Articulated Structure from Motion and Segmentation

Zhen Zeng

May 1, 2013

Abstract

In this paper, we aim to analyze and reconstruct an articulated object structure given an input video. The articulated structure is represented by several rigid body parts connected by joints, where the joints are the overlapping area of each connected parts. Given a video of an articulated object, such as human, doing some actions in front of the camera, we are able to discover the underlying articulated structure and generate point based 3D reconstruction of the articulated object. The structure and reconstruction are solved jointly by optimizing a global energy function. The work is composed by two parts that iteratively optimize between the reconstruction error and articulated structure. Experiments results show that we can successfully generate reasonable articulated structure, and give good 3D reconstruction at the same time.

1 Introduction

Articulated object such as human demonstrates lots of variability, such as global strucuture variation induced by articulated motion, and local surface variation indeuced by deformation. Compared to other rigid objects or structure, articulated object generally needs more complex models to due to its large variability. One compact way to represent an articulated object is its underlying 3D skeleton structure [9]. A skeleton 1.1essentially captures the approximate rigid body parts of the object, and also the joints that connect different parts.

One interesting articulated object example is human. Identifying the human part information may provide more useful cues for further application like human pose estimation or human action recognition. Traditionally, human action recognition uses the labeled image that already define human part information beforehand. Although some promising results has been achieved, labeled data required people manually annotated which is not that practical. Furthermore, learning the body part information from images may not be robust enough to handle different kind of scenario. Thus in this paper, we try to discover the articulated object structure and apply it to videos of human, which are possible to be applied in the future human pose estimation.

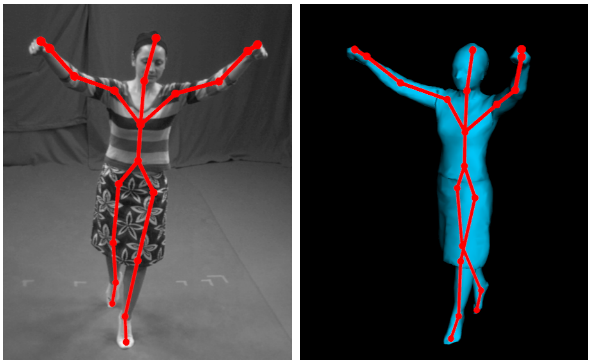


Figure 1.1: skeleton example.

Articulated structure from motion (A-SfM) in video aims at automatic reconstruction of the 3D points and the camera pose at each frame, and it also identifies which group of points belong to which rigid part, with the localization of joints connecting the parts. Thus articulated structure from motion can discover the underlying skeleton of an articulated object, but also with certain information on the 3D volumn each part possesses. In this work, we are going to investigate the state-of-art articulated structure from motion techniques in terms of its robustness to the noise of the tracked points, and background clutter.

2 Related Work

There were some previous works on 3D pose estimation of human, which is an instance of articulated object of the most interest. And since this kind of work targets particularly at the pose estimation of human, they usually have a predefined structure knowledge of human [8, 1, 3], which is not generally appicable when the object of interest changes, and experts are needed again for the design of other articulated objects.

Shotton et.al. [8] proposed a method to localize human body joint locations in real time given a single depth image. Their method needs device such as kinect to get the depth image, which is not as general as digital camera, and our method only needs video frames other than depth. And the robustness of their method depends heavily on their huge amount of training data. Their training dataset is composed by two parts: one generated automatically from predefined human structure with reasonable varaitions, which is designed by experts; the other contains realistic depth images of human, with manual labels of body joints. Although their efforts in 3D synthesized data of first set of dataset did save a lot of human efforts of labeling, they still need manual labels for the second dataset.

Previous works on articulated strcture from motion assumes the 2D trajectories of the points that belong to the articulated object is given [6][13], or easily gotten from pre-processing of the video by existing background subtraction work, or even manually selected [4]. Yan et.al. [13] discovered the articulated structure by identifying the ranks of the point trajectories, and joints that are connecting two rigid parts are localized as the intersection of two motion subspaces representing the two parts. Ross et.al. [4] proposed a probabilistic approach to learn the articulated structure, but their initial location of the joints are determined by an initial motion segmentation. Thus their method is prone to a poor initialization.

Fayad et.al. [6] is the first to solve the full body reconstruction and articulated structure discovery simultaneously. Unlike previous methods, they started with an excessive models, where each point and its neighbours are treated as a rigid part in the initialization stage. Thus they don't rely on any initial motion segmentation results, instead, they iteratively solve the 3D reconstruction of each rigid body part and the segmentation problem until the global energy stops decreasing. In their context, segmentation means assigning points to different rigid parts, with certain overlapping between different parts. Thus the segmentation result implies the underlying skeleton structure, as shown in fig 2.1 (4), each overlapping of parts are treated as a joint.

Yet the dataset being used by these works are relatively simple, with limited motion of camera. The scenario that we are interested in are the cases where the camera motion is not restricted, the foreground-background subtraction itself becomes a hard problem. There are recent state-of-the-art work on foreground subtraction by Elqursh et.al. [5] , Lee et.al. [10], Liu et.al. [7] and [14]. [7, 10]search for reliable keysegments by loading the whole video in first, where keysegments are regions that demenstrate distringuished motion from its surroundings. Then they learn color, shape or locality model of those key-segments as their foreground model. [5] learn both foreground and background model, and use Bayesian filtering to estimate the posterior of each pixel being foreground or background. Although all of them showed very promising results already on videos with severe motion involed, they still cannot provide pure foreground trajectories as input to those existing A-SfM works, without including single background noise. And the quality of the tracked points might also suffer due to motion blur.

Thus we are going to replicate the work by [6] first, and then experiment on how much the final articulated structure is going to suffer due to the noises of the point trajectories due to moving camera. We are going to evaluate it both qualitatively and quantitatively, examing on the structure discovery, joint localization and the reprojection error of 3D reconstruction.

3 Technical Part

As introduced in [6], the problem can be solved in an iterative fashion: a) as explained in 3.1, if the number of rigid body parts are known, and the point assignment to each rigid parts are known, the motion of the camera can be

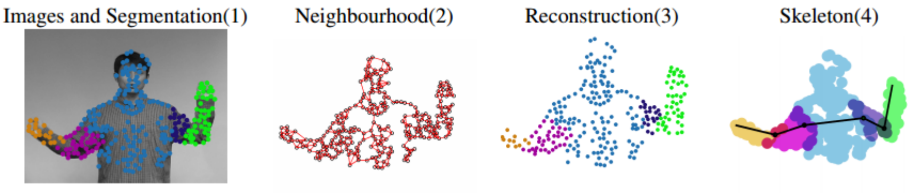


Figure 2.1: Reconstruction result[6].

estimated by traditional rigid structure from motion on each individual rigid part; b) as explained in 3.2, if the motion is known, each point can be assigned to the motion that explains it best, while encouraging neighbored points shared the same motion.

Note that there is no explicit localization for joint, but since each point can be assigned to multiple parts, so joints actually lie on those points being assigned to multiple parts. For example, as in fig 2.1(4), the points on the right shoulder are assigned to both torso and also right arm, so it indicates the location of the right shoulder.

3.1 Rigid SfM

Given a set of 2D trajectories that belong to the same rigid part, we can directly apply existing work on rigid SfM to get the reconstruction result. The work being used is [11]. They assumed the camera to be the orthographic model. Like other factorization method in rigid structure from motion, given the the 2D point matrix Z 3.1, where P is the number of points of each frame, and F is the number of frames,

$$Z = \begin{bmatrix} x_1^1 & \dots & x_1^P \\ y_1^1 & \dots & y_1^P \\ \vdots & \ddots & \vdots \\ x_F^1 & \dots & x_F^P \\ y_F^1 & \dots & y_F^P \end{bmatrix} \tag{3.1}$$

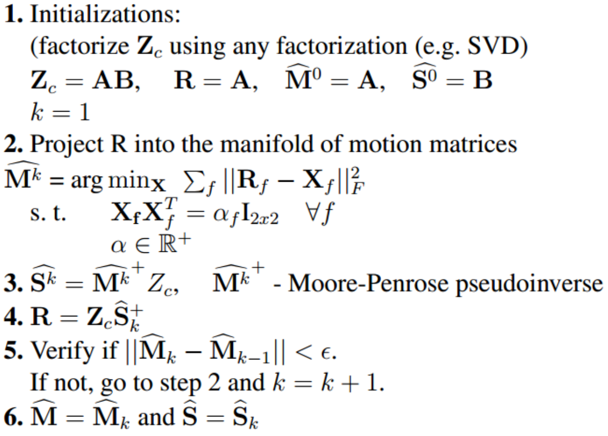
They first translate it by the mean of the point locations such that the center lies at the origin 3.2, where M, S are the motion parameters (only rotation in this case) and 3D point coordinates respectively,

$$Z_c = Z - center(Z) = MS \tag{3.2}$$

where

$$M = \begin{bmatrix} M_1 \\ \dots \\ M_F \end{bmatrix}, \, with \, M_f = \begin{bmatrix} i_x^f & i_y^f & i_z^f \\ j_x^f & j_y^f & j_z^f \end{bmatrix} = \begin{bmatrix} i^f \\ j^f \end{bmatrix} \tag{3.3}$$

Algorithm 1 Rigid Factorization [11]



```
1. Initializations:
   (factorize $\mathbf{Z}_c$ using any factorization (e.g. SVD)
   $\mathbf{Z}_c = \mathbf{AB}$,   $\mathbf{R} = \mathbf{A}$,   $\widehat{\mathbf{M}}^0 = \mathbf{A}$,   $\widehat{\mathbf{S}}^0 = \mathbf{B}$
   $k = 1$
2. Project R into the manifold of motion matrices
$\widehat{\mathbf{M}}^k$ = arg min$_\mathbf{X}$ $\sum_f ||\mathbf{R}_f - \mathbf{X}_f||_F^2$
   s. t.      $\mathbf{X_f X}_f^T = \alpha_f \mathbf{I}_{2x2}$   $\forall f$
              $\alpha \in \mathbb{R}^+$
3. $\widehat{\mathbf{S}^k} = \widehat{\mathbf{M}^k}^+ Z_c$,   $\widehat{\mathbf{M}^k}^+$ - Moore-Penrose pseudoinverse
4. $\mathbf{R} = \mathbf{Z}_c \widehat{\mathbf{S}}_k^+$
5. Verify if $||\widehat{\mathbf{M}}_k - \widehat{\mathbf{M}}_{k-1}|| < \epsilon$.
   If not, go to step 2 and $k = k + 1$.
6. $\widehat{\mathbf{M}} = \widehat{\mathbf{M}}_k$ and $\widehat{\mathbf{S}} = \widehat{\mathbf{S}}_k$
```

Then they iteratively project the left factor of Z_c on the motion manifold that is determined by the motion constriants 1. The motion constriants state that the motion matrix should be composed by two scaled orthonormal rows, as defined below

$$\begin{array}{rcl} i^f i^{fT} & = & \alpha^f \\ j^f j^{fT} & = & \alpha^f \\ i^f j^{fT} & = & 0 \end{array} \Rightarrow M_f M_f^T = \alpha^f I_{2x2} \tag{3.4}$$

The benefit of their method is that they can deal with 2D trajectories with missing data, which is a common case in realistic applications. With some random initialization for the lost data in Z, they can iteratively optimize over M and S to minimize the projection error. The algorithm is detailed in 2

3.2 Articulated Structure Optimization

Given the set of motions (or models) M for each rigid part, each point is assigned to one or multiple parts based on the corresponding projection error, and a regularization term controlling the number of parts. The tracked points are P. And the assignment of points to parts are denoted by $\mathbf{m} = \{m_1, m_2, ..., m_p\}$, where p is the number of points. And the energy to be optimized is

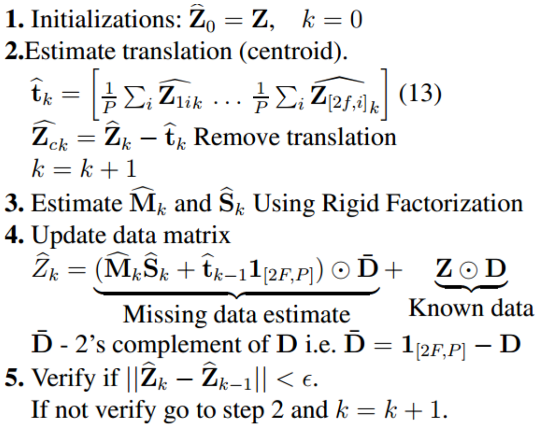
$$argmin_{\mathbf{m}\in(2^M)^p} C(\mathbf{m}) = \sum_{p\in P} (\sum_{\alpha \in m_p} U_p(\alpha)) + MDL(\mathbf{m}) \tag{3.5}$$

subject to

$$\forall p \in P \, \exists \alpha \, : \, p \in I_\alpha \tag{3.6}$$

$$\forall q \in N_p \, \wedge \, q \in I_\alpha \Longrightarrow \alpha \in m_p \tag{3.7}$$

Algorithm 2 Rigid Factorization with Missing Data [11]



```
1. Initializations: $\widehat{\mathbf{Z}}_0 = \mathbf{Z}, \quad k = 0$
2.Estimate translation (centroid).
  $\widehat{\mathbf{t}}_k = \left[ \frac{1}{P} \sum_i \widehat{\mathbf{Z}_{1ik}} \ldots \frac{1}{P} \sum_i \widehat{\mathbf{Z}_{[2f,i]}}_k \right]$ (13)
  $\widehat{\mathbf{Z}}_{ck} = \widehat{\mathbf{Z}}_k - \widehat{\mathbf{t}}_k$ Remove translation
  $k = k + 1$
3. Estimate $\widehat{\mathbf{M}}_k$ and $\widehat{\mathbf{S}}_k$ Using Rigid Factorization
4. Update data matrix
  $\widehat{Z}_k = \underbrace{(\widehat{\mathbf{M}}_k \widehat{\mathbf{S}}_k + \widehat{\mathbf{t}}_{k-1} \mathbf{1}_{[2F,P]}) \odot \bar{\mathbf{D}}}_{\text{Missing data estimate}} + \underbrace{\mathbf{Z} \odot \mathbf{D}}_{\text{Known data}}$
  $\bar{\mathbf{D}}$ - 2's complement of $\mathbf{D}$ i.e. $\bar{\mathbf{D}} = \mathbf{1}_{[2F,P]} - \mathbf{D}$
5. Verify if $||\widehat{\mathbf{Z}}_k - \widehat{\mathbf{Z}}_{k-1}|| < \epsilon$.
  If not verify go to step 2 and $k = k + 1$.
```

where I_α is the interior of model α, a point is in the interior of a model means that the point and all its neighbors belong to α. And constriant 3.7 puts hard constraint on that point p and its neighbours should have at least one shared model. This problem cannot be solved by a standard MRF framework, since the points can belong to multiple model. A recent work [12] uses modified α-expansion to solve multiple model assignment problem. Currently we are using quadratic function for MDL term, which is increasing as the number of models increases. And the cost $U_p(\alpha)$ is defined as the reprojection error 3.8

$$U_p(\alpha) = \|Z_p - (R_\alpha S_p + T_\alpha)\|^2 \tag{3.8}$$

3.3 Stitching Rigid Parts

After the rigid SfM for each part, we need to stitch them together based on the overlapping points between parts. Since the result of rigid SfM is based on the local coordinate reference to each rigid part, we need to estimate the rotation matrix between two parts for the stitching. Currently we use standard Kabsch algorithm for the alignment of 3D points. Basically, first we search for pairs that overlapping with each other with the most number of points, and align them by minimizing the euclidean distances between the overlapping points after transformation.

Note that when we aligned the pair of points together, their coordinates changed from the individual orignal references to the global one, which requires aF transformation in the motion parameters correspondingly, so that the 2D projections stay the same. For example, for rigid part i, the reconstruction and motion estimation result from 3.1 are S_i,R_i,T_i, and after the stitching, it is transformed into S_i' by R_k,T_k derived by Kabsch algorithm, then we need to update R_i,T_i to R_i',T_i' accordingly, as derived in 3.9

$$
\begin{aligned}
Z_i &= R_i S_i + T_i \\
S'_i &= R_k S_i + T_k \\
so\ Z_i &= R_i R_k^{-1} S'_i - R_i R_k^{-1} T_k + T_i \\
\Rightarrow \quad R'_i &= R_i R_k^{-1} \\
T'_i &= T_r - R_i R_k^{-1} T_k
\end{aligned}
\tag{3.9}
$$

The 3D coordinatees of the overlapping points are averaged between the pair. Thus the 3D reconstruct result of the first pair determines the global principal coordinates for the rest points, then we incrementally align more and more parts together to generate the full body reconstruction.

3.4 Initialization

As mentioned previously, in order to avoid errors due to poor initialization, we used excessive models at initializatoin stage. Instead of relying on motion segmentation which usually asks for the number of clusters expected, we treat every point together with its neighbours as a rigid part, and apply the rigid factorization to get the 3D reconstruction. As expected, there will be a lot of overlapping between many pairs of rigid parts, so it will take some time to stitch all of them together. Once all the reconstructed 3D points are unified in a single reference as explained in 3.3, given the transformed motion paramters as motion models, we assign each 3D point to the best model based on the energy minimization. After we get the new assignment of points to rigid parts, we are able to again apply the rigid reconstruction, and carry on the iterations until the defined global energy stops decreasing.

The neighbourhood of each points is defined as the nearest k points in the similarity measurement based on location and motion across all the frames. Inspired by [2], the distance or dissimilarity between two points are defined as the largest distance between them over the frames that both of the points exist in. In order to be able to reconstruct a rigid part at the initialization stage, each part should contain at least 3 points, which indicates that k should be at least 3.

4 Experiments

The regularizing term defined by MDL 3.5 is a critial element that determines the final optimal solution, it should not be increasing too quickly as the number of active models increase, otherwise the algorithm will prefer as simple models as possible, resulting in possible more errors in reconstruction; on the other hand, it should not be increasing too slowly as the number of active models increase, otherwise we will end up with more models than actually needed, reducing the compactness of the representation of the articulated structure.

We tested our algorithms on various videos provided by [6], and the corresponding 2D projection error together with discovered number of rigid parts of the optimal solution are as shown in 1.

datasets	dancing	booklet	head	puppet
# rigid parts	7	3	3	8
reprojection error	2.5672	2.9993	4.4332	3.8857
	toy_truck	yellow_crane	two_cranes	
# rigid parts	3	8	5	
reprojection error	2.2123	3.2546	2.8325	

Table 1: 2D projection error of 3D reconstruction & rigid parts

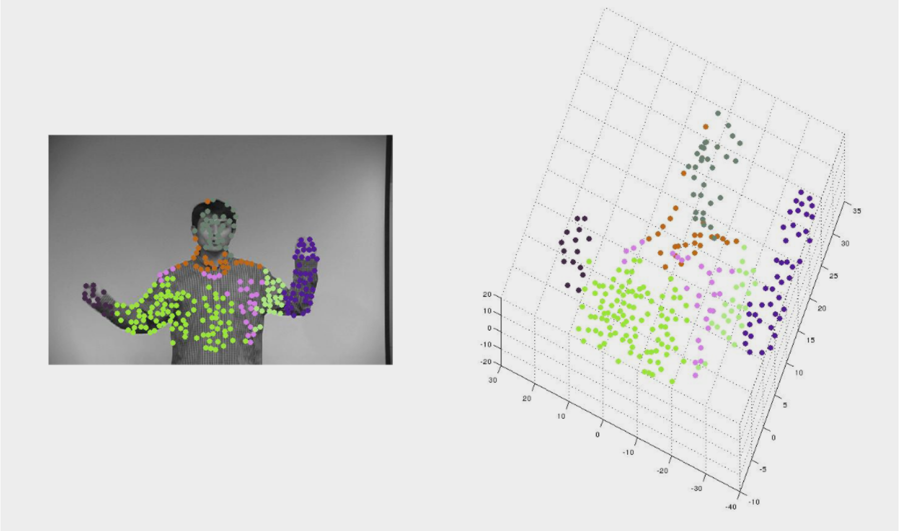


Figure 4.1: Qualitative Results.

As we can see, the partition of different articulated object is not as consistent as we expected it to be from human standard perspective, and in order to make it more consistent to what we expected, we can try to learn the weights for the qudratic terms in MDL using machine learning methods in the future. And results from appearance based methods such as superpixel could be a useful cue in the future to group points together as one rigid part.

And some qualitative results are shown in 4.1, different color means different rigid parts. And the assignment in 2D is on the left, and the 3D reconstruction result is on the right.

5 Conclusion

When the point trajectories of the foreground object is given, the replicate method works well at both identifying rigid parts, joints and 3D reconstruction. One remianing problem is that when the point trajectories also contain the background points, it needs further analysis in both 2D and 3D to evaluate whether the meethod can effectively separate independtly moving foreground object from the background, and avoiding separating the background into multiple rigid parts at the same time.

Once the problem of foreground and background subtraction can be solved effiecently in this framework, one interesting future work is to automatically learn general 3D and 2D model of an articulated object category given a set of videos with the object of interest demonstrating articulate motion. In this way, intead of predefining a general 3D model and augment it by 2D appearance, we can learn the 3D and 2D model at the same time.

References

[1] C. Bregler and J. Malik. Tracking people with twists and exponential maps. In *Proceedings. 1998 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No.98CB36231)*, pages 8–15. IEEE Comput. Soc.

[2] Thomas Brox and Jitendra Malik. Object segmentation by long term analysis of point trajectories. pages 282–295, September 2010.

[3] Bill Triggs Cristian Sminchisescu. Estimating Articulated Human Motion With Covariance Scaled Sampling.

[4] Daniel Tarlow David A. Ross. Unsupervised learning of skeletons from motion.

[5] Ali Elqursh and Ahmed Elgammal. Online Moving Camera Background Subtraction. In Andrew Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato, and Cordelia Schmid, editors, *Computer Vision ECCV 2012*, volume 7577 of *Lecture Notes in Computer Science*, pages 228–241. Springer Berlin Heidelberg, 2012.

[6] Joao Fayad, Chris Russell, and Lourdes Agapito. Automated articulated structure and 3D shape recovery from point correspondences. In *2011 International Conference on Computer Vision*, pages 431–438. IEEE, November 2011.

[7] Feng Liu,Michael Gleicher. Learning color and locality cues for moving object detection and segmentation. pages 320 – 327, 2009.

[8] Andrew Fitzgibbon Jamie Shotton. Real-time human pose recognition in parts from single depth images.

[9] Carsten Stoll Juergen Gall. Motion capture using joint skeleton tracking and surface estimation.

[10] Yong Jae Lee, Jaechul Kim, and Kristen Grauman. Key-segments for video object segmentation. In *2011 International Conference on Computer Vision*, pages 1995–2002. IEEE, November 2011.

[11] Manuel Marques and João Costeira. Estimating 3D shape from degenerate sequences with missing data. *Computer Vision and Image Understanding*, 113(2):261–272, February 2009.

[12] Chris Russell, Joao Fayad, and Lourdes Agapito. Energy based multiple model fitting for non-rigid structure from motion. In *CVPR 2011*, pages 3009–3016. IEEE, June 2011.

[13] Jingyu Yan and Marc Pollefeys. A factorization-based approach for articulated nonrigid shape, motion and kinematic chain recovery from video. *IEEE transactions on pattern analysis and machine intelligence*, 30(5):865–77, May 2008.

[14] Yaser Sheikh,Omar Javed,Takeo Kanade. Background Subtraction for Freely Moving Cameras. pages 1219 – 1225, 2009.