Lost Shopping! Monocular Localization in Large Indoor Spaces

Shenlong Wang, Sanja Fidler, Raquel Urtasun
Department of Computer Science
University of Toronto
{slwang, fidler, urtasun}@cs.toronto.edu

Abstract

In this paper we propose a novel approach to localization in very large indoor spaces (i.e., shopping malls with over 200 stores) that takes a single image and a floor plan of the environment as input. We formulate the localization problem as inference in a Markov random field, which jointly reasons about text detection (localizing shop names in the image with precise bounding boxes), shop facade segmentation, as well as camera's rotation and translation within the entire shopping mall. The power of our approach is that it does not use any prior information about appearance and instead exploits text detections corresponding to shop names as a cue for localization. This makes our method applicable to a variety of domains and robust to store appearance variation across countries, seasons, and illumination conditions. We demonstrate our approach on our new dataset spanning two very large shopping malls, and show the power of holistic reasoning.

1. Introduction

Due to the development of cost-effective solutions such as the global positioning system (GPS), people can easily localize, navigate and route through the city by simply clicking on their phone. Localization in large indoor spaces is, however, much more difficult, since GPS typically cannot communicate with the satellites inside the buildings. As a consequence, indoor navigation is still an open, yet crucial problem with a huge potential impact on many commercial and public services. The goal of this paper is to perform localization in large shopping malls given only a single image and a floor plan of the mall.

Sensor-based approaches to indoor localization have been developed, but require either a large number of nearby anchor points [39, 10] (*e.g.*, WiFi access points or beacons with known positions) that are densely distributed within the scene or pre-assume initial absolute locations [2, 1]. Most vision-based localization systems rely on a pre-recorded dataset containing all places of interest, and localization involves indexing in the dataset by matching the visual appearance [18, 31, 23, 42] and/or geometry [9, 34, 5, 30]. These approaches have the disadvantage that requires a priori knowledge of how the world looks like and are thus not very robust to appearance and geometric changes. The former is particularly important in indoor scenarios such as shopping malls, where the shops vary their display regularly to show their new seasonal products.

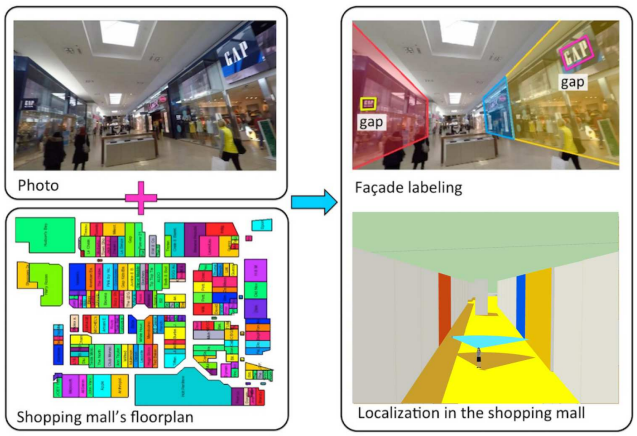


Figure 1. Given a monocular image and a shopping mall's floorplan, our goal is to estimate the 3D camera pose, and parse the facades of the visible shops.

SfM and SLAM methods [3, 34, 13] estimate camera pose and a map of the environment by capturing a large collection of images. Recently, [8] proposed to localize a car by matching the vehicle's trajectory to cartographic maps annotated with the road topology. This is appealing as it does not require knowledge of the world's appearance and only requires the cartographic map to be up to date. However, such an approach would likely fail in large indoor spaces since the corridor topology is not very discriminative as well as people move in less structured ways.

Instead, our work builds on the following observation: Suppose you are lost somewhere inside a big shopping mall and you need to meet your friend in Starbucks for a cup of overpriced coffee. The most common way of finding one's path is to find the name of one or two of the shops around you, and look them up in the shopping mall's floor-

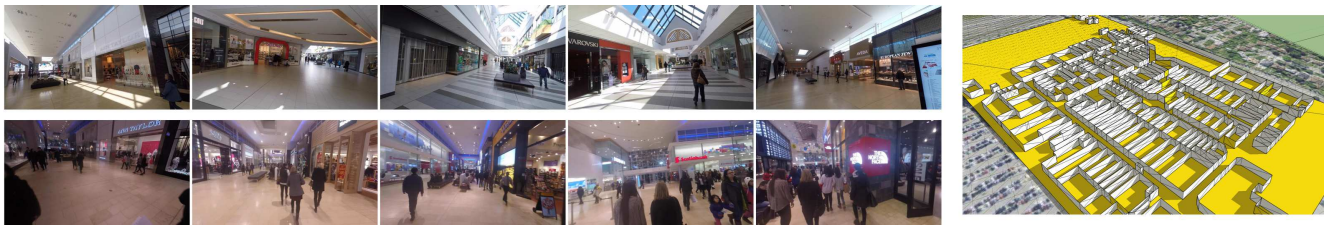


Figure 2. Overview of our dataset. **Left**: example images; **right**: 3D floorplan

Malls	Retailers	Images	Images with labeled texts	Images with labeled facades	Images with labeled locations
2	358	15984	3099	701	246

Table 1. Statistics about our shopping mall dataset, collected in two large malls (Promenade and Yorkdale in Toronto).

plan (typically available at the info point). Once you are localized, you can start planning a path to the desired destination based on the information provided by the map.

Following this intuition, our first contribution is a novel approach that can perform accurate localization within large 200+ store shopping malls using a single RGB image and the mall's floor plan. The floor plan contains rich information about which stores the mall has and their location, as well as the widths of the corridors and store facades. We formulate the indoor localization problem as inference in a Markov random field (MRF), which jointly reasons about text detection (localizing shop's names in the image with precise bounding boxes), shop facade segmentation, as well as camera's rotation and translation within the entire shopping mall. The power of our approach is that it does not use any prior information about appearance (such as image examples of particular stores) and instead only exploits text detections corresponding to the shop names. This makes our method applicable to a variety of domains and robust to store appearance variation across countries, seasons, and illumination conditions. Our second contribution is a new dataset containing precise ground-truth annotations for all these tasks for two large shopping malls. Our experiments show that our holistic model achieves good accuracy, outperforming the baselines that solve the individual tasks.

All our code and data is publicly available at: `http://www.cs.toronto.edu/~slwang/lostShopping/`. We are also planning to set up an online benchmark in order to inspire the community to work on and push forward the performance of this challenging problem.

2. Related Work

The most popular solution to indoor positioning are sensor-based methods [10, 39, 40], which rely on measuring distance to nearby anchor nodes (*e.g.* WiFi AP, Bluetooth iBeacon). They require distributed geolocated anchor nodes and sometimes specific sensors for users. For instance, a WiFi-based positioning system [10] measures the intensity of the received signal from the surrounding WIFI access points for which the location is known. As a consequence it relies heavily on maintaining a geolocalized wifi dataset which can be out-of-date very quickly. Moreover, the localization accuracy may fluctuate due to changes in signal strength and only works for regions with a sufficient number of sensors to enable trilateration and triangulation.

Simultaneous localization and mapping (SLAM) [13, 16] and structure-from-motion (SfM) [3, 34] are used for mobile robot localization [37]. These methods require a large collection of images to perform localization and build a 3D map of the environment. Recently, [8, 15] proposed to use the ego-trajectory and a cartographic map to localize a car. Similar ideas have been used in Project Tango, which localizes based on the estimated trajectory and the manually annotated starting point. However, the difficult conditions in shopping malls (e.g., moving crowds, occlusion, specularities, transparency), present a challenge for visual odometry and SfM. In contrast, in our work we aim to localize from a single RGB image and a floor plan. This makes a very natural setting where a user can simply take a photo and get her/his geolocation. Our work is also related to [38] who performed 3D parsing of a single outdoor image by exploiting a crowd-sourced cartographic map. However, in this work the authors assumed that localization was given by the vehicle's GPS and IMU.

Place recognition approaches [18, 42, 23, 36, 7] localize by matching an image against a large database. These approaches rely on large collections of pre-registered images with a dense coverage, which is especially difficult to acquire for indoor scenes. Retrieval approaches in indoor environments have typically been limited to relatively constrained spaces [29, 24]. For indoor scenes like shopping malls, they would require very frequent updates, making them impractical. In [5], the authors proposed an appearance free approach to outdoor localization that matches buildings' corner-points in the elevation map. For indoors, [4] perform localization by matching objects such as chairs/doors to configurations in the floorplan. This is a very interesting approach, however, it cannot be used in our setting since the shopping mall's floorplans do not contain this information. Our work also falls in the domain of egocentric computer vision. Previous work on egocentric cameras tackle scene [35, 28] and action recognition [14], field-of-

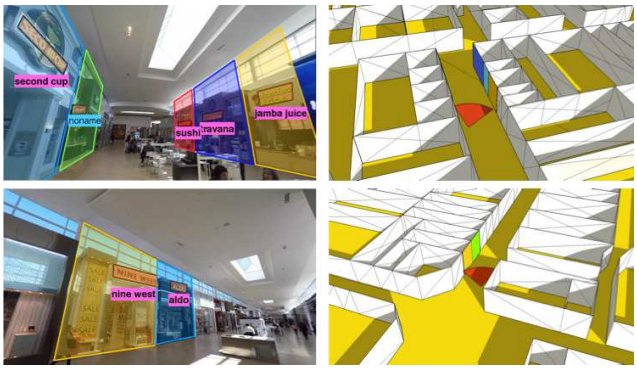


Figure 3. Ground-truth examples. **Left**: text and shop facade labeling; **right**: ground-truth camera location.

view localization [7] and story summarization [26].

Our approach is also related to work on room layout estimation [20, 32, 22] in indoor images and facade labeling in outdoor scenes [11]. We go beyond this line of work by jointly reasoning about layout and localization in large 3D spaces. A variety of approaches also make use of floor plans in indoor scenarios. These approaches typically tackle the problem of 3D reconstruction [17, 41, 27], where single or multi-camera rigs are used to record a site. In [27], indoor tourist sites are reconstructed by exploiting photos from the web, SfM and floorplans.

Perhaps the closest to our work is [25], which tries to reconstruct rental apartments from non-overlapping monocular images using the apartment's floorplan. They perform joint room layout estimation and camera localization with wall and window information. Here, we tackle localization in much larger spaces which requires a very different model formulation as well as different image cues (such as text and shop facade boundaries, *etc*.). It is also worth noting that we generalize the layout from the typical 3D box assumption to a set of facades that are oriented with the three dominant orientations. This is necessary due to the complex shape of shopping malls, see Fig. 2 for an illustration.

3. Data Collection and Overview

We collected a new dataset since there is no freely available datasets of large indoor spaces. Our aim was to get a set of geo-registered images in shopping malls with additional annotations for shop facades – facade segmentation as well as bounding boxes around text that indicates the shops' names. We visited two large shopping malls: Yorkdale and Promenade in Toronto, and collected data in two phases with time spanning from December 2014 to March 2015, in order to capture appearance, illumination, and lighting changes. For both of these malls high quality floor plans are available in an easy parseable, vector format. We recorded several video sequences with a GoPro camera mounted on the head to mimic one's field of view during shopping. In both visits, we recorded the whole shopping mall. We collected videos instead of single images in order to have a larger set of images based on which we could test robustness to viewpoint changes, etc.

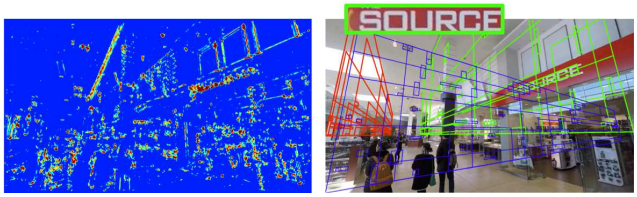


Figure 4. Our candidate text detections: We compute a saliency map (**left**) and evaluate scores for bounding boxes along directions of left, frontal, and right wall (**right**). We rectify the patches within each hypothesis to be fronto-parallel to avoid perspective distortion in the text.

We used AMT to label (tight) 2D polygons enclosing shops' names in the images, and curated the labels if needed. The shop facades were labeled by in-house annotators, who were asked to mark a quadrilateral indicating the left and right vertical boundaries of the shop as well as the bottom and top. An example of these annotations is shown in Fig. 3. Based on the corner points of the annotated facades and its corresponding corner points in the floor plan we computed ground-truth camera pose within the shopping mall. We note that due to the imperfect camera intrinsic parameters (estimated via the vanishing points) and noise in human labeling, the GT camera poses are not perfect. We manually corrected the shop's corner locations to improve the quality of our ground-truth. Fig. 2 illustrates a few image examples as well as 3D models that we estimate from the 2D correspondences. Note that resources such as wikipedia contain ceiling height information for our shopping malls.

Overview: In this paper we formulate the localization problem as inference in a Markov random field, which jointly reasons about text detection (localizing shop's names in the image with precise bounding boxes), shop facade segmentation, as well as camera's rotation and translation within the entire shopping mall. In the following we first present how to localize text in indoor scenes, followed by our holistic formulation.

4. Text Detection in Indoor Scenes

When arranging a meeting with a friend in a mall, one typically says e.g. "I'm in front of the Adidas store, come and meet me there". For humans one of the most important cues to localization inside a shopping mall is the name of the shops that surround you. Our first goal here is thus to detect store names in monocular images. The main difficulty is the presence of large perspective distortions.

Towards this goal, we extend the text detector of [21] to reason about text in a Manhattan world. [21] first runs an image through a convolutional neural network to obtain the probability of each pixel being text or background. It

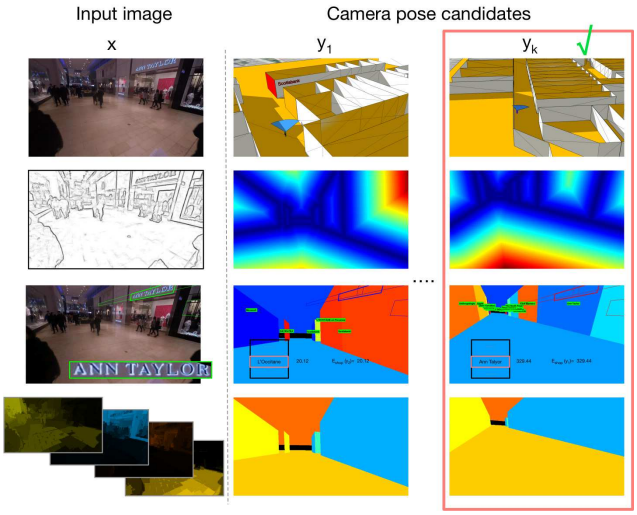


Figure 5. Our energy terms encode agreement of the camera pose and the input image (**1st row**) in terms of edges (**2nd row**), text detections, shop names (**3rd row**), and layouts (**4th row**). The best camera pose is selected among all sampled candidates.

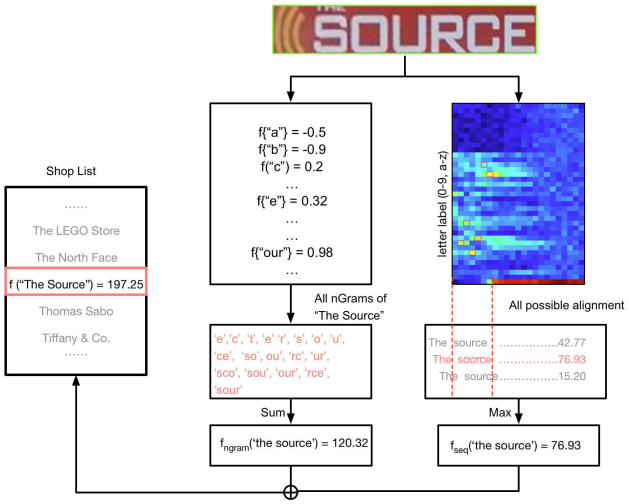


Figure 6. Shop name recognition with N-gram and sequential character classification with partial observations.

then employs a running length smoothing algorithm to detect horizontal lines, followed by an Otsu thresholding to separate the individual words. This approach works well when text is fronto-parallel or has little viewpoint distortion, but would fail in our challenging imagery.

Motivated by the observation that most text in a shopping mall is aligned with the three Manhattan directions, we incorporate vanishing directions in our text detection pipeline. We first estimate the vanishing points (VP) corresponding to three dominant orthogonal directions using [19]. We then cast rays from the three VPs and generate quadrilaterals by intersecting rays from different VPs. The box proposals are then extracted from the set of quadrilaterals on the left, right and frontal walls only, due to the fact that text very rarely appears on ceilings and floors. See Fig. 4 for an illustration. The score of each possible bounding box is computed by summing the pixel-wise convnet score for all pixels that form the quadrilateral. Note that this can be very efficiently computed via integral geometry [33]. This score is then normalized by a factor which is a function of aspect ratio of the quadrilateral. Intuitively we prefer bounding boxes with a reasonable aspect ratio (from 1:2 to 5:1) and penalize boxes outside this interval linearly.

We then rectify the sub-image inside each bounding box $\mathbf{b}$ via homography, which can be computed from the corresponding pair of vanishing points. Text classification is then conducted on the rectified image by scoring letter N-grams and sequential character classifiers of [21]. The output of the N-gram classifier is a 10000-dimensional vector indicating the probability of the presence of an N-gram without considering orders. The sequential classifier generates a 37×23 probability matrix representing the probability of the i-th letter in the box (23 in total) to have label j (0-9, a-z, null; 37 in total). For each candidate box $\mathbf{b}$ and each shop s (characterized by its name), we compute a weighted score based on the sum of these two probability matrices:

$$f_{\text{shop}}(s, \mathbf{b}) = f_{\text{seq}}(s, \mathbf{b}) + \lambda f_{\text{ngram}}(s, \mathbf{b})$$

where the sequential score is:

$$f_{\text{seq}}(s, \mathbf{b}) = \begin{cases} \max_j \sum_{i=1}^{|s|} p_{\text{seq}}(i+j, s(i)), & \text{if } |s| \leq |\mathbf{b}| \\ \max_j \sum_{i=1}^{|\mathbf{b}|} p_{\text{seq}}(i, s(i+j)) & \text{if } |\mathbf{b}| < |s| \end{cases}$$

This score represents the highest matching score between the shop's name and the bounding box considering incomplete observations, where $|\mathbf{b}|$ is the maximum length of a word in the candidate detection $\mathbf{b}$ and $|s|$ is the length of the shop's name. Given the sequential probability matrix [21] of a candidate detection, its maximum length $|\mathbf{b}|$ is decided by the number of characters with the highest classification score excluding the character class 'null'. For example, as shown in Figure 7, the maximum length of the candidate detection is seven, since the rest of the characters have highest score on the "null" class (the last row).

The N-gram score is defined as:

$$f_{\text{ngram}}(s, \mathbf{b}) = \sum_{g \in \mathcal{G}_s} p_{\text{ngram}}(g, \mathbf{b}),$$

where $\mathcal{G}_s$ is the N-gram set of the s shop's name. For instance, {'g', 'a', 'p', 'ga', 'ap', 'gap'} is the N-gram set for the shop "gap". We refer the reader to Fig. 6 for a visualization of our scoring function.

5. Indoor Self-localization

We are interested in performing self-localization in large indoor environments given a single image and a floor plan. Towards this goal, we exploit both geometric and semantic

cues and frame the problem as the one of inference in a Markov random field (MRF). We assume that the world is Manhattan and exploit the relationships between the layout and localization problems. This requires us to generalize the traditional layout representation from a 3D box [20, 33] to a set of facades, each of which is aligned with one of the three dominant orientations. This is due to the fact that the floor plans contain non-concave regions.

Since we have access to the mall's floor plan, then for a candidate camera location and orientation, the number of facades visible in the image is known and can be estimated by rendering. We can then score a candidate camera based on the agreement between the projected facades and several image cues. One of the most important sources of information is the fact that for most shops its name (or at least part of it) appears in the shop's facade (anywhere inside the facade). Our scoring function will exploit the relevant text detections and their classification scores. Our model also scores surface normals estimated from an image and how they agree with our camera (facade) hypothesis. Additionally, we exploit the fact that some edges in the image should correspond to facade boundaries between the different shops. We refer the reader to Fig. 5 for an illustration.

5.1. Problem Formulation

Given a single image $\mathbf{x}$ and a floorplan $\mathcal{M}$, we parameterize localization with a tuple $\mathbf{y} = \{\mathbf{t}, \mathbf{R}\}$, where $\mathbf{t} \in \mathbb{R}^3$ is the 3D camera center and $\mathbf{R} \in \mathcal{SO}(3)$ is the rotation matrix, as shown in Fig. 5. Given the estimated vanishing points, we can recover the camera intrinsic parameters as well as the camera rotation up to a flipping ambiguity along the horizontal direction. This ambiguity reflects which side of the corridor represents the left wall (see Fig. 7 for an illustration). As a consequence, the localization problem can be formulated with 3 degrees of freedom representing the translation as well as binary variable for the flip.

A key component necessary to compute many of our energy terms is the ability to project the shop facades into the image, taking into account occlusions. This is very simple if the layout is a cuboid, but it is more complex in real-world scenarios such as shopping malls where the layout contains several non-concave regions. Given a camera pose $\mathbf{y}$ and our floor plan $\mathcal{M}$, we can however employ rendering to compute the visible part of each shop facade. We conduct frustum culling and look-up-table based depth ordering to ensure a correct projection. This process is fairly efficient as it runs at 550 images per second. The output is a set of semantic 2D quadrilaterals visible in the image $\mathcal{Q}(\mathbf{y}, \mathcal{M}) = \{\mathbf{q}_i, s_i, \ell_i\}$ where $\mathbf{q}_i$ is the quadrilateral, s_i is the corresponding shop (*e.g.* Gap) and ℓ_i corresponds to the wall label (left/right/ceiling/floor).

Given the floor plan and a camera pose hypothesis, our energy scores the agreement between the facades and the surface normals estimated from the image, the agreement between the wireframe of visible shop facades and the image edges as well as the containment of text detection and text classification within the correct shop. Thus

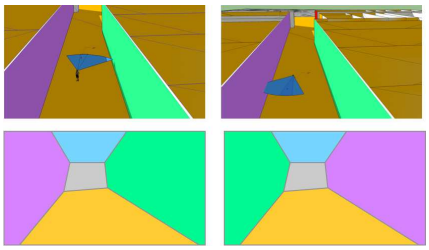


Figure 7. The left-right wall ambiguity comes from the fact that there exist two rotation matrices that agree with the vanishing points extracted from an image.

$$E(\mathbf{y};\mathbf{x},\mathcal{M}) = \begin{array}{l} E_{\text{layout}}(\mathbf{y};\mathbf{x},\mathcal{M}) + E_{\text{edge}}(\mathbf{y};\mathbf{x},\mathcal{M}) \\ +E_{\text{det}}(\mathbf{y};\mathbf{x},\mathcal{M}) + E_{\text{shop}}(\mathbf{y};\mathbf{x},\mathcal{M}) \end{array} \quad (1)$$

We next describe the potentials we employ in more detail.

Facade Surface Normals: Following recent work on room layout estimation [20, 33], we utilize geometric context (GC) [20] and orientation maps (OM) [22] as image features. This provides a 10-dimensional feature vector per pixel. We define the energy as the weighted sum of features in each facade. This encodes the agreement between the facades and the image evidence. Thus

$$E_{\text{facade}}(\mathbf{y}) = \sum_{i \in \mathcal{Q}(\mathbf{y},\mathcal{M})} \mathbf{w}_i^T \phi_{\text{nor},i}(\mathbf{q}_i, \ell_i) \quad (2)$$

and $\phi_{\text{nor}}(\mathbf{q}_i, \ell_i) = [\sum_{p \in \mathbf{q}_i} \phi_{\text{om}}(p); \sum_{p \in \mathbf{q}_i} \phi_{\text{gc}}(p)]$ is the sum of all features in i-th quadrilateral. The weights $\mathbf{w}_i$ are a set of $+1$ and -1 and encode the agreement. Fig. 5 illustrates the potentials. Note that since the facades are defined along Manhattan directions, integral geometry [33] can be used to compute these potentials in constant time.

Edge: This energy term encodes the fact that many image edges correspond to shop facades or wall junctions. We define the energy as the sum of minimum distances between the shop wireframe and the image edges

$$E_{\text{edge}}(\mathbf{y}) = w_{\text{edge}} \sum_{p \in \mathcal{E}} \min_{i \in \mathcal{Q}(\mathbf{y},\mathcal{M})} \text{dist}(p, \mathbf{q}_i), \quad (3)$$

where $\mathcal{E}$ is the set of all edge pixels extracted via the edge detector of [12] and $\text{dist}(p, \mathbf{q}_i)$ is the distance from a pixel to a polygon, which is defined to be the distance from a pixel to its nearest line segment on the polygon. This term can be efficiently calculated using the distance transform.

Detection: Motivated by the fact that the shop name usually appears inside the shop facade, we penalize text detections that cross shop boundaries

$$E_{\text{det}}(\mathbf{y}) = -w_{\text{det}} \sum_{\mathbf{b} \in \mathcal{B}} f_{\text{det}}(\mathbf{b}) \cdot \max_{i \in \mathcal{Q}(\mathbf{y},\mathcal{M})} IOB(\mathbf{b}, \mathbf{q}_i) \quad (4)$$

with $f_{\text{det}}(\mathbf{b})$ the score of the detection, $\mathcal{B}$ the set of all detected bounding boxes, and $IOB(\mathbf{b}, \mathbf{q}_i)$ the intersection of the bounding box and the i-th facade divided by the box size. Note that if a text bounding box $\mathbf{b}$ fully belongs to one facade, its corresponding energy is equal to 1, otherwise is proportional to its largest overlap with a facade. This energy evaluation can be computed very efficiently using an integral image for each bounding box.

Shop: This energy encodes consistency between the camera pose and the shop name classification (within the text bounding boxes). Similarly to the detection energy:

$$E_{\text{shop}}(\mathbf{y}) = -w_{\text{shop}} \sum_{\mathbf{b}\in\mathcal{B}} \sum_{i\in\mathcal{Q}(\mathbf{y},\mathcal{M})} f_{\text{shop}}(s_i, \mathbf{b}) \cdot IOB(\mathbf{b}, \mathbf{q}_i) \tag{5}$$

with $f_{\text{shop}}(s_i, \mathbf{b})$ the score of the text classifier for word (shop's name) s_i inside box $\mathbf{b}$. This term can be efficiently computed via integral images for each shop.

5.2. Inference

Since we use egocentric images, we can further assume a fixed height-above-ground position, and reduce the degrees of freedom of camera's translation from 3D to 2D. It is worth noting that our energy evaluation can be conducted very efficiently for a candidate camera pose, since all the energy potentials can be computed via distance transforms or accessing integral image/geometry accumulators. On average we can evaluate 400 candidate camera poses per second on a 16-core CPU, including both rendering and energy computation. We discretize the search space with a 1-meter step and conduct exhaustive search over the walkable corridors of the shopping mall. This yields 10200×2 camera poses to be explored, which makes the inference stage take 50 seconds. We can then compute the energy for each possible camera pose, sort them and take a list of top-k cameras. Computing all features takes around 100s, making the whole process run 3min for each image using Matlab. Note that simple pruning strategies can be utilized to reduce the search space significantly, *e.g.* only considering regions around the most confident detections of shops.

Since we prefer diversity in the top-k solutions, we do not sort the proposals merely according to the energy score, but instead employ the diverse k-best method of [6]. After obtaining the current best solution, we reweigh the energy according to the distance from the remaining points to this solution. In this way, the next immediate solution tends to be different from the previous one.

6. Experiments

In this section we report our experimental evaluation in our shopping mall dataset. We utilize 246 fully-labeled images to test our localization results, covering over 130 shops. We employ cross-validation to set all the weights. An additional validation set is used to choose models and hyper-parameters for text classification and detection. The test and validation sets are collected 4 months apart to demonstrate our ability to generalize.

	Detection		End-to-end	
	F-measure	AP	F-measure	AP
$\tau > 300$	54.05%	48.07%	44.58%	33.39%
$\tau > 900$	61.25%	56.85%	50.61%	40.30%
$\tau > 1600$	70.74%	67.09%	61.08%	50.10%

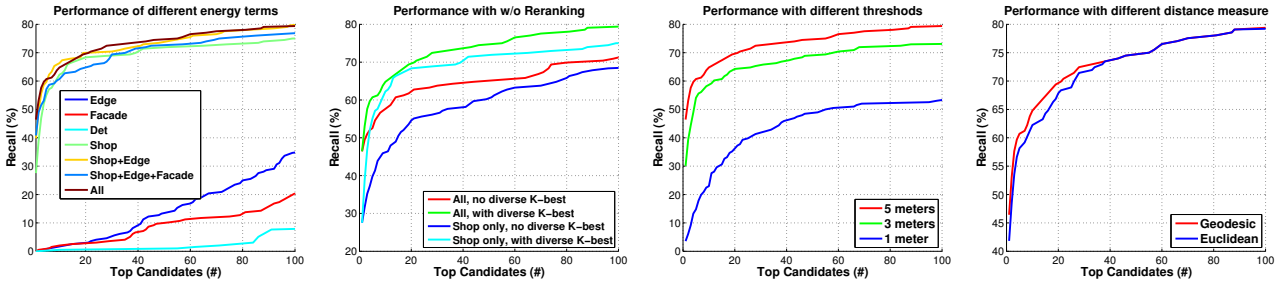
Table 3. Performance on text detection and end-to-end system τ represents the minimum size of the ground truth text box.

We adopt recall@k as our localization performance measure. We consider a localization hypothesis to be correct if its distance to the ground-truth location is smaller than a threshold. Two different distance measures are adopted: the euclidean distance as well as the geodesic distance over the corridor region. We evaluate our algorithm under three different thresholds, i.e., 1, 3 and 5 meters. We argue that 5 meters is still a reasonable threshold for this task, which is comparable with localization accuracy of most sensor-based methods. We also report the performance in terms of the median rank. This represents the median number of candidates required in order to get at least one successful location. Thus, the lower the median rank the better.

Importance of features: As shown in Tab. 2, shop energy has the highest performance amongst all features. Without semantic information from shop names, neither edge, facade nor detection can accurate localize. This is due to the fact that the structure of this type of man-made large indoor spaces is highly repetitive. Among these features, edges are most informative in terms of ranking and recall@k. However, when using all the features the accuracy improves significantly. Our holistic approach outperforms the shop-energy alone by more than 70% in terms of recall@1.

Diverse K-best: Fig. 8(b) shows that k-diverse re-ranking helps improve performance by 20% in terms of recall@5 and recall@10. This improvement becomes even larger with more than top-20 candidates. We refer the reader to Tab. 2 for an in depth comparison.

Text spotting: Tab. 3 and Fig. 9 show the performance of our text detection and classification algorithm described in section 4. Note that we evaluate the detector as well as our end-to-end pipeline. A detection hypothesis is considered to be a true positive if the intersection-over-union with a ground-truth box is over 0.3. We use a low threshold as it is a very difficult task. An end-to-end hypothesis is considered to be a true positive if the IoU is over 0.3 with one ground-truth and the top-1 shop name prediction is correct. For each ground-truth with multiple overlapping hypotheses, we only

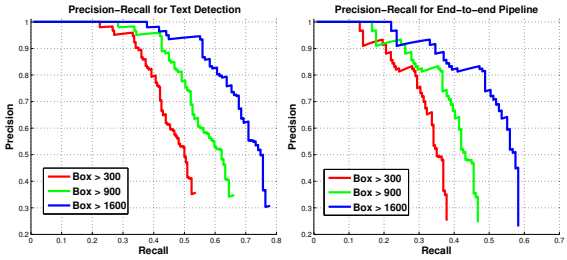


(a) Different energies (b) Reranking (c) Different threshold (d) Different distance

Figure 8. Recall@k curve under multiple configurations.

	With diverse k-best					No diverse k-best				
	Recall@k				Median rank	Recall@k				Median rank
Model	k=1	k=5	k=10	k=50	Median rank	k=1	k=5	k=10	k=50	Median rank
Random	-	-	-	-	-	0.00%	0.25%	3.16%	6.02%	245.0
Edge	0.00%	1.02%	3.02%	16.84%	172.0	0.51%	1.02%	1.02%	5.61%	614.5
Facade	0.00%	2.04%	3.06%	9.69%	246.5	0.00%	0.51%	1.53%	3.57%	779.5
Det	0.51%	5.61%	7.14%	31.12%	89.5	0.51%	1.53%	3.57%	22.96%	114.5
Shop	27.55%	55.10%	62.24%	71.43%	4.0	27.55%	39.80%	45.92%	60.20%	15.0
Shop+Det	29.08%	55.61%	62.24%	72.96%	4.0	29.08%	39.80%	45.92%	60.71%	15.0
Shop+Facade	35.20%	55.61%	59.69%	71.43%	3.0	35.20%	45.41%	53.57%	61.73%	8.0
Shop+Edge	39.80%	59.69%	**66.33%**	73.47%	2.5	39.80%	50.00%	55.61%	64.80%	5.0
Shop+Facade+Edge	40.82%	57.14%	60.71%	72.45%	3.0	40.82%	49.49%	54.08%	61.73%	5.5
All	**46.43%**	**60.71%**	64.80%	**74.49%**	**2.0**	**46.43%**	**52.55%**	**57.65%**	**64.80%**	**2.5**

Table 2. Quantitative performance under different configurations. (w/o diverse k-best)



(a) Detection (b) End-to-End

Figure 9. Precision-recall curve for text spotting.

consider the one with largest IOU as a true positive. We also consider three difficulty levels by choosing bounding box sizes (larger than 300 pixels, 900 pixels and 1600 pixels, respectively). If a text bounding box is smaller than 300 pixels (10×30), it is almost impossible to recognize it even for a human.

Qualitative results: We show multiple successful localization results in Fig. 10 which depicts an input image with GT labels, rendering with GT position, localization results in the floorplan, and rendering with our algorithm's output. Our method achieves very accurate localization results for images where texts are visible. Moreover, from the rendering results we can see that our localization results also match the geometric features well in terms of shop facades and layout. It is also worth noting that our method can even handle some non-manhattan cases as shown in the last example of Fig. 10. This is also due to the fact that our proposed approach does not rely on the 3D box assumption typical for the previous indoor (room layout) approaches.

Limitations: Our approach has several limitations. First, it is highly dependent on text detection and shop name recognition. We can hardly handle the case where there is no text in the scene or severe mis-classification errors (see Fig. 11). This is the reason for most failures cases. Furthermore, our approach assumes that the 3D floorplan is correct. In the last example of Fig. 11, the 'facade' of the kiosk is transparent and the actual height of the kiosk is lower than the ceiling height of the 3D floorplan, which makes our method fail.

7. Conclusions

We have presented an approach to localization in very large indoor spaces composed of shopping malls with 200+ stores. We formulate the problem as inference in a Markov random field, which jointly reasons about text detection (localizing shop's names in the image with precise bounding boxes), shop facade segmentation, as well as camera's rotation and translation within the entire shopping mall. Importantly, our approach only requires a single image and a floor plan as input and is able to localize accurately. In the future we plan to collect data and test our approach in other large indoor spaces such as airports and train stations. We also plan to investigate other cues, *e.g.* objects, video and logos to improve localization performance.

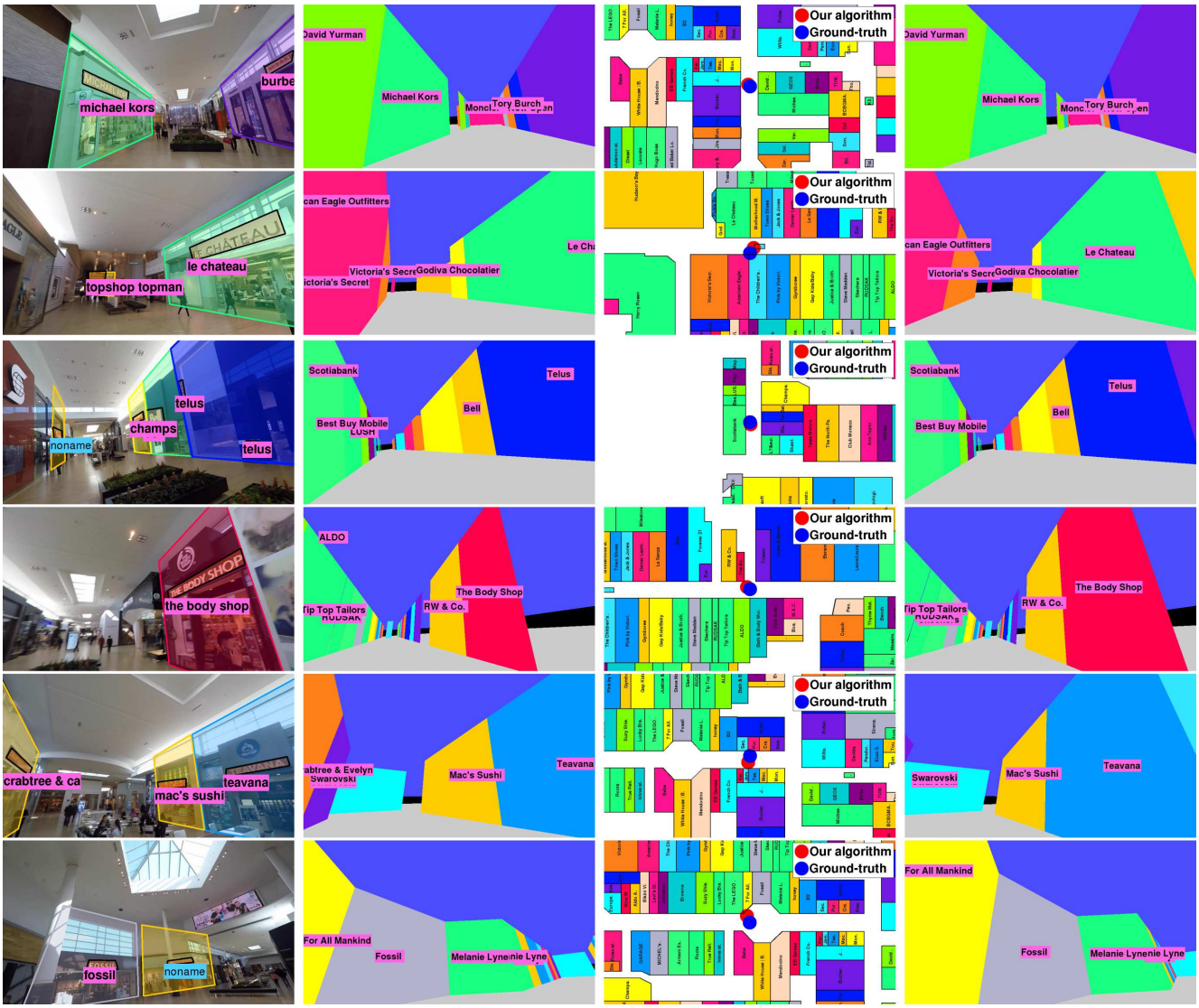


Figure 10. Visualization of success cases of our method. From left to right: input image with ground-truth labeling; rendering with ground-truth position; localization results in floorplan; rendering with our method's output.

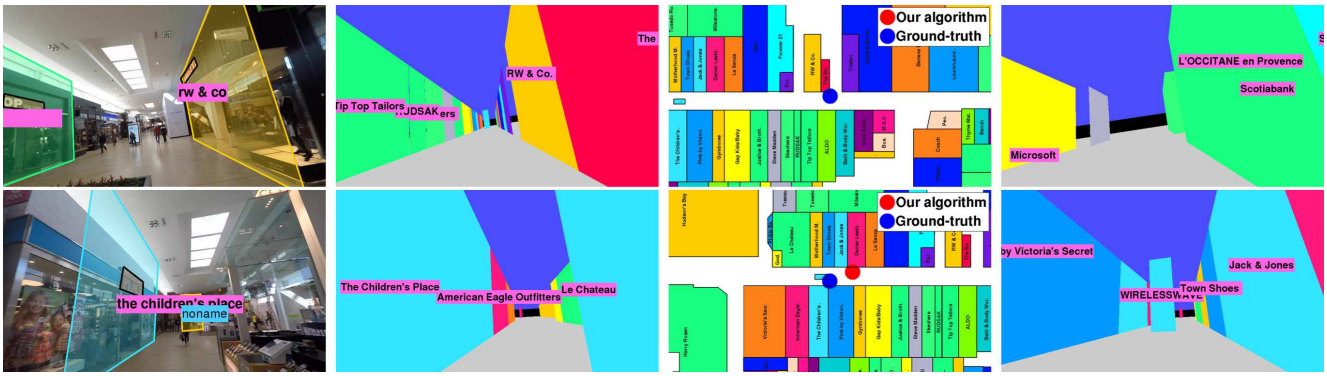


Figure 11. Failure cases of our method. From top to bottom: failed text classification due to extreme perspective distortion; wrong facade features due to transparent kiosk facade.

References

[1] Aislelabs. https://www.aislelabs.com/. 1

[2] Indoors. http://indoo.rs/. 1

[3] S. Agarwal, Y. Furukawa, N. Snavely, I. Simon, B. Curless, S. M. Seitz, and R. Szeliski. Building rome in a day. *Communications of the ACM*, 2011. 1, 2

[4] N. Atanasov, M. Zhu, K.Daniilidis, and G. Pappas. Semantic localization via the matrix permanent. In *RSS*, 2014. 2

[5] M. Bansal and K. Daniilidis. Geometric urban geo-localization. In *CVPR*, 2014. 1, 2

[6] D. Batra, P. Yadollahpour, A. Guzman-Rivera, and G. Shakhnarovich. Diverse m-best solutions in markov random fields. In *ECCV*. 2012. 6

[7] V. Bettadapura, I. Essa, and C. Pantofaru. Egocentric field-of-view localization using first-person point-of-view devices. In *WACV*, 2015. 2, 3

[8] M. A. Brubaker, A. Geiger, and R. Urtasun. Lost! leveraging the crowd for probabilistic visual self-localization. In *CVPR*, 2013. 1, 2

[9] S. Cao and N. Snavely. Graph-based discriminative learning for location recognition. In *CVPR*, 2013. 1

[10] N. Chang, R. Rashidzadeh, and M. Ahmadi. Robust indoor positioning using differential wi-fi access points. *Consumer Electronics, IEEE Trans. on*, 2010. 1, 2

[11] D. Dai, M. Prasad, G. Schmitt, and L. V. Gool. Learning domain knowledge for facade labelling. In *ECCV*, pages 710–723, 2012. 3

[12] P. Dollár and C. L. Zitnick. Structured forests for fast edge detection. In *ICCV*, 2013. 5

[13] J. Engel, T. Schöps, and D. Cremers. Lsd-slam: Large-scale direct monocular slam. In *ECCV*. 2014. 1, 2

[14] A. Fathi, Y. Li, and J. M. Rehg. Learning to recognize daily actions using gaze. In *ECCV*, 2012. 2

[15] G. Floros, B. van der Zander, and B. Leibe. Openstreetslam: Global vehicle localization using openstreetmaps. In *ICRA*, 2013. 2

[16] C. Forster, M. Pizzoli, and D. Scaramuzza. Svo: Fast semi-direct monocular visual odometry. In *ICRA*, 2014. 2

[17] Y. Furukawa, B. Curless, S. M. Seitz, and R. Szeliski. Reconstructing Building Interiors from Images. In *ICCV*, 2009. 3

[18] J. Hays and A. A. Efros. Im2gps: estimating geographic information from a single image. In *CVPR*, 2008. 1, 2

[19] V. Hedau, D. Hoiem, and D. Forsyth. Recovering the spatial layout of cluttered rooms. In *ICCV*, 2009. 4

[20] V. Hedau, D. Hoiem, and D. Forsyth. Thinking inside the box: Using appearance models and context based on room geometry. In *ECCV*, 2010. 3, 5

[21] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman. Deep structured output learning for unconstrained text recognition. *arXiv preprint arXiv:1412.5903*, 2014. 3, 4

[22] D. C. Lee, M. Hebert, and T. Kanade. Geometric Reasoning for Single Image Structure Recovery. In *CVPR*, 2009. 3, 5

[23] Y. Li, N. Snavely, and D. P. Huttenlocher. Location recognition using prioritized feature matching. In *ECCV*. 2010. 1, 2

[24] J. Liang, N. Corso, E. Turner, and A. Zakhor. Image based localization in indoor environments. In *Comp. for Geosp. Research and Appl.*, 2013. 2

[25] C. Liu, A. Schwing, K. Kundu, R. Urtasun, and S. Fidler. Rent3d: Floor-plan priors for monocular layout estimation. In *CVPR*, 2015. 3

[26] Z. Lu and K. Grauman. Story-driven summarization for egocentric video. In *CVPR*, 2013. 3

[27] R. Martin-Brualla, Y. He, B. C. Russell, and S. M. Seitz. The 3D Jigsaw Puzzle: Mapping Large Indoor Spaces. In *ECCV*, 2014. 3

[28] A. Quattoni and A. Torralba. Recognizing indoor scenes. In *CVPR*, 2009. 2

[29] N. Ravi, P. Shankar, A. Frankel, A. Elgammal, and L. Iftode. Indoor localization using camera phones. In *WMCSA*, 2006. 2

[30] T. Sattler, B. Leibe, and L. Kobbelt. Fast image-based localization using direct 2d-to-3d matching. In *ICCV*, 2011. 1

[31] G. Schindler, M. Brown, and R. Szeliski. City-scale location recognition. In *CVPR*, 2007. 1

[32] A. Schwing, S. Fidler, M. Pollefeys, and R. Urtasun. Box in the box: Joint 3d layout and object reasoning from single images. In *ICCV*, 2013. 3

[33] A. Schwing, T. Hazan, and R. Urtasun. Efficient structured prediction for 3d indoor scene understanding. In *ECCV*, 2012. 4, 5

[34] N. Snavely, S. M. Seitz, and R. Szeliski. Modeling the world from internet photo collections. *IJCV*, 2008. 1, 2

[35] R. Templeman, M. Korayem, D. Crandall, and A. Kapadia. Placeavoider: Steering first-person cameras away from sensitive spaces. In *Network and Distributed System Security Symposium (NDSS)*, 2014. 2

[36] A. Torii, J. Sivic, T. Pajdla, and M. Okutomi. Visual place recognition with repetitive structures. In *CVPR*, 2013. 2

[37] S. Treuillet and E. Royer. Outdoor/indoor vision based localization for blind pedestrian navigation assistance. *Intl. J. of Image and Graphics*, 10(4), 2010. 2

[38] S. Wang, S. Fidler, and R. Urtasun. Holistic 3d scene understanding from a single geo-tagged image. In *CVPR*, 2015. 2

[39] O. Woodman and R. Harle. Pedestrian localisation for indoor environments. In *International Conference on Ubiquitous Computing*, 2008. 1, 2

[40] O. Woodman and R. Harle. Pedestrian localisation for indoor environments. In *International Conference on Ubiquitous Computing*, 2008. 2

[41] J. Xiao and Y. Furukawa. Reconstructing the World's Museums. In *ECCV*, 2012. 3

[42] A. R. Zamir and M. Shah. Accurate image localization based on google maps street view. In *ECCV*. 2010. 1, 2