Learning the Missing Values in Depth Maps

Xuanwu Yin[a], Guijin Wang*[a], Chun Zhang[a], Qingmin Liao[a]
[a]Dept. of Electronic Engineering, Tsinghua University, Beijing, China 100084

ABSTRACT

In this paper, we consider the task of hole filling in depth maps, with the help of an associated color image. We take a supervised learning approach to solve this problem. The model is learnt from the training set, which contain pixels that have depth values. Then we apply supervised learning to predict the depth values in the holes. Our model uses a regional Markov Random Field (MRF) that incorporates multiscale absolute and relative features (computed from the color image), and models depths not only at individual points but also between adjacent points. The experiments show that the proposed approach is able to recover fairly accurate depth values and achieve a high quality depth map.

Keywords: depth refine, hole filling, MRF, supervised learning

1. INTRODUCTION

Recovering 3D depth information from images is a basic problem in computer vision, and has applications such as automatic driving, biomedical imaging, scene understanding and human-computer interaction.

Several approaches to do depth sensing have emerged over time. We can divide such systems into active and passive, according to the use or lack of controlled illumination. Passive approaches can exploit human vision depth cues such as stereo [1], which uses the disparity of object position in two images with different known viewpoints, and defocus, which uses the projected size of the optical pupil in the image plane or the spatial frequency content of the image to estimate depth. Active approaches [2] are developed by researchers to make depth estimation more robust in textureless regions and under variational illumination conditions. Some commercial products have been introduced, such as TOF (Time of Flight) cameras and Microsoft Kinect.

However, none of the techniques mentioned above can provide perfect depth data. There are always noises contained in the depth map, such as no-measured depth pixels, inaccurate object boundaries and random measurement error. These noises make the depth map incomplete and hard to use. Most of the time, the noisy depth data is used directly by applications, such as pose recognition and shape measurement, and affects the accuracy greatly. So the depth map is necessary to be preprocessed before being used.

Several methods have been developed to refine the depth maps. Directional 3D propagating method treats depth hole filling for RGBD images with a depth map that has different resolution from a color image of the same scene [3]. This method warps and up-samples the low-resolution depth image to have the same viewpoint and resolution as the color image. A depth-color cross-bilateral filtering is applied to interpolated holes in the up-sampled images. Another fusion based depth hole filling method uses motion compensated frames to temporal filtering and interpolates depth holes using color information [4]. In [5], adaptive cross-trilateral median filtering is introduced to improve depth maps. Camplani [6] presents a joint-bilateral filtering framework to inpaint the deph maps. An adaptive spatio-temporal filter [7] is presented to improve the accuracy and stability of Kinect depth. Yu [8] divided the holes into three types and focused on shadow removal. Yang [9] estimated the missing depth values using the depth distribution around depth holes. Besides, learning based depth estimation approach [10] has attracted more attention in recent years.

In this paper, a hole filling approach based on supervised learning is proposed, the associated color image is used to recover the missing values in the depth maps. We select the points that have depth values as the training set to train the parameters of our model. The missing values in the holes are predicted based on our regional MRF model. The experimental results show that the predicted values are both reasonable and reliable.

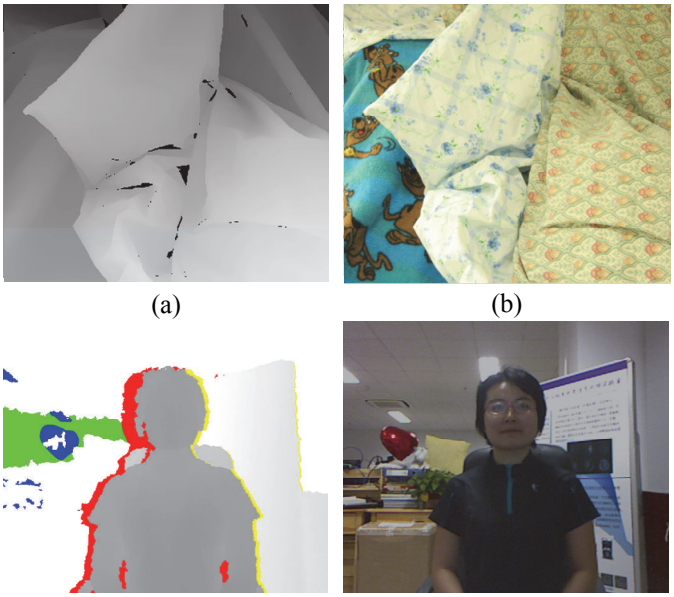
*wangguijin@tsinghua.edu.cn; phone +86-010-62781430; image.ee.tsinghua.edu.cn

2013 International Conference on Optical Instruments and Technology: Optoelectronic Imaging and Processing Technology, edited by Xinggang Lin, Jesse Zheng, Proc. of SPIE Vol. 9045, 904508

This paper is organized as follows: the cause of missing values and the learning based framework is introduced in section 2. The features used are described in section 3, and the MRF model is presented in section 4. Experimental results are presented in section 5. Finally, the conclusion and future works are discussed in section 6.

2. HOLE TYPES AND ALGORITHM FRAMEWORK

2.1 Cause of depth holes



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

Fig. 1. (a) & (b) Stereo's disparity map and its associated color image. (c) & (d) Active approach's depth map and its associated color image.

Fig.1 shows some sample depth maps. (a) is the disparity map obtained by stereo matching, which can be transformed into depth by triangulation easily, and (b) is the color image associated with (a). The holes in stereo are mostly caused by occlusion. Because of the difference in viewpoints, there are always mismatches between stereo pairs, leading to some missing pixels whose disparity (depth) cannot be estimated.

The holes in active approach's depth maps, see Fig.1(c) & (d), are caused by various factors. In this paper, we divide the causes into four types.

- Occlusion;
- Out-of-range;
- Non-Lambertian surface;
- Warping.

The four types above are illustrated in Fig.1(c) by red, green, blue and yellow. According to [2], the active depth sensing device is a structured sensor consisting of one infrared laser emitter, and one infrared camera. One other RGB camera is needed in order to obtain the associated color image. Occlusion is somewhat similar to that in stereo, which means that the objects prevent the infrared pattern from being projected to the background. If the reflected pattern is too weak(out-of-range) or there is little reflection(non-Lambertian surfaces), the depths can hardly be measured and holes are left. Otherwise, if you try to warp the depth image to make it have the same viewpoint with the color image, holes will occur in the warped depth image. The TOF cameras have similar defects with the active depth sensing approach except occlusion.

2.2 Algorithm framework

Although holes occur in the depth maps inevitably, the associated color images are always complete. Therefore we try to fill these holes by using the information of the color image. Firstly, we detect the holes in the depth map and take all the

pixels that have depth values as the training set. Meanwhile features are computed based on the color image. Then the model parameters are trained from the training set. Finally, the missing values in the depth map is predicted by our MRF model.

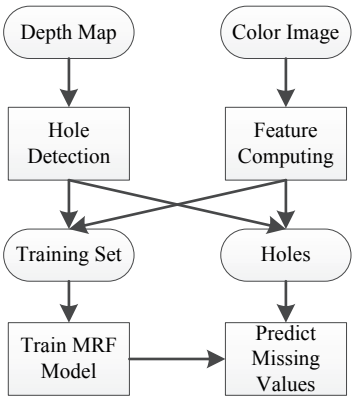


Fig. 2. Flowchart of depth hole filling.

3. FEATURES

Humans can judge depth from images by using cues such as texture variations, texture gradients, occlusion, known object sizes, haze, defocus, etc. Some of them are more suitable for judging relative depths under long range conditions, such as known object sizes, haze and defocus. Others are hard to be determined, such as occlusion. Since we are trying to predict the missing values from just the same image, simple features are preferred. We choose features that capture three types of local cues: texture variations, texture gradients, and color similarity.

3.1 Features for absolute depth

First, the color image is transformed from RGB into YCbCr color space, where Y represents the intensity channel, and Cb and Cr are the color channels. Information about textures is mostly contained in the variation of intensity. For each pixel, we apply Law's masks to measure texture energies. The nine Law's masks $M_1, \ldots, M_9$ are obtained by multiplying together pairs of three 1×3 masks, Fig.3. The texture variations and gradients can be expressed by applying the nine masks to the Y channel. Color similarity is captured by applying a local averaging filter (the first Law's mask) to the color channels.

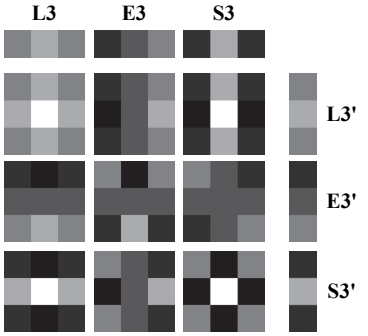


Fig. 3. Law's masks for texture energy. The one dimensional masks (Gaussian averaging L3, edge detection E3 and spot detection S3) are used to obtain nine two dimensional masks M_n.

For some pixel i in the image $I(x, y)$, the summary statistics is computed as follows. We use the output of each of the 11 (9 Law's texture masks on Y channel and 2 local average masks on color channels) filters $M_n(x, y)$, $n = 1, \ldots, 11$ as: $E_i(n) = \sum_{(x,y)\in N(i)} \left| I(x, y) * M_n(x, y) \right|^k$,where $k = \{1,2\}$ give the sum absolute energy and sum squared

energy respectively, where $N(i)$ is the smoothing neighborhood of i. This gives an initial feature vector of dimension 22.

To estimate the absolute depth, local features are insufficient, more global properties should be taken into account. We try to capture this information by using features extracted with different sizes of smoothing neighborhood. $N(i)$ is set to 5×5, 12×12 and 32×32 square window for different scales. Finally we get a $3 \times 22 = 66$ dimension absolute feature vector.

3.2 Features for relative depth

A different feature vector is used to model the dependencies between neighboring pixels. We use a full histogram (with 8 bins) for each filter output $\left|I(x, y) * M_n(x, y)\right|$ instead of computing summary statistics, giving a feature y_i of dimension 88. Our goal is to model the relationship between the depths of adjacent pixels. Hence we use the difference of the histograms between two neighboring pixels $y_{ij} = y_i - y_j$ as our relative feature.

4. PROBABILISTIC MODEL

We choose jointly Gaussian MRF as our optimization model. The posterior probability is given by the feature vectors and model parameters is as follows

$$P(d \mid X; \theta, \sigma) = \frac{1}{Z} \exp\left(- \sum_{i=1}^{K} \left(\frac{(d_i - f(x_i))^2}{2\sigma_i^2} + \sum_{j \in A(i)} \frac{(d_i - d_j)^2}{2\sigma_{ij}^2} \right) \right) \quad (1)$$

Where K is the total number of pixels in the image; x_i is the absolute feature vector; σ is the model parameters; $A(i)$ is the 8-neighborhood of i; Z is the normalization constant for the model; and $f(x_i)$ is a function that predicts the depth value with the absolute feature alone. In this paper, SVM is used as f. The SVM is trained with all the pixels that have depth values in the image.

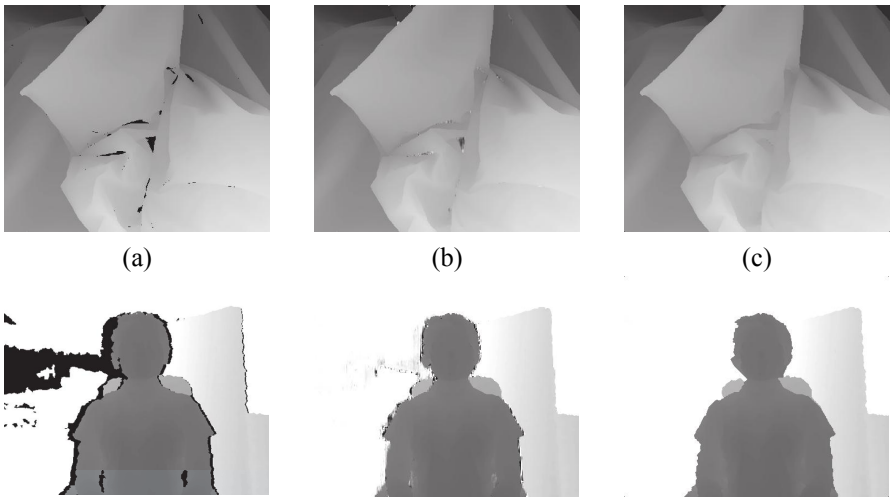
In practice, the variance term should be different for different images, we model the "variance" term $\sigma_i^2 = u^T x_i$ as a linear function of the absolute features. We determine u by fitting σ_i^2 to $(d_i - f(x_i))^2$ with constraint of $u \geq 0$, to keep σ_i^2 positive. This σ_i^2 gives a measure of uncertaintity of the first term, dependent on the absolute features. This is motivated by the observation that in some regions, depth cannot be estimated from local features reliably. Thus one has to rely more on neighboring pixels to estimate depth.

Similar to σ_i^2, we also model the variance parameter σ_{ij}^2 as a linear function of the pixels i and j's relative depth features y_{ij} as $\sigma_{ij}^2 = v^T \left| y_{ij} \right|$. This σ_{ij}^2 can help to determine which neighboring pixels to have similar depths. The smoothing effect is stronger if neighboring pixels are similar. v is chosen to fit σ_{ij}^2 to $(d_i - d_j)^2$ with constraint of $v \geq 0$, to keep σ_{ij}^2 positive.

After learning the parameters, we can obtain the MAP estimation of the depths by maximizing Eq.1 in terms of d. Since the model is Gaussian, the negative logarithm of Eq.1, which is often called energy, is quadratic in d, thus MAP estimation can be easily found in a closed form, by solving a set of linear equations. Although MRF is often treated as a global optimization model, we use it regionally in this paper. Since we focus on predicting the missing values instead of refining the whole depth map, we build a MRF model for each hole individually. This approach can greatly reduce computing complexity.

5. EXPERIMENTAL RESULTS

In order to evaluate the performance, we test our algorithm on depth maps acquired both by stereo matching and active depth sensing approach presented in [2]. The algorithm is implemented in Matlab and run on a desktop computer equipped with 2.50GHz Q9300 CPU and 4GB RAM. We don't have to set the parameters since all the parameters in this paper are computed adaptively with the input color image and depth map.



(a) (b) (c)

(d) (e) (f)

Fig. 4. Example results. Left: original input coarse depth map. Middle: hole filling with the absolute features alone. Right: hole filling by optimizing the MRF model.

Fig.4 shows two examples of depth hole filling. The top row is the sample depth map obtained by stereo matching, and the bottom row by active depth sensing. It can be seen from the left column that both these two depth maps contain holes, and there are more holes in the bottom depth map than the top one. The middle column presents the results by predicting the depth values with absolute features alone. In this case, the second term in Eq.1 is ignored. We can see that some of the values predicted by SVM are reasonable, while others remain unreliable. The right column shows the results when the entire model is used. It can be seen that the most of the predicted values are reasonable when the relative features are taken into account. More results are presents in Fig.6.

(a) (b) (c)

Fig. 5. Evaluation on manually corrupted depth map. (a) depth map with manmade holes (black squares); (b) hole filling with our method; (c) hole filling with the method described in [9].

We also test our algorithm on manually corrupted depth maps and compare our approach with the method in [9]. One example is presented in Fig.5. It can be seen that even though the holes pass through the boundaries, the proposed approach can achieve reasonable results. While the results in cross-boundary regions by [9] are unreasonable because the holes are assumed to always occur at rear objects. In addition, we test our approach on 100 manually corrupted depth maps. Over 80% of the predicted values are correct when the threshold is set to1 in intensity. If we loosen the threshold to 2, the accuracy will be over 87%.

The proposed algorithm is also very efficient because of the regional MRF optimization method. Most of the time, the computation can be completed within several seconds.

6. CONCLUSION

In this paper, we propose a learning based approach to fill the holes in depth maps. The absolute and relative features are firstly computed with the input color image. Then we build a MRF model for each hole individually. The parameters of the MRFs are determined by supervised learning with the pixels whose depth values are known. Finally, the missing depth values can be predicted by the MAP estimation of these MRFs. The experiments show that the predicted depth values are considerably reasonable, and our proposed approach can fill the holes in depth maps. The enhancement of depth maps can bring huge benefits to 3D vision applications, such as 3D segmentation, 3D object recognition and tracking, and human-computer interaction. This work is Partially supported by NSFC 61271390.

Fig. 6. More depth hole filling results. Top row: input color image. Middle row: input coarse depth map with holes. Bottom row: depth map after hole filling.

REFERENCES

[1] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," International journal of computer vision, vol. 47, 7-42 (2002).

[2] G. Wang, X. Yin, X. Pei, and C. Shi, "Depth estimation for speckle projection system using progressive reliable points growing matching," Applied optics, vol. 52, 516-524 (2013).

[3] Q. H. Nguyen, M. N. Do, and S. J. Patel, "Depth image-based rendering from multiple cameras with 3D propagation algorithm," in Proc. 2nd Int. Conf. Immersive Telecommunications, Berkelei, CA, 6-12 (2009).

[4] S. Matyunin, D. Vatolin, Y. Berdnikov, and M. Smirnov, "Temporal filtering for depth maps generated by Kinect depth camera," in Proc. 3DTV Conference: The True Vision-Capture, Transmission and Display of 3D Video, Antalya, Turkey, 1-4 (2011).

[5] M. Mueller, F. Zilly, and P. Kauff, "Adaptive cross-trilateral depth map filtering," in 3DTV-Conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON) 2010, 1-4 (2010).

[6] M. Camplani and L. Salgado, "Efficient spatio-temporal hole filling strategy for kinect depth maps," in IS&T/SPIE Electronic Imaging, 82900E-82900E-10 (2012).

[7] M. Camplani and L. Salgado, "Adaptive spatio-temporal filter for low-cost camera depth maps," in Emerging Signal Processing Applications (ESPA), 2012 IEEE International Conference on, 33-36 (2012).

[8] Y. Yu, Y. Song, Y. Zhang, and S. Wen, "A shadow repair approach for kinect depth maps," in Computer Vision–ACCV 2012, ed: Springer 2013, 615-626 (2013).

[9] N.-E. Yang, Y.-G. Kim, and R.-H. Park, "Depth hole filling using the depth distribution of neighboring regions of depth holes in the Kinect sensor," in Signal Processing, Communication and Computing (ICSPCC), 2012 IEEE International Conference on, 658-661 (2012).

[10] A. Saxena, J. Schulte, and A. Y. Ng, "Depth estimation using monocular and stereo cues," in Proceedings of the 20th international joint conference on Artifical intelligence, 2197-2203 (2007).