies and applications*. Artech House, 2007.

[2] L. Bo, X. Ren, and D. Fox. Kernel descriptors for visual recognition. In *Proc. NIPS*, volume 1, page 3, 2010.

[3] G. Carneiro, A. Chan, P. Moreno, and N. Vasconcelos. Supervised learning of semantic classes for image annotation and retrieval. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 29(3):394–410, 2007.

[4] D.M. Chen, G. Baatz, K. Koser, S.S. Tsai, R. Vedantham, T. Pylvanainen, K. Roimela, Xin Chen, J. Bach, M. Pollefeys, B. Girod, and R. Grzeszczuk. City-scale landmark identification on mobile devices. In *Proc. IEEE CVPR*, pages 737–744, June 2011.

[5] Y. Cheng, Y. Chawathe, A. LaMarca, and J. Krumm. Accuracy characterization for metropolitan-scale wi-fi localization. In *Proc. ACM MobiSys*, pages 233–245, 2005.

[6] N. Dalal and B. Triggs. Histogram of oriented gradient object detection. In *CVPR*, 2005.

[7] B. Girod, V. Chandrasekhar, R. Grzeszczuk, and Y.A. Reznik. Mobile visual search: Architectures, technologies, and the emerging mpeg standard. *IEEE MultiMedia*, 18(3):86–94, March 2011.

[8] T. Joutou and K. Yanai. A food image recognition system with multiple kernel learning. In *Proc. IEEE ICIP*, pages 285–288, 2009.

[9] Y. Kawano and K. Yanai. Foodcam: A real-time mobile food recognition system employing fisher vector. In *Proc. MMM*, pages 369–373, 2014.

[10] Roland Kwitt, Nuno Vasconcelos, and Nikhil Rasiwasia. Scene recognition on the semantic manifold. In *ECCV*, 2012.

[11] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.

[12] L.J. Li, H.Su, E.P. Xing, and L. Fei-Fei. Object bank: A high-level image representation for scene classification and semantic feature sparsification. In *NIPS*, 2010.

[13] Zhen Li and Kim-Hui Yap. Content and context boosting for mobile landmark recognition. *IEEE Signal Processing Letters*, 19(8):459–462, Aug 2012.

[14] DavidG. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60:91–110, 2004.

[15] Hui-Lan Luo, Hui Wei, and Loi Lei Lai. Creating efficient visual codebook ensembles for object categorization. *IEEE Trans. on Systems, man, and Cybernetics*, 41(2), 2011.

[16] F. Perronnin, J. Sanchez, and T. Mensink. Improving the fisher kernel for large-scale image classification. In *Proc. ECCV*. 2010.

[17] N. Rasiwasia and N. Vasconcelos. Bridging the gap: Query by semantic example. *IEEE Trans. on Multimedia*, 9(5):923–938, 2007.

[18] N. Rasiwasia and N. Vasconcelos. Holistic context models for visual recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 34(5):902–917, 2012.

[19] Nikhil Rasiwasia and Nuno Vasconcelos. Holistic context modeling using semantic co-occurrences. In *CVPR*, pages 1889–1895, 2009.

[20] J. Sivic and A. Zisserman. Video google: a text retrieval approach to object matching in videos. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 1470–1477, 2003.

[21] The Hartman Group. Clicks & cravings: The impact of social technology on food culture, 2012.

[22] Lorenzo Torresani, Martin Szummer, and Andrew Fitzgibbon. Efficient object category recognition using classemes. In *ECCV*, 2010.

[23] Jinjun Wang, Jianchao Yang, Kai Yu, Fengjun Lv, Thomas Huang, and Yihong Gong. Locality-constrained linear coding for image classification. In *Proc.IEEE CVPR*, 2010.

[24] K. Yaegashi and K. Yanai. Geotagged image recognition by combining three different kinds of geolocation features. In *Proc. IEEE ACCV*, 2010.

[25] Jianchao Yang, Kai Yu, Yihong Gong, and Thomas S. Huang. Linear spatial pyramid matching using sparse coding for image classification. In *CVPR*, 2009.

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

[27] Dell Zhang, Xi Chen, and Wee Sun Lee. Text classification with kernels on the multinomial manifold. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '05, pages 266–273, New York, NY, USA, 2005. ACM.

[28] Y. Zheng, M. Zhao, Y. Song, H. Adam, U. Buddemeier, A. Bissacco, F. Brucher, T. Chua, and H. Neven. Tour the world: building a web-scale landmark recognition engine. In *Proc. IEEE CVPR*, pages 1085–1092, 2009.