Unsupervised Learning of Visual Odometry with Depth Warp Constraints

IOP Conf. Series: Materials Science and Engineering **563** (2019) 042024 doi:10.1088/1757-899X/563/4/042024

Unsupervised Learning of Visual Odometry with Depth Warp Constraints

Haibin Shi*, Menghao Guo, Zhi Xu, Yuanbin Zou

College of Information Science and Engineering, Northeastern University, Shenyang 110819, China

E-mail: shihaibin@ise.neu.edu.cn

Abstract. Visual Odometry (VO) is one of the important components of Visual SLAM system. Some impressive work on the end-to-end deep neural networks for 6-DoF VO has appeared. We propose two-part cascade network structure to learn depth from binocular image and to infer ego-motion from consecutive frames. We propose depth warp constraints to make the Network learning more geometrically information. A lot of experiments on KITTI data set show that our model is superior to previous unsupervised methods and has comparable results with the supervised method, verifying that such a depth warp constraints perform successfully in the unsupervised deep method which is an important complement to the geometric method.

1. Introduction

With the rise of mobile robot technology, people have begun extensive research on this. Research on mobile robot technology involves the field of environment-aware technology, navigation and decision-making control science. Among them, environment-aware technology is the core of the whole mobile robot technology. Therefore, visual simultaneous localization and mapping (V-SLAM) technology has received extensive attention from researchers at home and abroad. The entire V-SLAM system can be divided into front-end and back-end. The front-end is equivalent to visual odometry, which study the relationship between frame and frame. The back-end is mainly to optimize the results of the front-end, and use the filtering theory (EKF, UKF, PF), or the optimization theory TORO, G2O to optimize the tree or graph. Finally, the optimal pose estimation is obtained.

The traditional visual odometry method [1, 2] estimates the motion of the camera by tracking the feature points between the sequence of image frames, estimates the pose of the current moment by accumulating motion between frames, and then transmits it to the back-end to reconstruct the environment. Visual odometry has a wide range of applications, can be applied to unmanned vehicle [3], drones [4, 5, 6], augmented reality [7] and so on.

Compared with the traditional inter-frame estimation method based on sparse features or dense features, the deep learning-based method does not require feature extraction, and does not require feature matching and complex geometric operations, making the method based on deep learning more intuitive and concise, resulting in far-reaching significance for mobile robot navigation and location.

In the deep learning method, Clark R et al. proposed using CNN and RNN to construct a VINet [8], which inputs image and IMU information and directly outputs the estimated pose. They use the deep learning method on the visual odometry to be very novel. But because their deep convolutional neural networks are supervised, a significant drawback of supervised deep learning methods is needed to use a large amount of manually labeled data for training. So Garg R [9] et al. proposed an unsupervised framework for depth prediction using a deep convolutional neural network without the need for prior

training and annotated ground-truth depths. Their network, under the same performance, has less than half the training time of other supervised methods. But they predict lower depth accuracy due to the use of a single image. Zhan et al. [10] differ from the popular network that gets depth from a single image, proposed a novel feature reconstruction loss to unsupervised predict single view depth and frame-to-frame odometry without scale ambiguity. But the disadvantage is that assumes no occlusion and the scene are rigid. Moreover, the depth warp constraints information of the image is not used.

In this paper, we present a novel end-to-end visual odometry architecture with depth warp constraints loss based on unsupervised deep convolutional neural networks. We trained a convolutional network end-to-end to calculate depth and ego-motion from a continuous, unlabeled pair of images. The ego-motion is estimated with image projection constraints and depth warp constraints as supervisory information.

2. Preliminaries

Our deep learning network consists of two parts:

The first part of our construction is the depth convolutional neural network (Depth ConvNet). The input is the left and right RGB images of the binocular camera at the same time, and the output is the depth map corresponding to the pixel coordinates of the two RGB images. The depth map here is a combination of two constraints. The first constraints are LR Warp Loss and the second constraints are the joint constraints Depth Warp Loss.

The second part of our construction is the visual odometry convolutional neural network (VO ConvNet). The input is two consecutive frames of RGB images from the right side of the binocular camera. The output is ego-motion between two consecutive frames of RGB images in the right eye of the binocular camera. Here VO ConvNet consists of two constraints, the first constraints are 2D Warp Loss and the second constraints are the joint constraints Depth Warp Loss.

Finally, our network test in a single camera for pure frame-to-frame VO estimation without any mapping, and can predict ego-motion without any scale ambiguity. We performed a comprehensive evaluation of our model in the KITTI dataset [11, 12].

3. Algorithm Framework

This section introduces our algorithm framework (shown in Figure 1), which learning the depth image $D_{R,t1}$ and $D_{R,t2}$ from the left and right images of the binocular camera at the same time and learning $T_{21} \in SE3$ from two consecutive frames.

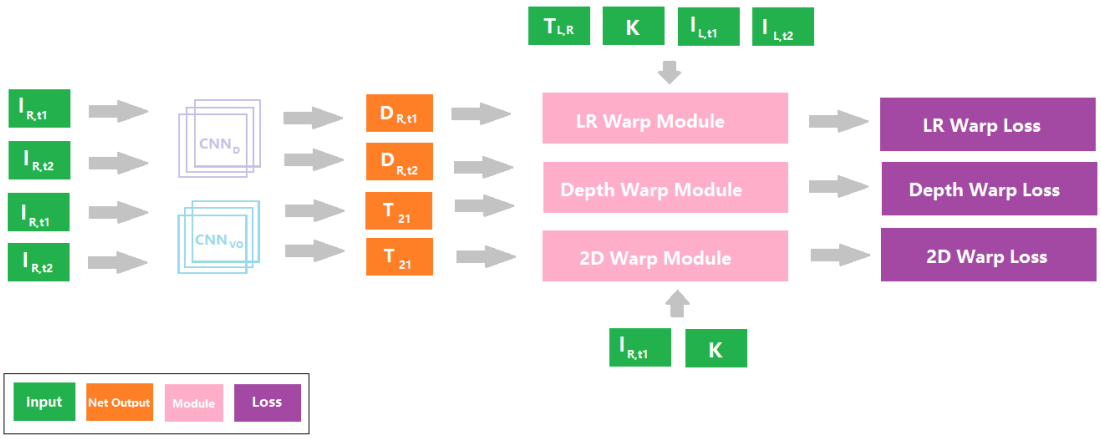

Figure 1. Algorithm Framework.

3.1 Network Architecture

Our network architecture is divided into two parts:

CNN_D is the Depth ConvNet for predicting the depth map, which using the encoder and decoder structure. For the encoder, in order to calculate the cost, we use the variant convolution network and the half filter (ResNet50-1by2) of ResNet50 [13].

CNN_{VO} is a VO ConvNet for predicting 6-DoF visual odometry. The network consists of 6 convolutional layers of two steps, followed by three fully connected layers. The last fully connected layer gives the 6D vector, which defines the transition T_{21} from the reference view to the warp view.

3.2 LR Warp Module

For a binocular camera, two frames of images at time $t1$ are defined as: $I_{L,t1}$ and $I_{R,t1}$. In addition, we define from $I_{L,t1}$ through the polar line geometry warp to $I_{R,t1}$. The warped image is called $I_{R,t1}^{(LR\_Warp)}$ (The upper right corner symbol is generated in the LR Warp Module, which is distinguished from other modules. The other upper right corners have the same meaning). The time $t2$ warp process is the same as time $t1$.

$$I_{R,t1}^{(LR\_Warp)} = f(I_{L,t1}, K, T_{LR}, D_{R,t1}) \tag{1}$$

$$I_{R,t2}^{(LR\_Warp)} = f(I_{L,t2}, K, T_{LR}, D_{R,t2}) \tag{2}$$

The image construction loss LR warp loss of the LR Warp Module is represented by the following formula:

$$L_{LR\_Warp} = \sum \left(\left| I_{R,t1} - I_{R,t1}^{(LR\_Warp)} \right| + \left| I_{R,t2} - I_{R,t2}^{(LR\_Warp)} \right| \right) \tag{3}$$

3.3 2D Warp Module

For a binocular camera, the two consecutive frames of the right-eye camera at time $t1$ and time $t2$ are defined as $I_{R,t1}$ and $I_{R,t2}$. We define from $I_{R,t1}$ through the direct method warp to $I_{R,t2}$. The warped image is called $I_{R,t2}^{(2D\_Warp)}$.

$$I_{R,t2}^{(2D\_Warp)} = f(I_{R,t1}, K, T_{21}, D_{R,t2}) \tag{4}$$

The image construction loss called 2D Warp Loss of the 2D Warp Module is represented by:

$$L_{2D\_Warp} = \sum \left(\left| I_{R,t2} - I_{R,t2}^{(2D\_Warp)} \right| \right) \tag{5}$$

3.4 Depth Warp Module

We have learned from Fang at el [14] that the constraint formula between the depth maps corresponding to two consecutive frames is:

$$A\xi = D_{R,t1} - D_{R,t2}^{(Depth\_Warp)} \tag{6}$$

Where,

$$A = \begin{bmatrix} -Y - Z_y f_y - Z_x XY \frac{f_x}{Z^2} - Z_y Y^2 \frac{f_y}{Z^2} \\ X + Z_x f_x + Z_x X^2 \frac{f_x}{Z^2} + Z_y XY \frac{f_y}{Z^2} \\ -Z_x Y \frac{f_x}{Z} + Z_y X \frac{f_y}{Z} \\ Z_x \frac{f_x}{Z} \\ Z_y \frac{f_y}{Z} \\ -1 - Z_x X \frac{f_x}{Z^2} - Z_y Y \frac{f_y}{Z^2} \end{bmatrix}^T$$

$$\xi = \begin{bmatrix} u & , & v \end{bmatrix} \in se3$$

The gradient of the depth image of $D_{R,t2}$ in A is expressed as $\nabla Z = (Z_x, Z_y)$, which is obtained by the Sobel operator; where f_x and f_y are normalized focal lengths, obtained by K. The 3D point $R = (X, Y, Z)$ is obtained by $D_{R,t2}$ through the camera pinhole model:

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = Z_{D_{R,t2}} K^{-1} \begin{bmatrix} u_{D_{R,t2}} \\ v_{D_{R,t2}} \\ 1 \end{bmatrix} \tag{7}$$

For binocular cameras, it is assumed that the $D_{R,t1}$ represents the depth map corresponding to $I_{R,t1}$, which is learned by CNN_D. $\xi \in se3$ is a Lie algebra representation of the relative camera pose transformation $T_{21} \in SE3$, which is learned by CNN_{VO}. Then bring these variables into formula (6) to get:

$$D_{R,t2}^{(Depth\_Warp)} = D_{R,t1} - A\xi \tag{8}$$

The image reconstruction loss between the warp view $D_{R,t2}^{(Depth\_Warp)}$ and the real view $D_{R,t2}$ is calculated as a supervised signal for training CNN_D and CNN_{VO}. Depth Warp Module's image construction loss Depth Warp Loss is represented by:

$$L_{Depth\_Warp} = \sum \left(\left| D_{R,t2} - D_{R,t2}^{(Depth\_Warp)} \right| \right) \tag{9}$$

3.5 Training loss

As described in Section 3.2, Sections 3.3 and 3.4, the main monitoring signals in our framework come from image reconstruction losses, while LR warp loss acts as an auxiliary supervisor. In order to obtain smooth depth prediction, edge-aware smoothing loss is the formula:

$$L_{ds}=\sum_{m,n}^{W,H}\left|\partial_x D_{m,n}\left|e^{-|\partial_x I_{m,n}|}+\left|\partial_y D_{m,n}\right|e^{-|\partial_y I_{m,n}|}\right|\right. \quad (10)$$

Where, $\partial_x(.)$ and $\partial_y(.)$ are gradients in horizontal and vertical direction respectively. The final loss function becomes:

$$L=\lambda_{LR\_Warp}L_{LR\_Warp}+\lambda_{2D\_Warp}L_{2D\_Warp}+\lambda_{Depth\_Warp}L_{Depth\_Warp}+\lambda_{ds}L_{ds} \quad (11)$$

Where, λ is the loss weight of each loss item and is obtained through training and fine tuning.

4. Result and Analysis

To validate the performance of our depth warp Constraints-based network, we evaluate the performance of our network by following the Odometry Split. We first compared the results with the very popular SLAM system ORB-SLAM [17] (with and without closed loop) and then compared the results with the one-eye training network [18] of Zhou et al. Last but not least, the results are compared to the no depth warp constraints network of Zhan et al [10]. Regarding the KITTI Visual Odometry dataset evaluation criteria, we routinely used subsequences of length (100, 200 ... 800) meters and reported the average translation and rotation errors of test sequences 09 and 10 in Table 1.

Table 1. Visual Odometry Result.

	Seq. 09 Terr(%)	Seq. 10 Rerr(°/100m)	Seq. 10 Terr(%)	Seq. 10 Rerr(°/100m)
ORB_SLAM(LC)	16.23	1.36	/	/
ORB_SLAM	15.30	0.26	3.68	0.48
Zhou et al.	17.84	6.78	37.91	17.78
Zhan et al.	11.92	3.60	12.62	3.43
Ours	5.73	2.66	8.54	2.74

Visual odometry result evaluated on Sequence 09, 10 of KITTI Odometry dataset as shown in Table 1. Terr is average translational drift error. Rerr is average rotational drift error. It can be seen from Table 1 that even without any further post-processing to repair the warp scale, our stereo-based ranging learning method can be much better than the method [10]. Our unsupervised method is superior to previous unsupervised methods and has comparable results with the supervised method, which reflects the effectiveness and advantages of our approach.

5. Conclusion

We propose an unsupervised learning framework based on depth warp constraints, which is used to train on binocular image data and then predict the ego-motion using visual odometry network on the monocular image. Our experimental results show that the accuracy and robustness of using binocular stereo sequences to learn these two tasks, using a binocular image to predict the depth map, and using a monocular image at different times to predict the relative pose of the camera. We also show the advantages of the single-view depth map predicted by the Depth Network using LR Warp Constraints, and the importance of the predicted depth map to the relative pose of the two frames of the subsequent predicted time series. In addition, we have proposed a novel Depth Warp Loss with the most advanced unsupervised single view depth and frame-to-frame odometry without scale ambiguity.

In addition, although our results are better than the existing method of using unsupervised deep learning, the current results are not comparable to the most advanced SLAM system. However, the use of deep learning methods can avoid constructing image features, and the use of unsupervised methods can use a large amount of unlabeled data. This is a promising and challenging research direction, and is expected to bring more inspiration to SLAM research.

Acknowledgments
This work is supported by National Natural Science Foundation (NNSF) of China under Grant 61873306.

References

[1] Scaramuzza D, Faundorfer F. Visual odometry: Part I: the first 30 years and fundamentals. *IEEE Robotics and Automation Magazine*, 2011, 18(4): 80−92
[2] Fraundorfer F, Scaramuzza D. Visual odometry: Part II: matching, robustness, optimization, and applications. *IEEE Robotics and Automation Magazine*, 2012, 19(2):78−90
[3] Craighead J, Murphy R, Burke J, Goldiez B. A survey of commercial and open source unmanned vehicle simulators. In: *Proceedings of the* 2007 *IEEE International Conference on Robotics and Automation. Roma, Italy:* IEEE, 2007.852−857
[4] Faessler M, Mueggler E, Schwabe K, Scaramuzza D. A monocular pose estimation system based on infrared LEDs. In: *Proceedings of the* 2014 *IEEE International Conference on Robotics and Automation (ICRA). Hong Kong, China:* IEEE, 2014. 907−913
[5] Meier L, Tanskanen P, Heng L, Lee G H, Fraundorfer F, Pollefeys M. PIXHAWK: a micro aerial vehicle design for autonomous flight using onboard computer vision. *Autonomous Robots*, 2012, 33(1−2): 21−39
[6] S. Weiss, M. W. Achtelik, S. Lynen, M. C. Achtelik, L. Kneip, M. Chli, and R. Siegwart, Monocular Vision for Long-term Micro Aerial Vehicle State Estimation: A Compendium, *Journal of Field Robotics, vol*. *30*, no. 5, 2013.
[7] Jianing Li. Research on Key Technologies of Augmented Reality System Based on RGB-D Camera [D]. Zhejiang University, 2017.
[8] Clark R, Wang S, Wen H, et al. VINet: Visual-Inertial Odometry as a Sequence-to-Sequence Learning Problem [C] AAAI. 2017: 3995-4001.
[9] Garg R, Vijay K B G, Carneiro G, et al. Unsupervised CNN for Single View Depth Estimation: Geometry to the Rescue [J]. 2016:740-756.
[10] Zhan H, Garg R, Weerasekera C S, et al. Unsupervised Learning of Monocular Depth Estimation and Visual Odometry with Deep Feature Reconstruction [J]. 2018.
[11] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? The kitti vision benchmark suite. *In Conference on Computer Vision and Pattern Recognition* (CVPR), 2012.
[12] A. Geiger, P. Lenz, C. Stiller, and R.Urtasun. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research* (IJRR), 2013.
[13] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *In Proceedings of the IEEE conference on computer vision and pattern recognition, pages* 770–778, 2016.
[14] Fang Z, Scherer S. Real-time onboard 6DoF localization of an indoor MAV in degraded visual environments using a RGB-D camera[C]// 2015 *IEEE International Conference on Robotics and Automation* (ICRA). IEEE, 2015.
[15] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos. Orb-slam: a versatile and accurate monocular slam system. *IEEE Transactions on Robotics*, 31(5):1147–1163, 2015.
[16] T. Zhou, M. Brown, N. Snavely and D. G. Lowe. Unsupervised learning of depth and ego-motion from video. *In CVPR*, 2017.