2011 10th International Conference on Machine Learning and Applications

Street View Challenge: Identification of Commercial Entities in Street View Imagery

Amir Roshan Zamir, Alexander Darino, Ryan Patrick and Mubarak Shah
Electrical Engineering & Computer Science
University of Central Florida
Orlando, USA
Email: aroshan@cs.ucf.edu

***Abstract*—This paper presents our submission to the *Street View Challenge* of identifying commercial entities in street view imagery. The provided data set of the challenge consists of approximately $129K$ street view images tagged with GPS-coordinates. The problem is to identify different types of businesses visible in these images. Our solution is based on utilizing the textual information. However, the textual content of street view images is challenging in terms of variety and complexity, which limits the success of the approaches that are purely based on processing the content. Therefore, we use a method which leverages both the textual content of the images and business listings, in order to accomplish the identification task successfully. The robustness of our method is due to the fact that the information obtained from the different resources is cross-validated leading to significant improvements compared to the baselines. The experiments show approximately 70% of success rate on the defined problem.**

***Keywords*-Street View Challenge; Commercial Entity; Store Front; Street View;**

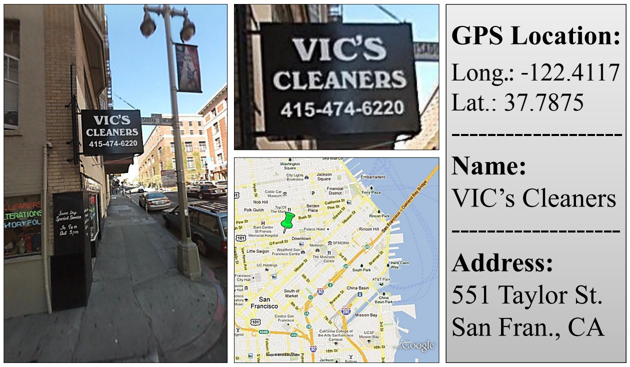


Figure 1. An example of identifying a business in street view imagery. The left and right portions show the input image and the results of applying our method respectively. The middle portion shows the sign of the business which contains the information used for the identification task along with the actual location of the business in San Francisco area.

I. Introduction

Street View Services have gained considerable popularity among internet users over the past few years. The service typically consists of 360° views of the streets at every few tens of meters. Currently, the major use of the Street View service is to provide the user with a virtual view of the streets. However, several other useful applications can be defined for such valuable resource of visual data, for example [6] and [8]. One such application would provide the user with a higher level of knowledge such as exact address and ratings\reviews about businesses in the images. Such information can be used for aligning the street view images with other available geographical resources that includes geo-tagged businesses as well. However, extracting such information from street view images in an automatic manner requires a system which is capable of coping with several complications such as unconstrained appearance of businesses in images, occlusion, low-quality of input data, and inaccuracies in the utilized resources. The problem defined in the challenge is to identify commercial entities visible in the provided set of street view images.

In order to perform this task, we use a method which is based on processing the textual information of the input images. Several efforts to recognize the text in natural scene have been made to date[3], [1], [2]. However, utilizing the current methods of deciphering text to recognize the content of natural images has achieved limited success. In this paper we use additional available resources in order to perform the recognition task properly and efficiently. As the input images are associated with GPS-coordinates, we utilize a detailed listing of businesses, such as the Yellow Pages[9], in order to generate a list of potential businesses which may be visible in the street view image. Then the textual content of the image is extracted and the result is cross validated with the list of potential businesses we generated. The cross-validation based framework makes our approach robust to inaccuracies in the results of each step. The procedure of processing the content includes a text detection step as well as recognizing each character individually. Therefore, the results will be the potential words visible in the image. These words will then be matched to the business listings using Levenshtein distance. The business which achieves the lowest distance to the potential string, i.e. the result of content processing, will be identified as the visible business in the image. Our experiments using the aforementioned framework achieves a detection rate of 70% in the street view data set.

978-0-7695-4607-0/11 $26.00 © 2011 IEEE
DOI 10.1109/ICMLA.2011.181

IEEE computer society

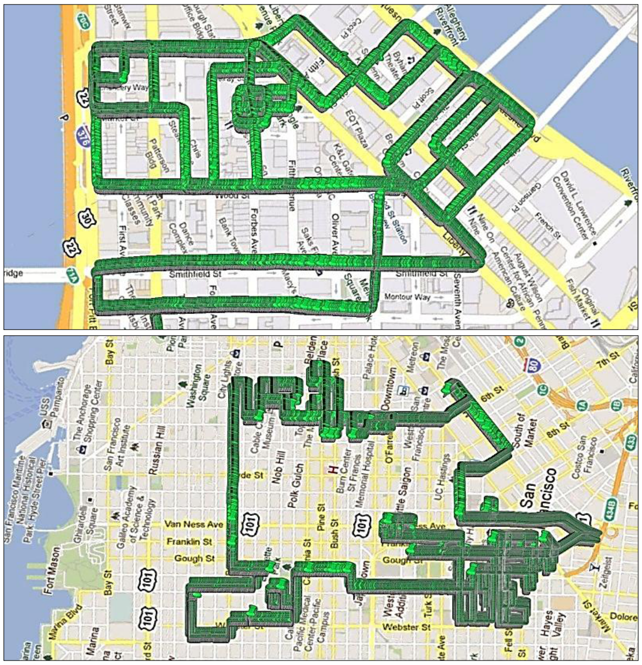


Figure 2. Upper and lower maps show the covered area in the provided data set for the cities of Pittsburgh and San Francisco, respectively. Each green mark represents one street view place mark.

II. Challenge Data set

The provided data set from the challenge consists of approximately $129K$ street view images collected in Pittsburgh, PA and San Francisco, CA. The images were collected using a camera which was fixed on top of a car. The data set has 4 images each showing approximately $90°$ of the view for each place mark. The distance between consecutive place marks is approximately 1 meter. Figure 2 shows the covered area in the provided data set for the aforementioned cities. Each green mark in figure 2 represents one street view place mark. Since the data set covers a large part of two cities in different geographical regions, the businesses visible in the images possess a wide range of appearance features.

III. Identifying Commercial Entities

Figure 3 shows the pipeline of the method. Each step will be explained in more detail in the rest of this section:

A. Using Commercial Listings

As explained in the preceding section, each street view image is tagged with GPS-coordinates. We use the GPS-coordinate of each image to extract a list of nearby commercial entities. For this purpose, we query the business listing [5], [9] based on the GPS-coordinates and collect the information for the businesses within a specified distance. The list is expected to include the business seen in the selected image, and have some textual information in common with it. We use the generated list to narrow down our search and compare it to the the results obtained from processing the content of the image as explained in the following subsection.

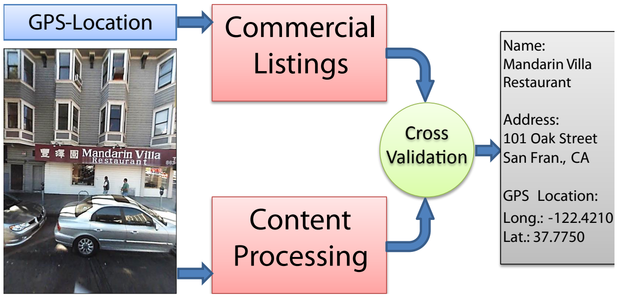


Figure 3. The pipeline of our method. The framework is based on cross-validating the results from listings of businesses and processing the textual content of the image. The input image and the information of the identified business are shown in the left and right part of the figure respectively.

B. Processing Image Content

In order to process the content of the image, we apply a text detection algorithm to the input image [10]. The output of the text detector is the potential sections of the image which may contain words and characters [10]. Then the potential section is resized and a collection of Gabor filters is applied to it for feature extraction, as proposed in [3]. The output of the filters for each potential section form the feature vector for that particular patch after concatenation and reduction in number of dimensions. For the training step, we apply the same process to a list of 52 synthetically generated images of English lowercase and upper case letters. The Gabor features extracted from the synthetic images are our reference features since they are labeled with the letter they belong to. Once the Gabor features are extracted from a region in the input images, we use a nearest neighbor classifier in order to find the most similar letter in the alphabet to the potential letter shown in the image region. This process is repeated for all the regions identified by the text detector. Therefore, we will have a potential match for each potential letter shown in the input image. However, up to this point of the framework, processing has been done purely based on the content of the image, and no external resource has been utilized yet. Therefore, the recognition results of this step are expected to be inaccurate.

In order to identify the right business from the list generated using business directories, we find the Levenshtein distance between the potential string computed in content processing and all the words in the list. Levenshtein distance is a measure of similarity between two strings of characters. Therefore, we expect the correct business to be the one which has the lowest distance to the potential string computed by processing the content. Note that we are not expecting either the listings or the results of content

Figure 4. Sample query images from the street view data set.

processing to be free of error, however we are expecting the right business to be the most similar one after finding the minimum Levenshtein distance. The business which achieves the lowest distance will be considered to be the matching business to that particular street view image.

IV. Experiments

We have evaluated our method on the provided data set of approximately $129K$ street view images. Regarding the large size of the data set, we used a random subset of $1/3$ of the images as the evaluation set. Figure 4 shows a few examples of the query images in the selected set. The ground truth for the sampled images was extracted manually. Since in this paper we are primarily focused on improving the recognition aspect of the framework by using the information in a cross-validated manner, we adjusted the results of text detection manually if it missed the right region of the image. However, we used the same text detection results for all the baselines that we compare our method to.

Figure 5 shows several examples of the application of our method to the images in the provided data set. For each example, the left portion of the figure shows the input image. The middle portion shows the section of the image which contains the textual information and the map showing the actual location of the business. The right part shows the results of identifying the business, as well as additional information as bi-products of the recognition task, such as address and exact GPS-coordinates. The additional information for each business in obtained from the business listings.

Table 1 shows the quantitative results of our framework, along with a comparison to a few other applicable methods. The OCR used is Ocrad-GNU [4], which achieves an acceptable performance in document processing. The poor performance of the OCR software shows the complexity of text recognition in natural scenes compared to document processing. Another method listed in the table 1 is proposed by Chen et al. [3]. They propose a method for detecting and recognizing the text in natural images. The detection rate of applying our method to the provided data set is 59%, which is significantly better than the baselines.

Method	Accuracy
OCR [4]	0%
Chen et al. [3]	4%
Our Algorithm	59%
Our Algorithm - Improved Quality	**70%**

Table I
Qualitative comparison of the results in terms of detection rate.

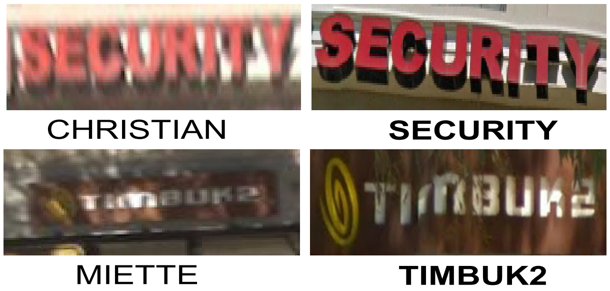
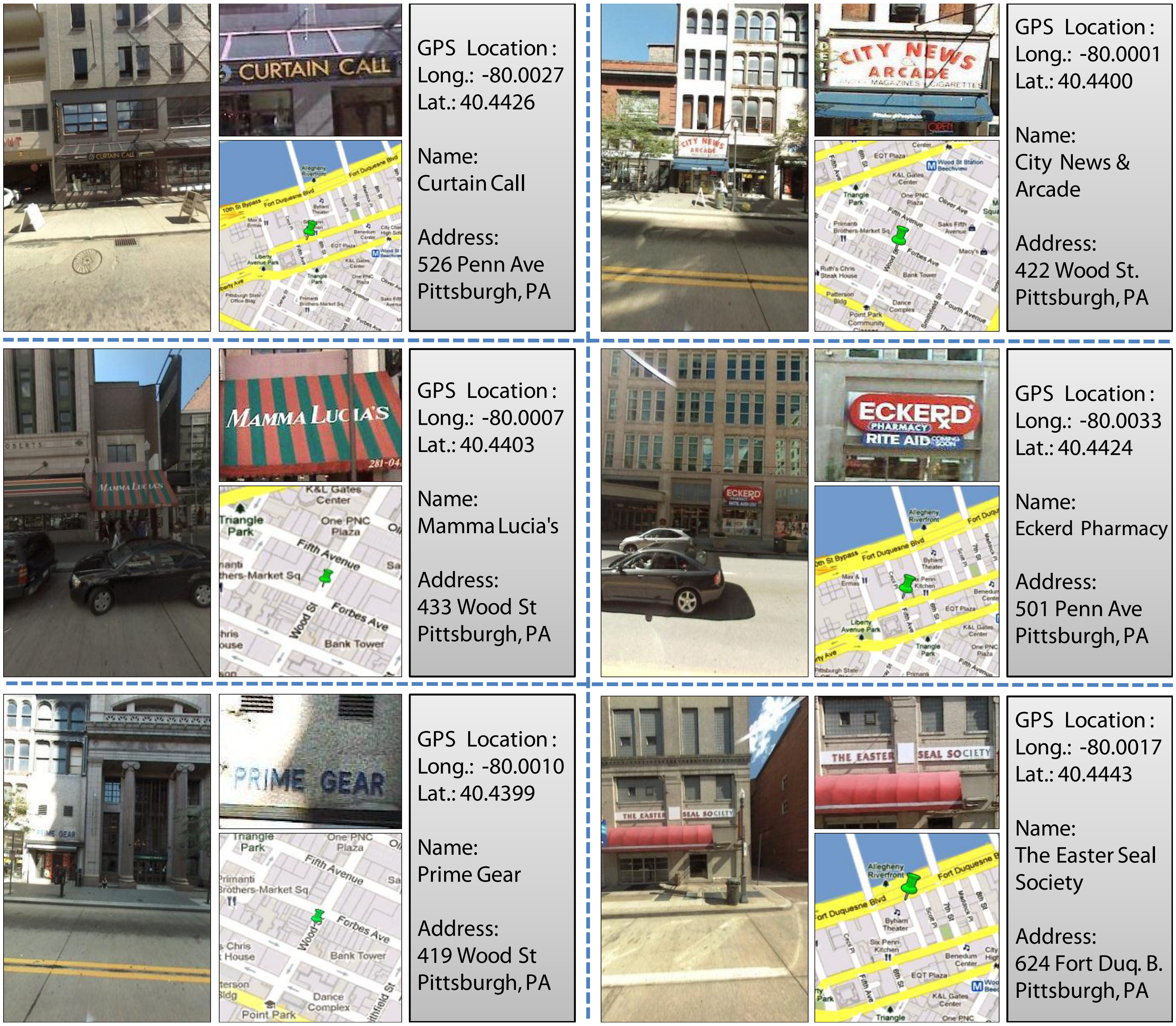


Figure 6. Examples of failure cases which turned into success upon using a higher quality input image. The recognized text is written below each image.

We identified the poor quality of some of the input images as the reason of failure of our method in many cases. This was because the input images were over-exposed, blurred or too dim for text recognition purpose. Therefore, we performed another experiment by collecting a set of input images with higher quality for the cases where our algorithm failed on the provided data set. The higher quality images are also street view images but they were more recently collected. We observed about 26% of the failure cases turned into success upon using a higher quality image. This implies, that we can further improve the detection rate of our method essentially by using input images with higher quality. Examples of such cases are shown in figure 6. As demonstrated on table 1, the performance of our method on the improved set is 70%.

V. Conclusion

In this paper we used a method based on text recognition to solve the problem of *Street View Challenge* to identifying business entities in street view images. Since the general recognition of text in natural images achieves limited success using the current methods, we apply a method which leverages both the content of the image and additional resources, such as business directories, in order to increase the accuracy of detection. Our method achieved a detection rate of up to 70% on the Challenge data set.



GPS Location:
Long.: -80.0027
Lat.: 40.4426

Name:
Curtain Call

Address:
526 Penn Ave
Pittsburgh, PA

GPS Location:
Long.: -80.0001
Lat.: 40.4400

Name:
City News &
Arcade

Address:
422 Wood St.
Pittsburgh, PA

GPS Location:
Long.: -80.0007
Lat.: 40.4403

Name:
Mamma Lucia's

Address:
433 Wood St
Pittsburgh, PA

GPS Location:
Long.: -80.0033
Lat.: 40.4424

Name:
Eckerd Pharmacy

Address:
501 Penn Ave
Pittsburgh, PA

GPS Location:
Long.: -80.0010
Lat.: 40.4399

Name:
Prime Gear

Address:
419 Wood St
Pittsburgh, PA

GPS Location:
Long.: -80.0017
Lat.: 40.4443

Name:
The Easter Seal
Society

Address:
624 Fort Duq. B.
Pittsburgh, PA

Figure 5. Sample results of our framework. Each part shows the input image, the sign, the actual location on the map, and additional information about the detected commercial entity.

REFERENCES

[1] Kim, K.C.; Byun, H.R.; Song, Y.J.; Choi, Y.W.; Chi, S.Y.; Kim, K.K.; Chung, Y.K.; *Scene text extraction in natural scene images using hierarchical feature combining and verification*, ICPR, 2004.

[2] J. J. Weinman, E. Learned-Miller and A. R. Hanson, *Scene Text Recognition Using Similarity and a Lexicon with Sparse Belief Propagation*, TPAMI, 2009.

[3] Chen, X. Yang, J. Zhang, J. Waibel, A., *Automatic Detection and Recognition of Signs From Natural Scenes*, IEEE TRANSACTIONS ON IMAGE PROCESSING2004, VOL 13; PART 1, pages 87-99.

[4] GNU-Orcad OCR, *http://www.gnu.org/s/ocrad/*.

[5] Yelp, *http://www.yelp.com/*.

[6] Schindler, G.; Brown, M.; Szeliski, R., *City-scale location recognition*, CVPR, 2011.

[7] Xilin Chen; Jie Yang; Jing Zhang; Waibel, A, *Automatic detection and recognition of signs from natural scenes*, Image Processing, IEEE Transactions on, 2004.

[8] Amir Roshan Zamir, Mubarak Shah, *Accurate Image Localization Based on Google Maps Street View*, ECCV, 2010.

[9] YellowPages, *http://www.yellowpages.com/*.

[10] Epshtein, B.; Ofek, E.; Wexler, Y.; , *Detecting text in natural scenes with stroke width transform* , CVPR, 2011.