A novel sparse-to-dense depth map generation framework for monocular videos

Runze Zhang*[a], Zhiguo Cao[a], Qian Zhang[b], Yang Xiao[a], Ruibo Li[a]
[a]National Key Lab of Sci. & Tech. on Multispectral Information Processing, School of Automation, Huazhong Univ. of Sci.& Tech. Wuhan
[b]School of Resources and Environmental Science, Hubei University. Wuhan

ABSTRACT

Generally, the process of monocular depth map generation consists of two stages: structure from motion and multi-view stereo. The multi-view stereo relies on the accuracy of the estimated camera pose and the photo-consistency assumption. However, the current methods cannot tackle the multi-view matching problem well because of the dependence on the accurate camera pose as well as the matching uncertainty. In this paper, to handle these issues, a new sparse-to-dense diffusion framework is put forward. First, the scene information is reconstructed from SFM (Structure From Motion) and the sparse point cloud is available instead of the camera pose. Secondly, the sparse depth point is re-projected to every frame as the depth label. Finally, the depth label is spread to the remaining pixels through a diffusion process. In addition, the edge detectors are used to make the propagation better-regulated. Experimentally, the results show that the proposed framework can robustly generate the depth map from monocular videos.

Keywords: Structure From Motion, Colorization, Sparse Depth Points, Edge detector

1. INTRODUCTION

Depth map is widely used with wide-ranging applications including 3D modelling, 3DTV and virtual reality, to name just a few. To recover the depth map from monocular videos, methods based on multi-view geometry are vogue for their effectiveness in view-changing videos. Even though current methods [12],[13],[14] can estimate the camera pose, the qualification of the depth map is hard to guarantee because of the contradiction between the camera pose and the photo-consistency assumption[4], which greatly relies on the estimated camera pose.

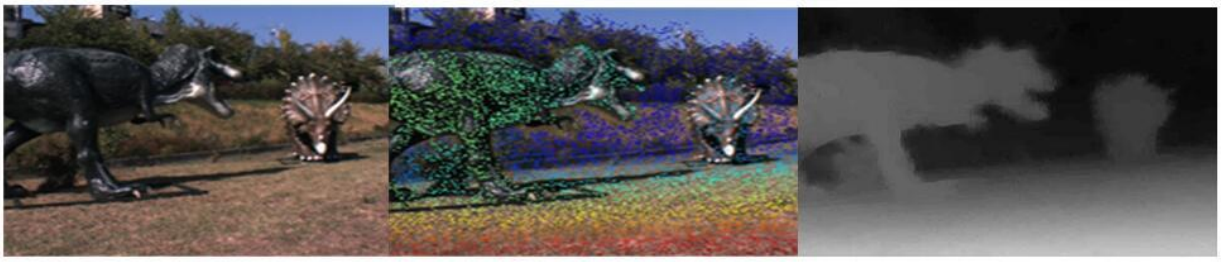
Generally, the process of monocular depth map generation consists of two stages: structure from motion[8](SFM) and multi-view stereo[4](MVS). SFM estimates camera poses and reconstructs the scene. MVS relies on the accuracy of the estimated camera pose and the photo-consistency assumption. Traditionally, local-based methods[2] utilize the winner-takes-all strategy to generate the depth map while global-based methods, such as graph-cut[3] or loopy belief propagation[3], treat this problem as optimizing an energy function. However, both of them cannot tackle the multi-view matching problem well because the accurate camera pose is necessary as well as the satisfying photo-consistency assumption.

In this paper, we propose a new framework to solve the problem. To overcome the inaccuracy of camera pose of some certain frames, we employ the reconstructed scene information from SFM rather than the camera pose. SFM can reconstruct the scene and generate the sparse point clouds. Then we reproject point clouds of the scene to every frame and obtain sparse depth points in every view. As far as we concern, point clouds are able to guide the procedure of generating the depth map. Therefore, we use the single-view diffusion in place of the multi-view matching. Inspired by the colorization[9] insights, we use sparse depth points in place of interactive color scribbles. According to the assumption that nearby pixels in space in a frame with similar colors would have similar depths, we can construct relevant energy function with a closed-form solution. Finally, we propagate the depth to the remaining pixels. In addition, we use the edge detectors[10] to make the propagation better-regulated. Figure 1 shows the baseline of our method without the edge detection. Figure 2 show that detected edges can improve the performance in certain cases.

*Further author information: (Send correspondence to Runze Zhang)
Runze Zhang: E-mail: zrzsgsg@163.com

Automated Visual Inspection and Machine Vision II, edited by Jürgen Beyerer, Fernando Puente León, Proc. of SPIE Vol. 10334, 103340H · © 2017 SPIE
CCC code: 0277-786X/17/$18 · doi: 10.1117/12.2269885

The remaining of this paper is organized as follows. The related work is discussed in Sec. II. Then the sparse-to-dense depth map framework is illustrated in Sec. III. Experiments and discussions are conducted in Sec. IV. Sec. V concludes the whole paper.



(a) (b) (c)

Figure 1. The baseline of the proposed method without edge guidance

(a) The original image. (b) The sparse depth points added to the original image, while different colors mean different depths. (c) The generated depth map through energy function solution.

2. RELATED WORKS

Since our framework consists of the narrow-baseline structure from motion and the colorization-based diffusion, we will firstly discuss the SFM & MVS relevant works and then discuss the Colorization section.

Structure From Motion& Multi-view stereo. Generally, the process of monocular depth map generation consists of two stages: structure from motion and multi-view stereo.

Structure From Motion The structure from motion estimates camera poses and reconstructs the scene. SFM [14][12][13] has been proved successfully to recover the camera pose and reconstruct the scene for unordered images. For videos, Zhang[11] put forward a robust approach for long video sequence with various focal lengths. Fisher Yu and David Gallup[15] use small motion assumption to estimate the camera pose and obtain good results for small-motion videos.

Multi-view stereo The multi-view stereo relies on the accuracy of the estimated camera pose and the photo-consistency assumption. The existing methods of multi-view stereo methods can be categorized into two main groups: local-based methods and global-based methods. Local-based methods[1],[2], like stereo matching, used the local region information to compute the cost volume and utilize the winner-takes-all strategy to generate the depth map. Hyowon Ha[7] uses the plane-sweep[17] to obtain the initial disparity map and employs the minimum spanning tree[16] to obtain the final result. However, the matching inaccuracy could lead to the depth incontinuity. Global-based methods[3],[4],[5], such as graph-cut or loopy belief propagation, treated depth map generation problem as optimizing an energy function. Zhang[6] put forward a coarse-to-fine scheme by first initializing the disparity maps using a segmentation prior and then refining the disparities by means of bundle optimization. However, Zhang cannot handle the narrow-baseline cases, for the estimated camera pose is wrong for narrow-baseline videos. Also the global-based methods use complex methods and the efficiency is low.

Colorization. In recent years, the colorization has been proposed for adding color to a given grayscale image for a few pixels which have color label. Levin[9]'s colorization algorithm is based on the simple premise: neighboring pixels that have similar intensities should have similar colors. Inspired by the colorization, Shunsuke Ono[18] employed the colorization-based coding to perform image compression task. Levin and Dani Lischinsk[19] used the premise to perform the Interactive digital matting. All the approaches obtain excellent results.

3. METHODOLOGY

Our approach is divided into two primary steps as an edge-assisted sparse-to-dense diffusion process, which uses the reconstructed scene information from SFM rather than the camera pose. The former SFM procedure is applied to estimate the camera pose and generate the sparse point clouds. In our approach, the sparse point clouds instead of the camera pose are used as the labels which can be considered with depth prior information. The point clouds of the scene are reprojected to every frame. According to the assumption that nearby pixels in space in a frame with similar colors would have similar depths, we can construct relevant energy function with a closed-form solution. Finally, we propagate the depth to the remaining pixels. In addition, we use the edge detectors to make the propagation better-regulated. The experimental results show that detected edges can improve the performance in some cases.

3.1 SPARSE POINT CLOUDS GENERATION

In the following sections, we provide an introduction of structure from motion, which attempts to estimate the camera pose and generate the sparse point clouds.

3.1.1 STRUCTURE FROM MOTION

For a monocular video sequence $\hat{I}$ with n frames, we usually denote $\hat{I} = \{I_t \mid t = 1, \ldots, n\}$,where $I_t(x)$ expresses the color (or intensity) of pixel x in frame t . The goal of SFM is to estimate the camera pose and reconstruct the scene. For videos, the baseline is narrower than the unordered pictures. Therefore, we intend to use narrow-baseline SFM framework. In our method, if we would like to recover the depth map of the i-th image, we combine the i-th image with its neighboring frame as a video clip. The narrow-baseline SFM solves the problem of every single clip. The key difficulty is the narrow-baseline for video clips which makes feature matching easier, but also causes much calculation difficulty. Here, we use the practical clues introduced by Yu and Gallup[15] .

First, the camera extrinsic matrix can be simplified as

$$P = [R(r) \mid t], R(r) = \begin{pmatrix} 1 & -r^z & r^y \\ r^z & 1 & -r^x \\ -r^y & r^x & 1 \end{pmatrix} \tag{1}$$

where $r = [r^x, r^y, r^z]^T$ is the rotation vector and $t = [t^x, t^y, t^z]^T$ is the translation vector. The matrix R means the transformation from rotation vector r into approximated rotation matrix.

Second, to make the optimization easier, each 3D point is parameterized by its inverse depth. So we have $P_j = (x_j, y_j, 1) / \omega_j$,where (x_j, y_j) is the projection of P_j in the reference image.

It is shown that the first clue helps to reduce the complexity of the rotation models and the second helps to regularize the scale of variables during the bundle adjustment.

Considering the distortion cases, our method uses the D-U model[7],which projects the 3D points from distorted image domain coordinates to undistorted image domain coordinates. We estimate a reasonable approximation of the camera model using one focal length f and two radial distortion parameters k_1 , k_2 . As done in [7], the principle point and radial distortion center are assumed to be equal to the image center. If u_{kj} is the distorted coordinates for the j-th feature in the k-th image relative to the image center, its undistorted coordinates can be calculated as $u_{kj} \cdot G(u_{kj} / f)$, where G is the D-U radial distortion function that is defined by:

$$G(\cdot) = 1 + k_1 \|\cdot\|^2 + k_2 \|\cdot\|^4 \tag{2}$$

As for the reference image, we define $k = 0$ .So the back-projection of the feature u_{0j} to its 3D coordinates x_j is defined using its inverse depth ω_j by:

$$x_j = \left[\frac{u_{0j}}{f \cdot \omega_j} \cdot G(\frac{u_{kj}}{f}) \right] \tag{3}$$

The following function describes the projection of x_j onto the i-th image plane as:

$$\pi(x_j, r_i, t_i) = \left\langle R(r_i) \cdot x_j + t_i \right\rangle \tag{4}$$

whereas the notation $\langle \bullet \rangle$ means the projection mapping from 3D point coordinate to planar 2D point coordinate. The final bundle adjustment is formulated to minimize the reprojection errors as:

$$\arg\min_{K,R,T,W} \sum_{k=1}^{n-1} \sum_{j=0}^{m-1} \rho(u_{kj} \cdot G(\frac{u_{kj}}{f}) - f \cdot \pi(x_j, r_k, t_k)) \tag{5}$$

Where m is the number of features, n is the number of images, $\rho(\cdot)$ is the element-wise Huber loss function[21], K, R, T are the camera parameters, and W is the set of inverse depth values.

As for the feature correspondences, in order to generate more 3D points, we use the Harris corner detector[22]. We first extract the local features in the reference image and find the corresponding feature locations in the other images using the Kanade-Lukas-Tomashi (KLT) algorithm[20]. So in order to obtain more inliers, we select the middle frame rather than the first image as the reference frame. Each tracking is performed forwards and backwards to reject outlier features with bidirectional error greater than 0.4 pixel.

For the estimation of the initial parameters, all the parameters are set to zero except that the focal length is the larger one between the image width and depth. For the inverse depths, we set a random value to each feature.

3.1.2 SPARSE POINT CLOUDS MAPPING

After the SFM process, we obtain the sparse reconstruction of the scene. The multi-view stereo methods use the camera pose again to do dense matching and receive the final result through complex optimization. Instead we just use the sparse point clouds to guide the single-view optimization. The camera pose is just used to project the 3D point from the world coordinate system to the i-th viewpoint coordinate system.

$$X_i = R_i \cdot X + \mathrm{t}_i \tag{6}$$

where X is the 3D point coordinates from the world coordinate system, R_i, t_i are the extrinsic parameters of the i-th frame relative to the reference frame and X_i is the 3D point coordinates from the i-th viewpoint coordinate system.

To reject the outliers of the sparse point clouds, we take the distribution of the intensities from the pixels in the images into account. We compute the variance of every 3D point through the whole video clip as:

$$C(x) = VAR([I_0(x), \cdots, I_{(n-1)}(x)]) \tag{7}$$

where x means single 3D point, and $I_{ik}(x)$ means the intensity of the planar 2D point from relative 3D point reprojection to the i-th image. $VAR(n)$ means the variance of vector n. Each selection is performed to reject the outlier 3D points with $C(x)$ value greater than 0.00005.

3.2 DEPTH MAP GENERATION

Most depth map generation process relies on the multi-view stereo. It greatly relies on the accuracy of the estimated camera pose and the photo-consistency assumption, which means that the matching points between multi-frames should have the similar intensities. Here, we propose a new effective generation process, which releases the dependence on the camera pose and the assumption. As far as we concern, point clouds are able to guide the procedure of generating the depth map.

Inspired by the colorization insights, we use sparse depth points in place of interactive color scribbles. According to the assumption that nearby pixels in space in a frame with similar colors would have similar depths, we can construct relevant energy function with a closed-form solution. In addition, we use the edge detectors to make the propagation better-regulated.

3.2.1 DIFFUSION ALGORITHM

Our diffusion methods are similar to the colorization. Here, we use the sparse depth point as the label to model the function. We denote these sparse depth points as representative depth pixels(RDP), and RDP can be represented by the positions and depth values of these pixels.

Let N be the number of pixels in the RGB image and r be an identifier of the pixels in raster-scan order ($1 \le r \le N$). $x(x \in R^N)$ is assumed to be a one-dimensional vector that contains RDP values, and x assigns non-zero values only for RDP. $D(D \in R^N)$ is assumed to be a one-dimensional vector that contains a depth component restored by our diffusion methods and is arranged in column in raster-scan order. $x(r)$ and $D(r)$ are the r-th elements of vector *x* and D respectively. Let Ω be the sets of the RDP and $|\Omega|$ be the number of the RDP. $s \in N(r)$ means the the s-th pixel belongs to the neighbor of the r-th pixel (normally r-th pixel is not included).

Therefore, our cost function is defined as:

$$J(D) = \sum_{r \notin \Omega} (D(r) - \sum_{s \in N(r)} \omega_{rs} \cdot D(s))^2 + \sum_{r \in \Omega} (D(r) - x(r))^2 \tag{8}$$

As for ω_{rs} , we select the first weighting functions in traditional colorization[9] as

$$\omega_{rs} \propto e^{-(y(r)-y(s))^2/2\sigma_r^2} \tag{9}$$

$y(r)$ means the intensity component of the r-th pixel. ω_{rs} is a weighting function that sums to one.

We express the energy function above in a matrix form:

$$J(D) = \|x - A \cdot D\|^2 \tag{10}$$

where $A = I - W$ (I is the $N \times N$ identity matrix) is an affinity matrix and W is an $N \times N$ matrix that contains ω'_{rs}, which is

$$\omega'_{rs} = \begin{cases} 0 & if \quad r \in \Omega \\ \omega_{rs} & otherwise \end{cases} \tag{11}$$

when $|\Omega| \neq 0$, A is regular matrix and D that fulfills $x = A \cdot D$ always exists. Thus, D is obtained using the following equation

$$D = A^{-1} \cdot x \tag{12}$$

3.2.2 EDGE-ASSISTED PROPAGATION

The depth map generation actually depends on not only the accuracy of the depth point label, but also the number of the labels. More and more accurate 3D point labels can guide the diffusion process more easily. But the change of accuracy is against the amount. Sometimes the diffusion may be unordered for lack of sufficient 3D point labels. So we put forward an edge-assisted method to better regulate the diffusion process. Here, we use the edge detectors[10] to detect the edges of the images and select the probability greater than 0.15 as the edges. If the number of pixels that belong to the edges are small, we consider the image is less-textured and do not use the edge detectors. Otherwise, we add the edge detectors to the energy function. We change ω'_{rs} as ω''_{rs} .

$$\omega''_{rs} = \begin{cases} 0 & if \ \ r \in E \\ \omega'_{rs} & otherwise \end{cases} \tag{13}$$

E is the sets of the pixels which are detected by the edge detectors. Through ω''_{rs} ,we can use the edges to help better regulate the diffusion process.

4. EXPERIMENTS

In our experiments, we first show the results with and without the edge-assisted guidance. The results are shown as Figure1. Here, we could find that the edge information can help to guide the diffusion process. The pole region is separated from the trees with the edge information guidance.

Then we compare our results with the state-of-the-art approaches. Here, the results are shown with disparity map and the disparity range is set from 0 to 202.The first two rows are from Zhang[6]'s ACTS datasets, while the last two rows are from Hyowon Ha[7]'s datasets. The results are shown as Figure 2. The black map for the last two cows means that Zhang's ACTS software cannot generate the results. The results show that Zhang cannot handle the narrow-baseline cases. Also, Hyowon Ha[7] cannot generate good results for the multi-view matching error, especially for the occlusion regions. Especially, there exist the regions which have depth discontinuity. Instead, our method can both handle these two cases and the results are comparable. It is demonstrated that our method is more robust.

Finally, we compare the efficiency of the approaches. We select the road datasets from Zhang[6] and put every seven images as a clip. We eventually obtain 20 clips. We compare the average time for a clip. The time includes the SFM and the depth map generation process. The results are shown as table 1. We use a notebook computer equipped with an Intel i7-6700HQ 2.6Ghz CPU and 24GB RAM. It is demonstrated that our method is a little slower than the Hyowon Ha[7]'s while much faster than Zhang[6]'s. But it is accepted because our diffusion process is implemented by MATLAB.

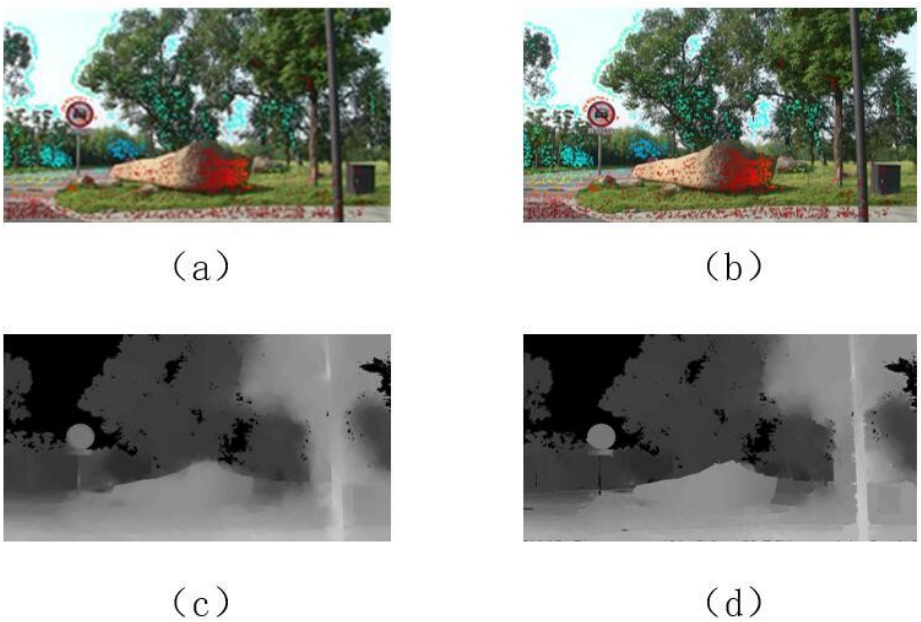


Figure 2. The experiment results without or with edge information.

(a) The reference image with depth point label. (b) The edge map. (c) The result without the edge guidance. (d) The result with the edge guidance

Table 1. The efficiency comparison with the state-of-the art

	SFM(s)	Depth map generation(s)	Total time(s)
Zhang[6]	134.28	420.24	554.52
Hyowon Ha[7]	5.53	55.18	60.71
Ours	6.45	73.16	79.61

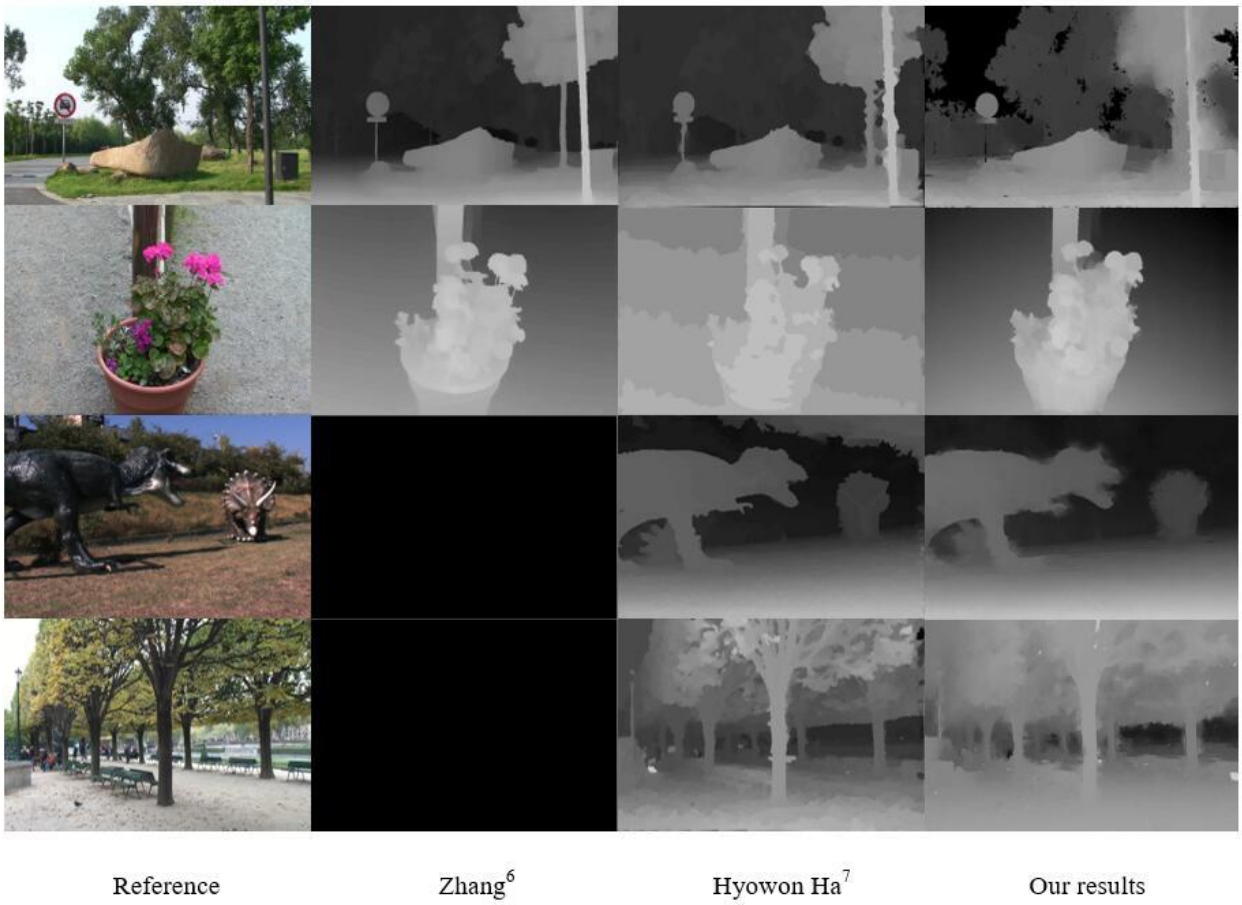


Figure 3. Comparison with state-of-the-art approaches.

5. CONCLUSION

We have put forward a new framework that recovers the depth map from a monocular video. Our method employs the sparse point cloud as label instead of using the camera pose. Also the edge detectors help to regulate the sparse-to-dense diffusion. The results show that our methods are more robust. In the meantime, the efficiency of our method is comparable with the state-of the-art.

ACKNOWLEDGMENTS

This work is jointly supported by the National High Technology Research and Development Program of China (863 Program) under Grant No. 2015AA015904 and the 2015 annual foundation of China Academy of Space Technology (CAST).

REFERENCES

[1] Okutomi, M. and Kanade, T., "A Multiple-Baseline Stereo," IEEE Trans. Pattern Analysis and Machine Intelligence 15(4), 353-363 (1993).
[2] Collins, R. T. ,"A Space-Sweep Approach to True Multi-Image Matching," Proc. IEEE Conference on Computer Vision and Pattern Recognition, 358-363(1996).
[3] Kolmogorov, V. and Zabih, R., "Computing Visual Correspondence with Occlusions Via Graph Cuts," Proc. IEEE International Conference on Computer Vision, 508-515(2001).
[4] Seitz, S. M., Curless, B., Diebel, J., Scharstein, D. and Szeliski, R., "A Comparison and Evaluation of Multi-View Stereo Reconstruction Algorithms," Proc. IEEE Conference on Computer Vision and Pattern Recognition, 519-528(2006).
[5] Kang S B. and Szeliski R., "Extracting View-Dependent Depth Maps from a Collection of Images," International Journal of Computer Vision 58(2), 139-163(2004).
[6] Zhang, G., Jia, J., Wong, T. T. and Bao H., "Consistent Depth Maps Recovery from a Video Sequence," IEEE Trans. Pattern Analysis and Machine Intelligence 31(6), 974-988(2009).
[7] Ha, H., Im, S., Park, J., Jeon, H. G. and Kweon, I. S., "High-quality depth from uncalibrated small motion clip," Proc. IEEE Conference on Computer Vision and Pattern Recognition, 5413-5421(2016).
[8] Hartley, R. and Zisserman, A., [Multiple view geometry in computer vision], Cambridge University Press, 151-561(2003).
[9] Levin, A., Lischinski, D. and Weiss, Y., "Colorization using optimization," ACM Transactions on Graphics 23(3), 689-694(2004).
[10] Zitnick, C. L. and Dollar, P., "Edge boxes: Locating object proposals from edges," Proc. European Conference on Computer Vision, 391-405(2014).
[11] Zhang, G., Qin, X., Hua, W., Wong, T. T., Heng, P. A. and Bao, H., "Robust metric reconstruction from challenging video sequences," Proc. IEEE Conference on Computer Vision and Pattern Recognition, 1-8(2007).
[12] Snavely, N., Seitz, S. M. and Szeliski, R., "Modeling the world from internet photo collections," International Journal of Computer Vision 80(2), 189-210(2008).
[13] Wu C., "Towards linear-time incremental structure from motion," Proc. IEEE International Conference on 3DTV-Conference, 127-134(2013).
[14] Moulon, P., Monasse, P. and Marlet, R., "Global fusion of relative motions for robust, accurate and scalable structure from motion," Proc. IEEE International Conference Computer Vision,3248-3255(2013).
[15] Yu, F. and Gallup, D., "3d reconstruction from accidental motion," Proc. IEEE Conference on Computer Vision and Pattern Recognition,3986-3993(2014).
[16] Yang, Q., "A non-local cost aggregation method for stereo matching," Proc. IEEE Conference on Computer Vision and Pattern Recognition,1402–1409(2012).
[17] Collins, R. T., "A space-sweep approach to true multi-image matching," Proc. IEEE Conference on Computer Vision and Pattern Recognition, 358–363(1996).
[18] Ono, S., Miyata, T. and Sakai, Y., "Colorization-based coding by focusing on characteristics of colorization bases," Proc. IEEE Picture Coding Symposium, 230-233(2010).
[19] Levin, A., Lischinski, D. and Weiss, Y., "A Closed-Form Solution to Natural Image Matting," IEEE Trans. Pattern Analysis and Machine Intelligence 30(2),228-242(2008).
[20] Lucas, B. D. and Kanade, T., "An iterative image registration technique with an application to stereo vision," Proc. International Joint Conference on Artificial Intelligence, 674–679(1981).
[21] Huber, P. J., "Robust estimation of a location parameter," Annals of Statistics 53(1),73–101(1964).
[22] Harris, C., and Stephens, M., "A combined corner and edge detector," Proc. Alvey Vision Conference, 147-151(1988).