Jumping Manifolds: Geometry Aware Dense Non-Rigid Structure from Motion Supplementary Material

Suryansh Kumar
Australian National University, CECS, Canberra, ACT, 2601
suryansh.kumar@anu.edu.au

Abstract

This document provides mathematical derivation to the objective function proposed in the main paper [2]. Additionally, we provide some more qualitative results and statistical evaluations of our algorithm. Lastly, we made a brief comment on the challenges associated with handling temporal grassmannians for NRSfM problem.

1. Mathematical derivation to the optimization of the objective function

In this section, we provide mathematical derivation of the following optimization proposed in the paper. Some of the derivations are similar to the Kumar *et al*. work [4, 5, 7]

$$\begin{aligned}&\underset{\mathtt{Z},\tilde{\mathtt{C}},\mathtt{S},\mathtt{S}^\sharp}{\text{minimize}}\frac{1}{2}\|\mathtt{W}-\mathtt{RS}\|_\mathtt{F}^2+\beta_1\|\chi-\chi\tilde{\mathtt{C}}\|_\mathtt{F}^2+\beta_2\|\mathtt{S}^\sharp\|_*+\\&\qquad\frac{\rho}{2}\|\mathtt{S}^\sharp-f(\mathtt{S})\|_\mathtt{F}^2+<\mathbf{L_1},\mathtt{S}^\sharp-f(\mathtt{S})>+\beta_3\|\mathtt{Z}\|_*+\\&\qquad\frac{\rho}{2}\|\tilde{\mathtt{C}}-\mathtt{Z}\|_\mathtt{F}^2+<\mathbf{L_2},\tilde{\mathtt{C}}-\mathtt{Z}>\\&\text{subject to: }\xi=f_g(\mathbf{P},\mathtt{S}),\tilde{\xi}=f_h(\Delta,\xi),\\&\mathtt{S}=f_s(\xi,\Sigma,\xi_\mathtt{v}),\mathbf{P}=f_p(\tilde{\xi},\tilde{\mathtt{C}},\mathbf{P}_\circ)\end{aligned}\tag{1}$$

The constraints in the Eq:(1) are invoked over iteration. The solution to each sub-problem is obtained by taking the derivative of the above ALM form w.r.t the concerned variable and equating it to zero.

1.1. Solution to '$\mathtt{S}$'

$$\begin{aligned}&\equiv\underset{\mathtt{S}}{\text{argmin}}\frac{1}{2}\|\mathtt{W}-\mathtt{RS}\|_\mathtt{F}^2+\frac{\rho}{2}\|\mathtt{S}^\sharp-f(\mathtt{S})\|_\mathtt{F}^2+\\&\qquad<\mathbf{L_1},\mathtt{S}^\sharp-f(\mathtt{S})>\\&\equiv\underset{\mathtt{S}}{\text{argmin}}\frac{1}{2}\|\mathtt{W}-\mathtt{RS}\|_\mathtt{F}^2+\frac{\rho}{2}\|\mathtt{S}-\Big(f^{-1}(\mathtt{S}^\sharp)+\frac{f^{-1}(\mathbf{L_1})}{\rho}\Big)\|_\mathtt{F}^2\end{aligned}\tag{2}$$

Taking the derivative of the above equation w.r.t '$\mathtt{S}$' and equating it to zero gives

$$(\mathtt{R}^\mathtt{T}\mathtt{R}+\rho\mathtt{I})\mathtt{S}=\mathtt{R}^\mathtt{T}\mathtt{W}+\rho\Big(f^{-1}(\mathtt{S}^\sharp)+\frac{f^{-1}(\mathbf{L_1})}{\rho}\Big)\tag{3}$$

We used MATLAB $\mathbf{mldivide}()$ function to solve it during our implementation. You may use any linear algebra package to solve the above well known form.

1.2. Solution to '$\mathtt{S}^\sharp$'

Similar to previous derivation, we can write the ALM form for the variable $\mathtt{S}^\sharp$

$$\begin{aligned}&\equiv\underset{\mathtt{S}^\sharp}{\text{argmin}}\beta_2\|\mathtt{S}^\sharp\|_*+\frac{\rho}{2}\|\mathtt{S}^\sharp-f(\mathtt{S})\|_\mathtt{F}^2+<\mathbf{L_1},\mathtt{S}^\sharp-f(\mathtt{S})>\\&\equiv\underset{\mathtt{S}^\sharp}{\text{argmin}}\beta_2\|\mathtt{S}^\sharp\|_*+\frac{\rho}{2}\|\mathtt{S}^\sharp-\Big(f(\mathtt{S})-\frac{\mathbf{L_1}}{\rho}\Big)\|_\mathtt{F}^2\end{aligned}\tag{4}$$

The above sub-problem is well-known form for nuclear norm minimization. By defining the soft-thresholding operator $\mathcal{S}_\tau(v)=\text{sign}(v)\max(|v|-\tau)$, the solution of $\mathtt{S}^\sharp$ can be obtained by

$$\mathtt{S}^\sharp=\mathtt{U_s}\mathcal{S}_{\frac{\beta_2}{\rho}}(\Sigma_\mathtt{s})\mathtt{V_s}\tag{5}$$

where, $[\mathtt{U_s},\Sigma_\mathtt{s},\mathtt{V_s}]=\text{svd}\Big(f(\mathtt{S})-\frac{\mathbf{L_1}}{\rho}\Big)$

1.3. Solution to '$\mathtt{Z}$'

$$\begin{aligned}&\equiv\underset{\mathtt{Z}}{\text{argmin}}\beta_3\|\mathtt{Z}\|_*+\frac{\rho}{2}\|\tilde{\mathtt{C}}-\mathtt{Z}\|_\mathtt{F}^2+<\mathbf{L_2},\tilde{\mathtt{C}}-\mathtt{Z}>\\&\equiv\underset{\mathtt{Z}}{\text{argmin}}\beta_3\|\mathtt{Z}\|_*+\frac{\rho}{2}\|\mathtt{Z}-\Big(\tilde{\mathtt{C}}+\frac{\mathbf{L_2}}{\rho}\Big)\|_\mathtt{F}^2\end{aligned}\tag{6}$$

Using the soft-thresholding function as mentioned before, the solution to $\mathtt{Z}$ is given by

$$\mathtt{Z}\equiv\mathtt{U}_z\mathcal{S}_{\frac{\beta_3}{\rho}}(\Sigma_\mathtt{z})\mathtt{V_z}\tag{7}$$

where $[\mathtt{U}_z,\Sigma_\mathtt{z},\mathtt{V_z}]=\text{svd}\Big(\tilde{\mathtt{C}}+\frac{\mathbf{L_2}}{\rho}\Big)$

1.4. Solution to '$\tilde{C}$'

Deriving the solution for '$\tilde{C}$' from the sub-problem involving the variable '$\tilde{C}$' is not straight forward rather, it's a bit involved and therefore, we first derive an equivalent form for the error term that involves tensor χ. The equivalent form is easy to handle and program on computers. Lets consider the following error term:

$$\|\chi - \chi\tilde{C}\|_F^2 \tag{8}$$

Using the notation from our paper, for any i^{th} Grassmann point this error term in Eq:(8) can be written as

$$\mathrm{Tr}\left(\left((\Theta_i\Theta_i^T) - \sum_{j=1}^{K} c_{ij}(\Theta_j\Theta_j^T)\right)^T\left((\Theta_i\Theta_i^T) - \sum_{j=1}^{K} c_{ij}(\Theta_j\Theta_j^T)\right)\right) \tag{9}$$

Expanding the above form gives

$$\begin{aligned}&\equiv \mathrm{Tr}\big((\Theta_i\Theta_i^T)^T(\Theta_i\Theta_i^T)\big) - 2\sum_{j=1}^{K} c_{ij}\mathrm{Tr}\big((\Theta_i\Theta_i^T)^T(\Theta_j\Theta_j^T)\big) + \\ &\sum_{l=1}^{K}\sum_{m=1}^{K} c_{il}c_{im}\mathrm{Tr}\big((\Theta_l\Theta_l^T)^T(\Theta_m\Theta_m^T)\big)\end{aligned} \tag{10}$$

From our definition $\Theta \in \mathbb{R}^{\tilde{d}\times p}$ as an orthonormal matrix. Using it simplifies the above equation to:

$$\begin{aligned}&\equiv p - 2\sum_{j=1}^{K} c_{ij}\Gamma_{ij} + \sum_{l=1}^{K}\sum_{m=1}^{K} c_{il}c_{im}\Gamma_{lm} \\ &\text{where, } \Gamma_{ij} = \mathrm{Tr}\big((\Theta_j^T\Theta_i)(\Theta_i^T\Theta_j)\big) \quad \{\text{using trace cyclic property}\}\end{aligned} \tag{11}$$

Let $\Gamma = (\Gamma_{ij})_{ij=1}^{K} \in \mathbb{R}^{K\times K}$. Its easy to verify that Γ is symmetric positive semi-definite. Therefore, using cholesky factorization of $\mathrm{chol}(\Gamma) = LL^T$, we can re-write the above equation as

$$\begin{aligned}&\equiv p - 2\mathrm{Tr}(\tilde{C}LL^T) + \mathrm{Tr}(\tilde{C}LL^T\tilde{C}^T) \\ &\equiv \text{const} + \|L - L\tilde{C}\|_F^2 \\ &\text{where, const. means constant w.r.t } \tilde{C}\end{aligned} \tag{12}$$

By substituting the result from Eq:(12) to the sub-problem w.r.t $\tilde{C}$, we get the following form:

$$\begin{aligned}&\equiv \underset{\tilde{c}}{\mathrm{argmin}}\ \beta_1\|L - L\tilde{C}\|_F^2 + \frac{\rho}{2}\|\tilde{C} - Z\|_F^2 + < \mathbf{L_2}, \tilde{C} - Z > \\ &\equiv \underset{\tilde{c}}{\mathrm{argmin}}\ \beta_1\|L - L\tilde{C}\|_F^2 + \frac{\rho}{2}\|\tilde{C} - \left(Z - \frac{\mathbf{L_2}}{\rho}\right)\|_F^2\end{aligned} \tag{13}$$

Taking the derivative of the Eq:(13) w.r.t $\tilde{C}$ and equating it to zero.

$$(2\beta_1 LL^T + \rho I)\tilde{C} = 2\beta_1 LL^T + \rho\left(Z - \frac{\mathbf{L_2}}{\rho}\right) \tag{14}$$

2. Solution to $E(\Delta)$

$$\begin{aligned}&E(\Delta) \equiv \underset{\Delta}{\text{minimize}} \sum_{(i,j)}^{K} w_{ij}\frac{1}{2}\|\Delta^T(\Lambda_{ij})\Delta\|_F^2 \\ &\text{subject to:} \\ &\mathrm{Tr}\Big(\Delta^T\big(\sum_{i=1}^{K}\lambda_{ii}\Omega_i\Omega_i^T\big)\Delta\Big) = 1\end{aligned} \tag{15}$$

The optimization equation proposed for $E(\Delta)$ is a well-studied optimization form and Riemann Conjugate gradient toolbox can be employed to achieve the solution. Nevertheless, we can also derive augmented lagrangian form to solve the same problem. By letting $X = \big(\sum_{i=1}^{K}\lambda_{ii}\Omega_i\Omega_i^T\big)$ and expanding the Frobenius norm term, we can re-write the equation as:

$$\begin{aligned}&E(\Delta) \equiv \underset{\Delta}{\text{minimize}} \sum_{(i,j)}^{K} \frac{w_{ij}}{2}\mathrm{Tr}\big(\Delta^T\Lambda_{ij}\Delta\Delta^T\Lambda_{ij}\Delta\big) \\ &E(\Delta) \equiv \underset{\Delta}{\text{minimize}}\ \mathrm{Tr}\Big(\Delta^T\sum_{(i,j)}^{K}\frac{w_{ij}}{2}\Lambda_{ij}\Delta^{t-1}\Delta^{(t-1)T}\Lambda_{ij}\Delta\Big) \\ &\text{subject to:} \\ &\mathrm{Tr}\Big(\Delta^T X\Delta\Big) = 1\end{aligned} \tag{16}$$

Here, $t-1$ refers to its known value before the current iteration. Now, by assuming $Y = \frac{w_{ij}}{2}\Lambda_{ij}\Delta^{t-1}\Delta^{(t-1)T}\Lambda_{ij}$, the above equation simplifies to standard eigen value decomposition problem *i.e.*

$$\begin{aligned}&E(\Delta) \equiv \underset{\Delta}{\text{minimize}}\ \mathrm{Tr}(\Delta^T Y\Delta) \\ &\text{subject to:} \\ &\mathrm{Tr}\Big(\Delta^T X\Delta\Big) = 1\end{aligned} \tag{17}$$

The equivalent Lagrangian function form is given by

$$\mathrm{Tr}(\Delta^T Y\Delta) + \lambda\Big(1 - \mathrm{Tr}\big(\Delta^T X\Delta\big)\Big) \tag{18}$$

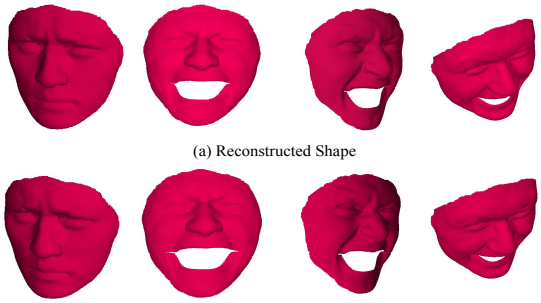
The Eq:(18) is of the standard form to generalized eigen value problem. You may use any standard linear algebra package to solve it.

3. More Results and Analysis

We provide some more qualitative results on Garg *et al*. benchmark dataset [1]. Figure (1) and Figure (2) show the 3D reconstruction results of our algorithm on real face and synthetic face sequence. Next, we provide one more statistical analysis of the our algorithm.

1. Dependence of the algorithm on variable $\tilde{d}$: While reducing the dimension for grouping the grassmann points, one of the critical aspect is to determine the dimension to which we should project for better results. We used well-known procedure of cumulative energy of eigen vectors to get the value of $\tilde{d}$. Mathematically, let Ω be the set that stack all the Grassmannians and σ_i

Figure 1: 3D reconstruction results on real face sequence [1]



(a) Reconstructed Shape

(b) Ground-Truth Shape

Figure 2: 3D reconstruction results on synthetic face sequence [1].

be the $\mathtt{i}^{\text{th}}$ singular value of $\Omega\Omega^{\mathtt{T}}$, then

$$\tilde{\mathtt{d}} = \underset{\mathtt{d}_{\text{opt}}}{\text{argmin}} \frac{\sum_{\mathtt{i}=1}^{\mathtt{d}_{\text{opt}}} \sigma_{\mathtt{i}}}{\sum_{\mathtt{i}=1}^{\mathtt{d}} \sigma_{\mathtt{i}}} \geq \tau \tag{19}$$

where τ can vary from 0 to 1 and $\mathtt{d}_{\text{opt}}$ (optimal dimension) is a positive integer. We put $\tau = 0.97$ for all our experiment. Figure (3) show the variations in the reconstruction error with the value of τ. It is observed that for different dataset the value of suitable $\tilde{\mathtt{d}}$ is different. The point to note is that if the reduced dimension is less than the intrinsic dimension, the samples may lose important information for better grouping of Grassmannians.

3.1. Why we opt not to distrub the temporal continuity for this problem?

Although clustering of frames into smaller groups (Grassmannians) allows simpler model and be handy if prior informations about shapes/activities are available. However, its quite possible that there will be repeat of certain activities or expression in the video sequence (say facial expression). In such cases the Grassmannians at frame 'f' and frame 'f+n' will be assigned to same group ('n' is the time instant at which activities repeat or is similar). As a result, such representation procedure may disturb the overall time continuity of the sequence. Also, these group of frames may form high-dimensional grassmannians, in order to project it into low-dimension using neighboring Grassmannians will get extremely difficult, for example, how to decide neighboring grassmannians using temporal grassmann samples?. On the other hand, grouping of trajectories (spatial) does not disturb the temporal continuity of the trajectory and we can easily define the neighbors using spatial information *i.e.*, spatial neighbors tend to be neighbors throughout the sequence, for a single deforming object (unless breaks or disassociate, which is very rare). But in shape space, we don't have any prior knowledge to define neighboring relation.

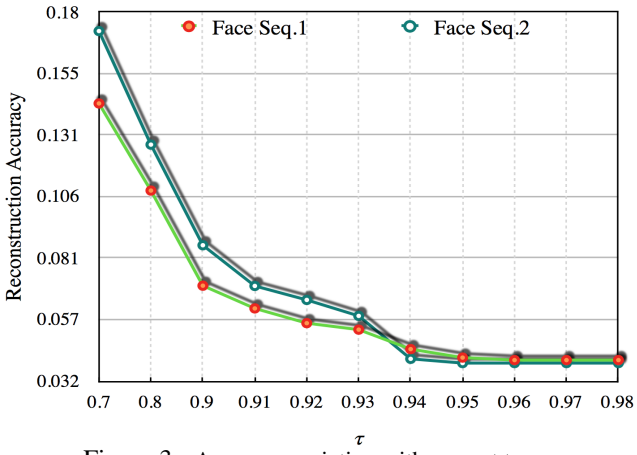


Figure 3: Accuracy variation with respect to τ.

3.2. What is gained with the added extra complexity in the algorithm (Ablation Test)?

By adding Eq:(7) we can have more discriminative Grassmannian representation which is useful in practice. The performance on noisy sequence clearly demonstrates this. Also, by adding this local representation constraint we have better control over local shapes with very minimal increment in the processing time.

Data	WO_NC (Noiseless)	WO_NC (Noise5%)	W_NC (Noiseless)	W_NC (Noise5%)
Seq.1	0.1083	0.1592	0.0404	0.0482
Seq.2	0.0972	0.1491	0.0392	0.0429
Seq.3	0.0913	0.1354	0.0280	0.0365
Seq.4	0.0924	0.1433	0.0327	0.0402

Table 1: Ablation test showing the (e_{3D}) comparison on dense synthetic face sequence. WO_NC means WithOut Neighboring Constraint and W_NC means With Neighboring Constraints.

3.3. Mean processing time per frame?

The mean processing time per frame is 10.57743s in comparison to [4] which is 9.53093s.

3.4. What we gained over [4]?

Temporal clustering does help improve reconstruction accuracy as shown in [4] but [4] algorithm does it by assuming that they know how many frames to select to represent a local temporal Grassmannian. Now such information is not available in practice. Our new algorithm show that without such temporal knowledge we can have an equivalent or better performance by using a better

representation of the grassmannians. This can be verified using the reconstruction accuracy table in the main paper.

References

[1] Ravi Garg, Anastasios Roussos, and Lourdes Agapito. Dense variational reconstruction of non-rigid surfaces from monocular video. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1272–1279, 2013.

[2] Suryansh Kumar. Jumping manifolds: Geometry aware dense non-rigid structure from motion. *arXiv preprint arXiv:1902.01077*, 2019.

[3] Suryansh Kumar. Non-rigid structure from motion: Prior-free factorization method revisited. *arXiv preprint arXiv:1902.10274*, 2019.

[4] Suryansh Kumar, Anoop Cherian, Yuchao Dai, and Hongdong Li. Scalable dense non-rigid structure-from-motion: A grassmannian perspective. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[5] Suryansh Kumar, Yuchao Dai, and Hongdong Li. Multi-body non-rigid structure-from-motion. In *3D Vision (3DV), 2016 Fourth International Conference on*, pages 148–156. IEEE, 2016.

[6] Suryansh Kumar, Yuchao Dai, and Hongdong Li. Monocular dense 3d reconstruction of a complex dynamic scene from two perspective frames. In *IEEE International Conference on Computer Vision (ICCV)*, pages 4649–4657, Oct 2017.

[7] Suryansh Kumar, Yuchao Dai, and Hongdong Li. Spatio-temporal union of subspaces for multi-body non-rigid structure-from-motion. *Pattern Recognition*, 71:428–443, May 2017.

[8] Suryansh Kumar, Ram Srivatsav Ghorakavi, Yuchao Dai, and Hongdong Li. Dense depth estimation of a complex dynamic scene without explicit 3d motion estimation. *arXiv preprint arXiv:1902.03791*, 2019.