DEPTH ESTIMATION FROM SINGLE IMAGES USING MODIFIED STACKED GENERALIZATION

H. Mohaghegh[1], S. Samavi[1], N. Karimi[1], S.M.R. Soroushmehr,[2, 3], K. Najarian[2,3,4]

[1]Department of Electrical and Computer Engineering, Isfahan University of Technology, Iran
[2]Emergency Medicine Department, University of Michigan, Ann Arbor, USA
[3]U-M Center for Integrative Research in Critical Care (MCIRCC), University of Michigan, Ann Arbor, USA
[4]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, USA

ABSTRACT

Despite the rapid growth of 3D displays in the last few years, insufficient supply of 3D contents has led to considerable effort in devising 2D to 3D conversion algorithms. Inferring associated depth from single 2D image is still a controversial issue in these algorithms. In this paper we propose an algorithm, which unlike previous strategies, aggregates both global and local information from a pool of images with known depth maps. Hence, we propose to extract a set of features from the image patches of globally similar images in a large 3D image repository. These features describe powerful monocular depth perception cues. Using these relevant and robust features and using modified stacked generalization learning scheme, our scheme directly extracts an accurate depth map from given images. Experimental results demonstrate that our method has surpassed state-of-the-art algorithms in both quantitative and qualitative analysis.

Index Terms— 2D to 3D conversion, Depth map, Monocular cues, Modified stacked generalization

1. INTRODUCTION

Today there is a rapid increase in production of 3D capable hardware such as TVs, gaming consoles and smart phones, but 3D contents cannot keep up with this pace, which results in a gap between 3D displays and 3D contents. To bridge this gap there is a need to convert available 2D contents to 3D. The conversion process consists of two main stages: depth extraction from single monocular image and depth image based rendering (DIBR) to synthesize virtual view images [1]. Recently numerous researches have focused on the first stage, which is a more challenging problem, since one single image can be produced from numerous possible scenes [2]. Moreover depth estimation from a single image enables further applications including video surveillance, 3D modeling [3], robotics and target tracking.

Generally, there are two main methods for depth estimation from a single image: semi-automatic methods [4] that expect a human operator interaction and automatic methods in which no operator intervention is required. Some of monocular cues used in conventional automatic methods include focus/defocus [5], texture gradient [6], shading [7], atmospheric effects and geometric perspective [8]. Among the learning-based methods, Saxena in [9] proposed a supervised learning method using linear regression and a Markov Random Field. Later on, they improved their method by using max-margin parameter learning [10]. In [11], Lin et al. proposed an SVM based framework which uses the spatial and frequency descriptors. Konrad et al. in [1] implemented a global estimation of whole depth map of a single image using nearest neighbor search based on Histogram of Oriented Gradients (HOG) features. Furthermore depth of query image derived from the median of depth of retrieved images where applying median operator leads to a depth map with low accuracy. Another strategy for fusion of depth candidates was proposed in [12]; this study uses weighted combination of representative depth fields. Moreover in [13] weighted median statistics is applied after retrieving matching candidates.

Karsch et al. devised a non-parametric depth transfer strategy where depth map was obtained by transferring the depth of a number of matching candidates followed by global optimization [14]. The main drawback of this method is that it is time consuming and its outputs are over smooth due to global optimization. By aggregating semantic labels, Liu et al. in [15] improved performance of depth estimation process. Requiring additional information such as semantic labels is the major disadvantage of this method. Eigen et al. trained a deep network to estimate the global structure of image [2]. While their results are more accurate, their relying on a large training set is a drawback of this method. In another research, authors propose a parameter transfer scheme which models the correlation between images and their depths using a set of parameters [16].

In this paper we present a simple but effective framework for automatically estimating an accurate depth map from a single 2D outdoor image. To this end, we propose a multiple level learning model to locally predict depth of image patches by exploiting a set of effective features. These features are highly correlated to distance in natural scenes. We only apply our algorithm on globally

similar images, rather than applying it to the whole dataset, and this improves the performance of our depth estimation.

In the rest of this paper Section 2 describes the proposed method. The experimental results are presented in Section 3, followed by the conclusion in Section 4.

2. PROPOSED METHOD

Our single image depth estimation has four main steps (i) given a database of RGBD images and an input image, we retrieve similar candidate images in terms of photometric properties from the database; (ii) a set of spatial domain features are extracted locally from image patches; (iii) depth estimation module predicts the depth map of the image by a mapping process from feature domain to depth domain; and (iv) predicted depth is then refined by using an edge-aware median filter. This results in the final inferred depth map.

2.1. Similar image retrieval

In the first stage of our scheme, a number of images with similar scenes are selected to be sent into the rest of algorithm. This pruning strategy helps to reduce or even remove the potential outlier data to be used in the learning phase. This is all based on the assumption that, images with analogous photometric appearance are expected to have similar depth structures.

In order to find these matching candidates, GIST features are used, which are high level features that represent global properties of a scene [17]. To obtain the matching score between two images, in terms of GIST features, the sum of squared differences (SSD) is defined as $SSD = \|G - G_i\|^2$.

Denoting G and G_i as the GIST feature vectors for input image and i^{th} candidate image in 3D repository, respectively. Now the candidate images which match the input in terms of global structure of the scene are available. Relying on only global information of input images and applying median operator across the associated depths, lead to inaccurate and spatially smooth depth map [1]. Instead, we try to employ local properties as well by dividing the input image into non overlapping 16×16 patches and extracting local features from each patch.

2.2. Image feature extraction

Perceiving depth of a single 2D image arises from a variety of monocular depth cues such as texture variations, blurriness, color/haze, etc [18]. Here we focus on a set of features which characterize these cues containing rich structural and statistical information from image patches.

We exploit these reliable features to learn the relationship between the image patches and the corresponding depth values. Employing such robust features which contain useful and important depth information helps our depth estimation module accurately learn and predict depth patches. Hence, each patch with its vertical coordinate, as an effective depth related feature, enters this stage and the features introduced below are extracted.

2.2.1. Texture

Texture variations of 2D images can provide human visual system with a good depth perception cue. Usually, texture of objects may seem different at different distances from the viewer [18]. To represent texture, we employ a powerful feature descriptor, called local binary pattern (LBP) [19]. For this purpose, a 256 bin histogram of LBP is computed for each patch. Gray-value variance and entropy of patches are also used for analyzing texture cue [20].

2.2.2. Blurriness

It is known that in a good quality photo, captured by an ideal camera, all points in the depth of field (DoF) are sharp and clear. However, slightly blurry pixels do exist due to possible depth variations in the image. This small defocus blurriness, named as just noticeable blur (JNB), proportionally changes with distance [21]. Hence, this feature can provide information about object's depth even in images which do not have a narrow DoF. To represent blurriness, average value of just noticeable blur map, is computed in each patch.

2.2.3. Color and lightness

Color information can be employed by human visual system to infer depth. This is especially true in outdoor images, which contain limited range of colors and each color represents a different entity [22]. For example the sky, with farthest distance, is usually blue or white. We use HSV color space which allows measurement of perceptual color characteristics. We use the average value of each color component (H, S and V). We also use 5-bin histogram and 3-bin histogram of H and S component respectively as color descriptors [23].

Potetz et al. in [24] indicate that light coming from farther distance is commonly absorbed by atmosphere. Therefore, it can be inferred that in outdoor images distance is correlated to brightness. By transforming color image to CIELAB color space, the average of luminance of each patch is used to represent its brightness.

By itself, any of the mentioned monocular cues have a little to do with depth perception, but applying them in various networks and combining network's output with each other, make them a set of powerful cues.

2.3. Depth estimation

Artificial neural networks are computational models inspired by biological human neural networks used for nonlinear approximation. In order to improve the performance of prediction, we use modified stacked generalization learning method [25]. This model combines

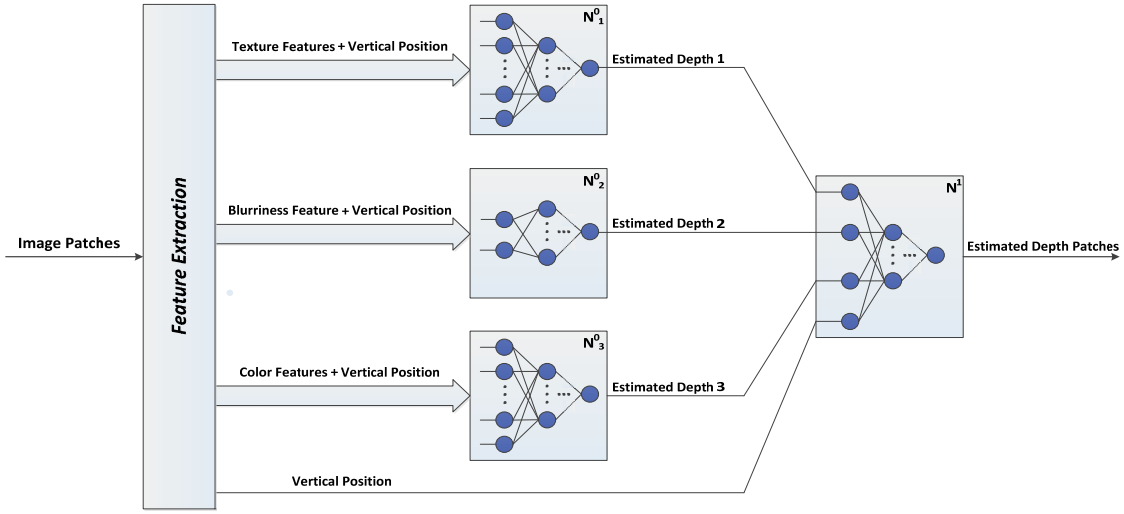

Fig. 1. Block diagram of depth estimation stage

prediction of multiple different estimators to achieve higher predictive accuracy. As can be seen in Fig.1, in the first level (level-0) of our stacked generalization scheme three neural networks are used. In order to have diverse networks, each of them is trained and tested on a set of features which describe one of the mentioned cues. The output of each network is the predicted depth of each patch. Outputs of level-0 generalizer are combined with the original input data and used as the inputs of the level-1 generalizer. Experimental results indicate that stacked generalization model improves generalization accuracy as compared to conventional single level learning schemes.

2.4. Depth refinement

Predicted depth map obtained up to this point, has local inconsistencies with query image due to the block based nature of framework. To preserve the global properties of the estimated depth and also to consider the edges of the input color image, a weighted median filter is applied [26]. Weighted median filter is a type of edge preserving smoothing filter that performs smoothing via L_1 norm minimization. Using the ability of this filter to capture strong edges from the input image, accuracy and quality of the filtered depth map are improved significantly.

3. EXMERIMENTAL RESULTS

Our proposed method has been tested on the Make3D range image dataset [27]. This is a challenging dataset for outdoor images, due to the variety of environment that is included. Make3D dataset consists of 534 images at 2272×1704 resolution and their associated depth maps captured by a laser scanner with a 55 × 305 resolution. For the sake of fair comparison with other algorithms and for computational efficiency, color images and ground-truth depths have been resized to a 345×460 resolution. We train our algorithm over 400 training images and depth estimation has been done for the 134 test images in this dataset. Each of single neural networks, in depth estimation stage, is a three layer pattern recognition network. The number of hidden layer neurons of the networks is empirically set to 40 and 30. Resilient back propagation is also used as training function in networks of each layer. In order to evaluate our proposed method quantitatively, we use two types of commonly used metrics to measure different aspects of the depth results. Three common metrics are reported to illustrate error between ground truth depth D^* and estimated depth D, defined as follow:

Relative error (Rel) = $\frac{1}{N}\sum_x \frac{|D_x - D_x^*|}{D_x^*}$

Root Mean Square Error (RMSE) = $\sqrt{\frac{1}{N}\sum_x (D_x - D_x^*)^2}$

log10 error = $\frac{1}{N}\sum_x |\log_{10}(D_x) - \log_{10}(D_x^*)|$

In addition, in order to measure the overall similarity of D^* and D, we also use normalized cross-covariance between ground-truth depth and estimated depth defined as:

$$NCC = \frac{1}{N\sigma_D \sigma_{D^*}} \sum_x (D_x - \mu_D)(D_x^* - \mu_{D^*})$$

where σ_{D^*} and σ_D are the standard deviation of D^* and D respectively, while μ_{D^*} and μ_D are the corresponding mean values and N refers to the number of pixels in image. It is worth mentioning that our research, unlike other researches, uses these two types of metrics along with each other to evaluate the accuracy of estimated depth map.

In Table 1, we compare our experimental results with those reported by different algorithms on this dataset. The results shown for [13] and [14] are obtained by running their source codes, while for other references we used the numbers reported in their papers. All of these algorithms were trained with 400 training images and tested on 134 images using the Make3D dataset. The empty entries in the table suggest that such measures were not available in the

Table1. Quantitative comparison with competing algorithms

Method	Lower is better			Higher is better	
	RMSE	log10	Rel	NCC (Average)	NCC (Median)
[15]	-	0.148	0.379	-	-
[10]	15.8	0.168	0.362	-	-
[14]	15.1	0.148	**0.361**	0.69	0.71
[13]	15.94	0.161	0.376	0.66	0.68
[12]	14	-	-	-	-
[16]	16.9	0.182	0.489	-	-
Ours	**13.46**	**0.145**	0.407	**0.74**	**0.77**

mentioned reference and their source codes were not publicly available. Some algorithms, besides using the input image, use additional information for depth perception. For instance Liu et al.'s semantic-based scheme [15] uses additional semantic labels in the conditional random field. However, our proposed method uses simple learning-based framework with no additional information. Still, we outperform many depth estimation algorithms in at least three of the four metrics. This superior achievement is mainly the result of the extraction of effective features. Correctly distributing these features into multiple neural networks enables the networks to accurately perform depth prediction. In order to evaluate the effect of each set of features, which are fed into a single network, we test it by removing one of the networks one at a time. It is observed that depth estimation errors increase by removing each network. Moreover the low error of applying multiple level model, instead of single network, confirms that adopting the mixture of features in separated networks in multiple level scheme, is beneficial for the depth estimation process.

As mentioned in Section2, with a nearest neighbor search, a selection of similar images are retrieved from the whole training set and only these selected candidates are used for feature extraction and the rest of depth estimation scheme. Numerical results indicate that applying the training phase only on images which have similar structures, rather than applying it on the whole dataset, leads to 3.2% reduction of RMSE, 13.2% reduction of log10 and 20.7% reduction of rel. But this increases the running-time. In other words this is a good evidence to prove the advantage of using global properties prior to our block based scheme. The number of similar images is determined by trial and error and is set to 50 in this work.

As further evaluation, in Fig. 2 we provide a qualitative comparison of our predicted depth maps with those recovered by different methods such as depth transfer [14] and Make3D [3] on some sample images of Make3D dataset. It can be seen in Fig.2 that, global optimization leads to over-smoothness of the predicted depth maps. It causes merging of the foreground into background. Thanks to our block-based method with appropriate block-size and also edge-aware median filter, better depth discontinuities are preserved and more detailed structures of the scene are recovered.

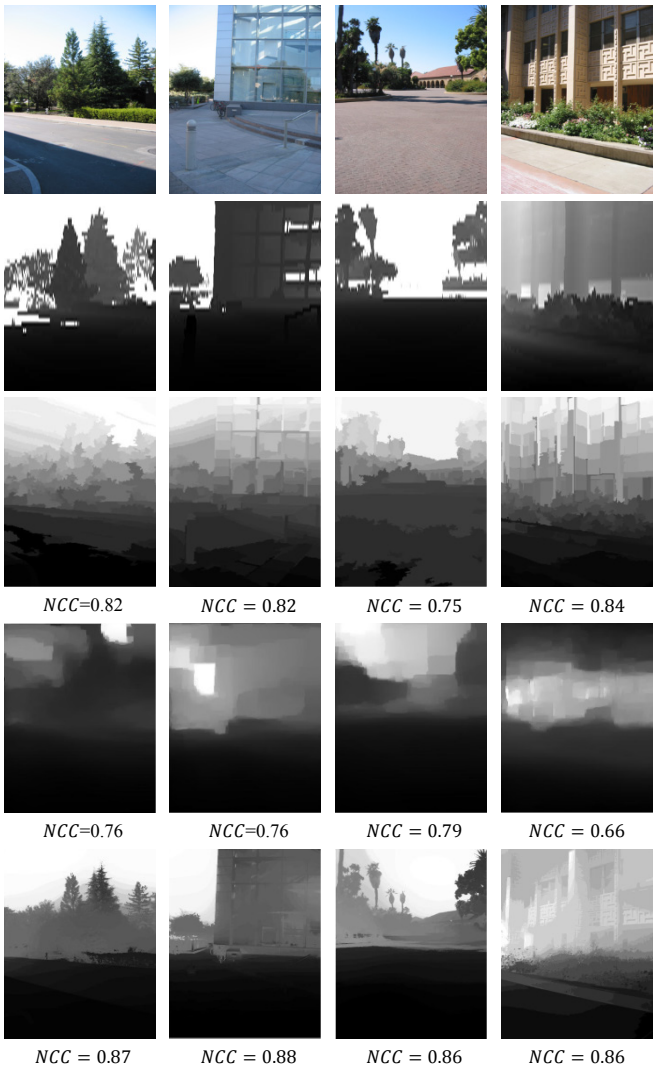

Fig. 2. Qualitative comparison with competing algorithms

4. CONCLUSION

We have devised a fully automatic method to estimate depth maps from single outdoor images. To this end we have introduced a set of local features which are highly correlated with depth. A multiple level learning method has been adopted that predicts depth of image patches by use of these reliable features. Finally, an edge aware filter is applied on the predicted depth map to diminish blockiness effect. Experiments prove that our block based framework, by using both global and local information of query image, gave better results than previous works. This is true both in terms of qualitative and quantitative accuracies.

5. ACKNOWLEDGMENT

The authors would like to thank Youngjung Kim for providing the source code of [13] used for the comparison.

5. REFERENCES

[1] J. Konrad, M. Wang, P. Ishwar, C. Wu, and D. Mukherjee, "Learning-based, automatic 2D-to-3D image and video conversion," *IEEE Transactions on Image Processing*, vol. 22, no. 9, pp. 3485 – 3496, September 2013.

[2] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," *Advances in Neural Information Processing Systems*, pp. 2366-2374, 2014.

[3] A. Saxena, M. Sun, and A.Y. Ng, "Make3D: learning 3D scene structure from a single still image," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, no. 5, pp. 824-840, 2009.

[4] R. Phan, and D. Androutsos, "Robust semi-automatic depth map generation in unconstrained images and video sequences for 2D to stereoscopic 3D conversion," *IEEE Transactions on Multimedia,* vol. 16, no. 1, pp. 122 – 136, January 2011.

[5] J. Lin, X. Ji, W. Xu, and Q. Dai, "Absolute depth estimation from a single defocused image," *IEEE Transactions on Image Processing*, vol. 22, no. 11, pp. 4545 – 4550, November 2013.

[6] D.A. Forsyth, "Shape from texture without boundaries," *Computer Vision*, Springer Berlin Heidelberg, pp. 225-239, 2002.

[7] J. Atick, P. Griffin, and N. Redlich, "Statistical approach to shape from shading: reconstruction of three-dimensional face surfaces from single two-dimensional images," *Neural Computation*, vol. 8, no. 6, pp. 1321-1340, 1996.

[8] T.Y. Kuo, Y.C. Lo, and C.C. Lin, "2D-to-3D conversion for single-view image based on camera projection model and dark channel model," *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1433-1436, 2012.

[9] A. Saxena, S.H. Chung, and A.Y. Ng, "Learning depth from single monocular images," *Advances in Neural Information Processing Systems*, pp. 1161-1168, 2005.

[10] D. Batra, and A. Saxena, "Learning the right model: efficient Max-Margin Learning in Laplacian CRFs," *IEEE Conference on Computer Vision and Pattern Recognition,* pp. 2136-2143, 2012.

[11] Y.H. Lin, W.H. Cheng, H. Miao, T.H. Ku, and Y.H. Hsieh, "Single image depth estimation from image descriptors," *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 809-812, 2012.

[12] J. Herrera, C.R. Blanco, and N. Garcfa, "Learning 3D structure from 2D images using LBP features," *IEEE International Conference on Image Processing,* pp. 2022-2025, 2014.

[13] Y. Kim, S. Choi, and K. Sohn, "Data-driven single image depth estimation using weighted median statistics," *IEEE International Conference on Image Processing,* pp. 3808-3812, 2014.

[14] K. Karsch, C. Liu, and S.B. Kang, "Depth transfer: depth extraction from video using non-parametric sampling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, no. 11, pp. 2144-2158, 2014.

[15] B. Liu, S. Gould, and D. Koller, "Single image depth estimation from predicted semantic labels," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1253-1260, 2010.

[16] X. Li, H. Qin, Y. Wang, Y. Zhang, and Q. Dai, "DEPT: depth estimation by parameter transfer for single still images," *Asian Conference on Computer Vision,* Springer International Publishing, pp. 45-58, 2014.

[17] A. Oliva, and A. Torralba, "Modeling the shape of the scene: a holistic representation of the spatial envelope," *International Journal of Computer Vision*, vol. 42, no. 3, pp. 145-175, 2001.

[18] A. Saxena, S.H. Chung, and A.Y. Ng, "3-d depth reconstruction from a single still image," *International journal of computer vision,* no. 1, pp.53-69, 2008.

[19] T. Ojala, M. Peitikaninen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," *Elsevier, Pattern recognition,* vol. 29, no. 1, pp. 51-59, 1996.

[20] S. Kang, and S. Kim, "Patch classifier of face shape outline using gray-value variance with bilinear interpolation," *Journal of Sensors,* February 2015.

[21] J. Shi, L. Xu, and J. Jia, "Just noticeable defocus blur detection and estimation," *IEEE Conference onComputer Vision and Pattern Recognition,* pp. 657-665, 2015.

[22] T. Troscianko, R. Montagnon, J.L. Clerc, E. Malbert, P.L. Chanteau, "The role of colour as a monocular depth cue," *Elsevier, Vision Research,* vol.31, no. 11, pp. 1923-1929, 1991.

[23] D. Hoiem, A.A. Efros, and M. Hebert, "Recovering surface layout from an image," *International Journal of Computer Vision,* no. 1, pp. 151-172, 2007.

[24] B. Potetz, and T.S. Lee, "Statistical correlations between two-dimensional images and three-dimensional structures in natural scenes," *Journal of the Optical Society of America*, no. 7, pp. 1292-1303, 2003.

[25] R. Ebrahimpour, M. Amini, and F. Sharifzadehi, "Farsi handwritten recognition using combining neural networks based on stacked generalization," *International Journal on Electrical Engineering and Informatics*, vol. 3, 2011.

[26] Z. Ma, K. He, Y. Wei, J. Sun, and E. Wu, "Constant time weighted median filtering for stereo matching and beyond," *IEEE International Conference on Computer Vision,* pp. 49-56, 2013.

[27] http://make3d.cs.cornell.edu/data.html