Visual Landmark Sequence-based Indoor Localization

Qing Li
School of Computer Science,
University of Nottingham
Nottingham, UK
psxql2@nottingham.ac.uk

Jiasong Zhu
College of Civil Engineering,
Shenzhen University
Shenzhen, China
jiasong.zhu@szu.edu.cn

Tao Liu
School of Geodesy and Geomatics,
Wuhan University
Wuhan, China
liuzimo@whu.edu.cn

Jon Garibaldi
School of Computer Science,
University of Nottingham
Nottingham, UK
jon.garibaldi@nottingham.ac.uk

Qingquan Li
College of Civil Engineering,
Shenzhen University
Shenzhen, China
qingquan.li@szu.edu.cn

Guoping Qiu
Shenzhen University, China
qiu@szu. edu. cn
University of Nottingham, UK
guoping.qiu@nottingham.ac.uk

ABSTRACT

This paper presents a method that uses common objects as landmarks for smartphone-based indoor localization and navigation. First, a topological map marking relative positions of common objects such as doors, stairs and toilets is generated from floor plan. Second, a computer vision technique employing the latest deep learning technology has been developed for detecting common indoor objects from videos captured by smartphone. Third, second order Hidden Markov model is applied to match detected indoor landmark sequence to topological map. We use videos captured by users holding smartphones and walking through corridors of an office building to evaluate our method. The experiment shows that computer vision technique is able to accurately and reliably detect 10 classes of common indoor objects and that second order hidden Markov model can reliably match the detected landmark sequence with the topological map. This work demonstrates that computer vision and machine learning techniques can play a very useful role in developing smartphone-based indoor positioning applications.

CCS CONCEPTS

•Information systems →Mobile information processing systems;

KEYWORDS

Indoor landmark; Localization and navigation; Convolutional neural network (CNN); Second order hidden Markov model

ACM Reference format:
Qing Li, Jiasong Zhu, Tao Liu, Jon Garibaldi, Qingquan Li, and Guoping Qiu. 2017. Visual Landmark Sequence-based Indoor Localization. In *Proceedings of GeoAI'17:1st ACM SIGSPATIAL Workshop on Artificial Intelligence and Deep Learning for Geographic Knowledge Discovery , Los Angeles Area, CA, USA, November 7–10, 2017 (GeoAI'17),* 11 pages.
DOI: 10.1145/3149808.3149812

1 INTRODUCTION

Location based services can benefit consumers greatly and have witnessed rapid development in the past decade. However, these services mainly focus on outdoors because there is still a lack of a robust indoor positioning technology. GPS does not function in indoor environments because its signal is weakened or even blocked by the wall or the surface of the building. How to obtain location information quickly and reliably still remains a big challenge for indoor localization based service.

There have been attempts to use smartphone camera for indoor localization [16, 19, 21, 27]. These methods exploit computer vision techniques to estimate people's location. They fall into two categories: methods based on image retrieval and methods based on structure from motion. The former uses images captured by smartphone camera to search for similar images in the image dataset whose position and orientation are already known. It requires huge offline efforts and is easy to get stuck in scene ambiguity in which similar scenes appear in different locations. The latter estimates location by solving vision geometry but does not work in the low texture environment.

In this paper, we propose an image-based approach to obtaining location information by recognizing landmarks in the indoor environment. Instead of directly comparing images, landmark sequence are extracted from video sequence and match landmarks with a topological map generated from floor plan map. The idea behind this is that for indoor way-finding scenario in an unfamiliar environment, precise localization might not be needed and coarse localization is enough to navigate people to their destinations. Since landmark sequence is used to map to a topological map, geo-tagged image database is avoided and much offline labour cost is saved. Also, our method does not require indoor environment to have highly textured surface.

To achieve it, a landmark detection approach based on convolutional neural network (CNN) is developed which aims to extract landmark sequence from a smartphone video, and second order hidden Markov model is used to match the detected landmark sequence to landmarks on the topological map. The main contributions of the paper are two-folds:

(1) A novel landmark detector that is able to recognize two distinctive types of indoor landmarks, indoor objects and

GeoAI'17, Los Angeles Area, CA, USA
© 2017 ACM. 978-1-4503-5498-1/17/11…$15.00
DOI: 10.1145/3149808.3149812

indoor scenes (intersections and corners) in a unified CNN network.

(2) A landmark matching algorithm based on a second order hidden Markov model is developed. It solves scene ambiguity problem where traditional methods have failed.

2 RELATED WORKS

The proposed method is mainly consisted of two processes, landmark detection and localization. We briefly review related works in these two areas.

2.1 Visual Landmark Detection

Landmarks can be divided into two categories: natural and artificial. Artificial landmarks are usually designed to tackle challenges of varying illuminations, view points and scales. They have many advantages. They are easily and precisely detected since they are designed based on known rules. Those rules not only aid handling variation of objects in the image but also act as guidance to develop detection strategies. However, the drawbacks are difficult to overcome. Deploying those landmarks changes building decoration which might not be feasible due to economic or owners' tastes, resulting in limitation in its application. Natural landmarks avoid changing indoor infrastructure by exploiting physical objects or scenes of the indoor environment. Common objects like doors, elevators, fire extinguishers and interesting locations like corners and turns are able to act as landmarks. They usually remain unchanged in a relatively long period and are able to be seen frequently in the indoor environment. The main challenge of natural indoor landmark detection is that an accurate and robust method is lacking. Many techniques based on handcrafted features derived from their color gradient or geometric information. [12] viewed planar and quadrangular objects as landmarks and detected them based on geometric shape. [6] viewed along-path objects as landmarks and they are used for localization. [11] proposed a landmark-based algorithm in which landmarks are represented with a set of SURF features. [1] used SIFT features to compare a landmark with landmarks pre-stored in the database. [24] presented a localization and navigation system based on landmarks, they exploited doors, stairs and tags in the environment as landmarks. SURF features and lines are used to recognize those landmarks. [25] presented localization methods based on indoor objects like doors, elevators, and cabinets based on geometric shape using edges and corners. Their approach provided high performance when dealing with certain landmarks. Current approaches either focus on certain types of object or fail to work with landmarks whose background are of high texture information. Besides, few researchers have tried to detect landmarks that are made up of corners and turns.

In this paper, we view landmark detection as a classification problem. Unlike previous approaches that recognize indoor objects relying on handcrafted features, deep learning neural network is chosen to recognize indoor objects and interesting locations at the same time. CNN have proved its high performance in classification[9] and indoor scene recognition[28] and outperformed approaches based on handcrafted feature.

2.2 Localization

Many positioning algorithms have introduced landmark information for indoor localization. Basically, landmarks are taken as supporting information to reduce the error drift of dead reckon approaches [7, 8, 10]. In this paper, we focus on topological localization with visual landmarks.

Many approaches performed landmark-based localization under geometric scheme. [4] have proposed a method using more than 3 landmarks for localization to estimate the user's position by applying geometric triangulation. [13] proposed a localizing algorithm based on a single landmark and the accurate position was estimated based on an affine camera model between 3-dimentional space and projected image space.

Another localization scheme is based on landmark retrieval. A detected landmark is matched to landmarks on a topological map and the location is assigned with the location of most similar one on the map. Many visual representations of landmarks are directly used to perform the match. [16] used omnidirectional panoramic images taken in different positions to represent landmarks and PCA-SIFT was applied to perform image matching. [2] developed a landmark-based navigation system using QR codes as landmarks and user's location determined and navigated by recognizing quick response code registered in the landmark's location. Other presentations are developed based on prior color distribution[13], shape[3, 30], light strength[22] or region connection relations[20]. However, in indoor environment, it is sometimes not feasible to match landmarks just based on visual feature due to duplication of objects and structure. [25] exploited text information around doors to handle this challenge. However, it is not able to apply to other indoor objects without tags around them. [5, 17, 29] exploited contextual information using hidden Markov model (HMM) to recognize landmarks and achieved good result. Common HMM model fails in situations of high ambiguity because it only considers current landmark to recognize the next landmark. In this paper, we develop a second order hidden Markov model to match landmarks to the map. It considers previous two landmarks when match current one. In this manner, more contextual information is taken into account to recognizing the landmarks and indoor scene ambiguity is greatly reduced.

3 LANDMARK DETECTION USING CONVOLUTIONAL NEURAL NETWORK

3.1 Landmark Definition

Indoor objects like doors, fire extinguishers, and stairs can be used as landmarks, some examples are shown in Figure 1. In this paper, three types of landmarks are defined: single object landmarks, multiple object landmarks, and scene landmarks. Single object landmarks consist of one object like a fire extinguisher or an elevator. Multiple object landmarks combine more than one objects together to identify a single location. Examples of this type include a door object and a tag object on the door. Multiple objects together can enhance the landmarks uniqueness and reduce ambiguity. Scene landmarks are key locations of the indoor structure such as corners, intersections or halls.

These three types of landmarks are detected relying on the indoor object and scene recognition result. The process is divided into two

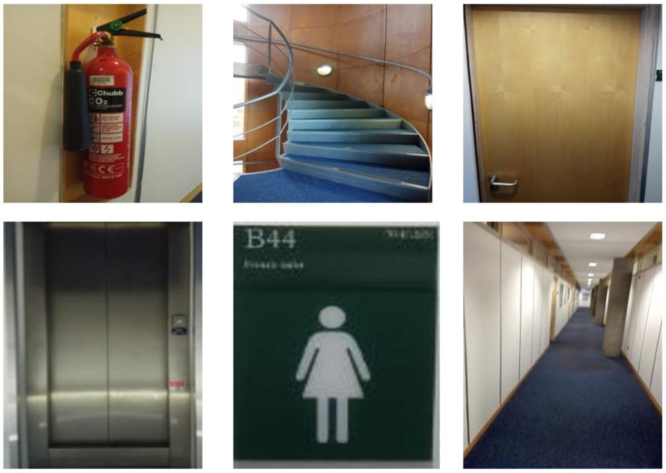

Figure 1: Common indoor objects and interesting locations

phases: offline and online phase. During offline phase, a CNN network is trained to recognize indoor objects and indoor scenes. In online phase, images are extracted from video sequence first, then region proposal algorithm is used to generate image patches containing the indoor objects, and finally the image patches are inputted to a trained CNN network to recognize objects. Landmark types are determined with indoor object recognition result. Figure 2 illustrates the process.

3.2 Convolutional Neural Networks for Object Detection

Recent years has seen the great success of convolutional neural networks in computer vision tasks including object recognition and classification. Real-time performance has been achieved in object detection with high accuracy [23]. In our landmark detection application, we retrain AlexNet[18] to recognize indoor objects and scenes. There are 5 convolutional layers and each is tailed by a max pooling layer in it. Two full convolutional layers are used to concentrate on global features after convolutional layers. The input layer takes image pixel as input. The output layer provides the probability of each predefined classes that input image belongs. Therefore, the number of neurons of the output layer is the same as the number of classes to classify. AlexNet is selected for two reasons. The first is its high performance in image classification. Secondly, it has relatively fewer layers and thus is computationally efficient. AlexNet is originally designed for ImageNet competition, which aims to recognize 1000 types of objects. However, not all indoor objects are included. Therefore, network architecture needs to be modified to adapt to landmark detection. Pre-trained CNN network is capable to extract key image information. We directly use their weights and only learn output layer for landmark detection. The number of neurons in output layer is the same as the indoor object and scene classes. Softmax function is chosen as the activation function of output layer neurons.

3.2.1 Frame Extraction. In the online phase, the first step is video frame extraction. For image extraction, sampling rate is vital for landmark detection accuracy and efficiency. If the rate is set very low, successive images have low overlap or even have no overlap at all. This can make an object not be completely seen in an image or some objects appear in the images thus missing some landmarks. High sampling rate leads to information redundancy, resulting in low landmark detection efficiency. Empirically, overlap between two successive images should be over 90%, in order to avoid missing landmarks. Overlap can be roughly estimated using equations 1 and 2. They are applied in two scenarios: walking along a line and turning to another direction.

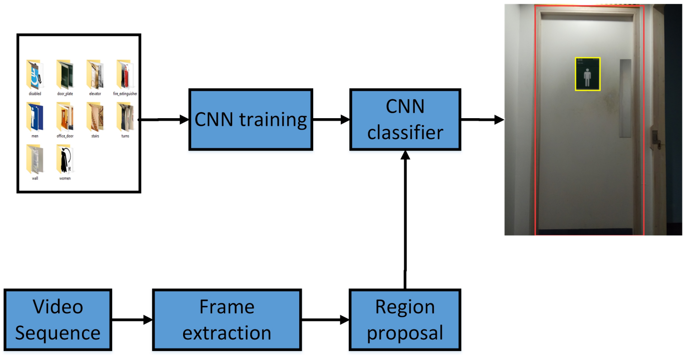

Figure 2: Flowchart of indoor landmark detection

$$Overlap = 1 - \frac{V}{2H\tan(\frac{\theta}{2})Hz} \times 100\% \tag{1}$$

$$Overlap = 1 - \frac{V_{ang}}{Hz\theta} \times 100\% \tag{2}$$

where V represents walking speed and H is the average distance between camera and surrounding environment. θ is the field of view of camera in each mobile phone. Hz represents sampling rate. V_{ang} is the angular velocity.

Generally speaking, humans walking speed is about 1. 4-2 m/s and turning 90°in 0. 8 second. In order to achieve over 90% overlap, empirically 3-5 frames per second would satisfy the requirement.

3.2.2 Region Proposal. Frequently more than one indoor entities (such as floor, chairs, tables, etc), whether they are of interest or not, appear in the images captured in an indoor environment. Their appearances affect the performance of recognizing targets. Objects are usually detected depending on their color and texture. Appearance of distracting objects decreases recognition accuracy. Therefore, finding a region that only contains an interesting target can greatly increase object detection accuracy. Instead of generating patches based on object salience[15], here we choose a selective search algorithm to generate interesting regions from images[26]. The process contains two steps. At first, an over-segmentation algorithm is chosen to generate massive initial regions in a variety of color space with a range of different parameters. Then a hierarchical grouping strategy based on diverse similarity measurements like color, texture, size and fill, also various starting points, is applied.

In this way, a set of candidate regions are generated of various sizes. Note that users usually are close to the landmark in indoor environment. Landmarks cover certain space of captured image.

Those regions of small size are of low possibility to have interesting indoor objects in them. Therefore, setting a threshold to filter them increases detection efficiency. Here the threshold is set to 50. (The value was determined empirically based on our data).

3.2.3 Landmark Detection. For an image extracted from video, after indoor object recognition stage, there are two results: having indoor object and no indoor object. Images with no indoor object are useless and are discarded. For images having objects, trained CNN network recognizes object type. If object is used for defining single object landmark, then a landmark is detected. If objects are components of multiple objects landmarks, more information are needed to determine landmark type.

In order to avoid separating indoor objects or missing objects, video is sampled with high overlap rate. The disadvantage it brings is that a single object is seen in more than one images. Therefore, instead of determining landmark-based on a single image, a successive image sequence is used to recognize landmarks. Another reason for this is that multiple indoor objects can be seen in an image sequence to determine multiple objects landmark, which might not be detected in a single image due to their size and position.

After landmark detection, the video sequence is divided into landmark segments. A landmark segment starts with the first frame when a landmark type is detected and ends with the last frame containing the current landmark.

4 VISUAL SEQUENCE LOCALIZATION USING SECOND ORDER HIDDEN MARKOV MODEL

Given a sequence of landmark types detected from a video, a matching algorithm is needed to match the landmarks and locations on the topological map. Indoor scenes are usually very similar in the visual space. Directly matching visual features of detected landmarks and those recorded on the floor map can cause ambiguity. We leverage both visual semantic information and contextual information using second order hidden Markov model to perform topological localization.

4.1 Second Order Hidden Markov Model

Hidden Markov model is a Markov model whose states can not be directly measured but can be estimated from observation indirectly. However, observations usually are not sufficient to precisely determine state alone. They generally satisfy certain probability distribution given a state. For landmark-based indoor localization problem, it also can be modelled with hidden Markov model. Detected landmark sequence is observation and landmark locations on the map are the states. The localization process can be regarded as the problem of finding most possible location sequence, given a sequence of landmark observations.

In practice, people are unlikely to walk back and forth between two locations. Hidden Markov model is incapable of excluding such probability. Therefore, we introduced second order hidden Markov model (HMM2) to cope with it.

4.1.1 Transition Matrix of HMM2. Unlike transition matrix of hidden Markov model which is a 2 dimensional matrix, the transition matrix of HMM2 is 3 dimensional. Its value $a_{i,j,k}$ means the probability that next state is k, given the condition that previous state is i and current state is j. For landmark-based indoor localization problem, it represents probability of going through certain landmark position given previous two landmarks positions.

To eliminate the possibility that users walk back and forth, matrix values are assigned to zero, which indicate that next state and previous state are the same. The rest of the matrix values are set based on topological map. Each edge from the same node is given the same probability.

With the constructed HMM2 model and the observed landmark sequence, we aim to find the most probable match between the observed landmark sequence and a location sequence on the topological map. This task can be solved using Viterbi algorithm. Traditional Viterbi algorithm is developed for HMM. It needs to be extended for HMM2.

4.2 Extended Viterbi Algorithm

For landmark-based indoor localization, we aim to recognize the landmark with the highest probability, given previous observation sequence and the hidden Markov model parameters. Assume that a hidden Markov model is known with states set called S. Initial probability π_i represents the probability that the process starts from state i. A_{ij} is the transition probability that the process move from i to j. Observation sequence is $U = Y\{y1, y2 \ldots yi\}$, the most likely state sequence of U is $X = \{x1, x2 \ldots xi\}$. We aim to find the sequence of states that has the maximum probability given the observation sequence. From Bayesian theory,

$$P(X|U) = \frac{P(U|X)P(X)}{P(U)} \tag{3}$$

where $P(U|X)$ denotes, the probability distribution of observation U, given state X. In hidden Markov model, it is represented as emission matrix. $P(X)$ is the prior probability distribution of X. For hidden Markov model, it represents the probability distribution of state sequence X. $P(U)$ is the probability distribution of observation sequence. It is a constant value. Hence the solution to maximizing $P(X|U)$ and maximizing $gu(X)$ are the same.

$$gu(X) = P(U|X)P(X) \tag{4}$$

Suppose that we have n observations. Taking logarithm of $gu(X)$, equation 4 is changed to equation 5.

$$lgu(X) = log(gu(X)) = \sum_{j=1}^{n} logP(y_i|X_i) + logP(x_1, x_2, \ldots\ x_n) \tag{5}$$

Since logarithm function is monotonically increasing, $lgu(X)$ and $gu(X)$ share the same solution for the maximization problem. Note that hidden Markov model requires next state only depends on current state. $LogP(x_1, x_2 \ldots, x_n)$ can be changed as follows.

$$logP(x_1, x_2, \ldots\ x_n) = (\sum_{j=2}^{n} logP(x_j|x_{j-1})) + logP(x_1) \tag{6}$$

The Viterbi algorithm is used to find the largest states sequence that maximizes $logP(x_1, x_2, \ldots,\ x_n)$. For any step k, M values are maintained that represent the largest probability of path end at each state, based on observation y_k and previous path. For each state, we compute the largest possible move and update the probability cost and record the move start. The process are as follows.

$$V_{1,k} = P(y_1|k) \times \pi_k \tag{7}$$

$$V_{t,k} = max(P(y_t|k) \times a_{x,k} \times V_{t-1,x}) \tag{8}$$

$$Ptr(k,t) = \arg\max_{k}(P(y_t|k) \times a_{x,k} \times V_{t-1,x}) \tag{9}$$

where $V_{t,k}$ is the probability of the most probable state sequence given t observations that have k as its final state. Since second order HMM considers previous state and current state to predict next step, thus equation 6 has to be changed to equation 10.

$$logP(x_1, \ldots, x_n) = \sum_{j=3}^{n} logP(x_j|x_{j-1}, x_{j-2}) + logP(x_2|x_1) + logP(x_1) \tag{10}$$

Initial stage and recursive stage are needed to modify. Initial cost is initialized with equations 11 and 12. Recursive stage is changed to equations 13 and 14.

$$V_(1,k) = P(y_1|k) \times \pi_k \tag{11}$$

$$V_2(l,m) = V_{1,k} \times a(l,m) \times P(y_2|k) \tag{12}$$

$$V_t(m,n) = max(V_{t-1}(l,m) \times a_2(l,m,n)) \times P(y_t|k) \tag{13}$$

$$Ptr_t(m,n) = \arg\max_{l} \left(V_{t-1}(l,m) \times a_2(l,m,n)\right) \tag{14}$$

The extended algorithm is summarized in algorithm 1.

Algorithm 1: Extended Viterbi finds the location sequence of maximum probability

```
Input: A sequence of observations $Y$, transition Matrix $A_1, A_2$,
       emission matrix $B$, initial location $\pi$
Output: A sequence of States
Def: $N$, number of locations; $M$, number of landmark type; $K$,
  number of observations
Initialization:
    $V_1 = A_1 \times \pi \times B$
Recursion:
    $V_t = V_{t-1} \times A_2 \times B_t$
    $Ptr_t = \arg\max (V_{t-1} \times A_2)$
Back trace:
    $X_K = \arg\max_n (V_N)$ n column index of the V
    $X_{K-1} = \arg\max_m (V_N)$ n row index of the V
    $X_t = Ptr_{t+1}(X_{t+1}, X_{t+2})$
Return $X$;
```

4.3 Localization Schemes

Given a sequence of landmark type, Viterbi algorithm searches the most probable path by comparing probabilities of possible path candidates. The number of path candidates is calculated to indicate localizing result. When the number of path candidate converges to 1, it means that route is localized. In real applications, users need to be notified once the path is localized. There is no need to wait until they finish the path. A real-time positioning report is required while walking. There are two localization schemes. *Online scheme*: Whenever a new observed landmark type is added to the observation sequence, the Viterbi algorithm is performed to find the matching landmark on the topological map for the new observation. *Offline scheme*: After all landmarks are detected, the observed landmark sequence is matched to a most probable landmark sequence on the topological map using the the Viterbi algorithm.

5 EVALUATION

5.1 Setup

To evaluate the proposed method, we conduct our experiment on the B floor of computer science school in the University of Nottingham. This is a typical office environment with many corridors and office rooms. Its floor plan map is shown in Figure 3. The topological map is produced from floor plan and is shown in Figure 4. It shows the distribution of landmarks of the environment. Node color represents the type of landmark and edge indicate adjacent relationship between two landmarks. There are 65 landmarks in the environment belonging to 8 types which are: office room, stair, elevator, fire extinguisher, man's toilet, woman's toilet, disabled toilet and intersection (corner). Among them, fire extinguisher, stair and elevator belong to single object landmark. Office rooms and toilets are multiple objects landmarks. Intersection is scene landmark. In the topological map, there exist some landmarks only appeared once, which we refer as unique landmarks. They include man's toilet, woman's toilet, disabled toilet and elevator. Landmarks appear more than once are referred as common landmarks.

A participant was asked to walk along 5 routes observing various types and number of landmarks. An Honor android mobile phone was fixed on the arm of the participant with a side view while the participant walked. We collected a video for each route and obtained 5 videos in total. For the 5 videos, the first two are taken with the camera on the left arm and the last three with camera on the right arm.

Route 1: This route goes through 15 landmarks. It starts from office door (node 52) and ends in the intersection (node 14) in topological map. It walks through a sequence of office door, containing a corner, an up turn and a left turn. In this route, there are all common landmarks.

Route 2: Route 2 starts from the left stair and goes straight to the end corner of the corridor. 10 landmarks exist in this route which includes 3 unique landmarks.

Route 3: Route 3 contains 14 landmarks which consists of both unique landmarks and common landmarks. It begins from an intersection (node 16) and goes through a sequence of office doors, turn and elevator and reaches left stair. In this route, unique landmarks exist at the end of the landmark sequence.

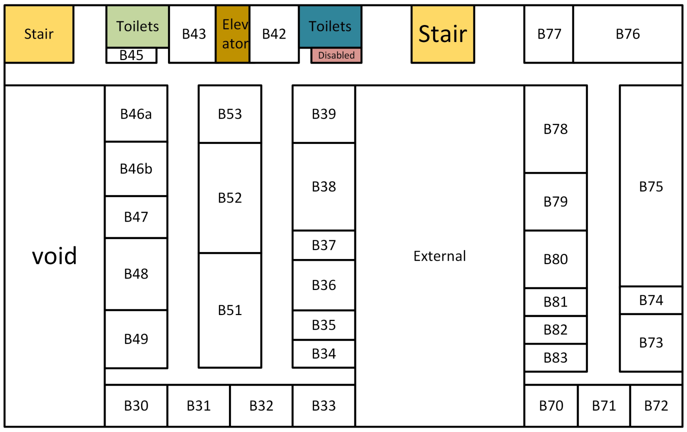

Figure 3: Floor plan map of B floor

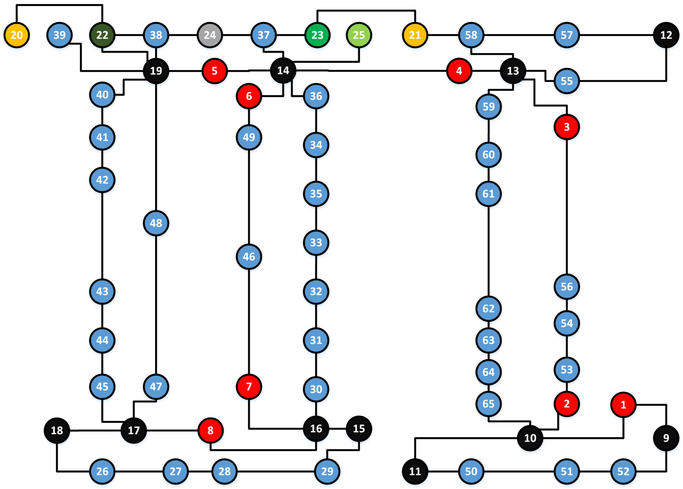

Figure 4: Corresponding landmark topological map of B floor. Color codes different landmark types. Fire extinguisher (Red), intersection (Black), stair (Yellow), office (Blue), elevator (Silver), man’s toilet (Green), woman’s toilet (Dark Green) and disabled toilet (Light Green)

Table 1: Confusion matrix of trained CNN network

Type	DTT	DP	ELV	FE	MTT	DR	ST	TN	WLL	WMTT
DTT	10	0	0	0	0	0	0	0	0	0
DP	0	20	0	0	0	0	0	0	0	0
ELV	0	0	10	0	0	0	0	0	0	0
FE	0	0	0	47	0	0	0	0	0	0
MTT	0	0	0	0	9	0	0	0	0	1
DR	0	0	1	0	0	55	0	0	3	1
ST	0	0	0	0	0	0	19	0	0	0
TN	0	0	0	0	0	0	0	7	0	0
WLL	0	0	0	0	0	0	0	0	12	0
WMTT	0	0	0	0	1	0	0	0	0	9

Route 4: This route starts from a turn (node 16) and ends at office (node 65) , going through an upturn, a left turn and a downturn, containing 17 landmarks. Route 4 just contains common landmarks but the quantity is larger than Route 1.

Route 5: Route 5 begins from a turn named node 16 in the topological map and goes up to the end of the corner then turns left. It goes strait until reaching the turn (node 19). It goes down to the turn (node 17). There are 22 landmarks in this route. In this route, unique landmarks encounter in the middle of the landmark sequence.

5.2 Training Indoor Object Classifier

Our experimental environment consists of 10 classes of objects. These are, 8 classes of indoor objects including door (DR), woman’s toilet tag (WMTT), man’s toilet tag (MTT), disable toilet tag (DTT), fire extinguisher (FE), door plate (DP), elevator (ELV), and stair (ST); one class of scene object (corner or intersection); and one type of background object (walls). Together, they form 7 types of landmarks.

We obtained about 500 images containing these 10 types of indoor objects and about half of these are were used for training (fine-tuning a CNN pre-trained on ImageNet data) and the rest for testing. These data comes from two sources, images on the Internet and video frames from scenes of the B floor corridor of the Computer Science building at the University of Nottingham, UK. We collected images from the Internet because images for certain classes are relatively fewer than others. For instance, images of toilets tags are far less than that of doors. Training dataset should be balanced in every class in order to achieve high generalizing ability of the classifier.

5.2.1 Training and Testing. We selected Alexnet as the basic network and fine-tuned it for our application. The output layer was modified by changing the number of neurons from 1000 to 10. The network is initialized with weights that won ImageNet classification Champion in 2012 except the output layer. The output layer was initialized with normal Gaussian distribution. CNN network was trained under Caffe[14]. It was trained in an MSI laptop in GPU mode. The laptop is installed with windows 10 system and its processor is Intel i7 and with a RAM of 8GB. The graphic card is Nvidia GTX970M. Parameters of the convolutional layers and fully connected layers are kept fixed and only learned the parameter of our output layer.

5.2.2 Result. The confusion matrix in table 1 is generated based on testing images. The overall accuracy is 96. 6%. The classification accuracy of each class are 100%, 100%, 100%, 90. 9%, 100%, 90%, 100%, 100%, 100%, 80% and 81. 81% respectively. It shows that the trained CNN network performed very well in recognizing these indoor objects although some misclassification happened. Small number of wrongly classified object appeared for certain reason. For disable tag, door plate and fire extinguisher, stair and turn, all the testing image are correctly classified since they are quite different among other classes in color, shape or scene. An elevator image is classified as door, a woman’s toilet tag is detected as man’s toilet tag and a man’s toilet tag is seen as the woman’s. It was because they are similar in shape. 3 wall images are classified as office door. It is caused by the fact that the wall in these scenes

Table 2: Landmark detection performance in real test

Route	number of landmark	detected
1	15	15
2	10	10
3	14	14
4	18	18
5	22	22

are made of white board with metallic edge, which is very similar to white door from the environment. In our case, 96. 6% accuracy is high enough for detecting indoor landmarks. However, better performance can be achieved by adding more training samples on those easily mistaken classes.

5.3 Landmark Sequence Detection Performance

All 5 videos were sampled at the rate of 3 frames per second. This rate is selected depending on walking and turning speed. These images are processed with selective search algorithm to generate patches. Image patches are gone through pre-processing like resizing and transformed to the format of Caffe detect indoor objects. Landmark is determined from the classification result according to the strategy in section 3.2.3.

Figure 5 shows ground truth and landmark detection result. Each line segment represents a landmark. The result shows that all landmarks of 5 videos are found, given the fact that number of red and green line segments are the same in every route. The detection result of each landmark sequence is shown in Figures 6, 7, 8, 9 and 10. Table 2 shows statistical result of landmark detection performance of these 5 videos. It demonstrates that our detector has correctly recognized the landmarks in this scene. However, it has to be admitted that in other more cluttered scenes result may not be as good.

5.4 Evaluation of HMM2 for Localization

5.4.1 Online Performance. With the landmark sequence detection result, HMM2 is applied to find the location of the users when a new landmark type is detected from video frames. In this part, we compare the online performances of our proposed method with HMM in two situations: with initialization and without initialization.

Figure 11 shows the performance of HMM and the proposed HMM2 on 5 routes when the staring position is known and unknown respectively.

Route 1 consists of common landmarks. In unknown starting position condition, curve of HMM fluctuates while curve of HMM2 shows a converging trend and finally converges when the 15th landmark is detected. In known starting position case, HMM converges with 7 landmarks and begins to increase from the 8th landmark. HMM2 remains convergent until 14th landmark is observed, and becomes convergent with 15 landmarks.

Route 2 contains several unique landmarks and is hard to be confused with other routes. The curves in Figure 11 also proves that both methods in the two cases converge.

Route 3 begins from common landmarks and ends with unique landmarks. The curves all converge when unique landmarks are observed. Given a common landmarks sequence, HMM2 shows a trend to converge with more landmarks observed, while HMM fluctuates. When starting position is provided, HMM2 converged for all routes.

Route 4 is also made up of common landmarks but more than Route 1 in quantity. It shows similar trend as in Route 1. HMM tends to have more path candidates while HMM2 has a tendency to converge. Knowing starting position helps the algorithm to converge.

Route 5 contains both common and unique landmarks and has unique landmarks in the middle of the sequence. Both HMM and HMM2 achieved good performance before passing the unique landmarks. HMM2 converged when more landmarks were detected while HMM failed to converge.

HMM fails to localize routes consisting of long common landmarks while HMM2 provides better results. Knowing starting positions and unique landmarks both aid to filter out wrong candidates and speed up convergence.

The HMM2 curve fluctuates when certain landmarks are observed. It is due to that landmark candidates have the same observation and both connect with the current landmark. For instance, in route 1, HMM2 changes from convergence to divergence. It is because node 13 is connected to two nodes 3 and 4, and their observations are both fire extinguishers. Same things happen in route 5 at node 2, 8 and 16.

5.4.2 Offline Performance. Offline matching is done after all landmarks are detected and we match the whole landmark sequence with the topological map. The ground truth routes and predicted routes are shown in Table 3. The result shows that the proposed method is capable of localizing users accurately except for Route 4 with no starting information. It demonstrates the effectiveness of the proposed method. In route 4, the proposed method did not converge to a single path. It is because Route 4 only involved common landmark types. It can localized with more landmarks.

6 CONCLUSION

This paper presents a novel landmark sequence-based indoor localization method using a smartphone camera. A new landmark approach is proposed based on a deep learning neural network and second order hidden Markov model is applied to recognize a consecutive locations of indoor topological map given detected landmark sequence from the user's traveling path. The advantage of the proposed method are 1) CNN neural network is introduced to detect both object and indoor scene landmarks using the same network; 2) context information and visual appearance are both exploited to recognize landmarks for localization. Experiment result demonstrated the effectiveness of proposed method. In this paper, only visual information is considered. Geometric information or Wi-Fi information can help with landmarks-based localization by providing geometric and wireless signal description of landmarks. In the future work, we tend to combine visual landmarks with other technologies like Wi-Fi and IMU to increase localization efficiency.

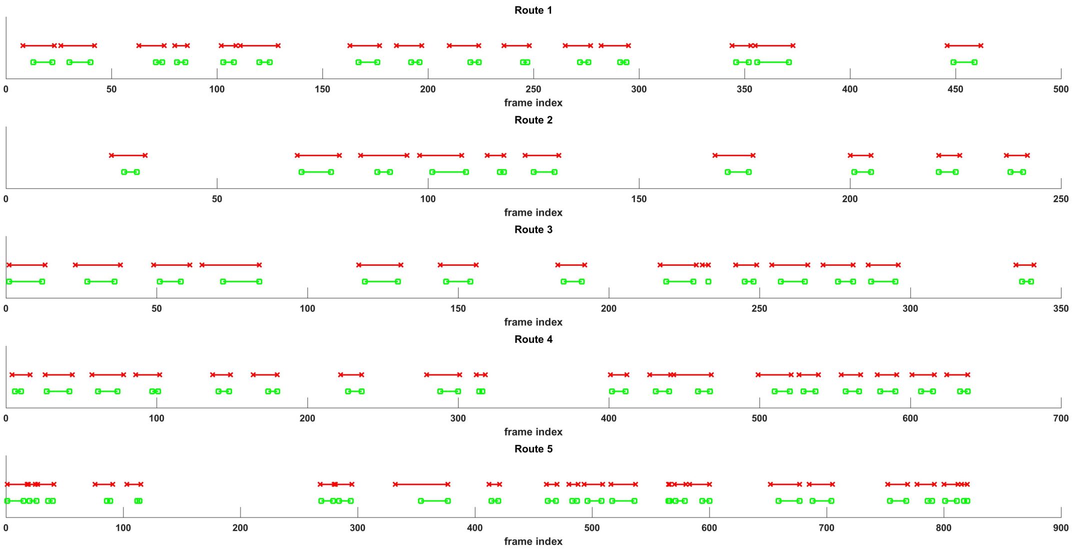

Figure 5: Landmark detection result. Each line segment indicates a video segment of a landmark. Red line represents ground truth and green line represents the detection result

Table 3: Routes localization Results

Route	Situation	Route Chain
Route1	True Route	52 → 51 → 50 → 11 → 10 → 65 → 64 → 63 → 62 → 61 → 60 → 59 → 13 → 4 → 14
	Predict Route without Initialization	52 → 51 → 50 → 11 → 10 → 65 → 64 → 63 → 62 → 61 → 60 → 59 → 13 → 4 → 14
	Predict Route with Initialization	52 → 51 → 50 → 11 → 10 → 65 → 64 → 63 → 62 → 61 → 60 → 59 → 13 → 4 → 14
Route2	True Route	20 → 22 → 38 → 24 → 37 → 23 → 21 → 58 → 57 → 12
	Predict Route without Initialization	20 → 22 → 38 → 24 → 37 → 23 → 21 → 58 → 57 → 12
	Predict Route with Initialization	20 → 22 → 38 → 24 → 37 → 23 → 21 → 58 → 57 → 12
Route3	True Route	16 → 30 → 31 → 32 → 33 → 35 → 34 → 36 → 14 → 37 → 24 → 38 → 22
	Predict Route without Initialization	16 → 30 → 31 → 32 → 33 → 35 → 34 → 36 → 14 → 37 → 24 → 38 → 22
	Predict Route with Initialization	16 → 30 → 31 → 32 → 33 → 35 → 34 → 36 → 14 → 37 → 24 → 38 → 22
Route4	True Route	16 → 30 → 31 → 32 → 33 → 35 → 34 → 36 → 14 → 4 → 13 → 59 → 60 → 61 → 62 → 63 → 64 → 65
	Predict Route without Initialization	16 → 30 → 31 → 32 → 33 → 35 → 34 → 36 → 14 → 4 → 13 → 59 → 60 → 61 → 62 → 63 → 64 → 65
	Predict Route without Initialization	10 → 65 → 64 → 63 → 62 → 61 → 60 → 59 → 13 → 4 → 14 → 36 → 34 → 35 → 33 → 32 → 31 → 30
	Predict Route with Initialization	16 → 30 → 31 → 32 → 33 → 35 → 34 → 36 → 14 → 4 → 13 → 59 → 60 → 61 → 62 → 63 → 64 → 65
Route5	True Route	10 → 2 → 53 → 54 → 56 → 3 → 13 → 58 → 21 → 23 → 37 → 24 → 38 → 22 → 19 → 40 → 41 → 42 → 43 → 44 → 45 → 17
	Predict Route without Initialization	10 → 2 → 53 → 54 → 56 → 3 → 13 → 58 → 21 → 23 → 37 → 24 → 38 → 22 → 19 → 40 → 41 → 42 → 43 → 44 → 45 → 17
	Predict Route with Initialization	10 → 2 → 53 → 54 → 56 → 3 → 13 → 58 → 21 → 23 → 37 → 24 → 38 → 22 → 19 → 40 → 41 → 42 → 43 → 44 → 45 → 17

ACKNOWLEDGMENTS

The Qing Li is supported by the Horizon Centre for Doctoral Training at the University of Nottingham (RCUK Grant No. EP/L015463/1)and by the Shenzhen University and would like to thank Staff for providing floor plan map of the environment.

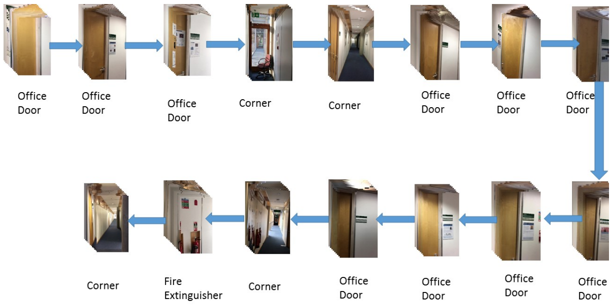

Figure 6: Landmark detection Result of Route 1

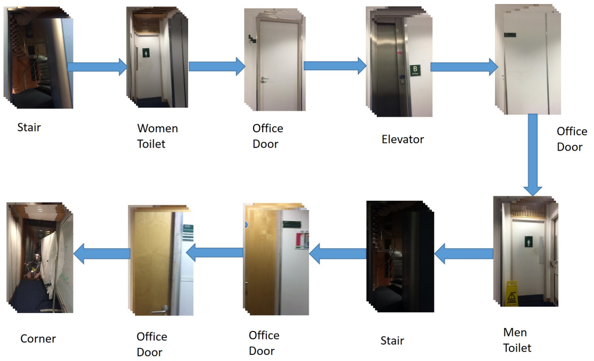

Figure 7: Landmark detection Result of Route 2

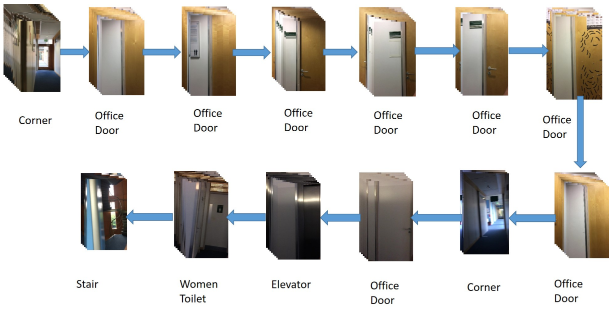

Figure 8: Landmark detection Result of Route 3

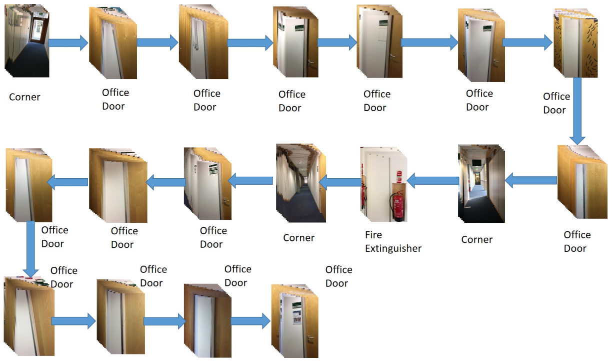

Figure 9: Landmark detection Result of Route 4

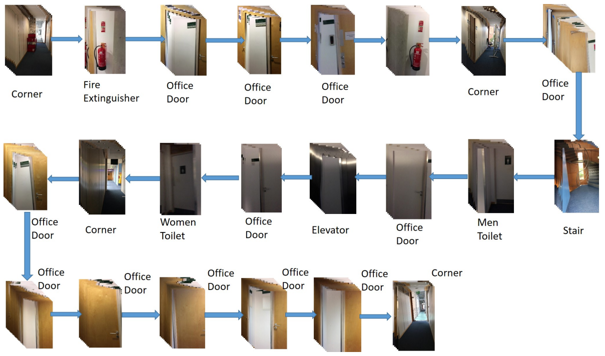

Figure 10: Landmark detection Result of Route 5

REFERENCES

[1] Yicheng Bai, Wenyan Jia, Hong Zhang, Zhi-Hong Mao, and Mingui Sun. 2014. Landmark-based indoor positioning for visually impaired individuals. In *Signal Processing (ICSP), 2014 12th International Conference on*. IEEE, 668–671.

[2] Anahid Basiri, Pouria Amirian, and Adam Winstanley. 2014. The use of quick response (qr) codes in landmark-based pedestrian navigation. *International Journal of Navigation and Observation* 2014 (2014).

[3] Craig Becker, Joaquin Salas, Kentaro Tokusei, and J-C Latombe. 1995. Reliable navigation using landmarks. In *Robotics and Automation, 1995. Proceedings., 1995 IEEE International Conference on*, Vol. 1. IEEE, 401–406.

[4] Margrit Betke and Leonid Gurvits. 1997. Mobile robot localization using landmarks. *IEEE transactions on robotics and automation* 13, 2 (1997), 251–263.

[5] Beatriz L Boada, Dolores Blanco, and Luis Moreno. 2004. Symbolic place recognition in voronoi-based maps by using hidden markov models. *Journal of Intelligent & Robotic Systems* 39, 2 (2004), 173–197.

[6] Kuan-Chieh Chen and Wen-Hsiang Tsai. 2010. Vision-based autonomous vehicle guidance for indoor security patrolling by a SIFT-based vehicle-localization technique. *IEEE transactions on vehicular technology* 59, 7 (2010), 3261–3271.

[7] Zhenghua Chen, Han Zou, Hao Jiang, Qingchang Zhu, Yeng Chai Soh, and Lihua Xie. 2015. Fusion of WiFi, smartphone sensors and landmarks using the Kalman filter for indoor localization. *Sensors* 15, 1 (2015), 715–732.

[8] Zhi-An Deng, Guofeng Wang, Danyang Qin, Zhenyu Na, Yang Cui, and Juan Chen. 2016. Continuous indoor positioning fusing WiFi, smartphone sensors and landmarks. *Sensors* 16, 9 (2016), 1427.

[9] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 580–587.

[10] Fuqiang Gu, Kourosh Khoshelham, Jianga Shang, and Fangwen Yu. 2016. Sensory landmarks for indoor localization. In *Ubiquitous Positioning, Indoor Navigation and Location Based Services (UPINLBS), 2016 Fourth International Conference on*. IEEE, 201–206.

[11] Kai Guan, Lin Ma, Xuezhi Tan, and Shizeng Guo. 2016. Vision-based indoor localization approach based on SURF and landmark. In *Wireless Communications and Mobile Computing Conference (IWCMC), 2016 International*. IEEE, 655–659.

[12] Jean-Bernard Hayet, Frédéric Lerasle, and Michel Devy. 2002. A visual landmark framework for indoor mobile robot navigation. In *Robotics and Automation, 2002. Proceedings. ICRA'02. IEEE International Conference on*, Vol. 4. IEEE, 3942–3947.

[13] Gijeong Jang, Sungho Lee, and Inso Kweon. 2002. Color landmark based self-localization for indoor mobile robots. In *Robotics and Automation, 2002. Proceedings. ICRA'02. IEEE International Conference on*, Vol. 1. IEEE, 1037–1042.

[14] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. 2014. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 675–678.

[15] Markus Kattenbeck. 2015. Empirically Measuring Object Saliency for Pedestrian Navigation. (2015).

[16] Hisato Kawaji, Koki Hatada, Toshihiko Yamasaki, and Kiyoharu Aizawa. 2010. Image-based indoor positioning system: fast image matching using omnidirectional panoramic images. In *Proceedings of the 1st ACM international workshop on Multimodal pervasive video analysis*. ACM, 1–4.

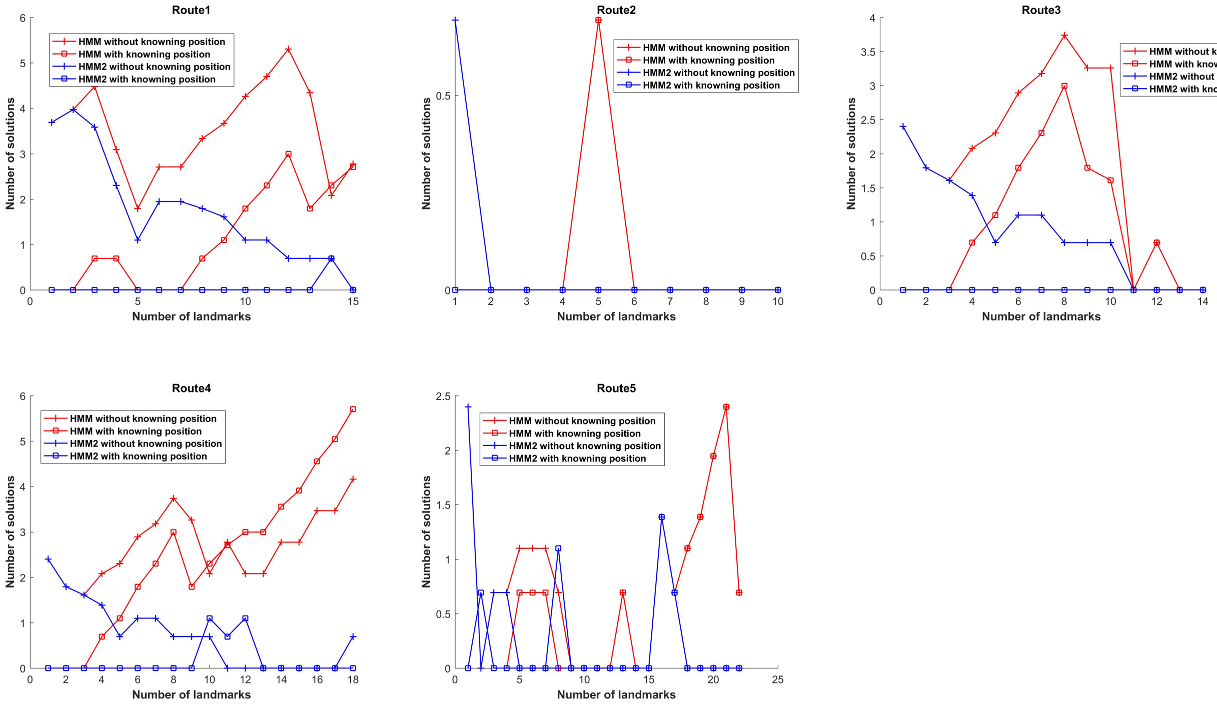

Figure 11: Comparison of HMM and HMM2 with unknown and known starting position. The axis X represents the number of observed landmarks and axis Y is the Logarithm of the number of candidates path with X landmarks detected. The red curves represent performance of HMM and blue curves are performance of HMM2. The cross indicates performance is achieved under unknown starting position condition and square for known starting position condition.

[17] Jana Kosecká and Fayin Li. 2004. Vision based topological Markov localization. In *Robotics and Automation, 2004. Proceedings. ICRA'04. 2004 IEEE International Conference on*, Vol. 2. IEEE, 1481–1486.

[18] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*. 1097–1105.

[19] Jason Zhi Liang, Nicholas Corso, Eric Turner, and Avideh Zakhor. 2013. Image based localization in indoor environments. In *Computing for Geospatial Research and Application (COM. Geo), 2013 Fourth International Conference on*. IEEE, 70–75.

[20] Guoyu Lin and Xu Chen. 2011. A Robot Indoor Position and Orientation Method based on 2D Barcode Landmark. *JCP* 6, 6 (2011), 1191–1197.

[21] Guoyu Lu and Chandra Kambhamettu. 2014. Image-based indoor localization system based on 3d sfm model. In *IS&T/SPIE Electronic Imaging*. International Society for Optics and Photonics, 90250H–90250H.

[22] Andry Maykol G Pinto, A Paulo Moreira, and Paulo G Costa. 2012. Indoor localization system based on artificial landmarks and monocular vision. *TELKOMNIKA (Telecommunication Computing Electronics and Control)* 10, 4 (2012), 609–620.

[23] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*. 91–99.

[24] M Serrão, João MF Rodrigues, JI Rodrigues, and JM Hans du Buf. 2012. Indoor localization and navigation for blind persons using visual landmarks and a GIS. *Procedia Computer Science* 14 (2012), 65–73.

[25] YingLi Tian, Xiaodong Yang, Chucai Yi, and Aries Arditi. 2013. Toward a computer vision-based wayfinding aid for blind persons to access unfamiliar indoor environments. *Machine vision and applications* 24, 3 (2013), 521–535.

[26] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. 2013. Selective search for object recognition. *International journal of computer vision* 104, 2 (2013), 154–171.

[27] Martin Werner, Moritz Kessel, and Chadly Marouane. 2011. Indoor positioning using smartphone camera. In *Indoor Positioning and Indoor Navigation (IPIN), 2011 International Conference on*. IEEE, 1–6.

[28] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. 2014. Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*. 487–495.

[29] Baoding Zhou, Qingquan Li, Qingzhou Mao, Wei Tu, and Xing Zhang. 2015. Activity sequence-based indoor pedestrian localization using smartphones. *IEEE Transactions on Human-Machine Systems* 45, 5 (2015), 562–574.

[30] Barbara Zitová and Jan Flusser. 1999. Landmark recognition using invariant features. *Pattern Recognition Letters* 20, 5 (1999), 541–547.