ST-Sem: A Multimodal Method for Points-of-Interest Classification Using Street-Level Imagery

Shahin Sharifi Noorian(✉), Achilleas Psyllidis, and Alessandro Bozzon

Delft University of Technology, Delft, The Netherlands
{s.sharifinoorian,a.psyllidis,a.bozzon}@tudelft.nl

Abstract. Street-level imagery contains a variety of visual information about the facades of Points of Interest (POIs). In addition to general morphological features, signs on the facades of, primarily, business-related POIs could be a valuable source of information about the type and identity of a POI. Recent advancements in computer vision could leverage visual information from street-level imagery, and contribute to the classification of POIs. However, there is currently a gap in existing literature regarding the use of visual labels contained in street-level imagery, where their value as indicators of POI categories is assessed. This paper presents *Scene-Text Semantics* (ST-Sem), a novel method that leverages visual labels (e.g., texts, logos) from street-level imagery as complementary information for the categorization of business-related POIs. Contrary to existing methods that fuse visual and textual information at a feature-level, we propose a late fusion approach that combines visual and textual cues after resolving issues of incorrect digitization and semantic ambiguity of the retrieved textual components. Experiments on two existing and a newly-created datasets show that ST-Sem can outperform visual-only approaches by 80% and related multimodal approaches by 4%.

Keywords: Points of Interest · Street-level imagery · Convolutional Neural Networks · Word embeddings · Semantic similarity

1 Introduction

An increasing amount of new data sources revolve around the concept of Point of Interest (POI) [2,3]. From a computational perspective, POIs represent real-world places as geometric point entities, in which people perform various activities. Online sources, such as geo-enabled social media and mapping applications, provide information about POI attributes (e.g. name, address, accessibility, provided facilities) and their functionality (e.g. opening hours, popularity, land use). A more recent addition to the spectrum of POI-related sources is street-level imagery. Unlike the well-established satellite imagery, this alternative source provides panoramic views of – primarily urban – environments at ground level.

M. Bakaev et al. (Eds.): ICWE 2019, LNCS 11496, pp. 32–46, 2019.
https://doi.org/10.1007/978-3-030-19274-7_3

A wealth of recent studies have used street-level imagery to analyze various aspects of urban environments [1,9,32]. Street-level imagery can be extracted from both proprietary (e.g. Google Street View) and public (e.g. Mapillary) online repositories. The information contained in street-level imagery is essentially visual. Thereby, POIs can be described based on morphological characteristics (e.g. height, color, materials, geometry) of their facades. A common feature which is found on several building facades, especially those of commercial stores and other business-related facilities, is *signs* or *visual labels*. These contain the name, logo, and other related information that help people identify businesses while navigating through physical space. Therefore, they could be a valuable source of information about the type and identity of a POI.

Advancements in computer vision, especially deep Convolutional Neural Networks (CNNs), have been successful in face detection, image segmentation, and scene recognition [22,30,31]. CNNs have also been used in the characterization of urban land uses [26,32], but several open challenges exist. Specific to the problem of POI classification, the performance of systems purely based on visual features is limited by issues of intra-class and inter-class differences in the appearance of store fronts across business categories [14]. Our hypothesis is that by incorporating semantically rich information from visual labels – e.g. the text from storefronts' signs – it is possible to substantially improve classification performance. The analysis and use of visual labels also present several challenges, e.g. their identification in street-level imagery, their transformation into textual information, and their disambiguation with regard to language and meaning.

This paper introduces a novel method, named *Scene-Text Semantics* (`ST-Sem`), that leverages visual labels from street-level imagery as complementary information for the classification of business storefronts. `ST-Sem` comprises three main components: (1) a *scene recognition* module classifies the POI type based on common visual characteristics of storefronts belonging to the same type; (2) a *scene-text semantic recognition* module detects textual information (e.g. labels, signs) in street-level imagery, transcribes it into a bag of words, and measures their semantic similarity against a reference corpus of business type descriptions; and (3) a *class rank module* which combines the prediction scores of the two aforementioned modules, and generates a final score for each candidate class. Contrary to existing approaches [13] that predominantly fuse visual and textual information at the feature level, `ST-Sem` combines visual and textual cues *after* resolving semantic ambiguities and incorrect digitization of the detected textual labels. Thanks to this late-fusion approach, `ST-Sem` can be more easily configured to work on different datasets (e.g. storefronts from different countries) with improved modularity and minimal visual models re-training.

Through extensive experiments, we show that `ST-Sem` is able to outperform state-of-the-art visual-only approaches for POI classification by 80%, and multimodal approaches by 4%. The experiments have been performed on the *Places* and *Con-text* datasets; and on a newly created dataset (that we contribute to the community) containing 1100 images from 22 classes of storefronts located in two countries. In addition to the quantitative comparison, we qualitatively analyze the performance of our method in an number of edge cases, to highlight limitations and suggest future directions of improvement.

The remainder of the paper is organized as follows: Sect. 2 discusses related work. In Sect. 3, we describe the ST-Sem method for POI classification from street-level imagery. Section 4 presents the experimental setup and discusses the obtained results. Finally, Sect. 5 summarizes the conclusions and discusses future lines of research.

2 Related Work

Capturing the semantics of POIs (e.g. bar or museum) is essential for facilitating location-based applications. While some location-based online social network services (e.g. Foursquare, Yelp) allow users to characterize the places they visit, the process is not automated and, instead, requires the users' help. In recent years, a considerable amount of methods are proposed by researchers for assigning semantic labels to unlabeled places [6,29]. Approaches on semantic place labeling mainly focus on user activity data to characterize POIs through their fine-grained visiting patterns.

Street-level imagery could be a useful data source in providing information about POI attributes (e.g. the type of POI) with higher spatial coverage. In recent years, researchers have demonstrated the feasibility of utilizing street-level imagery in mapping urban greenery [16], assessing physical changes in urban area [1], estimating city-level travel patterns [9], or in inferring subjective properties of urban areas such as safety, liveliness, and attractiveness [7]. Other works applied computer vision techniques to Google Street View images for finding morphological characteristics to distinguish European cities [5] or to infer the socioeconomic attributes of neighbourhood in US [8]. However, computer vision methods for POI types classification require large annotated datasets for training, which are currently available only for limited cities and countries.

Deep Convolutional Neural Networks (CNNs) have been successful in various vision-related tasks such as face detection, image segmentation, and scene recognition [22,31]. However, such breakthroughs in visual understanding do not imply that these models are suitable for fine-grained POI classification based on the visual appearance of store fronts from street-level imagery. This is due to the high degree of intra-class and the low degree of inter-class differences in the appearance of store fronts across business categories [14]. Yan et al. [28] take Spatial Context (i.e. nearby places) into account as complementary information to boost the performance of CNN models for classifying business places. Text in scene images, which frequently appears on the shop fronts, road signs, and billboards, usually conveys a large amount of valuable semantic information about the object or the scene in the same image. When it comes to the fine-grained classification of storefronts based on their business type, this textual information plays a crucial role in making more accurate predictions [21]. Most similar to our work, is that of Karaoglu et al. [13]. The latter proposed a multimodal approach that combines visual features and textual information from the imagery data in a single feature space as input for a SVM classifier. Our work differs from the existing literature in that we incorporate multilingual word embeddings trained on a large corpus to measure semantic relatedness between spotted textual information in street-level imagery and the candidate types of storefronts. Then, we propose a late fusion approach to leverage the obtained prediction scores of both modalities, and generate a final score for each candidate class. We compare ST-Sem against the approach of Karaoglu et al. [13] in Sect. 4, showing improved performance.

3 Methodology

This section presents ST-Sem, a novel multimodal approach for improving the fine-grained classification of business storefronts by leveraging visual and textual features in street-level imagery. The architecture of ST-Sem is depicted in Fig. 1. The image processing pipeline mainly consists of three components. First, the *Scene Recognition* module predicts the type of storefront at context level based on the common visual characteristics of each storefront type. Next, the *Scene-text Semantic Recognition* module detects textual data on the image and transcribes it into a bag of words and measures the semantic similarity between the bag of words, which usually represent the type of storefront, and each of the candidate types. Ultimately, the *Class Rank* module generates the final score for each candidate class by using a Linear Bimodal Fusion (LBF) method which combines the prediction scores from first and second modules. In the following paragraphs, we describe each component in detail.

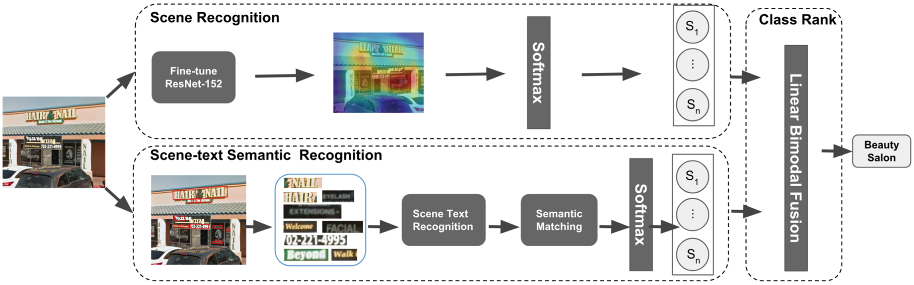


Fig. 1. Overview of the multimodal storefront image classification approach.

3.1 Scene Recognition

The scene recognition component is designed to classify images into one of the several candidate storefront types based on common morphological features. We use deep Convolutional Neural Networks (CNNs) which have been successful in several computer vision tasks [22,31].

Among a number of CNN models introduced for image classification tasks, we have employed the Residual Network (ResNet), which has shown superior performance on ImageNet classification [10].

We use the pre-trained `ResNet152-places365` model provided by [30], which includes the place categories[1]. *Places365* offers 365 classes, including some (e.g. `cliff` or `coral`) that do not qualify as POI type (e.g. `restaurant`) and, therefore, are not relevant in our setting. Without losing generality, we select 22 place types as our candidate class labels. Then, we fine-tune the pre-trained `ResNet152-places365` with a 22-way softmax classifier on the subset of the *Places* dataset. The last fully connected layer of the pre-trained network was removed and replaced by a new fully connected layer with 22 nodes to solve the storefront type classification problem. Then, the weights of the added fully connected layers were randomly generated from a Gaussian distribution with zero mean and standard deviation of 0.01.

3.2 Scene-Text Semantic Recognition

The scene-text extraction module is composed of three sub components. In the following paragraphs, we describe each sub-component in detail.

Scene-Text Detection. This sub-module aims at localizing and cropping text in images in the form of word boxes. Scene-text detection is challenging because scene texts have different sizes, width-height aspect ratios, font styles, lighting, perspective distortion, and orientation. In this work, we incorporate one of the state-of-the-art methods, named TextBoxes++ [17], which is an end-to-end trainable fast scene-text detector. The reason for choosing TextBoxes++ is that it outperforms state-of-the-art methods in terms of text localization accuracy and runtime issues of the *IC15* dataset [15] from Challenge 4 of the ICDAR 2015 Robust Reading Competition[2]. The *IC15* dataset is composed of 500 test images containing incidental scene text captured by Google Glass. Therefore, it is a good benchmark dataset to evaluate the required scene-text detector for storefront type classification. We adopt the pre-trained model parameters which are provided by the authors.

Scene-Text Recognition. The task of this sub-module is to transcribe cropped word images into machine-readable character sequences. However, it is considerably difficult to accurately recognize scene texts on street-level imagery because of the various shapes and distorted patterns of irregular texts. In order to tackle this problem, we adopt a multi-object rectified attention network (MORAN), proposed by [19]. MORAN consists of a multi-object rectification network (MORN) that rectifies images and an attention-based sequence recognition network (ASRN) that reads the text. In terms of reading rotated, scaled and stretched characters in different scene texts, this approach outperforms state-of-the-art methods on several standard text recognition benchmarks [19],

[1] https://github.com/CSAILVision/places365/.

[2] http://rrc.cvc.uab.es/?ch=4.

including the SVT-Perspective dataset [24] which contains 645 cropped images from Google Street View. In training the Scene-text recognition on the MJSynth dataset [11], which is dedicated for Natural Scene Text Recognition, we set the batch size to 64, the learning rate to 0.01 as suggested by the author. The model is trained for 10 epochs.

Semantic Matching. The semantic matching approach follows the assumption that textual information on the storefront indicates the type of business place. Given this assumption, the goal of the semantic matching module is to predict the type of storefront based on the semantic distance between the words extracted from the image and the standard name of each candidate storefront type, as defined in *ImageNet* synset[3], such as `cafe`, `bakery` etc. However, not all the words in street-level imagery should necessarily have semantic relations to the place type. Some words may have high similarity with one of the candidate classes, others may be completely irrelevant. For instance, words such as `hair`, `nail` or `beauty` on storefront images are likely to be related to a `Beauty Salon`. On the contrary, `OPEN/CLOSE` signs do not give any information about the type of storefront.

The text recognition module could result in some noisy texts, which need to be discarded. Before representing a word spotted by the word vector representation, we use a spell detection tool employing the Levenshtein Distance algorithm[4] to find permutations within an edit distance of 2 from the original word, and therefore remove noisy words. To further remove irrelevant words, we manually curated a blacklist of common – yet irrelevant – words, including verbs like `open`, `close`, `push`, `pull`, etc.

After reducing potential noise, we need to detect the language which the input word belongs to. In order to tackle this problem, we incorporate in our experiments the `polyglot` open source tool[5], which makes language prediction with a corresponding confidence score. If no language can be identified for the input word, English will be chosen as the default language.

Once the target language is determined, the recognized word must be transformed into a word vector representation. While there can be many implementations for capturing semantic relatedness [18], previous studies have shown that *word embeddings* [20,23] perform this task particularly well by measuring the cosine similarity of the word embedding vectors. These vector-based models represent words in a continuous vector space where semantically similar words are embedded close to one another. In our experiments, we adopt FastText [4] to transform recognized texts into a word vector representation. The main reason for incorporating FastText is its promising performance in overcoming the problem of out-of-vocabulary words, by representing each word as a bag of character n-grams. We use pre-trained word vectors for 2 languages (English and German), trained on Common Crawl and Wikipedia[6].

[3] http://www.image-net.org/synset.

[4] https://github.com/barrust/pyspellchecker.

[5] https://github.com/aboSamoor/polyglot.

[6] https://fasttext.cc/docs/en/crawl-vectors.html.

According to the detected language l, the corresponding pre-trained word vector V_l is selected; then, each recognized word is represented by the pre-trained 300-dimensional word vector as v_i. Finally, we use the method proposed by [25] to align the V_l in the same space as the English word vector for multilingual semantic matching. Similarly, each candidate class of storefront type C is represented by a word vector c_j with an English word embedding as reference. Then, we calculate the cosine similarity between each class label (c_j) and each spotted text (v_i) as follows:

$$\cos\left(\Theta_{ij}\right) = \frac{{v_i}^T c_j}{|v_i|\,|c_j|} \tag{1}$$

The probability scores P_i for each candidate storefront type is calculated by averaging similarity scores of all spotted words:

$$P_j = \frac{\sum_{i=1}^{K} \cos\left(\Theta_{ij}\right)}{K} \tag{2}$$

Then, a softmax function is used to normalize the probability scores for each candidate storefront type by the sum of the N candidate ranking scores so that they sum up to 1. The softmax function can be formulated as follows:

$$\sigma\left(Z\right)_j = \frac{e^{Z_j}}{\sum_{n=1}^{N} e^{Z_n}} \tag{3}$$

where Z is a vector of probability scores, N is the number of candidate classes, $j = 1, 2, \ldots, N$ is the index of each probability score in the probability vector Z, and $i = 1, 2, \ldots, K$ is the index of each spotted text. Similar to the scene recognition module, the scene-text extraction module results in a probability score for each candidate storefront type which is between 0 and 1.

3.3 Class Rank

Inspired by search re-ranking algorithms in information retrieval, we use a Linear Bimodal Fusion (LBF) method (here essentially a 2-component convex combination), which linearly combines the ranking scores provided by the CNN model and the semantic similarity scores from the scene-text semantic recognition module, as shown in Eq. 4.

$$S_{mixed}(d) = w_v.S_v(d) + (1 - w_v).S_t(d) \tag{4}$$

where S_{mixed}, $S_v(d)$, and $S_t(d)$ refer to the final ranking score, visual recognition score, and semantic similarity score for storefront type d respectively, w_v and w_t are the weights for the scene recognition component and scene-text extraction component, and $w_v + w_t = 1$. The weights are determined according to the relative performance of the individual components. Specifically, the weight for the scene recognition module is determined using the following equation:

$$w_v = \frac{acc_v}{acc_v + acc_t} \tag{5}$$

where acc_v and acc_t are the measured top@1 accuracy of scene recognition component and scene-text semantic recognition component, respectively.

4 Evaluation

This section reports on two experiments designed to assess the performance of our proposed approach in providing fine-grained classification of business storefronts using street-level imagery:

Comparison to Visual-Only Baselines. To show the utility and effectiveness of semantically rich textual information embedded in street-level imagery, we compare the classification performance of `ST-Sem` to two visual-only baselines, namely `GoogLeNet` [27] and `ResNet152` [10].

Comparison with Multimodal Methods. We also compare `ST-Sem` with a state-of-the-art method for fine-grained classification of business places [13], to highlight the advantages of incorporating multilingual word embeddings for semantic matching between textual cues on the image and different types of storefronts. To the best of our knowledge, the work of [13] is currently the best performing method in literature.

4.1 Experimental Setup

Datasets. We performed the experiments on three datasets: *Storefront*, a manually created collection of storefront images; *Places* [30], and *Con-text* [13]. Images in the *Storefront* and *Con-text* contain textual information, while *Places* is a mixture of images with and without text.

The *Storefront* fills a gap in existing datasets for the evaluation of solutions based on street-level imagery. Unlike other datasets, all images in *Storefront* are taken from a street-side perspective with complex scenes and objects with less visual differences, thus making it more difficult to classify the type of storefronts. We populated the *Storefront* dataset using the Street View API[7] from Google. We have randomly collected storefront images from two countries: USA and Germany. Those images are used in the storefront business type classification. The dataset consists of 1100 images, equally distributed in 22 categories: `Motel`, `Shoe Shop`, `Pharmacy`, `Jewelry`, `Candy Shop`, `Restaurant`, `Bakery`, `Cafe`, `Gift Shop`, `Pizzeria`, `Butcher`, `Bookstore`, `Icecream`, `Fastfood`, `Toy Shop`, `Auto Repair`, `Beauty Salon`, `Pet Shop`, `Supermarket`, `Bar`, `Clothing`, and `Hardware Store`.

To further evaluate the model, we carry out two additional experiments on two benchmarks that are used for fine-grained classification of places, including the business places we are interested in: (1) The *Places* dataset, introduced by [30]. It contains more than 10 million images divided into 365 unique scene categories with 5000 to 30,000 images each, and 100 testing images per category. As not all 365 categories are relevant for our study, we have selected a subset of images based on the categories described above. And (2) the *Con-text* dataset [13]. It is a sub-class (i.e. building and business places)

[7] https://developers.google.com/streetview/.

of the *ImageNet* dataset which consists of 24,255 images divided into 28 categories: `Steak House`, `Bistro`, `Pawn Shop`, `Cafe`, `Tavern`, `Tobacco`, `Dry Cleaner`, `Tea House`, `Country Store`, `Packing Store`, `Restaurant`, `Funeral`, `Repair Shop`, `Pharmacy`, `School`, `Computer Store`, `Medical Center`, `Diner`, `Pizzeria`, `Bakery`, `Hotspot`, `Massage Center`, `Pet Shop`, `Barber`, `Theatre`, `Bookstore`. The properties of each dataset are described in Table 1.

Table 1. Dataset statistics: number of images and place types

Dataset	#Categories	Training	Testing
Storefront	22	-	1,100
Places	22	12,500	2,200
Con-text	28	24,255	2,800

Implementation Details. Our method is built upon several pre-trained models: `ResNet152` [10] to extract image features, *TextBoxes++* [17] to localize texts in images, *MORAN* [19] to transform spotted texts into machine-readable character sequences, and 2 pre-trained word embeddings in two languages (English and German), trained on Common Crawl[8] and Wikipedia, to measure semantic relatedness between the transcribed words with the standard name of each candidate storefront type as defined in the *ImageNet* synset. A pre-trained *ResNet*152 model on *ImageNet*, is fine-tuned with a 28-way softmax classifier for the 28 candidate business types of the *Con-text* dataset. Similarly, a pre-trained *ResNet*152 model on the whole *Places* image database [30], is fine-tuned with a 22-way softmax classifier for *Places* and *Storefront*, which consist of 22 candidate types each. We used the same settings for fine-tuning the visual feature extractor on all datasets: the learning rate is set to 0.01 for the first 7K iterations, and then is divided by 10 every 10K iterations. The network is fine-tuned for 20 epochs. For the text representation, we only consider spotted words having more than 3 letters, under the assumption that words having less than 3 letters are not likely to provide relevant information for our classification task. The proposed method is implemented using the Caffe [12] deep learning framework. All the training and experiments are conducted on a NVIDIA Tesla P100 GPU. The source code of the pipeline, as well as the *Storefront* dataset, are available on GitHub[9].

Evaluation Metrics. Performance is evaluated using the following metrics: Average Precision (AP) of each category; and the mean of AP (mAP) over all

[8] http://commoncrawl.org/the-data/.
[9] https://github.com/shahinsharifi/ST-Sem.

categories. To calculate the Average Precision for each category, we use the formula used in [13,14] which is described as:

$$AP = \sum_{k=1}^{n} P(k)\Delta r(k) \tag{6}$$

where k is the rank in the sequence of classified images, n is the number of images in the current category, $P(k)$ is the precision at cut-off k, and $\Delta r(k)$ is the change in recall from items $k-1$ to k in the sorted list.

4.2 Results

Comparison with Visual-Only Baselines. We conduct an experiment to compare the performance of the proposed method with two visual-only baselines. We fine-tune the pre-trained GoogLeNet with a 28-class softmax classifier on *Con-text* and a 22-class softmax classifier on *Places*. The fine-tuned models on *Places* are also used for the experiment on *Storefront* as both datasets consist of the same categories. Our goal is to show the influence of leveraging textual information from imagery on the classification of business-related places. Table 2 displays the quantitative comparisons between two visual-only baselines and ST-Sem on all three datasets described above. The experiment shows that leveraging textual information from imagery in addition to visual features significantly outperforms visual-only baselines on all three datasets. As shown in Table 2, improvement varies between datasets due to the different amount of images that contain text.

Table 2. Classification performance of visual-only baselines and the proposed ST-Sem method on three datasets.

Datasets	Method	mAP(%)
Storefront	GoogLeNet	39.4564
	ResNet152	42.1785
	ST-Sem	**70.0524**
Places	GoogLeNet	83.5912
	ResNet152	85.7122
	ST-Sem	**87.0775**
Con-text	GoogLeNet	60.2569
	ResNet152	63.2546
	ST-Sem	**78.0221**

We observe that both visual baselines have relatively weak performance on *Storefront*. We attribute the difference to the visual layout of images from the *Storefront* dataset, as they are all produced from a street-level imagery source that differs from the training data of the respective network.

This scenario clearly highlights the advantages of ST-Sem, where by reusing pre-trained classifiers it is possible to achieve remarkable classification performance (79% and 66% increase on GoogLeNet and ResNet152, respectively) even when abundant training data are not available. Notice that by exploiting textual information, ST-Sem we are able to improve the classification performance also on the *Places* and *Con-text* datasets, with a respective 2% and 20% increase on GoogLeNet and ResNet152.

Table 3 breaks down the results by business category on the *Storefront* dataset. For some categories such as Motel or Pharmacy, all methods show relatively good performance. We account this result to the morphological dissimilarity of these POIs when compared to other types of business-related places. On the contrary, as most of bakeries have quite similar facades compared to other types of storefronts, both visual baselines show lower performance. We also observe that both visual baselines have relatively equal performance across all classes, which points out that the effectiveness of visual similarity is generally limited when the training data differ from the scenes to be predicted. We assume that the cross-class variations in the performance of ST-Sem are mainly accountable to the semantic ambiguity of the text used in their storefronts (arguably, hotel signs are less ambiguous than bookstores or beauty salons). We leave such an analysis to future work.

Table 3. Classification performance in mAP (%) for visual baselines and the proposed model for top 10 business type on the storefront dataset. Legend: MO – Motel; SH – Shoe Shop; PH – Pharmacy; JW – Jewelry; CSH – Candy Shop; RE – Restaurant; BA – Bakery; BS – Beauty Salon; CA –Cafe; BS – Bookstore

Method	MO	SH	PH	JW	CSH	RE	BA	BS	CA	BS
GoogLeNet	75.7	18.3	71.6	40.2	67.6	38.5	11.5	10.0	39.9	21.7
ResNet152	74.2	23.0	75.6	33.7	69.1	40.9	9.6	9.3	44.7	23.7
ST-Sem	**94.5**	**72.5**	**87.3**	**66.2**	**88.6**	**87.0**	**72.7**	**74.0**	**91.5**	**61.6**

Comparison with Multimodal Methods. We compare ST-Sem with Karaoglu et al. [13], the best performing state-of-the-art method that addresses the problem of business type classification by leveraging textual information from images. We perform experiments on three datasets. The results of Karaoglu et al. [13] are replicated by using the source code provided by the author[10]. The CNN models which are used in both methods to extract visual features are fine-tuned on *Con-text* and *Places*, respectively. As shown in Table 4, ST-Sem outperforms the state-of-the-art results from 67.5% to 70.05% (∼+4%) on the *Storefront* dataset. There is also a slightly better performance on *Context* (∼+1%) and *Places* (∼+2%).

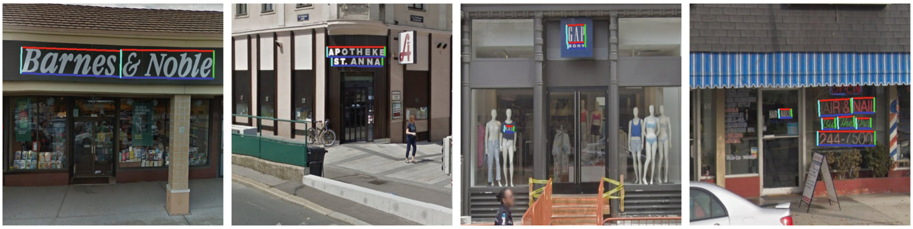
[10] https://github.com/taotaoorange/words-matter-scene-text-for-image-classification.

The results show that, in addition to obvious advantages in terms of language independence and modularity, `ST-Sem` provides comparable or superior performance across a variety of datasets.

Table 4. Classification performance mAP (%) for `ST-Sem` and Karaoglu et al. [13]

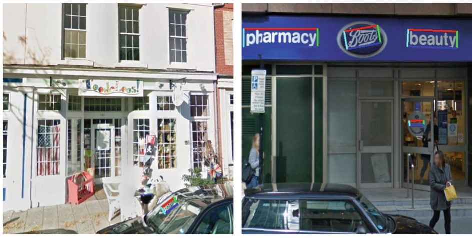
Datasets	Method	mAP(%)
Storefront	Karaoglu et al.	67.55
	ST-Sem	**70.05**
Places	Karaoglu et al.	84.35
	ST-Sem	**86.08**
Con-text	Karaoglu et al.	70.7
	ST-Sem	**71.35**

Discussion. In this section we discuss examples of scenes where `ST-Sem` provides non-obvious correct (Fig. 2) and incorrect (Fig. 3) predictions. As shown in Fig. 2a and c, `ST-Sem` can recognize the type of storefront, even when there is no word having direct relation to their types (e.g. *book* or *clothes*); the proposed semantic matching module is able to infer that texts such as *Barnes & Noble* and *GAP* are, respectively, semantically close to *bookstore* and *clothing* in the vector space, thus enabling correct classification. As depicted in Fig. 2b, the proposed method is also capable of measuring the semantic similarity between different languages. More specifically, *Apotheke* is recognized as a German scene-text on the image and then, it is transformed into a multilingual word vector which is semantically similar to *Pharmacy*.



(a) Bookstore (b) Pharmacy (c) Clothing (d) Beauty Salon

Fig. 2. Examples of correct classifications, with ground-truth label (GT) and probability score on the Storefront dataset. (a) GT: *Bookstore*, Predicted: *Bookstore - 0.71*; (b) GT: *Pharmacy*, Predicted: *Pharmacy - 0.83*; (c) GT: *Clothing*, Predicted: *Clothing - 0.75*, (d) GT: *Beauty Salon*, Predicted: *Beauty Salon - 0.67*



(a) Toy Shop (b) Pharmacy

Fig. 3. Examples of incorrect classification results on the storefront dataset. (a) GT: *Toy Shop*, Predicted: *Bookstore - 0.58* (b) GT: *Pharmacy*, Predicted: *Beauty Salon - 0.42*

Figure 3 shows examples of incorrect prediction. As shown in Fig. 3a, the scene-text detector failed to detect textual information on the corresponding image due to the uncommon font used in the signs. Therefore, the classification is only based on the visual features. This failure shows an obvious limitation of our method, i.e. that the overall performance is highly dependent on the performance of the scene-text detection module. Without textual information, the system simply relies on visual features. Figure 3b shows that the scene-text recognition module recognized two informative words (`pharmacy` and `beauty`) on the image, but the storefront type is not correctly classified. The reason of failure is likely to be that the semantic similarity scores of `pharmacy` and `Beauty Salon` are almost equal for this particular storefront. Therefore, similarly to the previous failure case, classification was only based on the morphological features of the storefront, which can indeed be erroneous.

5 Conclusion

In this work, we have introduced `ST-Sem`, a multimodal method that integrates visual and text cues extracted from street-level imagery for fine-grained classification of POIs. `ST-Sem` is able to incorporate multilingual word embeddings that measure the semantic similarity between detected textual information in street-level imagery and the candidate types of POIs. Moreover, `ST-Sem`'s late-fusion approach improves the performance of fine-grained storefront type classification, which is easily configurable for new datasets without requiring to re-train the models from scratch. Clearly, the effectiveness of `ST-Sem` is limited by the quality of the scene-text detection and the semantic similarity components. However, `ST-Sem` can be easily extended to include more methods.

In future work, we plan to extend the `ST-Sem` pipeline by incorporating additional semantically-rich information in the classification process, such as contextual information and user reviews, where available. We also plan to extend the scope of our experiments to other cities and languages, to further demonstrate the generalizability of the approach.

References

1. Alcantarilla, P.F., Stent, S., Ros, G., Arroyo, R., Gherardi, R.: Street-view change detection with deconvolutional networks. Auton. Robots **42**(7), 1301–1322 (2018)
2. Balduini, M., Bozzon, A., Della Valle, E., Huang, Y., Houben, G.J.: Recommending venues using continuous predictive social media analytics. IEEE Internet Comput. **18**(5), 28–35 (2014)
3. Bocconi, S., Bozzon, A., Psyllidis, A., Titos Bolivar, C., Houben, G.J.: Social glass: a platform for urban analytics and decision-making through heterogeneous social data. In: Proceedings of the 24th International Conference on World Wide Web, pp. 175–178. WWW 2015 Companion. ACM, New York (2015)
4. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Trans. Assoc. Comput. Linguist. **5**, 135–146 (2017)
5. Doersch, C., Singh, S., Gupta, A., Sivic, J., Efros, A.: What makes Paris look like Paris? ACM Trans. Graph. **31**(4) (2012)
6. Falcone, D., Mascolo, C., Comito, C., Talia, D., Crowcroft, J.: What is this place? inferring place categories through user patterns identification in geo-tagged tweets. In: 2014 6th International Conference on Mobile Computing, Applications and Services (MobiCASE), pp. 10–19. IEEE (2014)
7. Fu, K., Chen, Z., Lu, C.T.: Streetnet: preference learning with convolutional neural network on urban crime perception. In: Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, pp. 269–278. ACM (2018)
8. Gebru, T., Krause, J., Wang, Y., Chen, D., Deng, J., Aiden, E.L., Fei-Fei, L.: Using deep learning and google street view to estimate the demographic makeup of the us. arXiv preprint arXiv:1702.06683 (2017)
9. Goel, R., et al.: Estimating city-level travel patterns using street imagery: a case study of using Google street view in britain. PloS One **13**(5), e0196521 (2018)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
11. Jaderberg, M., Simonyan, K., Vedaldi, A., Zisserman, A.: Reading text in the wild with convolutional neural networks. Int. J. Comput. Vis. **116**(1), 1–20 (2016)
12. Jia, Y., et al.: Caffe: convolutional architecture for fast feature embedding. In: Proceedings of the 22nd ACM International Conference on Multimedia, pp. 675–678. ACM (2014)
13. Karaoglu, S., Tao, R., van Gemert, J.C., Gevers, T.: Con-text: text detection for fine-grained object classification. IEEE Trans. Image Proc. **26**(8), 3965–3980 (2017)
14. Karaoglu, S., Tao, R., Gevers, T., Smeulders, A.W.: Words matter: scene text for image classification and retrieval. IEEE Trans. Multimed. **19**(5), 1063–1076 (2017)
15. Karatzas, D., et al.: ICDAR 2015 competition on robust reading. In: 2015 13th International Conference on Document Analysis and Recognition (ICDAR), pp. 1156–1160. IEEE (2015)
16. Li, X., Ratti, C., Seiferling, I.: Mapping urban landscapes along streets using Google street view. In: Peterson, M.P. (ed.) ICACI 2017. LNGC, pp. 341–356. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-57336-6_24
17. Liao, M., Shi, B., Bai, X.: Textboxes++: a single-shot oriented scene text detector. IEEE Trans. Image Proc. **27**(8), 3676–3690 (2018)
18. Lofi, C.: Measuring semantic similarity and relatedness with distributional and knowledge-based approaches. Inf. Media Technol. **10**(3), 493–501 (2015)

19. Luo, C., Jin, L., Sun, Z.: Moran: A multi-object rectified attention network for scene text recognition. Pattern Recognition (2019)
20. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems, pp. 3111–3119 (2013)
21. Movshovitz-Attias, Y., Yu, Q., Stumpe, M.C., Shet, V., Arnoud, S., Yatziv, L.: Ontological supervision for fine grained classification of street view storefronts. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1693–1702 (2015)
22. Parkhi, O.M., Vedaldi, A., Zisserman, A., et al.: Deep face recognition. In: BMVC, vol. 1, p. 6 (2015)
23. Pennington, J., Socher, R., Manning, C.: Glove: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543 (2014)
24. Quy Phan, T., Shivakumara, P., Tian, S., Lim Tan, C.: Recognizing text with perspective distortion in natural scenes. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 569–576 (2013)
25. Smith, S.L., Turban, D.H., Hamblin, S., Hammerla, N.Y.: Offline bilingual word vectors, orthogonal transformations and the inverted softmax. arXiv preprint arXiv:1702.03859 (2017)
26. Srivastava, S., Vargas Muñoz, J.E., Lobry, S., Tuia, D.: Fine-grained landuse characterization using ground-based pictures: a deep learning solution based on globally available data. Int. J. Geogr. Inf. Sci. 1–20 (2018)
27. Szegedy, C., et al.: Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9 (2015)
28. Yan, B., Janowicz, K., Mai, G., Zhu, R.: xnet+sc: Classifying places based on images by incorporating spatial contexts. In: 10th International Conference on Geographic Information Science (GIScience 2018). Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik (2018)
29. Yang, D., Li, B., Cudré-Mauroux, P.: Poisketch: semantic place labeling over user activity streams. Technical Report, Université de Fribourg (2016)
30. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: a 10 million image database for scene recognition. IEEE Trans. Pattern Anal. Mach. Intell. **40**(6), 1452–1464 (2018)
31. Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., Oliva, A.: Learning deep features for scene recognition using places database. In: Advances in Neural Information Processing Systems, pp. 487–495 (2014)
32. Zhu, Y., Deng, X., Newsam, S.: Fine-grained land use classification at the city scale using ground-level images. IEEE Trans. Multimed. (2019)