3-D Motion Recovery via Low Rank Matrix Analysis

Meiyuan Wang [1], Kun Li [1*], Feng Wu [2], Yu-Kun Lai [3], Jingyu Yang [1]

[1] *Tianjin University, Tianjin, China*
[2]*University of Science and Technology of China, Hefei, China*
[3]*Cardiff University, Wales, UK*
[1*] Corresponding author: lik@tju.edu.cn

***Abstract*—Skeleton tracking is a useful and popular application of Kinect. However, it cannot provide accurate reconstructions for complex motions, especially in the presence of occlusion. This paper proposes a new 3-D motion recovery method based on low-rank matrix analysis to correct invalid or corrupted motions. We address this problem by representing a motion sequence as a matrix, and introducing a convex low-rank matrix recovery model, which fixes erroneous entries and finds the correct low-rank matrix by minimizing nuclear norm and ℓ_1-norm of constituent clean motion and error matrices. Experimental results show that our method recovers the corrupted skeleton joints, achieving accurate and smooth reconstructions even for complicated motions.**

***Index Terms*—Skeleton, Kinect, occlusion, motion recovery, low rank**

I. Introduction

3-D analysis of human motions has always been an active research topic in computer graphics and computer vision, involving a variety of research problems such as 3-D reconstruction [1], pose estimation [2], [3], [4], motion capture [5], [6], skeleton tracking [7], 3D model deformation [8], etc. Previous motion capture systems are challenging to implement, because they are expensive, difficult to maintain, and in need of abundant manual operations. Microsoft Kinect has shed a light on low-cost human motion capture. With its advantages of being convenient and inexpensive [9], Kinect gains its popularity in numerous works [10], and therefore has been widely used in the motion capture field [11]. However, skeletons captured by Kinect suffer from severe joint drifting and motion jitter, especially in the case of self-occlusion or object occlusion [12]. The accuracy of joint estimation is more satisfactory in controlled scenarios with simple non-occluded motions, such as standing upright, walking forward, which limits its wide applicability.

It is of great interest to researchers to reconstruct human motion via different methods. Many work focus on human motion estimation from RGB images. Menier *et al.* [13] estimate skeletal poses from foreground silhouettes. Li *et al.* [14] introduce sparse representation to estimate 3-D poses and camera motions. With the rapid development of neural networks and deep learning, some works bring deep learning models into motion estimation. Toshev *et al.* [3] propose a Deep Neural Networks (DNNs) to estimate human poses from RGB images by formulating it as a joint regression problem. Ouyang *et al.* [4] present a deep model to solve the pose estimation problem by utilizing information sources of mixture type, appearance score and deformation. Since the advent of depth cameras such as Kinect, pose estimation can be better addressed through depth images. Shotton *et al.* [15] introduce *body part classification* (BPC) and *offset joint regression* (OJR) algorithms to estimate human poses with robustness and efficiency from a single depth image. Wei *et al.* [5] develop an automatic motion capture system by integrating depth data, full-body geometry, etc. together using a single depth camera. However, these methods have limitations in recovering skeletons with occlusions. Saito *et al.* [7] solve this problem by finding subspace of valid motion, projecting corrupted skeletons into the motion manifold and finally rebuilding valid motion through inverse projection. However, this method needs a time-consuming training procedure.

3-D motion recovery from corrupted skeletons is challenging. The key is to impose proper priors to make the problem well-posed. As skeleton sequence has high temporal correlation, the motion should lie in a low-dimensional subspace. In this paper, we propose a new 3-D motion recovery method based on low rank matrix analysis. The skeleton sequence is rearranged into a (corrupted) matrix that contains errors and noise. Then, the clean motion matrix is recovered from the corrupted matrix by minimizing the nuclear norm of the clean matrix and the ℓ_1-norm of the error matrix, exploiting the characteristics of both matrices. Through this method, the corrupted matrix that has high percentages of noise and errors can be corrected. Experimental results show that our method successfully corrects the corrupted motion captured by Kinect v2.0, especially for complex motions.

The remainder of this paper is organized as follows. Section II gives the proposed 3-D motion recovery method. Validation experiments are presented in Section III, and the paper is concluded in Section IV.

IEEE VCIP'16, Nov. 27 - Nov. 30, 2016, Chengdu, China

II. The Proposed Method

A. Matrix Recovery Model

The human skeleton is represented by a collection of joints, which are easily influenced by noises and have drifting problems. Given the skeleton sequence captured by Kinect v2.0, an observation matrix $D \in \mathbb{R}^{m\times n}$ is formed by stacking the 3-D positions of all the joints together, where m is $3\times$ the number of frames of the input skeleton sequence and n is the number of joints (21 in our case, ignoring the finger joints). Let $A \in \mathbb{R}^{m\times n}$ be the recovered clean matrix, and let $E \in \mathbb{R}^{m\times n}$ be the error matrix. We have

$$D = A + E. \tag{1}$$

Classical works in the literature suggest that human motions lie in a subspace therefore can be effectively represented in lower dimensions [16]. This inspires us to focus on the time coherence of skeleton data in a motion sequence–the rank of the clean matrix A should be low. To verify this, as shown in Fig.1, the rank of our input skeleton matrix for *Dancing* sequence is 5 using singular value decomposition, validating our low-rank hypothesis, which forms the basis of our matrix recovery model.

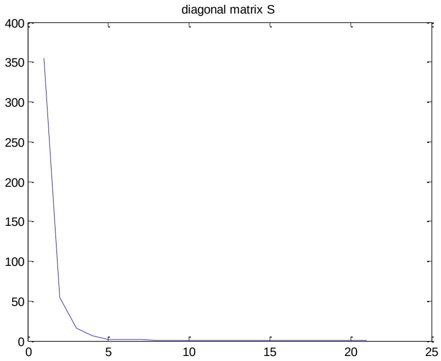

Fig. 1. Singular value decomposition of input skeleton matrix for sequence *dancing*.

Given that the latent matrix A is low-rank and the error matrix E is sparse, the problem can be formulated as

$$\begin{aligned} &\min rank(A) + \gamma \left\|E\right\|_0 \\ &\text{s.t.} \quad D = A + E, \end{aligned} \tag{2}$$

where $rank(A)$ is the rank of matrix A, $\left\|E\right\|_0$ is the ℓ_0 norm of matrix E, i.e. the number of non-zero entries in the matrix and $\gamma > 0$ is a weighting parameter that balance the importance of low-rank matrix A and sparsity of matrix E.

Since the problem in Eq.(2) is NP-hard, we use the nuclear norm of A as an approximate substitute for $rank(A)$, and the ℓ_1 norm as an approximate substitute for ℓ_0 norm of matrix E [17]. Thus we obtain a semidefinite programing problem:

$$\begin{aligned} &\min \left\|A\right\|_* + \lambda \left\|E\right\|_1 \\ &\text{s.t.} \quad D = A + E, \end{aligned} \tag{3}$$

where $\lambda > 0$ is a weighting parameter, $\left\|E\right\|_1 = \sum_{ij} |E_{ij}|$, $\left\|A\right\|_* = tr((AA^{\mathrm{T}})^{1/2}) = \sum_i \sigma_i$, where σ_i's are singular values of matrix A. Theoretical considerations in [18] suggest that λ must be of the form $\frac{C}{\sqrt{\max(m,n)}}$, where C is a constant, typically set to unity. Additionally, we evaluate the influence of the weighting parameter λ in our experiments, which achieves the balance between accuracy and smoothness of the recovered skeletons. The value of λ is chosen small enough to threshold away the noise (by keeping the variance low to obtain high stability), and large enough not to overshrink the original matrix (by keeping the bias low to ensure flexible motion).

B. Augmented Lagrangian Algorithm

Algorithm 1 :ALM Algorithm

```
**Input:** observed skeleton matrix $D \in \mathbb{R}^{m\times n}$
**Initialize:** $E_0 = 0$, $Y_0 = 0$, $\mu > 0$ $maxIter = 1000$
**while** not converged **do**
    $A_{\mathrm{k+1}} = M_{1/\mu}(D - E_{\mathrm{k}} + \frac{1}{\mu}Y_{\mathrm{k}})$;
    $E_{\mathrm{k+1}} = S_{\lambda/\mu}(D - A_{\mathrm{k+1}} + \frac{1}{\mu}Y_{\mathrm{k}})$;
    $Y_{\mathrm{k+1}} = Y_k + \mu(D - A_{\mathrm{k+1}} - E_{\mathrm{k+1}})$;
    $\mu_{k+1} = \rho\mu_k, \rho > 1$;
**end while**
**Output:** $A$, $E$
```

There are many algorithms to solve this minimization problem, e.g.,singular value thresholding (SVT) [19], augmented lagrangian method (ALM) [20], and accelerated proximate gradient algorithm (APG) [21]. In this paper, we choose augmented lagrangian method (ALM) in an iterative framework, due to its high efficiency and accuracy. The augmented Lagrangian of Eq.(3) is

$$\begin{aligned} L(A, E, Y, \mu) = &\left\|A\right\|_* + \lambda \left\|E\right\|_1 + \langle Y, D - A - E\rangle \\ &+ \frac{\mu}{2}\left\|D - A - E\right\|_F^2, \end{aligned} \tag{4}$$

where $\left\|M\right\|_F$ represents the Frobenious norm of a matrix M, Y is the Lagrangian multiplier, and $\langle\cdot,\cdot\rangle$ denotes the inner product of two matrices considered as long vectors. Then, we solve two optimizations: $\min_A L(A, E, Y)$ and $\min_E L(A, E, Y)$ respectively.

$$\arg\min_A L(A, E, Y) = S_{\lambda/\mu}(D - A + \frac{1}{\mu}Y) \tag{5}$$

$$\arg\min_E L(A, E, Y) = M_{1/\mu}(D - E + \frac{1}{\mu}Y), \tag{6}$$

where $S_\delta(x) = \mathrm{sgn}(x)\max(|x| - \delta, 0)$ denotes a shrinkage operator. Similarly, $M_\delta(X) = US_\delta(\Lambda)V$ denotes a singular value thresholding operator.

To address these minimization problems, we use a classical approach: First we minimize function $L(A, E, Y)$ with A fixed, then we similarly minimize $L(A, E, Y)$ with E fixed, and finally we update the Lagrange multiplier matrix Y. The ALM algorithm is summarized in Algorithm 1.

III. Experimental Results

We use real captured Kinect skeletons of 21-joints as our input $D \in \mathbb{R}^{m \times n}$, which suffers from severe joint drifting and motion jitter. Four motion sequences are tested: *marking time* (463 frames), *crossing and bending* (1245 frames), *facing aside* (421 frames) and *dancing* (1432 frames). Detailed demo video is presented in the supplementary material.

First, we test the proposed method on a simple non-occluded motion sequence: *marking time* (the first sequence in our demo video). As shown in Fig.2 and the video, the input skeleton has joint drifting in some frames where the person lifts a leg or waves arms. The recovered skeleton has detected these disturbances and rectified the noisy skeleton as shown in Fig.2 (c). In the sequence of *crossing and bending* (the second sequence in our demo video), human motions are more complicated and have higher percentages of self-occlusion. As shown in Fig.3 , the Kinect skeleton is obviously corrupted, while our method reconstructs a valid skeleton structure.

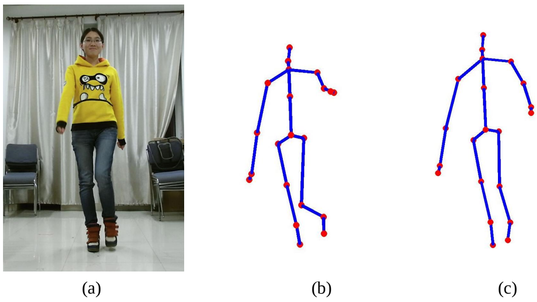

Fig. 2. Estimated skeleton of frame 171 in *marking-time* sequence. (a) captured color image, (b) Kinect skeleton, (c) recovered skeleton by our method.

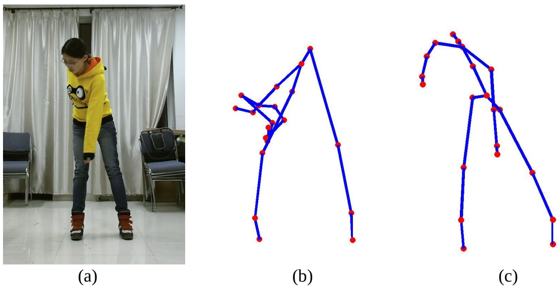

Fig. 3. Estimated skeleton of frame 1178 in *crossing and bending* sequence. (a) captured color image, (b) Kinect skeleton, (c) recovered skeleton by our method.

Then, we evaluate our method on more complex motion sequences containing various human movements and high percentages of self-occlusion: *facing aside* and *dancing* (the third and fourth sequence in the demo video). Note that in Fig.4, in order to present skeletons more legibly, we additionally render the skeleton in a side view, since the outstretched arms tend to be elusive in a single frame in the front view. From the side view, we can more obviously observe that the occluded arm is corrected estimated. The captured Kinect skeleton has been corrupted largely as shown in Fig.4(b)(c)(g)(h)(l)(m) and Fig.5(d)-(f). This shows that our method gets reasonable skeletons for these complex motions.

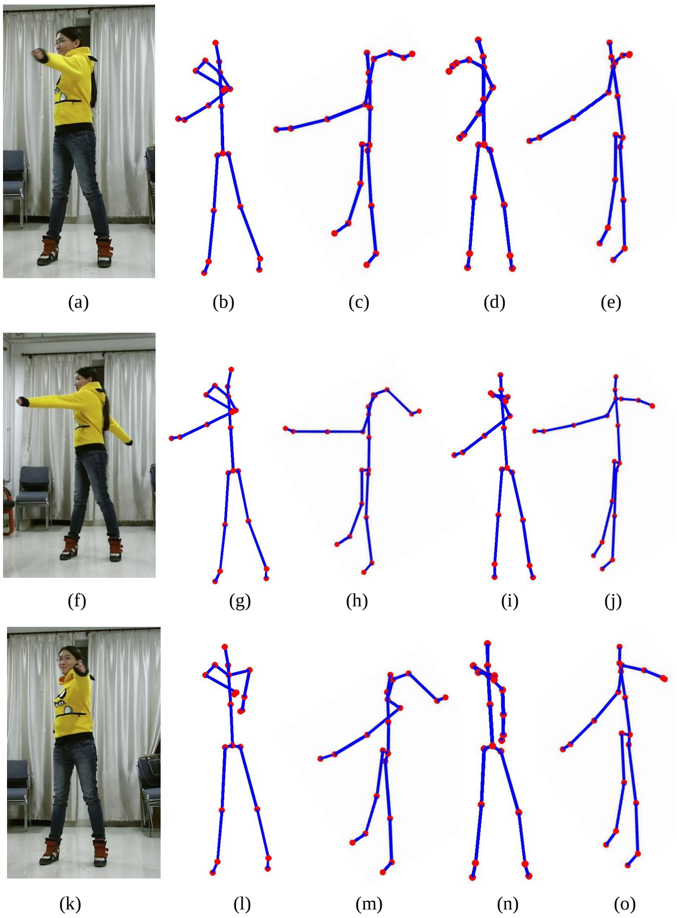

Fig. 4. Estimated skeleton of frame 91, 104 and 148 in *facing aside* sequence. Frame 104: (a)(f)(k) captured color image, (b)(g)(l) Kinect skeleton in the front view, (c)(h)(m) Kinect skeleton in the side view, (d)(i)(n) recovered skeleton by our method in the front view, (e)(j)(o) recovered skeleton by our method in the side view.

We test the running time of every input sequence on a desktop with an Intel Core i5-4690K CPU and 8GB RAM. The results are reported in Table I. Our method is fast and has the potential to achieve real-time online skeleton recovery.

TABLE I
RUNNING TIME

Skeleton Sequence	Number of Frames	Running Time(s)
Facing Aside	241	0.6190
Marking Time	463	0.8001
Crossing and Bending	1245	1.1470
Dancing	1423	1.3740

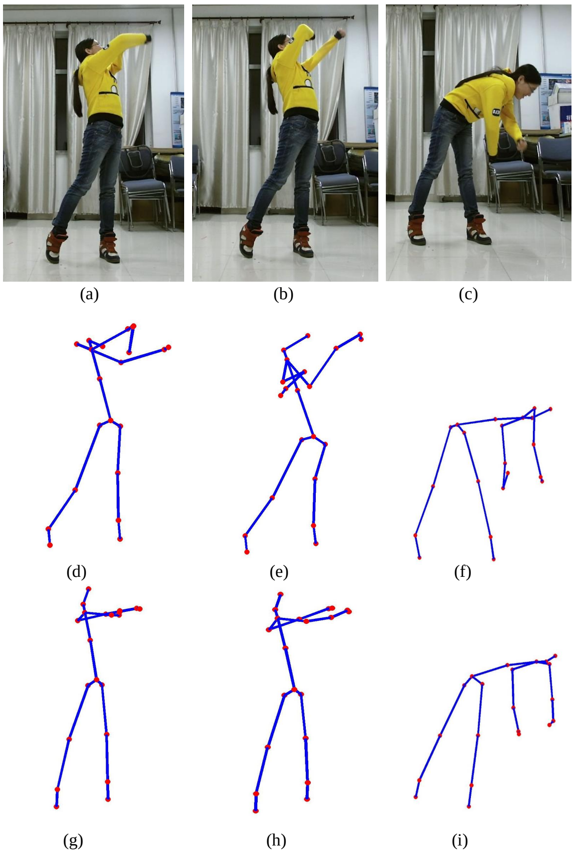

Fig. 5. Estimated skeleton of frame 141, 166 and 244 in *dancnig* sequence. (a) - (c) captured color image, (d) - (f) Kinect skeleton, (g) - (i) recovered skeleton by our method.

IV. Conclusion

This paper proposes a novel approach to motion reconstruction and skeleton recovery via low-rank matrix analysis. We exploit the time coherence in a skeleton sequence which is formulated as a low-rank configuration in a mathematical model. Matrix recovery method shows its efficiency in addressing the skeleton recovery problem. An ALM algorithm is devised to solve the optimization problem. We evaluate our method on real skeletons captured by Kinect v2.0, which contain severe errors. Experimental results show that our method accurately recovers high quality skeletons from the invalid corrupted motion data in high efficiency.

Similar to the algorithm of Kinect, our method does not guarantee the invariance of bone lengths. One can obtain the motion capture result with changeless bone length using an IK (inverse kinematics) algorithm based on our recovered skeletons [22]. In the future work, we will perform the low-rank matrix recovery on a graph to address this problem.

V. Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (Grant 61571322, 61302059 and 61372084).

References

[1] D. Alexiadis, D. Zarpalas, and P. Daras, "Fast and smooth 3d reconstruction using multiple rgb-depth sensors," in *Visual Communications and Image Processing(VCIP),2014 IEEE*, 2014, pp. 173–176.

[2] C. P. Chen, Y. T. Chen, P. H. Lee, Y. P. Tsai, and S. Lei, "Real-time hand tracking on depth images," in *Visual Communications and Image Processing (VCIP), 2011 IEEE*, 2011, pp. 1–4.

[3] A. Toshev and C. Szegedy, "Deeppose: Human pose estimation via deep neural networks," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 1653–1660.

[4] W. Ouyang, X. Chu, and X. Wang, "Multi-source deep learning for human pose estimation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2337–2344.

[5] X. Wei, P. Zhang, and J. Chai, "Accurate realtime full-body motion capture using a single depth camera," *ACM Transactions on Graphics*, vol. 31, no. 6, pp. 439–445, 2012.

[6] C. Wang, Z. Liu, and S. C. Chan, "Superpixel-based hand gesture recognition with kinect depth camera," *IEEE Transactions on Multimedia*, vol. 17, no. 1, pp. 29–39, 2015.

[7] J. Saito, D. Holden, and T. Komura, "Learning motion manifolds with convolutional autoencoders," in *SIGGRAPH Asia Technical Briefs*, 2015.

[8] M. Dou, J. Taylor, H. Fuchs, A. Fitzgibbon, and S. Izadi, "3d scanning deformable objects with a single rgbd sensor," in *Computer Vision and Pattern Recognition*, 2015, pp. 493–501.

[9] M. A. Livingston, J. Sebastian, Z. Ai, and J. W. Decker, "Performance measurements for the microsoft kinect skeleton," in *IEEE Virtual Reality*, 2012, pp. 119–120.

[10] B. Wang, Z. Chen, and J. Chen, "Gesture recognition by using kinect skeleton tracking system," in *International Conference on Intelligent Human-Machine Systems and Cybernetics*, 2013, pp. 1119–1119.

[11] D. S. Alexiadis and P. Daras, "Quaternionic signal processing techniques for automatic evaluation of dance performances from mocap data," *IEEE Transactions on Multimedia*, vol. 16, no. 5, pp. 1391–1406, 2014.

[12] S. Obdrzalek, G. Kurillo, F. Ofli, R. Bajcsy, E. Seto, H. Jimison, and M. Pavel, "Accuracy and robustness of kinect pose estimation in the context of coaching of elderly population." in *Proc. International Conference of the IEEE Engineering in Medicine and Biology Society*, 2012, pp. 1188–93.

[13] C. Menier, E. Boyer, and B. Raffin, "3d skeleton-based body pose recovery," in *Proc. International Symposium on 3D Data Processing Visualization and Transmission*, 2006, pp. 389–396.

[14] K. Li, J. Yang, and J. Jiang, "Nonrigid structure from motion via sparse representation." *IEEE Transactions on Cybernetics*, vol. 45, no. 8, pp. 1401–1413, 2015.

[15] J. Shotton, R. Girshick, A. Fitzgibbon, T. Sharp, M. Cook, M. Finocchio, R. Moore, P. Kohli, A. Criminisi, and A. Kipman, "Efficient human pose estimation from single depth images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 12, pp. 2821–2840, 2013.

[16] J. Chai and J. K. Hodgins, "Performance animation from low-dimensional control signals," *ACM Transactions on Graphics*, vol. 24, no. 3, pp. 686–696, 2005.

[17] E. J. Cands, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis," *Journal of the ACM*, vol. 1, no. 1, pp. 1–73, 2000.

[18] J. Wright, A. Ganesh, S. Rao, and Y. Ma, "Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization. submitted to the," *Journal of the ACM*, 2009.

[19] J. F. Cai, Cand, E. J. S, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *Siam Journal on Optimization*, vol. 20, no. 4, pp. 1956–1982, 2010.

[20] Z. Lin, M. Chen, and Y. Ma, "The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices," *Eprint Arxiv*, vol. 9, 2010.

[21] K. C. Toh and S. Yun, "An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares problems," *Pacific Journal of Optimization*, vol. 6, no. 615-640, p. 15, 2010.

[22] P. Baerlocher and R. Boulic, "An inverse kinematics architecture enforcing an arbitrary number of strict priority levels," *Visual Computer*, vol. 20, no. 6, pp. 402–417, 2004.