A Parametric Model for Describing the Correlation Between Single Color Images and Depth Maps

Yangang Wang, Ruiping Wang, *Member, IEEE*, and Qionghai Dai, *Senior Member, IEEE*

***Abstract*—This letter introduces a new approach for modeling the correlation between a single color image and its depth map with a set of parameters. The proposed model treats the color image as a set of patches and describes the correlation with a kernel function in a non-linear mapping space. We also present how to estimate the model parameters from sampled color image patches as well as the corresponding depth values. The proposed approach is tested on different color images and experimental results are comparable to the state-of-the-art, which demonstrates the power of the proposed method. Furthermore, we validate the efficiency of the proposed parametric model by evaluating each of its component, including the filters optimization, the choice of the patches and the kernel function.**

***Index Terms*—Color image and depth, filters optimization, kernel function, parametric model.**

I. Introduction

PERCEIVING color images and their corresponding depth maps (which is also known as range images) has attracted lots of attentions due to the great demands of 3D applications such as 3D reconstruction [1] and free view video (FVV) [2]. It can also improve the performance of traditional 2D computer vision tasks such as image enhancing [3], scene classification and recognition [4]. Although several small commercial hardwares (e.g., Xbox 360™ [5], Swissranger™ [6]) now can easily obtain the depth and color images simultaneously, they can not be widely used to capture natural images because of the short distance perception as well as the low resolution of the captured depth, which strongly impedes the further applications. Therefore, exploring the potential correlation between the color images and their corresponding depth maps, thus estimating the depth maps from color images has been an interesting research topic in computer vision.

Furthermore, due to the availability of large number of independently captured color images in recent years, it gains more and more attentions to use a single still color image to estimate the depth map [7], [8], [9]. The possibility of estimating the depth from a single color image is that there exist many depth related visual cues in a single still color image, such as texture gradients, occlusions and shading. One typical depth estimation approach [10] acquires the depth for the target color image by transferring the depths of similar color images from a large scale database. The approach first selects the most similar color images from the database and then transfers their corresponding depth maps to the target image by integrating the scene classification and optical flow techniques. Another kind of approaches [7], [8] try to build a parametric model to describe the correlation between color images and depth maps. Although depth transfer based approaches can always obtain better performance since they can fit specific natural images with a huge database, parametric model based methods do not need to search a large database which makes them computationally more efficient and more flexible in real applications.

In this letter, we focus on the the problem of modeling the correlation between a color image and its depth map. Previous parametric model [11] tries to describe the correlation with several pre-defined features. Its purpose is to find sufficiently good color-depth features, where several types of complex features are used to construct the feature vector, e.g., multi-scale feature and spatial consistency feature. However, it is hard to assess the individual contribution of each feature, and the model thus has to balance the weights of different features. We present a new parametric model which regards a color image as a set of patches and models the correlation between color patches and their depth values with more flexibilities. Different from previous methods, the basic elements of the features, i.e. the filters, can also be optimized in the proposed model. Our model targets to search better parameters for describing the correlation between color images and depth maps.

II. Problem Statement and Model

In this section, we first introduce our proposed model for building the correlation between color images and depth maps. Then we describe the algorithm of estimating the model parameters.

A. The proposed model

Given a color image I and its corresponding depth map D, our purpose is to build the correlation between I and D with a set of parameters. Similar as previous image processing approach [11], we treat the color image as a set of overlapped fixed-size (e.g. 15×15) color patches. In order to avoid the over-fitting problem, we only sample a small set of color

Manuscript received July 18, 2013; accepted September 21, 2013. Date of publication September 27, 2013; date of current version April 22, 2014. Part of this work was performed while R. Wang was a Postdoc at Tsinghua University. This work was supported by the NSFC under Grants 61035002, 60932007, and U0935001. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Xiaokang Yang.

Y. Wang and Q. Dai are with the Department of Automation, Tsinghua National Laboratory for Information Science and Technology (TNList), Tsinghua University, Beijing 100084, China (e-mail: ygwang.thu@gmail.com; qionghaidai@tsinghua.edu.cn).

R. Wang is with the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China (e-mail: wangruiping@ict.ac.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LSP.2013.2283851

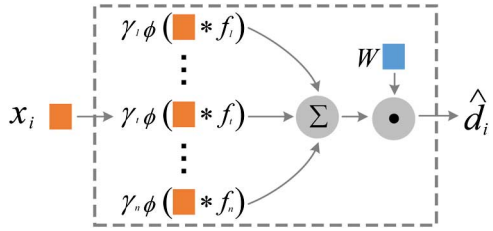


Fig. 1. The pipeline of the proposed model. For each color patch x_i, we sum all the convoluted color patches and compute the dot product with weighted matrix W to obtain the corresponding depth values.

patches $\{x_1, x_2, \ldots, x_p\}$ and their corresponding depth values $\{d_1, d_2, \ldots, d_p\}$ from image I and depth D respectively, where p is the number of samplers. The column vector $\mathbf{d}$ is used to denote the corresponding depth values of all the sampled color patches, e.g, $\mathbf{d} = [d_1, d_2, \ldots, d_p]^T$.

We model the correlation of color images and depth maps by measuring the sum squared errors between the color patches mapped depth values $\hat{\mathbf{d}}$ and the real depth values $\mathbf{d}$. The procedure of mapping the color patches is as follows (also shown in Fig. 1). Suppose we have a matrix $F = [f_1, f_2, \ldots, f_n]$, where each column is an individual filter and n is the number of filters. For each color patch x_i, we use a kernel function $\phi(\cdot)$ to map the n convoluted color patches and sum them up to one "color patch" by multiplying parameters γ_i for each of the convoluted patches. The column vector $\gamma = [\gamma_1, \gamma_2, \ldots, \gamma_n]^T$ encodes the variances of the convolution results from different filters. A weight matrix W, which has the same size as the convoluted patches, sums all the elements of the "color patch", and thus obtains the mapped patch corresponding "depth value". Therefore, our model is written as

$$E = \sum_{i=1}^{p} |tr(W^T \sum_{j=1}^{n} \gamma_j \phi(x_i * f_j)) - d_i|^2. \quad (1)$$

In Eq. (1), E is the estimation error and W, F, γ are the parameters of the proposed model. We could also rewrite Eq. (1) as

$$E = \sum_{i=1}^{p} |\mathbf{w}^T \phi(X_i F)\gamma - d_i|^2, \quad (2)$$

where X_i is a matrix reshaped from color patch x_i, each row of X_i is a filter-sized row vector; $\mathbf{w}$ is a column vector obtained by concatenating all the entries of the weight matrix W.

We will explain how to estimate them in the next subsection. In the following text, we first give further discussions about these parameters. After that, the strategy of kernel function $\phi(\cdot)$ involved in the proposed model is described.

The filters F extract the specific frequency information which indeed describes the texture gradient cues of the color images. Its accompaniment parameter γ is used to describe the variance of different filters for avoiding one filter to dominate the results. A good initialization of all the filters can be obtained by Principle Component Analysis (PCA) or Independent Component Analysis (ICA). After initialization, estimating the filters will be more effective as the experiments will demonstrate afterwards.

Another issue is about the size of filters. In general, small sized filters (e.g., 3×3) could accelerate the estimating procedure while considering less texture neighboring information, and large sized filters (e.g., 7×7) encode more neighboring information of color images but is time consuming. In our model, the existence of the weight matrix W makes it more feasible to use small sized filters since the matrix integrates the overall information from each color patch.

The kernel function $\phi(\cdot)$ maps the convolution results of the color image patches with the filters to be comparable with the ground truth depth values. In fact, how to define the explicit form of the kernel function $\phi(\cdot)$ is very important to the estimation accuracy of the proposed model. We have evaluated several different kernel functions and found that the simple but effective kernel function is $\phi(x) = \log(1+x^2)$. Details will be presented in Sec. III.

B. Model parameters estimation

Estimating the parameters of the proposed model is not trivial because of not only the number of parameters but also the existing non-linear kernel function. A natural approach to estimate the parameters is to alternate among all the variables, minimizing over one while keeping the others fixed, as proposed by Mairal *et al.* [12].

For the convenience of deducing the gradient of the parameters in E, we rewrite Eq. (2) in the following two matrix forms:

$$E = \|M\phi(XF)\gamma - \mathbf{d}\|_2^2, \quad (3)$$

or

$$E = \|\Gamma\phi(F^T\hat{X})\mathbf{w} - \mathbf{d}\|_2^2. \quad (4)$$

In Eq. (3), X is obtained by concatenating all the X_i; each row of M is $\mathbf{w}^T$. In Eq. (4), $\hat{X}$ is obtained by concatenating all the X_i^T; each row of Γ is γ^T. We can see Eq. (3) and Eq. (4) are respectively the least square problems of γ and $\mathbf{w}$, thus it is very easy to obtain the closed form solutions for updating the parameters γ and $\mathbf{w}$.

As for updating the filters F, we could compute the gradient of Eq. (1) and Eq. (3) w.r.t. the filter f_i at the t-th iteration, i.e. $f_{i,t}$, as

$$\frac{\partial E}{\partial f_{i,t}} = \gamma_i X^T J(Xf_{i,t}) M^T \left(\sum_{j=1}^{n} \gamma_j M\phi(Xf_j) - \mathbf{d} \right). \quad (5)$$

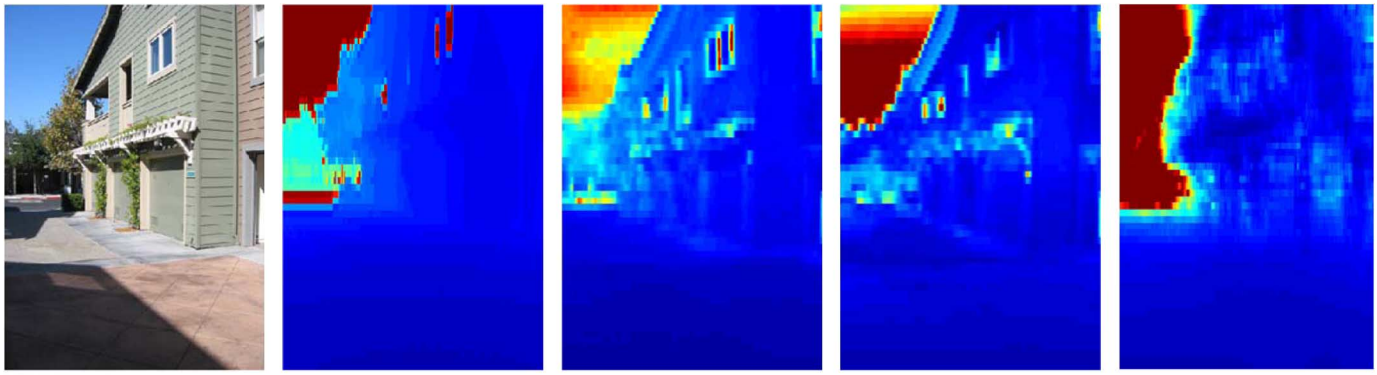
$J(Xf_{i,t})$ is the Jacobian matrix of vector $\phi(Xf_{i,t})$, it is a square matrix spanned by the gradient vector, i.e., $\phi'(Xf_{i,t})$.

Since $\phi(\cdot)$ is a non-linear kernel function, we use Taylor expansion of matrix $\phi(Xf_i)$ at filter $f_{i,t}$ and obtain

$$\phi(Xf_i) = \phi(Xf_{i,t}) + J(Xf_{i,t})X(f_i - f_{i,t}). \quad (6)$$

Let $L_{i,t} = MJ(Xf_{i,t})X$ be a matrix storing the triple-matrix product and we substitute it to Eq. (5). With the combination of Eq. (6), we can obtain the gradient of the proposed model w.r.t. the filter $f_{i,t}$ as

$$\frac{\partial E}{\partial f_{i,t}} = \gamma_i L_{i,t}^T \begin{bmatrix} \gamma_i L_{i,t} f_i + \sum_{j=1, j\neq i}^{n} \gamma_j L_{i,t} f_j - \mathbf{d} \\ + \sum_{j=1}^{n} (\gamma_j M\phi(Xf_{j,t}) - \gamma_j L_{j,t} f_{j,t}) \end{bmatrix}. \quad (7)$$



(a) (b) (c) (d) (e)

Fig. 2. Validation of the proposed parametric model: (a) the original color image; (b) the ground truth depth map; (c) the depth reconstruction result using 2000 randomly sampled pairs of color patches and depth values ($\mathrm{RE} = 0.3071, \mathrm{LE} = 0.1285$); (d) the depth estimation result with the parameters estimated from the same class images ($\mathrm{RE} = 0.3059, \mathrm{LE} = 0.1268$); (e) the depth estimation result with the model based approach proposed in [11] ($\mathrm{RE} = 0.5242, \mathrm{LE} = 0.1565$).

For convenience, we use the symbol $G_{i,t}$ and $K_{i,t}$ to represent

$$G_{i,t} \triangleq \gamma_i^2 L_{i,t}^T L_{i,t},$$
$$K_{i,t} \triangleq \gamma_i L_{i,t}^T \begin{pmatrix} \sum_{j=1, j\neq i}^{n} \gamma_j L_{i,t} f_j + \\ \sum_{j=1}^{n} \gamma_j M \phi(X f_{j,t}) \\ - \sum_{j=1}^{n} \gamma_j L_{j,t} f_{j,t} \end{pmatrix} - \gamma_i L_{i,t}^T \mathbf{d}. \quad (8)$$

It is noted that the gradients of all the filters should be $\mathbf{0}$ when they reach the minimum. We can obtain closed form solutions of the filters, that is, each of the filters $F = [f_1, \ldots, f_i, \ldots, f_n]$ should satisfy the equation $G_{i,t} f_i + K_{i,t} = 0$. We find that it is hard to get stable solution to solve this linear equation directly. On the contrary, we use the warm-start gradient descent method for the filters optimization in our experiments.

III. Experiments and Evaluations

In this section, we present several experiments to evaluate the proposed approach for reconstructing the depth maps with the estimated parameters.[1]All the color images and corresponding depth maps are used from [11]. In order to validate the efficiency of our method, we conducted the depth reconstruction comparisons against the subspace filter optimization, the sampling strategies of patches according with depth values and the different kernel functions selection.

We use two methods for the quantity comparisons. Denoting the reconstructed depth map as $\hat{d}$ and the ground truth depth map as d, we compute the relative error (**RE**) with $1/p * \sum_i |\hat{d}(i) - d(i)|/d(i)$; the log10 error (**LE**) with $1/p * \sum_i \log_{10}(\hat{d}(i)/d(i))$, where p is the number of pixels.

A. Validation of the proposed parametric model

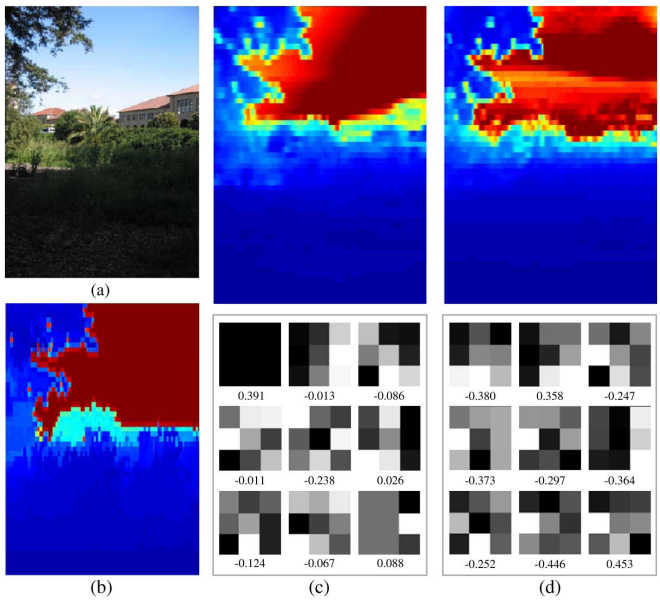
In order to validate the proposed parametric model, we estimate the parameters for a single color image using the ground truth depth map. We randomly select 2000 pairs of color patches and depth values, which is about 0.52‰ of the total pairs in one single color image (the size of the original color image is 2272×1704 and the overlapped patch size is 15×15). The estimated parameters are used to reconstruct its depth map. Furthermore, we compared the result with the depth estimation result by using the estimated parameters from the same class images. This is often useful for the single color image depth estimation applications, which we call **parameters transfer**. Compared against the previous model based approach [11], we find that the proposed model can obtain better performance. Fig. 2 shows the comparison result. From Fig. 2, we also can see: 1) the parametric reconstruction (Fig. 2(c)) result is a little different from the ground truth depth map. It means that the original color patches do not completely accord with the statistics revealed by the parameters; 2) (Fig. 2(c)) and (Fig. 2(d)) have almost the same estimation results, which indicates that using the parameters of the similar class images is suitable for depth estimation problem.

B. Validation of key components of our method

The filters optimization. To our knowledge, all existing parametric models for depth estimation have fixed the filters in the feature selection phase. In order to demonstrate the necessity of optimizing the filters, we evaluate the effectiveness of the filters optimization. All the filters are optimized in the subspace, that is, they are described as $F = B\tilde{F}$ where $\tilde{F}$ is the filters coefficients in the subspace basis B. We initialize $\tilde{F}$ as an identity matrix. Fig. 3 shows the evaluation results without/with filters optimization as well as the selected filters. Compared against Fig. 3(b) and Fig. 3(c), filters optimization can outperform the estimation. Besides, the optimized γ parameters correspond to the fixed non-optimized filters indicate that the average values of images (i.e., luminance) have the dominant impact in the estimation procedure.

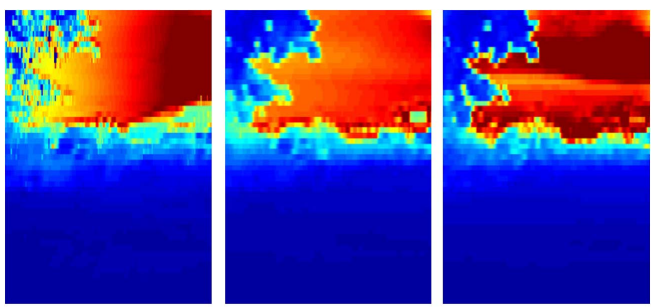
The selection of pairs. In the parameters optimization procedure, we found that different sampling approaches for choosing the pairs of color image patches and depth values made the optimized parameters converge to different results. We tested three sampling methods: one is to sample the pairs randomly from the whole image; another one considers the sampling noise and rejects the pairs that contradict with the luminance statistic curve of the whole image; the third one separates the image into different grids and sample the pairs uniformly from different grids.

[1]The code for all the experiments is available online athttp://media.au.tsinghua.edu.cn/ygwang/spl2013depth.jsp.



(a) (b) (c) (d)

Fig. 3. Evaluation on the filters optimization. (a) is the original color image; (b) is the ground truth depth map; (c) shows the estimation result (top row) by fixed filters ($\mathrm{RE} = 0.4914$ $\mathrm{LE} = 0.1893$), the selected filters and corresponding γ (bottom row); (d) shows the estimated depth map (top row) with optimized filters ($\mathrm{RE} = 0.4317$ $\mathrm{LE} = 0.1563$), the selected filters and corresponding γ (bottom row).

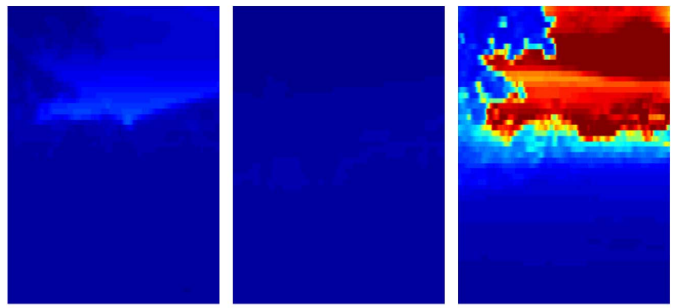


(a) (b) (c)

Fig. 4. The depth estimation results with different pairs selections methods: (a) pairs are randomly sampled from the whole image ($\mathrm{RE} = 0.8265$, $\mathrm{LE} = 0.2311$); (b) pairs are randomly sampled from the whole image considering the luminance statistic curve for rejection ($\mathrm{RE} = 0.6844$, $\mathrm{LE} = 0.1835$); (c) pairs are uniformly sampled from the pre-defined grids ($\mathrm{RE} = 0.4317$ $\mathrm{LE} = 0.1563$).

Fig. 4 shows the depth estimation results with different converged parameters. In order to obtain good performance of our model, it is better to separate the image into grids and sample the pairs uniformly in each grid.

The kernel function. We also conducted the evaluation on the kernel functions. We explored the kernel functions used in many computer vision problems [13] and three different kernel function forms which have similar shapes are tested. They are: $\phi_1(x) = \sqrt{|x|}$, $\phi_2(x) = x^2$ and $\phi_3(x) = \log(1 + x^2)$. The estimated depth maps in Fig. 5 demonstrate the effectiveness of the function (i.e., ϕ_3) finally used exploited in our proposed model. From the figure, we can see that ϕ_1 and ϕ_2 have the similar results but they fail the estimation. The reason may lie in that the gradient of ϕ_1 is sensitive to the small x and the gradient of ϕ_2 can not converge for large x.



(a) (b) (c)

Fig. 5. Evaluation on the kernel functions: (a) is the estimation result with ϕ_1 ($\mathrm{RE} = 1.0843$, $\mathrm{LE} = 0.7673$), (b) is the estimation result with ϕ_2 ($\mathrm{RE} = 1.2658$, $\mathrm{LE} = 0.8711$) and (c) is the estimation result with ϕ_3 ($\mathrm{RE} = 0.4317$ $\mathrm{LE} = 0.1563$).

IV. Conclusion

In this letter, we propose a new parametric model for constructing the relationship between color images and depth maps. We also present algorithms for estimating the parameters. The capability and effectiveness of our approach are demonstrated by reconstructing the depth maps in several testing images. With the optimized filters, our model outperforms previous parametric models. In the future work, we would like to analyze these parameters, thus to understand the inherent connections between color images and depth maps flexibly.

References

[1] S. Izadi, D. Kim, O. Hilliges, D. Molyneaux, R. Newcombe, P. Kohli, J. Shotton, S. Hodges, D. Freeman, and A. Davison, "Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera," in *Proc. 24th Annu. ACM Symp. on User Interface Software and Technology*, 2011, pp. 559–568.

[2] Q. Liu, Y. Yang, R. Ji, Y. Gao, and L. Yu, "Cross-view down/up-sampling method for multiview depth video coding," *IEEE Signal Process. Lett.*, vol. 19, no. 5, pp. 295–298, 2012.

[3] F. Li, J. Yu, and J. Chai, "A hybrid camera for motion deblurring and depth map super-resolution," in *IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2008.

[4] A. Torralba and A. Oliva, "Depth estimation from image structure," *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 24, no. 9, pp. 1226–1238, 2002.

[5] "Microsoft, xbox 360," [Online]. Available: http://www.xbox.com/.

[6] "Mesa imaging, swissranger," [Online]. Available: http://www.mesa-imaging.ch/.

[7] E. Delage, H. Lee, and A. Y. Ng, "A dynamic bayesian network model for autonomous 3d reconstruction from a single indoor image," in *IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2006.

[8] B. Liu, S. Gould, and D. Koller, "Single image depth estimation from predicted semantic labels," in *IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2010.

[9] C. Su, L. Cormack, and A. Bovik, "Color and depth priors in natural images," *IEEE Trans. Image Process.*, vol. 22, no. 6, pp. 2259–2274, 2013.

[10] K. Karsch, C. Liu, and S. B. Kang, "Depth extraction from video using non-parametric sampling," in *Eur. Conf. Computer Vision (ECCV)*, 2012.

[11] A. Saxena, S. Chung, and A. Ng, "3-d depth reconstruction from a single still image," *Int. J. Comput. Vis.*, vol. 76, no. 1, pp. 53–69, 2008.

[12] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online dictionary learning for sparse coding," in *Int. Conf. Machine Learning (ICML)*, 2009.

[13] R. Szeliski, *Computer Vision: Algorithms and Applications*. Berlin, Germany: Springer, 2010.