깊이추정을 위한 경량 스트라이드 컨볼루션 신경망

하템 이브라함, 강현수*
충북대학교

hatem@cbnu.ac.kr, *hskang@cbnu.ac.kr

Lightweight Strided Convolutional Neural Network for Depth Estimation

Hatem Ibrahem, Hyun-Soo Kang*
Chungbuk National University

Abstract

As the current computer vision applications requires high-speed algorithms, we propose a light-weight convolutional neural network architecture for depth estimation. The proposed architecture is simple and optimized for high speed applications while it attains a high depth estimation accuracies. It attains a high-speed of 25 fps on a GPU and 15 fps on a CPU. The proposed architecture can achieve high accuracy results thanks to the pixel-shuffling algorithm which perform the process of converting large number of small scale features into a large scale spatial map through pixels reordering. We train and evaluate the model on the challenging NYU depth v2 indoor dataset and compare our results with the state of the art methods on depth estimation.

Ⅰ. Introduction

The recent advances in depth estimation techniques showed the superior performance of the convolutional neural networks (CNN) in performing that task efficiently. Although, most recent CNN architectures depend on max-pooling (MP) for encoding the image data in a down-sampled representation, the MP operation is inefficient as it introduce a loss in the data. Recent research showed the effectiveness of the strides values higher than 1 in applying the convolutional layer as it down-samples the images using learnable parameters instead of mathematical operation such as MP. We also apply the pixel-shuffling operation [1] as a decoding step to up-sample the encoded features by the strided CNN in a single step using pixel reordering technique. The proposed depth estimation model is efficient in terms of accuracy as well as speed since it can perform the depth estimation task in real time (15~25 fps). An overview of the proposed method is shown in fig 1. The depth estimation is an important task since it is employed in 3D reconstruction, medical diagnosis, robotics, virtual, and augmented reality.

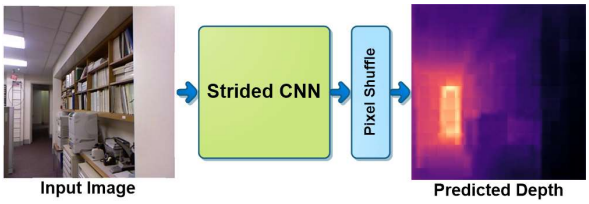

Fig 1. Overview of the proposed method.

Ⅱ. Proposed method

The proposed method depends on a sequence of 3x3 convolutional layers and batch normalization layers. The encoder consists of five down-sampling blocks each consist of 3 convolutional layers with 3x3 Kernel size, the first two convolutions with a stride of 1 and the third one is with a stride of 2 to down-sample the input features by half of the width and half of the height.

The decoding layer consists of a 1x1 convolutional layer to adjust the number of filters to 1024 to up-sample them directly using the pixel-shuffle operation which maps the input features of size H x W x (Cxr^2) to an up-sampled representation of the size rH x rW x C by reordering the pixel by mapping the corresponding pixel from each feature channel in the encoded representation to a super pixel in the output representation. This technique is inspired from a super resolution technique [1] in which they used the pixel shuffle operation to up-sample small feature maps of a low resolution image to from a higher resolution image. The pixel shuffle operation is shown graphically in fig 2, and the proposed CNN architecture with the pixel shuffle up-sampling is shown in detail in fig 3.

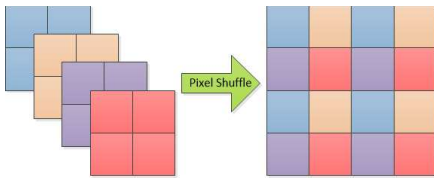

Fig 2. The pixel shuffle operation.

Fig 3. The detailed CNN architecture of the proposed method with the exact kernel size and filter count of each convolutional layer.

Ⅲ. Experimental results

We trained and evaluated the proposed method on NYU depthv2 [2] which is an indoor depth estimation dataset with images captured from bedrooms, kitchens, living rooms, and path rooms. The dataset consist of 1449 images with their corresponding depth map captured by Kinect RGB-D camera. We train the proposed method on 795 images and evaluate the model on the remaining 654 images. We constructed the proposed model using Tensorflow Keras and trained it using Adam's optimizer for 1500 epoch on a desktop computer with NVidia RTX3090 GPU.

We evaluate the predicted depth performance using the relative error (REL $=\frac{1}{N}\sum_P^N \frac{|y_t-y_p|}{y_t}$), the root mean squared of log error (RMSLE $=\sqrt{\frac{1}{N}\sum_P^N(\log(y_t)-\log(y_p))^2}$), and the delta depth accuracy $\delta = \max\left(\frac{y_t}{y_p}, \frac{y_p}{y_t}\right) < t$ for t threshold values 1.25, 1.252, and 1.253, where y_p and y_t are the predicted and target pixel values, respectively, and N is pixel's count in the depth map. The proposed method attains REL of 0.120, RMSLE of 0.385, and δ1 accuracy of 0.928 which outperform all the other recent depth estimation methods in Table I.

Table I. comparison between the proposed method and conventional depth estimation methods.

Model	REL	RMSLE	δ1	δ2	δ3
Laina et al [3]	0.127	0.573	0.811	0.953	0.988
Xu et al. [4]	0.121	0.586	0.811	0.954	0.987
Lee et al. [5]	0.131	0.538	0.837	0.971	**0.994**
SharpNet [6]	0.139	0.502	0.836	0.966	0.993
Zhang et al. [7]	0.144	0.501	0.815	0.962	0.992
Ours	**0.120**	**0.385**	**0.928**	**0.972**	**0.994**

Sample results obtained from our proposed method are shown in Fig. 4.

Ⅳ. Conclusions

The proposed method using the strided convolutional encoder and the pixel shuffle decoder is able to learn the depth estimation task efficiently as it could attain REL and δ1 accuracy of 0.120 and 0.928, respectively while working in a relatively high speed (~25 fps).

ACKNOWLEDGMENT

This work was supported by the Research Projects of "Development of automatic screening and hybrid detection system for hazardous material detecting in port container" funded by the Ministry of Oceans and Fisheries.

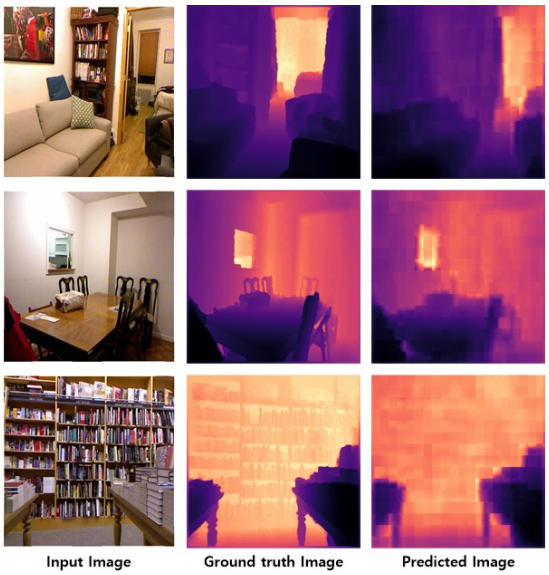

Fig 4. Sample results obtained from our proposed method on NYU depthv2.

REFERENCE

[1] W. Shi, J. Caballero, F. Huszar, J. Totz, ´A. P. Aitken, R. Bishop, D. Rueckert, Z. Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. IEEE Conf. Comput. Vis. Pattern Recog CVPR 2016.

[2] P. Kohli, N. Silberman, D. Hoiem, R Fergus. Indoor segmentation and support inference from rgbd images. In Eur. Conf. Comput. Vis. (ECCV), 2012.

[3] David Eigen, Christian Puhrsch, and Rob Fergus. Depthmap prediction from a single image using a multi-scaledeep network. In Advances in Neural Information Processing Systems (NeurIPs),2015.

[4] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, N. Navab. Deeper depth prediction with fully convolutional residual networks. International Conference on 3D Vision, 2016.

[5] J. Lee and C. Kim. Monocular depth estimation using relative depth maps. In IEEE Conf. Comput. Vis. Pattern Recog. (CVPR) 2019.

[6] M. Ramamonjisoa and V. Lepetit. Sharpnet: Fast and accurate recovery of occluding contours in monocular depth estimation. In proc. of IEEE Int. Conf. Comput. Vis. (ICCV) 2019.

[7] Z. Zhang, Z. Cui, C. Xu, Z. Jie, X. Li, J. Yang. Joint task-recursive learning for semantic segmentation and depth estimation. In Proc. of the Eur. Conf. Comput. Vis. (ECCV), 2018.