Learning-based depth estimation on wide angle images with non linear distortion

Julie Buquet [1 2], Jinsong Zhang[1], Patrice Roulet[3], Jean-Francois Lalonde[1], Simon Thibault[1]

[1] Faculty of Science and Engineering, Laval University 1045 Avenue de la médecine local 1033, Québec, QC G1V 0A6, Canada

[2] *Insititut d'Optique Graduate School, 2 Rue Augustin Fresnel, 91120, Palaiseau, France*

[3] *Immervision 2020 Boulevard Robert Bourassa local 2320, Montréal, Qc, H3A 2A5*

Julie.buquet.1@ulaval.ca, simont.thibault@phy.ulaval.ca, jean-francois.lalonde@gel.ulaval.ca, jinsong.zhang.1@ulaval.ca

Abstract: We developed a Convolutional Neural Network to estimate depth on wide-angle images using panomorph lens with controlled distortion. We simulated three different lens model and compared their performances based on their zone of augmented resolution.

1. Introduction

In computer vision, learning-based approaches have been proven very efficient for a lot of different tasks. Trained Convolutional Neural Networks allow us to extract information that surpasses Human Vision. From a single image, we can now realize tasks such as estimating distance, identifying objects and getting 3D contents that is relevant for a lot of fields: autonomous driving, medical operation assistance, surveillance …

1. Our Work

2.1 Dataset Generation

We used images from panomorph lenses based on controlled distortion produced by Immervision. This property leads to augmented resolution on chosen zones of the image

To simulate this effect, we generated 3 datasets of interior scenes from existing ones (Matterport 3D and SUNCG) with 3 different distortions : one linear (Fisheye), one with an augmented resolution at the center of the image and one at the edges of the image We wraped panoramic images by creating a mesh on the final image with (u,v) coordinates and realizing a three steps transformation. First, we express φ=arctan2(v,u) and Θ that represent respectively the angle between u and the camera and the field of view. We can express Θ as a function of the radius on the image r ($= \sqrt{v^2 + u^2}$) that include the distortion. Then, we come back to real world cartesian coordinates.

These coordinates are projected using the Latlong projection to get the original panorama mapping we obtained the desired datasets presented in Fig. 1. The images we worked on were 256x256 pixels. We generated 10 800 images from Matterport 3D and 289 from SUNCG.

Fig. 1. distortion simulation for each lens model : edges augmented, center augmented, Fisheye

2.2 Network architecture

The network we trained with these datas is composed of 4 modules. SqueezeNet realizes a multi-scale sampling of the image using dilated convolutions with very low cost and small architecture. Then the information extracted is rescaled with deconvolution blocs. After a fusion of all the layers, we use AlexNet to extract the depth estimation.

For this CNN, we have 11 981 649 parameters to train for a total memory cost of 596.45 MB.

2.3 Training of the CNN

We used 8591 images from Matterport and 164 from SunCG. We trained our network with different choices of hyperparameters and chose 450 epochs, a Batch size of 32 and a learning rate of 0,0001.
For the loss function, use *Ldepth* on the difference $e_i = \|d_i - g_i\|$ between the values of each pixel from our result and from the ground truth. The log function aims at giving more importance to the pixels closer in the image for the same error value. We combined it with *Lgrad* and *Lnorm* as in [1] that evaluate the error along image directions (x and y) focusing on edges reconstruction at different scales.

$$Lnorm = \frac{1}{n}\sum_{i=1}^{n} F(ei) \text{ with } F(ei) = ln(ei + \alpha) \text{ and } \alpha = 10^{-5} \quad (1)$$

$$Lgrad = \frac{1}{n}\sum_{i-1}^{n} F(\nabla x(ei)) + F(\nabla y(ei)) \quad (2)$$

$$Lnorm = \frac{1}{n}\sum_{i=1}^{n}\left(1 - \frac{\langle n_i^d, n_i^g \rangle}{\sqrt{\langle n_i^d, n_i^d \rangle}\sqrt{\langle n_i^g, n_i^g \rangle}}\right) \text{ with } n_i^{d/g} = [-\nabla x(di), -\nabla y(di), 1] \quad (3)$$

Each network was trained on GPU TitanX (RAM 12 G- 30 hrs of training) with the same parameters but different lens models and they all behaved the same during training.

2.4 *Test and comparison*

We evaluated the efficiency of the CNNs on a test dataset composed of 1072 images from Matterport and 13 from SUNCG. In order to compare the network, we had to dewarp each image obtained to compare it to the original ground truth.

We first evaluated individually the performances of all networks using RMS and REL. To compare the models we did this evaluation spatially and we added new edges reconstruction-based estimators as in [2]. Precision is presented in Table 1 and higher is better.

Table 1. Precision results obtained at the edges and center of the image.

	Edges	**Center**
center augmented lens	0.749	**0.570**
edges augmented lens	0.745	0.550
Fisheye	0.775	0.550

3. Conclusions

We created a CNN for depth map estimation and trained it with three types of non-linear distortion wide angle images. We observed better results on edges reconstruction with the centered augmented resolution lens. This result was not observed on the edges of the image probably because of the poor resolution at the beginning and the loss of information due to successive Warping and Dewarping. However it would be interesting to pursue the study with high-resolution images directly obtained with these lens. Also, as the results were more relevant for edges reconstruction, it would be interesting to apply the same experiment for another task such as object detection and identification.

4. References

[1] C. Wang K. Batmanghelich H. Fu, M. Gong and D. Tao, "*Deep ordinal regression network for monocular depth esti-mation*," (2018).

[2] Yan Zhang Takayuki Okatani Junjie Hu, Mete Ozay. *"Revisiting single image depth estimation : toward higher resolution maps with accurate object boundaries"*, (2018).