Neurocomputing 275 (2018) 1861–1870

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Deep EndoVO: A recurrent convolutional neural network (RCNN) based visual odometry approach for endoscopic capsule robots

Mehmet Turan[a,d,*], Yasin Almalioglu[b], Helder Araujo[c], Ender Konukoglu[d], Metin Sitti[a]

[a] Physical Intelligence Department, Max Planck Institute for Intelligent Systems, Stuttgart, Germany
[b] Computer Engineering Department, Bogazici University, Istanbul, Turkey
[c] Institute for Systems and Robotics, University of Coimbra, Coimbra, Portugal
[d] Department of Information Technology and Electrical Engineering, Computer Vision Laboratory, ETH Zurich, Zurich, Switzerland

ARTICLE INFO

Article history:
Received 16 February 2017
Revised 23 August 2017
Accepted 6 October 2017
Available online 7 November 2017

Communicated by Wei Wu

Keywords:
Endoscopic capsule robot
Visual odometry
Sequential deep learning
RCNN
CNN
LSTM
Localization

ABSTRACT

Ingestible wireless capsule endoscopy is an emerging minimally invasive diagnostic technology for inspection of the GI tract and diagnosis of a wide range of diseases and pathologies. Medical device companies and many research groups have recently made substantial progresses in converting passive capsule endoscopes to active capsule robots, enabling more accurate, precise, and intuitive detection of the location and size of the diseased areas. Since a reliable real time pose estimation functionality is crucial for actively controlled endoscopic capsule robots, in this study, we propose a monocular visual odometry (VO) method for endoscopic capsule robot operations. Our method lies on the application of the deep recurrent convolutional neural networks (RCNNs) for the visual odometry task, where convolutional neural networks (CNNs) and recurrent neural networks (RNNs) are used for the feature extraction and inference of dynamics across the frames, respectively. Detailed analyses and evaluations made on a real pig stomach dataset proves that our system achieves high translational and rotational accuracies for different types of endoscopic capsule robot trajectories.

1. Introduction

Following the advances in material science in last decades, untethered pill-size, swallowable capsule endoscopes with an onboard camera and wireless image transmission device have been developed and used in hospitals for screening the gastrointestinal tract and diagnosing diseases such as the inflammatory bowel disease, the ulcerative colitis and the colorectal cancer. Unlike standard endoscopy, endoscopic capsule robots are non-invasive, painless and more appropriate to be employed for long duration screening purposes. Moreover, they can access difficult body parts that were not possible to reach before with standard endoscopy (e.g., small intestines). Such advantages make pill-size capsule endoscopes a significant alternative screening method over standard endoscopy [1–5]. However, current capsule endoscopes used in hospitals are passive devices controlled by peristaltic motions of the inner organs. The control over capsule's position, orientation, and functions would give the doctor a more precise reachability of targeted body parts and more intuitive and correct diagnosis opportunity [6–10]. Therefore, several groups have recently proposed active, remotely controllable robotic capsule endoscope prototypes equipped with additional functionalities such as local drug delivery, biopsy and other medical functions [2,11–19]. However, an active motion control needs feedback from a precise and reliable real time pose estimation functionality. In last decade, several localization methods [4,20–23] were proposed to calculate the 3D position and orientation of the endoscopic capsule robot such as fluoroscopy [4], ultrasonic imaging [20–23], positron emission tomography (PET) [4,23], magnetic resonance imaging (MRI) [4], radio transmitter based techniques and magnetic field based techniques [16]. The common drawback of these localization methods is that they require extra sensors and hardware design. Such extra sensors have their own deficiencies and limitations if it comes to their application in small scale medical devices such as space limitations, cost aspects, design incompatibilities, biocompatibility issue and the interference of sensors with activation system of the device.

* Corresponding author at: Physical Intelligence Department, Max Planck Institute for Intelligent Systems, Stuttgart, Germany.
E-mail addresses: mturan@student.ethz.ch (M. Turan), yasin.almalioglu@boun.edu.tr (Y. Almalioglu), helder@isr.uc.pt (H. Araujo), ender.konukoglu@vision.ee.ethz.ch (E. Konukoglu), sitti@is.mpg.de (M. Sitti).

https://doi.org/10.1016/j.neucom.2017.10.014

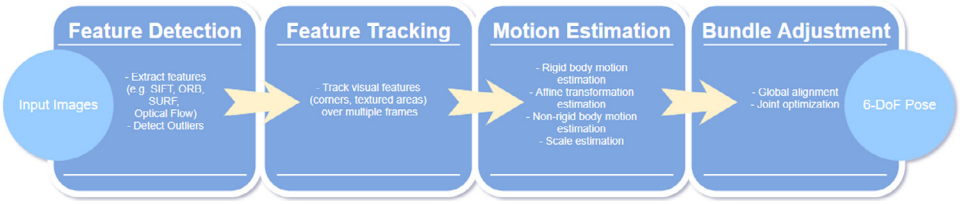


Fig. 1. Traditional visual odometry pipeline.

As a solution of these issues, a trend of visual odometry methods have attracted the attention for the localization of such small scale medical devices. A classic visual odometry pipeline typically consisting of camera calibration, feature detection, feature matching, outliers rejection (e.g. RANSAC), motion estimation, scale estimation and global optimization (bundle adjustment) is depicted in Fig. 1. Although some state-of-the-art algorithms based on this traditional pipeline have been applied for the visual odometry task of the hand-held endoscopes in the past decades, their main deficiency is tracking failures in low textured areas. In last years, deep learning (DL) techniques have been dominating many computer vision related tasks with some promising result, e.g. object detection, object recognition, classification problems etc. Contrary to these high-level computer vision tasks, VO is mainly working on motion dynamics and relations across sequence of images, which can be defined as a sequential learning problem. With that motivation, we propose a novel monocular VO algorithm based on deep recurrent convolutional neural networks (RCNNs). Since it is designed in an end-to-end fashion, it does not need any module from the classic VO pipeline to be integrated. The main contributions of our paper are as follows:

- To the best of our knowledge, this is the first monocular VO approach through deep learning techniques developed for the endoscopic capsule robot and hand-held standard endoscope localization.
- Neither prior knowledge nor parameter tuning is needed to recover the absolute trajectory scale contrary to monocular traditional VO approach.
- A novel RCNN architecture is introduced which can successfully model sequential dependence and complex motion dynamics across endoscopic video frames.
- A real pig stomach dataset and a synthetic human simulator dataset with 6-DoF ground truth pose labels and 3D scan are recorded, which we are considering to publish for the sake of other researchers in that area.

The proposed method solves several issues faced by typical visual odometry pipelines, e.g. the need to establish a frame-to-frame feature correspondence, vignetting, motion blur, specularity or low signal-to-noise ratio (SNR). We think that DL based endoscopic VO approach is more suitable for such challenge areas since the operation environment (GI tract) has similar organ tissue patterns among different patients which can be learned by a sophisticated machine learning approach easily. Even the dynamics of common artefacts such as vignetting, motion blur and specularity across frame sequences could be learned and used for a better pose estimation.

As the outline of this paper, Section 2 introduces the proposed RCNN based localization method in detail. Section 3 presents our dataset and the experimental setup. Section 4 shows our experimental results, we achieved for 6-DoF localization of the endoscopic capsule robot. Section 5 gives future directions.

2. System overview and analysis

Our architecture makes use of inception modules for feature extraction and RNN for sequential modelling of motion dynamics to regress the robot's orientation and position in real time (5.3 ms per frame). It takes two consecutive endoscopic RGB Depth frames each with timestamp and regresses the 6-DoF pose of the robot without need of any extra sensor. For the depth image creation from RGB input images, we used shape from shading (SfS) technique of Tsai and Shah, which is based on the following assumptions [24]:

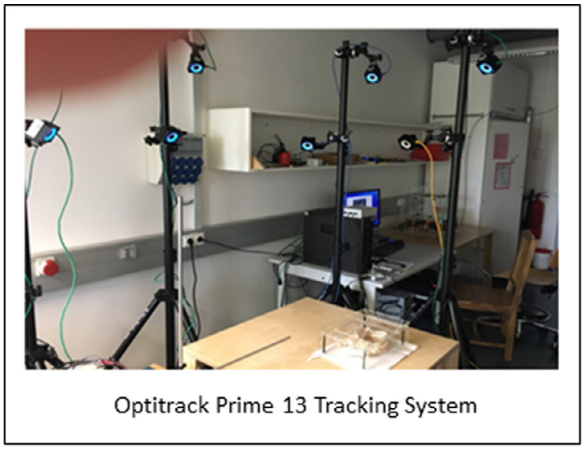
- The object surface is Lambertian;
- The light comes from a single point light source;
- The surface has no self-shaded areas.

For more details of the Tsai–Shah SfS method, the reader is referred to the original paper of the authors. In past couple of years, some powerful CNN architectures, such as GoogleNet [25], VGG16 [26], ResNet50 [27] have been developed and evaluated for various high level computer vision tasks, e.g. object detection, object recognition and classification [25,28–30]. One major drawback of CNN architectures is the fact that they only analyse just-in-moment information, whereas VO is rather dependent on the correlative information across frames. Unlike traditional feed-forward artificial neural networks, RCNN can use its internal memory to process arbitrarily long sequences by its directed cycles between the hidden units. Therefore, we think that RCNN architectures are more suitable than CNN architectures for VO tasks. The proposed deep EndoVO (endoscopic visual odometry) approach works as follows:

Algorithm 1 Deep EndoVO.

1: Take two consecutive input RGB images.
2: Create the depth images from RGB images using Tsai–Shah SfS method.
3: Subtract mean RGB Depth value of the training set from the RGB Depth images.
4: Stack the preprocessed RGB Depth frame pair to form a tensor.
5: Serve the tensor into the stack of inception modules to create the feature vector.
6: Feed the feature representation into the RNN layers.
7: Estimate the 6-DoF relative pose.

The proposed DL network consists of three inception layers and two LSTM layers concatenated sequentially. The inception layers, imitating visual cortex of human beings, are basically extracting multi-level features; i.e, features of different sizes such as small details, middle-size or larger features (see Fig. 3b). The final inception layer passes the feature representation into the RNN modules (see Fig. 3a). RNNs are very suitable for modelling the dependencies across image sequences and for creating a temporal motion model since it has a memory of hidden states over time and has directed cycles among hidden units, enabling the current hidden state to be a function of arbitrary sequences of inputs (see Fig. 3a).



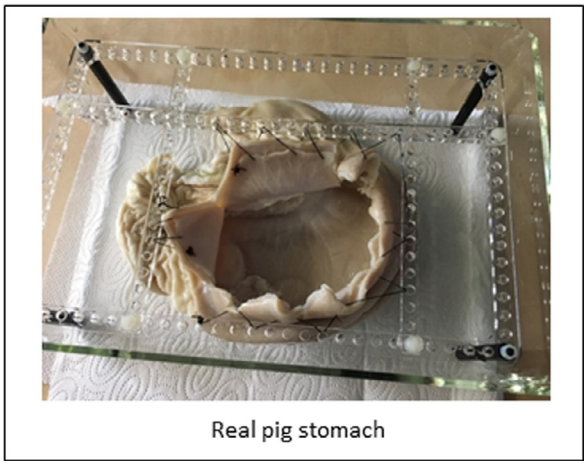
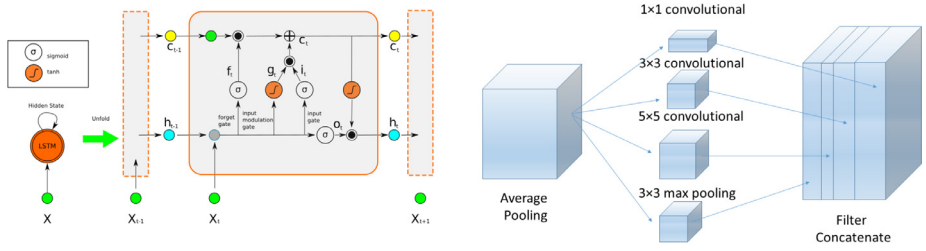


Fig. 2. Experimental overview.



(a) Information flow through the hidden units of the LSTM [31].

(b) Inception layer[25]

Fig. 3. The structure of the LSTM and inception layers of the proposed model is shown.

Thus, using RNN, the pose estimation of the current frame benefits from information encapsulated in previous frames [32,33]. Given a set of inception features x_k at time k, RNN updates at time step k, W denote corresponding weight matrices of the hidden units, b the bias vector, and H an element-wise hyperbolic tangent based activation function. Long short-term memory (LSTM) is more suitable than RNN to exploit longer trajectories since it avoids the vanishing gradient problem of RNN resulting in a higher capacity of learning long-term relations among the sequences by introducing memory gates such as input, forget and output gates and hidden units of several blocks. The input gate controls the amount of new information flowing into the current state, the forget gate adjusts the amount of existing information that remains in the memory and the output gate decides which part of the information triggers the activations. The folded LSTM and its unfolded version over time are shown in Fig. 3a along with the internal structure of a LSTM memory cell. It can be seen that unfolded LSTMs correspond to timestamps. Given the input vector x_k at time k, the output vector h_{k-1} and the cell state vector c_{k-1} of the previous LSTM unit, the LSTM updates at time step k according to the following equations, where σ is sigmoid non-linearity, tanh is hyperbolic tangent non-linearity, W terms denote corresponding weight matrices, b terms denote bias vectors, i_k, f_k, g_k, c_k and o_k are input gate, forget gate, input modulation gate, the cell state and output gate at time k, respectively [31]:

$$f_k = \sigma\left(W_f \cdot [x_k, h_{k-1}] + b_f\right) \qquad i_k = \sigma\left(W_i \cdot [x_k, h_{k-1}] + b_i\right)$$
$$g_k = \tanh(W_g \cdot [x_k, h_{k-1}] + b_g) \qquad c_k = f_k \odot c_{k-1} + i_k \odot g_k$$
$$o_k = \sigma\left(W_o | cdot [x_k, h_{k-1}] + b_o\right) \qquad h_k = o_k \odot \tanh(c_k)$$

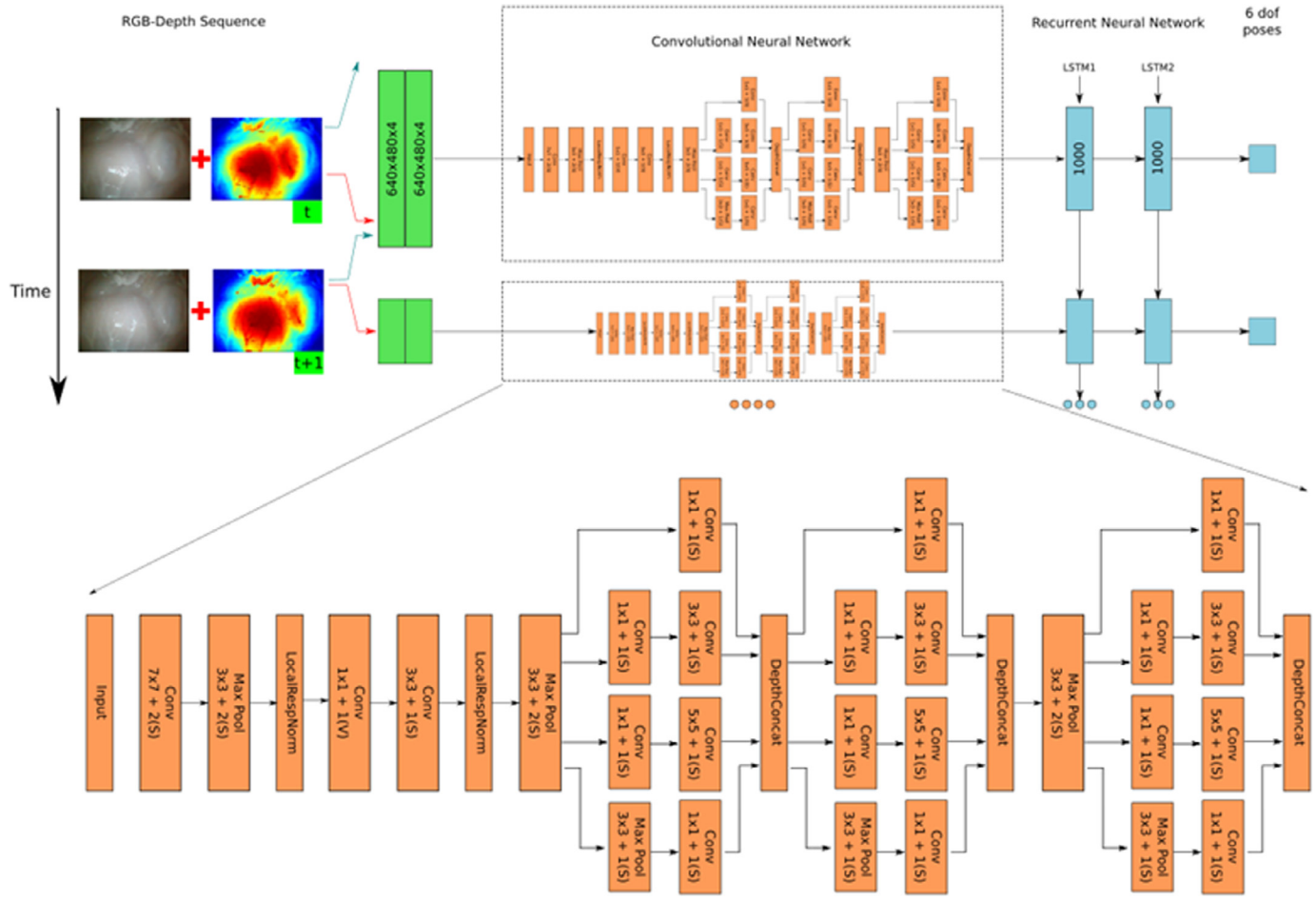


Fig. 4. Architecture of the proposed RCNN based monocular VO system.

Although the LSTM is prone to vanishing gradient problem of RNN and is capable to detect the long-term dependencies, its learning capacity can be increased further by stacking multiple LSTM layers vertically. Thus, our deep RNN consists of two LSTM layers with the output sequence of the first one forming the input sequence of the second one each containing 1000 hidden units, as illustrated in Fig. 4. The proposed system, which learns translational and rotational motions simultaneously to regress the 6-DoF pose, is trained on Euclidean loss using Adam optimization method with the following objective loss function:

$$\text{loss}(I) = \|\hat{x} - x\|_2 + \beta \|\hat{q} - q\|_2 \tag{1}$$

where x is the translation vector and q is the rotation vector. The pseudo-code to calculate the loss value is given in Algorithm 2. In our loss function, a balance β must be kept between the orientation and translation loss values which are highly coupled each other as they are learned from the same model weights. Experimental results show that the optimal β is given by the ratio between the loss values of predicted positions and orientations at the end of training session [30].

Algorithm 2 Pseudo code to calculate the loss over the network.

```
procedure CalculateLoss
    loss ← 0
    for layer in layers do
        for top, loss_weight in layer.tops, layer.loss_weights do
            loss ← loss + loss_weight × sum(top)
```

The back-propagation algorithm is used to calculate the gradients of RCNN weights, which are passed to the Adam optimization method to compute adaptive learning rates for each parameter employing the first-order gradient-based optimization of the stochastic objective function. In addition to saving exponentially decaying average of past squared gradients, v_t, Adam optimization keeps exponentially decaying average of past gradients, m_t that is similar to momentum. The update equations are given as

$$(m_t)_i = \beta_1 (m_{t-1})_i + (1 - \beta_1)(\nabla L(W_t))_i \tag{2}$$

$$(v_t)_i = \beta_2 (v_{t-1})_i + (1 - \beta_2)(\nabla L(W_t))_i^2 \tag{3}$$

$$(W_{t+1})_i = (W_t)_i - \alpha \frac{\sqrt{1 - (\beta_2)_i^t}}{1 - (\beta_1)_i^t} \frac{(m_t)_i}{\sqrt{(v_t)_i + \varepsilon}} \tag{4}$$

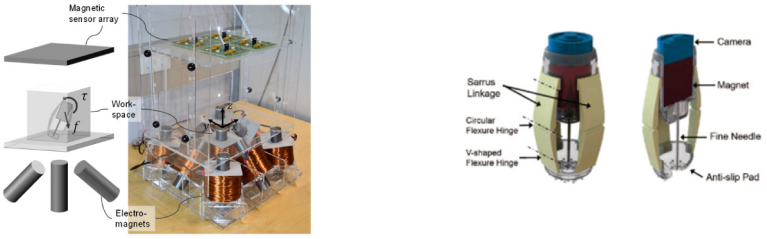
We used default values proposed by [34] for the parameters β_1, β_1 and ε: $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\varepsilon = 10^{-8}$.

3. Dataset

This section demonstrates the experimental setup of the proposed study, introduces our magnetically actuated soft capsule endoscopes (MASCE) and explains how the training and testing datasets were recorded.

3.1. Magnetically actuated soft capsule endoscopes (MASCE)

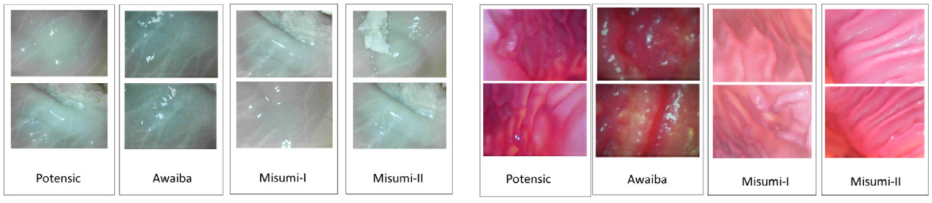
Our capsule prototype is a magnetically actuated soft capsule endoscope (MASCE) designed for disease detection, drug delivery



(a) Actuation system of the MASCE [19]

(b) Exterior (left) and section view (right) of MASCE [19]

Fig. 5. MASCE design features and actuation unit.



(a) Sample frames recorded on a real pig stomach

(b) Sample frames recorded on EGD simulator

Fig. 6. Sample frames from the datasets used in the experiments.

Table 1
Endoscopic camera specifications used for the experiments.

(a) Awaiba Naneye Endoscopic Camera		(b) Misumi-V3506-2ES Camera	
Resolution	250×250 pixel	Resolution	400×400 pixel
Footprint	$2.2 \times 1.0 \times 1.7$ mm	Diameter	8.2 mm
Pixel size	3×3 μm^2	Pixel size	5.55×5.55 μm^2
Frame rate	44 fps	Frame rate	30 fps
(c) Misumi-V3506-2ES Camera		(d) Potensic Mini Camera	
Resolution	640×480 pixel	Resolution	1280×720 pixel
Diameter	8.6 mm	Diameter	8.8 mm
Pixel size	6.0×6.0 μm^2	Pixel size	10.0×10.0 μm^2
Frame rate	30 fps	Frame rate	30 fps

and biopsy operations in the upper gastrointestinal tract. The prototype is composed of a RGB camera, a permanent magnet, a fine-needle and a drug chamber (see Fig. 5 for visual reference). The magnet exerts magnetic force and torque to the robot in response to a controlled external magnetic field [19]. The magnetic torque and forces are used to actuate the capsule robot and to release drug and deliver the needle through the hole in the bottom of the capsule. Magnetic fields from the electromagnets generate the magnetic force and torque on the magnet inside MASCE so that the robot moves inside the workspace. Sixty-four three-axis magnetic sensors are placed on the top, and nine electromagnets are placed in the bottom [19].

3.2. Training dataset

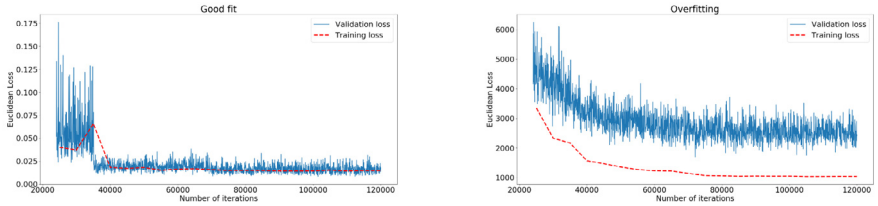
We created two groups of training datasets. The first training dataset was recorded on five different real pig stomachs (see Fig. 2), whereby the second dataset which was only used for training purposes, was captured using a non-rigid open GI tract model EGD (esophagus gastro duodenoscopy) surgical simulator LM-103 (see Fig. 2). To ensure that our algorithm is not tuned to a specific camera model, four different commercial endoscopic cameras were employed, specifications of which are shown in Table 1, accordingly. For each pig stomach-camera combination, 2000 frames were acquired which makes for four cameras and five pig stomachs 40, 000 frames, in total. Sample real pig stomach frames are shown in Fig. 6a for visual reference. As a second training dataset, for each of four cameras, we captured 10, 000 frames on an EGD human stomach simulator making 40, 000 frames, in total. Sample synthetic training frames are shown in Fig. 6b for visual reference. During video recording, Optitrack motion tracking system consisting of eight Prime-13 cameras and a tracking software was utilized to obtain 6-DoF localization ground truth data in a sub-millimeter precision (see Fig. 2) which was used as a gold standard for the evaluations of the pose estimation accuracy.

3.3. Testing dataset

We created a testing dataset recorded using five different real pig stomachs, which were not used for the training section. For each pig stomach-camera combination, 2000 frames are acquired making 40, 000 frames, in total. We did not capture any synthetic dataset for the testing session since it is less realistic due to obvious patterns of such artificial simulators. For all of the video records, again Optitrack motion tracking system was utilized to obtain 6-DoF localization ground truth.

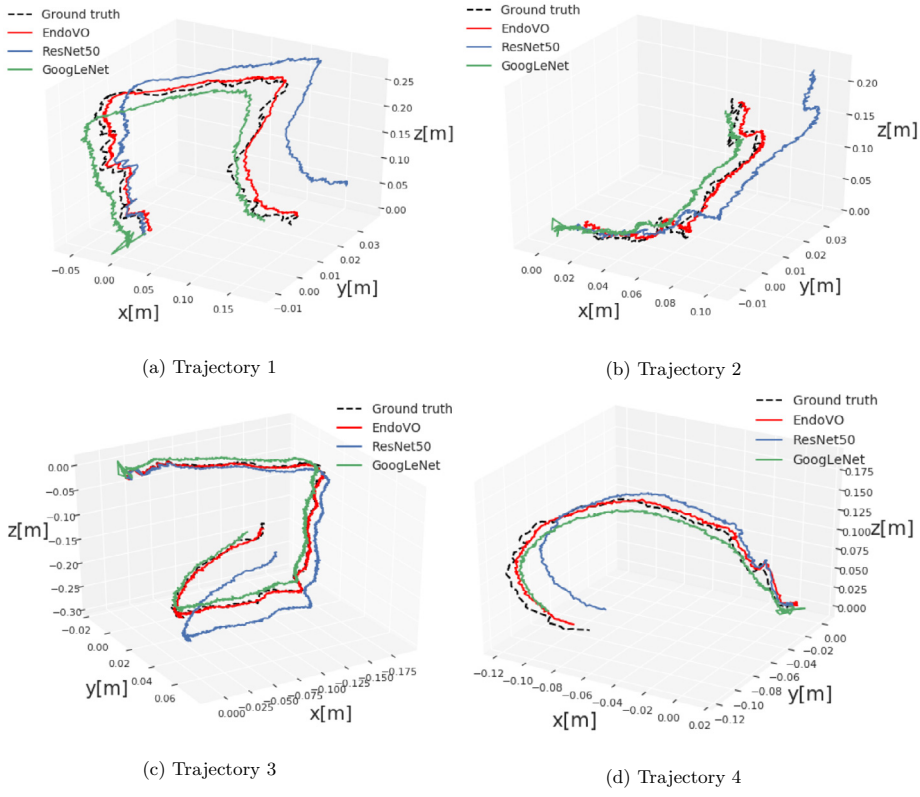
4. Evaluations and results

Architecture was trained using Caffe library and NVIDIA Tesla K40 GPU. Using back-propagation-through-time method, the weights of hidden units were trained for up to 200 epochs with an initial learning rate of 0.001. Overfitting meaning that the noise or random fluctuations in the training data are picked up and learned as concepts by the model, whereas these concepts do not apply to a new data and negatively affect the ability of the model to make generalizations, was prevented using dropout and early stopping techniques (see Fig. 10). Dropout regularization technique introduced by [35] is an extremely effective and simple method to avoid overfitting. It samples a part of the whole network and updates its parameters based on the input data. Early stopping is another widely used technique to prevent overfitting of a complex neural network architecture which was optimized by a gradient-based method. The approach is executed by splitting the dataset into a training and a validation set to evaluate the generalization capability of the model.



(a) Change in the loss values for a good fitting (b) Change in the loss values for overfitting

Fig. 7. The decrease in the training and validation loss values. In overfitting case, the training loss gets smaller than the validation loss. However, the loss values are balanced for a good fit.

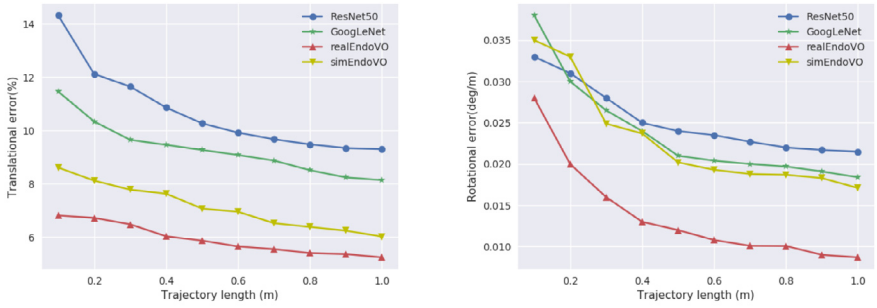


(a) Trajectory 1 (b) Trajectory 2

(c) Trajectory 3 (d) Trajectory 4

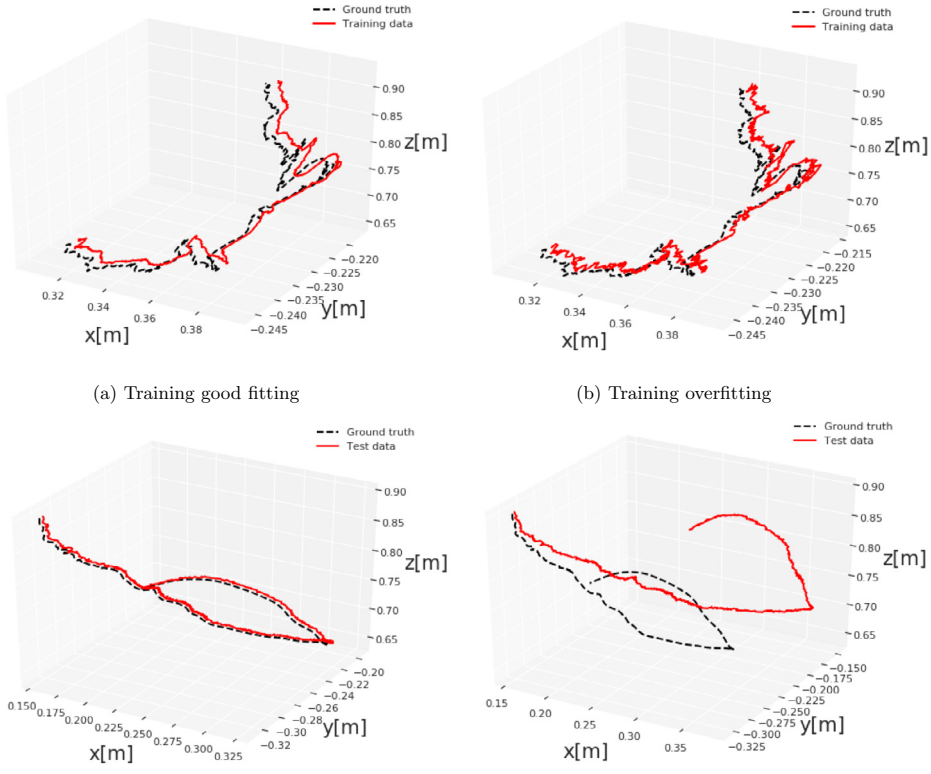
Fig. 8. Sample ground truth trajectories and estimated trajectories predicted by the DL based VO models. As seen, deep EndoVO is the closest to the ground truth trajectories. The scale is calculated and maintained correctly by the models.

For the testing sessions, only real pig stomach recordings were used to ensure real world conditions. Additionally, we strictly avoided to use any frame from the training session for the testing session. Two separate experiments were conducted, whereas training session of the first experiment was performed using only the synthetic training dataset (see Fig. 7b) which we call simEndoVO and training session of the second experiment was performed using frames from both synthetic and real pig stomach dataset (see Fig. 7b and a) which we call realEndoVO. The performance of the simEndoVO and realEndoVO approaches were analysed using averaged root mean square errors (RMSEs) for translational and rotational motions. For various trajectories with different complexity levels of motions, including uncomplicated paths with slow incremental translations and rotations, comprehensive scans with many local loop closures and complex paths with sharp rotational and translational movements, we performed testings on both simEndoVO and realEndoVO comparing them with GoogLeNet and ResNet50 architectures which were modified to regress 6-DoF pose values by removing softmax layer and integrating a fully-connected (FC) layer and an affine regressor layer. The average translational and rotational RMSEs for simEndoVO, realEndoVO, GoogLeNet and ResNet50 networks against different path lengths are shown in Fig. 9, respectively. The results depicted indicate, that realEndoVO clearly outperforms GoogLeNet and ResNet50, whereas simEndoVO slightly outperforms them. We presume that the effective use of LSTM in EndoVO architecture enabled learning motion dynamics across frame sequences, which is not feasible by architectures working with the principle of just-in-moment information processing; i.e. GoogleNet and ResNet50. The results in Fig. 9 also indicate that the training procedure including both simulator and real dataset was more informative than training only with simulator dataset. On the other hand, the accuracies achieved by the modified GoogLeNet are slightly better than accuracies achieved by the modified ResNet50, proving the superiority of inception layers over residual networks for feature extraction related tasks. Derived from RMSEs calculated, the rotational motion parameters seem to be more prone to overfitting compared to translational motion parameters (see Fig. 10 for visual reference). The reason for that observation could be the fact that inner organ scanning procedures generally contain more translational motions than rotational mo-



(a) Trajectory length vs translation error (b) Trajectory length vs totation error

Fig. 9. Deep EndoVO outperforms both of the other models in terms of translational and rotational position estimation.



(a) Training good fitting (b) Training overfitting

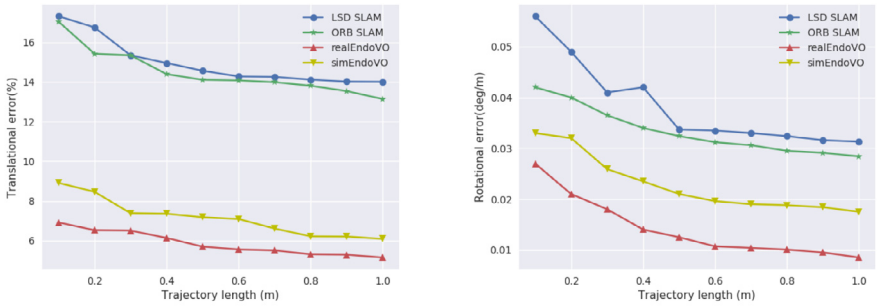
(c) Test good fitting (d) Test overfitting

Fig. 10. The affect of good fitting and overfitting. The first and the second rows show over-fitted and well-fitted models, respectively. As seen in subfigures, the model learns the details and noise in the training data to an undesired extent that it negatively impacts the performance of the model on the test data.

tions resulting in a better learning for translations. As the length of the trajectory increases, both the translational and rotational error of all the proposed models significantly decrease (see Fig. 9). Some sample ground truth and estimated trajectories for realEndoVO, GoogLeNet and ResNet50 are shown in Fig. 8 for visual reference. As seen in these sample trajectories, realEndoVO is able to stay close to the ground truth pose values for even sharp crispy motions, contrary to realEndoVO; GoogLeNet and ResNet50 path estimations which deviate drastically from the ground truth path values. Even for very fast and challenge paths such as Fig. 8a and c, the deviations of realEndoVO from the ground truth still remain in an acceptable range for medical operations. In addition to that, it is clearly seen that all of the three evaluated neural network architectures are able to estimate the scale very accurately without using any prior information or post alignment techniques contrary to traditional VO. Solving the scale ambiguity for monocular camera based VO makes our proposed DL based method more beneficial than traditional VO approach. As opposed to the traditional VO pipeline (see Fig. 1), the DL-based VO do not require any explicit feature extraction, matching, outlier detection or multi-scale bundle adjustment-like parameter tuning requiring operations, which can be seen as further benefits of the proposed approach.

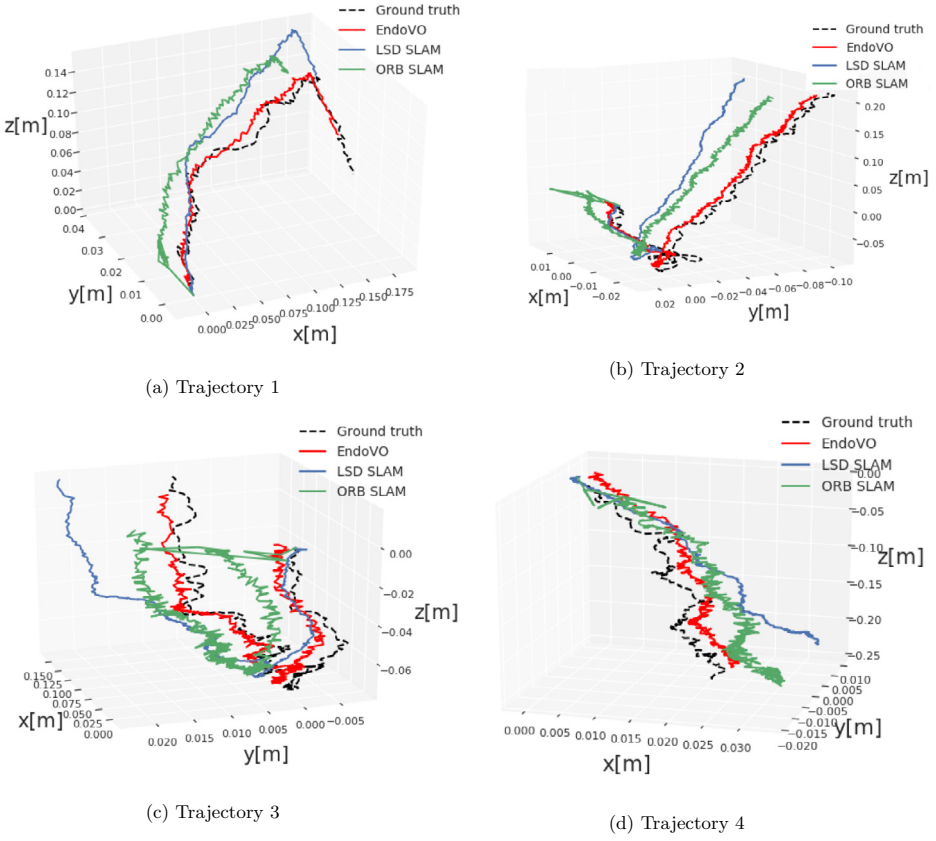
4.1. Comparisons of deep EndoVO with state-of-the-art SLAM methods

In this subsection, we compare the performance of the proposed deep EndoVO with two of the widely used state-of-the-art SLAM methods; i.e. large-scale direct monocular SLAM (LSD SLAM) [36] and the oriented fast and rotated brief SLAM (ORB SLAM)



(a) Trajectory length vs translation error (b) Trajectory length vs rotation error

Fig. 11. Deep EndoVO outperforms the state-of-the-art SLAM methods ORB SLAM and LSD SLAM in both the translation and orientation estimation.



(a) Trajectory 1 (b) Trajectory 2

(c) Trajectory 3 (d) Trajectory 4

Fig. 12. The ground truth and the trajectory plots acquired via deep EndoVO, LSD SLAM and ORB SLAM. Deep EndoVO is the closest to the ground truth trajectories compared to the state-of-the-art SLAM methods.

[37]. LSD SLAM is a direct image alignment-based method which optimizes the geometry using all of the image intensities. In addition to higher accuracy and robustness particularly in environments with little key points, this provides substantially more information about the geometry of the environment, which can be very valuable for medical robot applications, as well. ORB SLAM on the other hand, relies on feature point extraction and tracking to estimate camera pose and 3D map the environment. Even though it gives very promising results for feature-rich areas, its main deficiency appears once the robot enters poorly featured areas. Tracking failures are commonly observable for poorly featured GI tract tissues making ORB SLAM less proper for our case. We believe that our deep EndoVO architectures makes an optimal use of both direct and feature point information to estimate the pose. The average translational and rotational RMSEs for simEndoVO, realEndoVO, LSD SLAM and ORB SLAM, shown in Fig. 11 indicate that both simEndoVO and realEndoVO clearly outperform LSD SLAM and ORB SLAM in terms of pose accuracy. Sample trajectory estimations shown in Fig. 12 visualize clearly that the tracking capability of the proposed deep EndoVO is much more robust and reliable compared to LSD SLAM and ORB SLAM. In many parts of the trajectories, ORB SLAM and LSD SLAM deviate from the ground truth trajectory drastically, whereas deep EndoVO is still able to stay close to the ground truth values even for most challenge trajectory sections (see Fig. 12b and c).

5. Conclusion

In this study, we presented, to the best of our knowledge, the first deep VO method for endoscopic capsule robot and standard hand-held endoscope operations. The proposed system is able to achieve simultaneous representation learning and sequential mod-

elling of motion dynamics across frames by concatenating the inception modules with RNN layers. Many issues faced by traditional VO techniques such as feature correspondence establishment in low textured areas, high reflections, motion blur and low image quality are handled by the proposed deep EndoVO successfully. Since it is trained in an end-to-end manner, there is no need to carefully fine-tune the parameters of the system. As a future step, we consider to combine deep EndoVO with some functionalities from the traditional VO pipelines such as RANSAC for outlier detection and bundle fusion for globally consistent pose estimation etc to avoid drifts. Moreover, we consider to develop a stereo version of the proposed deep EndoVO approach.

References

[1] Z. Liao, R. Gao, C. Xu, Z.-S. Li, Indications and detection, completion, and retention rates of small-bowel capsule endoscopy: a systematic review, Gastrointestinal Endosc. 71 (2) (2010) 280–286.

[2] T. Nakamura, A. Terano, Capsule endoscopy: past, present, and future, J. Gastroenterol. 43 (2) (2008) 93–99.

[3] G. Pan, L. Wang, Swallowable wireless capsule endoscopy: progress and technical challenges, Gastroenterol. Res. Pract. 2012 (2012), Article ID 841691, 9 pages.

[4] T.D. Than, G. Alici, H. Zhou, W. Li, A review of localization systems for robotic endoscopic capsules, IEEE Trans. Biomed. Eng. 59 (9) (2012) 2387–2399.

[5] M. Sitti, H. Ceylan, W. Hu, J. Giltinan, M. Turan, S. Yim, E. Diller, Biomedical applications of untethered mobile milli/microrobots, Proc. IEEE 103 (2) (2015) 205–224.

[6] M. Turan, Y. Almalioglu, H. Araujo, E. Konukoglu, M. Sitti, A non-rigid map fusion-based RGB-depth slam method for endoscopic capsule robots, arXiv preprint, arXiv:1705.05444 (2017).

[7] M. Turan, Y. Almalioglu, E. Konukoglu, M. Sitti, A deep learning based 6 degree-of-freedom localization method for endoscopic capsule robots, arXiv preprint, arXiv:1705.05435 (2017).

[8] M. Turan, A. Abdullah, R. Jamiruddin, H. Araujo, E. Konukoglu, M. Sitti, Six degree-of-freedom localization of endoscopic capsule robots using recurrent neural networks embedded into a convolutional neural network, arXiv preprint, arXiv:1705.06196 (2017).

[9] M. Turan, Y.Y. Pilavci, R. Jamiruddin, H. Araujo, E. Konukoglu, M. Sitti, A fully dense and globally consistent 3d map reconstruction approach for GI tract to enhance therapeutic relevance of the endoscopic capsule robot, arXiv preprint arXiv:1705.06524 (2017).

[10] M. Turan, Y.Y. Pilavci, I. Ganiyusufoglu, H. Araujo, E. Konukoglu, M. Sitti, Sparse-then-dense alignment based 3d map reconstruction method for endoscopic capsule robots, arXiv preprint, arXiv:1708.09740 (2017).

[11] M.K. Goenka, S. Majumder, U. Goenka, Capsule endoscopy: present status and future expectation, World J. Gastroenterol. 20 (29) (2014) 10024–10037.

[12] F. Munoz, G. Alici, W. Li, A review of drug delivery systems for capsule endoscopy, Adv. Drug Delivery Rev. 71 (2014) 77–85.

[13] F. Carpi, N. Kastelein, M. Talcott, C. Pappone, Magnetically controllable gastrointestinal steering of video capsules, IEEE Trans. Biomed. Eng. 58 (2) (2011) 231–234.

[14] H. Keller, A. Juloski, H. Kawano, M. Bechtold, A. Kimura, H. Takizawa, R. Kuth, Method for navigation and control of a magnetically guided capsule endoscope in the human stomach, in: 2012 4th IEEE RAS & EMBS International Conference on Biomedical Robotics and Biomechatronics (BioRob), IEEE, 2012, pp. 859–865.

[15] A.W. Mahoney, S.E. Wright, J.J. Abbott, Managing the attractive magnetic force between an untethered magnetically actuated tool and a rotating permanent magnet, in: 2013 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2013, pp. 5366–5371.

[16] S. Yim, E. Gultepe, D.H. Gracias, M. Sitti, Biopsy using a magnetic capsule endoscope carrying, releasing, and retrieving untethered microgrippers, IEEE Trans. Biomed. Eng. 61 (2) (2014) 513–521.

[17] A.J. Petruska, J.J. Abbott, An omnidirectional electromagnet for remote manipulation, in: 2013 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2013, pp. 822–827.

[18] D. Son, S. Yim, M. Sitti, A 5-d localization method for a magnetically manipulated untethered robot using a 2-d array of hall-effect sensors, IEEE/ASME Trans. Mechatron. 21 (2) (2016) 708–716.

[19] D. Son, M.D. Dogan, M. Sitti, Magnetically actuated soft capsule endoscope for fine-needle aspiration biopsy, in: 2017 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2017, pp. 1132–1139.

[20] M. Fluckiger, B.J. Nelson, Ultrasound emitter localization in heterogeneous media, in: 2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, IEEE, 2007, pp. 2867–2870.

[21] J.M. Rubin, H. Xie, K. Kim, W.F. Weitzel, S.Y. Emelianov, S.R. Aglyamov, T.W. Wakefield, A.G. Urquhart, M. ODonnell, Sonographic elasticity imaging of acute and chronic deep venous thrombosis in humans, J. Ultrasound Med. 25 (9) (2006) 1179–1186.

[22] K. Kim, L.A. Johnson, C. Jia, J.C. Joyce, S. Rangwalla, P.D. Higgins, J.M. Rubin, Noninvasive ultrasound elasticity imaging (UEI) of crohn's disease: animal model, Ultrasound Med. Biol. 34 (6) (2008) 902–912.

[23] S. Yim, M. Sitti, 3-d localization method for a magnetically actuated soft capsule endoscope and its applications, IEEE Trans. Robot. 29 (5) (2013) 1139–1151.

[24] T. Ping-Sing, M. Shah, Shape from shading using linear approximation, Image Vision Comput. 12 (8) (1998) 487–498.

[25] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9.

[26] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, CoRR (2014). 1409.1556.

[27] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[28] N. Sünderhauf, S. Shirazi, F. Dayoub, B. Upcroft, M. Milford, On the performance of convnet features for place recognition, in: 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 2015, pp. 4297–4304.

[29] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., Imagenet large scale visual recognition challenge, Int. J. Comput. Vision 115 (3) (2015) 211–252.

[30] A. Kendall, M. Grimes, R. Cipolla, Posenet: a convolutional network for real-time 6-dof camera relocalization, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 2938–2946.

[31] F.A. Gers, J. Schmidhuber, F. Cummins, Learning to Forget: Continual Prediction with LSTM, 1999.

[32] F. Walch, C. Hazirbas, L. Leal-Taixé, T. Sattler, S. Hilsenbeck, D. Cremers, Image-based localization with spatial LSTMs, arXiv preprint, arXiv:1611.07890 (2016).

[33] S. Wang, R. Clark, H. Wen, N. Trigoni, Deepvo: towards end-to-end visual odometry with deep recurrent convolutional neural networks, in: 2017 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2017, pp. 2043–2050.

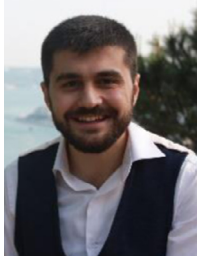
[34] D. Kingma, J. Ba, Adam: a method for stochastic optimization, arXiv preprint, arXiv:1412.6980(2014).

[35] N. Srivastava, G.E. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, J. Mach. Learn. Res. 15 (1) (2014) 1929–1958.

[36] J. Engel, T. Schöps, D. Cremers, Lsd-slam: large-scale direct monocular slam, in: European Conference on Computer Vision, Springer, 2014, pp. 834–849.

[37] R. Mur-Artal, J. Montiel, J.D. Tardós, Orb-slam: a versatile and accurate monocular slam system, IEEE Trans. Robot. 31 (5) (2015) 1147–1163.

Mehmet Turan received his diploma degree from the Information technology and Electronics Engineering Department of RWTH Aachen, Germany in 2012. He was a research scientist at UCLA (University of California Los Angeles) between 2013–2014 and a research scientist at the Max Planck Institute for Intelligent Systems between 2014–present. He is currently enrolled as a Ph.D. student at the ETH Zurich, Switzerland. He is also affiliated with Max Planck-ETH Center for Learning Systems, the first joint research center of ETH Zurich and the Max Planck Society. His research interests include SLAM (simultaneous localization and mapping) techniques for milli-scale medical robots and deep learning techniques for medical robot localization and mapping. He received DAAD fellowship between years 2005–2011 and Max Planck Fellowship between 2014–present. He has also received MPI-ETH Center fellowship between 2016–present.



Yasin Almalioglu received the B.Sc. degree with honours in computer engineering from Bogazici University, Istanbul, Turkey in 2015. He was a research intern at CERN Geneva, Switzerland and Astroparticle and Neutrino Physics Group at ETH Zurich, Switzerland in 2013 and 2014, respectively. He is currently pursuing the M.Sc. degree in computer engineering at Bogazici University, Istanbul, Turkey. His research interests include machine learning, Bayesian statistics, Monte Carlo methods, probabilistic graphical models, artificial neural networks and mobile robot localization. He received the Engin Arik Fellowship in 2013.

Helder Araujo is a professor at the Department of Electrical and Computer Engineering of the University of Coimbra. His research interests include computer vision applied to robotics, robot navigation and visual servoing. In the last few years he has been working on non-central camera models, including aspects related to pose estimation, and their applications. He has also developed work in active vision, and on control of active vision systems. Recently he has started work on the development of vision systems applied to medical endoscopy.

Ender Konukoglu, Ph.D., finished his Ph.D. at INRIA Sophia Antipolis in 2009. From 2009 till 2012 he was a post-doctoral researcher at Microsoft Research Cambridge. From 2012 till 2016 he was a junior faculty at the Athinoula A. Martinos Center affiliated to Massachusetts General Hospital and Harvard Medical School. Since 2016 he is an assistant professor of Biomedical Image Computing at ETH Zurich. His is interested in developing computational tools and mathematical methods for analysing medical images with the aim to build decision support systems. He develops algorithms that can automatically extract quantitative image-based measurements, statistical methods that can perform population comparisons and biophysical models that can describe physiology and pathology.

Dr. Metin Sitti received the B.Sc. and M.Sc. degrees in electrical and electronics engineering from Bogazici University, Istanbul, Turkey, in 1992 and 1994, respectively, and the Ph.D. degree in electrical engineering from the University of Tokyo, Tokyo, Japan, in 1999. He was a research scientist at UC Berkeley during 1999–2002. He has been a professor in the Department of Mechanical Engineering and Robotics Institute at Carnegie Mellon University, Pittsburgh, USA since 2002. He is currently a director at the Max Planck Institute for Intelligent Systems in Stuttgart. His research interests include small-scale physical intelligence, mobile microrobotics, bio-inspired materials and miniature robots, soft robotics, and micro-/nanomanipulation. He is an IEEE fellow. He received the SPIE Nanoengineering Pioneer Award in 2011 and NSF CAREER Award in 2005. He received many best paper, video and poster awards in major robotics and adhesion conferences. He is the editor-in-chief of the Journal of Micro-Bio Robotics.