IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

DATA VISUALIZATION AND PREDICTION USING MACHINE LEARNING TECHNIQUES

[1]Suraj Purushottam Gulhane, [2]Dr. A. B. Raut,

[1] Student, [2]Proffessor,

[1]Computer Science and Engineering,

[1]H.V.P.M.'s College of Engineering & Technology, Amravati, India

Abstract: Most of the people in our country love to travel, shopping, etc. and it is commonly seen that we don't have information about some places so it is difficult to ask any person about some place because of language barrier or may be any other problems. As the solution of this problem, we are developing an android application that will help user to get out details of any place such as shop, mall, hotel, restaurants, etc. The project is about optical character recognition with image processing and data mining. Optical character recognition is the core part of this model which reads each character through image. In this model a person who interested to buy something or wants to stay in hotel then by using smart phone he needs to capture the image of name of hotel or shop and then this application will extract text from image and will search for information related to respected place and get out entire details. This project will help them to get out entire details of any restaurants , hotels, place name, malls and any other places before entering into them this detail content all services they provide, product in shop or mall, current offers, types of room(AC/non AC),etc.

Index Terms **– Optical Character Recognition, Machine Learning, Image Processing.**

I. INTRODUCTION

Text recognition is the process of extracting text data from given images which is also considered as Optical Character Recognition (OCR) [1]. The problem of text recognition from images belongs to a class of computer vision, which may involve OCR or text detection or handwritten text detection. Simple text detection refers to find out the printed text on the images. Along with all the contents in images, text information is of immense interests, as it can be effortlessly understood by human and computer, hence finds wide applications in multimedia systems, digital libraries, and geographical information systems and many more [2]. The habitual recognition of text in natural images, is a significant challenge for visual understanding. Text, as the physical life of language, is one of the basic tools for preserving and communicating in pattern. Contemporary world is considered to be interpreted through the use of textual cues in addition to other labels, so text finds itself spread during many videos and images. Habitually, text recognition has been paying attention on document images, where OCR techniques are appropriate to digitize planar, manuscript based documents. Still, when applied to natural scene images, these document OCR techniques fail as they are tuned to the largely black along with white, line based background of written documents. The text that occurs in natural scene images is vastly changeable in exterior as well as layout, being drawn from a huge number of fonts moreover styles, distress from incompatible illumination, occlusions, orientations, noise in addition, and the presence of background objects causes unauthentic false-positive detections. Text recognition involves generating candidate character or word region detections, while word identification takes these proposals furthermore infers the words depicted.

Despite the fact that investigating OCR blunder modification, we came diagonally a range of examples of OCR text which was also tough for diverse online spell checkers to confirm plus correct. The suggestions generated were awfully distant from the truth for the majority of the words.

The chief reasons for this are:

1) Limited glossary is majorly imperfect.
2) Limited realization of language precise set of laws for authenticate appropriately spelled vocabulary.
3) The propose for such spell checkers is based on typing blunder, not on regular OCR n-gram confusions.
4) Ambiguities in suggestions from glossary due to huge number of basic word forms. A number of examples of mistaken OCR output for which the online spelling improvement systems were unsuccessful. We provide accurate suggestions for these words all the way through our proposed method.

II. RELATED WORK

With rising use of digital devices like digital cameras, mobile phones, PDAs, content based image analysis methods have caught serious concentration in the current past years. An image text information mining system is divided into four stages: text detection, text identification, text localization, text mining as well as text enhancement.

Among these stages, text detection and text localization are significant. Numerous methods are projected for text detection with text localization problems [3], a number of them achieved high-quality results for specific relevance. The presented methods of text detection with text localization can be generally categorized into two groups: Section based plus connected component (CC) based text detection. Various issues in addition to challenges specifically for Urdu text based on OCR have been discussed by Zaki et al. [4].

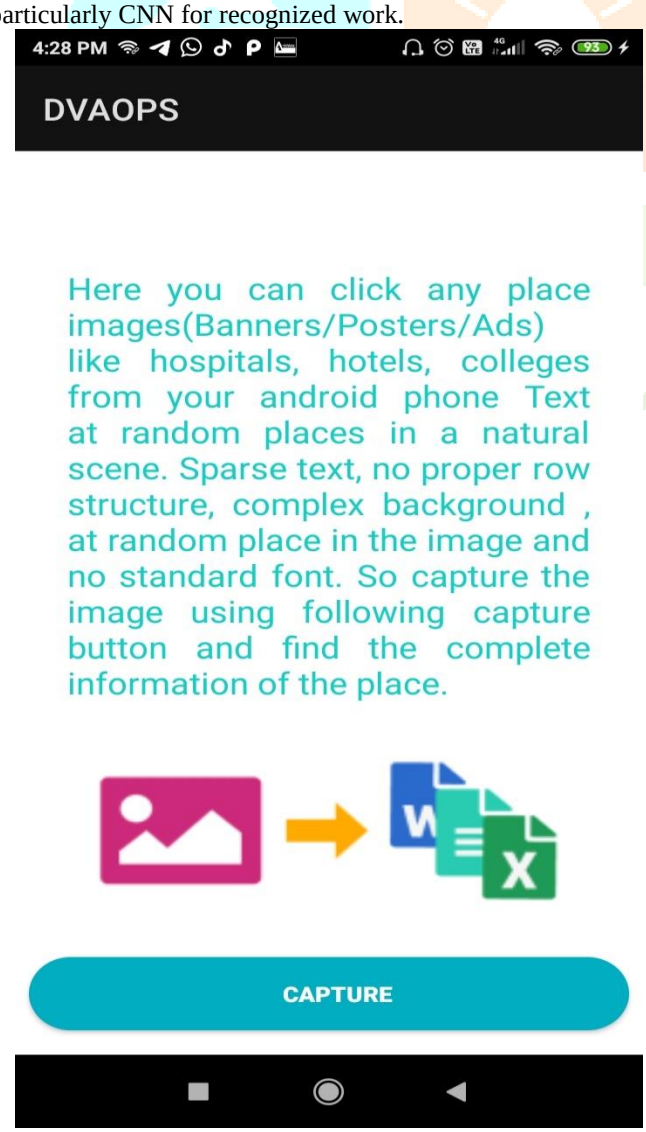
Sharma et al. [5] works on OCR procedure based on Convolutional Neural Network (CNN), their method uses this OCR method to mine information of a student filled in a particular form. This form contains 175 cells. A quantity of these cells are filled with capital alphabets, others are few numbers. It also explains how to identify each cell using the feature and CNN. Their method has achieved a good accuracy for both numerical as well as alphabetical data.

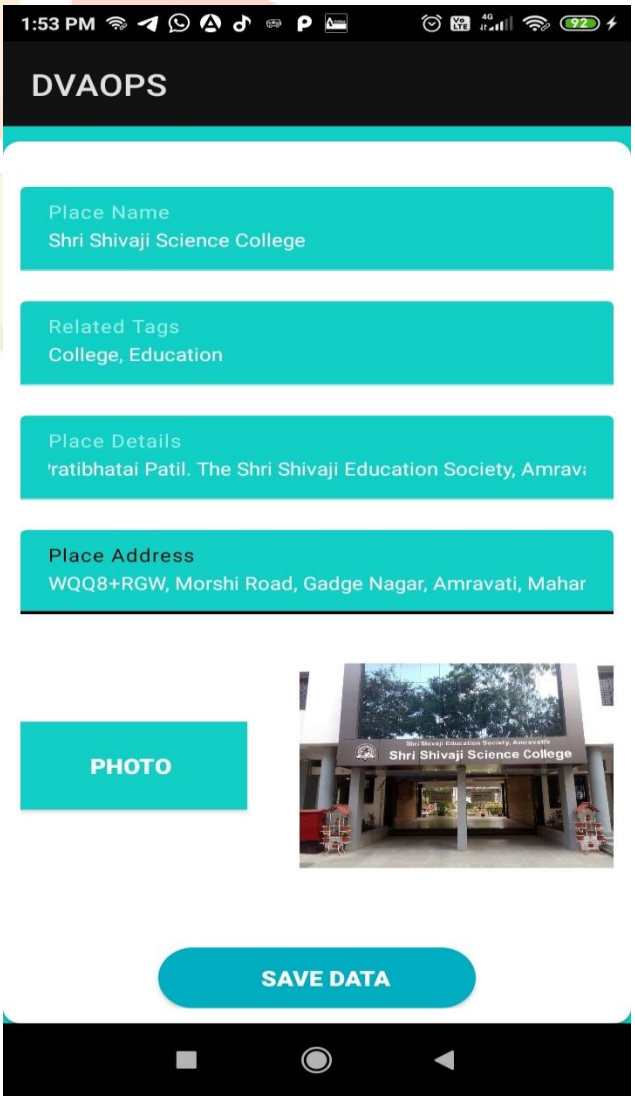
Prakash et al. [6], discussed with the OCR challenge for the Telugu, They have created a database for Telugu, a deep learning Technique, in addition to a consumer server solution for the online procedure of the algorithm.

Moteelal and Murthy [7] have proposed a framework of continuous content locations that usually deal with different ideas and recognize content. This technique depends on the relevant section. This approach can deal with the content on dispose image plus not bright districts. They build up windows based applications that can track content as well as isolate content in this field. They use Windows telephone of 8.1 rendition, OCR, C#. This approach is a direct and open source. Mol et al. [8], provides a method for text detection as well as recognition in pictures. This method emphasizes Fractional Poisson enhancement for removing Laplacian noise of the input image. Next, the Maximally Stable Extremal Regions (MSER) emphasized on the edge of the preprocessed image. Local filtering is used to filter areas of non-texture in addition to be recognized by an OCR. The results of this method are better than previous methods in terms of Peak Signal to Noise Ratio (PSNR) and Structural Similarity (SSIM) calculations. This method has used standard ICDAR dataset that have nearly real time images.

Purkaystha et al. [9], used Convolutional Neural Network (CNN) for Bengali handwritten character detection with satisfactory accuracy.

Razik et al. [10], equipped a database of handwritten Bangla numeral named SUST:BNHD. He used deep learning particularly CNN for recognized work.





The mobile device is computing cquipment used to connect with the world for various purposes. Users depend on mobile device to maintain information and update. 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 model for recognizing place names, a labeled dataset with a huge amount of real- scene images is required. However, precisely labeling categories and regions, 1.e., generating strong labels for Learning involves a significant amount of human labor, and thereby is usually not feasible for training a real-scene model. Instead of generating strong labels for real-cene images, an alternative learning technique, with the 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 which is a visually.consistent image for a brand and contains a mixture of text and graphics Conroy et al. Therefore here, proposed a probabilistic framework for learning and recognition of OPSs in real-world images. Real-world characteristics, such as viewing angles, arbitrary size, occlusions, varying lighting conditions, foreground and background clutter, etc., make logos, texts, or trademarks that fill a smaller area by other objects in real scene images. All these characteristics fail to identify texts or logos in place names of existing solution.

III. ANALYSIS OF PROBLEM

Learning based approaches are then adopted as promising solutions. Supervised learning is the main stream paradigm in modeling visual objects for recognition and localization. addressed the problem of concurrent object recognition and localization according to the data-dependent region hypothesis. Presented a semantics-preserving bag-of-words model by learning a distance metric to minimize the distance between the visual features with the same semantics. Learning based approaches are then adopted as promising solutions. Supervised learning is the mainstream paradigm in modeling visual objects for recognition and localization. addressed the problem of concurrent object recognition and localization according to the data-dependent region hypothesis. Presented a semantics-preserving bag-of-words model by learning adistance metric to minimize the distance between the visual features with the same semantics.

IV. PERPOSED WORK

Text detection and recognition has brought many challenges to the research community. Challenges in scenes such as complex backgrounds, small font sizes, inappropriate font styles, stable photos, emerging photos, bright lights have made the task difficult. This study provides a flexible method to extract texts from challenging images by employing basic image processing techniques. The system builds an application that can recognize the English content present in an image captured by using a smart phone and then display the recognized text into editable format on the mobile screen. In addition, the results of proposed app are compared with other Commercial mobile apps. The planned system has used the characters from figure with the assist of ANN for recognition. Following implementation steps are involved in general. Database gathering: In favor of the detection of text from a hidden image, initially we require to build up our dataset, which contains the description of the character recognized from the analysis to demonstrate the text enclosed by the picture. Pre Processing: For the period of this step preprocessing schemes such as quantization moreover resizing is used. As the major complexity in the image processing field is the range of the image which depends on the image resolution means number of pixel limited by the picture. Segmentation with fundamental filtering: After the preprocessing step the subsequent step is to mine the thought from the image that somewhere manuscript is near the image this can be finished by means of Edge Canny detector filter.

3.1 Binarization

Gray image undergoes thresholding and background deduction. After this stage, only the letters of the stroke are left as foreground represented by 1 and 0 for the remaining image background. A character accumulates the Boolean
matrix that is used to store either 1's or 0's. The dark pixel is stored as 1's and dark pixel is stored as 0's with the help of the image zoning method. The pseudocode for binarization is given below, Pseudo Code for Binarization:
1 Take input image for the noise removal process
2 Examine the intensity of a pixel inside the image.
3 IF intensity of a pixel is less than threshold value, Substitute the intensity with 0 Else Substitute the intensity with 1 End IF
4 Repeat the process for every pixel in the image.

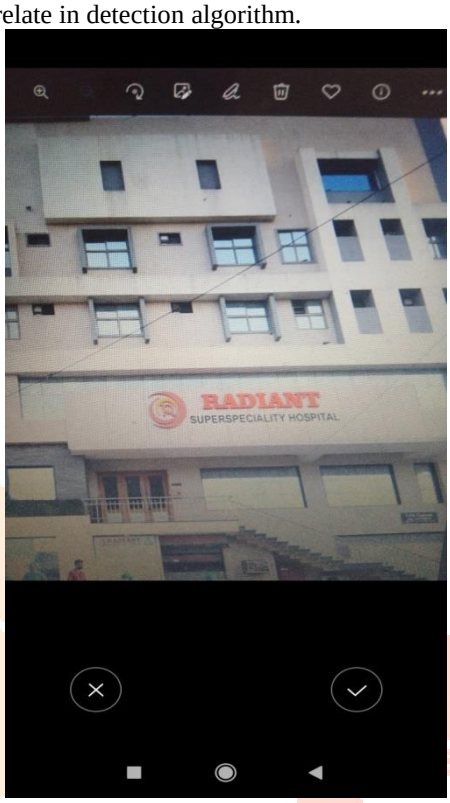
3.2 Segmentation

Segmentation is the process of separating characters from an image. When we get characters separately we can perform further processing on it. Segmentation is comprising of three groups: character, word and line segmentation. In this approach, character and line segmentation is applied. After binarization it is assumed that within each restricted section, gray level values of foreground pixels are lower or higher than the standard intensity, connected components with 255 value are extracted as the candidate text components 100 values are not considered.

3.3 Text recognition

This project focuses on mobile phones text recognition. For simplicity, we execute this step on a server with Google's open source OCR i.e. Tesseract (http://code.google.com/p/tesseract-ocr/) is used as an optical text recognition engine. Character recognition deals with cropping a particular region from an image, perform the correlation and write in the file. Before you perform the correlation, you need to load an existing saved model template so that it identifies the recognized character that has been found in the template. After achieving the character by character segmentation, save the image of the character. An

important field of digital image processing and Artificial Intelligence. It is widely used in data entry from books, bank documents, research papers, office papers or written all types of data. It is also used to save the old and moldy written materials like hand written books, documents etc. Early on versions are mandatory to be programmed with images of every character and worked on one or else multiple fonts at a single time. Therefore, smart systems based on Artificial Neural Network are considered which adjust themselves to the changeable fonts of characters. Characters are in fact the typical alphabets of any language. Character detection systems are chiefly classified into online and offline character detection system. In online character detection system, characters are detected at real time once it is written. Offline character detection can be additional divided into printed and handwritten characters detection. In offline character detection, handwritten characters are scanned from manual paper and transformed into the form of gray image to relate in detection algorithm.



3.4 Dataset

Instead of generating strong labels for real-scene images, we resort to an alternative learning technique, which is weakly supervised by a dataset with each image labeled with the OPS category it contains, i.e., our, learning involves a significant amount of human labor, and thereby is usually not feasible for training a real-scene model. Instead of generating strong labels for real-scene images, we resort to an alternative learning technique, which is weakly supervised by a dataset with each image labeled with the category it contains, i.e., a weakly labeled image, learning involves a significant amount of human labor, and thereby is usually not feasible for training a real-scene model. Instead of generating strong labels for real- scene images, we resort to an alternative learning technique, which is weakly supervised by dataset with each image labeled with the category it contain I.e., a weakly labeled image.

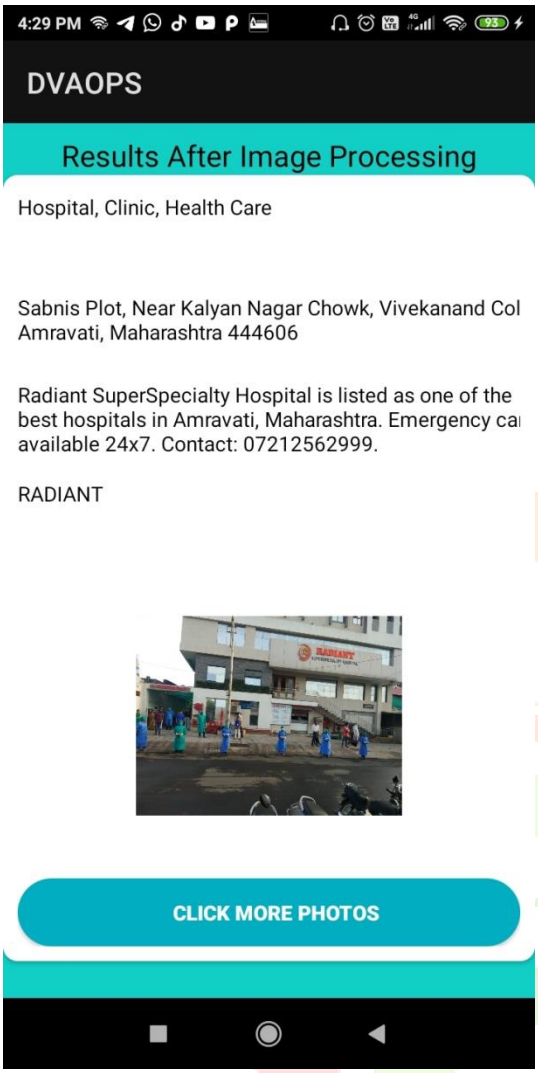
Most of the people in our country love to travel, shopping, etc. and it is commonly seen that we don't have information about some places so it is difficult to ask any person about some place because of language barrier or may be any other problems. As the solution of this problem, we are developing an android application that will help user to get out details of any place such as College, shop, mall, hotel, restaurants, etc.

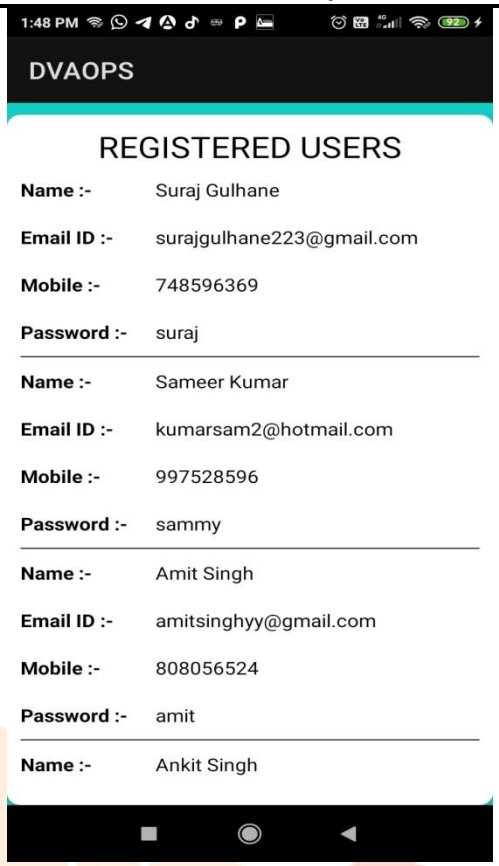
The project is about optical character recognition with image processing and data mining and machine learning technique. Optical character recognition is the core part of this model which reads each character through image. In this model a person who interested to buy something or wants to stay in hotel then by using smart phone he needs to capture the image of name of hotel or shop and then this application will extract text from image and will search for information related to respected place and get out entire details. This project will help them to get out entire details of any restaurants, hotels, name, malls and any other places before entering into them this detail content all services they provide, product in shop or mall, current offers.

V. RESULT AND DISCUSSION

To the authors' knowledge there is no any system available in the literature that extracts complete text. Furthermore, it has been observed that no solo algorithm is robust for recognition of an unrestrained range of text appearing in the wild. Various methods have been residential to mine text from composite color background pictures. Text eminence is low due to the existence of noise in addition to image decoding and encoding methods. Resolution of manuscript on the display is low, is also a problem as OCR's allow 300dpi resolution manuscript for getting high-quality text recognition results. Mining of text from wild defy over OCR of paper images. Challenges comprises of: worse resolution, indefinite text color, mysterious text sizes, location, orientation, and layout, unrestrained background, Low contrast can effect into impossible difference between background and character strokes. Recognition of sample in the intersections and distances to describe the rule that was one of the significant feature of our method. The distance outcome filtered with intersection outcome get better precision of detection of characters. In

view of the fact that, the method doesn't have segmentation complication, the time in use to identify a character is extremely small. The method mines the same quantity of aspect from all the letters. Consequently, overall accurateness is enhanced. For precision of 91% for a testing dataset of 140 characters confirm that the projected algorithm doesn't show any over appropriate anomalies. The method has only considered the minimum and maximum distances. The intersected distances in between the highest and lowest were unseen. Consequently, the method can be enhanced by calculating these to enhance its accuracy of detection. This method can be more enhanced by increasing the number of directions measured and introducing additional composite characters to the method. In addition, the method can be bringing in with a different parameter since the summation of intersections of a letter in every direction to raise the precision of the detection.





VI. CONCLUSION

In this project we proposed algorithm for solving the problem of character recognition. We had given the input in the form of images. The algorithm was trained on the training data that was initially present in the database. We have done preprocessing and segmentation and detect the line. In this way this project will help people to get information about new hotels , place name ,etc. in advance this will result in saving valuable time and will help to travel easily in new places by just clicking the on premises sign. This project is also helpful for getting the details of hospitals and college in case we want to know about facilities and services of this places. The project presents a brief survey of the applications in various fields along with experimentation into few selected fields. The proposed method is extremely efficient to extract all kinds of bimodal images including blur and illumination. The project will act as a good literature survey for researchers starting to work in the field of optical character recognition.

REFERENCES

[1] R. Ji, L.-Y. Duan, J. Chen, S. Yang, T. Huang, H. Yao, *et al.*,"PKUBench: A context rich mobile visual search benchmark," in *Proc. IEEE ICIP*, Jun. 2012, pp. 2545–2548.
[2] V. R. Chandrasekhar, D. M. Chen, S. S. Tsai, N.-M. Cheung, H. Chen, G. Takacs, *et al.*, "The stanford mobile visual search data set," in *Proc. 2nd ACM Conf. Multimedia Syst.*, Feb. 2011, pp. 117–122.
[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. 43–58, Jan. 2007.
[4] B. Girod, V. Chandrasekhar, N.-M. C. David M. Chen, R. Grzeszczuk, Y. Reznik, *et al.*, "Mobile visual search: Linking the virtual and physical worlds," *IEEE Signal Process. Mag.*, vol. 28, no. 4, pp. 61–76, Jul. 2011.
[5] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Proc. ANIPS*, 2007, pp. 545–552.
[6] Y. Zhang, L. Wang, R. Hartley, and H. Li, "Where's the weet-bix?" in *Proc. 8th ACCV*, 2007, pp. 800–810.
[7] 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.
[8] J. Park, G. Lee, E. Kim, J. Lim, S. Kim, H. Yang, *et al.*, "Automatic detection and recognition of Korean text in outdoor signboard images," *Pattern Recognit. Lett.*, vol. 31, no. 12, pp. 1728–1739, Sep. 2010.
[9] 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.
[10] L. Wu, S. Hoi, and N. Yu, "Semantics-preserving bag-of-words models and," *IEEE Trans. Image Process.*, vol. 19, no. 7, pp. 1908–1920, Jul. 2010.
[11] T. Yeh, J. Lee, and T. Darrell, "Fast concurrent object localization and recognition," in *Proc. IEEE Conf. CVPR*, Jun. 2009, pp. 280–287.