NON-RIGID STRUCTURE FROM MOTION VIA SPARSE SELF-EXPRESSIVE REPRESENTATION

Junjie Hu[1], *Terumasa Aoki*[1,2]

Tohoku University, Japan
[1]Graduate School of Information Sciences (GSIS)
[2]New Industry Creation Hatchery Center (NICHe)

ABSTRACT

To simultaneously recover 3D shapes of non-rigid object and camera motions from 2D corresponding points is a difficult task in computer vision. This task is called Non-rigid Structure from motion(NRSfM). To solve this ill-posed problem, many existing methods rely on low rank assumption. However, the value of rank has to be accurately predefined because incorrect value can largely degrade the reconstruction performance. Unfortunately, these is no automatic solution to determine this value. In this paper, we present a self-expressive method that models 3D shapes with a sparse combination of other 3D shapes from the same structure. One of the biggest advantages is that it doesn't need the rank to be predefined. Also, unlike other learning-based methods, our method doesn't need learning step. Experimental results validate the efficiency of our method.

***Index Terms*—** Non-rigid Structure from Motion, low rank, self-expressive, sparse combination

1. INTRODUCTION

Non-rigid Structure from motion(NRSfM) is a difficult problem to obtain the perfect solution because of the inherently high number of degrees of freedom. To solve this ill-posed problem, many efficient methods use low rank constraint such as modeling 3D shapes by using several 3D shape bases [1, 2, 3] or representing point trajectories with a linear combination of a fixed set of discrete cosine transform (DCT) trajectory bases [4, 5, 6]. Besides, Dai et al. proposed a well-known method (BMM) that minimizes the rank of 3D shapes [7]. BMM is one of the best approaches that achieves the most remarkable performance. However, the difficulty of low rank based methods is that the number of rank must be predefined accurately, because the incorrect number will largely degrade the algorithm's performance. Unfortunately, the simple way to find the optimal number has not been discovered, so we have to repeat numerous experiments to estimate it.

Recently, several learning-based methods also have been proposed. For example,[8] introduced a shape prior that constrains the reconstructed shapes to lie in the learned manifold. [9, 10, 11] proposed sparse representation methods that model 3D shapes or point trajectories by using learned dictionary. However, they need to learn shape bases or trajectory bases in advance. Since there are so many different deformable 3D objects in real situation, these methods are difficult to use in practice.

On the other hand, [12] proposed a self-expressive method to recover 3D points trajectories which doesn't need to learn shape bases or trajectory bases in advance. It represents 3D trajectory of each non-rigid 3D point with a linear combination of other 3D trajectories from the same structure. This method can recover 3D shapes of multi-objects at the same time, but its 3D reconstruction accuracy is lower than BMM. Also, [13] proposed a self-expressive method to recover 3D shapes. They introduced a significant symmetric constraint to sparse coefficient matrix which can improve reconstruction performance. In this paper, we combine the advantages of these two methods and propose a new self-expressive method. The proposed method doesn't rely on the choice of rank and can achieve state-of-the-art reconstruction performance.

2. PROPOSED METHOD

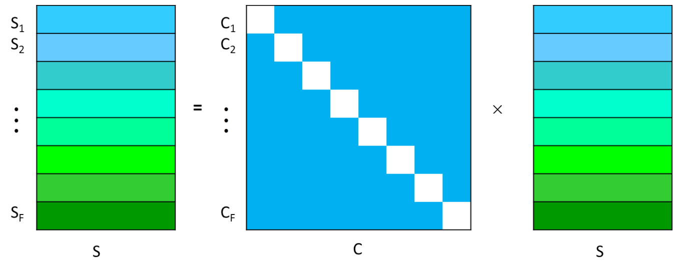
NRSfM problem can be represented as:

$$\begin{pmatrix} \mathbf{W}_1 \\ \vdots \\ \mathbf{W}_F \end{pmatrix} = \begin{pmatrix} \mathbf{R}_1 & & \\ & \ddots & \\ & & \mathbf{R}_F \end{pmatrix} * \begin{pmatrix} \mathbf{S}_1 \\ \vdots \\ \mathbf{S}_F \end{pmatrix} \quad (1)$$

And a self-expressive representation method for NRSfM attempts to solve the following function:

$$\begin{aligned} &\min_{\mathbf{C,S}} \quad \frac{1}{2}||\mathbf{W} - \mathbf{RS}||_F^2 + \lambda||\mathbf{C}||_1 \\ &\text{s.t.} \quad \mathbf{S} = \mathbf{CS}, \quad \mathbf{C}_{i,i} = 0, \quad \mathbf{C}_{i,j} \geq 0, \quad \mathbf{R}_i\mathbf{R}_i^T = \mathbf{I} \end{aligned} \quad (2)$$

where $\mathbf{R} \in \mathbb{R}^{2F \times 3F}$ is the orthogonal camera motion matrix, $\mathbf{R}_i$ denotes camera motion of *i*-th frame. $\mathbf{S} \in \mathbb{R}^{3F \times P}$ is the 3D non-rigid shapes matrix. $\mathbf{C}$ is a sparse coefficient matrix, $\mathbf{C}_{i,j}$ denotes the weight of *j*-th 3D shape for modeling *i*-th shape $\mathbf{S}_i$. Also, F and P denote the total number of frames



,

Fig. 1. Sparse self-expressive representation of 3D structures. i-th non-rigid 3D shape $\mathbf{S}_i$ is represented with a sparse combination of other 3D shapes from the same structure.

and points respectively. $\mathbf{W} \in \mathbb{R}^{2F \times P}$ is the projections of $\mathbf{S}$ in a set of 2D images.

It assumes that i-th 3D shape $\mathbf{S}_i$ can be well modeled by the sparse linear combination of other 3D shapes in $\mathbf{S}$. It can be represented as:

$$\mathbf{S}_i = \mathbf{C}_{i,1}\mathbf{S}_1 + \cdots + \mathbf{C}_{i,i-1}\mathbf{S}_{i-1} + \mathbf{0}\mathbf{S}_i + \\ \mathbf{C}_{i,i+1}\mathbf{S}_{i+1} + \cdots + \mathbf{C}_{i,F}\mathbf{S}_F \quad (3)$$

[13] introduced that $\mathbf{S}_i$ and $\mathbf{S}_j$ should have the same contribution to each other such that $\mathbf{C}_{i,j}$ must be equal to $\mathbf{C}_{j,i}$. Thus, $\mathbf{C}$ is a symmetric matrix.

$$\mathbf{C} = \mathbf{C}^T \quad (4)$$

Their experimental results validate it's a significant constraint. Thus, we also define $\mathbf{C}$ as a symmetric matrix based on this constraint. Then the self-expressive representation problem can be represented as Fig. 1. And like the method in [7, 12], we penalize the rank of $F \times 3P$ matrix P($\mathbf{S}$) to constrain the unrealistic deformations. Because the rank minimization problem is NP hard, we implement it by minimizing the nuclear norm which is the tightest convex relaxation of the rank of a matrix. Here, we use P($\mathbf{S}$) as an operator that re-arranges the entries of $\mathbf{S}$ into $F \times 3P$ matrix such that the i-th row of P(S) contains X, Y, Z coordinates of all points of the shape at frame i (i.e. all values of $\mathbf{S}_i$).

We propose to solve NRSfM via sparse self-expressive representation by minimizing an energy of the following form:

$$E = E_{shape} + \beta_1 E_{img} + \beta_2 E_{rank} \quad (5)$$

where:

$$E_{shape} = \frac{1}{2}||\mathbf{S} - \mathbf{CS}||_F^2 + \lambda_1||\mathbf{C} - \mathbf{C}^T||_F^2 + \lambda_2||\mathbf{C}||_1 \quad (6)$$

E_{shape} aims at modeling 3D shapes by a linear sparse combination of other shapes with additional constraint that $\mathbf{C}$ is a symmetric matrix. The second term minimizes reprojection error of all observed points, it can be represented as:

$$E_{img} = ||\mathbf{W} - \mathbf{RS}||_F^2 \quad (7)$$

The third term E_{rank} is to minimize the number of independent shapes used to represent the deformable object for compact representation.

$$E_{rank} = ||P(\mathbf{S})||_* \quad (8)$$

Thus, our energy function can be formulated as:

$$\min_{\mathbf{R,C,S}} \quad \frac{1}{2}||\mathbf{S} - \mathbf{CS}||_F^2 + \lambda_1||\mathbf{C} - \mathbf{C}^T||_F^2 + \lambda_2||\mathbf{C}||_1 \\ +\beta_1||\mathbf{W} - \mathbf{RS}||_F^2 + \beta_2||P(\mathbf{S})||_* \\ \text{s.t.} \quad \mathbf{C}_{i,i} = 0, \quad \mathbf{C}_{i,j} \geq 0, \quad \mathbf{R}_i\mathbf{R}_i^T = \mathbf{I} \quad (9)$$

The above energy function is non-convex, the initialization of variables is extremely important. In order to achieve good result, we initialize $\mathbf{R}$ using the method introduced in [4] and $\mathbf{S}$ is initialized by BMM. We solve the non-convex minimization problem by the following alternating scheme, as seen in Algorithm 1.

Algorithm 1 Self-expressive sparse representation for Non-rigid structure from motion

Input:
 2D correponding matrix W
Output:
 Camera motions R, 3D shapes S and sparse coefficient matrix C.

```
Initialize R, S;
while not converged do
    Fix R and S, update C;
    Fix R and C, update S;
    Fix C and S, update R;
end while
```

2.1. The solution to C

We compute $\mathbf{C}$ by minimizing (5). It's a convex problem. We solve it by Proximal Gradient Descent (PGD) algorithm [14]. The details are given in Algorithm 2.

Algorithm 2 The solution to sparse coefficient matrix C

Input:
3D shapes S, sparse coefficient matrix C.
Output:
Updated C.
Parameters:
λ_1, λ_2, step size μ.

while not converged **do**
  $\bigtriangledown f(C)$=$(CS-S)\ S^T$+$\frac{\lambda_1}{2}$ $(C$-$C^T)$;
  $C=C$-$\frac{1}{\mu}$($\bigtriangledown f(C)$+λ_2);
  if $C_{i,j}<0$ **then**
    $C_{i,j}=0$;
  end if
  $C_{i,i}=0$;
end while

2.2. The solution to S

S is updated by minimizing the following function:

$$\min_{\mathbf{S}} \quad \frac{1}{2}||\mathbf{S}-\mathbf{CS}||_F^2+\beta_1||\mathbf{W}-\mathbf{RS}||_F^2+\beta_2||P(\mathbf{S})||_* \quad (10)$$

We solve it by using fixed point iterative method [15]. The details of our solution are given in Algorithm 3.

Algorithm 3 Algorithm for updating S

Input:
3D shapes S, sparse coefficient matrix C, 2D corresponding points W, camera rotations R.
Output:
Updated 3D shapes S
Parameters:
β_1, β_2, τ and θ_0

for i = 1 to n **do**
  θ=θ_0;
  while not converged **do**
    $\bigtriangledown f(S)$=$(\mathbf{I}-C)^T(S-CS)$+$R^T(RS-W)$;
    $S=S$-$\tau\bigtriangledown f(S)$;
    $[U,D,V]$ = Singular Value Decomposition of $P(S)$;
    $D=max(D$-$\tau\theta\mathbf{I}$, 0);
    // where max(,) is an element-wise operator.
    S=$P^{-1}(UDV^T)$;
  end while
  θ = θ / 4;
end for

2.3. The solution to R

Camera rotations are updated by minimizing (6). **R** is an orthographic camera. We solve this as an instance of the Orthogonal Procrustes problem [16, 17]. By Using the Singular Value Decomposition (SVD), we can express

$$[\mathbf{U,D,V}]=svd(\mathbf{WS}^T(\mathbf{SS}^T)^{-1}); \quad (11)$$

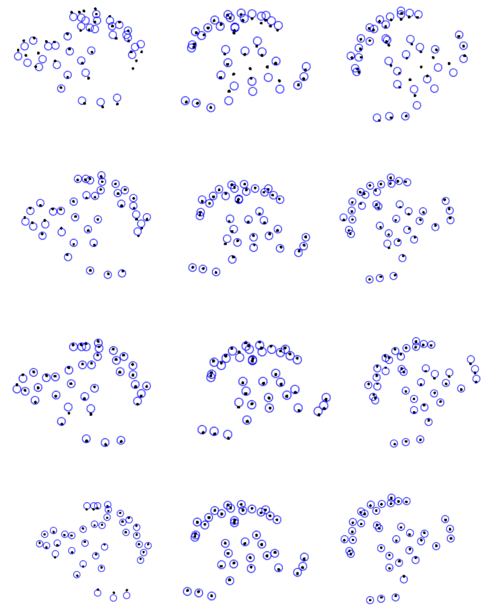


Fig. 2. From first row to last row: Front view of 3D shapes of Face1 recovered by PTA, CSF2, BMM, and the proposed method, respectively. Recovered shapes are blue circles and ground truth is dark dots.

Then we obtain **R** by setting $\mathbf{R}=\mathbf{UV}^T$.

3. EXPERIMENTAL RESULTS

In this section, we compare the proposed method with the state-of-the-art methods, such as the DCT trajectory based PTA [5]and CSF2 [4], Dai's nuclear minimization method BMM. It should also be pointed out that we can't compare our method against [13] because in this method the camera motion is known a priori.

We chose the same datasets that were chosen by the authors of the above methods, which contain Face1 (74/37) of [18]; Shark (240/91), Walking (260/55) and Face2 (316/40) of [3]; Drink (1102/41), Pickup(357/41), Yoga (307/41), Stretch (370/41), Dance(264/75) of [5]. Here (F/P) denotes the number of frames (F) and points (P). The 3D reconstruction error of non-rigid shapes is measured as:

$$e_{3D}=\frac{1}{\sigma FP}\sum_{f=1}^{F}\sum_{n=1}^{P}e_{fn}, \sigma=\frac{1}{3F}\sum_{f=1}^{F}(\sigma_{fx}+\sigma_{fy}+\sigma_{fz}) \quad (12)$$

where σ_{fx}, σ_{fy} and σ_{fz} are the standard deviations of the X, Y and Z coordinates of the original shape in frame *f*.

In our experiments, the value of rank is set in accordance with the original setting of the above low rank based methods for different datasets . Table 1 shows the 3D reconstruction error of each method on all datasets with recovered camera rotations. It is seen that the proposed method achieves

Table 1. Average 3D reconstruction error of PTA, CSF2, BMM, proposed method. (K) denotes the rank value which is used in low rank based methods. Blue cells are marked as the ones with the lowest error, and pink cells with the second lowest error.

Dataset	PTA(K)	CSF2(K)	BMM(K)	Proposed method
Face1	0.1247 (3)	0.0520 (5)	0.0502(11)	**0.0403**
Face2	0.0444(5)	0.0314(5)	0.0303(7)	**0.0266**
Walking	0.3954 (2)	**0.1035(5)**	0.1298(8)	0.1294
Pickup	0.2369(12)	0.2277(3)	**0.1731(12)**	0.2223
Stretch	0.1088(12)	0.0685(8)	0.1034(11)	**0.0647**
Yoga	0.1625(11)	0.1464(7)	**0.1150(10)**	0.1436
Drink	0.0250(13)	0.0223(6)	0.0266(12)	**0.0176**
Shark	0.1796(9)	**0.0444(5)**	0.2311(4)	0.1997
Dance	0.2958(5)	0.1983(7)	**0.1864(10)**	0.1889

Table 2. Average 3D reconstruction error of PTA, CSF2, BMM, and proposed method with ground truth camera rotations. Blue cells are marked as the ones with the lowest error, and pink cells with the second lowest error.

Dataset	PTA	CSF2	BMM	Proposed method
Pickup	0.0992	0.0814	0.0497	**0.0317**
Stretch	0.0822	0.0442	0.0456	**0.0273**
Yoga	0.0580	0.0371	0.0334	**0.0280**
Drink	0.0299	0.0215	0.0238	**0.0134**

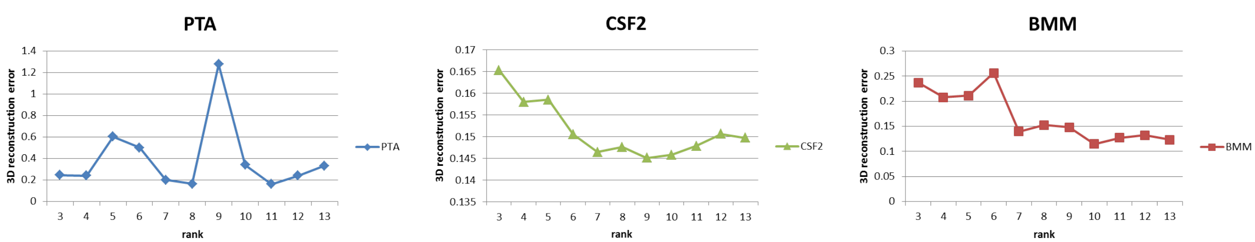


Fig. 3. 3D reconstruction error of Yoga recovered by low rank based methods with different rank. For better visualization, we show them respectively.

the best reconstruction performance on Face1, Face2, Stretch and Drink and gives second good reconstruction on Walking, Pickup, Yoga, Dance. Fig. 2 shows the visual evaluation of Face1.

The proposed method didn't achieve the best reconstruction on all datasets. We consider that this is because the camera rotations were not recovered accurately enough. To validate our assumption, we compare our proposed method against other methods with ground truth camera rotations. In this step, we evaluate algorithm performance on Drink, Pickup, Yoga, Stretch as they contain ground truth camera rotations. The results are shown in Table 2. It shows that the proposed method achieves the best reconstruction performance on each dataset. It proves that representing 3D shapes by self-expressive sparse representation can sufficiently improve algorithm's performance.

Fig. 3 shows the 3D reconstruction error of Yoga recovered by low rank based methods with different rank. It is clear that low rank based methods need the value of rank to be accurately predefined to achieve good reconstruction performance. If incorrect number is selected, their reconstruction performance can be largely degraded. Unlike PTA, CSF2 and BMM, our method doesn't rely on the choice of rank which is a big advantage in practice.

4. CONCLUSION

In this paper, we proposed a novel method for NRSfM via sparse self-expressive representation. One advantage of proposed method is that its reconstruction performance doesn't depend on the choice of rank. Also, compared to learning-based methods, it doesn't need training step. Experimental results show it can achieve state-of-the-art reconstruction performance on benchmark datasets. To further improve the performance of our method, future work will consider more accurate solutions for estimating camera rotations.

5. REFERENCES

[1] Christoph Bregler, Aaron Hertzmann, and Henning Biermann, "Recovering non-rigid 3d shape from image streams," in *Computer Vision and Pattern Recognition*. IEEE, 2000, vol. 2, pp. 690–696.

[2] Jing Xiao, Jinxiang Chai, and Takeo Kanade, "A closed-form solution to non-rigid shape and motion recovery," in *European Conference on Computer Vision*. 2004, pp. 573–587, Springer.

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

[4] Paulo FU Gotardo and Aleix M Martinez, "Non-rigid structure from motion with complementary rank-3 spaces," in *Computer Vision and Pattern Recognition*. IEEE, 2011, pp. 3065–3072.

[5] 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.

[6] Hyun Soo Park, Takaaki Shiratori, Iain Matthews, and Yaser Sheikh, "3d reconstruction of a moving point from a series of 2d projections," in *European Conference on Computer Vision*. Springer, 2010, pp. 158–171.

[7] Yuchao Dai, Hongdong Li, and Mingyi 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] Lili Tao and Bogdan J Matuszewski, "Non-rigid structure from motion with diffusion maps prior," in *Computer Vision and Pattern Recognition*. IEEE, 2013, pp. 1530–1537.

[9] Yingying Zhu and Simon Lucey, "Convolutional sparse coding for trajectory reconstruction," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 37, no. 3, pp. 529–540, 2015.

[10] Xiaowei Zhou, Spyridon Leonardos, Xiaoyan Hu, and Kostas Daniilidis, "3d shape estimation from 2d landmarks: A convex relaxation approach," in *Computer Vision and Pattern Recognition*. IEEE, 2015, pp. 4447–4455.

[11] Junjie Hu and Terumasa Aoki, "A convex approach for non-rigid structure from motion via sparse representation," in *International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, (VISIGRAPP)*, 2017, pp. 333–339.

[12] Suryansh Kumar, Yuchao Dai, and Hongdong Li, "Multi-body non-rigid structure-from-motion," *arXiv preprint arXiv:1607.04515*, 2016.

[13] Enliang Zheng, Dinghuang Ji, Enrique Dunn, and Jan-Michael Frahm, "Self-expressive dictionary learning for dynamic 3d reconstruction," *arXiv preprint arXiv:1605.06863*, 2016.

[14] Stephen Boyd and Lieven Vandenberghe, *Convex optimization*, Cambridge university press, 2004.

[15] Shiqian Ma, Donald Goldfarb, and Lifeng Chen, "Fixed point and bregman iterative methods for matrix rank minimization," *Mathematical Programming*, vol. 128, no. 1-2, pp. 321–353, 2011.

[16] Peter H Schönemann, "A generalized solution of the orthogonal procrustes problem," *Psychometrika*, vol. 31, no. 1, pp. 1–10, 1966.

[17] Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh, "Reconstructing 3d human pose from 2d image landmarks," in *European Conference on Computer Vision*. Springer, 2012, pp. 573–586.

[18] Marco Paladini, Alessio Del Bue, Marko Stosic, Marija Dodig, Joao Xavier, and Lourdes Agapito, "Factorization for non-rigid and articulated structure using metric projections," in *Computer Vision and Pattern Recognition*. IEEE, 2009, pp. 2898–2905.