Visual Business Recognition - A Multimodal Approach

Amir Roshan Zamir, Afshin Dehghan, and Mubarak Shah*
Center for Research in Computer Vision, UCF, Orlando, FL 32816, USA
aroshan@cs.ucf.edu, adehghan@cs.ucf.edu, shah@eecs.ucf.edu
http://crcv.ucf.edu/projects/Business-Recognition/

ABSTRACT

In this paper we investigate a new problem called *visual business recognition*. Automatic identification of businesses in images is an interesting task with plenty of potential applications especially for mobile device users. We propose a multimodal approach which incorporates business directories, textual information, and web images in a unified framework. We assume the query image is associated with a coarse location tag and utilize business directories for extracting an over complete list of nearby businesses which may be visible in the image. We use the name of nearby businesses as search keywords in order to automatically collect a set of relevant images from the web and perform image matching between them and the query. Additionally, we employ a text processing method customized for business recognition which is assisted by nearby business names; we fuse the information acquired from image matching and text processing in a probabilistic framework to recognize the businesses. We tested the proposed algorithm on a challenging set of user-uploaded and street view images with promising results for this new application.

Categories and Subject Descriptors: I.4 [Image Processing and Computer Vision]: Applications

Keywords: Business Recognition; Storefront; Location Based Service; Business Review; Map; Scene Text; Multi-hypotheses; Yelp.

1. INTRODUCTION

A business recognition system can provide smartphone or wearable computer users with extensive information about a particular business of interest in an automatic and convenient fashion. Such system can be used for enhancing the user experience in surfing maps or location-aware image understanding as well.

The existing methods for providing a smartphone user with information about a specific business primarily use *non-visual sensors* such as embedded GPS, digital compass, and gyroscope [1, 2]. These approaches are often based on matching the sensor data to a reference set; e.g. matching the location received from the GPS-chip along with the compass direction to a geo-tagged business directory such as Yelp. Such methods do not benefit from image content and typically achieve limited success as they generally require very precise sensors and accurately tagged business datasets. Additionally, other potential approaches such as existing scene text processing methods [3, 5, 6] cannot sufficiently cope with the complex appearance of text in business signs and logos.

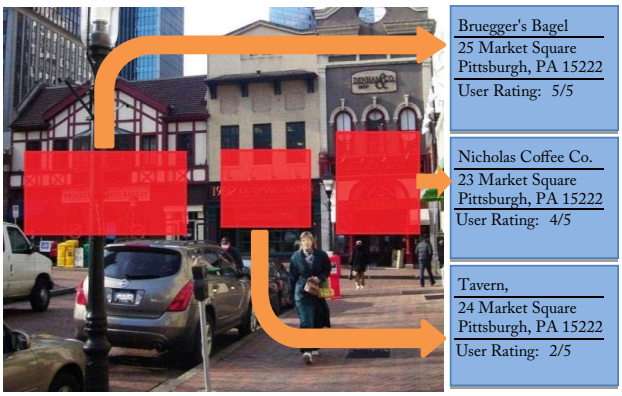

Figure 1: A business recognition system can automatically identify businesses in an image and provide additional relevant information such as reviews and similar nearby businesses.

Therefore, we propose a method which utilizes multimedia information obtained from both visual content, such as storefront appearance and text, and non-visual information, such as GPS and business directories. We show that our method achieves a notable rate of success by employing a multimodal approach and is capable of finding multiple businesses in an image and their spatial location.

2. FRAMEWORK OVERVIEW

The block diagram of the proposed method is shown in fig. 2. The images captured using smartphones are usually associated with a coarse geo-tag which typically comes from the inbuilt GPS-chip, cell tower signal or WPS. We use this approximate location for generating a list of nearby businesses by querying business directories.

To utilize the textual information (subsection 2.1), we perform text detection on the query image; then, we apply a multi-hypotheses text recognition approach assisted by the business lexicon which yields a PDF specifying how well a detected word in the query image matches the nearby businesses. Since the business in the image may include several words, we combine the PDFs of matching businesses to each word through marginalization to have a single PDF representing the textual information in the whole query image.

In order to leverage the images on the web in business recognition, we use the list of nearby business names as search keywords and collect a set of images from the web for each one. The query image is expected to share some similarity with the web images of the business which is visible in it. Therefore, we match the query image to the collected web images in order to identify the similar ones (subsection 2.2). This process yields a PDF which represents how well the web images of each nearby business match the query image. Lastly, we combine the two PDFs acquired from text pro-

*The authors would like to thank Oliver Nina.

MM'13, October 21–25, 2013, Barcelona, Spain.

http://dx.doi.org/10.1145/2502081.2502174.

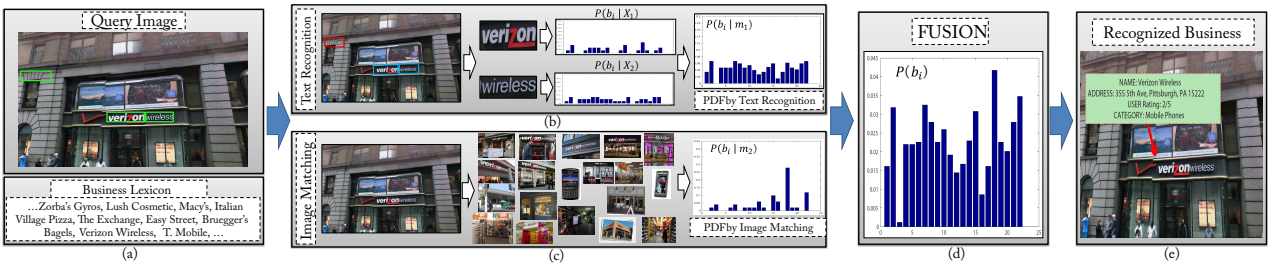

Figure 2: The block diagram of our method. (a) shows the query image, detected text and business lexicon. (b) illustrates the process of computing the PDFs of different words and marginalizing them into one PDF. (c) shows the query image, a subset of web images ordered based on how well they match the query, and the resulting PDF from image matching. (d) demonstrates the PDF obtained from the fusion process. (e) shows the business which achieves the highest probability after fusion, as the recognized business.

cessing and image matching in a probabilistic late fusion step to compute a PDF which utilizes both modalities (subsection 2.3).

Generating the Business Lexicon: We use the APIs of Yellow pages and Yelp to automatically retrieve and aggregate the nearby businesses within the distance of 150 meters to the approximate location. Regarding the inaccuracies in the business directories and the GPS-tag of the query, we set the radius to a large value to ensure the visible businesses in the query are among the retrieved results. $B = \{b_i | 1 \leq i \leq n_B\}$ represents the set of retrieved businesses where n_B denotes the number of nearby business. A business name may include more than one word, so $W = \{w_{i,j} | 1 \leq i \leq n_B, 1 \leq j \leq n_w(i)\}$ is the set of words in the name of all nearby businesses. $w_{i,j}$ represents the j^{th} word of i^{th} business's name, and $n_w(i)$ denotes the number of words the name of i^{th} business includes.

2.1 Business Recognition Using Text

Business recognition using textual information is inherently similar to the problem of text recognition in natural scene. However, the goal of business recognition is to establish a relationship between the reference businesses and the text in the query image and not necessarily recognizing it. Such relationship can be probabilistic or fuzzy, while text recognition aims at recognizing the text deterministically. Additionally, scene text recognition does not address other problems specific to business recognition such as combining the information obtained from different query words in order to perform the recognition of a single business. We employ the text processing method described in the rest of this section which is specifically customized for the task of business recognition and addresses the aforementioned issues. Additionally, it makes representing the matching results in a probabilistic manner feasible, as such representation is required in our fusion process.

Multi-hypotheses Character Recognition: We use Stroke Width Transform (SWT) [8] as our text detection method which identifies the regions of the image which might contain a word and each character therein. We use Gabor features for performing text recognition [3] on each character patch. In our training step, we generate 62 synthetic character patches comprised of lower and upper case English alphabet along with single digit numbers using the font `Arial`. Additionally, we compute six variations for each character using four consecutive image dilation and two erosions as we observed that the business signs in natural scenes tend to significantly vary in the width of characters compared the standard fonts. we apply a bank of 108 Gabor filters comprised of $n = 6$ frequencies and $m = 18$ scales to each synthetic character. Each character is then divided into 9 sub patches using a 3 by 3 grid. The Gabor feature of each sub patch is defined as the mean of Gabor features of the pixels therein. Therefore, each character is represented by a 972 dimensional vector which is reduced to 50 dimensional using PCA. The feature vectors of all 62 characters and their erosion-dilation variations form our reference set of character features.

During the test step, the same 108 Gabor filters are applied to a character patch returned by text detection and the size of the feature vector is reduced to 50 using the mapping found by PCA during the training. Then we use a k-nearest neighbor classifier to find the most similar k reference characters to the query patch. In other words, instead of assigning one character to the query patch, we nominate k characters as the possible matches. We employ this approach as the right character may not necessarily be the first match, while it usually appears among the top few matches. This is shown for the sample query word "**verizon**" in fig. 3 (b).

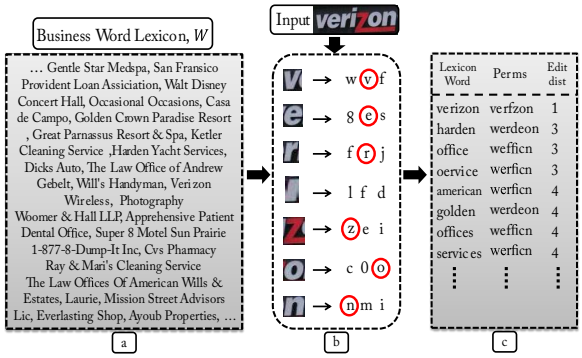

Figure 3: Illustration of the process of multi-hypotheses matching (eq. 1). (a): Business Lexicon. (b): the query word and nominated candidates for each query patch. The correct candidates are marked with red circles. (c): best matching permutations to each business word and their respective edit distance.

We denote a feasible permutation of the candidates for a query word by $X = \{\chi_a^1, \chi_b^2, \chi_c^3, \ldots\}$, which means the a^{th} candidate for the 1^{st} query patch, the b^{th} candidate for the 2^{nd} query patch, and so on are selected. Hence, each query word possesses a large number of feasible permutations of its character candidates; eight feasible permutations for a sample query are shown in figure 3 (c).

Matching Character Permutations to Business Lexicon: We solve the following optimization problem to identify the best permutation which matches a particular business word in the lexicon:

$$\hat{X}_{i,j} = \underset{X}{\operatorname{argmin}} ||X - w_{i,j}||, \tag{1}$$

where $\hat{X}_{i,j}$ represents the permutation which best matches the business word $w_{i,j}$. $||.||$ represents Levenshtein distance between two strings. We solve eq. 1 once for every word in the business words lexicon W in order to find the best matching permutation to each. This process is illustrated in fig. 3 for a sample case. Eight permutations, their respective matching words in the business lexicon, and the edit distance between them are shown in (c).

Bear in mind that the name of one business may include more than one word. Thus, we solve the following equation to find the best matching business word to the query for each nearby business:

$$\zeta(b_i) = \min_j ||\hat{X}_{i,j} - w_{i,j}||, \tag{2}$$

where b_i represents the ith businesses among the nearby businesses B. $\zeta(b_i)$ is the Levenshtein distance between the query word and

Figure 4: Sample web images for two businesses. The red margin marks the positive examples. Green, yellow and blue markers denote the keywords "business name","business name+city"and"business name+storefront".

the best matching word in the name of business b_i. Therefore, ζ can be interpreted as a distance function which represents how well business b_i matches the query word.

We would like to have a PDF which specifies a probability for each of the nearby businesses matching the query word represented by X. Therefore, the distances function $\zeta(b_i)$ is converted to a PDF using the following equation:

$$p(b_i|X) = \frac{sig(\zeta(b_i))}{\sum_i sig(\zeta(b_i))}, \quad (3)$$

where $p(b_i|X)$ is the probability of the business b_i to match the given query word X. sig is the sigmoid function with the standard form $sig(x) = \frac{1}{1+e^{-\tau x}}$, where τ is a constant which we set to -0.5 in our experiments. Therefore, a large edit distance corresponds to a small probability and vice versa.

Utilizing multiple words for recognizing a Business: The probability distribution function $p(b_i|X)$ acquired from eq. 3 specifies how well the nearby businesses match one query word. However, the business sign in the query image may include more than one word. Therefore, we need to associate the query words pertaining to one business in order to utilize all of them for recognizing the respective business. Usually the words which belong to one business in the image are spatially close and have similar appearance features. For instance, the words "**verizon**" and "**wireless**" in figure 2 (a) have similar colors and are located next to each other. Therefore, for each bounding box acquired from the text detector, we form a feature vector by concatenating its RGB color histogram with (x, y) spatial location of its center. Then, we perform mean shift clustering on the feature vectors of all the bounding boxes to associate the words which belong to one business. The number of resulting clusters is the number of businesses in the query image, and the elements in each cluster are the bounding boxes associated together. A sample case is shown in figure 2 (b) where the bounding boxes shown in the same color are associated together.

To leverage the associated query words in business recognition, we combine the PDFs each one yields through marginalization:

$$p_t(b_i) = \sum_{j=1}^{\alpha} p(b_i|X_j)p(X_j), \quad (4)$$

where $p(b_i|X_j)$ is the PDF obtained from eq. 3 for the query word X_j, and α is the number of associated query words. $p(X_j)$ is the probability of looking at the j^{th} query word for recognizing its respective business. We treat all the query words of one business sign equally by assigning equal chance to all: $p(X_j) = 1/\alpha$.

$p_t(b_i)$ in eq. 4 specifies the probability of each nearby business being visible in the query image based on the entire textual information in the query. In order to avoid confusing the PDFs obtained using text processing, image matching and fusion, we define $M = \{m_1, m_2\}$ as the set of approaches to business recognition which we employ. m_1 and m_2 represent text recognition and image matching respectively. Therefore, $p(b_i|m_1)$ represents the PDF obtained by employing text recognition which is equal to $p_t(b_i)$ of eq. 4. Fig. 2 (b) illustrates the described process for a sample query.

2.2 Business Recognition by Image Matching

Nowadays, for most of the businesses in urban area a number of images which show the storefront can be found on the web. Such images are typically uploaded by customers, business owners, or business directories for both franchise and non-franchise businesses.

In order to find the web images which pertain to a particular business, we generate four search keywords for each nearby business as: "business name", "business name+city" and "business name+storefront". We use the keywords to search for images on the web and download the retrieved ones using Ajax-based web image crawling. We save about 10 images per keywords which results in 40 images for each nearby business. We view the set of downloaded images as a reference dataset that each image therein is associated with a nearby business.

We employ bag of visual words (BoVW) model for matching the query image to the set of web images. We extract SIFT features from the web images and the query and compute their histogram of visual words using a vocabulary with 2000 words. The vocabulary is pre-computed on a set of 10000 random images. We employe tf-idf weighting scheme which reduces the contribution of less discriminative visual words [4]. We find the most similar web image of a business to the query using:

$$\psi(b_i) = \min_j |h_q - h_{i,j}|, \quad (5)$$

where h_q and $|.|$ represent the $BoVW$ histogram of the query, and L_2 distance respectively. $h_{i,j}$ represents the histogram of the j^{th} web image of the i^{th} business. Eq. 5 identifies the most similar image to the query for each nearby business. Therefore, the distance function $\psi(b_i)$ specifies how well the nearby business, b_i, matches the query based on the web images.

Using a method similar to the eq. 3 which was intended to convert edit distances to probability values, we convert the image matching distance function $\psi(b_i)$ to a PDF using, $p(b_i|m_2) = \frac{sig(\psi(b_i))}{\sum_i sig(\psi(b_i))}$, where $p(b_i|m_2)$ represents the probability of recognizing the business b_i in the query given the employed approach is image matching.

The procedure of downloading web images and computing their BoVW representation is relatively time consuming. However, since all the businesses in the broad area of interest, e.g. a city, are known, the web images can be downloaded and processed in an offline manner. That way, image matching between query and the web images of its nearby businesses can be done almost instantaneously.

2.3 Fusion of image matching and textual info

The purpose of the fusion step is to unify the information obtained from the two methods of text recognition and image matching to perform a more robust business recognition. Theoretically, the law of total probability is utilized for finding the probability of one event when it coincides with a random variable, so we employ it in fusing the PDFs acquired from text recognition and image matching. In our problem, the event is a nearby business, b_i, and the coinciding variable is m_i:

$$p(b_i) = p(b_i|m_1).P(m_1) + p(b_i|m_2).P(m_2) \quad (6)$$

where $P(m_1)$ and $P(m_2)$ are the probability of employing text recognition and image matching respectively. We define these two values using a training set of 50 query images. The training set consists of queries for which *only one* of the two methods worked successfully. We define $P(m_1)$ and $P(m_2)$ as:

$$P(m_1) = \frac{n_t}{n_t + n_i}, P(m_2) = \frac{n_i}{n_t + n_i} \quad (7)$$

where n_t is the number of images in the training set for which only text recognition identified the business. Similarly, n_i is the number

(a) (b) (c)

Figure 5: Business recognition results. (a): the query image, detected text and recognized words. (b): the PDFs found by text recognition, image matching and fusion along with the best matching web images. (c) the recognized businesses.

of images for which only image matching was successful. $n_t + n_i$ is the total number of images in the training set, i.e. 50.

The fusion process may not make a notable difference when both or none of the methods correctly recognize the business individually, regardless of the values of $p(b_i|m_1)$ and $p(b_i|m_2)$. However, when only one of the methods identifies the right business, proper values of $p(b_i|m_1)$ and $p(b_i|m_2)$ may results in successful overall recognition at the end. This is the reason our training set includes the query images for which only one of the methods worked. In other words, computing the values of $P(m_1)$ and $P(m_2)$ using the described method maximizes the chance of successful overall recognition for the cases where one of the methods fails.

If the word association method, explained in subsection 2.1, finds more than one business in the query, i.e. more than one cluster in mean-shift clustering, the text recognition and fusion process are repeated using the query words of each cluster in order to recognize multiple business. However, in case the best matching business in $p(b_i)$ has a low probability, typically < 0.10, we disregard it as it most likely corresponds to a false positive from the text detector.

3. EXPERIMENTS AND DISCUSSION

No dataset is currently available for evaluating the proposed framework as visual business recognition has not been studied to date. Therefore, we collected a data set of 1042 GPS-tagged images comprised of 642 user uploaded photos from Panoramio, Flickr and Picasa and 400 street view images for the cities of San Francisco, CA and Pittsburgh, PA. We manually filtered the images which do not show a business or have an incorrect GPS-tag. Each image may include up to four business. In case few businesses were retrieved by querying business directories for a particular query image, we added random businesses to make sure at least 20 businesses and 70 words existed in the lexicon to ensure each test is challenging enough. Fig. 5 shows sample business recognition results.[1]

Business recognition accuracy is defined as the number of correctly recognized businesses divided by the total number of businesses in the test set. We evaluated the proposed text processing and image matching methods on the test set individually to examine their performance in the single modal fashion; this resulted in the accuracy of 69% and 41% for text recognition and image matching respectively. However, when the two modalities were combined using the described fusion process, the accuracy increased to **75**% which signifies the effectiveness of our multimodal approach.

No other framework for visual business recognition has been proposed which we can use as a baseline. However, we compared the performance of our text processing method, which is customized for business recognition, with the state of the art scene text recognition algorithms in recognizing business *words*. The table in fig. 6 compares the performance of our approach to four baselines. The performance measure is the number of correctly recognized

[1] more results available at: http://crcv.ucf.edu/projects/Business-Recognition/

Method	Acc.
GNU-Ocrad OCR [9]	4%
Chen et al. [3]	7%
ABBYY OCR [7]	34%
Wang et al. [6]	42%
Proposed Method	**54%**

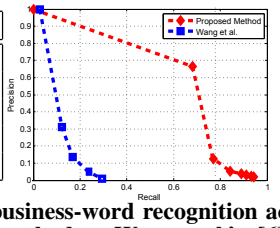

Figure 6: Left: Comparison of business-word recognition accuracies. Right: PR curve of our method vs. Wang et al.'s. [6]

words divided by the total number of words in our test images. Wang et al.'s [6] method which employs a pictorial structure for detection and recognition achieves the accuracy of 42% using the code of the authors with best parameters. Chen et al.'s [3] is another scene text recognition method for recognition of signs. Additionally, we compared our results with two OCR methods [7, 9] which have achieved notable success in document processing. In order to have a fair evaluation of text *recognition*, we manually adjusted the text detection results for our methods and the baselines, in case a word is completely missed in detection. For the baseline methods which do not need a separate text detection step [6, 7], we limited their search space to the adjusted text detection results to decrease their false positives. Fig. 6 (b) shows the precision-recall curves of our method (red) vs. Wang et al.'s [6] (blue).

Unlike the majority of existing scene text recognition methods [3, 6] which employ a heavy training process, e.g. by using a variety fonts and deformations, we used only one font and few deformations in our training. On the other hand, we leverage a more complex test step utilizing our multi-hypotheses character recognition approach. This is one of the reasons behind our superior performance in recognizing business words as they typically show a great deal of deformation and complexity which can not be effectively learnt in a training step. However, our multi hypotheses test step maximizes the use of business lexicon to alleviate this issue.

We observed that the majority of the failure cases of our method are due extreme deformation of characters and lack of a relevant image on the web for some businesses. Upon availability of an optimized parallel implementation of the framework, business recognition on a query can be done in no more than 3 seconds on average.

4. CONCLUSION

In this paper, we proposed a multimodal approach to a new application called *visual business recognition*. Our framework leverages textual information, web images and business directories and combines the results of each using a probabilistic late fusion process. We proposed a multi hypotheses approach to processing the text in images which is specifically customized for business recognition. The experiments showed the effectiveness of the proposed multimodal method for this new application.

5. REFERENCES

[1] Google Goggles, *https://sites.google.com/a/pressatgoogle.com/dec09searchevent/all-about-goggles*

[2] Nokia City Lens, *http://betalabs.nokia.com/trials/nokia-city-lens-for-windows-phone,http://tinyurl.com/nokia-city-lens*

[3] Chen et al., *Automatic detection and recognition of signs from natural scenes*. In: IEEE Trans. Image Processing, 2004.

[4] J. Philbin et al., *Object Retrieval with Large Vocabularies and Fast Spatial Matching*. In: CVPR, 2007.

[5] Amir Roshan Zamir, Alexander Darino, and Mubarak Shah, *Street view challenge: Identification of commercial entities in street view imagery*. In: ICMLA, 2011.

[6] K. Wang, B. Babenko and S. Belongie, *End-to-End Scene text recognition*. In: ICCV, 2011.

[7] ABBYY, *http://www.abbyy.com/*.

[8] B. Epshtein et al., *Detecting text in natural scenes with stroke width transform*. In: CVPR, 2011.

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