Hindawi
Discrete Dynamics in Nature and Society
Volume 2020, Article ID 8215389, 9 pages
https://doi.org/10.1155/2020/8215389

Research Article

RGBD Scene Flow Estimation with Global Nonrigid and Local Rigid Assumption

Xiuxiu Li,[1,2] **Yanjuan Liu**,[1,2] **Haiyan Jin**,[1,2] **Lei Cai**,[1,2] **and Jiangbin Zheng**[3]

[1]*Xi'an University of Technology, Xi'an 710048, China*
[2]*Shaanxi Key Laboratory for Network Computing and Security Technology, Xi'an 710048, China*
[3]*Northwestern Polytechnical University, Xi'an 710029, China*

Correspondence should be addressed to Xiuxiu Li; lixiuxiu@xaut.edu.cn

Received 27 March 2020; Accepted 30 May 2020; Published 29 June 2020

Guest Editor: Longzhuang Li

RGBD scene flow has attracted increasing attention in the computer vision with the popularity of depth sensor. To estimate the 3D motion of object accurately, a RGBD scene flow estimation method with global nonrigid and local rigid motion assumption is proposed in this paper. Firstly, the preprocessing is implemented, which includes the colour-depth registration and depth image inpainting, to processing holes and noises in the depth image; secondly, the depth image is segmented to obtain different motion regions with different depth values; thirdly, scene flow is estimated based on the global nonrigid and local rigid assumption and spatial-temporal correlation of RGBD information. In the global nonrigid and local rigid assumption, each segmented region is divided into several blocks, and each block has a rigid motion. With this assumption, the interaction of motion from different parts in the same segmented region is avoided, especially the nonrigid object, e.g., a human body. Experiments are implemented on RGBD tracking dataset and deformable 3D reconstruction dataset. The visual comparison shows that the proposed method can distinguish the motion parts from the static parts in the same region better, and the quantitative comparisons proved more accurate scene flow can be obtained.

1. Introduction

Vedula et al. [1] proposed the scene flow first, which describes a 3D motion field formed by the motion in 3D space scene. Scene flow is the fundamental input to high-level tasks such as scene understanding and analysis. With the development and applications of computer vision and artificial intelligence, the related technologies have been used in the object detection and segmentation [2, 3], depth interpolation, and 3D reconstruction in many dynamic scenes, such as autonomous driving [4, 5], high-speed video generation [6], and 3D reconstruction [7].

Some research efforts have been dedicated to the estimation of the scene flow, which involve different environments, monocular vision [8], stereo vision [1, 2, 9, 10], and RGBD [11–13]. Affordable RGBD cameras can directly capture both colour and depth information simultaneously, so we focus on the RGBD scene flow estimation. Among the existing methods, methods based on segmentation are attractive, which can deal with large displacement and occlusion better. For this method, the correlation of motion in the local area is considered, such as the assumption of local rigid area, which can improve the accuracy of the scene flow estimation. In the local rigid area, it is assumed that all pixels in a segmented region share a rigid motion.

However, if the segmented region is a nonrigid object, pixels with different motion degrees would affect each other and then affect the overall scene flow estimation effect. In this paper, the local rigid and global nonrigid assumption in segmented regions is introduced into the RGBD scene flow estimation. In this assumption, the local motion in a segmented object area is correlated, and the motion of the whole segmented object is nonrigid. With this assumption, the interaction of motion from different parts in the same segmented region is avoided.

2. Related Work

According to the difference of solving process, these approaches are divided into two categories roughly: the variational approaches [1, 8, 12, 13], which construct the objective function on scene flow directly, and the methods based on segmentation with the assumption of local rigid motion [14–16].

The variational approaches estimate the dense scene flow with constraints of the spatial-temporal vision information commonly. An objective function is constructed to estimate the dense scene flow [1, 8, 17]. Xiao et al. [8] construct an objective function on scene flow in a monocular camera environment, which includes a brightness constancy assumption, a gradient constancy assumption, a short time object velocity constancy assumption, etc. Jaimez et al. [17] considered the depth information from the RGBD camera and presented a dense real-time scene flow algorithm with brightness constancy and geometric consistency.

The methods based on segmentation estimate the rigid motion of each segmented region, and then the local rigid motion and nonrigid motion are mixed to get dense scene flow [16, 18–20]. In [20], an efficient RGBD PatchMatch was used to solve large displacement motion patterns and stage, and further occlusion model and spatial smoothness regularization were used to compute the RGBD scene flow field. Golyanik et al. [18] presented a multiframe scene flow approach that assumes scene transformations to be locally rigid in RGBD image sequences. Xiang et al. [19] used a 3D local rigidity assumption to estimate the dense scene flow in a variational framework. Schuster et al. [21] interpolated the sparse matches between stereoscopic image pairs to estimate scene flow, in which the initial sparse match is the local rigid assumption actually.

Sun et al [16] proposed a layered RGBD scene flow method, in which the depth ordering from RGBD is used to segment the scene, and solved the occlusions. The layered RGBD scene flow method is a promising method as spatial smoothness is separated from the model of discontinuities and occlusions, which can model occlusion boundaries by obtaining the relative depth order. Depth image is layered based on the depth information. In order to estimate the motion of the scene, it assumed that pixels belonging to the same layer have the same rigid motion.

The result of estimating scene flow directly is high dimensional, so the solution space is large and the calculation complexity is high. And methods with the assumption of local rigid motion reduce the solution space. However, for most of the methods, the assumption of local rigid motion, a local region is semantic, such as a superpixel or a specific object. So the assumption cannot be well applied to nonrigid objects because the internal motion of nonrigid object is not consistent. In this paper, we propose an assumption of global nonrigid and local rigid motion based on the study of Sun et al. [16], which can accurately estimate the motion of each segmented region by dividing each segmented region into different blocks. Besides, affordable RGBD cameras provide both colour and depth information simultaneously. Therefore, we would focus on approaches with colour and depth information.

3. Methodology

In this section, a framework for estimating RGBD scene flow is shown in Figure 1. In this framework, two steps are presented to get scene flow: the preprocessing and the scene flow estimation.

The preprocessing mainly performs basic processing on the input RGBD image sequence (the red box in Figure 1), thus providing materials for estimating scene flow efficiently, involving the colour-depth registration and depth image inpainting. Details will be introduced in Section 3.1. The scene flow estimation would present the calculating processing of scene flow. It includes two parts: depth image segmentation and scene flow estimation with the preprocessing result and spatiotemporal constraints from the RGBD image sequence (Section 3.2).

3.1. Preprocessing. In the preprocessing, the color-depth registration and the depth image inpainting are implemented. The color-depth registration is used to associate the depth image with the RGB image. And the depth image inpainting is used to repair holes and noises, which are from occlusions, lack of point correspondences, sensor imperfection, etc.

3.1.1. Colour-Depth Registration [22]. To register the RGB image and depth image, a projective matrix M is calculated as shown in equation (1). In equation (1), (x, y) is a pixel coordinate in the depth image, and (X, Y) is the corresponding coordinate in the RGB image:

$$\begin{bmatrix} X \\ Y \\ 1 \end{bmatrix} = M \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} m_1 & m_2 & m_3 \\ m_4 & m_5 & m_6 \\ m_7 & m_8 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}. \tag{1}$$

Furthermore, equation (1) can be rewritten as follows:

$$X = m_1 x + m_2 y + m_3 - m_7 x X - m_8 y X, \tag{2}$$

$$Y = m_4 x + m_5 y + m_6 - m_7 x Y - m_8 y Y. \tag{3}$$

In M, eight unknown parameters need to be solved, and four pairs of corresponding points in the depth image and RGB image are needed at least. In our paper, corners are used.

3.1.2. Inpainting. To process the holes and noises in the depth image, the inpainting algorithm with the guidance of RGB image information is used [23]. In this algorithm, holes and small noises are all regarded as noises, but holes have larger connected areas and the depth value is 0, while small noises have smaller connected areas. In this paper, holes are inpainted based on depth domain similarity and colour consistency from the aligned depth image and RGB image. And small noises are removed with the local bilateral filter.

3.2. Scene Flow Estimation. In order to estimate scene flow accurately, the depth image is segmented into different regions roughly since there is stronger local motion correlation in the same region. Based on the inpainted depth image, K-means clustering algorithm is used to segment and label the depth image, by which scene can be quickly and simply segmented based on the depth information. The value of K depends on the number of moving regions in the scene.

To calculate the RGBD scene flow, an assumption of global nonrigid and local rigid motion is proposed to describe the behaviours of the scene in this paper. In a segmented region, the pixels' motion of its inner local area is highly consistent, so it is assumed that the local motion of a segmented region is rigid.

3.2.1. Global Nonrigid and Local Rigid Assumption. Each segmented region is divided into a number of sufficiently small blocks and the size of the block is 3×3 (Figure 2). In the global nonrigid and local rigid assumption, pixels in each block share the common 3D rigid motion R, which includes the rotation and the translation relative to the camera coordinate system (local rigid assumption), and different blocks have different motions (global nonrigid assumption).

Let a 2D point $p_1 = (x_1, y_1)$ at frame t, and its corresponding 2D point $p_2 = (x_2, y_2)$ in frame $t+1$. The depth values of p_1 and p_2 are z_1 and z_2, which are from the depth images. According to the camera imaging principle and the 2D-3D transformation model in [16], the corresponding 3D point $P_1 = (X_1, Y_1, Z_1)$ and $P_2 = (X_2, Y_2, Z_2)$ of p_1 and p_2 are as follows:

$$\begin{cases} X_1 = z_1 \cdot \dfrac{(x_1 - c_x)}{f_x, Y_1} = z_1 \cdot \dfrac{(y_1 - c_y)}{f_y, Z_1} = z_1, \\ \\ X_2 = z_2 \cdot \dfrac{(x_2 - c_x)}{f_x, Y_2} = z_2 \cdot \dfrac{(y_2 - c_y)}{f_y, Z_2} = z_2, \end{cases} \tag{4}$$

where $(f_x, f_y)^T$ and $(c_x, c_y)^T$ represent the camera focal length and distortion coefficient, respectively.

The rigid motion R from P_1 to P_2 can be expressed as follows:

$$P_2 = R \cdot \begin{bmatrix} P_1 \\ 1 \end{bmatrix}. \tag{5}$$

In equation 5, the image coordinate p_2 corresponding to the spatial point P_2 is given by

$$p_2 = \left(f_x \frac{X_2}{z_2} + c_x, f_y \frac{Y_2}{z_2} + c_y \right). \tag{6}$$

The corresponding local rigid RGBD scene flow from p_1 to p_2 is as follows:

$$\begin{aligned} u^R(p_1) &= f_x \frac{X_2}{z_2} + c_x - x_1, \\ v^R(p_1) &= f_y \frac{Y_2}{z_2} + c_y - y_1, \\ w^R(p_1) &= z_2 - z_1, \end{aligned} \tag{7}$$

where u, v, and w are the horizontal motion, vertical motion, and depth change of p_1.

Furthermore, a term on spatial constraints for scene flow is presented as follows:

$$\begin{aligned} E_{\text{spa}}(u_{tk}, v_{tk}, w_{tk}, R_{tk}) = {} & E_{\text{spa\_}u}(u_{tk}, R_{tk}) + E_{\text{spa\_}v}(v_{tk}, R_{tk}) \\ & + E_{\text{spa\_}w}(w_{tk}, R_{tk}), \end{aligned} \tag{8}$$

where

$$\begin{aligned} E_{\text{spa\_}u}(u_{tk}, R_{tk}) &= \sum_p \left(\sum_{p' \in N_p} \left(u_{tk}(p) - u_{tk}^R(p') \right)^2 + \sum_{p' \in N_p} \left(\left(u_{tk}(p) - u_{tk}^R(p) \right) - \left(u_{tk}(p') - u_{tk}^R(p') \right) \right)^2 \right), \\ E_{\text{spa\_}v}(v_{tk}, R_{tk}) &= \sum_p \left(\sum_{p' \in N_p} \left(v_{tk}(p) - v_{tk}^R(p') \right)^2 + \sum_{p' \in N_p} \left(\left(v_{tk}(p) - v_{tk}^R(p) \right) - \left(v_{tk}(p') - v_{tk}^R(p') \right) \right)^2 \right), \\ E_{spa\_w}(w_{tk}, R_{tk}) &= \sum_p \left(\sum_{p' \in N_p} \left(w_{tk}(p) - w_{tk}^R(p') \right)^2 + \sum_{p' \in N_p} \left(\left(w_{tk}(p) - w_{tk}^R(p) \right) - \left(w_{tk}(p') - w_{tk}^R(p') \right) \right)^2 \right), \end{aligned} \tag{9}$$

where u_{tk}, v_{tk}, and w_{tk} are the scene flow of in directions x, y, and z for the segmented region k at frame t, and N_p is 4 nearest spatial neighbours of the pixel p.

$E_{\text{spa\_}u}$, $E_{\text{spa\_}v}$, and $E_{\text{spa\_}w}$ reflect motion correlation in different directions within the same segmented region.

3.2.2. Spatiotemporal Correlation. Referring to the objective function in [16, 24, 25], the spatiotemporal correlation of the RGBD image sequence is also considered besides the global nonrigid and local rigid assumption. The spatial-temporal correlation of the RGBD image sequence contains two

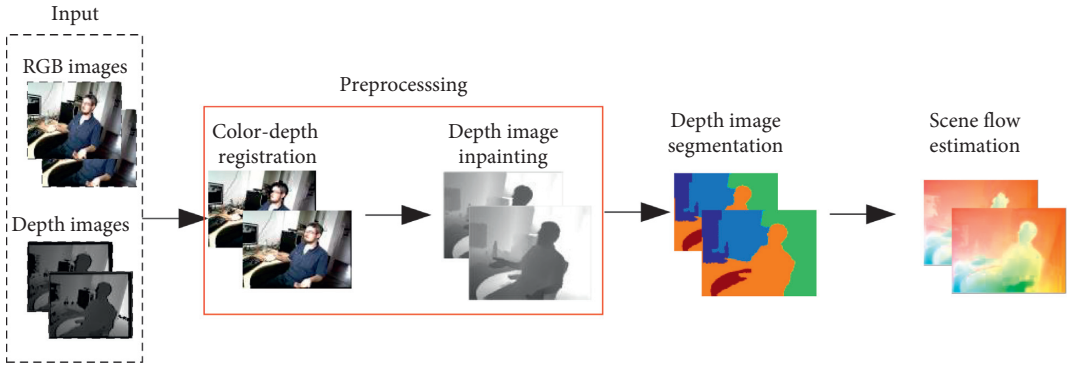


Figure 1: The framework for RGBD scene flow estimation. This framework shows the steps of RGBD scene flow estimation.

terms: the consistency of RGBD data and the coherence of the segmented regions.

(1) *The Consistency of RGBD Data.* If p is visible in frame t and $p + (u_{tk}(p), v_{tk}(p))$ is also visible in frame $t+1$ in the depth image and the aligned RGB image, the point has a constant appearance with the motion $(u_{tk}(p), v_{tk}(p), w_{tk}(p))$. The term consistency of RGBD data can be represented as follows:

$$E_{\text{data}}(u_{tk}, v_{tk}, w_{tk}) = \sum_{p} \Big((I_t(p) - I_{t+1}(p + (u_{tk}(p), v_{tk}(p))))^2 + (Z_t(p) + w_{tk}(p) - Z_{t+1}(p + (u_{tk}(p), v_{tk}(p))))^2 \Big), \tag{10}$$

(2) *The Coherence of the Segmented Region.* If p in frame t belongs to the segmented region k, $p + (u_{tk}(p), v_{tk}(p))$ in frame $t+1$ belongs to the segmented region k. The term coherence of the segmented region can be represented as follows:

$$E_{\text{sup}}(u_{tk}, v_{tk}, g_{tk}) = \sum_{p} \sum_{p' \in N_{(x,y)}} \left(g_{t,k}(p) - g_{t,k}(p') \right)^2 + \sum_{p} \left(g_{t,k}(p) - g_{t+1,k}(p + (u_{tk}(p), v_{tk}(p))) \right)^2, \tag{11}$$

where g_{tk} is a support function, which represents the probability size that a pixel belongs to the segmented region k in frame t.

According to equations (8), (10), and (11), a total objective function is constructed as follows:

$$E(u, v, w, g, R) = \sum_{t=1}^{T-1} \left(\sum_{k=1}^{K} \left(\lambda_{\text{data}} E_{\text{data}}(u_{tk}, v_{tk}, w_{tk}) + \lambda_{spa} E_{\text{spa}}(u_{tk}, v_{tk}, w_{tk}, R_{tk}) \right) \right) + \sum_{t=1}^{T} \sum_{k=1}^{K-1} \lambda_{\text{sup}} E_{\text{sup}}(u_{tk}, v_{tk}, g_{tk}), \tag{12}$$

where λ_{data}, λ_{spa}, and λ_{sup} represent the corresponding weight of E_{data}, E_{spa}, and E_{sup}, respectively.

The coordinate descent method is used to minimize the RGBD scene flow energy function in equation (12). Firstly, estimate the initial scene flow according to the interframe optical flow and segmentation of the depth image. Secondly, obtain the optimized scene flow by image warping while keeping the layering result fixed. Thirdly, calculate the optimized layered support function with coordinate descent method while keeping the scene flow fixed. Finally, get the final scene flow by looping the second and third operations.

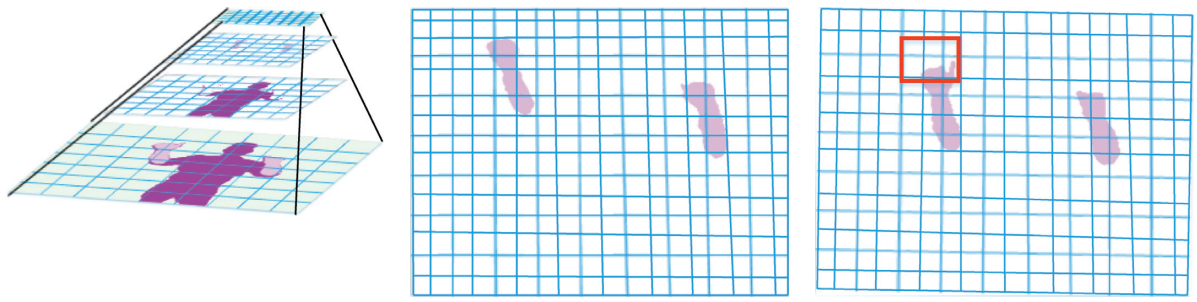
4. Experiments

In this section, the performance of the proposed method is evaluated by analysing the results without the assumption of global nonrigid and local rigid. Then, the method is implemented on Princeton Tracking Benchmark and Deformable 3D reconstruction dataset, and some qualitative or quantitative comparisons are presented.

4.1. Performance Evaluation on the Term on Spatial Constraints of Scene Flow. The term on spatial constraints of scene flow reflects the relationship between the scene flow of a pixel and its neighbourhood. To evaluate its performance, some experiments are implemented without this term.

In Figures 3 and 4, the scene flow is estimated without the spatial constraints of scene flow. For clarity, the figures for scene flow are not shrunk too much. It is obvious that scene flow loses smoothness in the same segmented region. That means the scene flow of pixels in the same region is discontinuous since the correlation of scene flow of pixels in the same region is not considered.

4.2. Princeton Tracking Benchmark. This dataset contains multiple independent moving targets and large areas of occlusion [28]. In this section, "Bear_back" sequence is used to test the method in this paper, and the results are shown in Figure 3. In the "Bear_back" sequence, the motion of the scene is produced by the opposite movement



(a) (b) (c)

Figure 2: Illustration for the assumption of global nonrigid and local rigid motion: (a) the depth layered image; (b) the layered result of the t^{th} frame k^{th} layer and (c) the layered result of the $t+1^{\text{th}}$ frame k^{th} layer. The block size is $3 * 3$.

of two hands mainly, in addition to some slight motion of the body.

In Figure 5, the first two columns are two consecutive images from Bear_back sequence, including RGB and depth images. The third column is the segmentation results, and K is set to 5 in the K-means clustering algorithm. The first and second rows of the fourth and fourth columns are the results of Sun's method [16] and our method, respectively. By comparing the scene flow in the red box between Sun's method and ours, it can be found that, under the same segmentation condition, the proposed method is closer to the moving region of the real image.

4.3. Deformable 3D Reconstruction Dataset [26]. Deformable 3D reconstruction dataset is a nonrigid dataset. In this paper, "Hat" and "Alex" sequence are used to test the proposed method, and different poses from different times are selected in these two sequences, respectively, to validate the proposed method is invariant to pose variation.

In "Hat" sequence, the motion is caused by the off-cap behaviour, and two poses are used, which is called Pose 1 and Pose 2. Pose 1 has small amplitude, involving the slight motion of hat, arm, and twist (Figure 6). Pose 2 includes the motion of hat mainly, and the direction of scene flow is the same basically (Figure 7). In Figures 6 and 7, the first two columns are the consecutive RGB and depth images, the third column is the segmented results with $K = 2$ in the K-means clustering algorithm, and the fifth column is the scene flow of Sun's method and ours. Occlusion calculation is an important part of Sun's method; therefore, the occlusions are also presented in this section.

In the fifth column of Figures 6 and 7, the estimation result of scene flow with Sun's method covers the whole human body which contains some stationary part. The reason for this problem may be that pixels in the same segmented region share a common rigid motion, which results in pixels without motion are also estimated scene flow. However, our method can estimate the scene flow of motion part, such as arm, head, and hat because each segmented region is divided into different blocks and the scene flow is estimated based on 3×3 block in each segmented region.

In "Alex" sequence, Pose 3 and Pose 4 are used. Pose 3 is produced by waving arms and some movement of clothes (Figure 8), and Pose 4 is obtained by the motion of arms (Figure 9). In the segmentation of "Alex" sequence, K is also set to 2. In Pose 3 and Pose 4, the motion amplitude of arms is greater than the rest of the human body. In Sun's method, the motion amplitude of the whole body is considered to be the same; however, the motion of arms is significantly greater than the rest of the body.

By comparing the scene flow estimation results visually (Figures 6~9), it can be found that our method can accurately estimate the scene flow of the nonrigid objects which involves different motion parts.

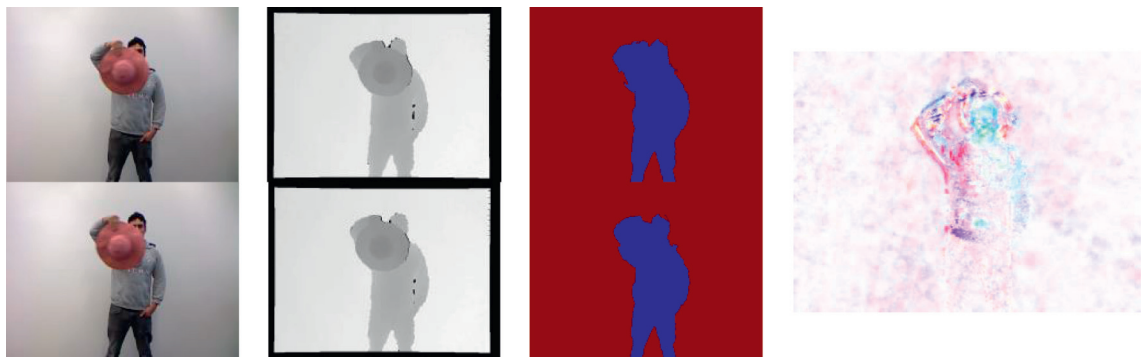
4.4. Evaluation Results. Quantitative results, RMS and AAE, are used to compare the proposed method and Sun's method.

RMS and AAE traverse all the pixels in the image, map the 3D scene flow acquired by the algorithm into a 2D optical flow, and compare it with the real optical flow value. The smaller the difference is, the more accurate the calculation is. Let that the estimated optical flow is $(u, v)^T$, and the true optical flow is $(u_{GT}, v_{GT})^T$, then the calculation formula of the RMS and AAE is as follows:

$$\text{RMS} = \sqrt{\frac{1}{N}\sum_{(x,y)}\left(\left(u_{GT}(x,y) - u(x,y)^2\right) + \left(v_{GT}(x,y) - v(x,y)\right)^2\right)},$$

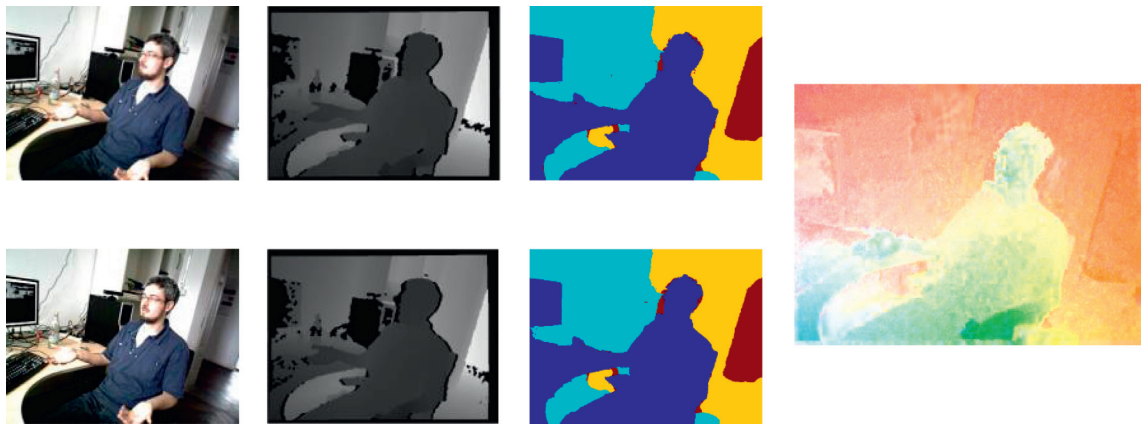
$$\text{AAE} = \frac{1}{N}\arccos\left(\frac{1 + u_{GT} \times u + v_{GT} \times v}{\sqrt{u_{GT}^2 + v_{GT}^2 + 1} \cdot \sqrt{u^2 + v^2 + 1}}\right), \tag{13}$$

where N is the number of pixels in the image.



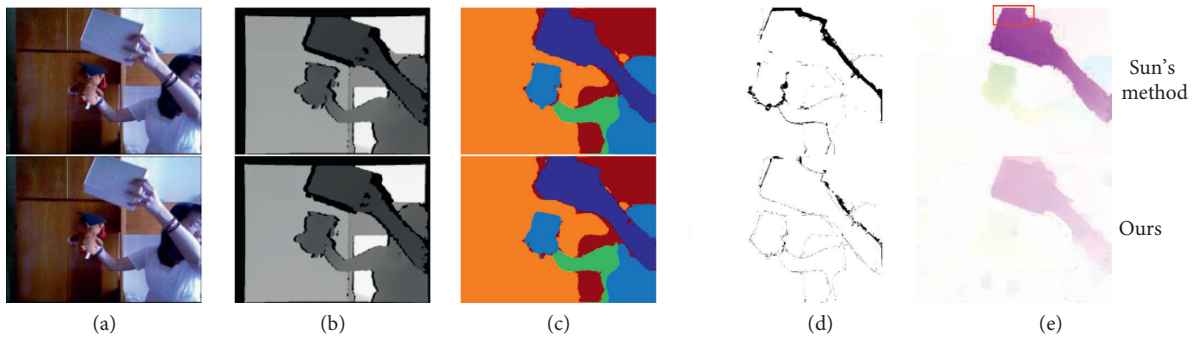
(a) (b) (c) (d)

Figure 3: "Hat" sequence [26] without spatial constraints of scene flow. Two consecutive frames from "Hat" sequence are input and segmented into 2 regions to estimate scene flow. (a) RGB images. (b) Depth images. (c) Segmentation $K = 2$. (d) Scene flow.



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

Figure 4: SRSF 20 sequence [27] without spatial constraints of scene flow. Two consecutive frames from SRSF 20 sequence are input and segmented into 4 regions to estimate scene flow. (a) RGB images. (b) Depth images. (c) Segmentation $K = 4$. (d) Scene flow.



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

Figure 5: "Bear_back" sequence test results. Two consecutive frames from "Bear_back" sequence are input and segmented into 5 regions to estimate occlusion and scene flow. (a) RGB images. (b) Depth images. (c) Segmentation $K = 5$. (d) Occlusions. (e) Motion.

Errors of the method in this paper and Sun's are shown, respectively, in Figures 10(a) and 10(b), where the blue bar represents the errors of ours and the orange bars are the errors of Sun's method. From Figure 9, it is obvious that the blue bars are shorter than the orange bars, that is, RMS and AAE of the proposed method are lower than those of Sun's method in the test datasets.

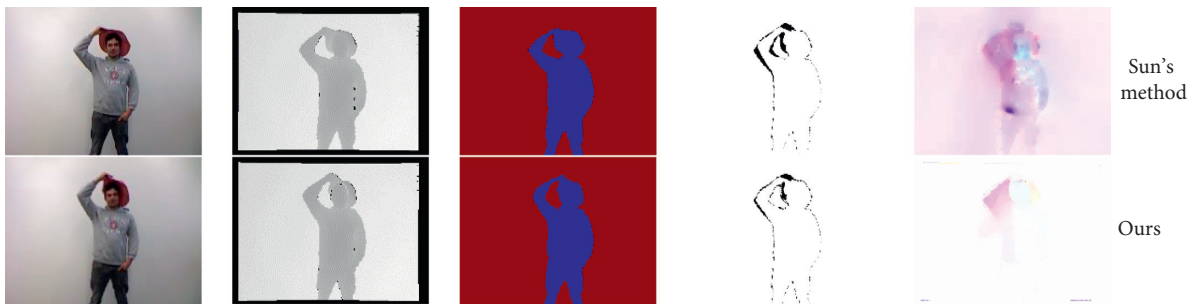


Figure 6: Pose 1 in "Hat" sequence." Input two consecutive frames about Pose 1 in "Hat" sequence, segment them into 2 regions (human body with hat and background), and estimate the occlusion and scene flow. (a) RGB images. (b) Depth images. (c) Segmentation $K = 2$. (d) Occlusions. (e) Motion.

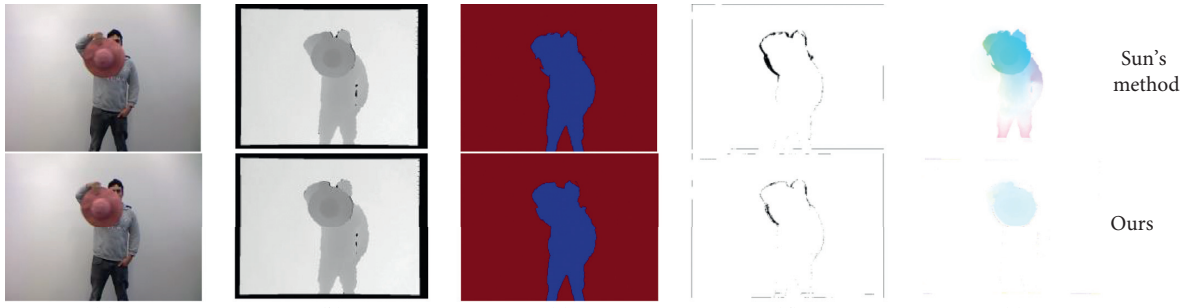


Figure 7: Pose 2 in "Hat" sequence." Input two consecutive frames about Pose 2 in "Hat" sequence. (a) RGB images. (b) Depth images. (c) Segmentation $K = 2$. (d) Occlusions. (e) Motion.

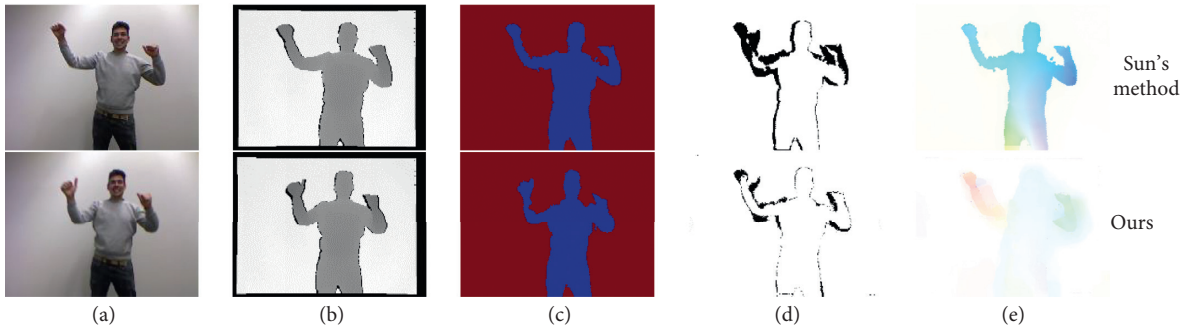


Figure 8: Pose 3 in "Alex" sequence. Input two consecutive frames about Pose 3 in "Alex" sequence. (a) RGB images. (b) Depth images. (c) Segmentation $K = 2$. (d) Occlusions. (e) Motion.

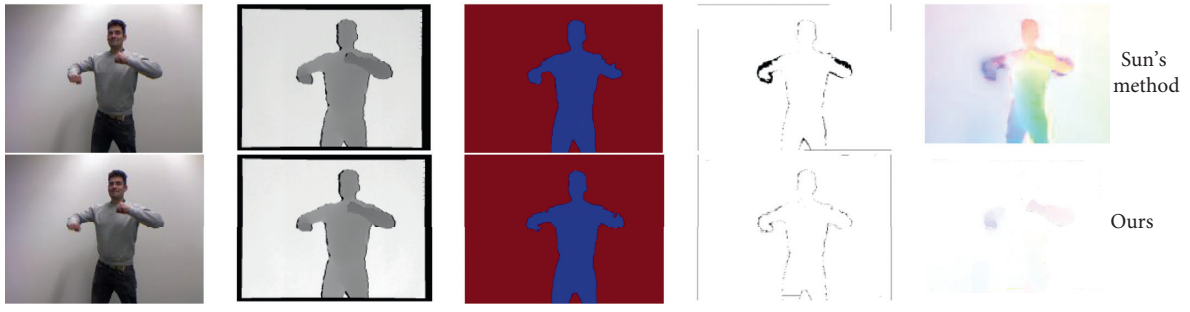


Figure 9: Pose 4 in "Alex" sequence. Input two consecutive frames about Pose 4 in "Alex" sequence. (a) RGB images. (b) Depth images. (c) Segmentation $K = 2$. (d) Occlusions. (e) Motion.

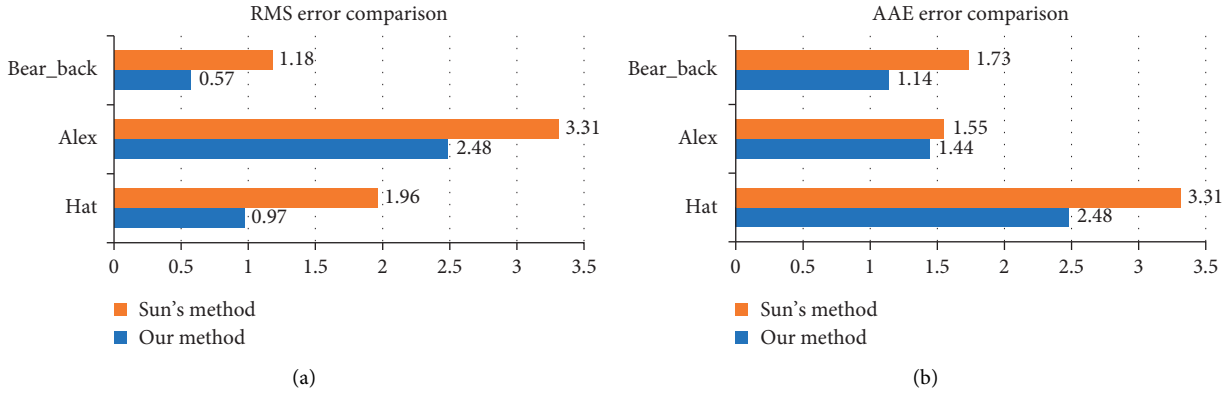


Figure 10: Comparisons of RMS and AAE. (a) RMS error. (b) AAE error.

5. Conclusions

In this paper, a RGBD scene flow estimation method with global nonrigid and local rigid motion assumption is presented. In this method, the preprocessing and the scene flow estimation are carried out. The preprocessing is used to get the registered RGB image and depth image, which would provide material for estimating scene flow. In the scene flow estimation, the K-means clustering algorithm is used to segment the depth image and process the occlusions, and then scene flow is estimated with the spatial-temporal correlation of the RGBD image sequence and global nonrigid and local rigid assumption in each segmentation region. To represent the global nonrigid and local rigid assumption, each segmented region is divided into a number of sufficiently small blocks since the pixels' motion in the same block is consistent and the pixels' motion in the different block is inconsistent. Experiments on different datasets and different poses show that the scene flow can be estimated more accurately with the proposed method.

However, the running time of the code is longer than [16] because each segmented region is divided into different blocks. In the future work, we will refer to the optimization of the model. For trained deep neural network methods can predict scene flow rapidly, we will refer to the existing methods to study learning-based methods.

Data Availability

The data used to support the findings of this study are available at http://tracking.cs.princeton.edu/dataset.html and http://campar.in.tum.de/personal/slavcheva/deformable-dataset/index.html.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The authors thank Deqing Sun et al. for providing the code. The authors would also gratefully acknowledge the support from the "International Conference on Brain Inspired Cognitive Systems-BICS" in 2019 for presenting our abstract which can be found in https://link.springer.com/chapter/10.1007/978-3-030-39431-8_21 [29]. This work was supported by the National Natural Science Foundation of China under grant nos. 6150238, 61501370, and 61703333.

References

[1] S. Vedula, S. Baker, P. Rander, R. Collins, and T. Kanade, "Three-dimensional scene flow," in *Proceedings of the Seventh IEEE International Conference on IEEE Computer Vision*, vol. 2, pp. 722–729, Kerkyra, Greece, September 1999.

[2] D. Zhou, V. Frémont, B. Quost, Y. Dai, and H. Dai, "Moving object detection and segmentation in urban environments from a moving platform," *Image and Vision Computing*, vol. 68, pp. 76–87, 2017.

[3] S. Javed, T. Bouwmans, M. Shah, and S. Jung, "Moving object detection on RGB-D videos using graph regularized spatio-temporal RPCA," in *Proceedings of the International Conference on Image Analysis and Processing*, pp. 1–11, Catania, Italy, September 2017.

[4] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3354–3361, Rhode Island, USA, June 2012.

[5] M. Menze and A. Geiger, "Object scene flow for autonomous vehicles," in *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (CVPR)*, pp. 3061–3070, Boston, MA, USA, June 2015.

[6] X. Zuo, S. Wang, J. Zheng, and R. Yang, "High-speed depth stream generation from a hybrid camera," in *Proceedings of the 2016 ACM International Conference on Multimedia (ACM MM)*, pp. 878–887, Klagenfurt Austria, May 2016.

[7] S. Wang, X. Zuo, C. Du, R. Wang, J. Zheng, and R. Yang, "Dynamic non-rigid objects reconstruction with a single RGB-D sensor," *Sensors*, vol. 18, no. 3, p. 886, 2018.

[8] D. Xiao, Q. Yang, B. Yang, and W. Wei, "Monocular scene flow estimation via variational method," *Multimedia Tools and Applications*, vol. 76, no. 8, pp. 10575–10597, 2017.

[9] F. Huguet and F. Devernay, "A variational method for scene flow estimation from stereo sequences," in *Proceedings of the the 11th International Conference on Computer Vision*, pp. 1–7, Rio de Janeiro, Brazil, October 2007.

[10] C. Vogel, K. Schindler, and S. Roth, "Piecewise rigid scene flow," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1377–1384, Sydney, Australia, December 2013.

[11] J. Quiroga, F. Devernay, and J. Crowley, "Scene flow by tracking in intensity and depth data," in *Proceedings of the CVPRW 2012–Computer Vision and Pattern Recognition Workshops IEEE*, pp. 50–57, Ostrava, Czech Republic, January 2012.

[12] S. Hadfield and R. Bowden, "Kinecting the dots: particle based scene flow from depth sensors," in *Proceedings of the 2012 IEEE Conference on Computer Vision*, pp. 2290–2295, Rhode Island, USA, June 2012.

[13] J-M Gottfried, J. Fehr, and C. S. Garbe, "Computing range flow from multi-modal kinect data," in *Proceedings of the International Symposium on Visual Computing*, pp. 758–767, Las Vegas, NV, USA, September 2011.

[14] T. Taniai, S. N. Sinha, and Y. Sato, "Fast multi-frame stereo scene flow with motion segmentation," in *Proceedings of the Computer Vision and Pattern Recognition (CVPR)*, pp. 6891–6900, Honolulu, HI, USA, July 2017.

[15] Z. Lv, C. Beall, P. F. Alcantarilla et al., "A continuous optimization approach for efficient and accurate scene flow," in *Proceedings of the European Conference on Computer Vision*, pp. 757–773, Amsterdam, The Netherlands, October 2016.

[16] D. Sun, E. B. Sudderth, and H. Pfister, "Layered RGBD scene flow estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 548–556, Boston, MA, USA, June 2015.

[17] M. Jaimez, M. Souiai, J. Gonzalez-Jimenez, and D. Cremers, "A primal-dual framework for real-time dense RGB-D scene flow," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pp. 98–104, Seattle, WA, USA, May 2015.

[18] V. Golyanik, K. Kim, R. Maier et al., "Multiframe scene flow with piecewise rigid motion," in *Proceedings of the International Conference on 3D Vision (3DV)*, pp. 273–281, Qingdao, China, October 2017.

[19] X. Xiang, M. Zhai, R. Zhang, W. Xu, and A. EI Saddik, "Scene flow estimation based on 3D local rigidity assumption and depth map driven anisotropic smoothness," *IEEE Access*, vol. 6, pp. 30012–30023, 2018.

[20] Y. Wang, J. Zhang, Z. Liu et al., "Handling occlusion and large displacement through improved RGB-D scene flow estimation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 7, pp. 1265–1278, 2016.

[21] R. Schuster, O. Wasenmuller, G. Kuschk, C. Bailer, and D. Stricker, "Sceneflowfields: dense interpolation of sparse scene flow correspondences," in *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1056–1065, Lake Tahoe, NV, USA, March, 2018.

[22] W. Song, A. V. Le, S. Yun, S-W. Jung, and C. S. Won, "Depth completion for kinect v2 sensor," *Multimedia Tools and Applications*, vol. 76, no. 3, pp. 4357–4380, 2017.

[23] Y. Zhang, J. Dai, H. Zhang, and L. Yang, "Depth inpainting algorithm of RGB-D camera combined with color image," in *Proceedings of the 2nd IEEE Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC)*, pp. 1391–1395, Xi'an, Shaanxi, China, May 2018.

[24] D. Sun, E. Sudderth, and M. Black, "Layered image motion with explicit occlusions, temporal consistency, and depth ordering," in *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*, pp. 2226–2234, Vancouver, British Columbia, Canada, December 2010.

[25] D. Sun, J. Wulff, E. Sudderth, H. Pfister, and M. Black, "A fully-connected layered model of foreground and background flow," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2451–2458, Portland, OR, USA, June 2013.

[26] M. Slavcheva, M. Baust, D. Cremers, and S. Ilic, "KillingFusion: non-rigid 3D reconstruction without correspondences," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5474–5483, Honolulu, HI, USA, July 2017.

[27] J. Quiroga, T. Brox, F. Devernay, and J. Crowley, "Dense semi-rigid scene flow esti- mation from rgbd images," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 567–582, Zürich, Switzerland, September 2014.

[28] http://tracking.cs.princeton.edu/dataset.html.

[29] X. Li, Y. Liu, H. Jin, L. Cai, and J. Zheng, "Layered RGBD scene flow estimation with global non-rigid local rigid assumption," in *Advances in Brain Inspired Cognitive Systems*, vol. 11691, BICS, Brussels, Belgium, 2019.