# Entrance Detection from Street-View Images

Jingchen Liu[1] Thommen Korah[2] Varsha Hedau[2] Vasu Parameswaran[2]
Radek Grzeszczuk[2] Yanxi Liu[1]
[1]The Pennsylvania State University, University Park, PA [2]Microsoft Corporation, Sunnyvale, CA

## Abstract

*We present a system for detecting building entrances in outdoor scenes, an important problem for urban scene understanding. While entrance detection in indoor scenes has received a lot of attention, tackling the problem in outdoor scenes is considerably more complicated and remains largely unexplored. The wide variety of door appearances and geometries, background clutter, occlusions, specularity, and other difficult lighting conditions together impose many challenges. In this paper, we propose a three stage system that starts with a high-recall entrance candidate extractor. The next stage classifies candidates based on local image features. The final stage fuses results from multiple views by using MCMC to solve a Bayesian inference problem, and to select the best set of entrances that explain the image of a facade. We achieve a precision of* 70% *at a recall of* 70% *on a challenging dataset of urban scene images. We will release this benchmark dataset to the public to facilitate future research on this topic.*

## 1. Introduction

Urban scene understanding, which entails parsing street-level imagery into constituent elements in the scene such as road, vehicles, buildings, facades, shops, entrances, etc, has received very little attention from the research community so far. In this paper, we study the problem of entrance detection from street-level imagery, an important sub-problem within urban scene, and which has many broad applications, especially in the areas of geo-coding, augmented reality, etc. While our full dataset includes imagery and LiDAR data, for this work, we focus on a vision based entrance detector, leaving LiDAR based entrance detection for a sequel. A vision based approach can complement LiDAR where it is insufficient, such as for closed entrances lying on the same plane as the building facade and transparent surfaces such as glass doors, which occur frequently in urban scenes. The input to the system is a collection of calibrated images and foreground masks, delineating an approximate region from the groundline up to roughly the first floor. These masks are generated via LiDAR, or can be derived from other sources such as digital surface maps. As part of the paper we will release a public dataset forming a benchmark that can be used to stimulate work on this challenging and interesting problem.

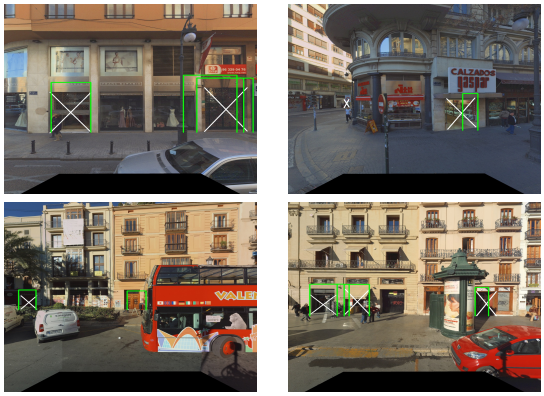


Figure 1. Entrance detections (green), Ground truth (white)

## 2. Approach

We first *detect* entrance candidates (rectangular image patches) based on edgelet distributions, then extract features to *classify* each candidate independently. Finally, we perform *joint inference in 3D* to resolve conflicts such as overlapping entrances, and to apply global constraints, to select the best set of entrance locations on a given facade.

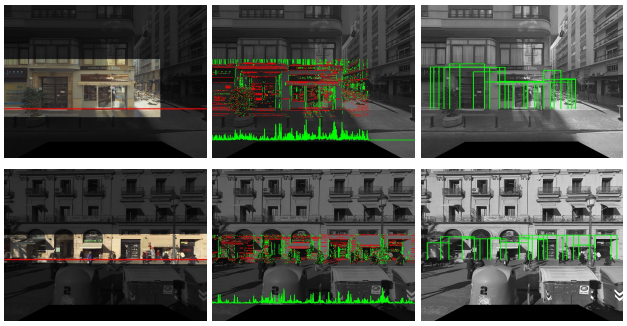
### 2.1. Candidate Extraction

As entrances are typically delineated by strong edges, we extract entrance candidates via 1D edgelet distributions. Entrance candidates are modeled as rectangular bounding boxes, whose bottom boundary aligns with the groundline. The left, right and top boundaries are determined by peak extraction from accumulated 1D edgelet distributions as shown in the two examples in Fig. 2. The parameters for this initial are set for high recall.

### 2.2. Entrance Classification

The goal for this stage is to perform local analysis around entrance candidates to prune away false positives. We ex-

perimented with four classes of feature: *Histogram of Gradient*(HoG), *Principal Component Analysis* (PCA), *Reflection Symmetry*(Sym), *Color Statistics*(CS). Our motivation for using HoG is its robustness to photometric variations and its ability to capture the overall structure around doors. We use the default setting of HoG feature as commonly used in the scenario of pedestrian detection ($16 \times 8$ cells and 9 histogram bins). We used PCA for reducing dimensionality of the HoG based feature vector. We learn dominant eigenvectors (we call *eigen-doors*) from positive samples and extract the low-dimensional reconstruction coefficients for both positive and negative candidate patches for discriminative training. Entrances are typically symmetric, which motivated us to use left-right-reflection symmetry. We apply a local scanning window with different scales, scanning through the left half of a candidate patch, compute its (histogram) difference to the corresponding patch on the right as a 'symmetry' feature. Finally we extract color statistic features (pixel mean and standard deviation) through a local scanning window under multiple scales. For classification, we used a Random Forest visual classifier learned from extracted features to produce a soft decision for each candidate patch independently [1].



(a) (b) (c)

Figure 2. (a) Input data with foreground facade masks and ground-lines; (b) Vertical (green) and horizontal (red) edgelets; (c) Entrance candidates

### 2.3. Joint Facade Analysis

Using calibration information, we back-project entrance candidates $c_n$ from multiple views to the same facade. From these, we select true and false positives. A solution 'hypothesis' $\mathbf{z}$ is a bit vector of length $n$ with (1,0) denoting (true,false) positives. Observations $\mathbf{O}$ are the classification scores. Using a Bayesian formulation: $P(\mathbf{z}|\mathbf{O}) \propto P(\mathbf{z})P(\mathbf{O}|\mathbf{z})$. For priors, we model entrance density i.e. number of entrances per meter as a Gaussian distribution. We incorporate hard non-overlapping constraint between 3D entrances into the prior.

$$P(\mathbf{z}) = \begin{cases} 0 & \exists(i,j) \text{s.t. } z_i = 1, z_j = 1, D_{ij}^{(h)} < \tau_1, \\ P(\mathbf{z}) = \mathcal{N}(\frac{\|\mathbf{z}\|_1}{L}|\mu,\sigma), & \text{otherwise} \end{cases}$$

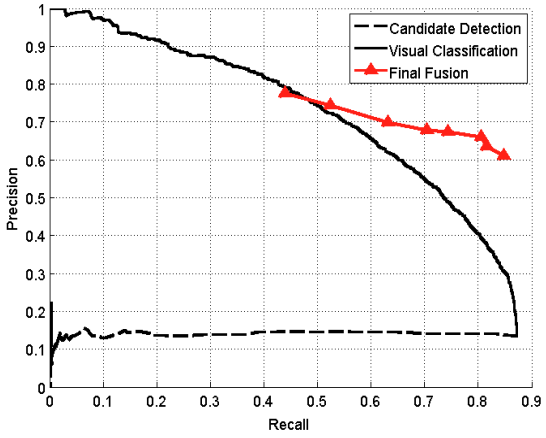



Figure 3. PR curve for detection, classification, and fusion.

where $L$ is the facade base length, $D_{ij}$ is the horizontal real-world distance between the center of candidate $i$ and $j$, and we use $\tau_1 = 1.5$(meter). $\mu$ and $\sigma$ are the average and standard deviation of entrance density learned empirically.

We assume independent observations and model each observation $\mathbf{O}_n$ as $P(\mathbf{O}_n|\mathbf{z}) = \mathbf{s}_n\mathbf{t}_n$ where $\mathbf{s}_n \in \{P_n^{(tp)}, P_n^{(fp)}\}$ is the likelihood of obtaining the observed classification score. $P_n^{(tp)}$, $P_n^{(fp)}$ is the probability that the visual classifier produces a true and false positive respectively. $\mathbf{t}_n \in \{1, P_n^{(fn)}\}$ where $P_n^{(fn)}$ is the probability that the visual classifier produces a false negative. A similar formulation has been used in the domain of multi-view pedestrian detection/association, *e.g.*, [2]. We adopt a stochastic optimization approach similar to *Markov chain Monte Carlo (MCMC)* and sample for the Maximum-a-Posterior solution of $\mathbf{z}$, during which we apply 3 types of balanced local moves: add/remove/shift an entrance.

## 3. Dataset and Experiments

We introduce a challenging benchmark for entrance detection from street-view images containing 250 panoramic images of resolution $1016 \times 1016$ and masks for the first floor region We evaluate the precision and recall performances in all 3 stages of our algorithm: candidate extraction, visual classification and detection fusion in 3D as given in Fig. 3. It can be seen that each succeeding stage improves the performance, and that the end-to-end system is able to achieve a precision and recall of about 70% each. For future work, we are working LiDAR analysis of the facade and fusion with image based observations.

## References


[1] L. Breiman. Random forests. *Machine Learning*, 45(1), 2001. 2

[2] L. Zhang, Y. Li, and R. Nevatia. Global data association for multi-object tracking using network flows. In *CVPR*, 2008. 2