PUSHING THE LIMIT OF NON-RIGID STRUCTURE-FROM-MOTION BY SHAPE CLUSTERING

Huizhong Deng *Yuchao Dai*

Research School of Engineering, College of Engineering and Computer Science
Australian National University, Australia

ABSTRACT

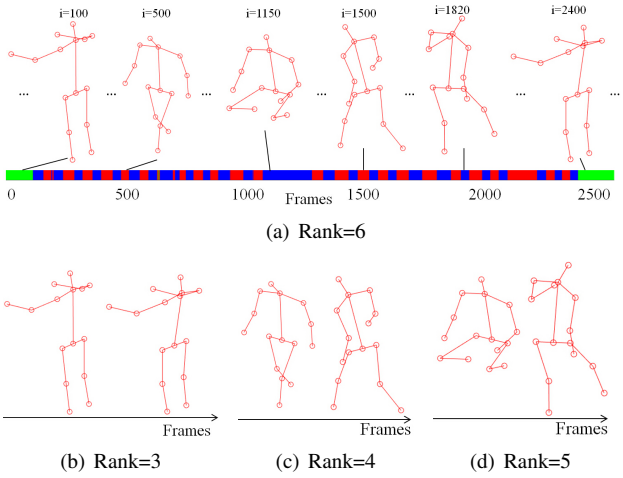
Recovering both camera motions and non-rigid 3D shapes from 2D feature tracks is a challenging problem in computer vision. Long-term, complex non-rigid shape variations in real world videos further increase the difficulty for Non-rigid structure-from-motion (NRSfM). Furthermore, there does not exist a criterion to characterize the possibility in recovering the non-rigid shapes and camera motions (*i.e.*, how easy or how difficult the problem could be). In this paper, we first present an analysis to the "*reconstructability*" measure for NRSfM, where we show that 3D shape complexity and camera motion complexity can be used to index the reconstructability. We propose an iterative shape clustering based method to NRSfM, which alternates between 3D shape clustering and 3D shape reconstruction. Thus, the global reconstructability has been improved and better reconstruction can be achieved. Experimental results on long-term, complex non-rigid motion sequences show that our method outperforms the current state-of-the-art methods by a margin.

Index Terms— Non-rigid structure-from-motion, shape clustering, reconstructability, 3D reconstruction.

1. INTRODUCTION

Non-rigid Structure-from-Motion (NRSfM) aims at estimating both camera motions and 3D dynamic shapes from 2D image measurements, which is central to dynamic scene understanding, motion capture and activity recognition. Despite the recent progress [1] [2] [3], NRSfM still lags far behind its rigid counterpart, which is well-developed and can be reliably solved. This is mainly due to the difficulty in modeling real world non-rigid variation [4] [5] [6] [7] [8] and difficulty in the corresponding minimization problem [9].

NRSfM can be roughly categorized into two classes: sparse methods and dense methods. Under sparse 3D reconstruction, a global model is generally used to regularize the otherwise highly under-determined problem. Linear combination model [4] has been widely used to capture the low-dimension structure of the 3D shapes. Following shape space approaches [5] [11] [12] [3] represent the non-rigid variation as a combination of basis shape variations. Meanwhile, the trajectory space based methods [1] aims at reconstructing each feature track's 3D trajectory using a pre-defined trajectory bases. The shape-trajectory approach [2] combines the two models and formulates the problems as revealing the trajectory of shape basis coefficients. Besides the linear combination model, Lee *et al*. [10] proposed a Procrustean Normal Distribution (PND) model, where the 3D shapes are aligned and fit into a normal distribution. In sparse reconstruction, the feature points are geometrically apart from each other, thus no spatial regularization can be enforced.



(a) Rank=6 (b) Rank=3 (c) Rank=4 (d) Rank=5

Fig. 1: Illustration of our method on the UMPM "Free" sequence. The top row shows the result by PND [10] using a global model while the bottom row shows the result by our method, where the whole sequence is clustered into 3 sub-sequences. The dimensionality of the subspace (rank) is shown alongside the corresponding results. Different colors are used to indicate the clustering result of the frames. Through iterative shape clustering and 3D shape reconstruction, we achieved an overall 3D reconstruction error as 0.2588 while the state-of-the-art method PND achieved 0.3887.

This work is in part supported by the Australian Research Council Grant (DE140100180) and National Natural Science Foundation of China (61420106007).

By contrast, dense NRSfM methods such as [13] [14] [15] aims at achieving 3D reconstruction for each pixel in the video sequence, where spatial constraint has been widely used to regularize the problem. In this paper, we will focus on sparse 3D reconstruction while the general principle proposed here can be applied to dense case with light modification.

Despite its success in reconstructing simple non-rigid deformations, NRSfM is still far from real world applications. Real world non-rigid reconstruction generally requires the ability to handle long, complex non-rigid shape variations. The long and complex motion not only increases the computational complexity but also adds difficulty in modeling various kinds of different motions (*e.g*., in motion capture, a human can sit, stand, walk, bend and dance inside a video). A complex non-rigid variation is hard to be correctly represented by a single subspace model or probabilistic model. Zhu *et al*. [16] represented complex motion as lying in a union of subspaces rather than sum of subspaces. However, the solution involves a complex non-convex optimization.

Furthermore, there does not exist a criterion to characterize the possibility in recovering non-rigid shape and camera motion given input video (*i.e*., how easy or how difficult the problem could be). This is a typical "chicken and egg" problem. At one hand, if the camera motion and 3D shape have been recovered, it is easy to define such a metric. At the other hand, if such metric is available before reconstruction, we can utilize the metric to design proper reconstruction methods. Park *et al*. [17] [18] looked into the theoretical aspect of 3D trajectory reconstruction and proposed a criterion called "reconstructibility", which measures the possibility and accuracy of reconstructing a 3D point from its 2D trajectory. This "reconstructability" is only valid in the trajectory reconstruction problem, where the camera motion is available.

To pave the way for NRSfM in real world applications and deal with long and complex motion in NRSfM, in this paper, we extend the concept of "reconstructibility" from trajectory reconstruction to the general NRSfM problem. Under our formulation, reconstructibility is defined on the recovered 3D shapes. To utilize this property and improve the reconstructibility in NRSfM, we propose an iterative method, which alternatively clusters a long, complex sequence into subsequences by using 3D shape similarity and reconstructs each subseqeunce. In this way, each subsequence has a much lower shape complexity and the global reconstructibility has been improved. Extensive experimental results on long, complex motion sequences show that our method outperforms the current state-of-the-art NRSfM methods by a margin, thus pushing the limit of NRSfM.

2. RECONSTRUCTABILITY FOR NRSFM

We consider a monocular camera observing a non-rigid shape. We assume an affine camera model and eliminate the translation as in [4]. The image measurement $\mathbf{w}_{ij} = [u_{ij}, v_{ij}]^T$ and 3D point $\mathtt{S}_{ij}$ on the non-rigid shape are related by the camera motion $\mathtt{R}_i$ as: $\mathbf{w}_{ij} = \mathtt{R}_i\mathtt{S}_{ij}$, where $\mathtt{R}_i \in \mathbb{R}^{2\times 3}$ denotes the first two rows of the i-th camera rotation. Using this representation, and stack all the F frames of measurements and all the P points in matrix form, we reach:

$$\mathtt{W} = \mathtt{RS}, \tag{1}$$

where $\mathtt{R} = \text{blkdiag}(\mathtt{R}_1, \cdots, \mathtt{R}_F) \in \mathbb{R}^{2F\times 3F}$ expresses the camera motion matrix. Factorization based NRSfM aims at factorizing the 2D *measurement matrix* $\mathtt{W} \in \mathbb{R}^{2F\times P}$ as the product of *camera motion* (projection) matrix $\mathtt{R}$ and a 3D non-rigid shape matrix $\mathtt{S} \in \mathbb{R}^{3F\times P}$, such that $\mathtt{W} = \mathtt{RS}$.

To characterize the possibility in recovering the non-rigid shape and camera motion given input feature tracks, we propose to analyze the camera motion and 3D shapes. In the following paragraphs, we first review the reconstructability proposed for trajectory reconstruction. Then we extend the concept to general NRSfM and propose our reconstructability evaluation based on 3D shape similarity.

Reconstructability in trajectory reconstruction: Given camera motions, trajectory reconstruction [17] aims at estimating a 3D point trajectory from 2D feature tracks. Park *et al*. proposed a measure on the possibility of reconstruction, namely "reconsructability", by analyzing the correlation between camera trajectory and moving point trajectory. Specifically, the reconstructability η, characterizing the relationship between camera motion, point motion and the trajectory basis, is defined as:

$$\eta(\Theta) = \frac{\|\Theta^{\perp}\beta_{\mathtt{C}}^{\perp}\|}{\|\Theta^{\perp}\beta_{\mathtt{X}}^{\perp}\|} \simeq \frac{\text{How poorly } \Theta \text{ describes } \mathtt{C}}{\text{How poorly } \Theta \text{ describes } \mathtt{X}}, \tag{2}$$

where Θ is the pre-defined trajectory bases, $\mathtt{C}$ is the camera trajectory, while $\mathtt{X}$ is the 3D point trajectory. In other words, a complex $\mathtt{C}$ and a simple $\mathtt{X}$ result in a high value of η. Note in trajectory reconstruction, the camera motion is available.

Reconstructability for NRSfM: To extend the concept of "reconsructability" from trajectory reconstruction to general NRSfM, we need to measure the complexity in both camera motion and 3D shape variation.

Shape complexity: Given a primitive non-rigid shape $\mathtt{S}$, its complexity (reconstructability) can be well characterized by the rank, *i.e*., $\eta_S = \text{rank}(\mathtt{S})$.

Motion complexity: Under our formulation, camera motion only consists of the rotation component. As camera rotation resides in a manifold as $\mathtt{R}_i \in \text{SO}(3)$, to define the complexity of camera motion, we need to characterize the distance on the manifold. To ease the computation, we use the chordal distance to evaluate the difference between rotations as: $d_{ij} = \|\mathtt{R}_i - \mathtt{R}_j\|_F$. In this way the global motion complexity could be defined as: $\eta_R = \sum_{i,j} d_{ij}^2$.

By putting the shape complexity and the motion complexity together, we obtain the "reconsructability" for general NRSfM as :

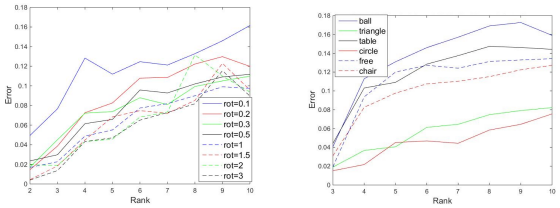
$$\eta(\mathbf{R}, \mathbf{S}) = \frac{\eta_R(\mathbf{R})}{\eta_S(\mathbf{S})}. \tag{3}$$

According to the definition, a larger motion complexity and a smaller shape complexity will generally result in a higher reconsructability, which is consistent with existing work in NRSfM [16] [17].

Numerical examples: To evaluate the correctness of our reconstructability for NRSfM, we set up a series of experiments on the UMPM sequences [19] to analyze the relationship between reconstructability and motion/shape complexity.

To obtain sequences with varying shape complexity, we project ground truth UMPM 3D shapes into low dimensional subspace with varying dimension K. Then we perform a Procrustean alignment to the sequence such that all frames are aligned to the first frame, thus eliminating the rigid component in non-rigid shape deformation. We applied two different kinds of camera motions in our experiments: 1). varying rotation speed (from 0.1 degree per frame to 3 degrees per frame, thus varying camera motion complexity) with a random direction following a Gaussian distribution at each frame; 2). completely random camera rotations at each frame, for which the camera motion complexity has been maximized.

Experimental results are illustrated in Fig. 2, where the two figures correspond to the two camera motion configurations. In the varying camera rotation speed case as shown in Figure 2(a), 3D reconstruction error generally increases with the increase of shape complexity (rank) and decreases with the increase of rotation speed. In the completely random camera motions case as shown in Figure 2(b), as shape complexity increases, the 3D reconstruction error increases correspondingly. All these experiments demonstrate that our new reconstructability clearly captures the essence in achieving better 3D reconstruction through evaluating shape complexity and motion complexity.



(a) (b)

Fig. 2: Numerical experiments analyzing the relationship between shape complexity, motion complexity and 3D reconstruction performance. (a) 3D reconstruction error on the "Triangle" sequence with varying shape complexity under different camera rotation speeds. (b) 3D reconstruction error for different UMPM sequences with varying shape complexity under complete random camera motions.

3. NRSFM BY 3D SHAPE CLUSTERING

According to the definition of reconstructability in Eq.-(3), a higher shape complexity will result in a lower reconstructability. For long, complex non-rigid variation sequences, shape complexity tends to increase with sequence length. Meanwhile, non-rigid variation in real world cases generally consists of local shape variations with low complexity. Therefore, we can increase the global reconstructability by clustering a long sequence into subsequences. In this section, we present an iterative shape clustering based NRSfM method.

Algorithm 1 Shape clustering based NRSfM.

Require: 2D feature tracks $\mathtt{W}$ (Complete or incomplete)
Initialize: 3D shape $\mathtt{S}^{(0)}$ from a factorization method.
while Not converged **do**
1). Compute similarity matrix $\mathtt{M}^{(it)}$ from 3D shapes $\mathtt{S}^{(it)}$.
2). Clustering: apply spectral clustering method to the similarity matrix $\mathtt{M}^{(it)}$, getting K subsequences.
3). Reconstruction: Each subsequence is reconstructed separately, and they are reassembled to $\mathtt{S}^{(it+1)}$.
end while
Ensure: Non-rigid shape $\mathtt{S}$, camera motion $\mathtt{R}$.

3.1. 3D shape similarity

To cluster a long sequence into subsequences, an initial 3D shape is required, as clustering on the 2D image measurements is unable to indicate the real shape similarity of the sequence [16]. The initialization is implemented by using PND [10]. The initial 3D reconstruction could depart from the ground truth. As explained later, our method does not need a very accurate initialization.

Given and initial 3D reconstruction $\mathtt{S}^{(0)}$, we can define a shape similarity matrix by comparing all the shapes against each other. The similarity matrix $\mathtt{M}$ is computed as $M(i,j) = M(j,i) = \exp(-\frac{||\mathbf{S}_i - \mathbf{S}_j||_F}{\sigma})$, where $||\mathbf{S}_i - \mathbf{S}_j||_F, i, j \in [1, 2, ..., P]$ denotes the Euclidean distance between two shapes, and σ is a scaling parameter.

3.2. Shape clustering

Once the similarity matrix $\mathtt{M}$ is obtained, spectral clustering [20] is used to cluster the whole sequence into subsequences. The benefit of spectral clustering is that it is designed to handle a similarity matrix directly, and can produce a stable clustering result. Clustering results are generally sensitive to cluster number K and the scaling parameter σ. Note that the subsequences do not necessarily consist of continuous frames.

3.3. Iterative reconstruction and clustering

After shape clustering, each subsequence is reconstructed separately by using off-the-shelf NRSfM methods. To get a refined and stable result, we perform the above method in an iterative manner. Each time we get a 3D reconstructed sequence, it is used to update the similarity matrix and corresponding shape clustering. The complete algorithm is

illustrated in Alg. 1 and also demonstrated in Fig.1.

4. EXPERIMENTAL RESULTS

We conducted extensive experiments on various long and complex motion sequences. As PND [10] is the state-of-the-art method, for the sake of space, we only compared our results with PND. 3D reconstruction error $e_{3D} = \|\mathtt{S} - \mathtt{S}^{\mathrm{GT}}\|_F / \|\mathtt{S}^{\mathrm{GT}}\|_F$ is used to evaluate the performance.

4.1. Datasets

UMPM Dataset: The Utrecht Multi-Person Motion (UMPM) benchmark [19] is a collection of video recordings of long and complex human motion sequences. In each sequence we extracted one human represented by 15 virtual joint positions at 50 fps frame rate. Six sequences are used in our experiment: *3_ball_12, p3_chair_16, p3_triangle_11, p4_circle_12, p4_free_11, and p4_table_11*, and the sequence lengths vary from 2537 frames to 3143 frames.

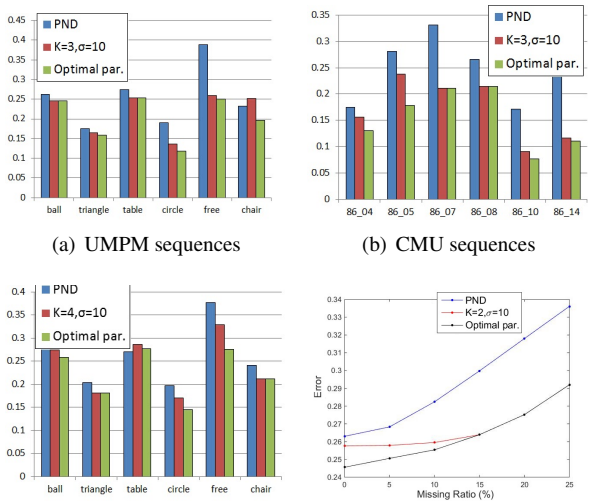
CMU Mocap dataset: The CMU Mocap dataset also contains long and complex human motions. In each sequence, 28 marker positions for one human are extracted at 40 fps. We used six CMU sequences: *CMU86_04, CMU86_05, CMU86_07, CMU86_08, CMU86_10, and CMU86_14*, whose lengths are between 2018 frames and 3359 frames.

4.2. Reconstruction results

For each UMPM and CMU sequence, we have a combination of the number of cluster K varying from 2 to 5 and scaling parameter σ of 10. Figure 3 shows the performance of our method and PND on various datasets and configurations. In all the figures, we compare three methods, namely PND, our method with fixed parameters and our method with optimal parameter for each sequence individually.

As shown in Fig. 3, on UMPM dataset our method outperforms PND on 5 out of the 6 sequences for fixed parameters. If we have the freedom to select parameters for each sequence individually, our method outperforms PND on all the 6 sequences. On CMU sequences, our method outperforms PND on all sequences under different configurations.

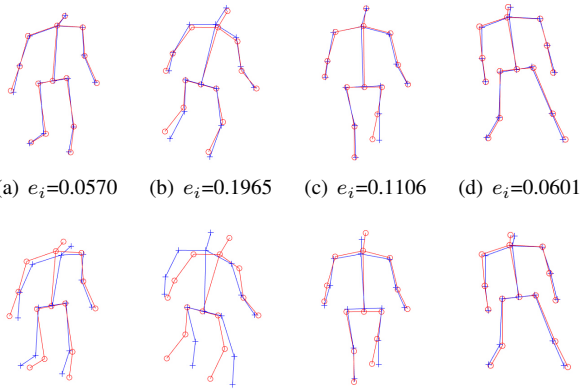
We also conducted experiments on noisy measurements case, where Gaussian noise was added to the UMPM sequences with a standard deviation of $\sigma_n = 0.01 \max\{\mathtt{W}\}$. In Figure 3(c), our method outperforms PND in 5 of the 6 sequences again. Finally, we evaluated our method under incomplete measurements case, where random missing data ratio ranges from 5% to 25%. As demonstrated in Figure 3(d), on the UMPM "Ball" sequence, our method shows superior performance compared with PND with random missing data. In both cases, our method achieves better results than PND, which proves that shape clustering does not affect the baseline method's capability of missing data handling.



(a) UMPM sequences (b) CMU sequences

(c) UMPM sequences with noise (d) "Ball" with missing data

Fig. 3: 3D reconstruction results on different sequences under different configurations.



(a) e_i=0.0570 (b) e_i=0.1965 (c) e_i=0.1106 (d) e_i=0.0601

(e) e_i=0.1788 (f) e_i=0.5052 (g) e_i=0.2963 (h) e_i=0.1492

Fig. 4: 3D reconstruction results of our method (top row) and PND (bottom row) on UMPM dataset. "○" indicates ground truth and "+" indicates 3D reconstruction points. Parameters: $K = 3, \sigma = 10$. Sequences from left to right: circle, free, table, triangle.

In Fig. 4, we illustrate the 3D reconstruction on the UMPM "Triangle" sequence for our method and PND. Clearly, our methods outperforms the current state-of-the-art NRSfM method PND by a margin.

5. CONCLUSION

In this paper, we present a novel reconstructability measure to general NRSfM and an iterative shape clustering based NRSfM method. Our method is easy to implement and pushes the performance of NRSfM methods to a new limit. Future work include extending our method to dense NRSfM case and automatic model selection in shape clustering.

6. REFERENCES

[1] Ijaz Akhter, Yaser Sheikh, Sohaib Khan, and Takeo Kanade, "Nonrigid structure from motion in trajectory space," in *Advances in Neural Information Processing Systems*, 2008, pp. 41–48.

[2] P.F.U. Gotardo and A.M. Martinez, "Non-rigid structure from motion with complementary rank-3 spaces," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2011, pp. 3065–3072.

[3] Yuchao Dai, Hongdong Li, and Mingyi He, "A simple prior-free method for non-rigid structure-from-motion factorization," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2012, pp. 2018–2025.

[4] Christoph Bregler, Aaron Hertzmann, and Henning Biermann, "Recovering non-rigid 3D shape from image streams," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2000, pp. 690–696.

[5] Jing Xiao, Jinxiang Chai, and Takeo Kanade, "A closed-form solution to non-rigid shape and motion recovery," *Int'l J. Computer Vision*, vol. 67, no. 2, pp. 233–246, 2006.

[6] Alessio Del Bue, João Xavier, Lourdes Agapito, and Marco Paladini, "Bilinear factorization via augmented lagrange multipliers," in *Proc. European Conf. Computer Vision*, 2010, pp. 283–296.

[7] Tomas Simon, Jack Valmadre, Iain Matthews, and Yaser Sheikh, "Separable spatiotemporal priors for convex reconstruction of time-varying 3d point clouds," in *European Conference on Computer Vision*, pp. 204–219. 2014.

[8] Antonio Agudo and Francesc Moreno-Noguer, "Simultaneous pose and non-rigid shape with particle dynamics," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2015, pp. 2179–2187.

[9] I. Akhter, Y. Sheikh, and S. Khan, "In defense of orthonormality constraints for nonrigid structure from motion," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2009, pp. 1534–1541.

[10] Minsik Lee, Jungchan Cho, Chong-Ho Choi, and Songhwai Oh, "Procrustean normal distribution for non-rigid structure from motion," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2013, pp. 1280–1287.

[11] Lorenzo Torresani, Aaron Hertzmann, and Chris Bregler, "Nonrigid structure-from-motion: Estimating shape and motion with hierarchical priors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 5, pp. 878–892, 2008.

[12] 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 *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2009, pp. 2898–2905.

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

[14] Ravi Garg, Anastasios Roussos, and Lourdes Agapito, "A variational approach to video registration with subspace constraints," *International Journal of Computer Vision*, vol. 104, no. 3, pp. 286–314, 2013.

[15] Chris Russell, Rui Yu, and Lourdes Agapito, "Video pop-up: Monocular 3d reconstruction of dynamic scenes," in *European Conference on Computer Vision*, pp. 583–598. 2014.

[16] Yingying Zhu, Dong Huang, Fernando De La Torre, and Simon Lucey, "Complex non-rigid motion 3d reconstruction by union of subspaces," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, June 2014, pp. 1542–1549.

[17] Hyun Soo Park, Takaaki Shiratori, Iain Matthews, and Yaser Sheikh, "3D reconstruction of a moving point from a series of 2D projections," in *Proc. European Conf. Computer Vision*, 2010, pp. 158–171.

[18] HyunSoo Park, Takaaki Shiratori, Iain Matthews, and Yaser Sheikh, "3d trajectory reconstruction under perspective projection," *International Journal of Computer Vision*, pp. 1–21, 2015.

[19] N.P. van der Aa, X. Luo, G.J. Giezeman, R.T. Tan, and R.C. Veltkamp, "Umpm benchmark: A multi-person dataset with synchronized video and motion capture data for evaluation of articulated human motion and interaction," in *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, Nov 2011, pp. 1264–1269.

[20] Andrew Y. Ng, Michael I. Jordan, and Yair Weiss, "On spectral clustering: Analysis and an algorithm," in *Advances in Neural Information Processing Systems*, pp. 849–856. 2002.