Neural Network for Dense Non-Rigid Structure from Motion

Subjects: Others

Contributor: Yaming Wang , Dawei Xu , Wenqing Huang , Xiaoping Ye , Mingfeng Jiang

Non-rigid Structure from Motion (NRSFM) is a significant research direction in computer vision that aims to estimate the 3D shape of non-rigid objects from videos.

dense Non-rigid Structure from Motion | self-attention | temporal smoothness

1. Introduction

Non-rigid Structure from Motion (NRSFM) is a significant research direction in computer vision that aims to estimate the 3D shape of non-rigid objects from videos. However, recovering 3D structure from 2D images poses a challenging inverse problem. A video can be considered as a collection of multi-frame images, and to achieve reconstruction, motion and deformation information between frames, along with prior assumptions, can be utilized.

Depending on the characteristics of the reconstructed object, the NRSFM problem can be divided into NRSFM with sparse feature points and NRSFM with dense feature points. Dense feature points are more numerous and accurate in rendering the surface of non-rigid objects compared to sparse feature points. Despite significant results achieved by existing sparse NRSFM methods, their performance in extending to the case of dense feature points remains unsatisfactory [1][2][3][4][5]. Recent progress has been made in dedicated algorithms for dense NRSFM problems [6][7][8][9].

The rise of deep learning has led to an increased focus on solving the challenges in NRSFM using neural networks, resulting in significant advancements. To address the challenge of dense feature points, Sidhu et al. [6] proposed the learnable neural network (N-NRSFM), which employs an automatic decoder model to assign a latent variable to each 3D shape and imposes constraints in the latent space to ensure similar 3D shapes have similar latent variables. This approach enhances robustness, scalability, and achieves lower 3D reconstruction errors in various scenarios. However, the method relies on the Tomas–Kanade decomposition [10] for solving the 3D mean shape and is sensitive to 2D trajectories with large errors. Inspired by Deng et al.'s study [11], where each frame does not exist independently but rather as part of a sequence, and the prior constraint should apply to the entire sequence rather than a single frame, the N-NRSFM method overlooks this point. Thus, it is important to revisit the NRSFM problem from this perspective to "translate" the input 2D image sequence into a 3D sequence structure.

2. Classical NRSFM Approaches

In the field of computer vision, considerable progress has been made over the past few decades in addressing the challenging problem of recovering 3D structures from 2D images. Several influential research directions have emerged, warranting attention in this context. One such direction is the shape space model, which is based on the low-rank assumption [12]. This model has been widely adopted to characterize the deformation of non-rigid objects, allowing for their recovery and analysis by constraining shape changes to a low-dimensional subspace. Another noteworthy approach is the trajectory space model [13], which describes the motion trajectory of non-rigid objects and infers three-dimensional structural information from it.

The probabilistic principal component analysis model (PPCA) [14] has also found important applications in computer vision. It employs probabilistic models to describe the data generation process and employs maximum likelihood estimation to infer the latent variables and model parameters. PPCA is advantageous in modeling non-rigid object deformation and motion.

The manifold hypothesis [15][16][17] constitutes another significant research direction, positing that object deformation and motion can be represented by low-dimensional manifolds. This idea has inspired subsequent methods such as the Grassmannian manifold method (GM) [9]. Moreover, the jump manifold method (JM) [18] serves as an extension of GM, considering the relationship between local surface deformation and point domains and utilizing high- and low-dimensional Grassmannian manifolds for modeling and reconstructing non-rigid object deformation.

In addition to the aforementioned approaches, the block matrix method (BMM) introduced by Dai [19] transforms the low-rank constraint into a semi-positive definite programming and kernel parametric minimization problem, providing a novel idea for non-rigid structure recovery. The SMSR method by Ansari [20] updates the input measurement matrix by applying trajectory smoothing constraints and employs the alternating direction multiplier method (ADMM) to optimize the objective function.

Furthermore, Lee et al. proposed the classical Expectation Maximization–Procrustean Normal Distribution (EM-PND) model [21] based on the generalized Procrustes analysis (GPA) [22]. This model does not require additional constraints or priors and can be applied to recover non-rigid objects at different time points. However, in practical scenarios, objects typically exhibit temporal variations, and enforcing smoothness becomes a crucial constraint for non-rigid 3D structure recovery. To address this, Lee et al. further proposed the Procrustean Markov Process (PMP) algorithm [23], which combines the PND assumption with a first-order Markov model to enable smooth recovery and modeling of non-rigid objects.

3. Neural-Network-Based Solutions for NRSFM

In recent years, the application of neural networks to solve the NRSFM problem has gained popularity among researchers [24][25][26]. Cha et al. [27] utilized a low-rank loss as the learning objective to constrain the shape output of their 2D–3D reconstruction networks. Novotny [26] proposed the C3DPO model, which employed low-rank factorization and consisted of two branches for viewpoint and shape prediction. The model achieved decoupling

and self-consistency of the branches by employing auxiliary neural networks to normalize the 3D shape of randomly rotated projections. On the other hand, Park et al. [28] achieved significant results by introducing a loss function that automatically determined the appropriate rotation for shape alignment, despite the relative simplicity of their underlying network structure.

Regarding the low-rank constraint, the shape basis (i.e., the rank parameter) plays a crucial role in reconstruction error. In previous NRSFM algorithms, the weights for low-rank, subspace, or compressed priors (e.g., rank or sparsity) often required tedious cross-validation for selection. However, Chen Kong et al. [24] proposed a deeply interpretable NRSFM network, known as Deep Neural Networks (DNNs), based on classical sparse dictionary learning and deep neural networks. This method eliminated the need for tedious cross-validation by simultaneously learning the prior weights and other parameters. They further extended their approach to handle occlusion and loss of feature points [29]. Wang et al. [30] advanced this approach by proposing the Deep-NRSFM++ model, which accounted for more realistic situations, including perspective projection cameras and critical occlusion. Ma et al. [31] built upon this approach by incorporating multi-view information and designing a simple yet effective loss function to ensure decomposition consistency. Subsequently, Zeng et al. [32] proposed a new residual recurrent network and introduced the Minimal Singular Value Ratio (MSR) as a metric for measuring shape rigidity between two frames. Based on this metric, they employed two novel pairwise loss functions to constrain the feature representation of 3D shapes, achieving advanced shape recovery accuracy in large-scale human motion and classified object reconstruction.

In contrast to classical NRSFM priors, some researchers have explored constraints provided by deep learning itself. Generative adversarial networks (GANs) have been utilized to predict lost depth information by enhancing 2D reprojection realism from different viewpoints [33][34][35][36]. However, due to the requirements of GAN learning, these methods are only applicable to large-scale datasets. Deng et al. [11] introduced the Sequence-to-Sequence (Sequence-to-Sequence) model to the NRSFM task, proposing the use of a multi-headed attention mechanism instead of a self-representation layer to impose priori subspace concatenation structures. Sidhu et al. [6] employed unsupervised neural networks for 3D reconstruction in dense NRSFM and achieved excellent results using automatic decoder deformation models and latent space constraints. Wang et al. [1] proposed the PAUL model, which argued that aligned 3D shapes could be compressed by depth under complete autoencoders. They further proposed the neural trajectory prior (NTP) for motion regularization [37]. The method can also be applied to other 3D computer vision tasks, including scene stream integration and dense NRSFM. Similar to N-NRSFM and PAUL, they introduced a bottleneck layer in the model to compress the generated trajectories into a low-dimensional space.

Despite the remarkable advancements in neural-network-based NRSfM in recent years, the emphasis has primarily been on the 3D reconstruction of sparse objects. The field of neural-network-based dense NRSfM is still in its nascent stage, leaving ample room for further exploration and development.

References

1. Wang, C.; Lucey, S. Paul: Procrustean autoencoder for unsupervised lifting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Online, 19–25 June 2021; pp. 434–443.
2. Russell, C.; Fayad, J.; Agapito, L. Dense non-rigid structure from motion. In Proceedings of the 2012 Second International Conference on 3D Imaging, Modeling, Processing, Visualization & Transmission, Li'ege, Belgium, 3–5 December 2012; pp. 509–516.
3. Golyanik, V.; Stricker, D. Dense batch non-rigid structure from motion in a second. In Proceedings of the 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), Santa Rosa, CA, USA, 24–31 March 2017; pp. 254–263.
4. Kumar, S.; Van Gool, L. Organic priors in non-rigid structure from motion. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; Springer: Berlin/Heidelberg, Germany, 2022; pp. 71–88.
5. Song, J.; Patel, M.; Jasour, A.; Ghaffari, M. A closed-form uncertainty propagation in non-rigid structure from motion. IEEE Robot. Autom. Lett. 2022, 7, 6479–6486.
6. Sidhu, V.; Tretschk, E.; Golyanik, V.; Agudo, A.; Theobalt, C. Neural dense non-rigid structure from motion with latent space constraints. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Part XVI 16. Springer: Berlin/Heidelberg, Germany, 2020; pp. 204–222.
7. Agudo, A.; Montiel, J.; Agapito, L.; Calvo, B. Online Dense Non-Rigid 3D Shape and Camera Motion Recovery. In Proceedings of the BMVC, Nottingham, UK, 1–5 September 2014.
8. Garg, R.; Roussos, A.; Agapito, L. Dense variational reconstruction of non-rigid surfaces from monocular video. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–27 June 2013; pp. 1272–1279.
9. Kumar, S.; Cherian, A.; Dai, Y.; Li, H. Scalable dense non-rigid structure-from-motion: A grassmannian perspective. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 254–263.
10. Tomasi, C.; Kanade, T. Shape and motion from image streams: A factorization method. Proc. Natl. Acad. Sci. USA 1993, 90, 9795–9802.
11. Deng, H.; Zhang, T.; Dai, Y.; Shi, J.; Zhong, Y.; Li, H. Deep Non-rigid Structure-from-Motion: A Sequence-to-Sequence Translation Perspective. arXiv 2022, arXiv:2204.04730.
12. Bregler, C.; Hertzmann, A.; Biermann, H. Recovering non-rigid 3D shape from image streams. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2000 (Cat. No. PR00662), Hilton Head, SC, USA, 13–15 June 2000; Volume 2, pp. 690–696.

13. Akhter, I.; Sheikh, Y.; Khan, S.; Kanade, T. Nonrigid structure from motion in trajectory space. In Proceedings of the Advances in Neural Information Processing Systems 21 (NIPS 2008), Vancouver, BC, Canada, 8–10 December 2008; Volume 21.
14. Torresani, L.; Hertzmann, A.; Bregler, C. Nonrigid structure-from-motion: Estimating shape and motion with hierarchical priors. IEEE Trans. Pattern Anal. Mach. Intell. 2008, 30, 878–892.
15. Rabaud, V.; Belongie, S. Re-thinking non-rigid structure from motion. In Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, USA, 23–28 June 2008; pp. 1–8.
16. Gotardo, P.F.; Martinez, A.M. Kernel non-rigid structure from motion. In Proceedings of the 2011 International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; pp. 802–809.
17. Hamsici, O.C.; Gotardo, P.F.; Martinez, A.M. Learning spatially-smooth mappings in non-rigid structure from motion. In Proceedings of the Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, 7–13 October 2012; Part IV 12. Springer: Berlin/Heidelberg, Germany, 2012; pp. 260–273.
18. Kumar, S. Jumping manifolds: Geometry aware dense non-rigid structure from motion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5346–5355.
19. Dai, Y.; Li, H.; He, M. A simple prior-free method for non-rigid structure-from-motion factorization. Int. J. Comput. Vis. 2014, 107, 101–122.
20. Ansari, M.D.; Golyanik, V.; Stricker, D. Scalable dense monocular surface reconstruction. In Proceedings of the 2017 International Conference on 3D Vision (3DV), Qingdao, China, 10–12 October 2017; pp. 78–87.
21. Lee, M.; Cho, J.; Choi, C.H.; Oh, S. Procrustean normal distribution for non-rigid structure from motion. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–27 June 2013; pp. 1280–1287.
22. Gower, J.C. Generalized procrustes analysis. Psychometrika 1975, 40, 33–51.
23. Lee, M.; Choi, C.H.; Oh, S. A procrustean Markov process for non-rigid structure recovery. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1550–1557.
24. Kong, C.; Lucey, S. Deep non-rigid structure from motion. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1558–1567.

25. Papyan, V.; Romano, Y.; Elad, M. Convolutional neural networks analyzed via convolutional sparse coding. J. Mach. Learn. Res. 2017, 18, 2887–2938.
26. Novotny, D.; Ravi, N.; Graham, B.; Neverova, N.; Vedaldi, A. C3dpo: Canonical 3d pose networks for non-rigid structure from motion. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 7688–7697.
27. Cha, G.; Lee, M.; Oh, S. Unsupervised 3d reconstruction networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 3849–3858.
28. Park, S.; Lee, M.; Kwak, N. Procrustean regression networks: Learning 3d structure of non-rigid objects from 2d annotations. In Proceedings of the European Conference on Computer Vision, Online, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 1–18.
29. Kong, C.; Lucey, S. Deep non-rigid structure from motion with missing data. IEEE Trans. Pattern Anal. Mach. Intell. 2020, 43, 4365–4377.
30. Wang, C.; Lin, C.H.; Lucey, S. Deep nrsfm++: Towards unsupervised 2d-3d lifting in the wild. In Proceedings of the 2020 International Conference on 3D Vision (3DV), Fukuoka, Japan, 25–28 November 2020; pp. 12–22.
31. Ma, Z.; Li, K.; Li, Y. Self-supervised method for 3D human pose estimation with consistent shape and viewpoint factorization. Appl. Intell. 2023, 53, 3864–3876.
32. Zeng, H.; Dai, Y.; Yu, X.; Wang, X.; Yang, Y. PR-RRN: Pairwise-regularized residual-recursive networks for non-rigid structure-from-motion. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 5600–5609.
33. Chen, C.H.; Tyagi, A.; Agrawal, A.; Drover, D.; Mv, R.; Stojanov, S.; Rehg, J.M. Unsupervised 3d pose estimation with geometric self-supervision. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 5714–5724.
34. Drover, D.; MV, R.; Chen, C.H.; Agrawal, A.; Tyagi, A.; Phuoc Huynh, C. Can 3d pose be learned from 2d projections alone? In Proceedings of the European Conference on Computer Vision (ECCV) Workshops, Munich, Germany, 8–14 September 2018.
35. Kudo, Y.; Ogaki, K.; Matsui, Y.; Odagiri, Y. Unsupervised adversarial learning of 3d human pose from 2d joint locations. arXiv 2018, arXiv:1803.08244.
36. Wandt, B.; Rosenhahn, B. Repnet: Weakly supervised training of an adversarial reprojection network for 3d human pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 7782–7791.

37. Wang, C.; Li, X.; Pontes, J.K.; Lucey, S. Neural prior for trajectory estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022; pp. 6532–6542.