An End-to-End Unsupervised Deep Network for Monocular Depth and Ego-Motion Estimation

Madhu Vankadari
Research Proposal

1 Motivation:

How do humans navigate? Are we trained with ground truth data while learning to perceive the geometry of the scene? Do we use some other external sensors to navigate at night time? How are we able to perceive the scene geometry even if it is dark, bright, sunny, foggy, rainy, etc.? How are we able to maintain long-term temporal dependencies? If a human brain can do all these with just stereo images, can it be replicated? If yes, what has the computer vision community done so far? Where are we, what more needs to be done, and how do we get there? In this proposal, I will try to address some of these questions with my own study of the literature and my own experience of trying to solve some of the related problems.

2 Introduction:

An understanding of the surroundings is required for a robot to successfully navigate in an environment. Currently, several sensors are used by robots to understand its own environment -ultrasonic proximity sensors, IR sensors, LIDARS, cameras (Stereo/Monocular), odometric sensors (wheel encoders, gyroscopes, accelerometers), etc. Among these, cameras are extremely cheap and provide much more information about the environment compared to other sensors. However, unlike humans, it is not an easy task for a robot to extract meaningful information from images. Nevertheless, things are changing fast, thanks to the rapid advancement that is being made in the field of computer vision, machine learning, and AI. In this proposal, I will restrict my discussion mostly to the use of cameras for autonomous navigation, which requires solving the SLAM problem. While images are better suited for identifying visual landmarks, the lack of depth information poses problems in understanding the geometry of the scene which is essential for creating reliable maps. Researchers leverage camera motion to compensate for the lack of direct depth information in monocular images. On the other hand, range sensors like Lasers and LIDARs provide depth information but lack visual features making it difficult to detect landmarks reliably. A combination of these two kinds of sensors usually yields the best results in SLAM, but at a higher cost which is not easily replicable in low-cost robotic applications.

3 Literature Survey:

As we can see, inferring depth and pose information from images lies at the bottom of truly solving the Visual SLAM problem where the robots can build reliable maps only from images. While a number of works have been reported that aim to solve the SLAM problem using only monocular images [4][15][16], the outcome is still far from what is desired. The estimation of depth from images can be solved to some extent by making use of multiple views (Multi-view / stereo geometry), again with an additional computational cost [17][5]. Even though these results are not as good as those obtained from lasers or LiDARs, they are getting closer with time.

The recent success of deep learning networks in solving several computer vision problems has created renewed interest in this field. Most of the works on depth estimation using deep learning can be divided into two categories - supervised and unsupervised algorithms. Supervised learning methods learn a CNN based regressor to directly learn depth from monocular images [3][13][12]. The predicted depth is compared against the ground truth depth to train the network. A similar kind of approach is used for ego-motion estimation [11][21] either by using fully connected layers or LSTMs. Although these methods enjoyed great success in the early days, they need ground truth information for all the training data which is practically impossible to collect and maintain. This motivated the computer vision community to look into other possible solutions such as making use of image geometry as a supervision to learn for depth and ego-motion. For instance, [6] proposed an approach that learns for disparity (inverse of depth) using the stereo geometry. In this method, the network predicts the disparity which is the corresponding pixel difference of a stereo pair of images. Using that, they reconstruct stereo images from one another. The photometric loss between the original and reconstructed is used to train the network. However, this method is very far from the supervised methods in terms of estimation accuracy. Later on, [8] [23] introduced a new objective and an image comparison metric to further improve results. On the other hand, there are some monocular methods [24],[22] that address this issue by using temporal image geometry inspired by direct-methods (traditional methods) to estimate depth and ego-motion.

4 My research and outcomes:

In this direction, we proposed a new deep learning architecture called UnDEMoN that tried to use both spatial and temporal geometry to further improve the efficacy of unsupervised depth and ego-motion estimation. This work was accepted for publication at IROS 2018 [1]. We also demonstrated that image based depth and pose estimation can be used for improving the accuracy of place recognition algorithms in visual SLAM applications in our paper published in ICRA 2019 [7]. We further improved the results for depth and pose estimation by using patchGAN paradigm where a discriminator is used to differentiate between the reconstructed images from the real ones. This work was published in IJCAI 2019 [19]. In addition, we proposed a method to enforce cycle consistency with adversarial learning during depth estimation which is shown to further improve the performance of UnDEMoN. This is currently under review at ICRA 2020.

5 Problems to be solved:

At present, I am investigating the effect of extreme photometric changes in the environment on the depth and pose estimation accuracy. It is observed that a model trained using day-time images fails miserably at night time due to a large domain shift between day and night. Moreover, the intensity gradient between corresponding pixels of temporally aligned images makes it impossible to use common photo-metric losses to train the network for night-time images. There has not been much work progressed in this line of research, though it is very crucial in autonomous driving applications. I believe that there is only style change between the day and night images and the context is the same (buildings, roads, cars, and vegetation, etc). In other words, if we can learn generic contextual embeddings that are robust to styles such as night, fog, rain, etc, then a single encoder-decoder network will be able to address all kinds of variations in the data [9]. As of now, I have tried two approaches to achieve the aforesaid,

1. Adversarial Discriminative Feature Adaptation (ADFA): In this approach, we used our pre-trained depth estimation model [19] as a reference for day-time features and trained another encoder from scratch to generate features which look like day-time from completely unpaired day-night image samples. The authenticity of the generated features is evaluated by a patchGAN discriminator. The results are shown to perform as comparable as an image translation method developed using CycleGAN. This work is under review at CVPR'20. However, this method fails to address the very low illumination areas and saturated image regions. I am further trying to solve these issues using Visual Place Recognition and Image translation methods.

2. The previous method needs to learn an individual encoder for every environmental condition. Currently, I have relaxed the assumption of unpaired data by generating paired samples with the help of GPS data. Using these paired samples, a more robust generic representation of the day-night pair can be learned with the aid of GAN based metric learning. This framework gives another flexibility of learning only a single encoder for both day and night images for depth and ego-motion estimation. This work is under progress and we are planning to submit to ECCV'20.

All of my methods do not talk about moving objects in the scene. These will never get reconstructed properly because of their relative motion to the camera. The improperly reconstructed pixels will result in wrong gradient flow while training if they are included in the loss function. There are few methods [22] [24] in literature that addresses this problem by adding another network named Flow-Net to estimate optical flow between two images or by predicting a mask to remove the moving pixels while calculating the loss. However, these methods are computationally expensive as they have to learn separate networks. Instead, we can solve this problem by borrowing the Spatio-temporal attention module [20][14][10] from NLP literature. This module gives the advantage of paying attention only to useful features and discarding the rest while training the model. Moreover, they are computationally far simpler than Flow-Net or explainability masks used so far in the literature.

In addition, my methods so far consider only a snippet of consecutive images instead of a full sequence of images due to which the predicted pose trajectories drift from the ground-truth. This is because the predicted poses can not maintain long term temporal dependencies. There are few methods in the current literature that solves the drift error obtained from the loop closure using traditional pose-graph optimization or bundle adjustment methods. There has been a lot of research in NLP to maintain long term dependencies between the subject and the total paragraph. One can borrow ideas from [2], [20], etc to make a complete end-to-end deep learning pipeline for depth and ego-motion estimation with long-term dependencies for a given sequence.

6 Conclusion:

In short, I would like to create a single monocular depth and ego-motion estimation network which can work for any time images (night, day, foggy and rainy, etc) during my Ph.D. This model should be able to address long-term dependencies, implicit place recognition, moving objects and occlusions as well.

References

[1] V Madhu Babu et al. "UnDEMoN: Unsupervised Deep Network for Depth and Ego-Motion Estimation". In: *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2018, pp. 1082–1088.

[2] Zihang Dai et al. "Transformer-xl: Attentive language models beyond a fixed-length context". In: *arXiv preprint arXiv:1901.02860* (2019).

[3] David Eigen, Christian Puhrsch, and Rob Fergus. "Depth map prediction from a single image using a multi-scale deep network". In: *Advances in neural information processing systems*. 2014, pp. 2366–2374.

[4] Jakob Engel, Thomas Schöps, and Daniel Cremers. "LSD-SLAM: Large-scale direct monocular SLAM". In: *European conference on computer vision*. Springer. 2014, pp. 834–849.

[5] Jakob Engel, Jörg Stückler, and Daniel Cremers. "Large-scale direct SLAM with stereo cameras". In: *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2015, pp. 1935–1942.

[6] Ravi Garg et al. "Unsupervised cnn for single view depth estimation: Geometry to the rescue". In: *European Conference on Computer Vision*. Springer. 2016, pp. 740–756.

[7] Sourav Garg, Niko Suenderhauf, and Michael Milford. "Semantic–geometric visual place recognition: a new perspective for reconciling opposing views". In: *The International Journal of Robotics Research* (2019), p. 0278364919839761.

[8] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. "Unsupervised monocular depth estimation with left-right consistency". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE. 2017, pp. 6602–6611.

[9] Xun Huang et al. "Multimodal unsupervised image-to-image translation". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 172–189.

[10] Saumya Jetley et al. "Learn to pay attention". In: *arXiv preprint arXiv:1804.02391* (2018).

[11] Alex Kendall, Matthew Grimes, and Roberto Cipolla. "Posenet: A convolutional network for real-time 6-dof camera relocalization". In: *IEEE International Conference on Computer Vision (ICCV), 2015*. IEEE. 2015, pp. 2938–2946.

[12] Fayao Liu et al. "Learning depth from single monocular images using deep convolutional neural fields". In: *IEEE transactions on pattern analysis and machine intelligence* 38.10 (2016), pp. 2024–2039.

[13] Yue Luo et al. "Single view stereo matching". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 155–163.

[14] Lili Meng et al. "Where and When to Look? Spatio-temporal Attention for Action Recognition in Videos". In: *arXiv preprint arXiv:1810.04511* (2018).

[15] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. "ORB-SLAM: a versatile and accurate monocular SLAM system". In: *IEEE transactions on robotics* 31.5 (2015), pp. 1147–1163.

[16] Richard A Newcombe, Steven J Lovegrove, and Andrew J Davison. "DTAM: Dense tracking and mapping in real-time". In: *2011 international conference on computer vision*. IEEE. 2011, pp. 2320–2327.

[17] Taihú Pire et al. "S-PTAM: Stereo parallel tracking and mapping". In: *Robotics and Autonomous Systems* 93 (2017), pp. 27–42.

[18] Eric Tzeng et al. "Adversarial discriminative domain adaptation". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 7167–7176.

[19] Madhu Vankadari et al. "Unsupervised Learning of Monocular Depth and Ego-Motion using Conditional PatchGANs". In: *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*. International Joint Conferences on Artificial Intelligence Organization, July 2019, pp. 5677–5684. DOI: 10.24963/ijcai.2019/787. URL: https://doi.org/10.24963/ijcai.2019/787.

[20] Ashish Vaswani et al. "Attention is all you need". In: *Advances in neural information processing systems*. 2017, pp. 5998–6008.

[21] Sen Wang et al. "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.

[22] Zhichao Yin and Jianping Shi. "GeoNet: Unsupervised Learning of Dense Depth, Optical Flow and Camera Pose". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Vol. 2. 2018.

[23] Huangying Zhan et al. "Unsupervised Learning of Monocular Depth Estimation and Visual Odometry with Deep Feature Reconstruction". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 340–349.

[24] Tinghui Zhou et al. "Unsupervised Learning of Depth and Ego-Motion from Video". In: *CVPR*. 2017.