GEOGRAPHIC INFORMATION USE IN WEAKLY-SUPERVISED DEEP LEARNING FOR LANDMARK RECOGNITION

Yifang Yin[†], *Zhenguang Liu*[‡], *Roger Zimmermann*[‡]

[†]Interactive and Digital Media Institute, National University of Singapore
[‡]School of Computing, National University of Singapore
[†]idmyiny@nus.edu.sg, [‡]{liuzheng,rogerz}@comp.nus.edu.sg

ABSTRACT

The successful deep convolutional neural networks for visual object recognition typically rely on a massive number of training images that are well annotated by class labels or object bounding boxes with great human efforts. Here we explore the use of the geographic metadata, which are automatically retrieved from sensors such as GPS and compass, in weakly-supervised learning techniques for landmark recognition. The visibility of a landmark in a frame can be calculated based on the camera's field-of-view and the landmark's geometric information such as location and height. Subsequently, a training dataset is generated as the union of the frames with presence of at least one target landmark. To reduce the impact of the intrinsic noise in the geo-metadata, we present a frame selection method that removes the mistakenly labeled frames with a two-step approach consisting of (1) Gaussian Mixture Model clustering based on camera location followed by (2) outlier removal based on visual consistency. We compare the classification results obtained from the ground truth labels and the noisy labels derived from the raw geo-metadata. Experiments show that training based on the raw geo-metadata achieves a Mean Average Precision (MAP) of 0.797. Moreover, by applying our proposed representative frame selection method, the MAP can be further improved by 6.4%, which indicates the promising use of the geo-metadata in weakly-supervised learning techniques.

1. INTRODUCTION

Recent research on supervised learning with convolutional neural networks has received significant success in a variety of visual tasks including image classification [1, 2], object detection and recognition [3, 4]. However, most of these methods require large numbers of training images with detailed annotations that are labeled manually. For example in complex scenes, it is highly beneficial for supervised object detection to train with bounding boxes of object locations [3]. As the process of labeling a large set of training images with class labels, object locations and attributes can be tedious and expensive, researchers have started utilizing other important contextual information that can be automatically retrieved with the help of various sensors. For example in landmark recognition, geographic metadata including the camera location and viewing direction is usually utilized as geographic constraints for efficient processing [5, 6, 7]. With the help of the ubiquitous sensor-equipped smartphones, nowadays users can easily take images and videos together with the aforementioned geo-metadata through the use of popular apps [8]. Not only the number of geotagged multimedia documents has been growing online rapidly [9], the spatial data about the geographic objects (*e.g.*, buildings) have also become increasingly available from online mapping services [10]. Both of these aspects make the use of geographic information in computer vision more encouraging.

Inspired by the above observations, we investigate the problem of transforming the geographic metadata into the training labels for landmark recognition. We build on the efficient geo-based landmark retrieval method [7] that retrieves frames of a landmark based on the camera's viewable scene model and the landmark's geometric information, but further improve the quality of the training labels by removing outliers while maintaining a good balance between the visual consistency and the visual diversity of the training data. In terms of visual consistency, we perform a Gaussian Mixture Model (GMM) based camera location clustering where the visual distance between the intra-cluster members is expected to be low. This is because the probability of a landmark having similar visual appearances is higher among frames taken at similar locations. Thereafter, we compute the mean feature vector for every cluster. Only the top $ratio \in (0, 1)$ frames that are the closest to the cluster centre will be selected as representatives, while the rest will be discarded due to the visual inconsistency. On the other hand, the visual distance between the inter-cluster members is likely to be high due to the different shooting angles, thus leading to high visual diversity of the training data. The experimental results demonstrate the effectiveness of our approach: training on the labels derived from the raw geo-metadata obtains a promising MAP of 0.797, which is further improved by 6.4% after applying our representative frame selection method.

978-1-5090-6067-2/17/$31.00 ©2017 IEEE

2. RELATED WORK

Content-based landmark recognition has been an important yet challenging task in computer vision. Unsupervised methods [11] do not require any ground truth labels, but good performance can only be achieved when classifying visually consistent objects. Fully supervised methods [12, 3] are generally more effective in object recognition by training classifiers with class labels or annotations of object locations in terms of bounding boxes [3] or even locations of object parts [13]. However, obtaining a massive amount of such well-labeled annotations is usually costly, tedious, or even biased. To solve the above issues, weakly supervised convolutional neural networks (CNNs) have been proposed recently, the goal of which is to locate objects based on the noisy image-level labels only [4, 14]. Oquab *et al.* [4] presented a novel CNN for object classification, which relies on image-level labels, yet can learn from cluttered scenes containing multiple objects. Xiao *et al.* [15] proposed a general framework to train CNNs with only a limited number of clean labels and millions of easily obtained noisy labels. Bekker and Goldberger [16] focused on the problem of training a neural network based on data with unreliable labels, and introduced an extra noise layer in a deep neural network to model the relationship between the true labels and the noisy observed labels.

As a supplement to visual features of image content, researchers have emphasized more on the use of contextual information in recent years. Several methods have been proposed for landmark recognition by utilizing geographic metadata alternatively [17, 18, 6, 19]. For example, Zheng *et al.* [5] presented a web-scale landmark recognition engine that organizes and recognizes landmarks on the scale of the entire planet Earth. Yin *et al.* [7] presented a pure geocontext-based method called the Geo Landmark Visibility Determination (GeoLVD), which computes the visibility of a landmark based on intersections of a camera's field-of-view and the landmark's geometric information obtained from online geographic information services. Such geocontext-based methods have the advantage of being highly efficient compared with content-based techniques, but the effectiveness can be susceptible to the sensor noise such as the errors and uncertainties intrinsically possessed by the geographic metadata collected from GPS, compass and other sensors.

3. TRAINING LABEL DERIVATION FROM FREE GEOGRAPHIC METADATA

Nowadays, sensor-equipped smartphones have made it possible for users to record geo-referenced videos all around the world. The location and viewing direction of the camera can be obtained from the built-in GPS and compass sensors, which are synchronized and associated with the video frames in a fine granularity (*e.g.*, with the sampling interval set to one second). Next, we briefly describe a geo-based landmark visibility determination algorithm in Section 3.1, and present a frame selection method that removes frames with inaccurate labels from the training set in Section 3.2.

3.1. Landmark Visibility Determination

To generate a training dataset for landmark recognition, we need to retrieve all the frames with presence of at least one target landmark. Here we adopt the GeoLVD method proposed by Yin *et al.* [7] that computes the visibility of a landmark in a frame based on efficient geometry calculations and occlusion checks. As input, the camera's field-of-view (FOV) is typically modeled by five parameters: camera location and orientation extracted from the geo-metadata, horizontal and vertical viewable angles, and the far visible distance estimated from camera optics [8]. Meanwhile, the geographic information (*e.g.*, footprint and height) of buildings can be easily retrieved from online mapping services such as the OpenStreetMap[1]. The GeoLVD method initially returns a set of visible angles of the landmark within the FOV. Let $y_k(x) \in \{-1, 1\}$ represent the absence or presence of landmark k in frame x, then we define,

$$y_k(x) = \begin{cases} -1 & VisibleR(k,x) = \varnothing \\ 1 & VisibleR(k,x) \neq \varnothing \end{cases} \tag{1}$$

where $VisibleR(k,x)$ denote the set of visible angles of landmark k within the FOV of frame x, returned by GeoLVD. The effectiveness of GeoLVD depends on the accuracy of the geo-metadata. Therefore, the training labels derived using Eq. 1 can be noisy and inaccurate [20]. To solve this problem, we propose a representative frame selection method to reduce the impact of the intrinsic sensor noise. The details are introduced in the next section.

3.2. Representative Frame Selection

Generally the labels derived using Eq. 1 can be affected by two major factors: (1) inaccurate geo-metadata and (2) occlusions caused by moving obstacles, both of which lead to visually inconsistent frames compared with the majority. Based on the observations that photos taken at similar geographic locations are more likely to show visually similar content, we propose to first cluster the frames based on the camera locations, and then remove outliers from each group based on the visual features. Let $g = (lat, lng)$ denote a camera location, the distribution of frames showing a landmark can be well described by a Gaussian Mixture Model (GMM) containing n bivariate normal distributions $\mathcal{N}(\mu, \Sigma)$ [21]:

$$P(g|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{i=1}^{n} \pi_i \mathcal{N}(g|\mu_i, \Sigma_i) \tag{2}$$

where $\boldsymbol{\pi}$, $\boldsymbol{\mu}$, and $\boldsymbol{\Sigma}$ are the unknown model parameters that represent the weights, mean values, and covariance matrices

[1] http://www.openstreetmap.org.

associated with the n normal distributions, respectively. Here we derive the log-likelihood function as,

$$\ln P(G|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{j=1}^{m} \ln P(g_j|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{3}$$

and adopt the expectation maximization (EM) algorithm to iteratively estimate the model parameters based on the camera locations, $G = \{g_1, g_2, ..., g_m\}$, we retrieved. The EM algorithm alternatively performs an E step that computes the posterior probability of location g belonging to the k-th normal distribution based on the current model parameters:

$$P(k|g) = \frac{\pi_k \mathcal{N}(g|\mu_k, \Sigma_k)}{\sum_{i=1}^{n} \pi_i \mathcal{N}(g|\mu_i, \Sigma_i)} \tag{4}$$

and an M step that updates the model parameters by maximizing the log-likelihood given in Eq. 3 as:

$$\begin{aligned} \pi_k &= \frac{\sum_{j=1}^{m} P(k|g_j)}{\sum_{k=1}^{n} \sum_{j=1}^{m} P(k|g_j)} \\ \mu_k &= \frac{1}{\sum_{j=1}^{m} P(k|g_j)} \sum_{j=1}^{m} P(k|g_j) g_j \\ \Sigma_k &= \frac{1}{\sum_{j=1}^{m} P(k|g_j)} \sum_{j=1}^{m} P(k|g_j)(g_j - \mu_k)(g_j - \mu_k)^{\mathsf{T}} \end{aligned} \tag{5}$$

We repeat the the steps until the algorithm converges, and a frame associated with location g is clustered to the group with the largest posterior probability $P(k|g)$.

Thereafter, we extract visual features (*e.g.*, the HOG descriptor [22] we used in this work) from all the frames, and compute the mean feature vector for each cluster. We define a parameter $ratio \in (0, 1)$ to control the number of selected frames. Let m_k denote the number of frames in cluster k. Then only the top $ratio \cdot m_k$ frames that are the closest to the mean feature vector in cluster k will be selected as representatives, while the rest of the frames will be discarded as outliers. We use the Euclidean distance between feature vectors as the measure of visual inconsistency. Later in the experiments, we will see that this two-step frame selection method is effective in reducing the noisiness of the training data and the classifier trained using the selected representatives is able to obtain significant MAP improvement. Next, we introduce the network architecture we used for weakly-supervised learning.

4. NETWORK ARCHITECTURE FOR WEAKLY-SUPERVISED LEARNING

A classifier can be trained based on the landmark annotations derived in Section 3. Here we adopt the state-of-the-art network architecture proposed by Oquab *et al.* [4], which is a weakly-supervised convolutional neural network for object classification that only requires image-level labels. The network architecture consists of seven convolutional feature extraction layers, two convolutional adaptation layers, and a single global max-pooling layer at the output that searches the highest scoring object position in the image. The authors modified the fully connected layers, which are commonly used in previous CNN architectures [1, 23, 24], into convolutional layers in order to deal with input images with different sizes. In a K class classification problem, let $f_k(x)$ denote the output of the network for input image x and class k, and $y_k(x) \in \{-1, 1\}$ indicate the absence or presence of class k in image x, then the loss function for the multi-label classification is given as,

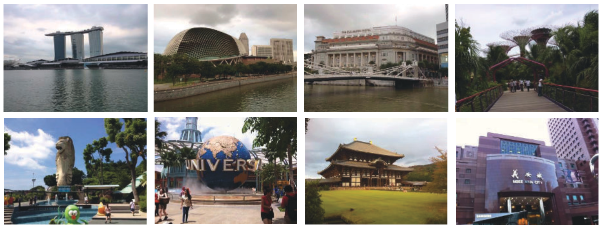


Fig. 1: Illustration of the frames showing each of the landmarks, sampled from the sensor-rich videos.

$$\ell\left(f_k(x), y_k(x)\right) = \sum_{k} \log\left(1 + e^{-y_k(x) f_k(x)}\right) \tag{6}$$

Subsequently, the posterior probability of class k being presented in image x can be estimated as below [4],

$$P(k|x) \approx \frac{1}{1 + e^{-f_k(x)}} \tag{7}$$

Usually a threshold of 0.5 is adopted, and the test images with a posterior probability equal to or larger than the chosen threshold are considered to be the positive instances of the target landmark.

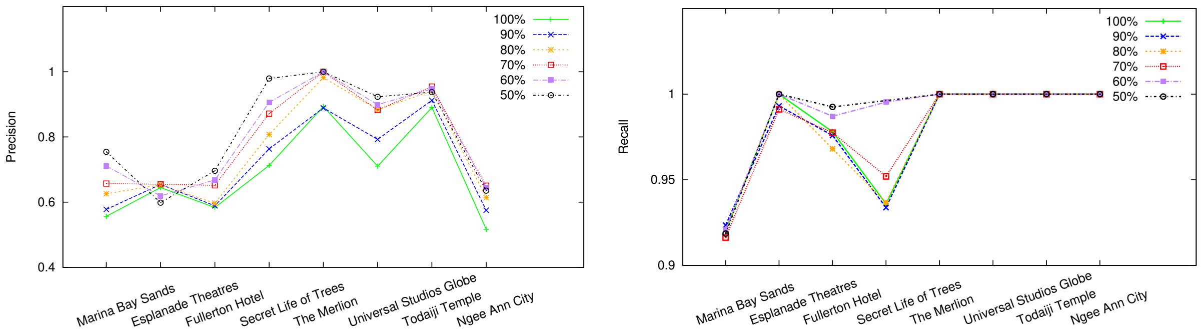
5. EVALUATION

We have trained convolutional neural networks introduced in Section 4 based on different training datasets. The evaluation consists of three steps. The first part introduces the experimental setup and the details of the training and the testing datasets. The second part evaluates our method to select representative frames. The third part compares the results of landmark classification and verifies the effectiveness of our proposed approach.

5.1. Experimental Setup

To prepare our training dataset, we retrieved 280 sensor-rich videos with the corresponding geo-metadata from the GeoVid[2] project. Users can record and upload videos using the GeoVid smartphone applications, watch videos via a

[2] http://geovid.org/



(a) Precision

(b) Recall

Fig. 2: Precision and recall comparison with different $ratio$ settings in the representative frame selection.

web browser, and download videos with the publicly available GeoVid APIs[3]. The training dataset is formed by 9,123 frames sampled every second from the sensor-rich videos. Additionally, we manually annotated the ground truth visibility of eight landmarks, namely the Marina Bay Sands hotel, the Esplanade theatres, the Fullerton Hotel, the Secret Life of Trees, the Merlion, the Universal Studios Globe, the Todaiji temple, and the Ngee Ann City. Fig. 1 illustrates examples of the representative frames in our training set for each of the eight landmarks. Furthermore, we collected 200 images of each landmark online to form a test set of 1600 images.

Next, we trained classifiers using the network architecture described in Section 4. We started with a pre-trained network from the Pascal VOC 2012 training and validation sets provided by Oquad *et al.* [4], and further trained the network using video frames with the ground truth annotations labeled manually and the noisy annotations derived from the geo-metadata. The classification results are compared and discussed in the next section.

5.2. Frame Selection

In order to evaluate the effectiveness of our representative frame selection method, we set the parameter $ratio$ in Section 3.2 to different values, and compute precision, recall and F1 measure for each of the parameter settings.

Fig. 2 illustrates the precision and recall of each landmark with $ratio \in \{100\%, 90\%, 80\%, 70\%, 60\%, 50\%\}$, where $ratio = 100\%$ means that the frames are annotated based on the raw geo-metadata as introduced in Section 3.1, and the rest indicate that we further select a subset of representative frames by keeping the top $ratio$ of the frames we formerly retrieved as introduced in Section 3.2. As can be seen, precisions are comparatively less satisfactory when $ratio = 100\%$. As the GeoLVD method computes a landmark's visibility based on the raw geo-metadata only without checking the image content, it is almost impossible for it to detect occlusions caused by obstacles such as people, vehicles or missing buildings in digital maps. In most of the cases, precisions have been greatly improved after removing visually inconsistent frames. But it is still difficult to make the precision close to one for some landmarks due to the semantic gap between low-level visual features and high-level concepts. The appearance of a landmark can be diverse even when taken at similar locations due to the change of illuminations, viewpoints, *etc.*

On the other hand, all landmarks have high recalls larger than 0.9, five of them even have recalls close to one when $ratio = 100\%$. This is because we only kept frames with the presence of at least one landmark. Positive instances of a landmark that are mistakenly labeled as negative were discarded, and thus did not have any impact on the recall. So unless two landmarks are located geographically close to each other (*e.g.*, Marina Bay Sands hotel and Secret Life of Trees) where frames of one landmark can be mistakenly labeled as the other one due to serious sensor errors, recalls mostly have a value close to one.

Table 1 reports the F1 measure and highlights the **best** score of each landmark with different $ratio$ values. The F1 measure is computed as $F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$, which can be considered as a good indicator for the quality of the frame set as it considers both precision and recall. Generally the F1 measure increases when setting $ratio$ to a smaller value, but the improvement of F1 tends to become stable after $ratio \leq 70\%$. We can say that parameter $ratio$ controls the tradeoff between the accuracy and the diversity of the selected frames, both of which are important indicators of a high quality training set. Based on the results shown in Table 1, $ratio = 70\%$ can be considered as a good threshold for the training dataset generation.

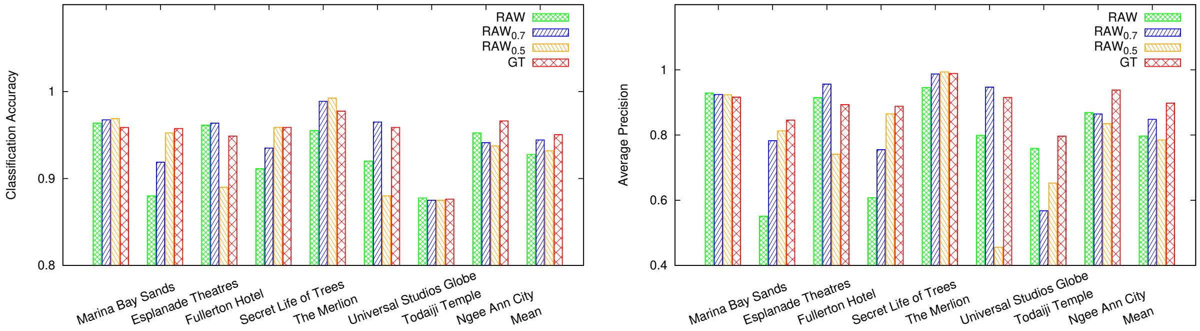
5.3. Landmark Classification

We trained four classifiers based on different training datasets and compared their effectiveness. Basically **GT** is equivalent

[3] http://api.geovid.org

Table 1: F1 measure comparison with different *ratio* settings in the representative frame selection.

ratio	100%	90%	80%	70%	60%	50%
Marina Bay Sands	0.693	0.711	0.745	0.765	0.802	**0.828**
Esplanade Theatres	0.784	0.789	**0.790**	0.789	0.765	0.749
Fullerton Hotel	0.731	0.735	0.738	0.782	0.797	**0.818**
Secret Life of Trees	0.809	0.840	0.867	0.910	0.949	**0.990**
The Merlion	0.944	0.941	0.991	**1.0**	1.0	1.0
Universal Studios Globe	0.831	0.88	0.938	0.938	0.947	**0.960**
Todaiji Temple	0.942	0.954	0.969	**0.977**	0.973	0.966
Ngee Ann City	0.682	0.730	0.761	**0.788**	0.784	0.777



(a) Classification Accuracy

(b) Average Precision

Fig. 3: Classification accuracy and average precision comparison per landmark using different training sets.

Table 2: Classification effectiveness comparison using different training sets.

Method	RAW	$\mathrm{RAW}_{0.7}$	$\mathrm{RAW}_{0.5}$	GT
Mean Accuracy	0.928	0.944	0.931	**0.950**
Mean Average Precision	0.797	0.848	0.785	**0.898**

to the advanced **WEAK SUP** method [4] by training with manual labels, and this provides a good comparison of our method to the state-of-the-art classification results.

- **GT:** a training set with ground truth labels.
- **RAW:** a training set annotated with labels derived from the raw geo-metadata.
- **$\mathbf{RAW}_{0.7}$:** a subset of **RAW** by keeping 70% of the frames as representatives.
- **$\mathbf{RAW}_{0.5}$:** a subset of **RAW** by keeping 50% of the frames as representatives.

The classification accuracy and the average precision for each landmark are illustrated in Fig. 3, and the comparison between the mean values over all the landmarks is reported in Table 2. Here we use average precision instead of precision, recall, and F1 measure as the former is considered as a better metric to evaluate a ranked list. We highlight the **best** and the second best results in Table 2. As can be seen, the classifier trained based on the ground truth labels achieved the best result and could be considered as an upper bound as reference. As we expected, $\mathrm{RAW}_{0.7}$ achieved the second best result and outperformed RAW by 1.72% and 6.4% in terms of mean accuracy and mean average precision, respectively, and therefore demonstrates the effectiveness of our frame selection approach. $\mathrm{RAW}_{0.5}$ performed less satisfactory than $\mathrm{RAW}_{0.7}$. One of the reasons might be that the gain in accuracy caused too much loss in diversity, and thus downgraded the effectiveness of the classifier. For example, there is an obvious decrease of $\mathrm{RAW}_{0.5}$ for the Fullerton Hotel and the Universal Studio Globe in Figure 3 most likely due to the diversity loss, while the rest of the statistics are generally consistent among different landmarks. It is also worth mentioning that even method RAW performed quite good, obtained a mean classification accuracy of 0.928 and a mean average precision of 0.797. It indicates the promising use of the geo-metadata, which are automatically recorded without any human efforts, in weakly-supervised deep learning techniques.

6. CONCLUSIONS AND FUTURE WORK

Given a collection of sensor-rich videos, we present an effective method that can automatically generate a large training dataset as the input of weakly-supervised learning for landmark recognition. The training labels (*i.e.*, the presence or absence of a landmark) of a frame can be efficiently determined based on the camera's field-of-view and the landmark's footprint and height. As the training labels can be noisy due to sensor errors and unpredictable occlusions, we further present a two-step approach that aims at removing frames with inaccurate labels considering both the distribution of camera locations and the consistency of visual content. We have compared the classification performance by using our automatically derived labels and the manually annotated ground truth labels. Promising results have been reported on the use of the geographic metadata in weakly-supervised learning.

In our future work we plan to evaluate the use of the geographic information in other components of supervised learning, *e.g.*, model label correlations based on both geo-information and visual features, to further improve the classification accuracy.

7. ACKNOWLEDGMENTS

This research was supported in part by the National Natural Science Foundation of China under Grant no. 61472266 and by the National University of Singapore (Suzhou) Research Institute, 377 Lin Quan Street, Suzhou Industrial Park, Jiang Su, People's Republic of China, 215123.

8. REFERENCES

[1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *NIPS*. 2012.

[2] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," in *CVPR*, 2014.

[3] Pedro F. Felzenszwalb, Ross B. Girshick, David McAllester, and Deva Ramanan, "Object Detection with Discriminatively Trained Part-Based Models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, 2010.

[4] Maxime Oquab, Léon Bottou, Ivan Laptev, and Josef Sivic, "Is Object Localization for Free? - Weakly-supervised Learning with Convolutional Neural Networks," in *CVPR*, 2015.

[5] Yan-Tao Zheng, Ming Zhao, Yang Song, Hartwig Adam, Ulrich Buddemeier, Alessandro Bissacco, Fernando Brucher, Tat-Seng Chua, and Hartmut Neven, "Tour the World: Building a Web-scale Landmark Recognition Engine," in *CVPR*, 2009.

[6] Zhen Li and Kim-Hui Yap, "Content and Context Boosting for Mobile Landmark Recognition," *IEEE Signal Processing Letters*, vol. 19, no. 8, 2012.

[7] Yifang Yin, Beomjoo Seo, and Roger Zimmermann, "Content vs. Context: Visual and Geographic Information Use in Video Landmark Retrieval," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 11, no. 3, 2015.

[8] Sakire Arslan Ay, Roger Zimmermann, and Seon Ho Kim, "Viewable Scene Modeling for Geospatial Video Search," in *ACM Multimedia*, 2008.

[9] Jiebo Luo, Dhiraj Joshi, Jie Yu, and Andrew Gallagher, "Geotagging in Multimedia and Computer Vision–a Survey," *Multimedia Tools and Applications*, 2011.

[10] Mordechai (Muki) Haklay and Patrick Weber, "OpenStreetMap: User-Generated Street Maps," *IEEE Pervasive Computing*, 2008.

[11] Josef Sivic, Bryan C. Russell, Alexei A. Efros, Andrew Zisserman, and William T. Freeman, "Discovering Object Categories in Image Collections," in *ICCV*, 2005.

[12] Jianchao Yang, Kai Yu, Yihong Gong, and Thomas Huang, "Linear Spatial Pyramid Matching using Sparse Coding for Image Classification," in *CVPR*, 2009.

[13] Thomas Brox, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik, "Object Segmentation by Alignment of Poselet Activations to Image Contours," in *CVPR*, 2011.

[14] Thibaut Durand, Nicolas Thome, and Matthieu Cord, "WELDON: Weakly Supervised Learning of Deep Convolutional Neural Networks," in *CVPR*, 2016.

[15] Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang, "Learning from Massive Noisy Labeled Data for Image Classification," in *CVPR*, 2015.

[16] Alan Joseph Bekker and Jacob Goldberger, "Training Deep Neural-networks based on Unreliable Labels," in *IEEE ICASSP*, 2016.

[17] Bo Zhang, Qinlin Li, Hongyang Chao, Bill Chen, Eyal Ofek, and Ying-Qing Xu, "Annotating and Navigating Tourist Videos," in *ACM SIGSPATIAL GIS*, 2010.

[18] Zhijie Shen, Sakire Arslan Ay, Seon Ho Kim, and Roger Zimmermann, "Automatic Tag Generation and Ranking for Sensor-rich Outdoor Videos," in *ACM Multimedia*, 2011.

[19] Yifang Yin, Zhijie Shen, Luming Zhang, and Roger Zimmermann, "Spatial-Temporal Tag Mining for Automatic Geospatial Video Annotation," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 11, no. 2, 2015.

[20] Yifang Yin, Guanfeng Wang, and Roger Zimmermann, "Automatic Geographic Metadata Correction for Sensor-rich Video Sequences," in *ACM SIGSPATIAL GIS*, 2016.

[21] Ying Zhang and Roger Zimmermann, "Camera Shooting Location Recommendations for Landmarks in Geo-space," in *IEEE MASCOTS*, 2013.

[22] Navneet Dalal and Bill Triggs, "Histograms of Oriented Gradients for Human Detection," in *CVPR*, 2005.

[23] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," in *CVPR*, 2014.

[24] Maxime Oquab, Leon Bottou, Ivan Laptev, and Josef Sivic, "Learning and Transferring Mid-level Image Representations Using Convolutional Neural Networks," in *CVPR*, 2014.