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

Shahin Sharifi, Achilleas Psyllidis and Alessandro Bozzon

s.sharifiNoorian@tudelft.nl

TU Delft Delft University of Technology

Points-of-Interest (POI) Data

- Keeping maps, in terms of POI data, up to date is very **expensive** and **time-consuming**

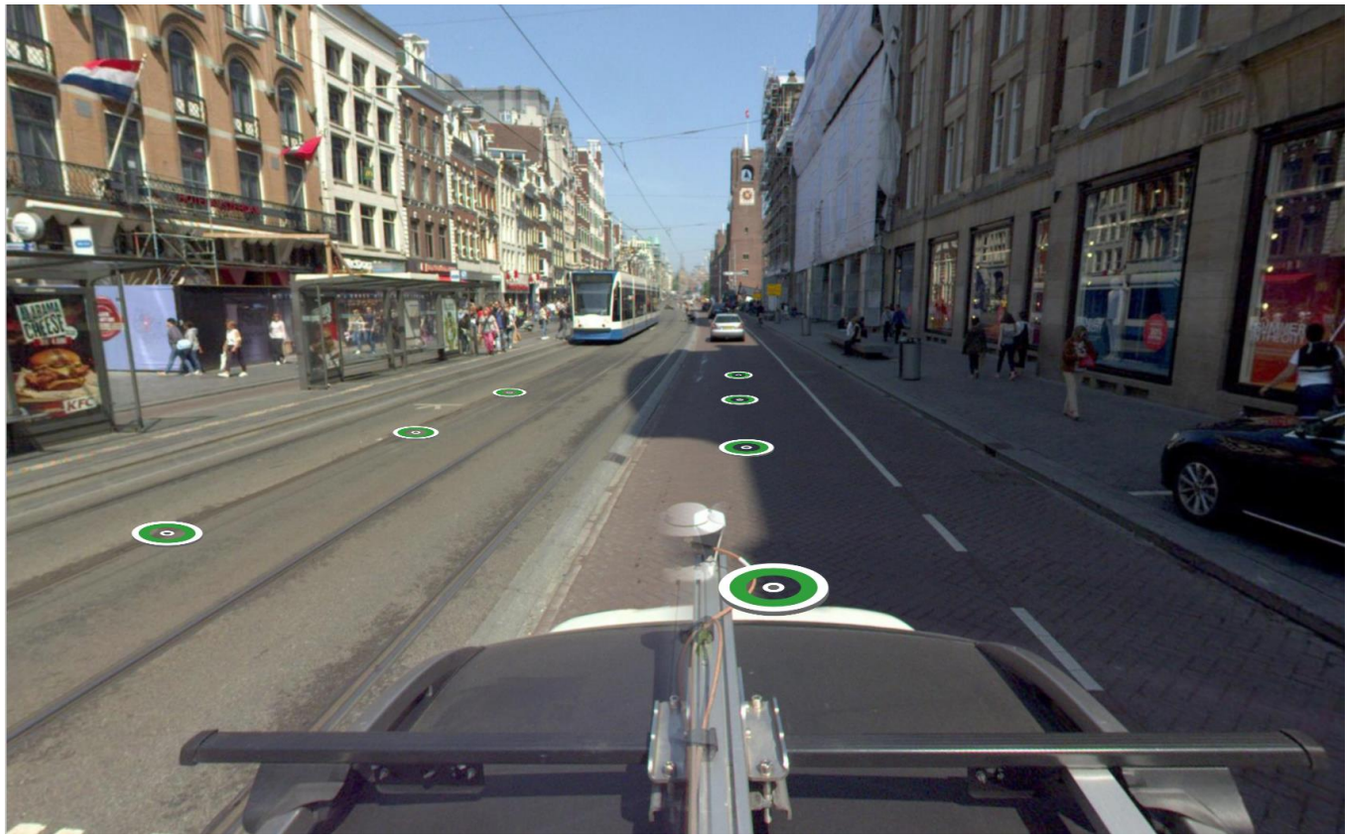
- The update process involves **continuously capturing vast amounts of information** about Points of Interest (POIs) from numerous sources around the world.

Street-level Imagery

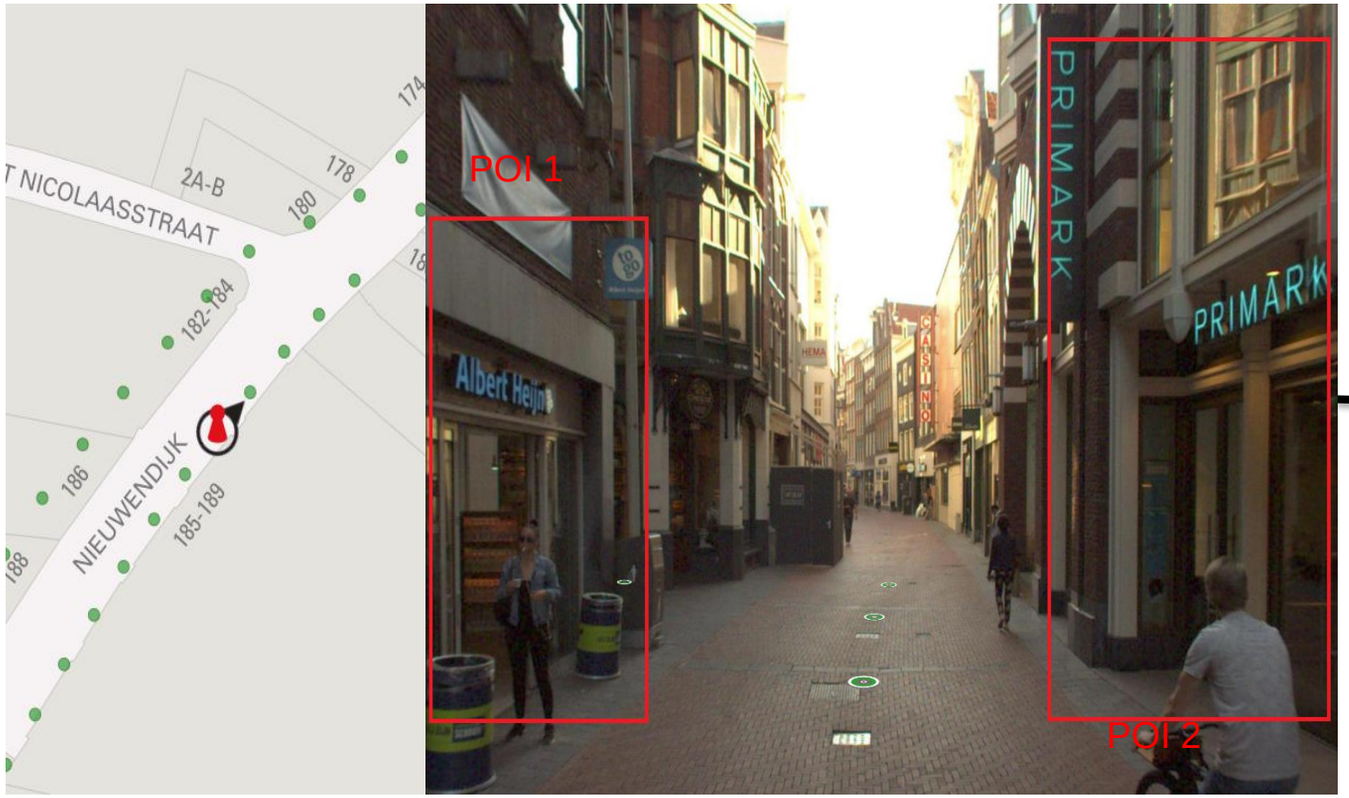
- Street-level imagery provides a variety of visual information about - primarily urban - environments at ground level, including the **facades of Points of Interest (POIs)**

- It has promising **update rate**

- It is available in a **big scale**



Problem



?

POI 1

Supermarket (80%)

Cafe (5%)

Bakery(2%)

Clothing(0.5%)

...

POI 2

Clothing (90%)

Book store (5%)

Cafe(2%)

Restaurant (0.5%)

...

Importance of non-morphological features

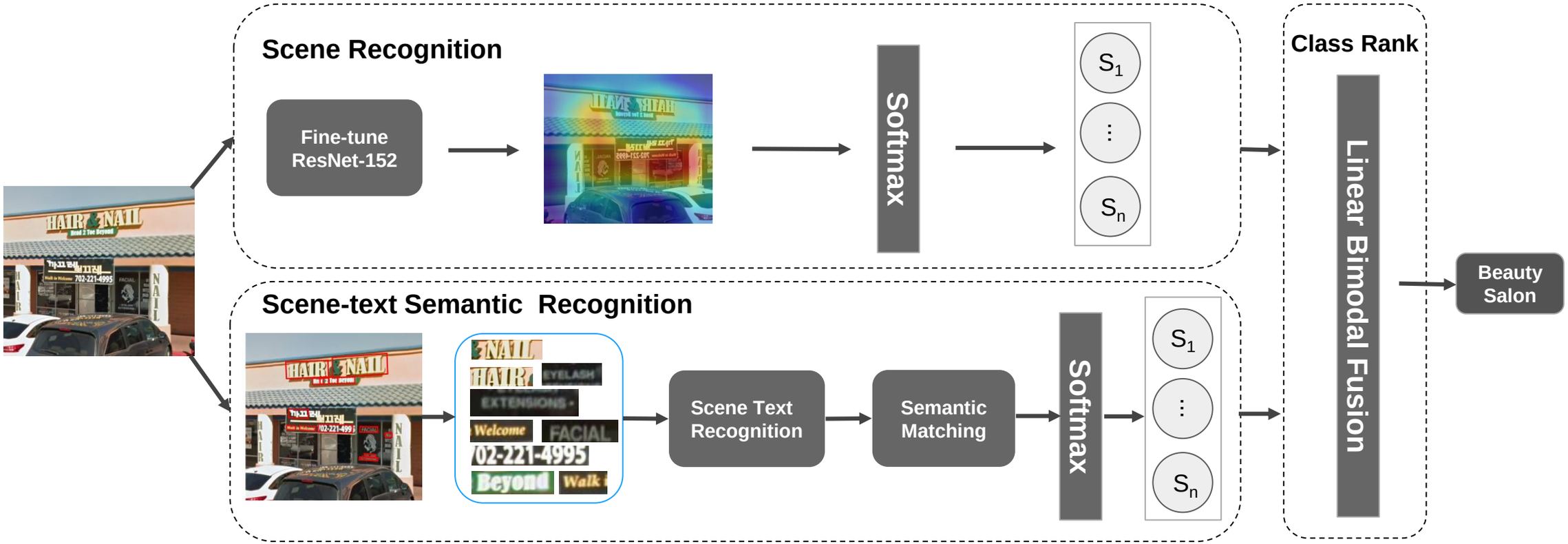
- Textual information on the facades of business-related POIs could be a valuable source of information which usually represents the type of POI.

Core Question

To what extent can we automatically extract the type of POIs from street-level imagery with minimal human effort?

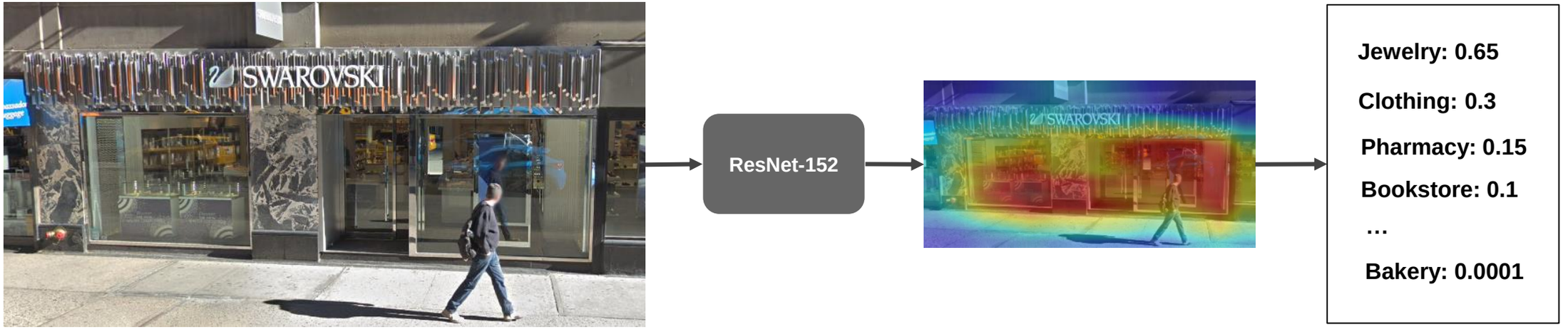
ST-SEM: Scene Text Semantics

A multimodal end-to-end approach which extracts **visual** and **textual** features from the input image and makes prediction on the type of POI.



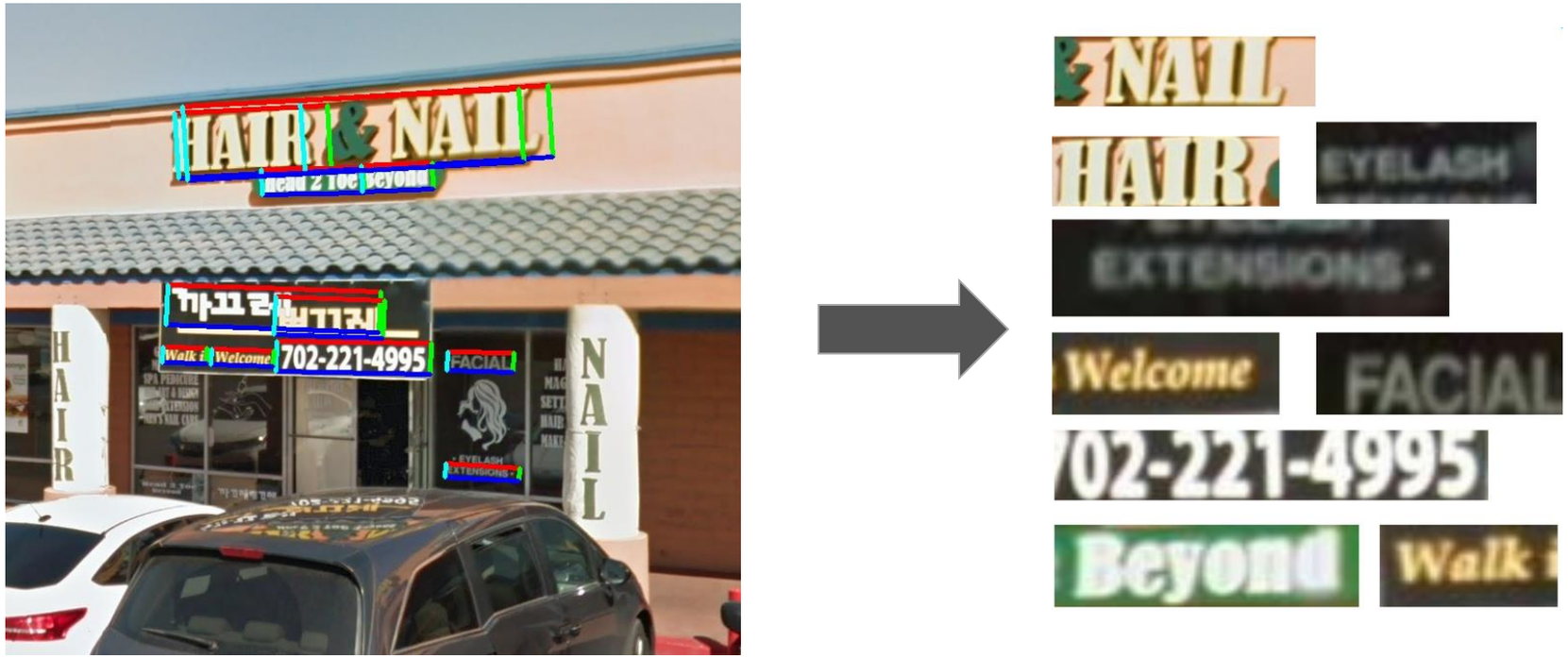
Scene Recognition

- Residual Network with 152 layers (**Resnet-152**)
- The model was trained on the **Places-365** dataset
- It is fine-tuned with a **22-way softmax** classifier according to 22 candidate types which are defined for the experiments



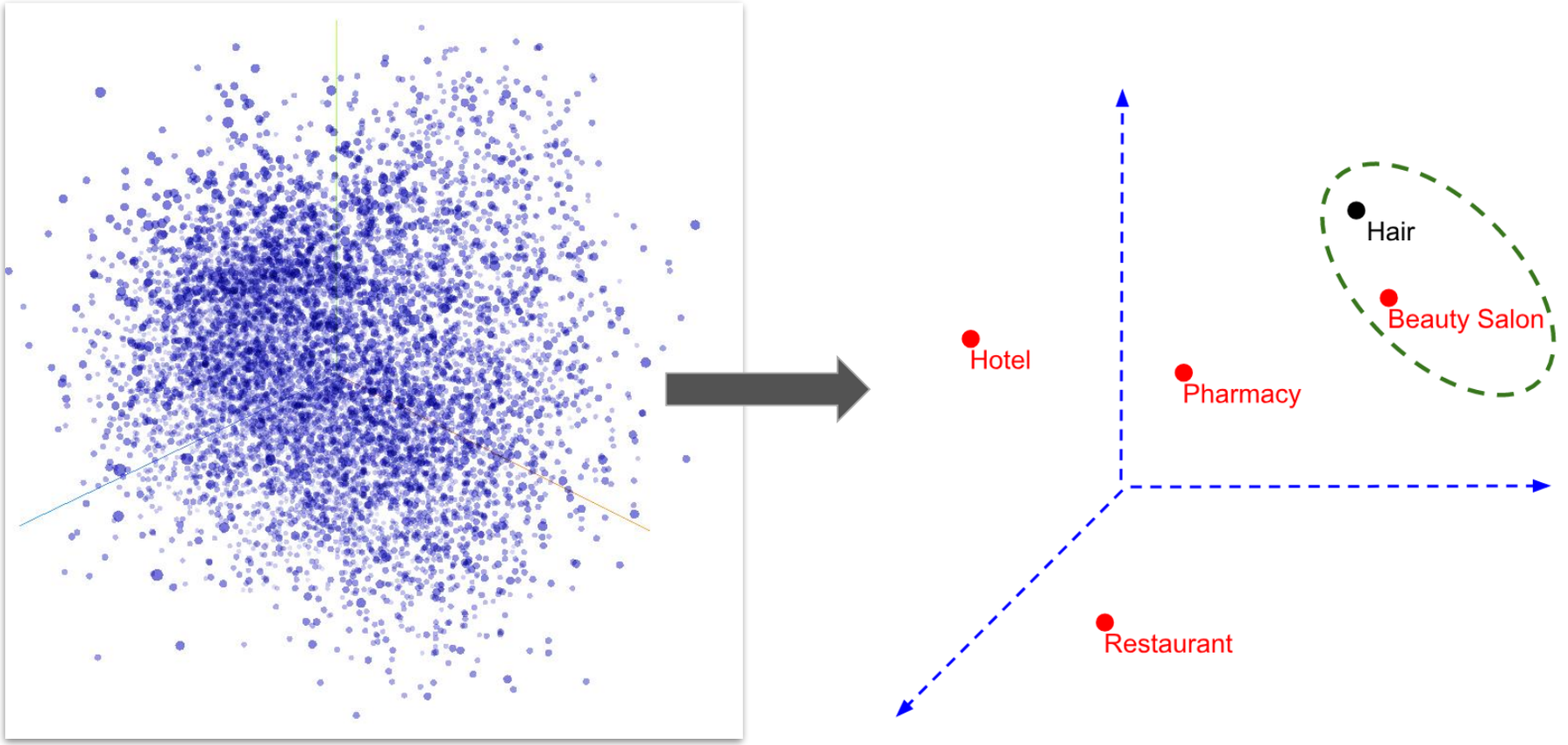
Scene-text Semantic Recognition

- The signs and textual information on the façade of POIs seems to be very effective for accurately predicting the type of POIs
- Convolutional Neural Network (**GoogleNet**) is trained on ICDAR dataset for scene-text detection
- A **multi-object rectified attention network** is used for transforming visual-text to machine-readable text



Word Embeddings

Word Embeddings are a vectorized form of words which are positioned in the vector space such that words that share common contexts in the corpus (e.g. Wikipedia) are located in close proximity to one another in the space



Example: GAP (Clothing)

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

nteraction

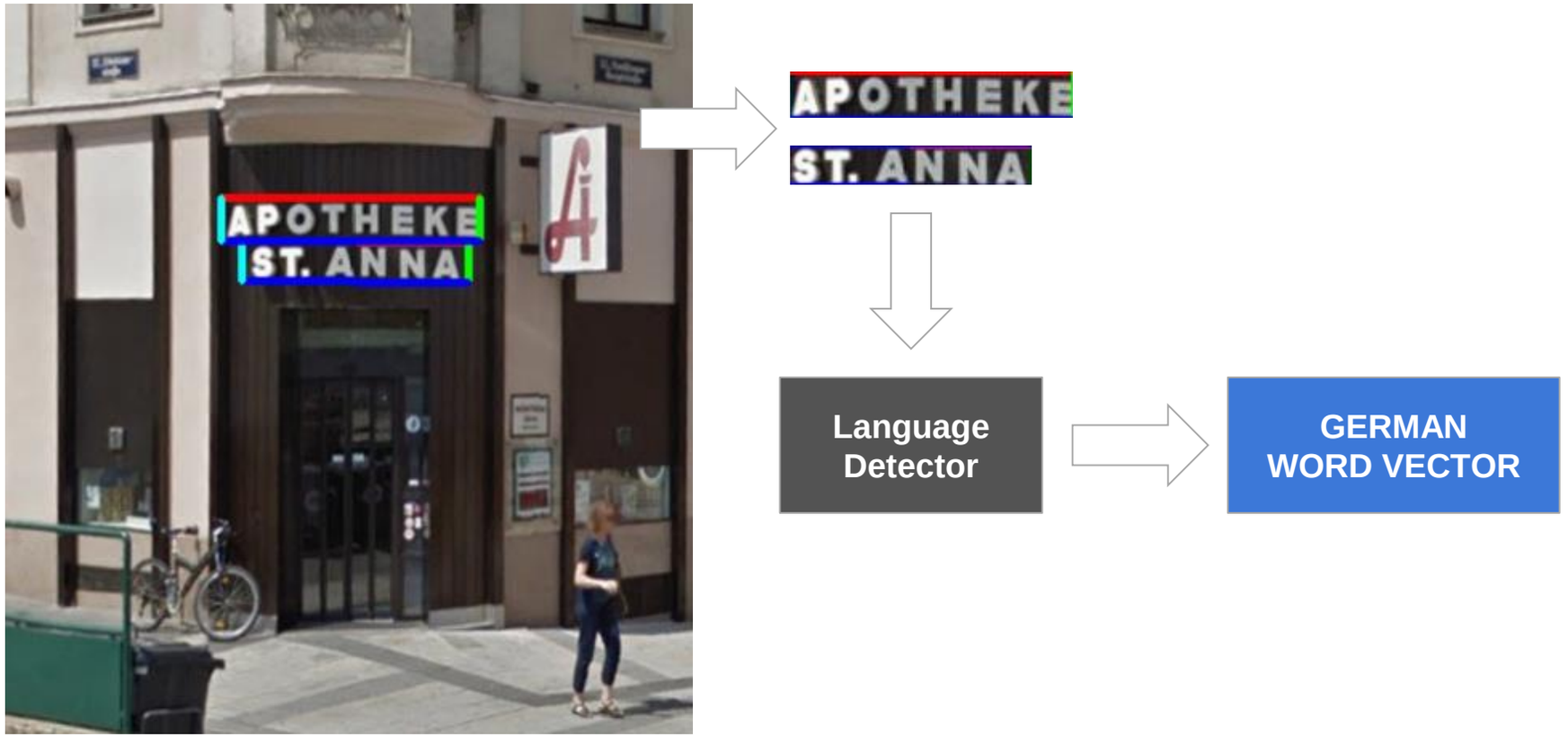
Gap Inc.

From Wikipedia, the free encyclopedia

The Gap, Inc.,[4] commonly known as **Gap Inc.** or **Gap**, (stylized as **GAP**) is an American worldwide clothing and accessories retailer.

It was founded in 1969 by Donald Fisher and Doris F. Fisher and is headquartered in San Francisco, California. The company operates six primary divisions: Gap (the namesake banner), Banana Republic, Old Navy, Intermix, Weddington Way, Hill City, and Athleta . Gap Inc. is the largest specialty retailer in the United States, and is 3rd in total international locations, behind Inditex Group and H&M.[5] As of

Cross-language Semantic Similarity



Experimental Setup

Datasets:

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

Evaluation Metric: The Average Precision (AP) of each category and the mean of AP (mAP) over all categories.

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

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
Δr(k) is the change in recall from items k-1 to k in the sorted list.

Results

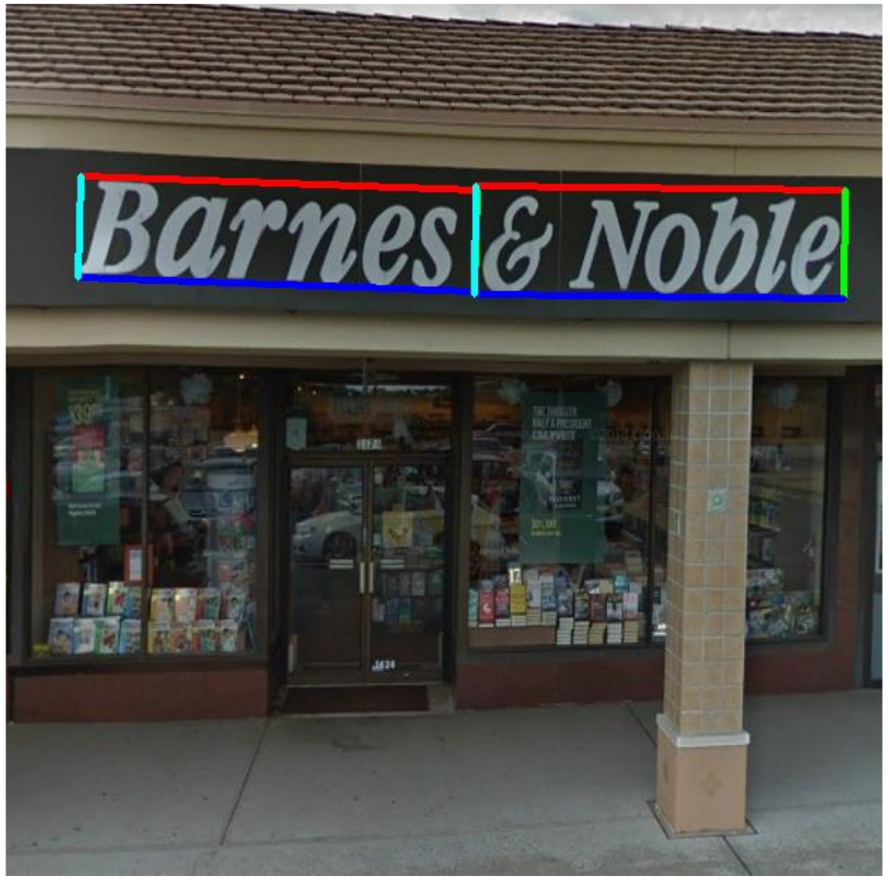
Comparing to the visual-only baseline:

Datasets	Method	mAP(%)
Places	Baseline	85.71
	ST-SEM	87.07
Con-text	Baseline	63.25
	ST-SEM	78.02
Storefront	Baseline	42.17
	ST-SEM	70.05

Comparing to the Multi-modal baseline:

Datasets	Method	mAP(%)
Places	Baseline	84.35
	ST-SEM	86.08
Con-text	Baseline	70.7
	ST-SEM	71.35
Storefront	Baseline	67.55
	ST-SEM	70.05

Correct Predictions

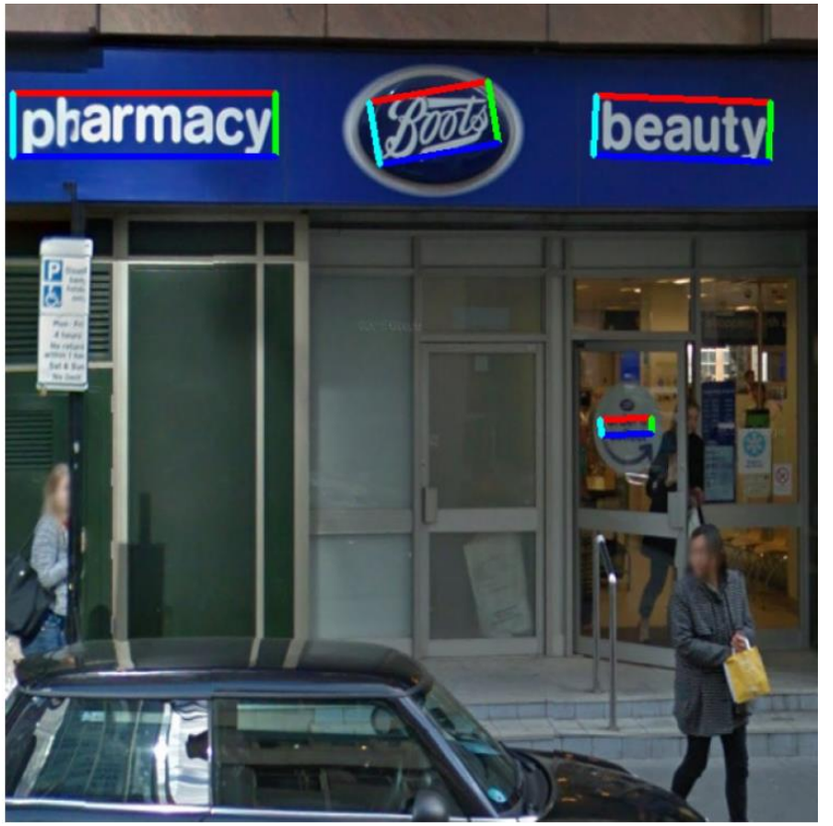


Actual: Bookstore
Predicted: Bookstore

Actual: Beauty Salon
Predicted: Beauty Salon

Limitations

Actual: Toy shop
Predicted: Bookstore



Actual: Pharmacy
Predicted: Beauty Salon

Future Work


SUBWAY
HAIR & NAIL
Head 2 Toe Beyond
NARA
EXPRESS
LEPIDOR
PUPPY TOWN
Scene Recognition
Fine-tune ResNet-152
Softmax
Scene-text Semantic Recognition
Scene Text Recognition
Semantic Matching
Softmax
Class Rank
Linear Bimodal Fusion
Beauty Salon

Summary

- Keeping maps, in terms of POI data, up to date is very expensive and time-consuming
- In order to alleviate the high cost of updating poi data, street-level imageries can be an option for extracting valuable information about points-of-interest ,mainly in urban area.
- Automatically extracting information about points-of-interests from street level imagery is very complicated.
- A multimodal end-to-end approach is proposed to leverage visual and textual features from the input image and predicts the type of POI.
- We have achieved promising results, but the proposed approach can be further improved.