Search-Based Depth Estimation via Coupled Dictionary Learning with Large-Margin Structure Inference

Yan Zhang[1], Rongrong Ji[2], Xiaopeng Fan[1](✉), Yan Wang[3], Feng Guo[2],
Yue Gao[4], and Debin Zhao[1]

[1] School of Computer Science and Technology,
Harbin Institute of Technology, Harbin, China
{y.zhang,fxp,dbzhao}@hit.edu.cn

[2] School of Information Science and Engineering,
Xiamen University, Xiamen, China
{rrji,betop}@xmu.edu.cn

[3] Microsoft, Redmond, USA
wanyan@microsoft.com

[4] School of Software, Tsinghua University, Beijing, China
gaoyue@tsinghua.edu.cn

Abstract. Depth estimation from a single image is an emerging topic in computer vision and beyond. To this end, the existing works typically train a depth regressor from visual appearance. However, the state-of-the-art performance of these schemes is still far from satisfactory, mainly because of the over-fitting and under-fitting problems in regressor training. In this paper, we offer a different data-driven paradigm of estimating depth from a single image, which formulates depth estimation from a search-based perspective. In particular, we handle the depth estimation of local patches via a novel cross-modality retrieval scheme, which searches for the 3D patches with similar structure/appearance to the 2D query from a dataset with 2D-3D mappings. To that effect, a coupled dictionary learning formulation is proposed to link the 2D query with the 3D patches, on the reconstruction coefficients to capture the cross-modality similarity, to obtain a rough depth estimation locally. In addition, consistency on spatial context is further introduced to refine the local depth estimation using a Conditional Random Field. We demonstrate the efficacy of the proposed method by comparing it with the state-of-the-art approaches on popular public datasets such as Make3D and NYUv2, upon which significant performance gains are reported.

Keywords: Single image depth estimation · Cross-modality retrieval · Coupled dictionary learning · Contextual refinement

1 Introduction

Depth estimation from a single monocular image [24] is a fundamental problem in computer vision, with various applications in stereo vision, robotics, and

B. Leibe et al. (Eds.): ECCV 2016, Part V, LNCS 9909, pp. 858–874, 2016.
DOI: 10.1007/978-3-319-46454-1_52

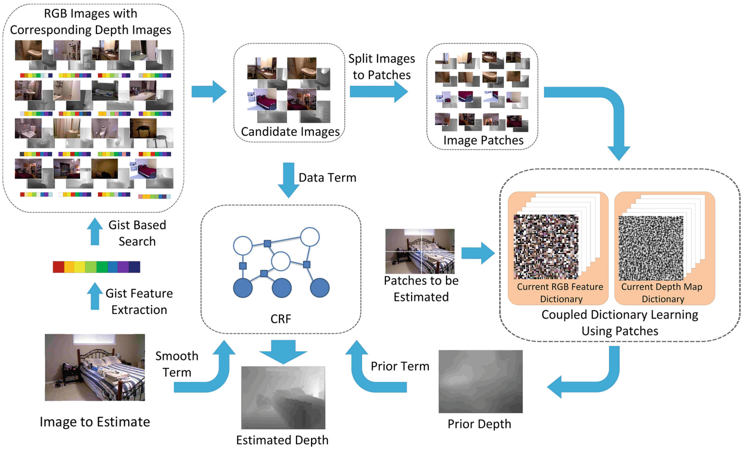

Fig. 1. The framework of our method.

scene understanding [17,26]. In a typical setting most approaches [1,9,17,25] use a standard regression or classification pipeline to predict the depth fitting, orientation and plane fitting. Such pipeline consists of the calculation of dense or sparse features, followed by an appearance feature representation and regressor training. The responses of a classifier or a regressor are combined in a probabilistic framework, and under very strong geometric priors the most probable scene layout is estimated. Despite promising progress achieved, these methods are still far from practical applications, with the conflict between the model capability and the data scalability, resulting in over-fitting or under-fitting for such learning-based paradigm.

Coming with the proliferation of 3D sensing devices *e.g.* Kinect and matured 3D modeling techniques *e.g.* Structure-from-Motion [5,7] and visual SLAM [21,30], massive-scale 3D data such as point clouds and depth maps are available nowadays, which can provide rich correspondences between 2D visual appearance and 3D depth structures. Therefore, is it possible to take advantage of such rich 2D-3D correspondences towards a search-based paradigm in depth estimation? In this work, we tackle the depth estimation from a different perspective with traditional methods [11–13,16,17,25,26]. In general, we adopt a search-based paradigm that leverages a dictionary-based cross-modality retrieval to robustly and efficiently find best-matches 3D depth given a 2D query patch as the local depth estimator. It is followed by a Taylor formula based contextual refinement to achieve a consistent yet accurate global depth estimation at the image-level.

In particular, unlike the traditional approaches [25,26] that learn a regressor from image to depth indirectly, we first perform joint dictionary learning to bridge the similarity gap between 2D image patches and 3D depth maps to facilitate cross-modal retrieval. Then, given an image patch, we search for the corresponding 3D patches from a large reference set with the depth information between 2D and 3D local patches. This approach provides key advantages in both online efficiency and generalization ability.

The above patch-wised local depth estimation is further integrated with spatial contextual constraints using Conditional Random Field (CRF), as was commonly adopted in existing works [6,11,17,26]. To evaluate the performance of the proposed method, we conduct experiments on the widely-used Make3D and NYUv2 datasets. We compare the proposed method with several existing state-of-the-art ones, including make3D [26], Semantic Labelling [17] and Depth Transfer [11]. We report significant performance gain to demonstrate the advantages of the proposed model. The main contributions of our work are three-fold:

- We propose a novel cross-modality retrieval paradigm that does not rely on training depth regressors to tackle over- and under-fitting issues previously existed;
- A novel coupled dictionary learning is introduced to bridge the similar gap between 2D query and 3D references, with detailed analytical solutions for fast yet accurate parameter learning;
- Adopting contextual refinement with Taylor expansion and CRF inference, which also improves the generalization capability. Compared with traditional methods [11,26], the proposed inference does not require parameter fitting on the training set.

2 Related Work

Previous works [4,26] in depth estimation from a single monocular image typically follow a regression setting. In this setting, the image is first over-segmented into superpixels, and then a pre-trained local depth regressor is applied on each individual superpixel to estimate the corresponding local depth. Subsequently, the Markov Random Field (MRF) or Conditional Random Field (CRF) [4] is frequently employed to impose spatial constraints on the estimated local depth. Such contextual cues usually include 3D location and orientation of the patch [26], as well as the global context [25] among patches. For instance in [17], Liu *et al.* partitioned depth estimation into two phrases, *i.e.*, semantic segmentation [27] and 3D reconstruction, with the semantic labels guiding the 3D reconstruction. In [20], Liu *et al.* modeled depth estimation as a discrete-continuous optimization problem, where the continuous variables encode the depth of superpixels in the input image, and the discrete ones represent relationships between neighboring superpixels. Karsch *et al.* [10,11] inferred the depth map by three stages: candidate images discovery, point-wise alignment and optimization procedure. More recently, deep learning [6] was introduced for single image depth estimation. For instance, Liu *et al.* [19] combined the Convolutional Neural Network (CNN) and the CRF model for depth estimation, where the CNN learns the geometric priors and the CRF model could further optimize the depth among adjacent superpixels. Similar to [19], the method in [15] also extracted deep CNN features for depth regression, which was combined with the CRF-based post processing. The limitation of the state-of-the-art methods for single image depth estimation is closely tied to the property of perspective geometry, which becomes a bottleneck for the current RGB-D based methods. In contrast, this limitation severely affects methods based on 3D model, since the 3D model can offer all stereo perspectives, which provides a new aspect to conquer this limitation.

Cross-modality retrieval has also attracted vast research focuses in recent years. In [34], Wang *et al.* built a cross-modality probabilistic graphical model to discover mutually consistent semantic information among different modalities. In [23], cross-modal correlations and semantic abstraction were employed to jointly model the text and image components. Zhuang *et al.* [37] proposed a SliM 2 model to formulate the multimodal mapping as a constrained dictionary learning problem, where the label information [3] is employed to discover the shared intra-modality structure. More recently, deep learning methods were further employed in cross-modality retrieval [31,34] for the tasks of text-to-image and image-to-text search. The main disadvantage of the existing cross-modality retrieval methods is that they can not learn the structure information from images and 3D models. However, we can rebuild the structure information of image patches by sharing the reconstruction coefficients with coupled dictionary learning.

Recent works have shown the effectiveness of coupled dictionary learning in exploring inherent correlations between two data channels. Here, we introduce the most relevant works to ours. Wang *et al.* [32] proposed a semi-coupled dictionary learning (SCDL) method to conduct cross-style image synthesis. Yang *et al.* [36] employed neural network to jointly learn dictionaries of different resolutions. To tackle the deblurring problem, Xiang *et al.* [35] trained dictionaries on both clean and blurred images jointly, and Wang *et al.* [33] learned a dictionary on deblurred intermediate results and blurred images jointly. Shekhar *et al.* [28] established the identity of multi-source information by joint sparse representation. And He *et al.* [8] jointly learned overcomplete dictionaries for one single super-resolution image. Note that the above methods suppose that both dictionaries are learned upon data with the same modality, which is very challenging to capture the cross-modality similarity using learned existing coupled dictionary learning methods.

3 Cross-Modality Retrieval for Local Depth Estimation

The first step is to infer local depth from a single image. To this end, we first train dictionaries from different modalities (*i.e.*, 2D image patches and 3D models) synchronously and then conduct cross-modality retrieval for each target patch. Based on the retrieval results, we estimate the depth from the most correlated 3D model directly. Section 3 presents the details of the above process.[1]

3.1 Coupled Dictionary Learning

Our basic assumption is that if an object can be decomposed into a set of 3D objects, its 2D projection should be able to decomposed in the same way, and

[1] Contextual refinement will be further introduced in Sect. 4.

vise versa. Therefore, given a set of 2D patches[2] $\boldsymbol{x}_{\text{im}}^{j}\,(j=1,\cdots,n)$ and the corresponding 3D model $\boldsymbol{x}_{\text{depth}}^{j}\,(j=1,\cdots,n)$, from a dictionary learning perspective, we aim to obtain a pair of *codes* y_{im}^{j} for x_{im}^{j} and y_{depth}^{j} for x_{depth}^{j} based on two dictionaries $\boldsymbol{D}_{\text{im}}$ and $\boldsymbol{D}_{\text{depth}}$. And these two codes are supposed to be similar after the proper projection. This intuition leads to the following formulation:

$$\begin{aligned}\min_{\boldsymbol{D}_{\text{im}},\boldsymbol{D}_{\text{dep}}} & \sum_{j}\left\|\boldsymbol{x}_{\text{im}}^{j}-\boldsymbol{D}_{\text{im}}\cdot\boldsymbol{y}_{\text{im}}^{j}\right\|_2^2+\alpha\left\|\boldsymbol{x}_{\text{dep}}^{j}-\boldsymbol{D}_{\text{dep}}\cdot\boldsymbol{y}_{\text{dep}}^{j}\right\|_2^2\\ & +\beta\left\|\boldsymbol{y}_{\text{dep}}^{j}-\boldsymbol{R}\cdot\boldsymbol{y}_{\text{im}}^{j}\right\|_2^2\\ s.t.\quad & \boldsymbol{R}^T\cdot\boldsymbol{R}=\boldsymbol{I},\end{aligned} \tag{1}$$

where $\boldsymbol{y}_{\text{dep}}^{j}$ and $\boldsymbol{y}_{\text{im}}^{j}$ are the reconstruction coefficients. $\boldsymbol{D}_{\text{im}}=[\boldsymbol{d}_{\text{im}}^{1},\boldsymbol{d}_{\text{im}}^{2},\cdots,\boldsymbol{d}_{\text{im}}^{c}]\in\Re^{p\times c}$ is the dictionary 2D patches, while $\boldsymbol{D}_{\text{dep}}=\left[\boldsymbol{d}_{\text{dep}}^{1},\boldsymbol{d}_{\text{dep}}^{2},\cdots,\boldsymbol{d}_{\text{dep}}^{c}\right]\in\Re^{q\times c}$ is the dictionary of 3D models.. The first term in Eq. 1 is the reconstruction error between the 2D image patches $\boldsymbol{x}_{\text{im}}^{j}$ and their corresponding representation results. The second term is 3D reconstruction error. And the third term is the projection error of coefficient $\boldsymbol{y}_{\text{dep}}^{j}$ and $\boldsymbol{y}_{\text{im}}^{j}$. Through the projection matrix $\boldsymbol{R}$, 3D and 2D coefficients are connected to enable cross-modality similarity matching.

However, it is not exactly proper to force projection matrix $\boldsymbol{R}$ to be orthogonal. Although orthogonality can guarantee $\boldsymbol{R}$ to be full rank and make the coefficient spaces equivalent, such strict constraint may leads to a suboptimal result. We therefore relax the constraint and merge the second and third terms in Eq. 2 which is equivalent to Eq. 1, but less restrict.

$$\min_{\boldsymbol{D}_{\text{im}},\boldsymbol{D}_{\text{dep}}}\left\|\boldsymbol{X}_{\text{im}}-\boldsymbol{D}_{\text{im}}\boldsymbol{Y}\right\|_F^2+\alpha\left\|\boldsymbol{X}_{\text{dep}}-\boldsymbol{D}_{\text{dep}}\boldsymbol{Y}\right\|_F^2. \tag{2}$$

where $\boldsymbol{Y}=\left[\boldsymbol{y}^1,\boldsymbol{y}^2,\cdots,\boldsymbol{y}^n\right]$ is the coefficient matrix and $\boldsymbol{X}_{\text{im}}=\{\boldsymbol{x}_{\text{im}}^{1},\boldsymbol{x}_{\text{im}}^{2},\cdots,\boldsymbol{x}_{\text{im}}^{n}\}$, $\boldsymbol{x}_{\text{im}}^{i}\in\Re^{p\times 1}$ is a set of n RGB image patches, whose corresponding depth image patches[3] are $\boldsymbol{X}_{\text{dep}}=\{\boldsymbol{x}_{\text{dep}}^{1},\boldsymbol{x}_{\text{dep}}^{2},\cdots,\boldsymbol{x}_{\text{dep}}^{n}\}$, $\boldsymbol{x}_{\text{dep}}^{i}\in\Re^{q\times 1}$.

To solve Eq. 2, an alternative minimization approach is designed.

1. Fix $\boldsymbol{D}$ to optimize $\boldsymbol{Y}$

$$\min_{\boldsymbol{Y}}\left\|\boldsymbol{X}_{\text{im}}-\boldsymbol{D}_{\text{im}}\boldsymbol{Y}\right\|_F^2+\alpha\left\|\boldsymbol{X}_{\text{dep}}-\boldsymbol{D}_{\text{dep}}\boldsymbol{Y}\right\|_F^2, \tag{3}$$

is an unconstrained optimization. And we can give the analytic solution as

$$\boldsymbol{Y}=\left(\boldsymbol{D}_{\text{im}}^T\boldsymbol{D}_{\text{im}}+\alpha\boldsymbol{D}_{\text{dep}}^T\boldsymbol{D}_{\text{dep}}\right)^{-1}\left(\boldsymbol{D}_{\text{im}}^T\boldsymbol{X}_{\text{im}}+\alpha\cdot\boldsymbol{D}_{\text{dep}}^T\boldsymbol{X}_{\text{dep}}\right), \tag{4}$$

2. Fix $\boldsymbol{Y}$ to update $\boldsymbol{D}$, then Eq. 2 can be reformed as:

$$\min_{\boldsymbol{D}_t}\left\|\boldsymbol{X}_t-\boldsymbol{D}_t\boldsymbol{Y}\right\|_F^2, t\in\{\text{im,dep}\}, \tag{5}$$

[2] We chose training patches by retrieving the most similar images from the database with gist.

[3] Without loss of generality, we take the depth image, the most popular 3D form in single monocular depth estimation, as an example.

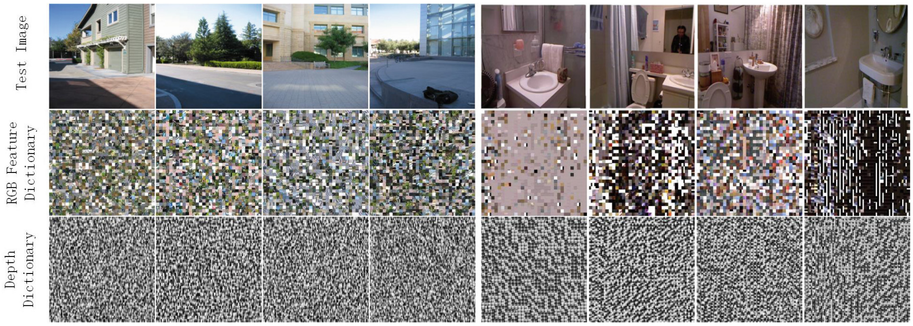

Fig. 2. Visualization results of Dictionaries: The left four columns are visualized from Make3D dataset and the right four columns from NYUv2 dataset. The first row consists of test images, the second row and third row consist of RGB feature dictionaries and depth dictionaries,respectively, which are trained by the candidate [22] images.

which can be solved by postmultiplication of Moore-Penrose generalized inverse matrix [2] of $\boldsymbol{Y}$[4] as

$$\begin{aligned} \boldsymbol{D}_t &= \boldsymbol{X}_t \text{Inverse}\left(\boldsymbol{Y}\right) \\ \text{Inverse}\left(\boldsymbol{Y}\right) &= \boldsymbol{Y}^T \left(\boldsymbol{Y}\boldsymbol{Y}^T + \boldsymbol{I}\epsilon\right)^{-1}, \end{aligned} \tag{6}$$

3.2 Cross-Modality Retrieval

So far, we have trained the dictionaries $\boldsymbol{D}_{\text{dep}}$ and $\boldsymbol{D}_{\text{im}}$. Given a set of queries $\boldsymbol{X}_{\text{im}}^{\theta}$ of 2D patches, our goal is to get the corresponding 3D model $\boldsymbol{X}_{\text{dep}}^{\theta}$. The optimal result can be obtained by Eq. 2 as

$$\min_{\boldsymbol{Y}^{\theta}, \boldsymbol{X}_{\text{dep}}^{\theta}} \left\| \boldsymbol{X}_{\text{im}}^{\theta} - \boldsymbol{D}_{\text{im}} \boldsymbol{Y}^{\theta} \right\|_F^2 + \alpha \left\| \boldsymbol{X}_{\text{dep}}^{\theta} - \boldsymbol{D}_{\text{dep}} \boldsymbol{Y}^{\theta} \right\|_F^2 . \tag{7}$$

To accelerate the convergence in Eq. 7, we can initialize parameters using

$$\begin{aligned} \hat{\boldsymbol{Y}}^{\theta} &= \min_{\boldsymbol{Y}} \left\| \boldsymbol{X}_{\text{im}}^{\theta} - \boldsymbol{D}_{\text{im}} \boldsymbol{Y}^{\theta} \right\|_2^2 + \alpha \left\| \boldsymbol{X}_{\text{dep}}^{\theta} - \boldsymbol{D}_{\text{dep}} \boldsymbol{Y}^{\theta} \right\|_F^2 , \\ \hat{\boldsymbol{X}}_{\text{dep}}^{\theta} &= \boldsymbol{D}_{\text{dep}} \hat{\boldsymbol{Y}}^{\theta}. \end{aligned} \tag{8}$$

After obtaining the reconstruction coefficient $\hat{\boldsymbol{Y}}^{\theta}$ and the related 3D model $\hat{\boldsymbol{X}}_{\text{dep}}^{\theta}$ of image patches, we can optimize the entire image by setting initial depth value $\hat{\boldsymbol{X}}_{\text{dep}}^{\theta}$[5].

[4] Generally, $\text{Inverse}\left(\boldsymbol{Y}\right) = \boldsymbol{G}^H \left(\boldsymbol{G}\boldsymbol{G}^H\right)^{-1} \left(\boldsymbol{M}^H \boldsymbol{M}\right)^{-1} \boldsymbol{M}^H$ takes too much time, and is replaced by Eq. 6. $\boldsymbol{G}$, $\boldsymbol{M}$ are the row and column full rank matrices computed by a full rank decomposition of $\boldsymbol{Y}$,respectively.

[5] Further process will be explained in Sect. 4.

4 Large Margin Structure Inference

We gather the initial depth patches (Sect. 3.2) to form the initial depth of the entire image $\boldsymbol{I}^0_{\mathrm{dep}}$. There are N images $\boldsymbol{I}^i_{\mathrm{im}}\,(i = 1, \cdots, N)$ from dataset that are similar [22] with the query image $\boldsymbol{I}^0_{\mathrm{im}}$ in RGB space, whose depth images are $\boldsymbol{I}^i_{\mathrm{dep}}\,(i = 1, \cdots, N)$. And the depth image, we want to infer, is $\widetilde{\boldsymbol{I}}_{\mathrm{dep}}$.

The Algorithm 1. the Proposed Method⋆

Input

Query Image $\boldsymbol{I}^0_{\mathrm{im}} \in \Re^{w\times h}$,
Candidate Images $\boldsymbol{I}^i_{\mathrm{im}} \in \Re^{w\times h}\,(i = 1, \cdots, N)$,
Corresponding Candidate 3D Models $\boldsymbol{I}^i_{\mathrm{dep}} \in \Re^{w\times h}\,(i = 1, \cdots, N)$.

1. Cross-Modality based Prior Depth Inference

   (a) Extract overlapped image patches $x^j_{\mathrm{im}} \in \Re^{p\times 1}\,(j = 1, \cdots, n)$ from $\boldsymbol{I}^i_{\mathrm{im}}$, and corresponding depth map patches $x^j_{\mathrm{dep}} \in \Re^{q\times 1}\,(j = 1, \cdots, n)$ from $\boldsymbol{I}^i_{\mathrm{dep}}$, using Eq. 4 and Eq. 6 to calculate Dictionary $\boldsymbol{D}_{\mathrm{im}} \in \Re^{p\times c}$ and $\boldsymbol{D}_{\mathrm{dep}} \in \Re^{q\times c}$

   (b) Extract non-overlapped image patches $x^k_{\mathrm{im}} \in \Re^{p\times 1}\,(k = 1, \cdots, m)$ from $\boldsymbol{I}^0_{\mathrm{im}}$, using Eq. 8 to calculate the corresponding initial depth map patches $x^k_{\mathrm{dep}} \in \Re^{p\times 1}\,(i = 1, \cdots, m)$.

   (c) Obtain the prior depth $\boldsymbol{I}^0_{\mathrm{dep}}$ of the entire image $\boldsymbol{I}^0_{\mathrm{im}}$

2. Large Margin Structure Inference
   To minimize Eq. 16, is equivalent to minimize

$$\begin{aligned}\ln \Psi_d\left(\widetilde{\boldsymbol{I}}_{\mathrm{dep}}, \boldsymbol{I}^i_{\mathrm{dep}}, \widetilde{\boldsymbol{I}}_{\mathrm{im}}, \boldsymbol{I}^i_{\mathrm{im}}\right) &= \ln \Psi_{ds}\left(\widetilde{\boldsymbol{I}}_{\mathrm{dep}}\right) + \ln \Psi_{dp}\left(\boldsymbol{I}^0_{\mathrm{dep}}, \widetilde{\boldsymbol{I}}_{\mathrm{dep}}\right)\\ &\quad + \ln \Psi_{dd}\left(\widetilde{\boldsymbol{I}}_{\mathrm{dep}}, \boldsymbol{I}^i_{\mathrm{dep}}, \widetilde{\boldsymbol{I}}_{\mathrm{im}}, \boldsymbol{I}^i_{\mathrm{im}}\right).\end{aligned} \tag{9}$$

Eq. 16 can be transformed into the following format

$$\ln \Psi_d\left(\widetilde{\boldsymbol{I}}_{\mathrm{dep}}, \boldsymbol{I}^i_{\mathrm{dep}}, \widetilde{\boldsymbol{I}}_{\mathrm{im}}, \boldsymbol{I}^i_{\mathrm{im}}\right) = \sum_r \left\|\boldsymbol{A}_r \widetilde{\boldsymbol{I}}_{\mathrm{dep}} - b_r\right\|. \tag{10}$$

To minimize Eq. 10, we can get the lth iteration solution of $\widetilde{\boldsymbol{I}}_{\mathrm{dep}}$ by gradient descent

$$\widetilde{\boldsymbol{I}}^l_{\mathrm{dep}} = \left(\sum_{r,s} \frac{\boldsymbol{A}^T_{(r,s)}\boldsymbol{A}_{(r,s)}}{\sqrt{\left(\boldsymbol{A}_{(r,s)}\widetilde{\boldsymbol{I}}^{l-1}_{\mathrm{dep}} - b_{(r,s)}\right)^2 + \epsilon}}\right)^{-1}\left(\sum_{r,s} \frac{\boldsymbol{A}^T_{(r,s)} b_{(r,s)}}{\sqrt{\left(\boldsymbol{A}_{(r,s)}\widetilde{\boldsymbol{I}}^{l-1}_{\mathrm{dep}} - b_{(r,s)}\right)^2 + \epsilon}}\right) \tag{11}$$

where $\boldsymbol{A}_{(r,s)}$ is the sth row of $\boldsymbol{A}_r$, $b_{(r,s)}$ is the sth element of vector b_r and ϵ is 10^{-6}

Output

The optimized depth map $\widetilde{\boldsymbol{I}}^*_{\mathrm{dep}}$ of image $\boldsymbol{I}^0_{\mathrm{im}}$

⋆The overall time complexity is $O(mn + p^3 + p^2q + N)$, in which m is the average number of patches in a training image, n is the number of images used for dictionary learning, p is the size of the codebook, and N is the size of the database to search, assuming the patch size is $q \times q$.

The Taylor expansion of $\widetilde{I}_{\text{dep}}$ and I^i_{dep} at point (a, b) are

$$\begin{aligned}\widetilde{I}_{\text{dep}}(x,y) = {} & \widetilde{I}_{\text{dep}}(a,b) + \boldsymbol{\nabla}_x \widetilde{I}_{\text{dep}}(a,b)\cdot(x-a) + \boldsymbol{\nabla}_y \widetilde{I}_{\text{dep}}(a,b)\cdot(y-b)\\ & + \frac{1}{2}\boldsymbol{\nabla}_x^2 \widetilde{I}_{\text{dep}}(a,b)\cdot(x-a)^2 + \frac{1}{2}\boldsymbol{\nabla}_y^2 \widetilde{I}_{\text{dep}}(a,b)\cdot(y-b)^2\\ & + \frac{1}{2}\boldsymbol{\nabla}_{x,y} \widetilde{I}_{\text{dep}}(a,b)\cdot(x-a)(y-b)\\ & + \frac{1}{2}\boldsymbol{\nabla}_{y,x} \widetilde{I}_{\text{dep}}(a,b)\cdot(x-a)(y-b) + R_n(x,y)\end{aligned} \tag{12}$$

and

$$\begin{aligned}I^i_{\text{dep}}(x,y) = {} & I^i_{\text{dep}}(a,b) + \boldsymbol{\nabla}_x I^i_{\text{dep}}(a,b)\cdot(x-a) + \boldsymbol{\nabla}_y I^i_{\text{dep}}(a,b)\cdot(y-b)\\ & + \frac{1}{2}\boldsymbol{\nabla}_x^2 I^i_{\text{dep}}(a,b)\cdot(x-a)^2 + \frac{1}{2}\boldsymbol{\nabla}_y^2 I^i_{\text{dep}}(a,b)\cdot(y-b)^2\\ & + \frac{1}{2}\boldsymbol{\nabla}_{x,y} I^i_{\text{dep}}(a,b)\cdot(x-a)(y-b)\\ & + \frac{1}{2}\boldsymbol{\nabla}_{y,x} I^i_{\text{dep}}(a,b)\cdot(x-a)(y-b) + L_n(x,y),\end{aligned} \tag{13}$$

where $R_n(x,y)$ and $L_n(x,y)$ are the higher order infinitesimals. To make $\widetilde{I}_D$ and I^i_D similar, Eqs. 12 and 13 should also be similar. Then we can get the expression of G_{sim} and G_{sel} as

$$\begin{aligned}G_{sim} = {} & \sum_{i=1}^{N}\left\|\boldsymbol{W}_i\cdot\left(\widetilde{I}_D - I^i_D\right)\right\| + \alpha\left\|\boldsymbol{W}_i\cdot\left(\boldsymbol{\nabla}_x\widetilde{I}_D - \boldsymbol{\nabla}_x I^i_D\right)\right\|\\ & +\alpha\left\|\boldsymbol{W}_i\cdot\left(\boldsymbol{\nabla}_y\widetilde{I}_D - \boldsymbol{\nabla}_y I^i_D\right)\right\| + \beta\left\|\boldsymbol{W}_i\left(\boldsymbol{\nabla}_x^2\widetilde{I} - \boldsymbol{\nabla}_x^2 I^i_D\right)\right\|\\ & +\beta\left\|\boldsymbol{W}_i\left(\boldsymbol{\nabla}_y^2\widetilde{I} - \boldsymbol{\nabla}_y^2 I^i_D\right)\right\| + \beta\left\|\boldsymbol{W}_i\left(\boldsymbol{\nabla}_{x,y}\widetilde{I} - \boldsymbol{\nabla}_{x,y} I^i_D\right)\right\|\\ & +\beta\left\|\boldsymbol{W}_i\left(\boldsymbol{\nabla}_{y,x}\widetilde{I} - \boldsymbol{\nabla}_{y,x} I^i_D\right)\right\|,\end{aligned} \tag{14}$$

and

$$\begin{aligned}G_{sel} = {} & \gamma\left\|\widetilde{I}_D - I^0_D\right\| + \alpha\left(\left\|\boldsymbol{W}_0\cdot\boldsymbol{\nabla}_x\widetilde{I}_D\right\| + \left\|\boldsymbol{W}_0\cdot\boldsymbol{\nabla}_y\widetilde{I}_D\right\|\right)\\ & +\beta\left(\left\|\boldsymbol{W}_0\cdot\boldsymbol{\nabla}_x^2\widetilde{I}\right\| + \left\|\boldsymbol{W}_0\cdot\boldsymbol{\nabla}_y^2\widetilde{I}\right\| + \left\|\boldsymbol{W}_0\cdot\boldsymbol{\nabla}_{x,y}\widetilde{I}\right\| + \left\|\boldsymbol{W}_0\cdot\boldsymbol{\nabla}_{y,x}\widetilde{I}\right\|\right).\end{aligned} \tag{15}$$

where G_{sim} is used to calculate similarity between input RGB image and candidate images and G_{sel} is the self control item which guarantees that adjacent points in an image have similar depth value.

Similar to the regular CRF, G_{sim} and G_{sel} can be reformed as traditional MRF *i.e.* the smoothing term $\Psi_{ds}\left(\widetilde{\boldsymbol{I}}_{\text{dep}}\right)$, the data term $\Psi_{dd}\left(\widetilde{\boldsymbol{I}}_{\text{dep}}, \boldsymbol{I}^i_{\text{dep}}, \widetilde{\boldsymbol{I}}_{\text{im}}, \boldsymbol{I}^i_{\text{im}}\right)$ and the prior depth term $\Psi_{dp}\left(\boldsymbol{I}^0_{\text{dep}}, \widetilde{\boldsymbol{I}}_{\text{dep}}\right)$, defined as

$$\Psi_d\left(\widetilde{\boldsymbol{I}}_{\text{dep}}, \boldsymbol{I}^i_{\text{dep}}, \widetilde{\boldsymbol{I}}_{\text{im}}, \boldsymbol{I}^i_{\text{im}}\right) = \Psi_{ds}\left(\widetilde{\boldsymbol{I}}_{\text{dep}}\right)\Psi_{dd}\left(\widetilde{\boldsymbol{I}}_{\text{dep}}, \boldsymbol{I}^i_{\text{dep}}, \widetilde{\boldsymbol{I}}_{\text{im}}, \boldsymbol{I}^i_{\text{im}}\right)\Psi_{dp}\left(\boldsymbol{I}^0_{\text{dep}}, \widetilde{\boldsymbol{I}}_{\text{dep}}\right). \tag{16}$$

Data Term. Depending on our basic assumption that similar image should have similar depth map, we use similar [22] candidate images to infer our depth map $\widetilde{\boldsymbol{I}}_{\text{dep}}$. We claim that this "similarity" should not only happen in the original RGB images, but also in the gradient of RGB images. When comparing with pixels in $\boldsymbol{I}_{\text{im}}^{0}$ and $\boldsymbol{I}_{\text{im}}^{i}$, the more similar they are, the less weight they have. Then we give our formulation of $\mathit{\Psi}_{dd}\left(\widetilde{\boldsymbol{I}}_{\text{dep}}, \boldsymbol{I}_{\text{dep}}^{i}, \widetilde{\boldsymbol{I}}_{\text{im}}, \boldsymbol{I}_{\text{im}}^{i}\right)$ as

$$
\begin{aligned}
\mathit{\Psi}_{dd}\left(\widetilde{\boldsymbol{I}}_{\text{dep}}, \boldsymbol{I}_{\text{dep}}^{i}, \widetilde{\boldsymbol{I}}_{\text{im}}, \boldsymbol{I}_{\text{im}}^{i}\right) = \prod_{i=1}^{N} exp(& \left\|\boldsymbol{W}_i\left(\widetilde{\boldsymbol{I}}_{\text{dep}} - \boldsymbol{I}_{\text{dep}}^{i}\right)\right\| + \alpha\left\|\boldsymbol{W}_i\left(\boldsymbol{\nabla}_x\widetilde{\boldsymbol{I}}_{\text{dep}} - \boldsymbol{\nabla}_x\boldsymbol{I}_{\text{dep}}^{i}\right)\right\| \\
& +\alpha\left\|\boldsymbol{W}_i\left(\boldsymbol{\nabla}_y\widetilde{\boldsymbol{I}}_{\text{dep}} - \boldsymbol{\nabla}_y\boldsymbol{I}_{\text{dep}}^{i}\right)\right\| + \beta\left\|\boldsymbol{W}_i\left(\boldsymbol{\nabla}_x^2\widetilde{\boldsymbol{I}}_{\text{dep}} - \boldsymbol{\nabla}_x^2\boldsymbol{I}_{\text{dep}}^{i}\right)\right\| \\
& +\beta\left\|\boldsymbol{W}_i\left(\boldsymbol{\nabla}_y^2\widetilde{\boldsymbol{I}}_{\text{dep}} - \boldsymbol{\nabla}_y^2\boldsymbol{I}_{\text{dep}}^{i}\right)\right\| + \beta\left\|\boldsymbol{W}_i\left(\boldsymbol{\nabla}_{x,y}\widetilde{\boldsymbol{I}}_{\text{dep}} - \boldsymbol{\nabla}_{x,y}\boldsymbol{I}_{\text{dep}}^{i}\right)\right\| \\
& +\beta\left\|\boldsymbol{W}_i\left(\boldsymbol{\nabla}_{y,x}\widetilde{\boldsymbol{I}}_{\text{dep}} - \boldsymbol{\nabla}_{y,x}\boldsymbol{I}_{\text{dep}}^{i}\right)\right\|)
\end{aligned}
\tag{17}
$$

where $\boldsymbol{W}_i$[6]is the point-wise similarity diagonal matrix.[7]

Smoothing Term. We encourage neighborhood pixels have smooth depth estimations. This is achieved in $\mathit{\Psi}_{ds}\left(\widetilde{\boldsymbol{I}}_{\text{dep}}\right)$ by setting self-adapting coefficient of adjacent pixels smoothing. When the features of adjacent pixels are similar, then the smoothing coefficient of that pixel pair would achieve a low smoothing weight to make the pixel pair very smooth; meanwhile, when the adjacent pixel features are dramatically different, then the smoothing coefficient will be very high, which makes the smoothing term lose their efficacies. We come up with the following design to characterize the above intuitions.

$$
\begin{aligned}
\mathit{\Psi}_{ds}\left(\widetilde{\boldsymbol{I}}_{\text{dep}}\right) = & exp(\alpha\left\|\boldsymbol{W}_0\boldsymbol{\nabla}_x\widetilde{\boldsymbol{I}}_{\text{dep}}\right\| + \alpha\left\|\boldsymbol{W}_0\boldsymbol{\nabla}_y\widetilde{\boldsymbol{I}}_{\text{dep}}\right\| + \beta\left\|\boldsymbol{W}_0\boldsymbol{\nabla}_x^2\widetilde{\boldsymbol{I}}_{\text{dep}}\right\| \\
& + \beta\left\|\boldsymbol{W}_0\boldsymbol{\nabla}_y^2\widetilde{\boldsymbol{I}}_{\text{dep}}\right\| + \beta\left\|\boldsymbol{W}_0\boldsymbol{\nabla}_{x,y}\widetilde{\boldsymbol{I}}_{\text{dep}}\right\| + \beta\left\|\boldsymbol{W}_0\boldsymbol{\nabla}_{y,x}\widetilde{\boldsymbol{I}}_{\text{dep}}\right\|)
\end{aligned}
\tag{18}
$$

where the first two terms in Eq. 18 are of first-order gradient smooth, which cover the nearest four pixels neighbours; while the other terms in Eq. 18 are second-order gradient smooth, which cover more further area. $\boldsymbol{\nabla}_x, \boldsymbol{\nabla}_y, \boldsymbol{\nabla}_x^2, \boldsymbol{\nabla}_y^2, \boldsymbol{\nabla}_{x,y}, \boldsymbol{\nabla}_{y,x}$, are the gradient operator matrix, $\widetilde{\boldsymbol{I}}_{\text{dep}}$ is a column vector and $\boldsymbol{W}_0$ is the self-adapting smooth control (diagonal) matrix.

Prior Term. We also claim that the estimated prior should join in the depth consistency potential as

$$
\mathit{\Psi}_{dp}\left(\widetilde{\boldsymbol{I}}_{\text{dep}}, \boldsymbol{I}_{\text{dep}}^{0}\right) = exp\left(\gamma\left\|\widetilde{\boldsymbol{I}}_{\text{dep}} - \boldsymbol{I}_{\text{dep}}^{0}\right\|\right) \tag{19}
$$

Comparing with traditional methods [11,26], there is no pre-trained parameters in our model, which provides a highly generalization ability. Meanwhile, the

[6] $\boldsymbol{W}_i\left(j, j\right) = \text{sigmoid}\left(\frac{\left\|F_{\text{im}}^{0}(j) - F_{\text{im}}^{i}(j)\right\| - \mu_i}{\sigma_i}\right)$, F_{im}^{*} is the SIFT [18] feature of image I_{im}^{*}. And the elements of $\boldsymbol{W}_0$ in Eq. 18 are calculated with the same image but adjacent point.

[7] $\|\cdot\|$ is the one-norm.

larger neighbourhood have been considered,[8] without increasing time complexity. We show the proposed algorithm and the entire framework in Algorithm 1 and Fig. 1, respectively.

5 Experiments

In this section, we report our experimental results on single image depth estimation for both outdoor and indoor scenes. We use the Make3D [26] range image data set and the NYUv2 [29] Kinect data set, as they are the largest open data available at present.

5.1 Evaluation Protocols

For quantitative evaluation, we report errors obtained with the following error metrics, which have been extensively used in [11,14,17,26].

- Mean relative error (rel): $\frac{1}{L}\sum_i \frac{|\hat{d}_i - d_i|}{d_i}$;
- Mean $log10$ error ($lg10$): $\frac{1}{L}\sum_i |log_{10}\hat{d}_i - log_{10}d_i|$;
- Root mean squared error (rms):$\sqrt{\frac{1}{L}\sum_i \left\|\hat{d}_i - d_i\right\|_2^2}$

where d_i is the ground truth depth, $\hat{d}_i$ is the estimated depth, and L denotes the total number of pixels in all the evaluated images.

In the training stage, we select 10 similar [22] images from dataset with the query image, and use the patches(7 × 7 pixels, 3 pixels overlap) extracted from these similar images to train RGB feature [26] dictionary and depth dictionary simultaneously, whose dimensionality is 1024. And the balance parameter in Eq. 2 is 1. In the testing stage, we extract non-overlapping patches of query image to infer the prior depth image. And to optimize this prior depth image with Eq. 11, we fix the parameter of Eqs. 17, 18 and 19 as γ for 0.5, α for 10 and β for 0.1.

5.2 Performance on Make3D Dataset

The Make3D dataset consists of 534 images with corresponding depth maps. There are 400 training images and 134 testing images. All images are resized to 460×345 pixels. It is worth noting that this data set is published a decade ago, the resolution and distance range of the depth image is rather limited (only 55 × 305 pixels). Furthermore, it contains noise in the locations of glass window etc. These limitations have some influence on the training stage and the resulting error metrics. Therefore we report errors based on two different criteria in Table 1: (C1) Errors are computed in the regions with ground-truth depth less than 70;

[8] First-order derivation covers the nearest 4 points, and second-order derivation effect as a two-level first-order derivation.

Table 1. Result comparisons on the Make3D dataset.(C1) Errors are computed in the regions with ground-truth depth less than 70; (C2) Errors are computed in the entire image

Method	Error(C1) (lower is better)			Error(C2) (lower is better)		
	rel	lg10	rms	rel	lg10	rms
Make3D [26]	-	-	-	0.370	0.187	-
Semantic Labelling [17]	-	-	-	0.379	0.148	-
Depth MRF*[25]	-	-	-	0.530	0.198	16.7
Feedback Cascades*[16]	-	-	-	-	-	15.2
DepthTransfer [11]	0.355	**0.127**	**9.20**	0.361	0.148	15.10
Ours	**0.345**	**0.127**	9.41	**0.337**	**0.137**	**13.70**

*Results reported in DepthTransfer [11].

(C2) Errors are computed in the entire image. We compare our method with the state-of-the-art methods such as Make3D [26], Depth Transfer [11] and Semantic Labelling [17].

In Table 1, we present a quantitative comparison of the depth estimation between our method and these methods on representative images from Make3D data set. Table 1 demonstrates that, in most cases, our method outperforms those competing methods in terms of two evaluation criteria. Also, to make the result visible, we show the depth prediction results achieved by our method in Fig. 3. From Fig. 3, we can observe that the prediction results achieved by our method are very close to the ground truth images, and are much better than those obtained by the Make3D approach. To prove the validity of our methods, we also compare our method with state-of-the-art in the "Prior Depth Inference" and "Depth optimization", respectively (Table 2 and Table 3). At last, we show the influence of parameters in Table 4.

From Table 1, we can see that our method outperforms in most of the metrics. Furthermore, comparing "Error (C2)" criteria with "Error (C1)", our model achieves more gains in far distance objects than the near ones. And comparing with other methods, without pre-trained parameter [25,26] and supplementary information [17], our model can still work well.

We also test our model with state-of-the-art in each stage. In Table 2 we assess the effectiveness of the "Prior Depth Inference" stage, and test our model in learned-dictionary and random dictionary. From the result we can see that, even with a random dictionary our model still outperforms state-of-the-art methods. Compared with Table 4, random dictionary performance is similar to learned-dictionary of 5 pixels patch size. In Table 3, we assess the effectiveness of the "Entire image depth inference" stage with the same prior depth value of [11]. From this table we can see that, our model have lower rms but higher rel which means our method effective but slightly unstable.

In Table 4, we can see that the patch size parameter poses greater influence than the other two. Generally speaking, mapping RGB to depth is an ill-posed

Table 2. Result comparisons on the Make3D dataset without MRF to fine-tune.

Method	Error(C1) (lower is better)			Error(C2) (lower is better)		
	rel	lg10	rms	rel	lg10	rms
Make3D [26]	-	-	14.79	-	-	29.27
DepthTransfer [11]	0.936	0.217	12.01	0.903	0.247	20.49
Ours(with random dictionary)	0.936	0.216	**11.96**	0.862	0.218	16.45
Ours	**0.866**	**0.216**	11.99	**0.801**	**0.217**	**16.41**

Table 3. Result comparisons on the Make3D dataset, with the same prior depth estimation, different MRF to fine-tune.

Method	Error(C1) (lower is better)			Error(C2) (lower is better)		
	rel	lg10	rms	rel	lg10	rms
DepthTransfer [11]	**0.355**	**0.127**	9.20	**0.361**	0.148	15.10
Ours	0.375	**0.127**	**9.18**	0.364	**0.141**	**14.11**

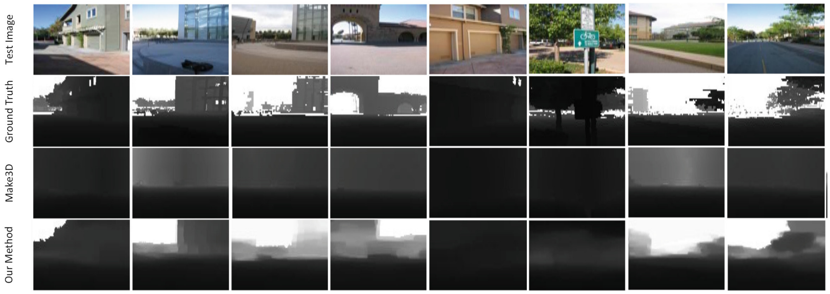

Fig. 3. Examples of depth predictions on the Make3D dataset.

problem that there may be many depth patches for a certain RGB patch. And the larger the patch is, the more details can be learnt. However, due to the lack of adequate images, the range of patch size is also limited. Based on this reason, the dictionary size effects a little, which can also be seen in Fig. 2 that there are lots of reduplicative feature in the trained dictionary.

From Fig. 3 we can see that our method reproduce the depth map well, especially at shape controlling.

5.3 Performance on NYUv2 Dataset

The NYUv2 dataset contains of 1449 images, where 795 images are used as a training set and 654 images are used as a testing set[9]. All images are resized

[9] We only compare the result of standard data partition, when the code is not available.

Table 4. Result comparisons on the Make3D dataset with different parameters. PatchSize is the size of extracted patches for "Coupled Dictionary Learning" and DictionarySize is the capacity of Dictionary $\boldsymbol{D}_{\mathrm{im}}$ and $\boldsymbol{D}_{\mathrm{dep}}$ in Eq. 2.

Parameter	Error(C1) (lower is better)			Error(C2) (lower is better)		
	rel	lg10	rms	rel	lg10	rms
PatchSize = 3	1.470	0.216	12.02	1.332	0.217	16.41
PatchSize = 5	0.936	0.216	11.99	0.862	0.217	16.41
PatchSize = 7	**0.866**	**0.216**	**11.99**	**0.801**	**0.217**	**16.41**
DictionarySize = 256	0.867	0.216	12.03	0.803	0.217	16.41
DictionarySize = 512	0.867	0.216	12.03	0.803	0.217	16.41
DictionarySize = 1024	**0.866**	**0.216**	**11.99**	**0.801**	**0.217**	**16.41**
$^{\star}\alpha = 0.1$	0.866	0.216	11.99	0.801	0.217	16.41
$\alpha = 1$	**0.866**	**0.216**	**11.99**	**0.801**	**0.217**	**16.41**
$\alpha = 10$	0.866	0.216	11.99	0.801	0.217	16.41

$^{\star}\alpha$ is the balance parameter in Eq. 2

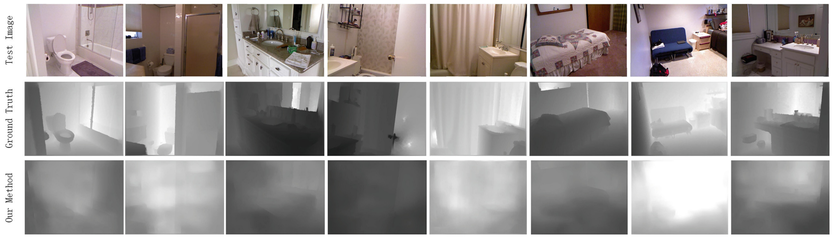

Fig. 4. Examples of depth predictions on the NYUv2 dataset.

to 460 × 345 pixels in order to preserve the aspect ratio of the original images. In Table 5, we compare our method with state-of-the-art methods, including Make3D [26], Depth Transfer [11] and so on.

As illustrated in Table 5, we present a qualitative comparison of the depth estimation with these methods on representative images from NYUv2 data set, which demonstrates the superior performance of our method. Also, to make the result visible, we show our method in Fig. 4. The set of parameter in our method is the same as in Sect. 5.2. Due to the similar experiment result (Sect. 5.2) and the limitation of pages.

5.4 Comparison with Deep Learning Methods

It is well known that deep learning methods have obtained remarkable achievement in many research areas, due to their greater learning ability than most of traditional methods. The proposed method has no advantage in model capability or complexity, compared to deep learning.

Table 5. Result comparisons on the NYUv2 dataset.

Method	Error (lower is better)		
	rel	lg10	rms
Make3D [26]	0.349	-	1.214
Depth Fusion*[12]	0.368	0.135	1.3
Depth Fusion(no warp)*[13]	0.371	0.137	1.3
DepthTransfer [11]	0.350	0.131	1.2
Ours	**0.342**	**0.130**	**1.18**

*Results reported in DepthTransfer [11].

However deep learning has a critical drawback that the training process usually takes a long time (weeks and even months), despite considerable efforts have been taken to alleviate this problem. Most deep neural networks also heavily rely on parameter tuning, with significant sensitivity on certain parameters such as learning rate. This prevents deep learning approaches from being applied in scenarios that require frequent and agile updating.

On the contrary, the proposed approach requires no traditional training stage and few parameters. This greatly reduces the effort of adapting to a new dataset, making the approach more flexible and reliable. Since these two methods are designed for different scenarios, we do not conduct the experimental comparison.

6 Conclusion

In this paper, we propose a novel cross-modality retrieval method to estimate the object depth value from a given 2D image. To our best knowledge, this is the first method to estimate depth value by cross-modality retrieval. And to solve the cross-modality problem, we propose a novel and effective coupled dictionary learning method. Based on the local depth estimation from the cross-modal retrieval using the dictionary, we further refine the depth of the entire image by solving a convex optimization problem. From the depth estimation result (Figs. 3 and 4), we can see that details are not well reserved in our method. Because our method highly depends on the candidate images. When the images do not describe the same scene as query image does, or the “bad” image win a high similar score on pixel level, our method can not work well. In the future, we plan to combine our model with the deep learning or other methods to improve the robustness in handling real-world image transformation. Furthermore, we plan to augment the performance by integrating the semantic information from the recent development in CNN framework.

Acknowledgement. This work was supported in part by the Major State Basic Research Development Program of China (973 Program) under Grant 2015CB351804, the National Key R&D Program (No. 2016YFB1001503), the National Science Foundation of China (No. 61422210, No.61472101, No. 61373076, No. 61402388), the CCF-

Tencent Open Research Fund, and the Open Projects Program of National Laboratory of Pattern Recognition.

References

1. Barinova, O., Konushin, V., Yakubenko, A., Lee, K.C., Lim, H., Konushin, A.: Fast automatic single-view 3-d reconstruction of urban scenes. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008. LNCS, vol. 5303, pp. 100–113. Springer, Heidelberg (2008). doi:10.1007/978-3-540-88688-4_8
2. Bellman, R., Bellman, R.E., Bellman, R.E., Bellman, R.E.: Introduction to Matrix Analysis, vol. 960. SIAM (1970)
3. Byeon, W., Breuel, T.M., Raue, F., Liwicki, M.: Scene labeling with lstm recurrent neural networks. In: Computer Vision and Pattern Recognition (CVPR), pp. 3547–3555. IEEE (2015)
4. Chellappa, R., Jain, A. (eds.): Markov Random Fields. Theory and Application. Academic Press, Boston (1993)
5. Dellaert, F., Seitz, S.M., Thorpe, C.E., Thrun, S.: Structure from motion without correspondence. In: Computer Vision and Pattern Recognition (CVPR), vol. 2, pp. 557–564. IEEE (2000)
6. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. In: Advances in Neural Information Processing Systems (NIPS), pp. 2366–2374 (2014)
7. Engel, J., Schöps, T., Cremers, D.: LSD-SLAM: large-scale direct monocular SLAM. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8690, pp. 834–849. Springer, Heidelberg (2014). doi:10.1007/978-3-319-10605-2_54
8. He, L., Qi, H., Zaretzki, R.: Beta process joint dictionary learning for coupled feature spaces with application to single image super-resolution. In: Computer Vision and Pattern Recognition (CVPR), pp. 345–352. IEEE (2013)
9. Hoiem, D., Efros, A.A., Hebert, M.: Geometric context from a single image. In: International Conference on Computer Vision (ICCV), vol. 1, pp. 654–661. IEEE (2005)
10. Karsch, K., Liu, C., Kang, S.B.: Depth extraction from video using non-parametric sampling. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7576, pp. 775–788. Springer, Heidelberg (2012). doi:10.1007/978-3-642-33715-4_56
11. Karsch, K., Liu, C., Kang, S.B.: Depth transfer: Depth extraction from video using non-parametric sampling. IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI) **36**(11), 2144–2158 (2014)
12. Konrad, J., Brown, G., Wang, M., Ishwar, P., Wu, C., Mukherjee, D.: Automatic 2d-to-3d image conversion using 3d examples from the internet. In: IS&T/SPIE Electronic Imaging, p. 82880F. International Society for Optics and Photonics (2012)
13. Konrad, J., Wang, M., Ishwar, P.: 2d-to-3d image conversion by learning depth from examples. In: Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 16–22. IEEE (2012)
14. Ladicky, L., Shi, J., Pollefeys, M.: Pulling things out of perspective. In: Computer Vision and Pattern Recognition (CVPR), pp. 89–96. IEEE (2014)

15. Li, B., Shen, C., Dai, Y., van den Hengel, A., He, M.: Depth and surface normal estimation from monocular images using regression on deep features and hierarchical CRFs. In: Computer Vision and Pattern Recognition (CVPR), pp. 1119–1127 (2015)
16. Li, C., Kowdle, A., Saxena, A., Chen, T.: Towards holistic scene understanding: feedback enabled cascaded classification models. In: Advances in Neural Information Processing Systems (NIPS), pp. 1351–1359 (2010)
17. Liu, B., Gould, S., Koller, D.: Single image depth estimation from predicted semantic labels. In: Computer Vision and Pattern Recognition (CVPR), pp. 1253–1260. IEEE (2010)
18. Liu, C., Yuen, J., Torralba, A.: Sift flow: dense correspondence across scenes and its applications. IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI) **33**(5), 978–994 (2011)
19. Liu, F., Shen, C., Lin, G.: Deep convolutional neural fields for depth estimation from a single image. In: Computer Vision and Pattern Recognition (CVPR), pp. 5162–5170. IEEE (2015)
20. Liu, M., Salzmann, M., He, X.: Discrete-continuous depth estimation from a single image. In: Computer Vision and Pattern Recognition (CVPR), pp. 716–723. IEEE (2014)
21. Mullane, J., Vo, B.N., Adams, M.D., Vo, B.T.: A random-finite-set approach to Bayesian slam. IEEE Trans. Robot. **27**(2), 268–282 (2011)
22. Oliva, A., Torralba, A.: Modeling the shape of the scene: a holistic representation of the spatial envelope. Intl. J. Comput. Vis. (IJCV) **42**(3), 145–175 (2001)
23. Rasiwasia, N., Costa Pereira, J., Coviello, E., Doyle, G., Lanckriet, G.R., Levy, R., Vasconcelos, N.: A new approach to cross-modal multimedia retrieval. In: International Conference on Multimedia (MM), pp. 251–260. ACM (2010)
24. Rock, J., Gupta, T., Thorsen, J., Gwak, J., Shin, D., Hoiem, D.: Completing 3d object shape from one depth image. In: Computer Vision and Pattern Recognition (CVPR), pp. 2484–2493. IEEE (2015)
25. Saxena, A., Chung, S.H., Ng, A.Y.: Learning depth from single monocular images. In: Advances in Neural Information Processing Systems (NIPS), pp. 1161–1168 (2005)
26. Saxena, A., Sun, M., Ng, A.Y.: Make3d: learning 3d scene structure from a single still image. IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI) **31**(5), 824–840 (2009)
27. Sharma, A., Tuzel, O., Jacobs, D.W.: Deep hierarchical parsing for semantic segmentation (2015)
28. Shekhar, S., Patel, V.M., Nasrabadi, N.M., Chellappa, R.: Joint sparse representation for robust multimodal biometrics recognition. Pattern Anal. Mach. Intell. (TPAMI) **36**(1), 113–126 (2014)
29. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from RGBD images. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7576, pp. 746–760. Springer, Heidelberg (2012). doi:10.1007/978-3-642-33715-4_54
30. Silveira, G., Malis, E., Rives, P.: An efficient direct approach to visual slam. IEEE Trans. Robot. **24**(5), 969–979 (2008)
31. Srivastava, N., Salakhutdinov, R.R.: Multimodal learning with deep boltzmann machines. In: Advances in Neural Information Processing Systems (NIPS), pp. 2222–2230 (2012)

32. Wang, S., Zhang, L., Liang, Y., Pan, Q.: Semi-coupled dictionary learning with applications to image super-resolution and photo-sketch synthesis. In: Computer Vision and Pattern Recognition (CVPR), pp. 2216–2223. IEEE (2012)
33. Wang, Y., Cho, S., Wang, J., Chang, S.-F.: Discriminative indexing for probabilistic image patch priors. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8692, pp. 200–214. Springer, Heidelberg (2014). doi:10.1007/978-3-319-10593-2_14
34. Wang, Y., Wu, F., Song, J., Li, X., Zhuang, Y.: Multi-modal mutual topic reinforce modeling for cross-media retrieval. In: International Conference on Multimedia (MM), pp. 307–316. ACM (2014)
35. Xiang, S., Meng, G., Wang, Y., Pan, C., Zhang, C.: Image deblurring with coupled dictionary learning. Int. J. Comput. Vis. (IJCV) 1–24 (2014)
36. Yang, J., Wang, Z., Lin, Z., Cohen, S., Huang, T.: Coupled dictionary training for image super-resolution. IEEE Trans. Image Process. (TIP) **21**(8), 3467–3478 (2012)
37. Zhuang, Y.T., Wang, Y.F., Wu, F., Zhang, Y., Lu, W.M.: Supervised coupled dictionary learning with group structures for multi-modal retrieval. In: Twenty-Seventh AAAI Conference on Artificial Intelligence (AAAI) (2013)