Dual CNN Models for Unsupervised Monocular Depth Estimation

Zhang, Liqiang

July 12, 2018

Abstract

A lot of progress has been made to solve the depth estimation problem in stereo vision. Though, a very satisfactory performance is observed by utilizing the deep learning in supervised manner for depth estimation. This approach needs huge amount of ground truth training data as well as depth maps which is very laborious to prepare and many times it is not available in real scenario. Thus, the unsupervised depth estimation is the recent trend by utilizing the binocular stereo images to get rid of depth map ground truth. In unsupervised depth computation, the disparity images are generated by training the CNN with an image reconstruction loss based on the epipolar geometry constraints. The effective way of using CNN as well as investigating the better losses for the said problem needs to be addressed.

1. Introduction

The image based depth estimation of scene is a very active research area in the field of computer vision. Several researchers have shown their great interest to work over this problem due to its wide and real scenario applications [4] such as autonomous and assistive driving human body pose estimation with 3D [3], robot assisted surgery , robot movement and grasping, learning human activities from RGBD videos, etc. The depth map from images can be estimated in various ways like structure from motion, monocular methods, etc. Such models are first trained in off-line mode with huge training database having monocular images as the inputs and corresponding depth maps as the labels.

Nowadays, the deep learning and convolutional neural networks (CNNs) based methods perform outstanding in most of the problems of computer vision such as image classification, object detection, semantic segmentation. The stereo surgery, robot movement and grasping [19], learning image pairs and ground truth disparity data are needed in human activities from RGBD videos [14], etc. The depth order to train the learning-based stereo models. In real scenario, creating such data is very difficult. Moreover, these methods generally create the artificial data which can not represent the real challenges appearing in natural images and depth maps.

In order to overcome the limitations of aforementioned supervised depth estimation techniques, some researchers have started working for unsupervised depth estimation which works reasonably good as well comparable to the supervised methods without need of any ground truth depth maps. The unsupervised methods utilize the underlying theory of epipolar constraints [2]. They computed a warp image from disparity map and right image. Finally, the error between original and reconstructed left image is used as the loss to train the whole setup in unsupervised manner.

2. Proposed Methodology

The proposed idea of dual network model (DNM) using CNN is illustrated in Fig. 1. This model is based on the 6 losses, thus referred as the DNM6 model. The DNM6 model has two CNN one for each left and right images of stereo pair. During training, the left image I^l and right image I^r are considered as the inputs to the left CNN named as CNN-L and right CNN named as CNN-R respectively. The $I_{i,j}$ refers to the $(i, j)^{th}$ co-ordinate of image I.

It is assumed that both I^l and I^r images are captured in similar settings. Both CNN's are based on the auto-encoder algorithm and combined these two networks named as dual network. The left image is reconstructed from the left disparity map d^l and input right image I^r, whereas the right image is reconstructed from the right disparity map d^r and input left image I^l as shown in the Fig. 1. The reconstructed left and right images are referred as I^l and I^r respectively throughout the paper.

To enforce the appearance of estimated images must be similar to the input image, a combination of L1 norm and Structural Similarity Index Metric (SSIM) loss term is used for both left and right images which is defined as Eq. 1:

$$C^{\beta}_{ap} = \frac{1}{N}\sum_{i,j}\alpha\frac{1 - SSIM(I^{\beta}_{i,j}, \hat{I}^{\beta}_{i,j})}{2} + (1-\alpha)\|I^{\beta}_{i,j} - \hat{I}^{\beta}_{i,j}\|. \tag{1}$$

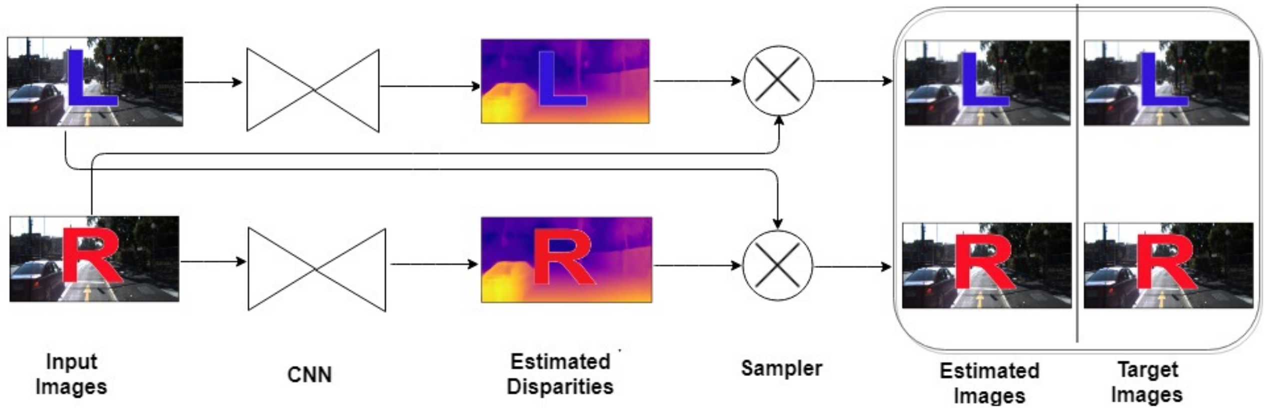

Figure 1. Pictorial representation of proposed Dual Network Model with 6 Losses (DNM6)

Method	Abs Rel	Sq	Rel	RMSE log	d1-all	a1	a2	a3
Godard et al. [1]No LR	0.123	1.417	6.315	0.220	30.318	0.841	0.937	0.973
Godard et al. [1]	0.124	1.388	6.125	0.217	30.272	0.841	0.936	0.975
DNM6 Model	**0.1157**	1.2037	5.830	**0.203**	**30.004**	**0.852**	**0.945**	**0.979**
DNM12 Model	**0.1157**	**1.1404**	**5.772**	**0.203**	30.342	0.848	0.944	**0.979**

Table 1. Experimental results by using proposed dual CNN based DNM6 and DNM12 models for unsupervised depth estimation over KITTI benchmark database. The training is done over KITTI training images and the evaluation is done over KITTI test images. The best results are highlighted in bold face.

where $\beta \in \{l, r\}, C^l_{ap}$ refers appearance matching loss between estimated left image and input left image and C^r_{ap} refers appearance matching loss between estimated right image and input right image and α represents the weight between SSIM and L1 norm. The image gradient based disparity smoothness loss is computed from both disparity maps to ensure the estimated disparity map should be smooth. Similar to [1], the disparity smoothness loss is given by the following as Eq. 2:

$$C^\beta_{ds} = \frac{1}{N}\sum_{i,j}|\partial_x d^\beta_{ij}|e^{-\|\partial_x I^\beta_{ij}\|} + |\partial_y d^\beta_{ij}|e^{-\|\partial_y I^\beta_{ij}\|}. \quad (2)$$

where $\beta \in \{l, r\}$, C^l_{ds} refers the disparity smoothness loss of left disparity map d^l estimated by CNN-L, C^r_{ds} refers the disparity smoothness loss of right disparity map d^r estimated by CNN-R and ∂ is the partial derivative.

3. Result

The proposed DNM6 and DNM12 models are used for the unsupervised monocular depth estimation over bench mark KITTI driving database. The results are reported in Table 1. The results of proposed DNM6 and DNM12 models are also compared with very recent state-of-the-art unsupervised method proposed by Godard et al. [1] with and without left-right (LR) consistency in Table 1.

References

[1] C. Godard, O. M. Aodha, and G. J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, pages 6602–6611, 2017. 2

[2] Y. Kuznietsov, J. Stuckler, and B. Leibe. Semi-supervised deep learning for monocular depth map prediction. In *CVPR*, pages 2215–2223, 2017. 1

[3] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, and R. Moore. Real-time human pose recognition in parts from single depth images. In *CVPR*, pages 1297–1304, 2013. 1

[4] A. Torralba and A. Oliva. Depth estimation from image structure. *IEEE TPAMI*, 24(9):1226–1238, 2002. 1