Advanced Mechanism in Learning and Recognition of OPS from Weakly Labeled Street View Images

Vaishali A. Mahale[1] Prof. P. M. Yawalkar[2]
[1]P.G. Student [2]Assistant Professor
[1,2]Department of Computer Engineering
[1,2]MET's Institute of Engineering, Adgoan, Nashik, SPPU, Maharashtra, India

Abstract— Mobile phones are powerful image and video processing device containing the other various features like high-resolution cameras, colour displays, and hardware-accelerated graphics. The various applications give rise to a key technique of daily life visual object recognition. On premise sign is a popular form of commercial advertising, widely used in our day to day life. The OPSs containing visual diversity associate with complex environmental conditions. Observing that such, real-world characteristics are lacking in most of the existing image data sets. In this, first proposed an OPS data set, namely OPS-62, in which totally 4649 OPS images of 62 different businesses are collected from Google's Street View. For addressing the problem of real-world OPS learning and recognition, developed a probabilistic framework based on the distributional clustering, in order to exploit the distributional information of each visual feature. Learning OPS images for more accurate recognitions and less false alarms. Experimental results shows that, using SURF descriptor technique applied on OPS-62 dataset outperform over the existing technique of using SIFT descriptor. Further, it shows marginal improvement in the average recognition rate. The proposed approach is simple, linear, and can be executed in a parallel fashion, making it practical and scalable for large-scale multimedia applications.

Key words: Real-World Objects, Street View Scenes, Learning and Recognition, Object Image Data Set

I. Introduction

The mobile device is computing equipment used to connect with the world for various purposes. Users depend on mobile device to maintain information and update. OPSs (On Premises Signs) shows great visual diversity accompanied with complex environmental conditions. Consider an example, user walk on the street and he simply point his mobile camera to a store to quickly access its related information, inquire special offers, and make reservations through his mobile without physically entering on that store.

Street view scenes are commonly captured by customers devices and they have more real-world characteristics lacking in most existing image datasets, e.g. perspective distortion, foreground and background clutter, etc. To learn a reliable OPS model for recognizing OPSs, a labeled dataset with a huge amount of real-scene images is required.

However, precisely labeling OPS categories and regions, i.e., generating strong labels for Learning involves a significant amount of human labor, and thereby is usually not feasible For training a real-scene OPS model. Instead of generating strong labels for real-scene images, an alternative learning technique, this is weakly supervised by a dataset with Each image labeled with the OPS category it contains, i.e., a weakly labeled image.

To create a recognizable image for a business to attract customers, each business has its own OPSs which is a visually consistent image for a brand and contains a mixture of text and graphics [5]. Therefore here, proposed a probabilistic framework for learning and recognition of OPSs in real-world images. Real-world characteristics of OPSs, such as viewing angles, arbitrary size, occlusions, varying lighting conditions, foreground and background clutter, etc., make logos, texts, or trademarks in OPSs fill a smaller area by other objects in real scene images. All these characteristics fail to identify texts or logos in OPSs of existing solutions. The main approach is to take advantage of probabilistic framework to extract discriminative visual words of each OPS category and therefore able to localize and recognize each OPS within images by using learnt OPS model.

II. Literature Review

The goal of learning and recognition of OPSs in real-world images can be shows as a problem of object recognition and localization. Firstly, understand the value of OPS, the importance of signage in the business community. D. Conroy studied and explained the role of the signage in businesses and organizations. [1]

Bag-of-features technique used for content based image classification gaining to their less complexity and good performance. E. Nwark shows that for random sampling gives equal or better classifiers than the sophisticated multi-scale interest operators that are in common use for a representative selection of commonly used test databases and for moderate to large numbers of samples. [2]

W. H. Cheng.et.al introduced a novel framework for video adaptation based on content re-composition. The objective was to provide effective small size videos which intentioned the important aspects of a scene while faithfully retaining the background context. A generic video attention model was developed to extract user-interest objects, in which a high-level combination strategy was proposed for fusing the adopted three types of visual attention features: intensity, color, and motion.[3]

J.harel and c.koch introduced a new visual model named as Graph-Based Visual Saliency (GBVS). It contains of two steps, first is creating activation maps on certain feature channels, and second is normalizing them in a manner which highlights clearness or brightness and to recognize combination with other maps. This model is less complex and biologically acceptable as it is naturally parallelized. [4]

To discover association rules at multiple resolutions in order to identify frequent spatial configurations of local features that correspond to classes of logos appearing in real world scenes spatial pyramid mining technique is used. [5]

G.Kim proposed a novel representation and formulation of the problem inspired by the tools commonly used for the analysis of complex networks such as the WWW and social networks. They proposed an approach for learning visual models of object categories in an unsupervised manner in which they first build a large-scale complex network which captures the interactions of all unit visual features across the entire training set and infer information.[6]

C. H. Lampert proposed an simple yet powerful branchand- bound scheme that allows efficient maximization of a large class of classifier functions over all possible subimages. It converges to a globally optimal solution typically in sublinear time.They shown that how their method was applicable to different object detection and retrieval scenarios.[7]

An Scale- Rotation invariant Pattern Entropy (SR-PE) algorithm, for the detection of near-duplicates in large-scale video corpus. SR-PE is a novel pattern evaluation technique capable of measuring the spatial regularity of matching patterns formed by local key-points. More importantly, the coherency of patterns and the perception of visual similarity, under the scenario that there could be multiple ND regions undergone arbitrary transformations respectively, was carefully addressed through entropy measure. To demonstrate their work in large-scale dataset, a practical framework composed of three components: bag-of-words representation, local key-point matching and SR-PE evaluation, is also proposed for the rapid detection of near-duplicates.[8]

T.yeh et.al studied an efficient method for concurrent object localization and recognition based on a data-dependent multi-class branch-and-bound formalism. They presented experimental results that demonstrate the merit of their algorithm in terms of recognition accuracy, localization accuracy, and speed, compared to baseline approaches including exhaustive search, implicit-shape model, and efficient sub-window search.[9]

F. Schroff et.al introduced to automatically generate a large number of images for a specified object class. A multi-modal approached employed text, meta-data and visual features used to gather many, high-quality images from the web. Candidate image obtained by a text based web search querying on the object identifier. The web pages and the images they downloaded. The task was then to remove irrelevant images and re-rank the remainder. The principal novelty was in combined text/meta-data and visual features in order to achieve a completely automatic ranking of the images.[10]

R.Fergus has been analyzed a simple approach to learning models of visual object categories from images gathered from Internet image search engines. They describe two simple approaches, derived from the pLSA technique for text document analysis that can be used to automatically learn object models from these data. They use two applications of the learned model: first, to re-rank the images returned by the search engine and second, to recognize objects in other image data sets [11].

Y.H. kuo has proposed supplementary visual words discovery through visual and textual clusters in a scalable and unsupervised fashion. This paper highly detailed study on the problems of current BoW model and the requirement for semantic visual words to improve the recall rate for image object retrieval [12].

S.Romberg has proposed a cascaded index for scalable multi-class recognition of logos. For the evaluation of their given system, they had construct and released a logo recognition benchmark which consists of manually labelled logo images, complemented with non-logo images, all posted on Flickr dataset [13].

To make effective and scalable framework used for the customers purpose introduce a new mobile queue-card management system which offers more freedom to customers by allowing image-based queue-card retrieving and service-information querying actions using mobile phones [14].

J.Revaud has been proposed to gain a statistical model for the distribution of incorrect detections output by an image matching algorithm. It results in a novel scoring criterion in which the weight of correlated key point matches is reduced, penalizing irrelevant logo detections [15].

III. PROBLEM DEFINITION

Several researchers research on detection of texts in objects for example in OPSs, products and images. Observing that when viewing angle of camera changes, significantly changed in texts in their shapes due to distortions or partially occluded and therefore cannot be well recognized by the existing solution.

IV. PROPOSED SYSTEM

The main motive of this research is to detecting the problem of real world OPS learning and recognition of weakly labeled street view images. Here, to tackle this problem by developing a framework based on distributional clustering, in which Tsung-hung tsai developed a framework which is used to exploit distributional information of each visual feature. OPS- 62 dataset demonstrated the performance of given approach over a latent semantic analysis model for more accurate recognition and also improvement in average recognition rate. To implement the proposed approach two main algorithms are used:

1) Visual saliency based codebook generation of OPS categories.
2) OPS modeling and recognition using distributional clustering.

A. System Architecture

In this system, user gives input as image which is captured through mobile cameras. System gives input images and perform the actual proposed framework on given input image. In this framework two basic algorithms are used: first is visual saliency based codebook generation of OPS categories. In this algorithm first, filter out the background region for minimizing the number of noisy visual word using visual saliency analysis. After removing the background noise, visual feature are extracted using dense sampling strategy and Opponent SIFT descriptor for codebook Generation. After acquiring a codebook for each

OPS category compute a discriminative subset and apply a distributional clustering to collecting of all the code words in OPS categories into two disjoint clusters. Then allow the concurrent OPS recognition and localization in super-pixel level using obtained OPS and background models. Second algorithm is OPS modelling and recognition using distributional Clustering here, super pixel segmentation performed on input image. After the segmentation visual feature extraction in super pixel level then recognized the OPS image with learned dataset.

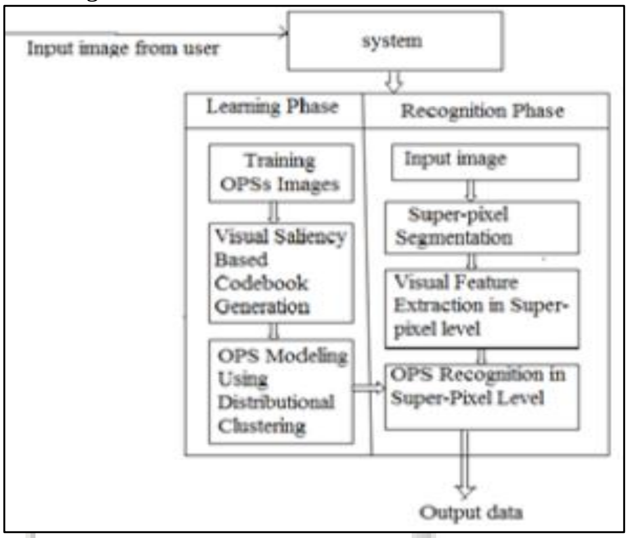

Fig 1: System Architecture

B. *Algorithms*

1) *Visual Saliency Based Codebook Generation of OPS Categories*

a) Visual Features

The Opponent-SIFT descriptor method is used to extract visual features in the training images. The opponent color space is a color model used for matching the human perception of color, which can be converted from the RGB color space:

$$\begin{pmatrix} O1 \\ O2 \\ O3 \end{pmatrix} = \begin{pmatrix} \frac{R-G}{\sqrt{2}} \\ \frac{R+G-2B}{\sqrt{6}} \\ \frac{R+G+B}{\sqrt{3}} \end{pmatrix} \quad (1)$$

Where,

O_3 -The intensity information,

O_1 &O_2 - the color information.

The Opponent-SIFT descriptor defined a 384-dimensional feature vector by concatenating three 128-dimensional SIFT descriptors extracted from all of the channels in the opponent color space, respectively.

b) Determination of Salient Regions

A MATLAB implementation of Graph-Based Visual Saliency (GBVS) is applied to compute saliency values of pixels, with the values normalized to a range between 0 and 1 for generating a saliency map for each image. A saliency value threshold θ, they can divide an OPS image into two disjoint regions, one of high saliency and another is low saliency. The problem now becomes how to determine accurate θ value. I is a set of OPS images, I = It and pixel-level masks to point where OPS pixels are, let ∅It(θ) be the ratio of the total OPS area in the high saliency region to the size of the high-saliency region, and ϕIt(θ) be the ratio of the total OPS area in the high saliency region to the total OPS area in the image It. They aim to select a threshold value θ that can increase It(θ) and ϕIt(θ) simultaneously. The objective function O_ can be calculated as follows:

$$\Theta^* = \arg_{\theta} \max \frac{1}{|I|} \frac{\emptyset It(\theta).\varphi It(\theta)}{\emptyset It(\theta)+\varphi It(\theta)} \quad (2)$$

To formulate the value of θ, randomly select out a small set of images and w.r.t discriminative values of θ. After finding the value of θ, the high-saliency regions of each image can be directly extracted through the visual saliency analysis.

2) *OPS Modeling and Recognition Using Distributional Clustering*

After getting a codebook Ω_{C_i} for OPS category Ci as the category alphabet, they calculate a discriminative subset $\Omega^{+}_{C_i}$ from $\Omega^{-}_{C_i}$ as the visual model of the corresponding OPS category using the distributional clustering.

a) OPS Modeling Using Distributional Clustering:

An OPS image is considered to be generated by a parametric model π. Initially, the probability of generating an image It can be written as a sum of total probability over all the OPS categories:

$$P(I_t|\pi) = \sum_{\forall C_k \in X} P(I_t|C_k;\pi)P(C_k|\pi), \quad (3)$$

Where X is the set of all the adopted OPS categories, $P(Ck|\pi)$ denotes the category priors, and $P(It|Ck; \pi)$ is the probability of It given the OPS category Ck. The probability distribution of $P(Ck|\pi)$ is consider to be uniform, i.e. $P(Ck|\pi) = 1/|X|$. $P(It|Ck; \pi)$ is calculated. After the probability $P(Ck|It; \pi)$ for all OPS categories and then picking the category with the highest value. Discriminative OPS code words need to be first screened out for the purpose of stability and reliability. From a communication theory view as, a code word is discriminative of the OPS category *Ci* if it occurs much more often in the images related to *Ci* than those belonging to other OPS categories. Therefore, for each OPS category *Ci*, here apply the distributional clustering, here classified it into two disjoint clusters, i.e. the discriminative subset $\omega+ Ci$ and the less-discriminative subset $\omega\ Ci$.The distributional representation of a code word wj can be defined by the conditional probability mass function:

$$P_{C|w}(C = C_i|W = w_j) = \frac{\emptyset(w_j,C_i)}{\sum_{\forall C_k \in X} \emptyset(w_j,C_k)} \quad (4)$$

Where $\phi(wj, Ci)$ shows the frequence that the code word wj occurs in the images belonging to Ci . The code word wj with a sharper distribution $PC|W(C|W = wj)$ would have stronger discriminative capability. Since the numbers of generated code words belonging to the OPS and the background would be very unbalanced, instead of using the similarity based clustering like the k-means algorithm (k=2 in our case), they determine the cluster membership of a code word by a pre-trained threshold θ :

$$w_j \in \{\Omega^{+}_{C_i}$$
$$\{\Omega^{-}_{C_i} \quad (5)$$

In addition to the OPS models obtained by Equation, we can also build a background model $\omega+ BG$ by collecting all the less discriminative code words: The distributional information of the code words is ignored in the pLSA based approaches; such that it becomes extremely difficult even to determine which code words (latent topics) are related to the specified categories.

b) OPS Recognition:

Given an input image $\hat{I}$, they consider the concurrent OPS recognition and localization in super-pixel level using the obtained OPS and background model. For each super-pixel $\hat{S}$ in $\hat{I}$, a visual feature is extracted and they search for the most similar code word:

$$w^* = \text{Arg } w_k \min \mathrm{D}\left(w_k, \hat{f}_s\right), \forall w_k \in \Omega \qquad (6)$$

The super-pixel S is then allocating with the same category label of w_, either an adopted OPS category or the background. In comparison, the recognition scheme is simple, linear, and can be executed in a parallel fashion. This makes our approach used for practical use.

C. *Mathematical Model*

The Deterministic finite automaton consists of five tuples as follows:

M = {Q,Σ, δ, q0, f}

Q={q0, q1, q2, q3, q4 }

qΣ0: Initial State,

f: Finite state

(q4): {t, k, c, q}

δ: Transition Function

δ (q0, t)=q1, δ (q1, k)=q2, δ (q2c)=q3, δ (q3, q)=q4

where,

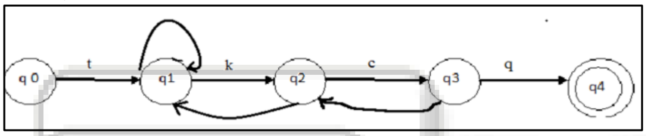

Fig. 2: Deterministic Finite Automata.

q0 - image as input.

q1 - Determine Salient region and visual features extraction for generating the codebook.

q2- segmentation and visual feature extraction on input image.

q3- input image recognition with training images.

q4- image with offers and related information.

t - codebook generate.

k - segmentation and visual feature extracted.

c - obtain recognize image.

q - image, related info and location.

V. Results and Discussion

A. *Dataset*

In real world images objects are visible in different arbitrary size, position, scale, viewing angles, accompanied with perspective of distortion, varying brightening condition, foreground and background clutters. Thus, it is observed that such a real world features are lacking in the most of existing dataset .So, we focus on the street view images which are weakly labelled and used for commercial advertising.

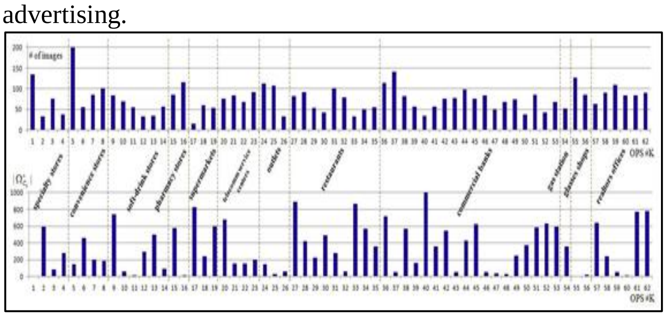

Fig. 3: Statistics of OPS-62 dataset

An OPS data set, namely OPS-62, in which totally 4649 OPS images of 62 different businesses are collected from Google Street View. The adopted OPS categories are classified into 12 business type.

B. *Performance Parameters for Clustering*

1) *Average Precision*

Precision and recall are the basic measures used in evaluating search strategies. There is a set of records in the database which is relevant to the search topic. Records are assumed to be either relevant or irrelevant (these measures do not allow for degrees of relevancy).

2) *Average Recall*

In Information Retrieval recall is the ratio of the number of relevant records retrieved to the total number of relevant records in the database. It is usually expressed as a percentage. For example for text search on a set of documents recall is the number of correct results divided by the number of results that should have been returned.

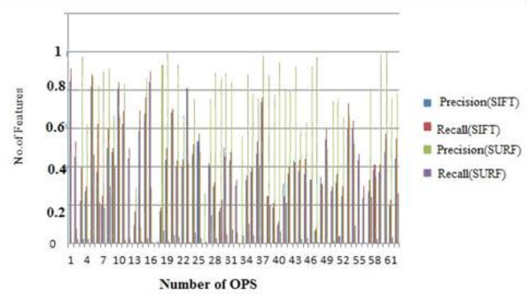

Fig. 4: Difference between SIFT & SURF on basis of precision & recall

Figure shows the graphical representation of OPS-62 model for precision the recall. X axis shows the number of OPS and Y axis shows the number of input images feature. Figure Shows better recognition using SURF algorithm comparatively SIFT algorithm.

C. *Comparison between SIFT and SURF Algorithms*

OPS	Previous System		OPS system		OPS	Previous system		OPS System	
	AP	AR	AP	AR		AP	AR	AP	AR
1	0	0	0.914286	0.85	31	0.8454	0.071	0.48	0.4333
2	0.079	0.021	0.533333	0.4545	32	0.057	0.002	0.333333	0.3043
3	0.977	0.021	0.4	0.2222	33	0.558	0.043	0	0
4	0.625	0.023	0.3	0.2727	34	0.882	0.101	0.357143	0.3333
5	0.874	0.463	0.882353	0.8167	35	0.782	0.042	0.4	0.375
6	0.823	0.212	0.625	0.375	36	0.76	0.408	0.533333	0.4706
7	0.901	0.187	0.25	0.2	37	0.98	0.005	0.763158	0.7381
8	0.914	0.2989	0.6	0.5	38	0.881	0.203	0.25	0.25
9	0.668	0.406	0.5	0.48	39	0.777	0.008	0.214286	0.1875
10	0.658	0.003	0.842105	0.8095	40	0.949	0.061	0.111111	0.1
11	0.833	0.018	0.692308	0.625	41	0.811	0.214	0.25	0.3125
12	0.029	0.005	0.5	0.4444	42	0.799	0.002	0.4	0.3636
13	0.293	0.01	0.166667	0.1	43	0.924	0.01	0.421053	0.4348
14	0.081	0.013	0.692308	0.5882	44	0.602	0.025	0.44	0.3793
15	0.862	0.03	0.761905	0.63	45	0.633	0.028	0.444444	0.3636
16	0.293	0.007	0.896552	0.8438	46	0.927	0.006	0.333333	0.3333
17	0.014	0.008	0	0	47	0.976	0.008	0.076923	0.0667
18	0.936	0.066	0.1875	0.1667	48	0.31	0.004	0.3125	0.35
19	0.994	0.015	0.5	0.4375	49	0.5	0.002	0.6	0.5455
20	0.711	0.042	0.7	0.6818	50	0.746	0.013	0.3	0.2727
21	0.934	0.032	0.434783	0.4	51	0.76	0.039	0.363636	0.32
22	0.667	0.004	0.4375	0.4	52	0.657	0.005	0.3	0.25
23	0.418	0.007	0.807692	0.8148	53	0.671	0.009	0.733333	0.6
24	0.755	0.059+	0.517241	0.4686	54	0.524	0.095	0.642857	0.6
25	0.48	0.029	0.571429	0.5313	55	0	0	0.470588	0.4324
26	0.265	0.008	0	0	56	0	0	0.3	0.24
27	0.753	0.147	0.409091	0.4167	57	0.811	0.243	0.333333	0.2727
28	0.893	0.026	0.32	0.2963	58	0.348	0.022	0.411765	0.381
29	0.866	0.229	0.190476	0.1667	59	0.989	0.004	0.413793	0.375
30	0.89	0.05	0.454545	0.05	60	1	0.002	0.571429	0.48
					61	0.759	0.035	0.227273	0.2
					62	0.783	0.265	0.55	0.4444
					Avg.	0.686	0.071	0.786	0.087

Table 1: Comparison between previous and proposed system

OPS system using SURF algorithm which clearly shows the improved result as compare to the SIFT algorithm. Following table shows that previous system work on SIFT technique and OPS system work on SIFT and SURF technique.

Table 1 shows the difference between previous work and OPS system using SURF algorithm distributional clustering. The table shows average precision 0.914286 and average recall 0.85 for improved system which clearly shows the improved result. Result show's SURF algorithm gives better OPS recognition rate as compare to SIFT algorithm.

VI. Conclusion

Learning and recognition technique have been solving the problem of real world street view challenges to identify business entities in street view images. Learning technique is exploited to benefit the selection of discriminative visual words and construct effective OPS model by using distribution clustering. The OPS - 62 dataset which contain more real world images for visual object recognition. The visual saliency based codebook generation and OPS recognition using distributional clustering algorithms will able to improve average recognition rate as compared to the previous system. Using SURF algorithm, the OPS recognition is improved. The experimental result shows that the SIFT algorithm gives 81.2 % of the OPS recognition rate and SURF gives 82.8 % OPS recognition rate. Hence, SURF algorithm gives better recognition rate as compare to SIFT algorithm. However, in view of the low average recall values relatively, the OPS recognition in real-world scenes is still challenging problem.

References

[1] D. Conroy," What's Your Signage (How On-Premise Signs Help Small Businesses Tap into a Hidden Profit Center).", New York, NY, USA: State Small Bus. Develop. Center, 2004.

[2] E. Nowak, F. Jurie, and B. Triggs, "Sampling strategies for bag-of features image classification", in Proc. ECCV, 2006, pp. 490-503.

[3] W.-H. Cheng, C.-W. Wang, and J.-L. Wu, Video adaptation for small display based on content recomposition, IEEE Trans. Circuits Syst. Video Technol., vol. 17, no. 1, pp. 4358, Jan. 2007.

[4] J. Harel, C. Koch, and P. Perona,"Graph-based visual saliency", in Proc. ANIPS, 2007, pp. 545- 552.

[5] J. Kleban, X. Xie, and W.-Y. Ma, "Spatial pyramid mining for logo detection in natural scenes", in Proc. IEEE ICME, Apr. 2008, pp. 1077-1080.

[6] G. Kim, C. Faloutsos, and M. Hebert, Unsupervised modeling of object categories using link analysis techniques, in Proc. IEEE Conf. CVPR, Jun. 2008, pp. 18.

[7] C. H. Lampert, M. B. Blaschko, and T. Hofmann, Beyond sliding windows: Object localization by efficient subwindow search,in Proc. IEEE Conf. CVPR, Jun. 2008, pp. 18.

[8] T. Yeh, J. Lee, and T. Darrell, Fast concurrent object localization and recognition,in Proc. IEEE Conf. CVPR, Jun. 2009, pp. 280287.

[9] F. Schroff, A. Criminisi, and A. Zisserman, Harvesting image databases from the web, IEEE Trans. Pattern Anal. Mach. Intell., vol. 33, no. 4, pp. 754766, Apr. 2010.

[10] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman, "Learning object categories from internet image searches" Proc. IEEE, vol. 98, no. 8, pp. 1453-1466, Aug. 2010.

[11] Y.-H. Kuo, H.-T. Lin, W.-H. Cheng, Y.-H. Yang, and W. H. Hsu,"Unsupervised auxiliary visual words discovery for largescale image object retrieval", in Proc. IEEE Conf. CVPR, Jun. 2011, pp. 905-912.

[12] S. Romberg, L. G. Pueyo, R. Lienhart, and R. van Zwol, "Scalable logo recognition in real-world images ", in Proc. 1st ACM ICMR, 2011, pp. 1-25.

[13] C.-W. You, W.-H. Cheng, A. W. Tsui, T.-H. Tsai, and A. Campbell, "MobileQueue: An image-based queue card retrieving system through augmented reality phones,", in Proc. 14th ACM Int. Conf. Ubiquitous Comput, 2012, pp. 1-2.

[14] J. Revaud, M. Douze, and C. Schmid, "Correlation-based burstiness for logo retrieval,", in Proc. 20th ACM Int. Conf. Multimedia, 2012, pp. 965-968.

[15] Tsung-Hung Tsai, Wen-Huang Cheng, Chuang-Wen You, Min-Chun Hu, Arvin Wen Tsui, and Heng-Yu Chi"Learning and Recognition of On-Premise Signs from Weakly Labeled Street View Images",, IEEE TRANSACTIONS ON IMAGE PROCESSING, VOL. 23, NO. 3, MARCH 2014.