Hindawi
International Journal of Digital Multimedia Broadcasting
Volume 2022, Article ID 3582037, 8 pages
https://doi.org/10.1155/2022/3582037

Research Article

A Convolutional Neural Network for Nonrigid Structure from Motion

Yaming Wang,[1,2] Xiangyang Peng,[1] Wenqing Huang,[1] and Meiliang Wang[2]

[1]*Pattern Recognition and Computer Vision Lab, Zhejiang Sci-Tech University, Hangzhou 310000, China*
[2]*Zhejiang Key Laboratory of DDIMCCP, Lishui University, No. 1 Xueyuan Road, Lishui 323000, China*

Correspondence should be addressed to Wenqing Huang; patternrecog@163.com

Received 23 December 2021; Revised 27 March 2022; Accepted 8 April 2022; Published 28 April 2022

Academic Editor: Sayyouri Mhamed

In this study, we propose a reconstruction and optimization neural network (RONN), a novel neural network for nonrigid structure from motion, which is completed by an unsupervised convolution neural network. Compared with the traditional method for directly solving 3D structures, our model focuses on depth information that is lost owing to projection. This mathematical model is developed using a convolutional neural network with three modules for integration, reconstruction, and optimization, as well as two prior-free loss functions. The proposed RONN achieves competitive accuracy on several tested sequences and high visual quality of various real video sequences.

1. Introduction

Nonrigid structure from motion (NRSfM) targets the recovery of a nonrigid structure and camera matrix from given 2D point tracks in monocular views. Unlike its rigid counterpart [1], NRSfM is a highly ill-posed problem with several inherent ambiguities. Moreover, solving this problem requires additional constraints or priors. Many methods assume that the movement of the camera is slow and smooth [2–6]; however, this limits its applicability to real sequences. Another assumption is that the deformation of nonrigid instances can be represented using the weighted sum of basic deformations in the trajectory space [4] and shape space [7]. With these assumptions, the NRSfM problem is transformed into solving the basic deformation and its coefficients.

Inspired by these assumptions, many researchers have used neural networks to solve the sparse NRSfM problem [8, 9], which learns shape representations through unsupervised networks, while maintaining good generalization ability in the face of unseen data. However, their models are incapable of handling dense situations.

Dense NRSfM has achieved remarkable progress over the last several years [2, 10–13]. In 2020, Sidhu et al. [13] proposed the first dense neural NRSfM (N-NRSfM) approach with mean shape and demonstrated state-of-the-art performance on widely used datasets. However, when confronted with a long sequence or drastic changes, the mean shape is unreasonable. Additionally, it requires a considerable amount of time to obtain high-performance results.

In this study, we introduce a reconstruction and optimization neural network (RONN) and two improved loss functions for dense NRSfM. RONN mainly includes a depth reconstruction module and a camera optimization module, which reconstruct the depth information lost due to projection and optimize the camera matrix, respectively. Inspired by recent advances in NRSfM [8], the proposed improved loss function is combined with the minimum singular value ratio, and experiments show that it improves the original loss function to varying degrees.

The main contributions of our study are as follows.

(1) We propose the first dense NRSfM network for reconstruction using depth information, namely, RONN. It is a convolutional neural network including reconstruction and optimization, which realizes the reconstruction of the 3D structure and the optimization of the camera matrix, respectively. Its specific structure will be introduced in Section 4.1.

Compared to directly solving the overall 3D structure in the method [13], RONN avoids the use of average shapes and reduces the amount of theoretical computation

(2) For the first time, we changed the input of the network from every frame to every point, enabling the network to cope with datasets of different sizes. Section 5.3 shows that RONN reconstructs dense and sparse 3D structures without 3D supervision and achieves competitive accuracy on multiple test sequences

(3) Compared with the original loss function, the weighted loss function using msr can handle complex deformation and further improve the reconstruction accuracy. The weighting method will be described in detail in Section 4.2. The comparative experiments in Section 5.2 show varying degrees of improvement

2. Related Work

2.1. NRSfM. NRSfM is inherently ill-conditioned and requires additional constraints or priors to guarantee solution uniqueness. We are concerned with the following additional limitations:

(1) Bregler et al. [14] proposed a low rank, where the rank of the rigid 3D structure fixed is three. Dai et al. [7] rearranged the rows of S as $S^{\#}$ to obtain stronger low rank priors, demonstrating state-of-the-art performance on sparse datasets at the time. Ansari et al. [10] proposed scalable monocular surface reconstruction (SMSR) with an improved low rank. Its scalability enables the achievement of competitive accuracy on both sparse and dense data

(2) Park et al. [15] proposed Procrustean regression, which is a regression problem based on Procrustes-aligned shapes. In [15], they proposed a novel regression framework for NRSfM, comprising Procrustes-aligned shape loss and low rank loss. The framework is versatile and can reconstruct a 3D structure under dense datasets. Additionally, Park et al. [16] proposed a novel framework for training neural networks with a Procrustean regression. Although the network structure is simple, it shows superior reconstruction performance compared to the state-of-the-art method. In [16], it was proven that Procrustes alignment could determine unique motions and eliminate the rigid motion components from reconstructed shapes

2.2. Neural NRSfM. There have been studies on combing NRSfM with neural networks. Supervised neural networks require large amounts of training data; however, only a few datasets are currently available for quantitative evaluation of NRSfM methods. In contrast, unsupervised networks are easier to implement. C3DPO [17] and deep NRSfM [9] learn basis shapes from 2D observations without 3D supervision in sparse data sets. C3DPO [17], which was proposed by Facebook's AI lab, uses a factorization network to replace the classical factorization step. Additionally, to ensure the effect of factorization, it collaborates with another canonicalization network to achieve a robust 3D reconstruction effect. This framework achieved high-performance reconstruction results for rigid and nonrigid datasets. Kong and Lucey [9] proposed a new a priori hypothesis, using multilayer sparse coding to represent 3D nonrigid shapes, and designed an innovative encoder-decoder neural network to realize an unsupervised network for NRSfM. They extended the classic sparse coding algorithm, ISTA, to block sparse scenarios and provide state-of-the-art performance through the proposed network. However, sparse coding also limits its application to dense datasets. Sidhu et al. [13] introduced the first dense neural NRSfM approach, namely, N-NRSfM, and achieved competitive performance on widely used dense datasets. They used the mean shape to achieve reconstruction, but this became a limitation. Once the mean shape is determined, the reconstruction result is obtained. Therefore, when confronted with large-scale deformation, the reconstruction results are not as good as expected.

3. Mathematical Model

Consider a monocular camera for observing a nonrigid object with a set of P feature points. Let S_f be the 3D shape matrix of the nonrigid object at the fth frame and W_f be its 2D matrix according to an orthographic projection. Specifically,

$$S_f = \begin{bmatrix} x_f^1 & \cdots & x_f^P \\ y_f^1 & \cdots & y_f^P \\ z_f^1 & \cdots & z_f^P \end{bmatrix}, S_f \in \mathbb{R}^{3\times P}, \tag{1}$$

$$W_f = \begin{bmatrix} u_f^1 & \cdots & u_f^P \\ v_f^1 & \cdots & v_f^P \end{bmatrix}, W_f \in \mathbb{R}^{2\times P}. \tag{2}$$

W_f and S_f are related by the full rotation matrix G_f as

$$W_f = \Pi \bullet G_f \bullet S_f, \Pi = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, G_f \in \mathbb{R}^{3\times 3}, \tag{3}$$

where W_f is already centralized; therefore, the camera matrix is reduced to pure rotation [14]. According to Formula (3), W_f contains both the camera and 3D structure information. In this study, a reasonable network architecture was designed to separate the required information.

Formula (3) can be changed to the following form.

$$S_f = G_f^T \bullet \begin{bmatrix} W_f \\ \varkappa_f \end{bmatrix}. \tag{4}$$

According to Formula (4), the full rotation matrix will have a certain effect on the reconstruction results; therefore, optimization of the full rotation matrix is necessary.

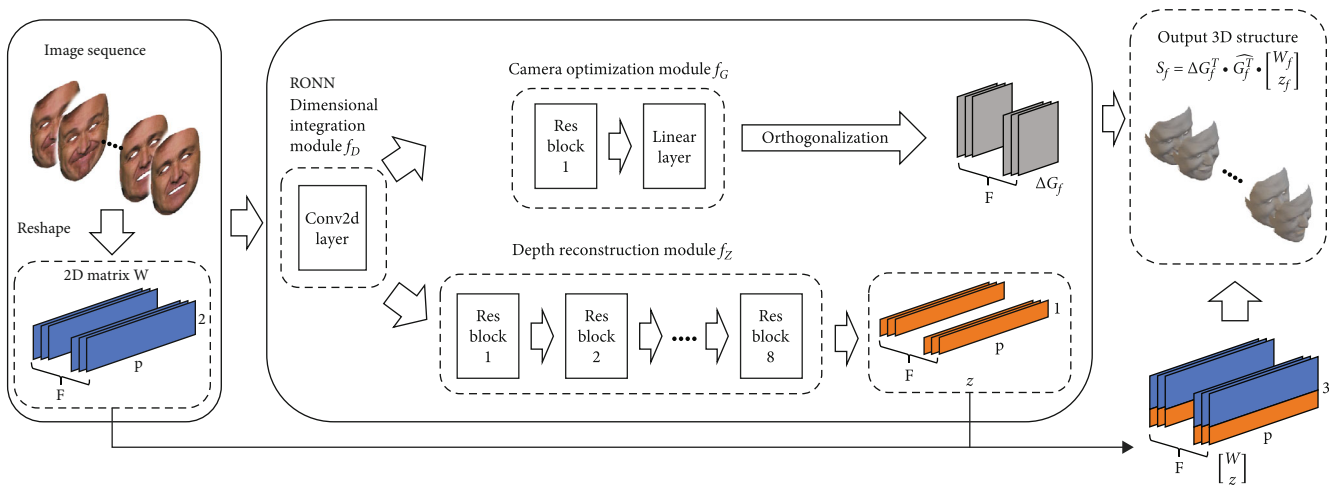


Figure 1: Overview of the proposed RONN.

4. RONN Model

In this section, we introduce the structure and loss functions of the RONN.

4.1. Network Structure. As illustrated in Figure 1, inspired by the trajectory space, our RONN contains three modules: dimensional integration module f_D, camera optimization module f_G, and depth reconstruction module f_z. The dimensional integration module integrates the information contained in u and v in the W matrix, f_G represents the optimization module for the camera matrix, and f_z represents the reconstruction module for the depth information z.

The 2D matrix W first passes through a dimensional integration module, which consists of a convolutional layer with a kernel size of 2×1 and a ReLU layer. The next two modules are f_G, which has B (set to 1 by default) residual blocks and a linear layer after rearranging the shape, from which we obtain $\widetilde{\Delta G_f} \in \mathbb{R}^{3\times 3}$, and f_z, which has $Kz_f \in \mathbb{R}^{1\times P}$. Specifically, the residual block contains two convolutional layers of kernel size 1×1 and a ReLU layer.

Through this network, our reconstruction result can be expressed as

$$S_f = \Delta G_f^T \bullet \widehat{G_f^T} \bullet \begin{bmatrix} W_f \\ z_f \end{bmatrix}, \Delta G_f = UV^T, \tag{5}$$

where $U\Sigma V^T = \widetilde{\Delta G_f}$ is the SVD of the camera matrix.

In this study, the method for initializing $\widehat{G}$ is the same as that in [1] on dense datasets and [7] on sparse datasets.

4.2. Loss Function. To solve the NRSfM problem, we propose minimizing the loss functions with the initial rotation matrix G as

$$E = \alpha E_{\text{temp}} + \beta E_{\text{PA}} + \gamma E_{\text{rank}}, \tag{6}$$

where $\{E_{\text{temp}}, E_{\text{PA}}, E_{\text{rank}}\}$ encode the additional constraints.

The temporal smoothness term E_{temp} is used to constrain the similarity of the reconstruction results of adjacent frames as

$$E_{\text{temp}} = \frac{1}{F-1} \sum_{f=1}^{F-1} \left\| \mu_f \left(S_{f+1} - S_f \right) \right\|_\epsilon, \tag{7}$$

where $\|\bullet\|_\epsilon$ denotes the Huber loss of the matrix. Weight μ_f is discussed later.

The Procrustean alignment term E_{PA} can determine unique motions and eliminate the rigid motion components from reconstructed shapes [15]; the term can be expressed as follows.

$$\begin{gathered} E_{PA} = \frac{1}{F} \sum_{f=1}^{F} \left\| S_f \bullet T - \bar{S}_f \right\|_\epsilon, \\ \bar{S}_f = \sum_{i \in A} \varphi_f^i \bullet S_f \bullet T, \sum_{i \in A} \varphi_f^i = 1, A = [1, f-1] \cup [f+1, F], \end{gathered} \tag{8}$$

where $T = (I - (1/P)11^T)$ is the translation matrix, centering the shape at the origin. Weight φ_f^i is discussed later in this paper. This function is aimed at minimizing the error between the 3D shape of each frame and the reference shape to optimize the rotation matrix.

A rearranged shape matrix is expressed as

$$S^{\#} = \begin{bmatrix} X_1^1 & \cdots & X_1^P & Y_1^1 & \cdots & Y_1^P & Z_1^1 & \cdots & Z_1^P \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ X_F^1 & \cdots & X_F^P & Y_F^1 & \cdots & Y_F^P & Z_F^1 & \cdots & Z_F^P \end{bmatrix}, \tag{9}$$

with an additional constraint rank $(S^{\#}) < K$. In [10], they assumed that the mean 3D component was dominant in $S^{\#}$ and could be removed in the temporal dimension. By combining both ideas, E_{rank} is defined as

Table 1: e_{3D} on baseline.

	Baseline		
	$\{E_{\text{temp}}\}$	$\{E_{\text{PA}}, E_{\text{rank}}\}$	$\{E_{\text{temp}}, E_{\text{PA}}, E_{\text{rank}}\}$
Traj.A	0.0506	0.0457	0.0467
Traj.B	0.0670	0.0456	0.0459

Table 2: e_{3D} on different loss function combinations.

	Baseline		RONN	
	$\{E_{PA}, E_{\text{rank}}\}$	$\{E_{PA}, E_{\text{rank}}, E_{\text{temp}}\}$	$\{E_{PA}, E_{\text{rank}}\}$	$\{E_{PA}, E_{\text{rank}}, E_{\text{temp}}\}$
Traj.A	0.0457	0.0467	0.0312	0.0309
Traj.B	0.0456	0.0459	0.0379	0.0359

$$E_{\text{rank}} = \left\| P \bullet S^{\#} \right\|_{*}, \tag{10}$$

where $P = (I - (1/F)11^T)$ is the orthogonal projection, and 1 is a vector of ones.

When using the optimized module f_G, E_{PA} must be used as the data term and E_{rank} as the regularization term to form a Procrustean regression.

Weights μ_f and φ_f^i are set using the minimal singular-value ratio [8]. Given two 2D matrices, W_i and W_j, let $A_j^i \in \mathbb{R}^{4\times P}$ be the stacked matrix of W_i and W_j as follows:

$$A_j^i = \begin{bmatrix} W_i \\ W_j \end{bmatrix}. \tag{11}$$

Then, the ratio of the minimal singular value of A is used to define the rigidity measure msr as follows:

$$\text{msr}\left(W_i, W_j\right) = \frac{\sigma_4^2}{\sum_{l=1}^{4} \sigma_l^2}, \tag{12}$$

where σ_l is the l-th singular value of A_j^i in descending order.

Then, weights μ_f and φ_f^i are defined as follows:

$$\mu_f = -\log\left(\text{msr}\left(W_f, W_{f+1}\right)\right), \tag{13}$$

$$\varphi_f^i = \frac{\log\left(\text{msr}\left(W_f, W_i\right)\right)}{\sum_{j\in A} \log\left(\text{msr}\left(W_f, W_j\right)\right)}, \quad A = [1, f-1] \cup [f+1, F]. \tag{14}$$

5. Experiment

In this section, the experimental results are described for several widely used benchmarks and real datasets. First, we introduce the datasets and experimental setups, then analyse and compare the proposed model with state-of-the-art dense and sparse datasets and, finally, use real data for experiments.

5.1. Datasets and Setups

5.1.1. Datasets. Three dense benchmark datasets are used in the comparison of methods: synthetic faces (two sequences with 99 frames and two different camera trajectories denoted by Traj.A and Traj.B, with 28,887 points per frame) [2], expressions (384 frames with 997 points per frame) [18], and actor mocap (100 frames with 36,349 points per frame) [19].

5.1.2. Evaluation Metrics. For algorithm performance indicators, the 3D error e_{3D} defined as follows.

$$e_{3D} = \frac{1}{F} \sum_{f=1}^{F} \frac{\left\| S_f^{GT} - S_f \right\|_F}{\left\| S_f^{GT} \right\|_F}, \tag{15}$$

where $\|\bullet\|_F$ denotes the Frobenius norm and S_f^{GT} denotes the ground truth 3D structure at the fth frame.

5.1.3. Training Details. The RONN was implemented in PyTorch [20]. We used the Adam optimizer with a learning rate of 0.0005 and trained for 2000 epochs. In the experiment, the weight was fixed at $\alpha = \beta = \gamma = 1$.

5.2. Model Analysis

5.2.1. Structure of RONN. The baseline was formed by removing the f_G module in the RONN. These experiments show the advantages of f_z and the necessity of f_G; it contains different combinations of loss functions on *synthetic face* sequences (Traj.A and Traj.B).

The advantages of f_z are listed in Table 1. Because the f_z network solves the depth information z, rather than the entire 3D structure, this allows the reconstruction to be achieved only with E_{temp}, although the error is relatively large.

The necessity of f_G is shown in Table 2. When using the combination of E_{PA} and E_{rank}, e_{3D} is reduced by 31.72% for Traj.A and 16.88% for Traj.B. When using the combination of E_{PA}, E_{rank}, and E_{temp}, e_{3D} is reduced by 33.83% for Traj.A and 22.00% for Traj.B.

Table 3: e_{3D} on improved loss functions.

(a)

	Baseline			
	$\{E_{\text{temp}}\}$	$\{E^*_{\text{temp}}\}$	$\{E_{\text{PA}}, E_{\text{rank}}\}$	$\{E^*_{\text{PA}}, E_{\text{rank}}\}$
Traj.A	0.0506	0.0535	0.0457	0.0456
Traj.B	0.0670	0.0735	0.0456	0.0450

(b)

	RONN			
	$\{E_{\text{PA}}, E_{\text{rank}}\}$	$\{E^*_{\text{PA}}, E_{\text{rank}}\}$	$\{E_{\text{PA}}, E_{\text{rank}}, E_{\text{temp}}\}$	$\{E^*_{\text{PA}}, E_{\text{rank}}, E^*_{\text{temp}}\}$
Traj.A	0.0312	0.0355	0.0309	0.0321
Traj.B	0.0379	0.0399	0.0359	0.0396

Table 4: e_{3D} for the synthetic face sequence [2] (Traj.A and Traj.B).

(a)

	JM* [21]	GM* [11]	SMSR [10]	PPTA [22]	CMDR [23]
Traj.A	0.0280	0.0294	0.0304	0.0309	0.0324
Traj.B	0.0327	0.0309	0.0319	0.0572	0.0369

(b)

	VA [2]	DSTA [25]	EM-FEM† [26]	N-NRSfM [13]	RONN
Traj.A	0.0346	0.0374	0.0389	0.032	0.0309
Traj.B	0.0379	0.0428	0.0304	0.0389	0.0359

Note: *denotes methods that use Procrustes analysis for the shape alignment, whereas most methods use orthogonal Procrustes. †Represents the sequential method.

Table 5: e_{3D} for the expression dataset [18].

	EM-LDS [3]	CSF2 [27]	KSTA [28]	GMLI [18]	N-NRSfM [13]	RONN
Expr.	0.044	0.048	0.035	0.026	0.026	0.026

5.2.2. Effectiveness of Improved Loss Functions. To understand the effectiveness of the improved loss functions, we conducted experiments with the following original loss functions.

$$E_{\text{temp}} = \frac{1}{F-1}\sum_{f=1}^{F-1}\left\|S_{f+1} - S_f\right\|_{\epsilon}, \tag{16}$$

$$E_{\text{PA}} = \frac{1}{F}\sum_{f=1}^{F}\left\|S_f \bullet T - \bar{S}\right\|_{\epsilon}, \bar{S} = \frac{1}{F}\sum_{f=1}^{F} S_f \bullet T. \tag{17}$$

In Table 3, in the case of RONN without f_G, compared with E^*_{temp}, e_{3D} of the E_{temp} is reduced by 5.4% for Traj.A and 8.8% for Traj.B.

Because the combination of E_{PA} and E_{rank} must be used when using the f_G network, an experiment without f_G is added to show the performance of different improved functions. However, the improvement in E_{PA} did not affect the error before and after. However, with the addition of f_G, the combination of E_{PA} and E_{rank} reduces the error by 12.11% on Traj.A and 5.01% on Traj.B, which also shows the necessity of optimizing the camera matrix.

5.3. Comparison of Methods

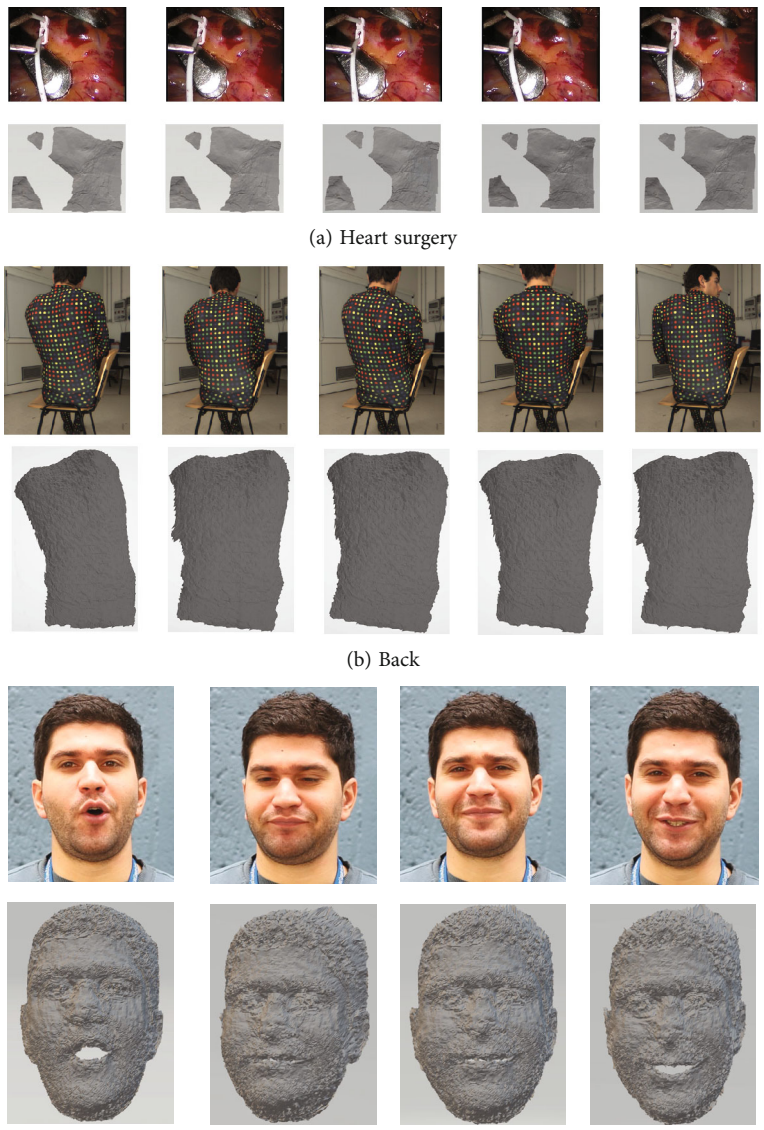
5.3.1. Synthetic Faces. e_{3D} for synthetic faces are listed in Table 4. Compared with jumping manifolds (JM) [21], Grassmannian manifold (GM) [11], SMSR [10], probabilistic point trajectory approach (PPTA) [22], consolidating monocular dynamic reconstruction (CMDR) [23, 24], variational approach (VA) [2], dense spatial-temporal approach (DSTA) [25], expectation-maximization finite element method (EM-FEM) [26], and N-NRSfM [13], the RONN achieves e_{3D} close to the best method on Traj.A and exhibits

Table 6: e_{3D} for the actor mocap dataset [19].

	SMSR [10]	CMDR [23]	RONN
Actor mocap	0.054	0.0257	0.0226

Table 7: e_{3D} for the sparse dataset.

	CSF2 [27]	PND [29]	BMM [7]	BMM-v2 [30]	RONN
Drink	0.0227	0.0037	0.0266	0.0119	0.0147
Pickup	0.1791	0.0372	0.1731	0.0198	0.028
Yoga	0.1179	0.0140	0.1150	0.0129	0.0237
Stretch	0.1136	0.0156	0.1034	0.0144	0.0234
Dance	0.1877	0.1454	0.1864	0.1060	0.1404
Shark	0.1117	0.0135	0.2311	0.0551	0.0332



(a) Heart surgery

(b) Back

(c) Real face

Figure 2: Visualizations of partial reconstruction result of real image sequences.

an average on e_{3D} on Traj.B. Compared with Traj.A, the reconstruction accuracy of Traj.B is poor, as reflected by many methods.

5.3.2. Expressions. e_{3D} for the expressions are presented in Table 5. Compared with the expectation-maximization linear dynamical system (EM-LDS) [3]; column space fitting, version 2 (CSF2) [27]; kernel shape trajectory approach (KSTA) [28]; global model with local interpretation (GMLI) [18]; and N-NRSfM [13], the RONN achieves $e_{3D} = 0.026$ on par with those of GMLI and N-NRSfM, which is currently the best method for this sequence. However, the number of iterations is significantly reduced (compared to 60,000 times in N-NRSfM).

5.3.3. Actor Mocap. e_{3D} for the expressions are listed in Table 6. Compared with CMDR [23, 24] and SMSR [10], the RONN achieves $e_{3D} = 0.0226$, which is better than SMSR and CMDR.

5.3.4. Sparse Reconstruction. Except for the linear layer in f_G, the RONN is composed of a convolutional network, and each feature point shares parameters so that the network can handle datasets with different numbers of feature points. When facing classic sparse data, comprising six standard sequences, namely, drink, pickup, yoga, stretch, dance, and shark, the RONN could also realize reconstruction. The number of frames (F) and number of points (P), i.e., the (F, P) set, for these datasets are (1102, 41), (357, 41), (307, 41), (370, 41), (264, 75), and (240, 91). As shown in Table 7, compared with the 3D reconstruction of dense data, in the sparse 3D reconstruction scene, the correlation between each point is relatively small owing to the few feature points. Compared with the classic sparse 3D reconstruction methods, the reconstruction results of the RONN are not as good as are expected. However, even if the reconstruction error of RONN is not the best, it is also not the worst.

5.4. Experiments with Real Data. We also reconstructed several real image sequences ,i.e., heart surgery [31], back [32], and real face [2] (see Figure 2). As for the real face, owing to the large amount of noise in matrix W, the final reconstruction result is not as smooth as expected. As for the back and heart surgery, the RONN achieved good visual reconstruction results.

6. Conclusion

This study proposes RONN and two improved loss functions. Our method can achieve reconstruction from 2D to 3D without supervision. One of the advantages of the RONN method is its scalability and consistent performance on datasets with different numbers of feature points.

As the first network to directly solve depth information to achieve reconstruction, RONN uses the depth reconstruction module f_z, which can achieve 3D structure reconstruction with only temporal smoothness loss. Procrustean regression is used to optimize the camera matrix and improve performance and use msr to weight the above loss function to further improve the network's ability to deal with complex deformation and experimentally demonstrate the improved loss function and the high performance of RONN.

Compared to the N-NRSfM approach, which is the first dense neural NRSfM, we do not need a mean shape and employ fewer training epochs. Compared to the classic sparse reconstruction method, the RONN shows better scalability. Transforming the input of the network from every frame to every point enables the network to better cope with the dense conditions.

Because of the direct use of W, the current limitation of the proposed method is its sensitivity to the 2D matrix W. As shown in Figure 2(c), the noise in W is directly shown in the 3D structure, causing the result to be unsmooth. Since the original rotation matrix is required, the reconstructed structure accuracy will be affected by the original rotation matrix. Moreover, the RONN cannot handle data loss.

The N-NRSfM provides a new perspective on dense NRSfM, which we further improved, achieving results. In future research, we will consider complex situations, such as denoising and data loss.

Data Availability

The Python code data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The author declares that there is no conflict of interest regarding the publication of this paper.

Acknowledgments

This research was supported by the Natural Science Foundation of Zhejiang Province (LZ20F020003, LY17F020034, LY17F020003, and LSZ19F010001) and the National Natural Science Foundation of China (61272311 and 61672466).

References

[1] C. Tomasi and T. Kanade, "Shape and motion from image streams under orthography: a factorization method," *International Journal of Computer Vision*, vol. 9, no. 2, pp. 137–154, 1992.

[2] R. Garg, A. Roussos, and L. Agapito, "Dense variational reconstruction of non-rigid surfaces from monocular video," in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1272–1279, Portland, OR, USA, 2013.

[3] L. Torresani, A. Hertzmann, and C. Bregler, "Nonrigid structure-from-motion: estimating shape and motion with hierarchical priors," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 5, pp. 878–892, 2008.

[4] I. Akhter, Y. Sheikh, S. Khan, and T. Kanade, "Nonrigid structure from motion in trajectory space," *Advances in Neural Information Processing Systems*, vol. 21, pp. 41–48, 2008.

[5] M. Paladini, A. Del Bue, M. Stosic, M. Dodig, J. Xavier, and L. Agapito, "Factorization for non-rigid and articulated structure using metric projections," in *2009 IEEE Conference on*

Computer Vision and Pattern Recognition, pp. 2898–2905, Miami, FL, USA, 2009.

[6] P. F. U. Gotardo and A. M. Martinez, "Computing smooth time trajectories for camera and deformable shape in structure from motion with occlusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 10, pp. 2051–2065, 2011.

[7] Y. Dai, H. Li, and M. He, "A simple prior-free method for non-rigid structure-from-motion factorization," *International Journal of Computer Vision*, vol. 107, no. 2, pp. 101–122, 2014.

[8] H. Zeng, Y. Dai, X. Yu, X. Wang, and Y. Yang, "PR-RRN: pairwise-regularized residual-recursive networks for non-rigid structure-from-motion," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5600–5609, Montreal, QC, Canada, 2021.

[9] C. Kong and S. Lucey, "Deep non-rigid structure from motion," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1558–1567, Seoul, Korea (South), 2019.

[10] M. D. Ansari, V. Golyanik, and D. Stricker, "Scalable dense monocular surface reconstruction," in *2017 International Conference on 3D Vision (3DV)*, pp. 78–87, Qingdao, China, 2017.

[11] S. Kumar, A. Cherian, Y. Dai, and H. Li, "Scalable dense non-rigid structure-from-motion: a Grassmannian perspective," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 254–263, Salt Lake City, UT, USA, 2018.

[12] C. Russell, J. Fayad, and L. Agapito, "Dense non-rigid structure from motion," in *2012 Second International Conference on 3D Imaging, Modeling, Processing, Visualization & Transmission*, pp. 509–516, Zurich, Switzerland, 2012.

[13] V. Sidhu, E. Tretschk, V. Golyanik, A. Agudo, and C. Theobalt, "Neural dense non-rigid structure from motion with latent space constraints," in *Lecture Notes in Computer Science Book Series*, vol. 12361, pp. 204–222, Springer, 2020.

[14] C. Bregler, A. Hertzmann, and H. Biermann, "Recovering non-rigid 3D shape from image streams," in *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No.PR00662)*, vol. 2, pp. 2690–2696, Hilton Head, SC, USA, 2000.

[15] S. Park, M. Lee, and N. Kwak, "Procrustean regression: a flexible alignment-based framework for nonrigid structure estimation," *IEEE Transactions on Image Processing*, vol. 27, no. 1, pp. 249–264, 2018.

[16] S. Park, M. Lee, and N. Kwak, "Procrustean regression networks: learning 3D structure of non-rigid objects from 2D annotations," in *European Conference on Computer Vision*, vol. 29, pp. 1–18, Glasgow, UK, 2020.

[17] D. Novotny, N. Ravi, B. Graham, N. Neverova, and A. Vedaldi, "C3DPO: canonical 3D pose networks for non-rigid structure from motion," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7688–7697, Seoul, Korea (South), 2019.

[18] A. Agudo and F. Moreno-Noguer, "Global model with local interpretation for dynamic shape reconstruction," in *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 264–272, Santa Rosa, CA, USA, 2017.

[19] L. Valgaerts, C. Wu, A. Bruhn, H. P. Seidel, and C. Theobalt, "Lightweight binocular facial performance capture under uncontrolled lighting," *ACM Transactions on Graphics*, vol. 31, no. 6, pp. 1–11, 2012.

[20] A. Paszke, S. Gross, F. Massa et al., "PyTorch: an imperative style, high-performance deep learning library," *Advances in Neural Information Processing Systems*, vol. 32, pp. 8026–8037, 2019.

[21] S. Kumar, "Jumping manifolds: geometry aware dense non-rigid structure from motion," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5346–5355, Long Beach, CA, USA, 2019.

[22] A. Agudo and F. Moreno-Noguer, "A scalable, efficient, and accurate solution to non-rigid structure from motion," *Computer Vision and Image Understanding*, vol. 167, pp. 121–133, 2018.

[23] V. Golyanik, A. Jonas, and D. Stricker, "Consolidating segmentwise non-rigid structure from motion," in *2019 16th International Conference on Machine Vision Applications (MVA)*, pp. 1–6, Tokyo, Japan, 2019.

[24] V. Golyanik, A. Jonas, D. Stricker, and C. Theobalt, "Intrinsic dynamic shape prior for fast, sequential and dense non-rigid structure from motion with detection of temporally-disjoint rigidity," 2019, https://arxiv.org/abs/1909.02468.

[25] Y. Dai, H. Deng, and M. He, "Dense non-rigid structure-from-motion made easy — a spatial-temporal smoothness based solution," in *2017 IEEE International Conference on Image Processing (ICIP)*, pp. 4532–4536, Beijing, China, 2017.

[26] A. Agudo, J. M. M. Montiel, L. Agapito, and B. Calvo, "Online dense non-rigid 3D shape and camera motion recovery," in *BMVC 2014- Proceedings of the British Machine Vision Conference 2014 (2014)*, Nottingham, United Kingdom, 2014.

[27] P. F. U. Gotardo and A. M. Martinez, "Non-rigid structure from motion with complementary rank-3 spaces," in *CVPR 2011*, pp. 3065–3072, Colorado Springs, CO, USA, 2011.

[28] P. F. U. Gotardo and A. M. Martinez, "Kernel non-rigid structure from motion," in *2011 International Conference on Computer Vision*, pp. 802–809, Barcelona, Spain, 2011.

[29] M. Lee, J. Cho, C. H. Choi, and S. Oh, "Procrustean normal distribution for non-rigid structure from motion," in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1280–1287, Portland, OR, USA, 2013.

[30] S. Kumar, "Non-rigid structure from motion: prior-free factorization method revisited," in *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 51–60, Snowmass, CO, USA, 2020.

[31] D. Stoyanov, "Stereoscopic scene flow for robotic assisted minimally invasive surgery," in *MICCAI'12 Proceedings of the 15th International Conference on Medical Image Computing and Computer-Assisted Intervention-Volume Part I*, vol. 15, pp. 479–486, Nice, France, 2012.

[32] C. Russell, J. Fayad, and L. Agapito, "Energy based multiple model fitting for non-rigid structure from motion," in *CVPR 2011*, pp. 3009–3016, Colorado Springs, CO, USA, 2011.