Exploring the use of line segments as an intermediate representation for depth estimation from a single image in indoor environments

Jose G Nava Zavala and Jose Martinez-Carranza
Instituto Nacional de Astrofisica, Optica y Electronica
Luis Enrique Erro No. 1, Sta. Ma. Tonantzintla, Pue
jose.nava@inaoep.mx, carranza@inaoep.mx

Abstract

We present a method for depth estimation from a single image using an intermediate representation. Rather than regressing depth from a chromatic image in RGB format, we propose to explore the use of line segments extracted from the original chromatic image aiming to assess whether low level features are useful to infer a depth image. Our proposed approach has been tested on the NYU-depth dataset for indoor scenes and on synthetic images created with Airsim. Our experiments show promising results confirming that it is possible to estimate a depth image from a single image containing line segments only.

1. Introduction

Since the pioneering work of Saxena [17], depth estimation from a single image has received wide attention in the last years, primarily due to deep learning techniques, which have contributed to the development of methods based on neural networks with impressive results [23]. However, several of these methods involve supervised training, thus requiring large amounts of labelled image data. One of the first datasets providing chromatic and depth images was that of KITTI [8], containing images captured from outdoor scenarios and intended to serve as a benchmark for robotic algorithms, e.g., visual odometry, visual SLAM and object detection algorithms. More specialised datasets have been made public such as the NYU depth dataset for indoor scenarios [18], and more recently, much larger datasets [16,19], whose particularity is that of providing thousand of synthetic images generated from virtual environments representing a wide set of scenarios, e.g., urban, neighbourhoods, woods, mountains, fields, farms, etc.

But, while the number of datasets could grow, one could wonder whether all the examples are necessary, this is, whether they enhance or bias the model. These are typical questions in machine learning, taking relevance in the context of depth estimation from a single image where a depth image may correspond to several images representing the same scene. Motivated by this, in this work, we explore the use of an intermediate representation that could be used as a prototype representing essential information of several image instances of the same scene, whose variability is produced by illumination changes. We propose to use an image containing line segments detected with the well-known Line Segment Detector (LSD) [9] as illustrated in Fig. 1. This proposal is inspired by the fact that humans are good at extracting complex information from sketches drawn on paper, such as recognising faces, objects and 3D structure. This has been explored in [3] for face recognition from sketches using deep learning, and the same concept can be extended to object and scene recognition [21].

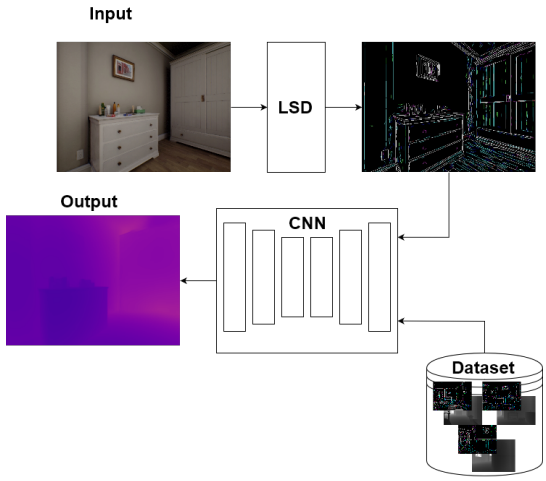

Figure 1. We propose to explore the use of line segments to estimate a depth image using a Convolutional Neural Network (CNN). We use the LSD [9] line detector to generate an image containing line segments and a conventional CNN for single image depth estimation [1] but trained with line segment images and their corresponding depth image.

2. Related work

There exist multiple methods for depth estimation, the state-of-the-art uses specialized hardware based methods such as LIDAR, Time of Flight or enhanced stereo-pair cameras such as Kinect and Kinect V2 or Intel Realsense. Hardware based methods are generally used directly, like LIDAR for autonomous cars, or, with the desire to remove this hardware barrier, to generate datasets such as the well known NYU-Depth and NYU-DepthV2 [18] and KITTI [7].

Monocular depth estimation is an ill-posed problem. Ambiguities can arise, scale ambiguities, translucent or reflective materials such as glass can have images where the geometry of a scene cannot be derived. This problem has been approached by many by using a CNN [1, 2, 5, 6, 10, 13, 15, 20] with a steady increase in performance as Vision tasks methods are applied and deeper networks are implemented, and changes in the network architecture like the use skip connections, have demonstrated that a monocular camera can have good results for this problem. Other methods include the use of supervised learning with MRF [17] with single image, and using this information to enhance stereo methods.

The use of Line Segments have been used to reconstruct 3D models from a series of images [14] to reconstruct 3d models from a series of photos, identifying the planes delimited by the line segments and fusing the planes to recreate the model. The use of lines for pose, Yu et al. [22] uses Line Segment detection matches 3D lines to captured 2D lines, this correspondence between captured 2d image from the camera and the 3D model allows for the estimation of pose.

LSD has been used in [11], an unsupervised method for depth estimation, but in contrast to us, this unsupervised method uses two adjacent images in an image training dataset to estimate relative motion and thus be able to reconstruct a relative 3D model of the views. The authors modified DepthNet [4] to learn the coefficients of 3D planar structures given the relative 3D model and in this context, segmented planes and lines are used, in the loss function, to evaluate the linear consistency of the projected 3D planes on the images. Thus, lines are used as part of the evaluation mechanism instead of being used as visual data to infer depth.

3. Methodology

We propose the use of Line Segments from a monocular chromatic camera as an intermediate representation to train a neural network that can estimate a corresponding depth map from a single image in a scene. Human architecture consists of planes with texture-less surfaces (e.g., walls, the floor) whose edges represent structural features that can be represented by line segments. For the neural network we propose the use of a state-of-the-art networks that produces high-quality depth maps with RGB images and process the input with the proposed representation.

For the LSD algorithm, we use OpenCV implementation which follows the implementation of Grompone et al. [9]. First, the RGB image is loaded in grayscale to apply LSD algorithm, the network used for training still requires three channel images, we used three different parametres for LSD initialization to find the Line Segments and each result is saved to a single image with each channel representing the different thresholds. We empirically found having three different parametres for LSD algorithm performed better than having the same parametres for the three channels.

The datasets used are NYU Depth Dataset V2 and a synthetic dataset. For the synthetic dataset capture we use Airsim, a simulator that allows an RGB-D free-floating camera, and a script with position a rotation to capture the chromatic and corresponding depth image. The camera was positioned on a point in the room and captured 512 images while rotating 0.7 degrees to capture around that point in the room, then moved to a different point and the process repeated to capture different points in all the rooms in the map. Both chromatic and depth images were saved with a resolution of 640 x 480.

For the neural network, we decided to use DenseDepth [1], a high-quality depth estimation from a monocular chromatic camera, other than changing the input data to train the network no other changes have been made, this allows us to directly compare the results for the same image, by converting to the LSD intermediary representation for training and prediction we can compare the results for the same input image.

For the quantitative evaluation to compare against state-of-the-art we use some of the same standard metrics used in those works average relative error (rel), root mean squared error (rms) and average $\log_{10}$ error (log10) [1, 5, 6, 12]

4. Experiments

For this work, we focused on assessing our approach on images representing indoor scenarios. We use the popular NYU-Depth dataset and a simulated dataset created with Airsim from which high quality depth images can be generated, but also to assess the performance of our approach in synthetic images.

For the simulated dataset, we used Airsim and a map from the Unreal Engine marketplace with minor changes, 129 different coordinates where taken and with a script for each scene 512 frames with chromatic and depth were saved as PNG, depth images where saved in range of 0-10m (Kinect ranges between around 15cm to 10m) and 16 bits per pixel. 103 of the scenes were used for training, 17 for testing and 9 from a different room were used for evalua-

Figure 2. Comparison among images with higher and lower brightness. Input images are shown in the first row. Depth image estimated with the network [1] trained with RGB images are shown in the second row. Depth images estimated with the same network but trained with line segments (LSD) [9] are shown in the third row. Note that our method is more consistent under these bright changes.

tion.

The results obtained with the NYU-Depth dataset are summarised in Table 1, which show that our method does not outperform the state-of-the-art, but it performs closely. Still, we demonstrate that the line segments of a single image contain enough information to estimate a depth image. When comparing with the simulated dataset a similar performance is observed for our method. For qualitative comparison, **??** shows some examples of depth image estimation with a network trained with RGB images and compared with the depth image estimated with the same network but trained with an image containing lines segments only. To evaluate the performance of our approach when evaluated with images whose visual texture may differ from that of the images used for training, we decided to train the DenseDepth network using images from NYU-Depth and then test it on synthetic images and vice-versa. This experiment was carried out using chromatic images as input for the training and also for images containing line segments. The results of this cross-over evaluation are shown in Table 2. RGB is used to indicate the use of chromatic images as input and LSD for images containing line segments.

5. Conclusion

We have presented initial results of a method that explores the use of line segments only to estimate a depth image using Convolutional Neural Networks. Inspired by methods performing neural inference from sketches such as face, object or scene recognition, we propose to use an image containing line segments as a form of sketch image. We have used the popular and robust LSD detector to extract line segments, carrying out experiments using the NYU dataset and a simulated dataset to asses the performance our approach with synthetic images. In both cases, our approach does not outperform the best methods in the state-of-the-art, but it obtains close results. In qualitative terms, depth images regressed from line segments are similar to those regressed with RGB images.

Method	rel↓	rms↓	log_{10} ↓
Eigen et al. [5]	0.158	0.641	-
Laina et al. [13]	0.127	0.573	0.055
MS-CRF [20]	0.121	0.586	0.052
Hao et al. [10]	0.127	0.555	0.053
Fu et al. [6]	**0.115**	0.509	**0.051**
Alhashim et al. [1]	0.123	**0.465**	0.053
Ours	0.162	0.675	0.068
Alhashim et al. [1]	**0.0976**	**0.515**	**0.047**
Ours	0.129	0.600	0.066

Table 1. Comparisons of different methods on the NYU Depth v2 dataset and our synthetic dataset. Top resultas are from NYU, bottom our synthetic dataset. The reported numbers are from the corresponding original papers. The best results are in bold, the higher the better.

Input Image	Training	Eval.	rel↓	rms↓	log_{10} ↓
RGB	NYU	Synth.	**0.293**	**1.331**	0.128
	Synth.	NYU	0.668	1.561	0.203
LSD	NYU	Synth.	0.304	1.348	**0.116**
	Synth.	NYU	0.861	1.755	0.233

Table 2. Comparison of generalization between the NYU-Depth V2 trained model and the synthetic dataset and with our proposed intermediate representation. The best results are bolded.

References

[1] Ibraheem Alhashim and Peter Wonka. High quality monocular depth estimation via transfer learning. *CoRR*,

abs/1812.11941, 2018. 1, 2, 3

[2] Yuanzhouhan Cao, Zifeng Wu, and Chunhua Shen. Estimating depth from monocular images as classification using deep fully convolutional residual networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(11):3174–3182, 2018. 2

[3] Shu-Yu Chen, Wanchao Su, Lin Gao, Shihong Xia, and Hongbo Fu. Deep generation of face images from sketches. *arXiv preprint arXiv:2006.01047*, 2020. 1

[4] Arun CS Kumar, Suchendra M Bhandarkar, and Mukta Prasad. Depthnet: A recurrent neural network architecture for monocular depth prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 283–291, 2018. 2

[5] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *CoRR*, abs/1406.2283, 2014. 2, 3

[6] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. *CoRR*, abs/1806.02446, 2018. 2, 3

[7] Alberto Garcia-Garcia, Sergio Orts-Escolano, Sergiu Oprea, Victor Villena-Martinez, and José García Rodríguez. A review on deep learning techniques applied to semantic segmentation. *CoRR*, abs/1704.06857, 2017. 2

[8] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012. 1

[9] Rafael Grompone von Gioi, Jérémie Jakubowicz, Jean-Michel Morel, and Gregory Randall. Lsd: a line segment detector. *Image Processing On Line*, 2:35–55, 2012. 1, 2, 3

[10] Zhixiang Hao, Yu Li, Shaodi You, and Feng Lu. Detail preserving depth estimation from a single image using attention guided networks. *CoRR*, abs/1809.00646, 2018. 2, 3

[11] Hualie Jiang, Laiyan Ding, Junjie Hu, and Rui Huang. Plnet: Plane and line priors for unsupervised indoor depth estimation. *CoRR*, abs/2110.05839, 2021. 2

[12] Lubor Ladicky, Jianbo Shi, and Marc Pollefeys. Pulling things out of perspective. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014. 2

[13] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. *CoRR*, abs/1606.00373, 2016. 2, 3

[14] Pierre-Alain Langlois, Alexandre Boulch, and Renaud Marlet. Surface reconstruction from 3d line segments. *CoRR*, abs/1911.00451, 2019. 2

[15] Fayao Liu, Chunhua Shen, and Guosheng Lin. Deep convolutional neural fields for depth estimation from a single image. *CoRR*, abs/1411.6387, 2014. 2

[16] Rafael Lopez-Campos and Jose Martinez-Carranza. Espada: Extended synthetic and photogrammetric aerial-image dataset. *IEEE Robotics and Automation Letters*, pages 1–1, October 2021. 1

[17] Ashutosh Saxena, Sung H. Chung, and Andrew Y. Ng. 3-d depth reconstruction from a single still image. *International Journal of Computer Vision*, 76(1):53–69, 2007. 1, 2

[18] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. *Computer Vision - ECCV 2012*, pages 746–760, 2012. 1, 2

[19] Wenshan Wang, Delong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. Tartanair: A dataset to push the limits of visual slam. 2020. 1

[20] Dan Xu, Elisa Ricci, Wanli Ouyang, Xiaogang Wang, and Nicu Sebe. Multi-scale continuous crfs as sequential deep networks for monocular depth estimation. *CoRR*, abs/1704.02157, 2017. 2, 3

[21] Peng Xu, Timothy M Hospedales, Qiyue Yin, Yi-Zhe Song, Tao Xiang, and Liang Wang. Deep learning for free-hand sketch: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 1

[22] Huai Yu, Weikun Zhen, Wen Yang, and Sebastian Scherer. Line-based 2-d–3-d registration and camera localization in structured environments. *IEEE Transactions on Instrumentation and Measurement*, 69(11):8962–8972, Nov. 2020. 2

[23] ChaoQiang Zhao, QiYu Sun, ChongZhen Zhang, Yang Tang, and Feng Qian. Monocular depth estimation based on deep learning: An overview. *Science China Technological Sciences*, pages 1–16, 2020. 1