Abstract

We present a robust deep learning based 6 degrees-of-freedom (DoF) localization system for endoscopic capsule robots. Our system mainly focuses on localization of endoscopic capsule robots inside the GI tract using only visual information captured by a mono camera integrated to the robot. The proposed system is a 23-layer deep convolutional neural network (CNN) that is capable to estimate the pose of the robot in real time using a standard CPU. The dataset for the evaluation of the system was recorded inside a surgical human stomach model with realistic surface texture, softness, and surface liquid properties so that the pre-trained CNN architecture can be transferred confidently into a real endoscopic scenario. An average error of 7.1% and 3.4% for translation and rotation has been obtained, respectively. The results accomplished from the experiments demonstrate that a CNN pre-trained with raw 2D endoscopic images performs accurately inside the GI tract and is robust to various challenges posed by reflection distortions, lens imperfections, vignetting, noise, motion blur, low resolution, and lack of unique landmarks to track.

Keywords: Capsule endoscope robot, deep learning, CNN, localization,
1. Introduction

In the past decade, the advances in material science enabled the fabrication of small, low cost devices in a variety of popular applications. Following these advances, untethered pill-size, swallowable capsule endoscopes with an on-board 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, 2, 3, 4].

However, current capsule endoscopes used in hospitals are passive devices controlled by peristaltic motions of the inner organs. The control over the 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. 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 [5, 2, 6, 7, 8, 9, 10, 11]. An active motion control is, on the other hand, heavily dependent on a precise and reliable real time pose estimation capability which makes the robot localization the key capability for a successful endoscopic capsule robot operation. In the last decade, several localization methods [4, 12, 13, 14, 15] were proposed to calculate the 3D position and orientation of the endoscopic capsule robot such as fluoroscopy [4], ultrasonic imaging [12, 13, 14, 15], positron emission tomography (PET) [4, 15], magnetic resonance imaging (MRI) [4], radio transmitter based techniques and magnetic field based techniques. The common drawback of these localization methods is that they require extra sensors and hardware design. Such extra sensors have
their own drawbacks 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 the sensors with the activation system of the device.

As a solution of these issues, a trend of vision-based localization methods have attracted the attention for the localization of such small scale medical devices. As a first attempt, structure from motion (SfM) methods have been proposed to deal with monocular endoscope localization [16, 5]. However, SfM methods are incapable of real time processing which makes them unsuitable for robotic localization. The use of visual simultaneous localization and mapping (VSLAM) in medical field has been researched by [17, 2], who modified the extended Kalman filter SLAM (EKF-SLAM) framework from [6] for medical applications. [18, 1] modified the breakthrough VSLAM method parallel tracking and mapping (PTAM) algorithm to a stereo-endoscope to build a denser 3D map than previous EKF based SLAM systems. Due to the non-rigid deformations inside human body, the use of only a monocular endoscope has remained as a challenge. [19, 9] provided extensive validation on in-vivo human sequences, e.g. demonstrating the efficiency of EKF-SLAM for hernia defect measurements in hernia repair surgery. Following the trend of PTAM modification, another milestone VSLAM method, ORB-SLAM was adjusted into medical field and proposed as ORB SLAM-based Endoscope Tracking and 3D Reconstruction method [20] for endoscopic localization.

All of the presented vision-based localization methods use landmark-based methods, where the underlying idea is to search for distinct landmark positions and track them across frames for localization. These methods show poor performance on endoscopic images mainly because such images lack distinct features that algorithms need to extract and track to perform a precise visual odometry. Another issue in front of mono camera based VSLAM algorithms is that the translation in z direction is mathematically not possible to determine up to a scale using only one camera. Moreover, all of the existing methods in literature are proposed for handheld standard endoscopic cameras. However, endoscopic
capsule robot differs from standard handheld endoscope due to its own characteristic problems such as space availability, limited energy source, low camera quality and resolution. Convolutional neural networks (CNN), such as deep neural networks, convolutional deep neural networks, deep belief networks and recurrent neural networks have been applied to fields such as computer vision, automatic speech recognition, natural language processing and bioinformatics where they have produced state-of-the-art results on various tasks and outperformed already existing methods in these fields. There has been especially a wide-spread adoption of various deep-neural network architectures for computer vision related tasks because of the apparent empirical success these architectures have shown in various image classification tasks. In fact, the seminal paper ImageNet Classification with Deep Convolutional Neural Networks has been cited over 8000 times [21]. With this intention, we propose in this paper a new method for 6-DoF localization of both endoscopic capsule robot and handheld standard endoscope using the CNN system. We demonstrate that CNN learns the most relevant feature vector representation related to 3D position and orientation estimation of the robot from 2D raw image inputs. The method we introduced solves several issues confronted with the challenge endoscopic image datasets faced by typical SLAM pipelines, such as the need to establish frame-to-frame feature correspondence which is practically not working on low textured inner organ tissues and the need to extract large baseline key frames. The main benefit of our deep learning based approach is that it does not require any additional sensor. And compared to vision based unsupervised localization methods, our supervised method making use of pre-trained information is more appropriate for endoscopic type of images since the operation environment remains more or less similar among different patients. For the first time in literature, we propose a deep learning based real time 6-DoF localization method for mono camera based endoscopic capsule robot and handheld standard endoscope. As the outline of this paper, section 2 introduces the proposed CNN based localization method in detail. Section 3 presents our dataset and the experimental setup. Section 4 shows our experimental results for 6-DoF localization of the
This paper addresses the vision-based localization problem of the endoscopic capsule robots. Localization capability is a significant functionality for mobile robots. We introduce a novel localization method using a CNN architecture that is trained end-to-end to regress the robot’s orientation and position in real time (5 ms per frame). The proposed system takes a single endoscopic RGB image and regresses the camera’s 6-DoF pose without a need of any extra sensor. A loss function is minimized using the back-propagation algorithm to optimize the parameters of the architecture.

Inspired from GoogLeNet [22], the proposed architecture consists of a stack of convolutional layers, max pooling layer, normalization layer, inception modules, pooling layer, fully-connected layer and affine regressor layer, a total of 23 layers. A general block diagram of the modified network architecture can be seen in Figure 1. We removed softmax layer from the GoogLeNet and integrated a fully-connected (FC) layer and an affine regressor layer to regress 6-DoF pose values. The hyperparameters of convolution layer such as kernel size, stride and number of filters are optimized via the training to get the most relevant information from the raw image for 6-DoF localization.

Rectified-linear
unit (RELU) layer, an activation unit of lower computational complexity, is robust to vanishing gradients during the training to ensure that the corresponding neural unit remains active and continues updating the learning weights. Pooling layer reduces spatial dimensions in order to lower the overfitting risk and gain computational performance. Fully connected layer connects each neuron to all activation units in the previous layer. Introduced by GoogLeNet, inception modules are the most important layers of the proposed CNN architecture increasing the learning capacity of the network. A block diagram for inception module is shown in Figure 2. Its task is the extraction of the large scale information using larger kernels while keeping to gather detail information using smaller kernels. The module basically acts as multiple convolution filter inputs, that are processed on the same input. It also does pooling at the same time. All the results are then concatenated. This allows the model to take advantage of multi-level feature extraction from each input.

For the implementation of the proposed architecture, Caffe library [23] was used which provides a modifiable framework, enables multi-GPU training and accelerates the computational procedure by CUDA supported GPU computation. The learning rate was initialized to 0.001 reducing as the epochs of the
training continues. Adaptive Moment Estimation (Adam) [24] optimization method was used to optimize the goal function. We run our algorithm on an Amazon EC2 p2.xlarge GPU compute instance. The list of the parameters is as follow:

- learning rate: 0.001
- momentum1: 0.9
- momentum2: 0.999
- epsilon: $10^{-8}$
- solver type: Adam
- batch size: 64
- GPU: NVIDIA K80

2.2. CNN-based 6-DoF Localization

Our proposed CNN system learns translational and rotational movements simultaneously. To regress pose, we trained the CNN architecture on Euclidean
loss using Adam with the following objective loss function

\[ \text{loss}(I) = \| \hat{x} - x \|_2 + \beta \| \hat{q} - q \|_2 \]  

(1)

where \( x \) is the translation vector and \( q \) is the rotation vector. The pseudocode to calculate the loss value is given in Algorithm 1. In our loss function, a balance \( \beta \) must be kept between the orientation and translation loss values which are highly coupled as they are learned from the same model weights \([25]\).

Experimental results showed that the optimal \( \beta \) was given by the ratio between expected error of position and orientation at the end of the training session.

Algorithm 1 Pseudo code to calculate the loss over the network

1: \textbf{procedure} \textsc{CalculateLoss}
2: \hspace{1em} loss \leftarrow 0
3: \hspace{1em} \textbf{for} layer in layers \textbf{do}
4: \hspace{2em} \textbf{for} top, loss\_weight in layer\_tops, layer\_loss\_weights \textbf{do}
5: \hspace{3em} loss \leftarrow loss + loss\_weight \times \text{sum(top)}

The back-propagation algorithm is used to calculate the gradients of CNN weights. These gradients are passed into the Adam optimization method which computes 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 keeps exponentially decaying average of past gradients, \( m_t \) that is similar to momentum. The update equations are given as

\[
(m_t)_i = \beta_1 (m_{t-1})_i + (1 - \beta_1)(\nabla L(W_t))_i \\
(v_t)_i = \beta_2 (v_{t-1})_i + (1 - \beta_2)(\nabla L(W_t))^2_i \\
(W_{t+1})_i = (W_t)_i - \alpha \frac{\sqrt{1 - (\beta_2)_i^t} - (m_t)_i}{\sqrt{(v_t)_i + \varepsilon}} (m_t)_i \\
\]

(2)  
(3)

(4)

We used default values proposed by \([24]\) for the parameters \( \beta_1, \beta_2 \) and \( \varepsilon \): \( \beta_1 = 0.9, \beta_2 = 0.999 \) and \( \varepsilon = 10^{-8} \).
2.3. Transfer Learning

We sidestep the common problem of CNN that it requires large amount of training data by using transfer learning for the first layers of the proposed architecture. For that aim, we started the pose training using the weights acquired by ImageNet [26]. According to the experimental results, we were able to conclude that a network trained to output pose-invariant classification labels is suitable as a starting point for a pose regressor in the context of endoscopic capsule robot localization. A possible explanation of that observation is that the network must have activation units that are invariant to different pose values.

During transfer learning on the first layers, we kept the learning rate low in order not to distort the already learned weights. During the training of the next layers, the learning rate of individual models were increased to enable a learning procedure from scratch. We aim to publish the weights of the trained model for the sake of other research groups working on endoscopic capsule robot localization.
3. DATASET, EQUIPMENT AND SPECIFICATIONS

In deep learning based applications, the existence of a large amounts of training data is very important. Since there is no publicly available dataset existing for 6-DoF endoscopic capsule robot localization to our knowledge, we created our own dataset and labels for that purpose using different endoscopic cameras mounted on our magnetically activated soft capsule endoscope (MASCE) system as seen in Figure 5.

For the labelling of the endoscopic videos, OptiTrack motion tracking system consisting of four infrared cameras and a tracking software was employed (see Figure 6). To make sure that our dataset was not narrowed down to just one specific endoscopic camera, three different endoscopic cameras were used to capture the endoscopic videos. The cameras used for recording the dataset were the AWAIBA Naneye, VOYAGER and POTENSIC endoscopic camera. Specifications of the cameras can be found in Table 1, Table 2 and Table 3 respectively.

The videos were recorded on a non-rigid, realistic surgical stomach model Ko-ken LM103 - EDG (EsophagoGastroDuodenoscopy) Simulator. We used paraffin oil to increase the reflection inside the simulator since paraffin oil has the same viscosity as the stomach fluid. For visual reference see Figure 3. A total of 15 minutes of stomach imagery was recorded for this research containing over...
Table 1: AWAIBA NANEYE MONOCULAR CAMERA SPECIFICATIONS

| Specification       | Value                  |
|---------------------|------------------------|
| Resolution          | 250 x 250 pixels       |
| Footprint           | 2.2 x 1.0 x 1.7 mm     |
| Pixel size          | 3 x 3 μm²              |
| Pixel depth         | 10 bit                 |
| Frame rate          | 44 fps                 |

10000 frames. Some sample frames of the dataset are shown in Figure 4.

To avoid overfitting of the proposed CNN architecture, we extended the dataset applying different image distortion techniques on to the raw images such as Gaussian blur, median blur, brightness distortion.

4. EXPERIMENTS AND RESULTS

Cross validation method was used to assess the performance of the proposed CNN based localization method. The dataset was divided into two parts of which the first part was used for training purposes and the second part for
Table 2: POTENSIC MINI MONOCULAR CAMERA SPECIFICATIONS

| Specification       | Details               |
|---------------------|-----------------------|
| Resolution          | 1280 x 720 pixels     |
| Footprint           | 5.2 x 4.0 x 2.7 mm    |
| Pixel size          | 10 x 10 μm²           |
| Pixel depth         | 10 bit                |
| Frame rate          | 15 fps                |

Table 3: VOYAGER MINI CAMERA SPECIFICATIONS

| Specification       | Details               |
|---------------------|-----------------------|
| Resolution          | 720 x 480 pixels      |
| Footprint           | 5.2 x 5.0 x 2.7 mm    |
| Pixel size          | 10 x 10 μm²           |
| Pixel depth         | 10 bit                |
| Frame rate          | 15 fps                |
validation purposes. For a quantitative analysis, we compared the estimated 6-DoF pose values with the ground truth data. The mean and standard deviation of the translational and rotational errors in x, y and z axis are shown in Figure 9.

We conducted two separate experiments for the training of the CNN network. The first experiment was conducted with 10000 frames of which approximately 7000 frames were used for training and 3000 frames for validation. In the second experiment, 70000 frames were used partially consisting of the raw images and partially consisting of the distorted data we generated from the original dataset. We run both experiments on an Amazon EC2 p2.xlarge machine instance. The duration of the first training session was 4 hours and 37 minutes and the duration of the second training session was 19 hours and 17 minutes. After the training sessions, we tested the proposed CNN based localization method on a test dataset that was never seen during the training session by the CNN. The 6-DoF pose estimation per image takes 5 ms on a standard CPU Intel i5 core desktop PC.

Figure 7 shows the training error (blue) and validation error (red) of translational and rotational movements both as a function of the number of training cycles. Both the validation and training error decrease and converge to a global
Figure 8: Estimated and ground truth 3D trajectories for two different cases.
minimum, where the learning cycle was stopped to avoid any overfitting and underfitting of the dataset.

Better results were obtained in the second experiment: The best rotation accuracy is achieved with respect to x axis with an error of 1.60% whereas the second rotational accuracy is achieved with respect to y axis with an error of 3.01%. The rotational error around the z axis is 5.71%. The translation errors are 4.72%, 9.16% and 7.44% in x, y and z axis, respectively. Figure 8 represents for two different challenge cases consisting of fast rotations and large translations the 3D trajectory plots acquired by the OptiTrack system (blue) and the 3D trajectory estimated by our method (red). As observed in both cases, the proposed CNN system is able to track the endoscopic capsule robot motion very accurately.

Figure 10 compares our proposed system with the existing popular SLAM methods including ORB SLAM, PTAM and LSD SLAM. To compare the algorithms, we calculated the root-mean-square error (RMSE) of the estimated results with respect to the OptiTrack ground truth. As shown in the Figure 10, PTAM shows the worst performance out of all methods with an RMSE of 2.6 cm while our proposed system outperforms all of them with an RMSE of 0.18 cm for a trajectory of 18 cm.
5. CONCLUSION

We present, to our knowledge, the first application of deep CNN to end-to-end 6-DoF localization for endoscopic capsule robots and handheld standard endoscopes. We have demonstrated that one can overcome the need for huge amount of training data for CNN with the help of transfer learning. We also concluded that the pre-trained weights on ImageNet dataset indeed reveals information about pose values and they are useful for stomach environment as well even though it is not directly related to the endoscopic environment and robot localization. Our method tolerates many issues faced by modern SLAM techniques such as feature correspondence establishment across frames in low textured areas, high reflections, motion blur and low image quality. Since huge amount of training dataset is not easy to create for endoscopic applications, we consider to publish our trained weights for 6-DoF endoscopic capsule robot localization online for the sake of other groups working in that area.

As future work, we aim to use our trained deep learning architecture as a deep feature extractor and serve such deep features as an input to modern SLAM
methods which will eliminate the need for long training sessions. Additionally, we consider to extend the proposed deep learning based localization method to stereo endoscopic capsule robot case.

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 endoscopy 71 (2) (2010) 280–286.

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

[3] G. Pan, L. Wang, Swallowable wireless capsule endoscopy: progress and technical challenges, Gastroenterology research and practice 2012.

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

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

[6] F. Munoz, G. Alici, W. Li, A review of drug delivery systems for capsule endoscopy, Advanced drug delivery reviews 71 (2014) 77–85.

[7] F. Carpi, N. Kastelein, M. Talcott, C. Pappone, Magnetically controllable gastrointestinal steering of video capsules, IEEE Transactions on Biomedical Engineering 58 (2) (2011) 231–234.

[8] 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.

[9] 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: Robotics and Automation (ICRA), 2013 IEEE International Conference on, IEEE, 2013, pp. 5366–5371.

[10] S. Yim, E. Gultepe, D. H. Gracias, M. Sitti, Biopsy using a magnetic capsule endoscope carrying, releasing, and retrieving untethered microgrippers, IEEE Transactions on Biomedical Engineering 61 (2) (2014) 513–521.

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

[12] 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.

[13] J. M. Rubin, H. Xie, K. Kim, W. F. Weitzel, S. Y. Emelianov, S. R. Aglyamov, T. W. Wakefield, A. G. Urquhart, M. O’Donnell, Sonographic elasticity imaging of acute and chronic deep venous thrombosis in humans, Journal of Ultrasound in Medicine 25 (9) (2006) 1179–1186.

[14] 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 in medicine & biology 34 (6) (2008) 902–912.

[15] S. Yim, M. Sitti, 3-d localization method for a magnetically actuated soft capsule endoscope and its applications, IEEE Transactions on Robotics 29 (5) (2013) 1139–1151.
[16] M. Sitti, H. Ceylan, W. Hu, J. Giltinan, M. Turan, S. Yim, E. Diller, Biomedical applications of untethered mobile milli/microrobots, Proceedings of the IEEE 103 (2) (2015) 205–224.

[17] P. Mountney, D. Stoyanov, A. Davison, G.-Z. Yang, Simultaneous stereoscope localization and soft-tissue mapping for minimal invasive surgery, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2006, pp. 347–354.

[18] B. Lin, A. Johnson, X. Qian, J. Sanchez, Y. Sun, Simultaneous tracking, 3d reconstruction and deforming point detection for stereoscope guided surgery, in: Augmented Reality Environments for Medical Imaging and Computer-Assisted Interventions, Springer, 2013, pp. 35–44.

[19] O. G. Grasa, J. Civera, A. Guemes, V. Munoz, J. Montiel, EKF monocular slam 3D modeling, measuring and augmented reality from endoscope image sequences, in: Medical image computing and computer-assisted intervention (MICCAI), Vol. 2, 2009.

[20] N. Mahmoud, I. Cirauqui, A. Hostettler, C. Doignon, L. Soler, J. Marescaux, J. Montiel, Orbslam-based endoscope tracking and 3D reconstruction, arXiv preprint arXiv:1608.08149.

[21] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in neural information processing systems, 2012, pp. 1097–1105.

[22] 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.

[23] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, T. Darrell, Caffe: Convolutional architecture for fast fea-
ture embedding, in: Proceedings of the 22nd ACM international conference on Multimedia, ACM, 2014, pp. 675–678.

[24] D. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980.

[25] 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.

[26] Z. Wu, Y. Zhang, F. Yu, J. Xiao, A gpu implementation of googlenet (2014).