Color-Introduced Frame-to-Model Registration for 3D Reconstruction

Fei Li(✉), Yunfan Du, and Rujie Liu

Fujitsu Research and Development Center Co., Ltd., Beijing, China
{lifei,duyunfan,rjliu}@cn.fujitsu.com

Abstract. 3D reconstruction has become an active research topic with the popularity of consumer-grade RGB-D cameras, and registration for model alignment is one of the most important steps. Most typical systems adopt depth-based geometry matching, while the captured color images are totally discarded. Some recent methods further introduce photometric cue for better results, but only frame-to-frame matching is used. In this paper, a novel registration approach is proposed. According to both geometric and photometric consistency, depth and color information are involved in a unified optimization framework. With the available depth maps and color images, a global model with colored surface vertices is maintained. The incoming RGB-D frames are aligned based on frame-to-model matching for more effective camera pose estimation. Both quantitative and qualitative experimental results demonstrate that better reconstruction performance can be obtained by our proposal.

Keywords: 3D reconstruction · Color mapping · Registration · Frame-to-model matching · Optimization

1 Introduction

As one of the most important tasks in computer vision and graphics, high-quality digitization of real-world objects has always been a hot research theme. Generally speaking, digital 3D objects cannot be directly acquired by some equipment, and they are often reconstructed based on the captured raw data, such as depth maps and color images. In the last decades, several 3D reconstruction methods have been explored. With the development of sensing technology, consumer-grade RGB-D cameras have appeared recently. Since they are often with low cost, easy portability and high streaming rate, these cameras have been widely used in various applications, such as augmented reality, computer games, and virtual shopping. Although with many advantages, consumer-grade RGB-D cameras are often with inevitable distortions, and the obtained data are usually not accurate enough, especially for the captured depth maps [1]. Therefore, how to obtain satisfactory 3D reconstruction performance with inaccurate capture devices has been paid more and more attention in recent years [2–12].

3D reconstruction with RGB-D cameras mainly resorts to the obtained depth information, and there are two main steps in the whole process: registration and

L. Amsaleg et al. (Eds.): MMM 2017, Part II, LNCS 10133, pp. 112–123, 2017.
DOI: 10.1007/978-3-319-51814-5_10

integration. The procedures can be simply described as follows. Since multiple depth maps are captured from different positions and directions, they are first aligned into the same coordinate space based on the estimated camera poses. Then the registered models are combined together to get the final reconstruction results.

The accuracy of model alignment has a major influence on the final reconstruction performance, hence most research work mainly focuses on the approach of registration. The most direct idea is to estimate the camera pose of each depth map as accurate as possible, and two frequently adopted ways are frame-to-frame matching [13] and frame-to-model matching [2,3]. Frame-to-frame matching only considers two consecutive frames at a time, and estimates the camera pose of each new depth map by aligning it to its last frame. While frame-to-model matching maintains a model constructed by all the frames coming before, and registers the incoming depth map to the growing model. Since the existing frames are made full use of, frame-to-model matching is often more effective and more robust than frame-to-frame matching [3]. No matter which kind of matching way is adopted, most registration approaches pay more attention to real-time 3D reconstruction, and only consider the frames before the incoming depth map. When the real-time requirement is not necessary, more useful information can be involved for more accurate results. Two-pass registration [4] is such an example. In the first pass, all the available frames are used to construct a whole model; and in the second pass, each frame is revisited again and aligned to the acquired model. As the frames after the incoming depth map are also taken into account, more stable camera pose estimation results can be obtained.

Besides kinds of approaches for camera pose estimation, camera distortions are also addressed in some recent research. Two categories of methods are often adopted. The first category of methods are mainly based on calibration, and try to estimate a specific distortion function with a pre-defined form for the given camera [14,15]. Since specialized calibration sequences are required, the applicable cases of these methods are largely confined. Moreover, the real distortion function is usually irregular and complicated, so the assumption about its form may be not exact. The second category of methods do not explicitly estimate camera distortions, but attempt to correct them by introducing non-rigid deformation, and the most suitable deformation parameters are often calculated by optimization. Elastic registration [5] finds the most appropriate mapping for each 3D point. Because of the lack of prior knowledge, it always leads to unnecessary warping in the final reconstruction results. Moreover, it is quite time-consuming. By factorizing the non-rigid deformation into a rigid localization component and a latent non-rigid calibration component, the method of SLAC [7] effectively conducts localization and calibration at the same time. As both of the two aspects are elaborated considered in one joint optimization framework, it achieves better reconstruction performance. Meanwhile, since the number of parameters to be optimized is much smaller, its overall computational cost is dramatically reduced.

Most representative systems only adopt depth maps for 3D reconstruction, and the methods of geometric alignment have been largely explored. However, as only distance information is recorded, the captured depth map has a certain limitations. Apart from its inaccuracy and missing data, depth-based reconstruction is prone to drift and failure in the presence of smooth surfaces. Therefore, some methods attempt to introduce color images and utilize photometric cue for performance improvement [16–20]. Considering real-time requirements, complicated image processing operations cannot be conducted, and pixel-level dense matching is often adopted. In the existing methods, only frame-to-frame matching is used for photometric consistency based registration. That is to say, for color-involved camera pose estimation, only the incoming image and the previous one are considered. Like the case for depth-based registration, it is hoped that frame-to-model matching will also outperform frame-to-frame matching when color information is taken into account.

In this paper, with the basic idea to utilize color images as well as depth maps for better reconstruction performance, a novel registration approach is proposed. The whole system is with the similar workflow to KinectFusion [2,3]. However, its maintained global model contains not only surface vertices and their normals, but also the corresponding color values. When a new pair of depth map and color image come, the idea of frame-to-model matching is adopted, and both geometric and photometric consistency are considered to develop a unified optimization problem. Since the relationship between the new captured data and the existing RGB-D frames is fully explored, more accurate camera pose estimation results are obtained. Moreover, by introducing non-rigid correction functions, our proposal can be easily extended to involve camera distortions in the same framework.

The rest of the paper is organized as follows. Section 2 describes our proposed color-introduced frame-to-model registration approach in detail. Our experimental results are illustrated in Sect. 3. Some conclusions and analysis of future work follow in Sect. 4.

2 Color-Introduced Frame-to-Model Registration

In this section, first we explain the approach to represent the maintained global model with color information. Then we present the optimization problem based on both geometric and photometric consistency for frame-to-model registration, and talk about how to effectively solve it. Finally, we discuss some extensions of our proposed method.

2.1 Colored Global Model Representation

In KinectFusion [2,3], the whole geometry information of the 3D object to be reconstructed is represented by a volumetric truncated signed distance function (TSDF). The TSDF value of each point is defined as the signed distance between its calculated depth and the value of its projection position in the depth map, and the distance is truncated into a pre-defined interval. For each iteration, when a new depth map comes, after the step of registration for camera pose estimation, its corresponding TSDF volume is calculated and aligned to the maintained global TSDF volume. Then in the step of integration, the calculated TSDF volume and the global one are combined together by a simple running weighted average [21].

Given the integrated TSDF volume, with the idea to find the points with zero-valued TSDF, pixel-level raycast [22] is performed to determine the vertices on the model surface, and the corresponding normals are obtained by gradient extraction. For the global TSDF volume constructed by $(k-1)$ depth maps, the calculated surface vertices and their normals are denoted as $\mathbf{V}_{k-1}^{g}(\mathbf{u})$ and $\mathbf{N}_{k-1}^{g}(\mathbf{u})$, respectively, where the superscript "g" indicates that they are defined in the global world space, and $\mathbf{u}$ is the pixel position in the image. These data are used as the "model" for depth-based frame-to-model matching [2,3].

In our proposal, TSDF volume is also adopted for maintaining the overall geometry information, the surface vertices $\mathbf{V}_{k-1}^{g}(\mathbf{u})$ and their normals $\mathbf{N}_{k-1}^{g}(\mathbf{u})$ are calculated in the same way as that in KinectFusion [2,3]. To further involve [redacted], each surface vertex is transformed into the camera space and projected to the available color images. Let the already estimated camera poses for the $(k-1)$ RGB-D frames be $\{\mathbf{T}_{g,1}, \mathbf{T}_{g,2}, \cdots, \mathbf{T}_{g,k-1}\}$, which indicate the transformation matrices from each camera space to the global world space. Thus in the m-th $(m = 1, 2, \cdots, k-1)$ camera space, the surface vertices are calculated as

$$\mathbf{p}_m = [x_m, y_m, z_m, 1]^T = \mathbf{T}_{g,m}^{-1}\mathbf{V}_{k-1}^{g}(\mathbf{u}) \tag{1}$$

where homogeneous coordinates are adopted, and $\mathbf{T}_{g,m}$ is a 4×4 matrix involving both rotation and translation. Let $H(\cdot)$ denote the projection operation from 3D space to 2D space, f_x and f_y be the focal lengths, and (c_x, c_y) be the coordinates for the principal point, then the projection position is

$$\mathbf{v}_m = H(\mathbf{p}_m) = \left[\frac{x_m f_x}{z_m} + c_x, \frac{y_m f_y}{z_m} + c_y\right]^T \tag{2}$$

Given the $(k-1)$ color images $\{\mathcal{F}_1, \mathcal{F}_2, \cdots, \mathcal{F}_{k-1}\}$, the color of each surface vertex can be calculated by averaging all the color values of the corresponding projection positions, namely

$$C_{k-1}^{g}(\mathbf{u}) = \frac{1}{k-1}\sum_{m=1}^{k-1}\mathcal{F}_m(\mathbf{v}_m) \tag{3}$$

The above discussion only considers the projection model for each separated point, but does not address its visibility. In fact, the surface vertex $\mathbf{V}_{k-1}^{g}(\mathbf{u})$ may be unseen from the viewpoint of the m-th camera, thus the color value of the projection position will be meaningless. Therefore, a constraint is introduced, and we only choose the vertices whose calculated depth in the m-th camera space and the depth value of the projection position are close enough. Let the $(k-1)$ depth maps be denoted as $\{\mathcal{D}_1, \mathcal{D}_2, \cdots, \mathcal{D}_{k-1}\}$, the constraint can be formulated as

$$|z_m - \mathcal{D}_m(\mathbf{v}_m)| < \theta \tag{4}$$

where θ is a pre-defined threshold. Since it is more likely that the surface vertex cannot be seen from the viewpoint of images captured a long time before, the color calculation is confined for only considering the previous M images to reduce the computational load. Therefore, the final color of each surface vertex is determined as

$$C^g_{k\text{–}1}(\mathbf{u}) = \frac{\sum_{i=1}^{M} \mathcal{F}_{k\text{–}i}(\mathbf{v}_{k\text{–}i}) \cdot I\left(|z_{k\text{–}i} - \mathcal{D}_{k\text{–}i}(\mathbf{v}_{k\text{–}i})| < \theta\right)}{\sum_{i=1}^{M} I\left(|z_{k\text{–}i} - \mathcal{D}_{k\text{–}i}(\mathbf{v}_{k\text{–}i})| < \theta\right)} \tag{5}$$

where $I(\cdot)$ is the indicator function, namely $I(A) = 1$ when A is true, otherwise $I(A) = 0$.

With the constantly updated TSDF volume, all the surface vertices may be totally changed with the new RGB-D frames, thus the color value of each surface vertex must be recalculated in each iteration. Fortunately, the calculation can be efficiently implemented in parallel by GPU, thus the computational time cost is quite low. In our proposal, the calculated colors $C^g_{k\text{–}1}(\mathbf{u})$, as well as the surface vertices $\mathbf{V}^g_{k\text{–}1}(\mathbf{u})$ and the corresponding normals $\mathbf{N}^g_{k\text{–}1}(\mathbf{u})$, are utilized as the global model in the step of registration.

2.2 Frame-to-Model Matching Based Optimization

When the k-th depth map $\mathcal{D}_k$ comes, its corresponding vertices $\mathbf{V}_k(\mathbf{u})$ and normals $\mathbf{N}_k(\mathbf{u})$ are obtained by the method in KinectFusion [2,3]. As they are expressed in the camera space, the superscript "g" is not used. The incoming color image $\mathcal{F}_k$, as well as the calculated $\mathbf{V}_k(\mathbf{u})$ and $\mathbf{N}_k(\mathbf{u})$, are treated as the "frame" information.

In our proposal, both geometric and photometric consistency are taken into account for registration, and the way of frame-to-model matching is adopted for the two aspects. By fully exploring the available depth maps and color images in a unified optimization framework, the camera pose $\mathbf{T}_{g,k}$ can be estimated more accurately.

For frame-to-model geometry matching, the case is the same as that in KinectFusion [2,3]. By finding the corresponding projection positions and involving point-to-plane distance, the cost term is defined as

$$E_1(\mathbf{T}_{g,k}) = \sum_{(\mathbf{u},\hat{\mathbf{u}})\in\mathcal{U}} \left\| \left(\mathbf{T}_{g,k}\mathbf{V}_k(\mathbf{u}) - \mathbf{V}^g_{k\text{–}1}(\hat{\mathbf{u}})\right)^T \mathbf{N}^g_{k\text{–}1}(\hat{\mathbf{u}}) \right\|^2 \tag{6}$$

where $\mathbf{u}$ and $\hat{\mathbf{u}}$ are corresponding projection positions for the same point in 3D space, and $\mathcal{U}$ is the set of corresponding position pairs found by considering both the vertex coordinates and the normal directions.

For frame-to-model color matching, the color of each 3D vertex in the global model is compared with the value of the 2D projection position in the k-th color image, and it is hoped that the two color values should be as close as possible. According to Eqs. (1) and (2), we calculate each surface vertex represented in the k-th camera space $\mathbf{p}_k$ and its corresponding projection position $\mathbf{v}_k$, and the cost term is described as

$$\begin{aligned} E_2(\mathbf{T}_{g,k}) &= \sum_{\mathbf{u}\in\mathcal{W}} \left\| C^g_{k-1}(\mathbf{u}) - \mathcal{F}_k(\mathbf{v}_k) \right\|^2 \\ &= \sum_{\mathbf{u}\in\mathcal{W}} \left\| C^g_{k-1}(\mathbf{u}) - \mathcal{F}_k\Big(H\big(\mathbf{T}^{-1}_{g,k}\mathbf{V}^g_{k-1}(\mathbf{u})\big)\Big) \right\|^2 \end{aligned} \qquad (7)$$

where $\mathcal{W}$ is the set of valid projection positions defined by considering a similar constraint as that in Eq. (4)

$$\begin{aligned} \mathcal{W} &= \left\{ \mathbf{u} \;\middle|\; |z_k - \mathcal{D}_k(\mathbf{v}_k)| < \theta \right\} \\ &= \left\{ \mathbf{u} \;\middle|\; \Big| z_k - \mathcal{D}_k\Big(H\big(\mathbf{T}^{-1}_{g,k}\mathbf{V}^g_{k-1}(\mathbf{u})\big)\Big)\Big| < \theta \right\} \end{aligned} \qquad (8)$$

where z_k is the z-coordinate of $\mathbf{p}_k$, indicating the calculated depth in the k-th camera space.

The aforementioned two cost terms are linearly combined in our proposal, and the final cost function involving both geometric and photometric consistency is defined as

$$\begin{aligned} E(\mathbf{T}_{g,k}) &= E_1(\mathbf{T}_{g,k}) + \lambda E_2(\mathbf{T}_{g,k}) \\ &= \sum_{(\mathbf{u},\hat{\mathbf{u}})\in\mathcal{U}} \left\| \left(\mathbf{T}_{g,k}\mathbf{V}_k(\mathbf{u}) - \mathbf{V}^g_{k-1}(\hat{\mathbf{u}})\right)^T \mathbf{N}^g_{k-1}(\hat{\mathbf{u}}) \right\|^2 \\ &\quad + \lambda \sum_{\mathbf{u}\in\mathcal{W}} \left\| C^g_{k-1}(\mathbf{u}) - \mathcal{F}_k\Big(H\big(\mathbf{T}^{-1}_{g,k}\mathbf{V}^g_{k-1}(\mathbf{u})\big)\Big) \right\|^2 \end{aligned} \qquad (9)$$

where λ is a balanced coefficient for the two terms. It may remain unchanged for all the RGB-D frames or vary with different values. An example for introducing variable coefficient is to consider the blurriness of each color image. λ can be set to a smaller value for more blurry images, as it is likely that the photometric cue obtained from blurry images is inaccurate. The camera pose $\mathbf{T}_{g,k}$ can be calculated by minimizing the overall cost function, which also means maximizing the geometric and photometric consistency.

2.3 Solution to Optimization Problem

As how to minimize the first cost term $E_1(\mathbf{T}_{g,k})$ in the optimization problem has been detailedly explained in KinectFusion [3], we pay more attention to the second term $E_2(\mathbf{T}_{g,k})$ in this section.

The solution of KinectFusion is an iterative approach, and the camera pose to be determined in one iteration is locally linearized around its value obtained in the last iteration. That is to say, in the n-th round of iteration for calculating $\mathbf{T}_{g,k}$, we have

$$\begin{aligned}\mathbf{T}_{g,k}^{(n)} &\approx \mathit{\Delta}\mathbf{T}\ \mathbf{T}_{g,k}^{(n-1)} \\ &= \begin{bmatrix} 1 & -\gamma & \beta & a \\ \gamma & 1 & -\alpha & b \\ -\beta & \alpha & 1 & c \\ 0 & 0 & 0 & 1 \end{bmatrix} \mathbf{T}_{g,k}^{(n-1)}\end{aligned} \tag{10}$$

Here a vector $\mathbf{x} = [\alpha, \beta, \gamma, a, b, c]^T \in \mathbb{R}^6$ is adopted for parameterizing the incremental transformation matrix $\mathit{\Delta}\mathbf{T}$, in which $[\alpha, \beta, \gamma]^T$ and $[a, b, c]^T$ are used for describing the tiny rotation and translation variations, respectively. To deal with $\mathbf{T}_{g,k}^{-1}$ in the second cost term $E_2(\mathbf{T}_{g,k})$, considering that all the six elements in the vector $\mathbf{x}$ are with small values, it can be easily obtained

$$\begin{aligned}\left(\mathbf{T}_{g,k}^{(n)}\right)^{-1} &\approx \left(\mathbf{T}_{g,k}^{(n-1)}\right)^{-1} (\mathit{\Delta}\mathbf{T})^{-1} \\ &\approx \left(\mathbf{T}_{g,k}^{(n-1)}\right)^{-1} \begin{bmatrix} 1 & \gamma & -\beta & -a \\ -\gamma & 1 & \alpha & -b \\ \beta & -\alpha & 1 & -c \\ 0 & 0 & 0 & 1 \end{bmatrix}\end{aligned} \tag{11}$$

In the iterative solution for minimizing the first cost term $E_1(\mathbf{T}_{g,k})$, the problem is transformed into a linear equation in KinectFusion. Since the derivation process is complicated, the final result is simply denoted as $\mathbf{Ax} = \mathbf{b}$, where $\mathbf{A}$ is a 6×6 symmetric matrix, and $\mathbf{b}$ is a 6×1 vector. For more details about the derivation, the readers can be referred to [3]. For minimizing the second cost term $E_2(\mathbf{T}_{g,k})$, since it is with the form of non-linear least squares, the Gauss-Newton method can be adopted. In the n-th round of iteration, we calculate the residual vector $\mathbf{r} = [\mathbf{r_u}]$ where

$$\mathbf{r_u} = C_{k-1}^{g}(\mathbf{u}) - \mathcal{F}_k\left(H\left(\left(\mathbf{T}_{g,k}^{(n-1)}\right)^{-1} \mathbf{V}_{k-1}^{g}(\mathbf{u})\right)\right) \tag{12}$$

as well as its Jacobian matrix $\mathbf{J}$ with respect to $\mathbf{x}$, then the parameterized vector $\mathbf{x}$ can be updated by $\mathbf{J}^T\mathbf{J}\mathbf{x} = -\mathbf{J}^T\mathbf{r}$. In our proposal, by taking both of the two cost terms into account, $\mathbf{x}$ is calculated by solving the equation

$$\left(\mathbf{A} + \lambda\mathbf{J}^T\mathbf{J}\right)\mathbf{x} = \left(\mathbf{b} - \lambda\mathbf{J}^T\mathbf{r}\right) \tag{13}$$

Like KinectFusion [3], we can also down-sample the depth maps and color images for multi-scale representations, and conduct a coarse-to-fine framework to effectively solve the optimization problem.

2.4 Extensions

In the above discussion, we only talk about the case when all the depth maps have their corresponding color images. While in the practical applications with RGB-D cameras, depth maps and color images may be captured with different frame rates, and it is more likely that color images are obtained with lower frame rate for larger resolution. To extend our method to manage the situation, our proposed color-introduced frame-to-model registration is only utilized when the pair of depth map and color image exist, otherwise only geometry matching is adopted for camera pose estimation.

Our proposal can also be easily extended for involving camera distortions. Like the method in [23], in order to deal with the optical aberrations, a non-rigid correction function $\mathbf{L}_k$ over the image plane is introduced, and the cost term for photometric consistency is modified as

$$E_2(\mathbf{T}_{g,k}) = \sum_{\mathbf{u}\in\mathcal{W}} \left\| C^g_{k-1}(\mathbf{u}) - \mathcal{F}_k\big(\mathbf{L}_k(\mathbf{v}_k)\big)\right\|^2 \tag{14}$$

where $\mathbf{L}_k$ is directly defined for some pre-given positions, and generalized to other positions in the image plane by bilinear interpolation. The camera pose $\mathbf{T}_{g,k}$ and the parameters of the correction function $\mathbf{L}_k$ can be iteratively calculated by joint optimization.

3 Experimental Results

To evaluate the performance of our proposed approach, some experiments are implemented on two data sets. For quantitative evaluation, camera pose estimation and 3D reconstruction are conducted on the RGB-D SLAM benchmark [24]. Three sequences "fr1/desk", "fr1/room", as well as "fr3/long_office_household" from the benchmark are adopted, and the first 100 pairs of depth maps and color images in each sequence are used. Some captured color images in the three sequences are shown in Fig. 1. Since the benchmark provides ground-truth trajectories obtained from a high-accuracy motion capture system, the estimated camera poses can be compared with the ground-truth data. The absolute translational root mean square error [24] is utilized as the performance measure.

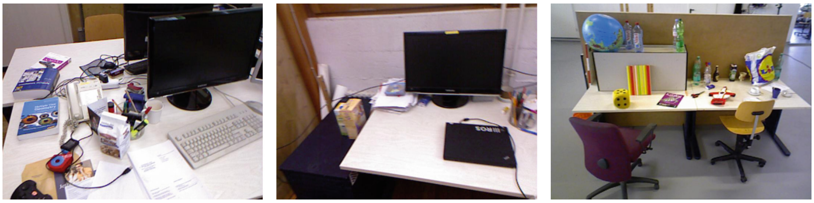


Fig. 1. Some captured color images in the sequences "fr1/desk", "fr1/room", and "fr3/long_office_household" from the RGB-D SLAM benchmark (Color figure online)

Three other registration approaches are used for comparison, including depth-based frame-to-frame matching (D_F2F) [13], depth-based frame-to-model matching (D_F2M) [2,3], as well as D_F2M further involving frame-to-frame color matching (D_F2M+C_F2F) [20]. The absolute translational root mean square errors for the three methods and our proposal are listed in Table 1, it can be seen that our approach produces the best results for all the sequences. D_F2F only adopts two consecutive depth maps for camera pose estimation. As much useful information is not well explored, its performance is the worst of all. Compared with D_F2M, both D_F2M+C_F2F and our proposal introduce photometric consistency in the step of registration. Their superiority demonstrates that it is reasonable to utilize color images as additional cue for obtaining more accurate trajectories. As far as the two color-involved approaches are concerned, like the case of geometric alignment, frame-to-model matching also outperforms frame-to-frame matching for color-introduced registration. Therefore, our proposal is more effective than D_F2M+C_F2F.

Table 1. Absolute translational root mean square errors (in centimeters) on difference sequences from the RGB-D SLAM benchmark

Sequence	D_F2F	D_F2M	D_F2M+C_F2F	Our proposal
fr1/desk	4.53	2.03	1.99	1.83
fr1/room	4.70	4.33	4.14	3.89
fr3/office	4.74	1.90	1.84	1.78

To demonstrate the difference of the estimated camera poses more clearly, as an example, the reconstruction results of D_F2M and our proposal for the sequence "fr1/desk" are placed in the same coordinate space and illustrated in Fig. 2. We can see that there are obvious displacements between the two results. Similar cases can be obtained for other sequences, and the displacements between the reconstruction results of our proposal and other approaches always exist as well. It is known that inaccurate estimated camera poses will hinder subsequent steps such as color mapping, thus effective registration approach is of great importance for 3D processing.

We also conduct experiments on our own data, which consists of 1328 pairs of depth maps and color images captured from various viewpoints for a toy teddy bear. All the depth maps are adopted in the process of depth-based reconstruction. For involving color information, 42 images with low blurriness are chosen to ensure the accuracy of the introduced photometric cue. In order to better compare different approaches, colored reconstruction results are illustrated here, and the color of each surface vertex is simply determined by averaging all the corresponding color values in the images. The reconstructed toy teddy bears by D_F2M and our proposal are shown in Fig. 3. We can see that even only about 3.2% (42/1328) color images are adopted for registration, our result is more clear than that by D_F2M, especially for the area of characters on the box, which indicates that the estimated camera poses by our proposal are more accurate. It should be noted that only simple color mapping is implemented here, thus the performance is not satisfactory. If more elaborate color mapping methods, such as [23], are utilized, better results can be achieved.

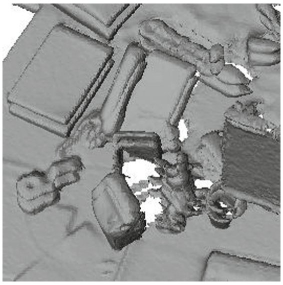
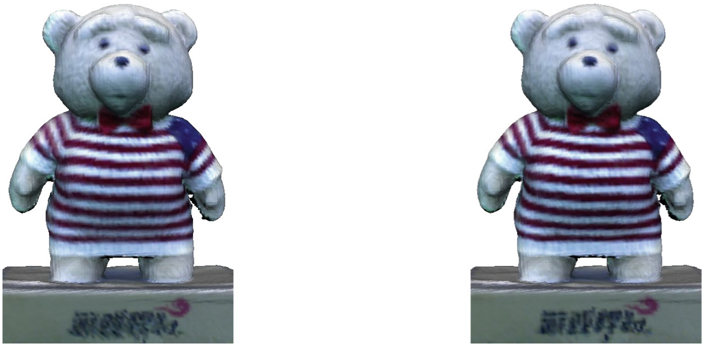


Fig. 2. Reconstruction results of D_F2M and our proposal for the sequence “fr1/desk” placed in the same coordinate space



(a) D_F2M

(b) Our proposal

Fig. 3. Reconstructed toy teddy bears by D_F2M and our proposal

4 Conclusions and Future Work

In this paper, by taking both geometric and photometric consistency into account, a novel registration approach for 3D reconstruction is proposed. Since they can provide additional information over depth maps, color images are reasonably introduced and largely explored. To make full use of the existing RGB-D frames, the maintained global model contains not only surface vertices and their normals, but also the corresponding color values. Frame-to-model geometry matching and color matching are simultaneously considered in a unified optimization framework, and an iterative solution is well developed. Furthermore, our method can be easily extended to deal with the case when depth maps and color images are captured with different frame rates, and it is convenient to further involve camera distortions. Experiments demonstrate that our proposal can achieve more accurate camera pose estimation results.

For the next research work, we will mainly focus on how to involve camera distortions more effectively. In the existing methods to address camera distortions by non-rigid transformation, the correction function is usually directly defined on a uniform lattice, and the lattice is kept the same for all the color images. In our proposal, only the projection positions of the surface vertices are useful for camera pose estimation. Generally speaking, they are not evenly distributed in the image plane, and their distributions are different for each image. Therefore, if only the same uniform lattice is adopted, the interpolation results may be inaccurate for the positions not on the lattice, and more complex lattices should be considered for better representation of non-rigid transformations. In the future, we will pay our attention to the problem of how to adaptively determine the most effective lattice for each color image, and attempt to efficiently finish its implementation.

References

1. Smisek, J., Jancosek, M., Pajdla, T.: 3D with kinect. In: Fossati, A., Gall, J., Grabner, H., Ren, X., Konolige, K. (eds.) Consumer Depth Cameras for Computer Vision, pp. 3–25. Springer, Heidelberg (2013)
2. Izadi, S., Kim, D., Hilliges, O., Molyneaux, D., Newcombe, R., Kohli, P., Shotton, J., Hodges, S., Freeman, D., Davison, A., Fitzgibbon, A.: KinectFusion: real-time 3D reconstruction and interaction using a moving depth camera. In: Proceedings of ACM Symposium on User Interface Software and Technology, pp. 559–568 (2011)
3. Newcombe, R.A., Izadi, S., Hilliges, O., Molyneaux, D., Kim, D., Davison, A.J., Kohli, P., Shotton, J., Hodges, S., Fitzgibbon, A.: KinectFusion: real-time dense surface mapping and tracking. In: Proceedings of IEEE International Symposium on Mixed and Augmented Reality, pp. 127–136 (2011)
4. Zhou, Q.Y., Koltun, V.: Dense scene reconstruction with points of interest. ACM Trans. Graph. **32** (2013)
5. Zhou, Q.Y., Miller, S., Koltun, V.: Elastic fragments for dense scene reconstruction. In: Proceedings of IEEE International Conference on Computer Vision, pp. 473–480 (2013)
6. Nießner, M., Zollhöfer, M., Izadi, S., Stamminger, M.: Real-time 3D reconstruction at scale using voxel hashing. ACM Trans. Graph. **32** (2013)
7. Zhou, Q.Y., Koltun, V.: Simultaneous localization and calibration: self-calibration of consumer depth cameras. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, pp. 454–460 (2014)
8. Choe, G., Park, J., Tai, Y.W., Kweon, I.S.: Exploiting shading cues in kinect IR images for geometry refinement. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, pp. 3922–3929 (2014)
9. Zollhöfer, M., Nießner, M., Izadi, S., Rhemann, C., Zach, C., Fisher, M., Wu, C., Fitzgibbon, A., Loop, C., Theobalt, C., Stamminger, M.: Real-time non-rigid reconstruction using an RGB-D camera. ACM Trans. Graph. **33** (2014)
10. Newcombe, R.A., Fox, D., Seitz, S.M.: DynamicFusion: reconstruction and tracking of non-rigid scenes in real-time. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, pp. 343–352 (2015)
11. Dou, M., Taylor, J., Fuchs, H., Fitzgibbon, A., Izadi, S.: 3D scanning deformable objects with a single RGBD sensor. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, pp. 493–501 (2015)