Literature Review

By Davit Buniatyan

Depth Reconstruction using Recurrent Neural Networks

Abstract

In computer vision, pixel-depth estimation from camera is considered to be a basic problem. There has been vast amount of research and various sets of methods to tackle it from binocular or monocular, single images or image sequences desceribed in the following literature review. The limitations of methods include fixed or strictly moving camera, scene separation (outdoor or indoor), lack of dataset and more. Our research concentrates on depth reconstruction from monocular images using recurrent neural networks to capture the spatiotemporal context. In the beginning, we will concentrate on single image depth reconstruction. It is believed that the ability of human to reconstruct depth of a single image has statistical nature instead of geometrical interpretation.

1. Introduction

In the following review, we are going to discuss previous methods that tackled the problem from different perspectives under various contexts. We will discuss widely used benchmarks to compare results and introduce deep learning techniques that succeeded in similar problems such as estimation of optical flow or image segmentation. While there is a much prior work on estimating depth based on stereo images or motions there has been relatively little on estimating depth from a single image.

2. Single Images

One of the first methods to capture 3d structure of the scene was 'Tour into the picture' with spidery mesh interface that allowed doing perspective morphing from 2D images [1]. R. Szelisk and P. Torr introduced geometrically constrained methodology that has been foundation of approaches later defined in this review [2]. Another approach was extraction of the depth from Shading [3]. In this section, we will provide most descriptive methods that have been used to tackle the problem. It is generally considered to be ill posed problem, as there are infinitely many possible solutions.

2.1 Outdoor or Indoor [4, 5]

One of the first full automatic methods that creates directly 3D model from a single photograph is considered to be Automatic Photo Pop-up by D. Hoiem, A. Efros and M. Hebert at Carnegie Mellon University. The method is limited to dealing with only outdoor scenes such that it is possible to reconstruct a coarse, scaled 3D model from a single image by classifying each pixel as ground, vertical or sky and estimating horizon positions using features such as color, texture, image location and geometry. Later set of similar neighborhood of superpixels are considered to be constellations, such that one superpixel might belong to multiple constellations. Decision tree is trained for classifying each constellation ('ground', 'vertical' and 'sky') and thus each super pixel has a probability of being in each label. Then, the intersection of ground, vertical and sky lines are considered for defining the 3d orientation of planes in the image based on inferred camera properties. Then simple model is constructed using cutting and folding for generating those planes. The method achieves 30% accuracy for reconstructing accurate scenes.

Contrariwise, one of the first automatically 3d reconstructing algorithms from single indoor images is considered to be the method proposed by E. Delage, H. Lee and Andrew Y.Ng at Stanford that uses a dynamic Bayesian network model. Assuming 'floor-wall' geometry, the model takes as input: multi-scale intensity gradients in both the horizontal and vertical directions, as well as their absolute values and their squares, in addition to that, the neighborhood pixels, similarity to the floor chroma. Overall 50 features considered for estimating if the given pixel belongs to the floor or not. Using projective mapping, the 3d position of every pixel is located and accordingly walls reconstructing. The algorithm demonstrates the ability to detect robustly the floor boundary and generate accurate 3d reconstructions.

2.3 Learning Depth from Single Monocular Images [6]

Saxena introduced a method for predicting depth as a function of the image. The supervised learning method uses discriminatively trained Markov Random Field (MRF) that incorporates multiscale local and global features. The image is divided into small patches and single absolute and relative value is estimated for each patch. According to the paper monocular cues such as texture variations, texture gradients, occlusion, known objects sizes, haze, defocus are specific features that are good indicators of depth. Three local cues (texture variations, texture gradients

and haze) are extracted as local representatives of patches (in total 17 filters). The sum absolute energy and sum squared energy is computed by multiplying the pixel intensity by the filter in the patch. To capture global information the same processing is done at different scales and neighborhood feature vectors are incorporated. For relative depth extraction, the feature vector is computed by histograms. As particular patch depends on the features of its neighbors and other parts, MRF is used to model the depth of the patch and depth of its neighboring patches. Gaussian and Laplacian MRFs are considered for maximizing log-likelihood. 425 image-depth map pairs are collected for training and testing the model by achieving average error 0.132.

2.4 Make3D [7]

Further on, Saxena extended the work to reconstruct 3D scenes from single still image using Markov Random Field to infer a set of plane parameters that capture both 3D location and 3D orientation of image depth cues as well as the relationship between different parts of image. The algorithm starts by segmenting the image into many small planar surfaces. Planar surface in their turn are segmented into superpixels with high probability of being on the same real plane. The features such as relation of the depth between superpixels, neighboring structures, Co-planar structures and co-linearity are captured in MRF model considering their 'confidence' level. For parameter learning Multi-Conditional Learning was used where the graphical model is approximated by product of several marginal conditional likelihoods and plane parameters of 3d location and orientations are estimated. Totally 534 images-depthmaps have been collected and achieved 64.9% qualitatively correct 3D models by outperforming other methods and becoming state-of-the-art technique in 3D scene reconstruction. Furthermore, the model is extended to capture full photorealistic model based on multiple images.

2.5 From Semantic Labels [8]

B. Liu proposed a method for single image depth estimation from semantic labels. The work addresses the problem of depth perception from a single monocular image through the incorporation of semantic information. The algorithm has two phases. The first phase solves the problem of multi-class image labeling. Pixels are labeled with general categories (for example: sky, tree, road etc.) similar to outdoor/indoor scenes. The second phase is to learn the depth estimations for each semantic class based on each location. For example if the label is sky then the pixel is infinitely. Embedding several more precise assumptions (for example the class 'road' and 'ground' are usually on the horizontal plane or it more probable that 'road' and 'building' are near to each other) into Markov random field the depth map from single image is estimated. The model was trained on Saxena's outdoor dataset and achieved state-of-the-art results on $\log_{10}$ metric and comparable performance to state-of-the-art for the relative error metric.

2.6 Depth Fusion [9,10]

By the availability of cloud computing and 3D content, J. Konrad purposed automatic method for 2D-to-3D conversion based on the dataset of existing stereoscopic videos such as Youtube 3D. Taking an assumption that there exists similar stereo pair in the dataset given arbitrary image, the method extracts depth information of the pair and uses for estimation of the input image and generates stereoscopic view. In details, the algorithm does, k nearest-neighbours (kNN) search to find k most similar 2D left images in this case from Youtube 3D, estimates the parameters for warping the query image to the left image of each kNN stereopair, warps extracted disparity of video pairs to the query images, does median filtering of the k warped disparity field including cross-bilateral smoothing and generates the right image.

The approach later was reconsidered on extracting depth information from still images, such that given an image the kNN searches over either stereopairs or image+depth repository, fuses the depth maps (instead of disparity) and similarly filters. At the end, the algorithm renders the stereo image based on median field followed by suitable processing of occlusions and newly exposed areas. The following approach was considered with Make3D and outperformed under indoor scenes, even though Make3D was originally trained on outdoor scenes. The training data included 1449 pairs of RGB and depth collected on Kinect camera.

2.7 Pulling things out of Perspective [11]

The paper mostly discusses about the significance using the properties of the perspective geometry that reduces the learning pixel-wise depth classifier to much simpler classifier predicting only the likelihood of a pixel being at an arbitrarily fixed canonical depth. It takes the advantage over other methods by constructing an image pyramid and in the result avoiding a pitfall, which requires the training set to have the same object at different depths in order to predict accurately. It states that that for stereo and multi-view images join semantic segmentation and 3D reconstruction leads to better results than performing each task separately.

3. Image Sequences

3.1 Recovery from video [12]

G. Zhang introduces a novel method for automatically constructing view-dependent depth map per frame by making depth values in multiple scenes consistent and assigning distinctive depth values for pixels. In order to meet objectives of consistency and distinction, the algorithm uses segmentation as the initial step and then iteratively refines disparities in a pixelwise manner. More precisely, the algorithm estimates depth maps for each pixel by belief propagation, incorporates segmentation, then initialize disparities with segmentation and plane fitting using a nonlinear continuous optimization, and refine the final output with bundle optimization. Even though lack of comparison with other algorithms, it was evaluated qualitatively and quantify and achieved visually very appealing results. However, if there is no sufficient camera motion, the recovered depths could be less accurate.

3.2 DepthTransfer [13]

Similar to Konrad's approach, depth extraction from video was proposed which is called DepthTransfer. The algorithm, given a database RGBD images, finds the similar images to the input image in RGB space. Then, found images and their depths are warped in order to align with the input image. Finally, an optimization procedure is used to interpolate and smooth the warped candidate depth values that result in inferred depth. In order to effectively find similar images their high-level image features are extracted and top 7 similar frames are selected for warping. The optical flow with motion estimations is used for consecutive frames depth maps to be matched continuously. Separate dataset is collected for training the algorithm and results outperformed both DepthFusion and Make3D on single images on Make3D dataset and outperformed Depth Fusion train and tested on NYU Depth dataset even though the results are not that perceptual. According to the paper, generalization to interior scenes is much more difficult than outdoor.

3.3 Intrinsic Depth: Improving Depth Transfer with Intrinsic Images (will be added)

3.4 Photo Tourism: Exploring photo collections in 3D (will be added)

4. Deep Learning

As Convolutional Neural Networks (CNN) proved to be successful method for the image feature extraction and, hence, was widely used in current the state-of-the-art solution in various computer vision applications most notably object detection and scene classification. Here we present state-of-the-art models for depth estimation from single images and other applications that CNNs have been particularly useful.

4.1 Multi-Scale Deep Network [14]

A novel approach has been developed due to increase of Neural Networks activity in Computer Vision tasks that estimates the depth from single image directly regressing on the image. The method has two core components. The first component captures global structure scene and the second captures the local context. The goal of global structure is to predict overall depth map and embed into the local component. It takes down scaled image and passes through 4 convolutional and 2 fully connected layers. The local structure initially processes the image through a convolutional layer and concatenates with output from global component. Finally, after two convolutional layers the final depth image is computed. In addition to NYU and Make3D dataset the, at that time the method achieved state-of-the-art performance on KITTY dataset.

4.2 Convolutional Neural Fields [15]

Deep Convolutional Neural Field has been proposed for depth estimation using CNN as a feature descriptor based on Multi-Scale Deep Network and continuous Conditional Random Field (CRF) as loss layer, which is minimized with negative log-likelihood. The model requires segmented image of superpixels to feed to a unary layer and a pairwise layer. Both then outputs to CRF layer. Unary layer is a CNN with 5 convolutional and 4 fully connected layers that outputs n-dimensional vector contained regressed depth values of the n superpixels. Simultaneously, pairwise layer takes similarity of all neighboring superpixel pairs as input, feed each of them to a fully-connected layer and outputs a vector containing all the 1-dimensional similarities. This is by far the most advanced method that uses deep learning and outperforms state-of-the-art methods in single image depth estimation trained and tested on NYU v2 and Make3D dataset.

4.3 FlowNet

4.4 AutoEncoder

A neural network called autoencoder is used for dimensionality reduction of information trained on the input, itself. The hidden layer is contains less cells than the input/output layer. Hence, by minimizing the loss, useful features are preserved and encoded in hidden layer. K. Konda and R. Memisevic introduced unsupervised deep learning model using autoencoder that captures the depth and motion from sequence of stereo image pairs [16]. They extended autoencoder model by taking two separate inputs of the stereo image patches and concatenating into one hidden layer. As depth/motion recovery is usually a middle step for further problem solving, action recognition has been successfully attempted using extracted features from autoencoder, which resulted in state-of-the-art performance in 3D activity analysis.

4.5 Recurrent Neural Networks

Recurrent Neural Networks (RNNs) is a type of neural network that processes sequences one step at a time by passing the hidden state of the previous timestep to the next one and hence generates output taking into consideration context. RNNs are used primarily for many-to-one (when the sequence of information is parsed and final single output is made), one-to-one (at each timestep RNN produces one output) and many-to-many (all data is processed) labeling types [17]. Recurrent Neural Networks achieved successful results on language modeling, translation and character recognition. Simple RNN with unit activation cell suffers from vanishing-gradient problem. The hidden state information either diverges or converges to zero. Long-Short Term Memory cells has been proposed by S. Hochreiter and J. Schmidhuber [18] to capture long term dependencies by learning when to remember and forget information. LSTM is based on four cells called gates: input, memory, forget, output. Another limitation of RNN is that the model has only past context at each timestep. Bi-directional RNN has been proposed which consists of simultaneous layers which processing the input from first to last and from last to first correspondingly and then concatenates the output to perform prediction with past as well as future context [19]. RNNs have been particularly useful in image or video captioning [20, 21] for generating descriptions after preprocessing of convolutional layers, however in order to be useful for more structured input in computer vision 1-dimensional limitation should be extended.

4.5.3 Multidimensional [22]

Alex Graves proposed multidimensional Recurrent Neural Network that extends temporal RNNs onto n-dimensional space. At each time-step the architecture takes an input of multi-dimensional data such as the pixel of image, video or MRI and the hidden state of the previous steps across all dimensions. The advantage of this model is that it scales linearly as the number of inputs grows and it is applicable to capture the long-term context of the data contrariwise to convolutional neural networks. As in bidirectional RNN model where the past and the future is encountered in the output by two simultaneous layers that parse the data from both sides at the same time, Graves introduced multi-directional model where the data is parsed from all possible dimensions. The model was tested on MNIST dataset. It showed equal results on the dataset, however outperformed state-of-the-art method CNN modeling in modeling image warping. Recent successful applications of multi-dimensional RNNs are presented below.

- **Scene Labeling [23]**

  Complete end-to-end learning based approach using 2D LSTM was proposed for scene labeling task. The network consists of 3 layers: input, hidden and output. The input layers takes RGB pixels (3xnxn) with sliding window, processes the information using 4 stacked multidirectional multidimensional layer to capture whole context and passes to the next two consecutive LSTM hidden layers at each timestep, which produces final result with softmax output layer. The model is the state-of-the-art technique for Stanford Background and Sift Flow dataset by outperforming other methods that embed CNNs. Even though the training takes 10 days on single CPU, the processing of each image is linear and takes 1.3s. Implemented on GPU. In conclusion, this paper supports the argument that RNNs are capable of handling pixel-level labeling tasks by capturing both local and global context.

- **Pixel Recurrent Neural Network [24]**

  Recent paper by Google DeepMind for modeling the distribution of natural images for predicting the image used enhanced version of 2D LSTM to recover unseen image parts without external information. It has been shown that PixelRNNs significantly improved results on the Binary MNIST and CIFAR-10 datasets. Furthermore, they introduced ImageNet as a new dataset. And final conclusion has been reached that the PixelRNN is capable of modeling spatially local and long-range correlations and are able to produce images that are sharp and coherent

4.5.4 Grid-LSTM [25]

Furthermore, Nal Kalchbrenner, Ivo Daihelka and Alex Graves extend multidimensional capability of Recurrent Neural Networks by introducing Grid Long-Short Term Memory, a network of LSTM cells arranged in a multidimensional grid that can be applied to vectors, sequences or higher dimensional data such as images. Grid-LSTM outperforms Multidimensional by resolving instability of large grids by adding cells along the depth dimension. Additionally unit cells are simplified and unified such that from all dimensions the input and output consists of two gates: the memory and hidden gate. Each block is consisted from LSTM cells along each dimension. The network achieves state-of-the-art results on Wikipedia character prediction benchmark among neural nets and outperforms a phrase-based reference system on a Chinese-to-English translation task. Additionally achieves near state-of-the-art results on MNIST dataset.

4.6 Video modeling

- Unsupervised video representation via LSTM [26] (will be added)
- ConvLSTM [27] (will be added)

5. Experimentation

5.1 Datasets

Available depth datasets are available.

- Make3D - make3d.cs.cornell.edu/data.html
- B3DO - kinnectdata.com
- NYU depth v1 and v2 - cs.nyu.edu/˜silberman/datasets
- RGB-D dataset - cs.washington.edu/rgbd-dataset
- DepthTransfer - kevinkarsch.com/depthtransfer
- SUN3D - http://sun3d.cs.princeton.edu

5.2 Benchmarking

In the below, available benchmarking criteria for comparison

- Average relative error (rel): $\frac{1}{T}\sum_p \frac{\left|d_p^{gt}-d_p\right|}{d_p^{gt}}$
- Root mean squared error (rms): $\sqrt{\frac{1}{T}\sum_p\left(d_p^{gt}-d_p\right)^2}$
- Average $\log_{10}$ error (log10): $\frac{1}{T}\sum_p\left|\log_{10} d_p^{gt}-\log_{10} d_p\right|$
- Accuracy with threshold *thr*: percentage (%) of d_p s.t: $max(\frac{d_p^{gt}}{d_p},\frac{d_p}{d_p^{gt}}) = \delta < thr$;

Where d_p^{gt} and d_p are the ground-truth and predicted depths respectively at pixel indexed by p, and T it is the total number of pixels in all the evaluated images.

5.3 Current Results

State-of-the-art results on NYU v2 dataset

Method	Error (lower is better)			Accuracy (higher is better)		
	rel	log10	rms	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Make3D	0.349	-	1.214	0.447	0.745	0.897
DepthTransfer	0.35	0.131	1.2	-	-	-
Discrete-continuous CRF	0.335	0.127	1.06	-	-	-
Ladicky et al.	-	-	-	0.542	0.829	0.941
Multi-Scale Deep Network	**0.215**	-	0.907	0.61	**0.887**	**0.971**
Convolutional Neural Field	0.230	**0.095**	**0.0824**	**0.614**	0.883	**0.971**

State-of-the-art results on Make3D dataset

Method	Error (C1) (Lower is better)			Error (C2) (Higher is better)		
	rel	log10	rms	rel	log10	rms
Make3D	-	-	-	0.370	0.187	-
Semantic Labelling	-	-	-	0.379	0.148	-
DepthTransfer	0.355	0.127	9.20	0.361	0.148	15.10
Discrete-continuous CRF	0.335	0.137	9.49	0.338	0.134	13.29
Convolutional Neural Field	**0.314**	**0.119**	**0.860**	**0.307**	**0.125**	12.89

*These data is taken from [15] and not all models in the paper are presented

6. Conclusion

7. Bibliography

[1] Horry, Youichi, Ken-Ichi Anjyo, and Kiyoshi Arai. "Tour into the picture: using a spidery mesh interface to make animation from a single image." Proceedings of the 24th annual conference on Computer graphics and interactive techniques. ACM Press/Addison-Wesley Publishing Co., 1997.

[2] Szeliski, Richard, and Philip HS Torr. "Geometrically constrained structure from motion: Points on planes." 3D Structure from Multiple Images of Large-Scale Environments. Springer Berlin Heidelberg, 1998. 171-186.

[3] Zhang, Ruo, et al. "Shape-from-shading: a survey." Pattern Analysis and Machine Intelligence, IEEE Transactions on 21.8 (1999): 690-706.

[4] Hoiem, Derek, Alexei A. Efros, and Martial Hebert. "Automatic photo pop-up." ACM Transactions on Graphics (TOG) 24.3 (2005): 577-584.

[5] Delage, Erick, Honglak Lee, and Andrew Y. Ng. "A dynamic bayesian network model for autonomous 3d reconstruction from a single indoor image." Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on. Vol. 2. IEEE, 2006.

[6] Saxena, Ashutosh, Sung H. Chung, and Andrew Y. Ng. "Learning depth from single monocular images." Advances in Neural Information Processing Systems. 2005.

[7] Saxena, Ashutosh, Min Sun, and Andrew Y. Ng. "Make3d: Learning 3d scene structure from a single still image." Pattern Analysis and Machine Intelligence, IEEE Transactions on 31.5 (2009): 824-840.

[8] Liu, Beyang, Stephen Gould, and Daphne Koller. "Single image depth estimation from predicted semantic labels." Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. IEEE, 2010.

[9] Konrad, J., et al. "Automatic 2d-to-3d image conversion using 3d examples from the internet." IS&T/SPIE Electronic Imaging. International Society for Optics and Photonics, 2012.

[10] Konrad, Janusz, Meng Wang, and Prakash Ishwar. "2d-to-3d image conversion by learning depth from examples." Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on. IEEE, 2012.

[11] Ladicky, Lubor, Jianbo Shi, and Marc Pollefeys. "Pulling things out of perspective." Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on. IEEE, 2014.

[12] Zhang, Guofeng, et al. "Consistent depth maps recovery from a video sequence." Pattern Analysis and Machine Intelligence, IEEE Transactions on 31.6 (2009): 974-988.

[13] Karsch, Kevin, Ce Liu, and Sing Bing Kang. "Depth extraction from video using non-parametric sampling." Computer Vision–ECCV 2012. Springer Berlin Heidelberg, 2012. 775-788.

[14] MLA Eigen, David, Christian Puhrsch, and Rob Fergus. "Depth map prediction from a single image using a multi-scale deep network." Advances in Neural Information Processing Systems. 2014.

[15] Liu, Fayao, Chunhua Shen, and Guosheng Lin. "Deep convolutional neural fields for depth estimation from a single image." arXiv preprint arXiv:1411.6387 (2014).

[16] Konda, Kishore, and Roland Memisevic. "Unsupervised learning of depth and motion." arXiv preprint arXiv:1312.3429 (2013).

[17] Graves, Alex. Supervised sequence labelling with recurrent neural networks. Vol. 385. Heidelberg: Springer, 2012.

[18] S. Hochreiter and J. Schmidhuber. Long Short-Term Memory. Neural Computation, 9(8):1735–1780, 1997.

[19] Schuster, Mike, and Kuldip K. Paliwal. "Bidirectional recurrent neural networks." Signal Processing, IEEE Transactions on 45.11 (1997): 2673-2681.

[20] Karpathy, Andrej, and Li Fei-Fei. "Deep visual-semantic alignments for generating image descriptions." arXiv preprint arXiv:1412.2306 (2014).

[21] Jeff Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell, Long-term Recurrent Convolutional Networks for Visual Recognition and Description, arXiv:1411.4389 / CVPR 2015

[22] Graves, Alex, Santiago Fernandez, and Ju3rgen Schmidhuber. "Advances in Neural Network Architectures-Multi-dimensional Recurrent Neural Networks." Lecture Notes in Computer Science 4668 (2007): 549-558.

[23] Byeon, Wonmin, et al. "Scene Labeling with LSTM Recurrent Neural Networks." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015.

[24] van den Oord, Aaron; Kalchbrenner, Nal; Kavukcuoglu, Koray, "Pixel Recurrent Neural Networks", arXiv:1601.06759 (2016)

[25] Kalchbrenner, Nal, Ivo Danihelka, and Alex Graves. "Grid long short-term memory." arXiv preprint arXiv:1507.01526 (2015).

[26] Srivastava, Nitish, Elman Mansimov, and Ruslan Salakhutdinov. "Unsupervised learning of video representations using lstms." arXiv preprint arXiv:1502.04681 (2015).

[27] Shi, Xingjian, et al. "Convolutional LSTM network: A machine learning approach for precipitation nowcasting." arXiv preprint arXiv:1506.04214 (2015).