2016 23rd International Conference on Pattern Recognition (ICPR)
Cancún Center, Cancún, México, December 4-8, 2016

Single Image Depth Estimation Using Joint Local-Global Features

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

[1]Department of Electrical and Computer Engineering, Isfahan University of Technology, 84156-83111 Iran
[2]Department of Emergency Medicine, University of Michigan, Ann Arbor, 48109 USA
[3]Michigan Center for Integrative Research in Critical Care, University of Michigan, Ann Arbor, 48109 USA
[4]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, 48109 USA

Abstract— **Inferring scene depth from a single monocular image is an essential component in several computer vision applications such as 3D modeling and robotics. This process is an ill-posed problem. To tackle this challenging problem, previous efforts have been focusing on exploiting only global or local depth aware properties. We propose a model that incorporates both of them to obtain significantly more accurate depth estimates than using either global or local properties alone. Specifically, we formulate single image depth estimation as a *K* nearest neighbor search problem at both image level and patch level. At each level, a set of rich depth aware features, describing monocular depth cues, is employed in a nearest-neighbor regression model. By comparing the results with and without patch based fusion, the importance of our joint local-global framework becomes clear. The experimental results also demonstrate superior performance compared with existing data-driven approaches in both quantitative and qualitative analyses with a significantly simpler algorithm than others.**

Keywords—Depth estimation, Monocular depth cues, Joint local-global framework, KNN regression model, Data-driven approaches.

I. INTRODUCTION

Understanding scene depth has a variety of applications in computer vision disciplines, including 3D model reconstruction, recognition, robot navigation and surveillance. For example, there are depth-based fall detection systems which are used for assisting elderlies and people with disabilities [1]. In addition, scene depth estimation plays a fundamental role in the process of 2D to 3D image/video conversion through depth image based rendering (DIBR) procedure. DIBR is the process of generating 3D content from a single image and its associated depth map [2].

Recently, emergence of 3D technology in various fields ranging from entertainment to medical imaging, enrich the user's viewing experience by creating the illusion of depth. Particularly in medical cases, lesser-invasive surgeries are enabled by providing surgeons 3D vision. On the other hand, getting 3D content which requires extra cameras is not as common as monocular imaging; this is the case especially where hardware size is an important issue such as in endoscopy in medical imaging [3]. To close this gap between 3D displays and 3D contents, many 2D to 3D image/video conversion algorithms have been proposed in the last few years. Single image depth estimation as the main step of this procedure is a technically ill-posed problem, due to the lack of reliable depth cues such as motion or stereo correspondence in a monocular viewpoint. Humans, however, are able to perceive scene depth easily in monocular situations, thanks to the prior knowledge they learned over the years. This observation has motivated many researchers to simulate human visual system (HVS) behavior in depth perception using monocular depth cues exploited from a 3D repository. As a pioneer work in this regard, Saxena *et al.* proposed the Make3D algorithm [4] in which both monocular depth cues and the relation between different parts of the image are modeled in a Markov random filed (MRF). They also presented an MRF model to capture 3D position and orientation of super pixels in an image [5]. Make3D algorithm was further improved by Batra and Saxena in [6]. They proposed max-margin parameter learning in conditional random fields (CRFs) with Laplacian potentials. In [7], Liu *et al.* integrated semantic labels with monocular depth cues to improve the 3D reconstruction process. Hoiem *et al.* constructed the surface layout of a scene by labeling of the images into geometric classes [8]. More recently, the algorithm in [9] transfers depth gradient as reconstruction cues, instead of directly selecting depth values from the training data.

Apart from learning-based methods for single image depth estimation, various conventional algorithms have been devised which are typically based on image content. These methods predict depth values by directly making use of monocular depth cues such as atmospheric effects [10], focus/defocus [11] and occlusions [12]. However, the main drawback of such approaches is that they usually tend to impose some strict assumption on the image content. Hence their application typically is limited up to some restricted scenes, such as images containing haze, defocus due to the limited depth of field (DOF) and etc. But practically, most of the time, real images do not provide such conditions. This is why data-driven approaches have come under the focus in recent years.

Some data-driven approaches are based on a reasonable assumption about existence of correlation between image visual appearance and depth values. The core idea of such methods is that scenes with similar appearance are expected to have similar depth values and hence, retrieving similar images to the input image from 3D repository is the mutual step between such algorithms. A set of high level image features such as histogram of oriented gradients (HOG) [13], local binary pattern (LBP) [14], GIST [15] or a combination of them is typically used in the retrieving process. Konrad *et al.* proposed a fast and simple way to fuse associated depths of *K*

candidate images [2]. They applied median operator on K depth maps to provide an initial estimate of a depth map. Herrera *et al.* replaced simple median operator with weighted averaging in which depth maps are weighted according to image's similarity [16, 17]. Weighted median statistics is another fusion strategy which has been devised in this regard [18]. Depth transfer algorithm devised by Karsch *et al*. warps candidate depths, based on SIFT (scale-invariant feature transform) flow [19]. This procedure is then followed by a global optimization that encourages smoothness across the estimated depth. Depth estimation by parameter transfer (DEPT) algorithm was proposed in [20] to estimate realistic depth map by modeling the correlation between images and their depth information using parameter transfer. The way in which associated depths of candidate images are fused is an important issue in such data-driven methods. In this paper we argue that having only a global perspective in fusion phase is insufficient to predict depth maps that are both visually pleasing as well as quantitatively accurate. Hence, we propose to capture both global and local information from various cues by exploiting depth relevant features from the entire image and image patches respectively. More specifically, we estimate depth value of an input image in a patch-based framework using a nearest neighbor regression type idea in both image level and patch level.

The rest of this paper is organized as follows. Section II presents the proposed depth estimation algorithm. The performance of our proposed method is evaluated in Section III and finally Section IV concludes the paper.

II. Proposed Method

As discussed in the previous section, we aim to use the merits of both global and local information of a scene for estimating its depth map. To this end as depicted in Fig. 1, given an input image and a database, we first perform a nearest neighbor search in *Similar Image Retrieval* stage to retrieve K images most similar to the input image, from 3D database. In Fig.1 we call a database containing image/depth pairs as RGBD. The K candidate images along with their corresponding depth maps act as our new 3D training set. Afterward, instead of globally fusing K candidate depths [2,16,17], we proceed by adopting a patch based framework to consider local aspects of images in addition to their global structure (to have both local and global perspective). Specifically we collect non-overlapping patches of size 16×16 from both candidate images and the input image in *Image Patch Formation* stage. Next, for each patch of the input image, a set of similar patches is retrieved from all the patches of our new 3D training set in *Similar Patch Retrieval* stage. The corresponding depth patch is estimated for each patch of the input image, by fusing candidate depth patches through *Depth Patch Fusion* stage. Then in *Stitching* stage, all the estimated depth patches are stitched together to form the overall initial depth map. Finally, we refine the initial depth map from patch level to pixel level to reconstruct the final depth map in *Depth Refinement* stage.

In the following, four main stages of our algorithm are discussed in detail.

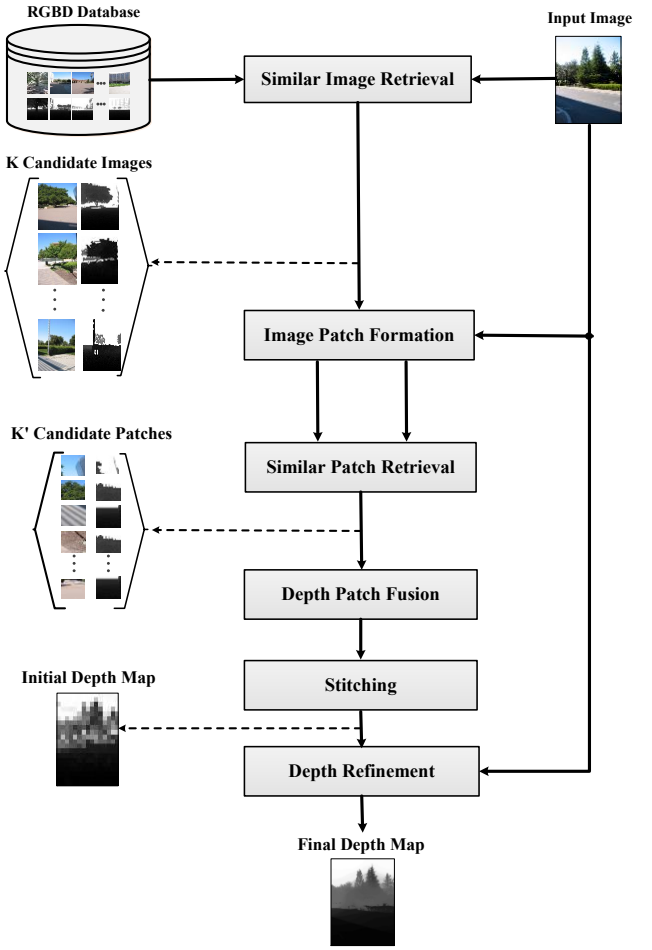

Fig. 1. Block diagram of our depth estimation model

A. Similar Image Retrieval

It has been known that in a large scale 3D dataset, there exist two kinds of images, those that have similar structure to the input image and therefore are relevant for estimation of its depth and those that are irrelevant. The latter case should be rejected from 3D repository to reduce or even remove the influence of potential outliers in the training set. GIST [15], as a set of high level image features is used to characterize the global structure of the images. The structure similarity between the input image (I) and n^{th} color image (I_n) in the training set is measured by computing the sum of squared differences (SSD) of the corresponding image feature descriptor as follows:

$$SSD = \|G(I) - G(I_n)\|^2 \quad (1)$$

where $G(X)$ is the GIST feature vector of image X. We then discard all image/depth pairs from dataset but the top K nearest neighbors with respect to SSD and consider them as the new training pairs, which are relevant for learning depth. In Fig. 2, we show four nearest neighbor search results for two outdoor input images from Make3D dataset [21] (see Section III), retrieved by comparing the SSDs of GIST features with each other. Despite the fact that none of the four candidates

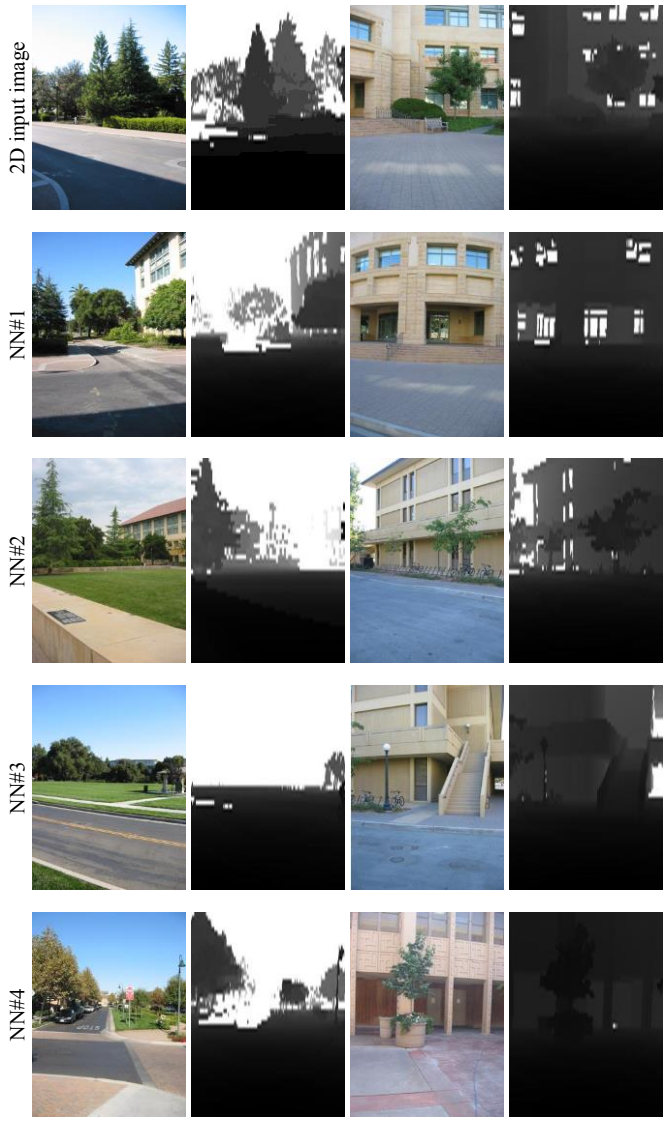

Fig. 2. Color image and depth field of two 2D input (first row) and their four nearest neighbors (rows 2-5).

completely matches the corresponding input image, the overall underlying depth is somewhat correlated to the associated ground truth depth.

B. Similar Patch Retrieval

Relying only on global properties of images is insufficient to predict absolute depth values of such images accurately. On the other hand, since the K retrieved images are not aligned semantically with the input image, applying global fusion strategies on the associated depth maps results in over-smoothing. To tackle the above shortcoming, considering image patches, we propose to find K' patches in the new 3D dataset (obtained in the first stage) that have most similar depth to the input patch. The KNN search in this stage is based on effective local features applied to capture two monocular cues typically used to perceive depth by humans: texture and relative height.

1) Texture

Texture variation is a prominent monocular cue mostly used in human visual system (HVS) to perceive depth since object's texture looks different depending on the distance from the viewer [4]. To capture this significant cue, histogram of local binary pattern (LBP) is computed per patch. LBP provides a robust means of describing patterns in a texture which has been widely used in various pattern recognition applications [14]. Entropy is another texture feature we used along with LBP to capture texture cue more effectively.

2) Relative height

In addition to texture, relative height provides a cue for depth perception in a sense that objects with higher vertical position in the image, appear to be more distant. To show the correlation between depth value of a point and its relative height, we conducted an experiment by computing the average value of all depth images in Make 3D dataset, depicted in Fig. 3. As can be inferred from the average depth map in Fig. 3, the bottom of the image corresponds to the regions that are closer to the camera; specifically the distance of a point from the viewer usually increases as its height in the image increases.

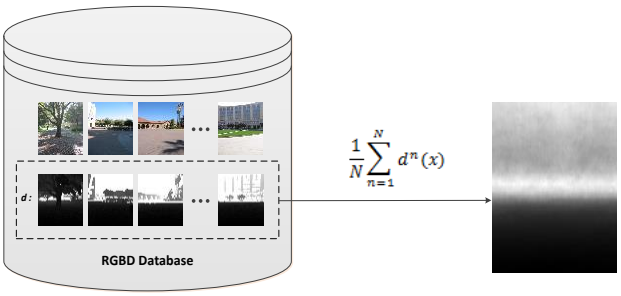

Fig. 3. The relation of depth values and relative height

At the end of this stage, we attempt to demonstrate that above-mentioned features are of critical importance in obtaining good performance in depth estimation. In other words, we illustrate the fact that image patches with similar features have roughly similar depth values. To this end we compute the correlation between our proposed feature vectors extracted from image patches. The root mean square errors (RMSEs) of associated depth patches are also calculated. Furthermore, the probability that the RMSE of depth patches being equal to zero, given various correlation values between feature vectors, is computed and reported in Table I. The results in this table show that the higher the correlation of image feature vectors, the higher the histogram peak at zero which are good evidence that the employed features are highly correlated with depth values.

TABLE I. THE RELATIONSHIP BETWEEN IMAGE FEATURE VECTORS AND DEPTH VALUES

Correlation	$\mathbf{P(RMSE = 0)}$
[0,0.2]	2.35×10^{-4}
(0.2,0.4]	3.66×10^{-4}
(0.4,0.6]	5.71×10^{-4}
(0.6,0.8]	0.001
(0.8,1]	0.021

C. Depth Patch Fusion

The selected depth patches D_k , $1 \le k \le K'$ obtained in the previous stage, should be combined to get the depth of input patch. Two fusion approaches were considered in this paper to reduce the influence of potential outliers: the median and weighted average operator on the K' candidate depth patches. In the former case, we compute the depth values of each patch (at pixel level) by applying median operator across the selected depth patches at each spatial location x according to (2):

$$D(x) = \text{median}\{D_k(x) \mid 1 \le k \le K'\} \tag{2}$$

Depth patch combination in the second strategy is inspired by the assumption that the patches with analogous image structure are expected to have similar depth distribution. Accordingly, the contribution of each selected depth patch in our final depth estimate is proportional to the similarity of the input patch with associated color patch. Specifically, the weighted averaging process across K' candidate depth patches can be defined as:

$$D = \sum_{k=1}^{K'} w_{\mathrm{k}} D_{\mathrm{k}} \tag{3}$$

where D_k is the average value of k^{th} candidate depth patch and w_k is a weight function representing the contribution degree of that patch. Denoting $F(P)$ and $F(P_k)$ as the feature vectors for input patch and k^{th}candidate patch (employed for matching patches), we define w_k as:

$$w_{\mathrm{k}} = \frac{1}{\|F(P) - F(P_k)\|^2} \tag{4}$$

Eventually, as the result of (3), D, a preliminary estimate of input patch, is obtained.

D. Depth Refinement

The initial depth map obtained up to this point is locally inconsistent with the color input image due to the patch-based nature of our proposed framework. As a result, this often leads to the lack of edges in estimated depth where sharp boundaries should occur and the lack of depth smoothness in some homogenous regions. Here, we refine it through a simple but effective post-processing, based on image-guided joint filtering. In particular, we employ weighted median filter (WMF), an edge-preserving smoothing filter proposed in [22], which accomplishes smoothing via L_1 norm minimization. As a result, applying this stage makes the initial depth map to be smoothed out, while keeping its edges sharp and aligned with those of the input image.

III. Experimental Results

In this section, we evaluate our single image depth estimation model on a popular dataset, which is available online: Make3D range image dataset [21]. The make3D dataset consists of 534 image/depth pairs depicting outdoor scenes with corresponding depth maps collected with laser scanner. The images are captured from various environments with different structures which makes Make3D dataset a challenging one for outdoor scenes. The color images and depth maps are of 2272×1704 and 55×305 resolution, respectively. Nevertheless, we resize them to 345×460 pixels for computational efficiency and preserving the aspect ratio of original images. The whole dataset was divided into 400 training images and 134 test images. We examined our proposed method on the images of the test set and used the training set to perform our K nearest neighbor search. It should be noted that all the experimental results are achieved by setting constant parameters K=20 and K'=10 empirically.

For quantitative evaluation, we report error for three common error metrics which have been used extensively in previous works. We compute RMS error, relative error (Rel) and error in log10 scale as follows:

$$\text{RMS} = \sqrt{\frac{1}{N}\sum_{x}(D(x) - D^*(x))^2}$$

$$\text{Rel} = \frac{1}{N}\sum_{x}\frac{|D(x) - D^*(x)|}{D^*(x)}$$

$$\text{log10} = \frac{1}{N}\sum_{x}|\log_{10}(D(x)) - \log_{10}(D^*(x))|$$

Here D^* and D are the ground-truth and estimated depth respectively and N is the total number of pixels in all the images. In addition, normalized cross covariance (NCC) which measures similarity between the estimated depth map and the ground-truth depth, is also used. Denoting μ_D and μ_{D^*} as the mean value of D and $D^*$and σ_D and σ_{D^*} as the corresponding standard deviations, NCC is defined as follows:

$$NCC = \frac{1}{N\sigma_D\sigma_{D^*}}\sum_{x}(D(x) - \mu_D)(D^*(x) - \mu_{D^*}) \tag{5}$$

We first present baseline comparisons to prove the superiority of jointly considering the global and local structures of images, compared with only reasoning globally. In order to compare our core idea for candidate depth fusion, with those of [2, 16, 17], all the stages, except the way of fusing K candidate depths in two experiments, were considered the same. We applied median (Med) and weighted average (WA) operator on the whole depth candidates, globally, to simulate the fusion idea of [2] and [16, 17] respectively. The performance of algorithms has been captured by the average (Avg), and median of NCC, along with the average of three other error metrics across all 134 test images. Table II shows the numerical results. As can be observed, our method achieves better performance by jointly using global and local properties of images in a unified framework and a significant drop in the performance is visible when our patch-based framework is neglected.

TABLE II. BASELINE COMPARISONS ON MAKE3D DATASET (BEST RESULTS ARE BOLDED).

Method		Lower is better			Higher is better	
		RMS	log10	Rel	NCC (Avg)	NCC (Med)
Global	*WA*	15.86	0.186	0.579	0.66	0.66
	Med	16.86	0.167	0.387	0.65	0.67
Global + Local	*WA*	**15.08**	**0.163**	0.447	**0.70**	**0.72**
	Med	16.58	**0.163**	**0.351**	0.66	0.67

We also compare our method with several popular state-of-the-arts methods in Table III. All methods were trained on 400 image/depth pairs and were tested on 134 images of the same dataset. We got the results of [18] and [19] by running the authors' source codes, which are publicly available[1] and the numerical results of other references are directly taken from their tables. Dash signs indicate that those numbers were not reported in the mentioned reference. Most of the works in the literature compare their results with some of the mentioned metrics. We quantitatively evaluate our approach by NCC along with three other error metrics.

TABLE III. QUANTITATIVE COMPARISON WITH COMPETING ALGORITHMS

Method	Lower is better			Higher is better	
	RMS	log10	Rel	NCC (Avg)	NCC (Med)
[4]	16.7	0.198	0.530	-	-
[5]	-	0.149	0.458	0.64	0.69
[6]	15.8	0.168	0.362	-	-
[7]	-	0.149	0.375	-	-
[8]	-	0.320	1.423	-	-
[18]	15.9	0.161	0.376	0.66	0.68
[19]	15.1	**0.148**	0.361	0.69	0.71
[20]	16.9	0.182	0.489	-	-
Ours (WA)	**15.08**	0.163	0.447	**0.70**	**0.72**
Ours (Med)	16.58	0.163	**0.351**	0.66	0.67

Despite the fact that most of these algorithms employ computationally expensive optimization [19], sophisticated graphical model [4,5,6] or even the additional knowledge of pixel labels during training phase [7,8], we outperform them in three of the four metrics. Most notably, our method (using median operator) obtains significantly better results in terms of relative error (Rel). This improvement of the results is attributed to the integration of both global and local information and extraction of robust and effective depth aware features in a patch-based framework.

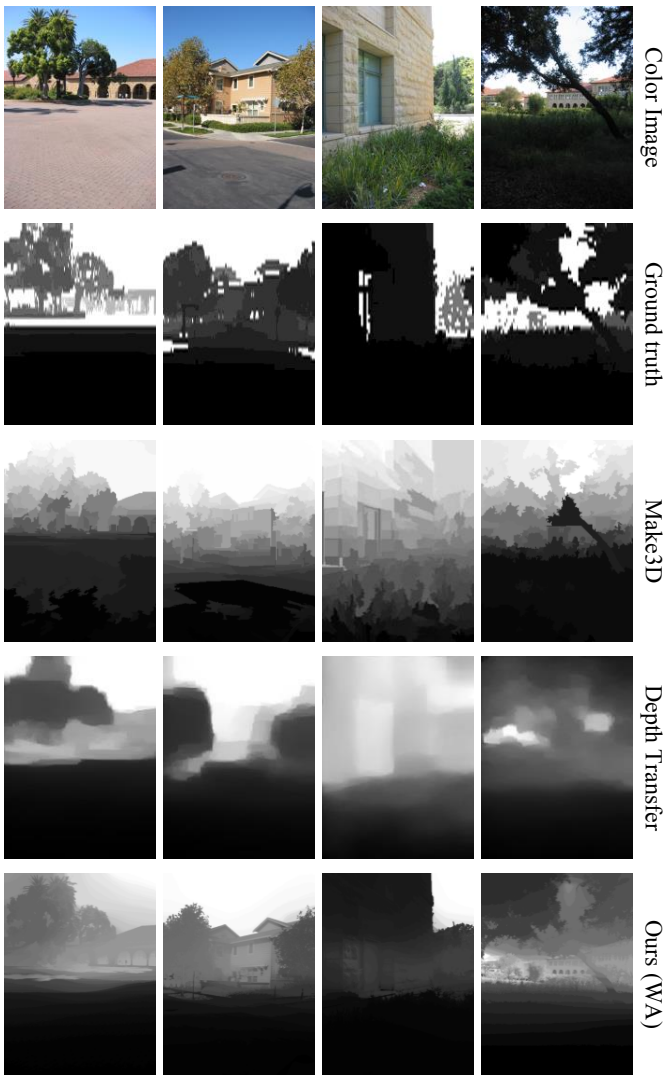

Fig. 4. Qualitative comparison with other approaches

We further present a qualitative comparison between our estimated depth maps with those recovered by *depth transfer* [19] and Make3D [5] algorithms on representative images from Make3D data set. Qualitative results are given in Fig. 4. It can be observed that Make3D algorithm fails to capture the complicated structures in some cases and do not recover a visually pleasing depth map. Moreover, results of *depth transfer* algorithm tend to be over-smoothed due to the global optimization employed in depth interpolation process. In contrast, depth boundaries in our results are better aligned to those of the input color image. Clearly, our method leads to more visually pleasant estimates with sharper transitions, demonstrating the superiority of applying patch-based framework followed by proper depth refinement algorithm.

[1] "http://www.kevinkarsch.com/" and "http://make3d.cs.cornell.edu/code.html"

IV. Conclusion

Estimating scene depth from a single monocular image is an inherently ambiguous task, requiring combination of both local and global information of the image. In this regard, we have presented a fully automatic technique using the merits of local features in addition to contextual information of the scene in a unified framework. Practically, we have formulated depth estimation as a nearest neighbor regression at two levels. After pruning the large scale 3D dataset with high level global features (KNN at image level), a set of depth related features which captures prominent monocular depth cues, is locally exploited from image patches for KNN procedure at patch level. The experiments showed that achieved results are superior compared with competing algorithms in both quantitative and qualitative comparisons.

References

[1] B. Kwolek and M. Kepski, "Improving fall detection by the use of depth sensor and accelerometer," Neuro computing, Elsevier, pp. 637-645, 2015.

[2] 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.

[3] Real-time 2d-3d image converting software with a monocular small camera toward the lesser invasive laparoscopic surgery - SAGES Abstract Archives, http://www.sages.org/

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

[5] 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(PAMI), no. 5, pp. 824-840, 2009.

[6] D. Batra, and A. Saxena, "Learning the right model: efficient max-margin learning in Laplacian CRFs," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2136-2143, 2012.

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

[8] D. Hoiem, A. Efros, and M. Hebert, "Recovering surface layout from an image", International Journal of Computer Vision (IJCV), vol. 75, no. 1, pp. 151-72, 2007.

[9] S. Choi, D. Min, B, Ham, Y. Kim, C. Oh and K. Sohn, "Depth analogy: data-driven approach for Single image depth estimation using gradient samples," IEEE Transactions on Image Processing, vol. 24, no. 12, pp. 5953-5966, December 2015.

[10] 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 (ICASSP), pp. 1433-1436, 2012.

[11] 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, 2013.

[12] G. Palou and P. Salembier, "Monocular depth ordering using t-junctions and convexity occlusion cues," IEEE Transactions on Image Processing, vol. 22, no. 5, pp. 1926-1939, 2013.

[13] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in Proc. IEEE Conference on Computer Vision and Pattern Recognition(CVPR), pp. 886-893, 2005.

[14] T. Ojala, M. Pietikainen, and T. Maenpaa "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), vol. 24, pp. 971-987, 2002.

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

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

[17] J. Herrera, C.R. Blanco, and N. Garcfa, "Fast 2D to 3D conversion using a clustering-based hierarchical search in a machine learning framework," 3DTV-Conference: The True Vision-Capture, Transmission and Display of 3D Video, pp. 1-4, 2014.

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

[19] 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 (PAMI), no. 11, pp. 2144-2158, 2014.

[20] 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 (ACCV), Springer International Publishing, pp. 45-58, 2014.

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

[22] 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 (ICCV), pp. 49-56, 2013.