MODELLING UNCERTAINTY OF SINGLE IMAGE INDOOR LOCALISATION USING A 3D MODEL AND DEEP LEARNING

Debaditya Acharya[1]*, Sesa Singha Roy[2], Kourosh Khoshelham[1], Stephan Winter[1]

[1]Department of Infrastructure Engineering, The University of Melbourne, Parkville, Victoria, Australia, 3010
[2] Institute for Sustainable Industries and Livable Cities, Victoria University, Werribee, Victoria, Australia, 3030
acharyad@student.unimelb.edu.au, sesa.singharoy@live.vu.edu.au, {k.khoshelham,winter}@unimelb.edu.au

KEY WORDS: Indoor localisation, Camera pose regression, 3D models, Deep learning, Bayesian systems, Uncertainty

ABSTRACT:

Many current indoor localisation approaches need an initial location at the beginning of localisation. The existing visual approaches to indoor localisation perform a 3D reconstruction of the indoor spaces beforehand, for determining this initial location, which is challenging for large indoor spaces. In this research, we present a visual approach for indoor localisation that is eliminating the requirement of any image-based reconstruction of indoor spaces by using a 3D model. A deep Bayesian convolutional neural network is fine-tuned with synthetic images generated from a 3D model to estimate the camera pose of real images. The uncertainty of the estimated camera poses is modelled by sampling the outputs of the Bayesian network fine-tuned with synthetic images. The results of the experiments indicate that a localisation accuracy of 2 metres can be achieved using the proposed approach.

1. INTRODUCTION

Indoor localisation is the key enabler of many applications like navigation guidance, location-based services and augmented reality. In the absence of global navigation satellite system signals (GNSS) in indoor environments, several approaches have emerged in the past two decades that include Wifi, ultra-wideband or radio frequency identification (Mautz, 2012). However, these approaches are dependent on a dedicated network of sensors and are often expensive. With the widespread availability of mobile devices having sensors, especially camera, visual approaches such as SLAM, visual odometry and model-based visual tracking have become a focus of research (Acharya et al., 2019b). However, the key constraint for these approaches is the requirement of an initial location at the start of localisation (Se et al., 2002).

The existing visual approaches that are independent of the initial location includes image retrieval approaches (Arandjelovic and Zisserman, 2012, Radenović et al., 2016) and direct pose regression approaches (Kendall et al., 2015, Shotton et al., 2013). The main drawback of such approaches is the creation of databases of images, depth images or point clouds that usually involves performing a 3D reconstruction of the indoor spaces before (Piasco et al., 2018). Sensing large indoor spaces becomes a challenge due to time and resources involved in the process. For example, structure-from-motion (SfM) requires capturing a large number of overlapping images to estimate the camera poses and reconstruct the environment. Besides, the SfM approaches are computationally expensive and are susceptible to errors.

In this research, a solution is proposed to eliminate the requirement of image-based reconstruction the indoor spaces, utilising a 3D indoor model that can be obtained from an existing building information model (BIM) of the building. This 3D model is used to generate synthetic images with known camera poses. The images are subsequently used for fine-tuning a deep Bayesian convolutional neural network (CNN) to regress the camera poses of real test images. Therefore, this approach eliminates the requirement of real images with their ground truth poses during the fine-tuning phase, which is usually derived from SfM approaches.

In Addition, we model the uncertainty of camera pose estimations by adopting a Bayesian CNN (Kendall and Cipolla, 2016). Uncertainty provides an indication of confidence and trust in an estimated position in the absence of ground truth. Uncertainty modelling in pose regression networks has been done based on real images by previous workers. In this work, we use synthetic images to fine-tune deep CNNs, and evaluate the uncertainty of pose regression by testing with real images. The results are compared with a benchmark that is created by using real images and SfM approaches. The following are the main contributions:

1. The proposed approach eliminates the requirement of performing image-based reconstruction of indoor spaces by utilising synthetic images rendered from a 3D model.
2. The uncertainty of the estimated camera poses from a network fine-tuned with synthetic images is modelled to show the correlation with actual errors.
3. A detailed analysis of the trajectory is performed to identify the factors contributing to large errors and uncertainties.

The paper proceeds with a review of visual approaches in Section 2. The related concepts are explained in Section 3. The details of the dataset, experiments and results are explained in Section 4, followed by discussions and future directions in Section 5 and conclusions in Section 6.

2. RELATED WORKS

The existing visual approaches that estimate the image location without an initial estimate are either image retrieval approaches or direct pose regression approaches. The image-based retrieval approaches use an existing database of images with known camera pose (Arandjelovic and Zisserman, 2012) to match with

*Corresponding author

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-IV-2-W5-247-2019 | © Authors 2019. CC BY 4.0 License.

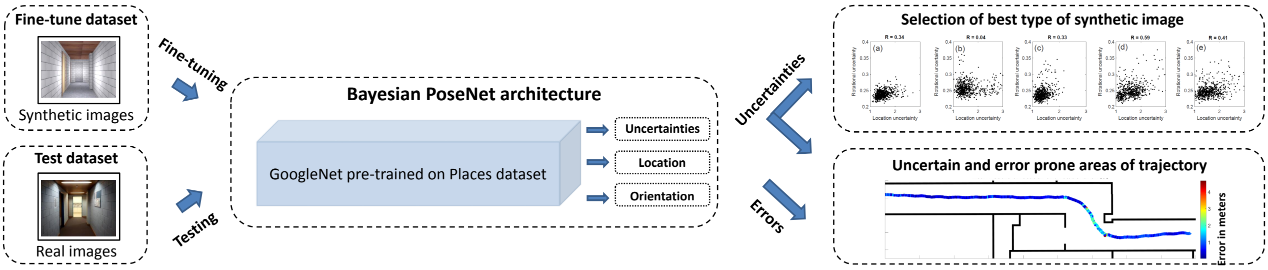


Figure 1. The design of the approach.

the query image to estimate camera pose. However, these approaches require a large database of images with known pose and are susceptible to viewpoint changes (Piasco et al., 2018).

The direct pose regression approaches include the works of (Irschara et al., 2009, Shotton et al., 2013), where point clouds and depth images are used respectively to regress the camera pose. Although these methods can provide accurate estimates of the camera pose they are limited by the requirement of a 3D point cloud that is usually created using SfM approaches or of a specific camera to sense the depth of the indoor scene.

In the last five years, deep CNNs have achieved outstanding performance in image classification (Krizhevsky et al., 2012, Acharya et al., 2018) and object detection and tracking (Acharya et al., 2017). There is a recent trend in the research community using deep CNNs for localisation. For example SLAM (Parisotto et al., 2018) and visual odometry (Wang et al., 2017) use a sequence of images, and direct pose regression approaches, such as PoseNet (Kendall et al., 2015), use single images only. PoseNet is a pre-trained deep CNN that is fine-tuned with images of known pose derived from SfM approaches. Similar to the approaches using SfM for localisation, the major challenge for PoseNet and its following approaches is the requirement of 3D reconstruction the environment. Geometrical constrains have been used to improve the pose regression ability of PoseNet by using a geometrical loss function (Kendall et al., 2017). Synthetic images derived from a BIM have been used to fine-tune PoseNet to localise with real images (Acharya et al., 2019a).

The uncertainty of a system plays a vital role in understanding the confidence of the estimations. In the literature, Bayesian systems have been used to model the uncertainty of a system, such as Kalman filtering (Groves, 2013). On the other hand, probabilistic approaches to direct pose regression such as the works of (Kendall and Cipolla, 2016), measure the uncertainty of the network for localisation. In Kalman filter uncertainty is estimated by propagation of variance, whereas in regression networks it is done by stochastic Monte Carlo sampling.

The estimated uncertainty correlates with the actual error obtained from ground truth camera poses for real images (Kendall and Cipolla, 2016). An additional finding of the study was that the location and the rotation uncertainties are also correlated, and thus can be used to form a single uncertainty value which represents the overall model uncertainty.

While uncertainty modelling in pose regression networks has been done on real images, in this work we use synthetic images that eliminate the requirement of image-based reconstruction of the indoor space or performing SfM approaches. We achieve this by fine-tuning the networks with synthetic images, and ground truth camera poses derived from the 3D model. Additionally, we explore the correlation of the uncertainty with the actual errors obtained from ground truth for a network fine-tuned with synthetic images and tested on real images.

3. METHODOLOGY

The proposed approach uses a pre-trained Bayesian CNN fine-tuned with different types of synthetic images with known camera poses to estimate the unknown camera poses of real images. Figure 1 shows the overview of the approach, where we use a modified architecture of GoogLeNet (Szegedy et al., 2015). We fine-tuned the network with a synthetic dataset containing several types of images generated by graphical rendering from the 3D model. Subsequently, the fine-tuned networks are tested with real images for its pose regression ability and uncertainty of the network is modelled to check the correlation with actual errors. Besides, we identify the factors contributing large errors and uncertainties for networks fine-tuned with synthetic images and tested on real images.

3.1 Rendering synthetic images from a 3D model

In the literature, the ground truth camera poses are estimated using SfM approaches for fine-tuning the networks (Kendall et al., 2015, Walch et al., 2017, Clark et al., 2017, Kendall and Cipolla, 2016). In this research, the ground truth camera poses are known, and a virtual camera is used to render synthetic images. The details of the creation of the synthetic dataset are described in Section 4.1.1.

The shallow layers of a deep CNN learn generic low-level image features, such as edges and textures, whereas the deeper layers can distinguish high-level features such as roads and sky. These high-level image features serve as image landmarks that is used to perform pose regression (Kendall et al., 2015). The image features in the deep layers of a pre-trained deep CNN are invariant to colour, texture, pose or context (background) for the task of classification (Peng et al., 2015). This leads us to believe that a deep CNN fine-tuned with synthetic images with known camera pose should be able to regress the camera pose of real images with different colour and texture.

A pre-trained deep CNN requires only a few fine-tuning images as compared to millions of images required for training from random weights. As generating millions of synthetic images is an overwhelming task, we fine-tune a pre-trained network with 1500 synthetic images and perform pose regression by leveraging transfer learning. A relevant question is whether the fine-tuned network can learn relevant high-level geometrical features from the synthetic images for camera pose regression with real images. The relevant high-level features in the context of indoor spaces could be doors, walls and ceilings.

Additionally, another question is whether processing the real images to make them similar to the synthetic images will improve the pose regression ability of the network. To examine the concept, we convert the real and the synthetic images to edge gradient magnitude (gradmag) images, hence transforming them both to a common image feature space. Consequently, we fine-tune the networks with edge gradmag of synthetic images and test with edge gradmag of real images.

3.2 Fine-tuning a Bayesian network

Bayesian neural networks model the uncertainty in neural networks by gathering distributions over the network weights (MacKay, 1992). Drop-out (Srivastava et al., 2014) is a common practice that is used to avoid over-fitting a network while training with limited training data. Sampling a Bayesian network with randomly dropped out connection at test time can be considered as a way of getting Monte Carlo samples from the posterior distribution of the estimations (Kendall and Cipolla, 2016). Moreover, as the posterior distribution is not directly traceable for such a network, a variational inference usually is used to approximate the distribution of the weights (Gal and Ghahramani, 2016).

We adopt a 24 layer deep Bayesian CNN (Kendall and Cipolla, 2016), where the softmax classifier (Szegedy et al., 2015) is replaced with a fully connected layer of dimension 2048. The fully connected layer is connected to a 7-dimensional affine pose regressor that regresses the camera pose p:

$$p = [x, q] \tag{1}$$

where x = location of camera [X, Y, Z]
q = quaternions defining camera rotation $[q_0, q_1, q_2, q_3]$

The weights are taken from a pre-trained network trained on Places dataset (Zhou et al., 2014). The Places dataset contains millions of images of different locations and is suitable for training networks to classify places. Using the weights of a pre-trained network eliminates the tedious task of training the network with millions of images. Drop-outs are performed only before the convolution layers that had randomly initialised weights, in order to improve the pose regression ability of the network. The network is subsequently fine-tuned with synthetic images using the objective loss function:

$$loss(I) = ||\hat{x} - x||_2 + \beta \left|\left| \hat{q} - \frac{q}{||q||} \right|\right|_2 \tag{2}$$

where $\hat{x}$ = estimated location
$\hat{q}$ = estimated orientation
I = Image
β = scaling factor to balance errors of x and q.

At the test time, the Bayesian network is sampled to obtain Monte Carlo drop-out samples, and the pose is estimated by taking a mean of the samples. The number of samples required for optimum performance of the network is 40, beyond which there is no evidence of improvement (Kendall and Cipolla, 2016). The distribution of the samples follows a unimodal Gaussian distribution for both location and rotation (Kendall and Cipolla, 2016), and the distributions contains only one peak corresponding to the maxima. The locational and rotational uncertainties are expressed as the trace of the unimodal Gaussian's covariance matrices and are represented by numbers. Note that the locational uncertainties have the unit of m^2 and the rotational uncertainties are dimensionless scalers (product of two quaternions).

4. EXPERIMENTS AND RESULTS

Two experiments were performed to model the uncertainty of the network's pose regression on real images when fine-tuned with synthetic images. The first experiment consists of creating a baseline performance using real images, and to determine the optimum value of weighting factor β. The second experiment was performed to measure the performance of the network fine-tuned with different types of synthetic images on real images. In addition, the experiment involved fine-tuning two networks with edge gradmag of synthetic images and testing with the gradmag of the real images.

Caffe (Jia et al., 2014) deep learning library was used for the experiments. The loss was minimised with a learning rate of 10^{-3} using Adagrad gradient descent optimisation for 160 epochs on an NVIDIA GTX980M graphics processor unit (GPU) with 4GB memory. The images were resized to 320x240 pixels, and a crop of 224x224 pixels was applied subsequent to mean subtraction during fine-tuning and testing.

4.1 Dataset

Two datasets were used in the experiments, namely a synthetic image dataset and a real image dataset. The synthetic image dataset was created from the 3D model. The real image dataset was collected by capturing images of a corridor of a building with a camera of a smartphone.

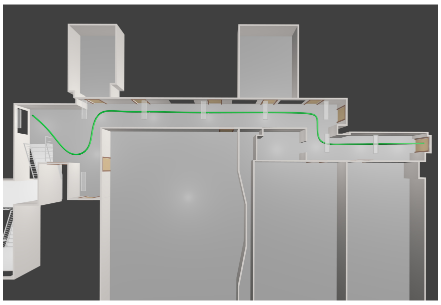


Figure 2. The 3D model derived from a BIM. The trajectory used for generating synthetic dataset is shown in green color. For better visualisation of the indoor space, the roof was removed.

4.1.1 Synthetic image dataset We used a 3D model derived from the UoM dataset, which is part of the ISPRS benchmark on indoor modelling (Khoshelham et al., 2017, Khoshelham et al., 2018, Ramezani et al., 2017). The dataset consists of a part of a corridor having coverage of $230m^2$ and is shown in Figure 2. In the literature, the ground truth camera poses of the real images are estimated using SfM approaches, and the images are captured along a trajectory in the indoor or outdoor spaces (Kendall et al., 2015, Walch et al., 2017, Kendall and Cipolla, 2016, Clark et al., 2017). Following the same idea, we define a trajectory of approximately 30 meters in the 3D model as shown in Figure 2.

We generated synthetic images by rendering images at an interval of 5 centimetres along the trajectory. Additional synthetic images were rendered by tilting the camera $\pm 10^o$ around Y and Z axis. Ideally, we should render synthetic images of the indoor environment in all possible locations and orientations (Wu et al., 2017). However, that will raise the rendering time to generate synthetic images substantially. Consequently, we generated a total number of 3000 synthetic images for each type of dataset.

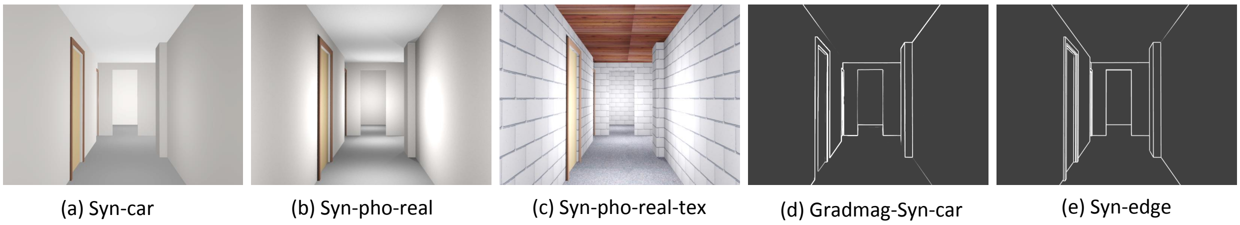


Figure 3. Different types of synthetic images generated from the 3D model.

Five different types of synthetic images were generated by moving a virtual camera along the predefined trajectory. The five datasets are called Cartoonish (Syn-car), Photo-realistic (Syn-pho-real), Photo-realistic textured (Syn-pho-real-tex), Gradmag of Cartoonish (Gradmag-Syn-car) and Edge render (Syn-edge). The names in the brackets are pseudonyms that we used for convenience. The naming convention of image types are as per Blender[1], which was used to render the images. Sample images of each dataset are shown in Figure 3.

The Syn-car images use a rendering model that roughly traces the path of light rays. The Syn-pho-real and Syn-pho-real-tex images use an advanced Physics-based rendering model that traces the light rays in a way very close to the real world. The main difference between Syn-pho-real and Syn-pho-real-tex is the presence of synthetic textures, such as textures of brick and carpet for the latter. Gradmag-Syn-car images were derived by taking the edge gradmag of the Syn-car images. Syn-edge images were created by rendering only the edges visible in the camera field-of-view, such as the edge of walls and floor. The Syn-edge images does not contain the effects of illumination of the environment as compared to the Gradmag-Syn-car images.

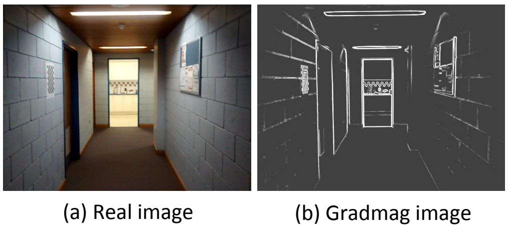


Figure 4. A real image and its gradient magnitude image.

4.1.2 Real image dataset A video sequence containing a total number of 1000 images of the corresponding corridor was captured by a smartphone camera having a field-of-view of 56.32^o. The images were captured at the rate of 30 frames per second and had a resolution of 640x480 pixels. We suppressed the weak edges that were below a threshold to produce the edge gradmag of the real images. A real image and its corresponding edge gradmag image is shown in Figure 4.

4.2 Baseline performance using real images

We created a baseline performance by fine-tuning a network with real images (not gradmag) of known camera poses and tested with real images again. The ground truth camera poses of the real images were estimated by using SfM approach of Agisoft PhotoScan Professional® software that uses bundle adjustment for estimating and refining camera poses. We provided the 3D coordinate of the reconstructed environment manually for performing accurate bundle adjustment, resulting in the re-projection error of 0.91 pixels, and the reconstruction error of 2.60 centimetres.

The image sequences were randomly portioned into fine-tuning and validation sets containing 500 images respectively. The suitable value of the parameter β was estimated by performing a grid search in the expected range of 120 to 750 for indoor environments (Kendall et al., 2015). Figure 5 shows the median location and rotation errors with the variation of β. The location errors are calculated as the Euler distances from the ground truth and the rotational errors in radians are calculated by $2cos^{-1}(\langle \hat{q}q \rangle)$, where $\langle . \rangle$ denotes the inner product of two vectors.

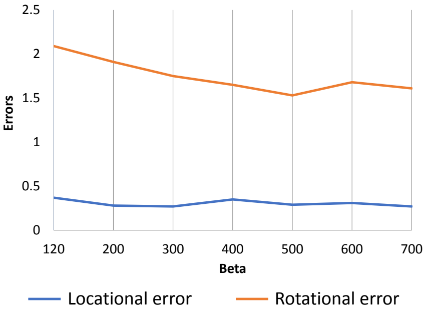


Figure 5. The variation of locational and rotational errors with β.

Figure 6 shows the distribution of location and rotation uncertainties corresponding to the different values of β. The correlation of the locational and rotational uncertainties are estimated by using the correlation coefficient R which is defined as:

$$R = \frac{cov(U_l, U_r)}{\sigma_{U_l}\sigma_{U_r}} \tag{3}$$

where cov = covariance of two random variables
U_l = Uncertainty of location
U_r = Uncertainty of rotation
σ_{U_l} = Standard deviation of U_l
σ_{U_r} = Standard deviation of U_r.

R has a range of [-1,1]. The sign of the coefficient denotes the positive or negative correlation, whereas the magnitude indicates the strength of the relation of two random variables. Figure 6 indicates that a strong positive correlation for the value of $\beta = 500$. Therefore, we select a β value of 500 for our experiments, as the ideal network should have correlated locational and rotational uncertainties. Figure 7 (a) shows the distribution of the estimated camera poses along with the locational and rotational errors and it is observed that the largest errors are near the 90^o turns of the corridor. Figure 7 (b) and (c) shows the variation of locational uncertainty with location errors and rotational uncertainty with rotational errors respectively.

[1] Blender is an open-source 3D computer graphics software for performing simulations and for creating visual art and videos. For more information visit: www.blender.org

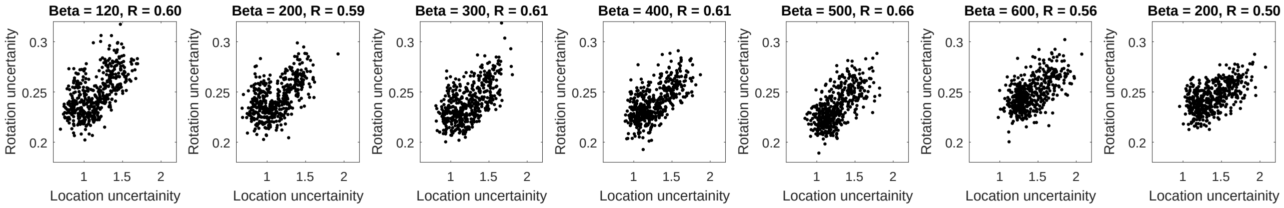


Figure 6. The variation of locational and rotational uncertainties with β. R denotes correlation coeffcient.

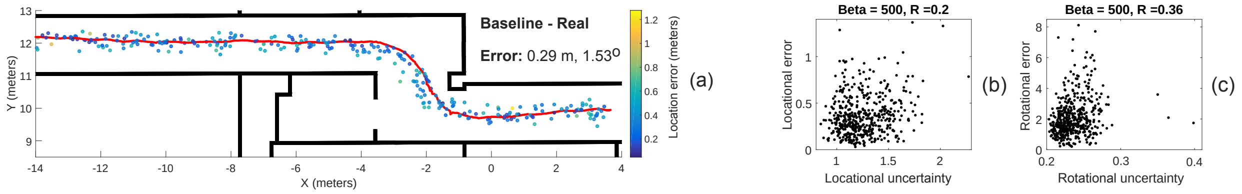


Figure 7. Baseline performence of a network fine-tuned and validated using real images with $\beta = 500$. (a) The distribution of the estimated camera poses. The red line shows the ground truth trajectory estimated using SfM approaches. The colours represent the magnitude of the errors corresponding to each point. The locational and rotational errors are in meters and degrees respectively. (b) locational uncertainty vs locational errors (c) rotational uncertainty vs rotational errors

Total synthetic images	fine-tuning images	validation images	real gradmag images for testing
3000	1500	1500	1000

Table 1. The number of synthetic fine-tuning, synthetic validation and real testing images for the five networks.

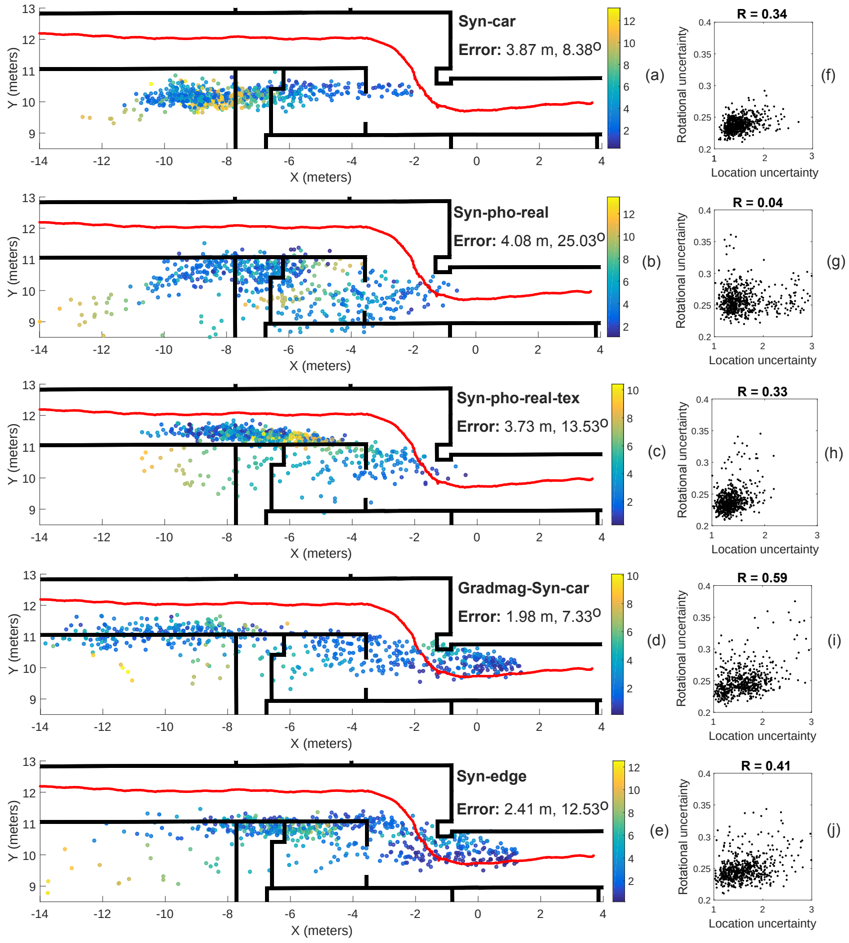
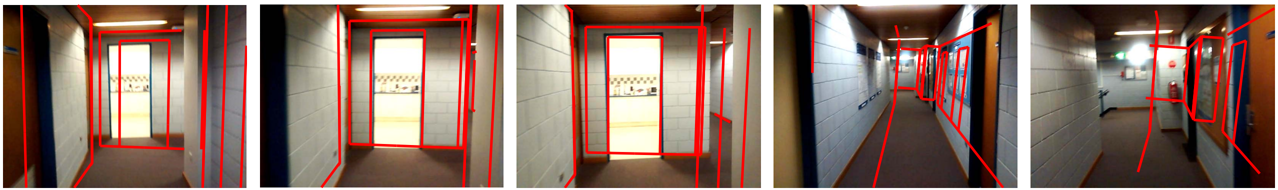


Figure 8. The results of the network (a) fine-tuned with Syn-car and tested on real images (b) fine-tuned with Syn-pho-real and tested on real images (c) fine-tuned with Syn-pho-real-tex and tested on real images (d) fine-tuned with Gradmag-Syn-car and tested on edge gradmag of real images (e) fine-tuned with Syn-edge and tested on edge gradmag of real images. Uncertainty plots for (f) Syn-car (g) Syn-pho-real (h) Syn-pho-real-tex (i) Gradmag-Syn-car and (j) Syn-edge datasets.



(a) Frame 77 (b) Frame 98 (c) Frame 113 (d) Frame 371 (e) Frame 541

Figure 9. Predicted camera poses by a network fine-tuned with Gradmag-Syn-car dataset and tested with edge gradmag of real images.

Frame	Location error (m)	Location uncertainty	Rotational error	Rotational uncertainty
77	0.55 m	1.01 m^2	5.74^o	0.22
98	0.24 m	1.44 m^2	2.77^o	0.24
113	0.46 m	1.21 m^2	5.31^o	0.22
371	1.61 m	1.42 m^2	5.30^o	0.25
541	1.05 m	1.48 m^2	7.70^o	0.23

Table 2. Errors and uncertainties corresponding to Figure 9

4.3 Estimating uncertainty with synthetic images

In this experiment, five networks were fine-tuned with synthetic datasets and were tested with two types of real images, to estimate the best type of image suitable for performing cross-domain pose regression. Moreover, we investigate whether the uncertainties of the estimated camera poses from a network fine-tuned with synthetic images are correlated with actual errors. The location and rotation errors, uncertainties and the distribution of the estimated points are accounted.

Three datasets namely Syn-car, Syn-pho-real and Syn-pho-real-tex are similar in appearance to the real images. Therefore we test networks fine-tuned with the three datasets directly with real images. However, the two networks fine-tuned with Gradmag-Syn-car and Syn-edge datasets are tested with edge gradmag of real images. Figure 8 (a) - (e) shows the distribution of the estimated camera poses for real images with errors by the networks. Figure 8 (f) - (j) shows the locational and rotational uncertainties of the estimated camera poses.

In Figure 8 the networks fine-tuned on edge gradmag images (Gradmag-Syn-car and Syn-edge) perform better than the other networks. There is approximately two-fold accuracy improvement for the networks fine-tuned on edge gradmag images (Gradmag-Syn-car and Syn-edge), compared to Syn-car, Syn-pho-real and Syn-pho-real-tex images. The best camera pose estimations (both locational and rotational) came from the network fine-tuned with Gradmag-Syn-car dataset and tested with edge gradmag of real images, followed by the network fine-tuned with Syn-edge images.

Additionally, the distribution of the estimated locations from the network fine-tuned with Gradmag-Syn-car images (Figure 8 (d)) is best amongst other image types, followed by Syn-edge images (Figure 8 (e)). Figure 8 (i) - (j) show higher correlation for locational and rotational uncertainties, for both the edge gradmag images, Gradmag-Syn-car being the most correlated.

It is observed from Figure 8 that the locational errors for the network fine-tuned with Syn-pho-real-tex is least, as compared to Syn-car and Syn-pho-real images, but the errors of all the three networks are comparable. The possible explanation could be the presence of the synthetic textures in Syn-pho-real-tex dataset that make this type of images appear similar to the real image. However, the rotational errors of network fine-tuned with Syn-car dataset are least, followed by networks fine-tuned with Syn-pho-real-tex and Syn-car images.

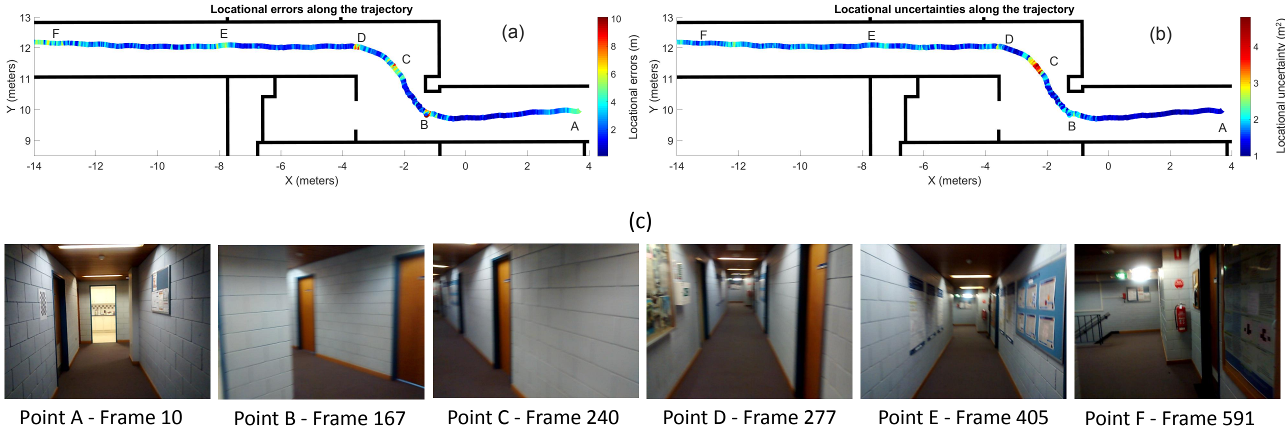
From Figure 8 (a) - (c) it is noted that the distribution of the estimated locations by the networks fine-tuned with Syn-car and Syn-pho-real are poor as compared to Syn-pho-real-tex images dataset. For all three networks, the estimated poses result in cluster formation away from the ground truth trajectory. Additionally, there exists a shift in the estimated camera locations, indicating the presence of a bias. Regarding uncertainties, from Figure 8 (f) - (h), it is seen that the network fine-tuned with Syn-car and Syn-pho-real-tex has the similar correlation for locational and rotational uncertainties, as compared to very low correlation for Syn-pho-real dataset. This fact explains the high errors of the Syn-pho-real images as compared to other image types.

Figure 9 shows the overlay of the 3D model on five selected test images using the camera poses estimated by the network fine-tuned with Gradmag-Syn-car dataset and Table 2 shows the errors and uncertainties of the corresponding camera poses. The locational and rotational errors in those frames varied from 0.24 - 1.61 meters and 2.77^o - 7.70^o respectively. The range of locational and rotational uncertainties was 1.01 - 1.48 square meters and 0.22 - 0.25 respectively.

Figure 10 (a) - (b) show the trend of the locational errors and uncertainties respectively for visualising the most error-prone and uncertain areas of the trajectory. Figure 10 (c) shows the images corresponding to the large estimation errors. In Figure 10 (a) - (b) it can be seen that the most substantial errors and uncertainties are present in part BC of the trajectory, which corresponds to a 90^o turn and a few large errors exists near Point D, E and F.

The images with large errors, as seen in Figure 10 (c), contain significant amount of either motion blur or artefacts. Motion blur results in low values of edge gradmag of the real images, which results in loss of information and explain the large errors in Points B, C and D of the trajectory. Artefacts are objects that are present in the real image but not in the 3D model, such as posters and notice boards and results in higher errors for Points A, D, E and F. Additionally, artefacts are also introduced due to the structural difference of the 3D model and the building, which are introduced as a result of errors in modelling the building.

Table 3 shows the errors and uncertainties of the frames shown in Figure 10 (c) with high errors. The locational and rotational errors in those frames varied from 6.74 - 10.10 meters and 9.01^o - 62.10^o respectively. The range of locational and rotational uncertainties was 1.45 - 3.09 square meters and 0.24 - 0.35 respectively. Comparing Table 3 with Table 2 indicate that larger uncertainties are associated with images having larger errors.



Point A - Frame 10
Point B - Frame 167
Point C - Frame 240
Point D - Frame 277
Point E - Frame 405
Point F - Frame 591

Figure 10. (a) shows the average normalised errors of the trajectory (b) shows the average normalised uncertainties of the trajectory and (c) shows corresponding points on the trajectory with high errors.

Frame	Location error	Location uncertainty	Rotational error	Rotational uncertainty
10	6.97 m	2.86 m^2	10.05^o	0.25
167	10.10 m	1.99 m^2	62.10^o	0.27
240	6.93 m	3.09 m^2	42.73^o	0.35
277	8.07 m	2.47 m^2	19.02^o	0.25
405	6.74 m	1.45 m^2	23.34^o	0.27
591	8.19 m	1.90 m^2	9.01^o	0.24

Table 3. Errors and uncertainties corresponding to Figure 10 (c)

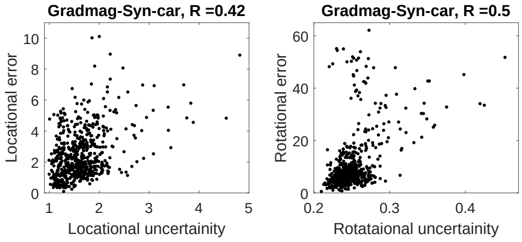


Figure 11. (a) Location uncertainty vs error (b) Rotational uncertainty vs errors for Gradmag-Syn-car images.

Figure 11 shows the correlation of the locational errors with locational uncertainties and rotational errors with rotational uncertainties. Figure 11 (a) indicates the locational uncertainties are more correlated with the location errors as compared to baseline experiment with real images in Figure 7 (b). Similar trend is observed for rotational errors and uncertainties when comparing Figure 11 (b) with 7 (c). Thus, the estimated uncertainties can be used as a measure of errors and are particularly useful in the absence of ground truth, as it provides the confidence with which we can trust the camera pose estimations.

5. DISCUSSION AND FUTURE DIRECTIONS

In indoor spaces, specifically with repetitive structures, two places might appear similar, and it is impossible to differentiate between them without some auxiliary information. This problem is referred to as perceptual aliasing (Lowry et al., 2016) and is also a limitation of the current approach. For example, consider frame 240 in Figure 10 (c), where the camera is close to the wall, and only one door is visible. As all the doors and walls in the building have the same appearance, it creates confusion for pose estimation. This phenomenon can explain the high errors and uncertainties near part Point C of the trajectory in Figure 10. Similarly, the high errors in the benchmark experiment with real images can be explained based on this fact.

The proposed approach is independent of any central processing server and hence does not need any additional infrastructure for operation, in addition, the network is very lightweight concerning storage (approximately 50 MB). The processing time for each image on a GPU is approximately 67 milliseconds and 9.07 seconds on an i7 CPU. Fortunately, most of the modern pervasive devices contain a GPU that can be used to accelerate the processing.

The achievable accuracy by the localisation approach is suitable for pedestrian navigation (Taneja et al., 2011) but is not sufficient for robotic applications; however, there is still room for improvement. A potential future direction to overcome the perceptual aliasing could be the use of Kalman filter to integrate the uncertainties derived from the network with the spatio-temporal information from a video sequence, or the use of long-short-term-memory (LSTM) to exploit spatio-temporal information alone. On the other hand, to address the computational challenge, another future direction for the research could be exploring mini-networks that are made explicitly for operation in mobile devices.

6. CONCLUSIONS

We present a visual indoor localisation approach that can provide the initial location of a camera for existing visual localisation approaches, such as SLAM, visual odometry and model-based visual tracking for pedestrian navigation. We address a challenge of the existing approaches by eliminating the requirement of performing SfM approaches before the operation. We fine-tune a pre-trained deep Bayesian CNN with different types of synthetic images generated from a 3D model and to regress the camera pose of real images. The experiments indicate that a localisation accuracy of approximately 2 meters can be achieved by representing real images as edge gradient magnitude images. The localisation errors are the result of the presence of motion blur, artefacts in the images and repetitive structures. Moreover, the uncertainties associated with the estimated locations are correlated with the localisation errors, thus providing the measure with which we can trust the estimated locations. The current research point towards interesting future directions that can improve the performance of such an approach.

REFERENCES

Acharya, D., Khoshelham, K. and Winter, S., 2017. Real-time detection and tracking of pedestrians in cctv images using a deep convolutional neural network. In: *Proceedings of the 4th Annual Conference of Research@Locate*, Vol. 1913, pp. 31–36.

Acharya, D., Khoshelham, K. and Winter, S., 2019a. BIM-PoseNet: Indoor camera localisation using a 3D indoor model and deep learning from synthetic images. *ISPRS Journal of Photogrammetry and Remote Sensing* 150, pp. 245–258.

Acharya, D., Ramezani, M., Khoshelham, K. and Winter, S., 2019b. BIM-Tracker: A model-based visual tracking approach for indoor localisation using a 3D building model. *ISPRS Journal of Photogrammetry and Remote Sensing* 150, pp. 157–171.

Acharya, D., Yan, W. and Khoshelham, K., 2018. Real-time image-based parking occupancy detection using deep learning. In: *Proceedings of the 5th Annual Conference of Research@Locate*, Vol. 2087, pp. 33–40.

Arandjelovic, R. and Zisserman, A., 2012. Three things everyone should know to improve object retrieval. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, Vol. 00, pp. 2911–2918.

Clark, R., Wang, S., Markham, A., Trigoni, N. and Wen, H., 2017. Vidloc: A deep spatio-temporal model for 6-dof video-clip relocalization. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Vol. 3.

Gal, Y. and Ghahramani, Z., 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: *international conference on machine learning*, pp. 1050–1059.

Groves, P. D., 2013. *Principles of GNSS, inertial, and multisensor integrated navigation systems*. Artech house.

Irschara, A., Zach, C., Frahm, J. M. and Bischof, H., 2009. From structure-from-motion point clouds to fast location recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2599–2606.

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S. and Darrell, T., 2014. Caffe: Convolutional architecture for fast feature embedding. In: *Proceedings of the 22nd ACM International Conference on Multimedia*, pp. 675–678.

Kendall, A. and Cipolla, R., 2016. Modelling uncertainty in deep learning for camera relocalization. In: *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4762–4769.

Kendall, A., Cipolla, R. et al., 2017. Geometric loss functions for camera pose regression with deep learning. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Vol. 3, p. 8.

Kendall, A., Grimes, M. and Cipolla, R., 2015. Posenet: A convolutional network for real-time 6-dof camera relocalization. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2938–2946.

Khoshelham, K., Tran, H., Vilariño, L. D., Peter, M., Kang, Z. and Acharya, D., 2018. An evaluation framework for benchmarking indoor modelling methods. *ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* Vol. XLII-4, pp. 297–302.

Khoshelham, K., Vilariño, L. D., Peter, M., Kang, Z. and Acharya, D., 2017. The ISPRS benchmark on indoor modelling. *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences* Vol. XLII-2/W7, pp. 367–372.

Krizhevsky, A., Sutskever, I. and Hinton, G. E., 2012. Imagenet classification with deep convolutional neural networks. In: *Proceedings of the Advances in neural information processing systems (NIPS)*, pp. 1097–1105.

Lowry, S., Sünderhauf, N., Newman, P., Leonard, J. J., Cox, D., Corke, P. and Milford, M. J., 2016. Visual place recognition: A survey. *IEEE Transactions on Robotics* 32(1), pp. 1–19.

MacKay, D. J. C., 1992. A practical bayesian framework for backpropagation networks. *Neural Computation* 4(3), pp. 448–472.

Mautz, R., 2012. Indoor positioning technologies (habilitation). *ETH Zurich, Department of Civil, Environmental and Geomatic Engineering, Institute of Geodesy and Photogrammetry Zurich.*

Parisotto, E., Chaplot, D. S., Zhang, J. and Salakhutdinov, R., 2018. Global pose estimation with an attention-based recurrent network. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 350–359.

Peng, X., Sun, B., Ali, K. and Saenko, K., 2015. Learning deep object detectors from 3d models. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1278–1286.

Piasco, N., Sidib, D., Demonceaux, C. and Gouet-Brunet, V., 2018. A survey on visual-based localization: On the benefit of heterogeneous data. *Pattern Recognition* 74, pp. 90 – 109.

Radenović, F., Tolias, G. and Chum, O., 2016. Cnn image retrieval learns from bow: Unsupervised fine-tuning with hard examples. In: *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 3–20.

Ramezani, M., Acharya, D., Gu, F. and Khoshelham, K., 2017. Indoor positioning by visual-inertial odometry. *ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences* VOL. IV-2/W4, pp. 371–376.

Se, S., Lowe, D. and Little, J., 2002. Global localization using distinctive visual features. In: *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, Vol. 1, pp. 226–231.

Shotton, J., Glocker, B., Zach, C., Izadi, S., Criminisi, A. and Fitzgibbon, A., 2013. Scene coordinate regression forests for camera relocalization in rgb-d images. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2930–2937.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research* 15(1), pp. 1929–1958.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. and Rabinovich, A., 2015. Going deeper with convolutions. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9.

Taneja, S., Akcamete, A., Akinci, B., Garrett Jr, J. H., Soibelman, L. and East, E. W., 2011. Analysis of three indoor localization technologies for supporting operations and maintenance field tasks. *Journal of Computing in Civil Engineering* 26(6), pp. 708–719.

Walch, F., Hazirbas, C., Leal-Taixe, L., Sattler, T., Hilsenbeck, S. and Cremers, D., 2017. Image-based localization using lstms for structured feature correlation. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 627–637.

Wang, S., Clark, R., Wen, H. and Trigoni, N., 2017. Deepvo: Towards end-to-end visual odometry with deep recurrent convolutional neural networks. In: *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2043–2050.

Wu, J., Ma, L. and Hu, X., 2017. Delving deeper into convolutional neural networks for camera relocalization. In: *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5644–5651.

Zhou, B., Lapedriza, A., Xiao, J., Torralba, A. and Oliva, A., 2014. Learning deep features for scene recognition using places database. In: *Advances in neural information processing systems*, pp. 487–495.