SEGMENTATION AND 3D RECONSTRUCTION OF NON-RIGID SHAPE FROM RGB VIDEO

Antonio Agudo

Institut de Robòtica i Informàtica Industrial, CSIC-UPC, 08028, Barcelona, Spain

ABSTRACT

In this paper we propose a unsupervised and unified approach to simultaneously recover time-varying 3D shape, camera motion, and temporal clustering into deformations, all of them, from partial 2D point tracks in a RGB video and without assuming any pre-trained model. As the data are drawn from a sequentially ordered images, we fully exploit this information to constrain all model parameters we estimate. We present an energy-based formulation that is efficiently solved and allows to estimate all model parameters in the same loop via augmented Lagrange multipliers in polynomial time, enforcing similarities between images at any level. Validation is done in a wide variety of human video sequences, including articulated and continuous motion, and for dense and missing tracks. Our approach is shown to outperform state-of-the-art solutions in terms of 3D reconstruction and clustering.

Index Terms— Non-Rigid Structure from Motion, Deformation Segmentation, Sequential Data, Optimization

1. INTRODUCTION

RGB videos are nowadays present in everyone's life thanks to the rapid development of recording cameras. In the last years, many solutions have been proposed to perceive the world in 3D without accounting for any extra sensor. While some degree of success has been achieved when the structure observed by the sensor is rigid [1, 2], recovering the 3D geometry of the vivid moving real world is still in its infancy. In this case, the fact that many different 3D structure configurations may have similar 2D projections produces severe ambiguities that can be only resolved by incorporating more sophisticated constraints than those utilized in the rigid case. In the community, this problem is denominated as Non-Rigid Structure from Motion (NRSfM), and consists in estimating motion and non-rigid 3D shape from 2D point tracks in a monocular video without the need for a pre-trained model. The results can be exploited in many application domains, including augmented reality, medical image, multimedia, or in human-computer interaction to name just a few.

The simultaneous recovery of non-rigid 3D shape and pose parameters usually results in a non-convex optimization problem, that in combination with the orthogonality constraints on the camera parameters, make the problem even more complicated. Maybe, the most popular priors are based on low-rank constraints over different modalities [3, 4, 5, 6, 7, 8, 9, 10] that induce the deformations. However, these approaches rarely exploit the full similarities that the sequential data can provide. In this work, we propose to fully exploit this physical temporal coherence, by implicitly enforcing smooth motion and deformation, smooth temporal similarities to infer the segmentation, and smooth 2D projections to complete missing entries; all of them, in combination with low-rank priors. Thanks to our formulation, we provide smooth patterns that produce more accurate segmentations and 3D reconstructions than state-of-the-art approaches.

2. RELATED WORK

Retrieving a non-rigid 3D structure together with the camera motion from solely the observation of 2D point tracks in a monocular video is an ill-posed problem that requires to exploit the art of priors. The most used idea to address the problem is to assume that the 3D shape lies in a low-rank subspace defined by shape [5, 11], trajectory [6, 7], shape-trajectory [12, 13, 14], or force [8] vectors. The main limitation of previous approaches is the dimensionality of the subspace is known in advance, making them very problem specific. Later, other approaches have imposed the low-rank constraint by directly minimizing the rank of a matrix representing the 3D shape, considering the data lie in a single [15, 16], in a union of temporal [9, 17], or in a dual union of spatio-temporal [18] subspaces. In combination with previous approaches, smoothness constraints have also been incorporated to provide robustness [3, 4, 5, 19, 20]. Unfortunately, the temporal coherence in video data has not been fully exploited in previous formulations. In this paper, we introduce temporal consistency in all model parameters we recover, providing clean, robust and accurate estimations that outperform state-of-the-art solutions. To achieve that, we present a novel unsupervised formulation where all model parameters are estimated in the same loop, penalizing deviations on consecutive frames by means of the extensive use of smoothness filters.

This work has been partially supported by the CSIC project R3OBJ 201850I099, the Spanish Ministry of Science and Innovation under project HuMoUR TIN2017-90086-R, and by the Spanish State Research Agency through the María de Maeztu Seal of Excellence to IRI MDM-2016-0656.

3. NON-RIGID STRUCTURE FROM MOTION

We now review the NRSfM formulation that will be later employed to introduce our approach. To this end, let us consider a set of N 3D points observed along F pictures represented by $\mathbf{x}_n^f = [x_n^f, y_n^f, z_n^f]^\top$ for the n-th point at frame f. Considering that the point is observed by an orthographic camera, its 2D projection in the f-th image plane can be denoted as $\mathbf{p}_n^f = [u_n^f, v_n^f]^\top$. After collecting all points in all images, the 3D-to-2D projection system can be defined as:

$$\underbrace{\begin{bmatrix} \mathbf{p}_1^1 & \cdots & \mathbf{p}_N^1 \\ \vdots & \ddots & \vdots \\ \mathbf{p}_1^F & \cdots & \mathbf{p}_N^F \end{bmatrix}}_{\mathbf{P}\in\mathbb{R}^{2F\times N}} = \underbrace{\begin{bmatrix} \mathbf{R}^1 & & \\ & \ddots & \\ & & \mathbf{R}^F \end{bmatrix}}_{\mathbf{G}\in\mathbb{R}^{2F\times 3F}} \underbrace{\begin{bmatrix} \mathbf{x}_1^1 & \cdots & \mathbf{x}_N^1 \\ \vdots & \ddots & \vdots \\ \mathbf{x}_1^F & \cdots & \mathbf{x}_N^F \end{bmatrix}}_{\mathbf{X}\in\mathbb{R}^{3F\times N}},$$

where $\mathbf{P}$ is a measurement matrix to group the 2D point tracks, $\mathbf{G}$ is a block diagonal matrix, made of the F truncated 2×3 camera rotations $\mathbf{R}^f$, and $\mathbf{X}$ is a shape matrix with the 3D point locations. It is worth noting that the previous expression is obtained after removing zero-mean measurements, i.e., the 2D translation. The NRSfM problem consists in factoring the measurement matrix $\mathbf{P}$ into the motion $\mathbf{G}$ and shape $\mathbf{X}$ factors, i.e., inferring camera pose and 3D reconstruction from 2D point tracks in a monocular video.

In order to address the problem, the most standard approach is to enforce a low-rank constraint over $\mathbf{P}$, since the time-varying configurations in $\mathbf{X}$ can lie in a linear subspace of rank $3K$ [21]. However, it was observed that another interpretation can be considered, removing the use of unnecessary degrees of freedom [15, 16, 22]. To achieve that, we can define the matrix $\hat{\mathbf{X}}$ that rearranges the entries of $\mathbf{X}$ into a new $3N \times F$ matrix, where a K-rank constraint could be imposed. Both matrices can be related by means of $\mathbf{X} = (\mathbf{I}_3 \otimes \hat{\mathbf{X}}^\top)\mathbf{A}$ and $\hat{\mathbf{X}} = (\mathbf{X}^\top \otimes \mathbf{I}_3)\mathbf{B}$, where $\otimes$ is a Kronecker product operator, $\mathbf{I}_3$ is an identity matrix, and $\mathbf{A}$ and $\mathbf{B}$ are binary matrices of size $9N \times N$ and $9F \times F$, respectively. Similarly, we can define the matrix $\hat{\mathbf{P}}$ that rearranges the entries of $\mathbf{P}$ into a $2N \times F$ matrix, and the relations $\mathbf{P} = (\mathbf{I}_2 \otimes \hat{\mathbf{P}}^\top)\mathbf{C}$ and $\hat{\mathbf{P}} = (\mathbf{P}^\top \otimes \mathbf{I}_2)\mathbf{D}$ with $\mathbf{C}$ and $\mathbf{D}$ other binary matrices. As we will see later, both interpretations are used in our formulation.

4. ART OF PRIORS FOR SEQUENTIAL DATA

As it was introduced in the last section, our input data consist in a RGB video, i.e., our input information follows a sequential pattern along the time, where most pictures are similar to their neighbors. Thanks to this observation, we can exploit the natural consistency of sequential data by incorporating some penalty terms into our model. Without loss of generality, we will use this type of penalties to enforce smooth relations in all model parameters we consider.

From a physical perspective, the motion of a camera should be faster than the deformation of a dynamic object. Considering that, the level of temporal regularization needs to be stronger in the shape parameters than in the camera ones. However, this does not mean that only the 3D reconstruction should be regularized, since its 2D projection is also related according to the projection equation. Taking inspiration in the theory of finite differences, for a sudden motion like that followed by a camera, we can introduce a first-order filter of the type $\mathbf{R}^{f+1} \approx \mathbf{R}^f$ to enforce smooth movements, i.e., the location of the f-th camera in two neighboring pictures does not change much. In contrast, for shape deformations we need even more regularization, extending the influence of the neighborhood in a temporal domain. This can be done by using, for instance, a fourth-order central difference, where five neighbors are considered for regularization as:

$$\frac{\partial^2 \mathbf{x}}{\partial t^2} = \frac{-\mathbf{x}^{i-2} + 16\mathbf{x}^{i-1} - 30\mathbf{x}^i + 16\mathbf{x}^{i+1} - \mathbf{x}^{i+2}}{12\Delta t^2}, \qquad (1)$$

where Δt denotes the temporal increase between frames, and the variable $\mathbf{x}$ represents a generic 3D point. Omitting the term Δt, the previous equation can be represented by means of a compact form by using a $F \times F$ matrix $\mathbf{F}$ as:

$$\mathbf{F}_{kj} = \begin{cases} -30 & \text{if } j = k, k = \{3, \ldots, F-2\} \\ -16 & \text{if } j = k, k = \{2, F-1\} \\ 16 & \text{if } j|k = k|j+1, k|j = \{2, \ldots, F-2\} \\ -1 & \text{if } j|k = k|j-2, k|j = \{3, \ldots, F\} \\ 1 & \text{if } j|k = k|j-1, k|j = \{2, F\} \\ 0 & \text{if otherwise} \end{cases},$$

where first- and third-order filters are included in the boundaries to achieve a smooth transition. It is worth pointing out that this matrix is highly sparse, allowing to enforce the temporal filter at low computational cost.

5. MOTION, 3D TIME-VARYING SHAPE AND CLUSTERING FROM 2D POINT TRACKS

Our goal is to simultaneously retrieve camera motion, 3D time-varying shape and deformation clustering from incomplete 2D point tracks in a monocular video. In this section, we formulate the full problem by considering the preliminary concepts defined in sections 3 and 4. We also present a unsupervised, efficient and unified optimization strategy to sort it out, without assuming any training data at all.

5.1. Problem Statement

To solve for motion, 3D shape and deformation clustering from incomplete sequential data, we assume the shape deformation can be modeled by means of a union of temporal subspaces. To do that, we consider a low-rank $F \times F$ similarity matrix $\mathbf{T}$ to encode higher entries for pairs of pictures of the same cluster, such that $\hat{\mathbf{X}} = \hat{\mathbf{X}}\mathbf{T} + \mathbf{N}$, where $\mathbf{N}$ is a $3N \times F$ residual noise. As it was introduced in section 3, $\hat{\mathbf{X}}$ is also low-rank, and therefore, both matrices can be computed by minimizing their rank without assuming any prior information. Since this is a non-convex NP-hard problem we used the nuclear norm instead, which is its convex relaxation [23]. Additionally, we exploit the sequential-data prior by adding extra constraints by means of the matrix $\mathbf{F}$ we introduced in section 4. This can be done enforcing a couple of hard constraints over the non-rigid 3D shape $\hat{\mathbf{X}}\mathbf{F} = \mathbf{0}$ (as in [9]) and over the 2D point tracks $\hat{\mathbf{P}}\mathbf{F} = \mathbf{0}$, respectively. Moreover, we also enforce this constraint in temporal similarities, using the term $\mathbf{TF}$, forcing consecutive columns of $\mathbf{T}$ to be similar.

With these ingredients, we denote the set of all model parameters to be recovered by $\mathbf{\Psi} \equiv \{\mathbf{P}, \hat{\mathbf{P}}, \mathbf{G}, \mathbf{X}, \hat{\mathbf{X}}, \mathbf{T}, \mathbf{N}\}$. Our input data consists of partial 2D point tracks in a RGB video $\bar{\mathbf{P}}$, and the corresponding observability matrix $\mathbf{O} \in \mathbb{R}^{F\times N}$, with $\{1, 0\}$ entries indicating whether a point in a specific frame is visible or not. Taking into account the orthonormality constraints on camera rotations, our problem is:

$$\begin{aligned}\arg\min_{\mathbf{\Psi}} \;& \|\left(\mathbf{O}\otimes\mathbf{1}_2\right)\odot\left(\mathbf{P}-\bar{\mathbf{P}}\right)\|_F^2+\beta\|\mathbf{P}\|_*+\zeta q(\mathbf{G}) \\ & +\gamma(\|\hat{\mathbf{X}}\|_*+\|\mathbf{T}\|_*)+\lambda(\|\mathbf{TF}\|_1+\|\mathbf{N}\|_{2,1}) \quad (2)\end{aligned}$$

$$\text{subject to}\quad \begin{aligned} &\mathbf{P}=\mathbf{GX}\\ &\mathbf{GG}^\top=\mathbf{I}_{2F}\\ &\hat{\mathbf{X}}=\hat{\mathbf{X}}\mathbf{T}+\mathbf{N}\\ &(\mathbf{I}_3\otimes\hat{\mathbf{X}}^\top)\mathbf{A}=\mathbf{X}\\ &(\mathbf{I}_2\otimes\hat{\mathbf{P}}^\top)\mathbf{C}=\mathbf{P}\\ &\hat{\mathbf{X}}\mathbf{F}=\mathbf{0}\\ &\hat{\mathbf{P}}\mathbf{F}=\mathbf{0}\end{aligned}$$

where $\mathbf{1}$ denotes a vector of ones, and $\odot$ represents a Hadamard product. $\|\cdot\|_*$, $\|\cdot\|_1$ and $\|\cdot\|_{2,1}$ indicate the nuclear norm, l_1-norm and $l_{2,1}$-norm, respectively, and $\|\cdot\|_F$ is the Frobenius norm. $\{\beta, \zeta, \gamma, \lambda\}$ are penalty coefficients. In Eq. (2), we denote by $q(\cdot)$ the function to impose smooth solutions on the camera rotation, as it was commented in section 4.

5.2. Optimization

The problem in Eq. (2) is non-convex, and it can be approximated by a two-step strategy that will be iteratively combined to achieve a solution. On the one hand, we have to compute the camera rotation $\mathbf{G}$ by considering the terms in Eq. (2) where that variable is implicated, writing the problem as:

$$\arg\min_{\mathbf{R}^f\in SO(3)} \frac{1}{2}\sum_{f=1}^{F}\sum_{p=1}^{P}\|\mathbf{p}_n^f-\mathbf{R}^f\mathbf{x}_n^f\|_F^2+\zeta\sum_{f=1}^{F-1}\|\nabla^f\mathbf{R}\|_{\mathcal{F}}^2, \quad (3)$$

where every $\mathbf{R}^f$ matrix lies in the $SO(3)$ manifold (see section 3). This problem can be solved by using the trust-region solver in the Manopt toolbox [24].

On the other hand, we need to solve for the rest of model parameters. To this end, we present the energy to be minimized by considering all model parameters except the camera rotation. Applying Augmented Lagrange Multipliers (ALM), the equivalent Lagrangian function is:

$$\begin{aligned}\arg\min_{\bar{\mathbf{\Psi}}} \;& \|\left(\mathbf{O}\otimes\mathbf{1}_2\right)\odot\left(\mathbf{P}-\bar{\mathbf{P}}\right)\|_F^2+\beta\|\mathbf{P}\|_* \quad (4)\\ & +\gamma(\|\hat{\mathbf{X}}\|_*+\|\mathbf{J}\|_*)+\lambda(\|\mathbf{U}\|_1+\|\mathbf{N}\|_{2,1})\\ & +\langle\mathbf{M}_1,\mathbf{P}-\mathbf{GX}\rangle+\frac{\alpha}{2}\|\mathbf{P}-\mathbf{GX}\|_F^2\\ & +\langle\mathbf{M}_2,\hat{\mathbf{X}}-\hat{\mathbf{X}}\mathbf{T}-\mathbf{N}\rangle+\frac{\alpha}{2}\|\hat{\mathbf{X}}-\hat{\mathbf{X}}\mathbf{T}-\mathbf{N}\|_F^2\\ & +\langle\mathbf{M}_3,(\mathbf{I}_3\otimes\hat{\mathbf{X}}^\top)\mathbf{A}-\mathbf{X}\rangle+\frac{\alpha}{2}\|(\mathbf{I}_3\otimes\hat{\mathbf{X}}^\top)\mathbf{A}-\mathbf{X}\|_F^2\\ & +\langle\mathbf{M}_4,(\mathbf{I}_2\otimes\hat{\mathbf{P}}^\top)\mathbf{C}-\mathbf{P}\rangle+\frac{\alpha}{2}\|(\mathbf{I}_2\otimes\hat{\mathbf{P}}^\top)\mathbf{C}-\mathbf{P}\|_F^2\\ & +\langle\mathbf{M}_5,\hat{\mathbf{X}}\mathbf{F}\rangle+\frac{\alpha}{2}\|\hat{\mathbf{X}}\mathbf{F}\|_F^2+\langle\mathbf{M}_6,\hat{\mathbf{P}}\mathbf{F}\rangle\\ & +\frac{\alpha}{2}\|\hat{\mathbf{P}}\mathbf{F}\|_F^2+\langle\mathbf{M}_7,\mathbf{U}-\mathbf{TF}\rangle+\frac{\alpha}{2}\|\mathbf{U}-\mathbf{TF}\|_F^2\\ & +\langle\mathbf{M}_8,\mathbf{T}-\mathbf{J}\rangle+\frac{\alpha}{2}\|\mathbf{T}-\mathbf{J}\|_F^2,\end{aligned}$$

where two dual variables $\mathbf{U}$ and $\mathbf{J}$ are included, and $\bar{\mathbf{\Psi}} \equiv \mathbf{\Psi}\cup\{\mathbf{U},\mathbf{J}\}\setminus\mathbf{G}$. In addition, we also introduce the Lagrange multipliers: $\{\mathbf{M}_1, \mathbf{M}_4\} \in \mathbb{R}^{2F\times N}$, $\{\mathbf{M}_2, \mathbf{M}_5\} \in \mathbb{R}^{3N\times F}$, $\mathbf{M}_3 \in \mathbb{R}^{3F\times N}$, $\mathbf{M}_6 \in \mathbb{R}^{2N\times F}$, and $\{\mathbf{M}_7, \mathbf{M}_8\} \in \mathbb{R}^{F\times F}$; and $\alpha > 0$ is a penalty weight to improve convergence. The problem in Eq. (4) can be efficiently sorted out by recovering each model parameter independently and in closed form while keeping fixed the rest of parameters, as it is proposed in [25, 26]. For initialization, we follow the first two steps proposed in [18]. That is, we first assume a low-rank prior to solve a matrix-completion problem to retrieve $\mathbf{P}$ from incomplete data $\bar{\mathbf{P}}$, and then computing the rotation $\mathbf{G}$. For simplicity, no temporal regularizations are enforced in these steps. After that, we alternatively solve Eqs. (3)- (4) until convergence.

6. EXPERIMENTAL EVALUATION

We now report our experimental evaluation on several human motion videos, including articulated and continuous deformation, several body configurations and scenarios with missing or dense entries. For quantitative evaluation, we apply our algorithm on the articulated human motion dataset introduced in [6], which includes five types of activities. As in the literature [7, 13, 15], we will provide the normalized mean 3D error e_S, and the mean rotation error e_R. For further details, we refer the reader to these papers. Additionally, we also report the object clustering error e_C as defined in [18], after applying spectral clustering [27] over the estimated matrix $\mathbf{T}$.

To establish a comparison, we consider eight state-of-the-art methods: EM-PPCA [5], MP [11], PTA [6], CSF [13], KSTA [12], BMM [15], PPTA [7], and URS [9]; under two situations: noise-free and noisy 2D point tracks as it was done in [7]. We do not consider modern unsupervised deep-learning approaches [28] as they require large amounts of training data to obtain competitive solutions [29]. It is worth

Data \ Met.	EM-PPCA [5]		MP [11]		PTA [6]		CSF [13]		KSTA [12]		BMM [15]		PPTA [7]		URS [9]			(Ours)		
	e_R	$e_S(R)$	e_R	$e_S(R)$	e_R	$e_S(R)$	e_R	$e_S(R)$	e_R	$e_S(R)$	e_R	$e_S(R)$	e_R	$e_S(R)$	e_R	e_S	$e_C[\%]$	e_R	e_S	$e_C[\%]$
Noise-free observations																				
Drink	.186	.261(7)	.330	.357(12)	.006	.025(13)	.006	.022(6)	.006	.020(12)	.007	.027(12)	.006	.011(30)	.006	**.009**	0.8(2)	.006	**.009**	0.6(2)
Stretch	.749	.458(7)	.832	.900(8)	.055	.109(12)	.049	.071(8)	.049	.064(11)	.068	.103(11)	.058	.084(11)	.058	.061	4.1(3)	.058	**.060**	4.1(3)
Yoga	.688	.445(8)	.854	.786(2)	.106	.163(11)	.102	.147(7)	.102	.148(7)	.088	**.115(10)**	.106	.158(11)	.106	.143	0.3(2)	.091	.133	0.2(2)
Pick-up	.417	.423(14)	.249	.429(5)	.155	.237(12)	.155	.230(6)	.155	.233(6)	.121	**.173(12)**	.154	.235(12)	.154	.221	3.7(3)	.147	.209	3.0(3)
Dance	–	.339(4)	–	.271(5)	–	.296(5)	–	.271(2)	–	.249(4)	–	.188(10)	–	.229(4)	–	.165	–	–	**.150**	–
Average error:		.385		.549		.166		.148		.143		.121		.143		.119			**.112**	
Relative error:		3.44		4.90		1.48		1.32		1.28		1.08		1.28		1.06			1.00	
Noisy observations																				
Drink	.231	.250(7)	.329	.517(12)	.043	.045(13)	.043	.044(6)	.043	.042(12)	.044	.056(12)	.042	.038(30)	.042	.044	3.6(2)	.036	**.034**	1.4(2)
Stretch	.819	.886(7)	.872	.975(8)	.091	.144(12)	.091	.121(8)	.091	.166(11)	.098	.183(11)	.091	.123(11)	.091	**.119**	8.4(3)	.091	**.119**	5.1(3)
Yoga	.700	.507(8)	.858	.791(2)	.124	.174(11)	.125	.168(7)	.125	.172(7)	.136	.195(10)	.124	.174(11)	.125	.167	0.0(2)	.112	**.162**	0.2(2)
Pick-up	.499	.807(14)	.250	.407(5)	.148	.228(12)	.148	.224(6)	.148	.222(6)	.141	.212(12)	.148	.228(12)	.148	.207	3.1(3)	.147	**.205**	2.5(3)
Dance	–	.336(4)	–	.282(5)	–	.299(5)	–	.266(2)	–	.248(4)	–	.236(10)	–	.222(4)	–	.164	–	–	**.157**	–
Average error:		.557		.594		.178		.165		.170		.176		.157		.140			**.135**	
Relative error:		4.12		4.40		1.32		1.22		1.26		1.30		1.16		1.04			1.00	

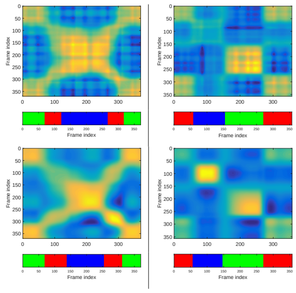

Table 1. Quantitative and qualitative evaluation on human motion capture videos. Left: We provide rotation e_R and 3D reconstruction e_S errors for competing techniques: EM-PPCA [5], MP [11], PTA [6], CSF [13], KSTA [12], BMM [15], PPTA [7], and URS [9]; and for our approach, considering both noise-free and noisy observations. For every solution, we also indicate in parentheses the rank K of the linear subspace that produced the lowest e_S error. Relative error is always computed with respect to our reconstruction. For ours, we also provide clustering error $e_C[\%]$, and the number of motion clusters in parentheses. The symbol "$-$" denotes that ground truth data is not available. **Right:** Our estimated similarity matrix $\mathbf{T}$ (top) and the ground truth (bottom), together with the associated clustering bar, for the sequences *Stretch* (left) and *Pickup* (right).

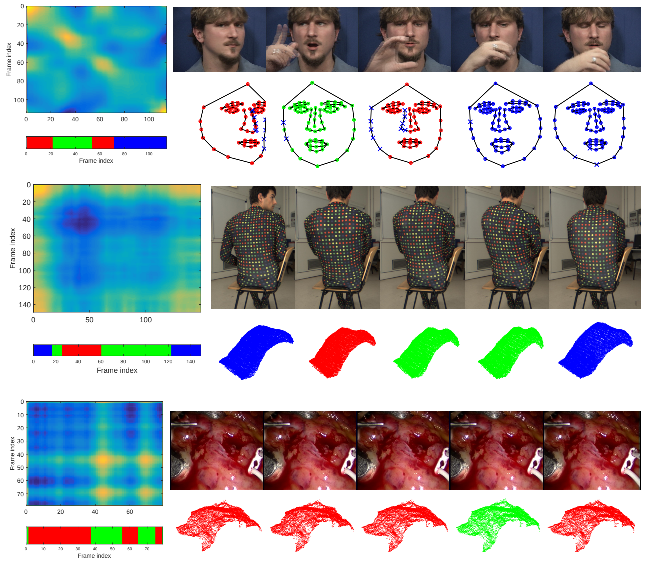

Fig. 1. Qualitative evaluation on real videos. In both cases, we display the same information, for (from top to bottom) *Face*, *Back* and *Heart* sequences. **Left:** Deformation similarity matrix we recover, and the corresponding clustering bar. **Right:** Images together with its 3D reconstruction using another point of view. Every color corresponds to a deformation cluster. Blue crosses represent missing points.

noting that for some methods, the subspace rank K needs to be tuned by hand, rather than our approach that does not require tuning any rank. Table 1-left summarizes the 3D reconstruction and rotation errors for all methods, sequences, and situations. Note that our method outperforms consistently the state of the art in terms of 3D reconstruction, reducing the 3D error of other approaches by large margins between the 6% and 490% for noise-free, and between the 4% and 440% for noisy observations, respectively. As our approach, is the only one that estimates all model parameters in the same loop, may become less robust to artifacts than the rest, but though, this is a key factor to achieve more accurate solutions. Regarding segmentation, our method provides more smooth and clean similarity matrices than [9], producing better segmentations. In table 1-right, we display a qualitative comparison between our similarity matrix $\mathbf{T}$ and the ground truth, along with the clustering bars. As it can be seen, though our estimation is somewhat noisy, the clustering we obtain is quite accurate.

We also show the robustness of our algorithm against occlusions, by processing an American-sign-language sequence, where a human face is moving and gesturing [8]. Figure 1-top shows some images and our 3D reconstruction even for missing point tracks, as well as the deformation similarity and its clustering bar (we detect three clusters: closed mouth, and open mouth with closed and open eyes). Finally, we also validate our method on dense data by running two video sequences with 20,561 and 68,295 2D point tracks taken from [16], where a back and a heart are moving and deforming, respectively. In Fig. 1-middle/bottom is displayed the 3D reconstruction we obtain for some pictures, along with the estimated similarities and clusters. In spite of being only qualitative, our 3D reconstruction seems to be very accurate and coherent with the deformation clusters. Again, our algorithm obtains cleaner similarity matrices than [9], producing better temporal-consistency segmentations.

7. CONCLUSION

We have proposed a novel formulation to solve the Clustering-NRSfM problem in a unified, efficient and unsupervised fashion. To do that, we have introduced an energy-based formulation that can be minimized in the same loop, where the sequential nature in video data is completely exploit to infer all model parameters we consider. Experimental results show our solution provides more accurate solutions than the rest of competing methods to retrieve human motion in terms of 3D reconstruction and clustering, being applicable in situations with dense and missing tracks. Our future work is oriented to extend our model to full perspective cameras.

8. REFERENCES

[1] J. Lim, J.M. Frahm, and M. Pollefeys, "Online environment mapping," in *CVPR*, 2011.

[2] R. Newcome and A. J. Davison, "Live dense reconstruction with a single moving camera," in *CVPR*, 2010.

[3] A. Bartoli, V. Gay-Bellile, U. Castellani, J. Peyras, S. Olsen, and P. Sayd, "Coarse-to-fine low-rank structure-from-motion," in *CVPR*, 2008.

[4] M. Lee, J. Cho, C. H. Choi, and S. Oh, "Procrustean normal distribution for non-rigid structure from motion," in *CVPR*, 2013.

[5] L. Torresani, A. Hertzmann, and C. Bregler, "Nonrigid structure-from-motion: estimating shape and motion with hierarchical priors," *TPAMI*, vol. 30, no. 5, pp. 878–892, 2008.

[6] I. Akhter, Y. Sheikh, S. Khan, and T. Kanade, "Nonrigid structure from motion in trajectory space," in *NIPS*, 2008.

[7] A. Agudo and F. Moreno-Noguer, "A scalable, efficient, and accurate solution to non-rigid structure from motion," *CVIU*, vol. 167, no. 2, pp. 121–133, 2018.

[8] A. Agudo and F. Moreno-Noguer, "Force-based representation for non-rigid shape and elastic model estimation," *TPAMI*, vol. 40, no. 9, pp. 2137–2150, 2018.

[9] A. Agudo and F. Moreno-Noguer, "Deformable motion 3D reconstruction by union of regularized subspaces," in *ICIP*, 2018.

[10] S. Kumar, Y. Dai, and H. Li, "Spatio-temporal union of subspaces for multi-body non-rigid structure-from-motion," *PR*, vol. 77, no. 11, pp. 428–443, 2017.

[11] 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 *CVPR*, 2009.

[12] P. F. U. Gotardo and A. M. Martinez, "Kernel non-rigid structure from motion," in *ICCV*, 2011.

[13] P. F. U. Gotardo and A. M. Martinez, "Computing smooth time-trajectories for camera and deformable shape in structure from motion with occlusion," *TPAMI*, vol. 33, no. 10, pp. 2051–2065, 2011.

[14] X. Xu and E. Dunn, "Discrete Laplace operator estimation for dynamic 3D reconstruction," in *ICCV*, 2019.

[15] Y. Dai, H. Li, and M. He, "A simple prior-free method for non-rigid structure from motion factorization," in *CVPR*, 2012.

[16] R. Garg, A. Roussos, and L. Agapito, "Dense variational reconstruction of non-rigid surfaces from monocular video," in *CVPR*, 2013.

[17] Y. Zhu, D. Huang, F. de la Torre, and S. Lucey, "Complex non-rigid motion 3D reconstruction by union of subspaces," in *CVPR*, 2014.

[18] A. Agudo and F. Moreno-Noguer, "DUST: Dual union of spatio-temporal subspaces for monocular multiple object 3D reconstruction," in *CVPR*, 2017.

[19] M. Lee, C. H. Choi, and S. Oh, "A procrustean Markov process for non-rigid structure recovery," in *CVPR*, 2014.

[20] S. Parashar, D. Pizarro, and A. Bartoli, "Isometric non-rigid shape-from-motion with riemannian geometry solved in linear time," *TPAMI*, vol. 40, no. 10, pp. 2442–2454, 2018.

[21] J. Xiao, J. Chai, and T. Kanade, "A closed-form solution to non-rigid shape and motion," *IJCV*, vol. 67, no. 2, pp. 233–246, 2006.

[22] A. Agudo and F. Moreno-Noguer, "Robust spatio-temporal clustering and reconstruction of multiple deformable bodies," *TPAMI*, vol. 41, no. 4, pp. 971–984, 2019.

[23] Y. Chen, H. Xu, C. Caramanis, and S. Sanghavi, "Robust matrix completion with corrupted columns," in *ICML*, 2011.

[24] N. Boumal, B. Mishra, P. A. Absil, and R. Sepulchre, "Manopt, a matlab toolbox for optimization on manifolds," *JMLR*, vol. 15, no. 4, pp. 1455–1459, 2014.

[25] J.F. Cai, E. Candes, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM JO*, vol. 20, no. 4, pp. 1956–1982, 2010.

[26] Z. Lin, M. Chen, L. Wu, and Y. Ma, "The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices," *UIUC Technical Report UILU-ENG-09-2215*, 2009.

[27] W. Y. Chen, Y. Song, H. Bai, C.J. Lin, and E. Chang, "Parallel spectral clustering in distributed systems," *TPAMI*, vol. 33, no. 3, pp. 568–586, 2010.

[28] D. Novotny, N. Ravi, B. Graham, N. Neverova, and A. Vedaldi, "C3DPO: Canonical 3D pose networks for non-rigid structure from motion," in *ICCV*, 2019.

[29] A. Agudo, "Unsupervised 3D reconstruction and grouping of rigid and non-rigid categories," *TPAMI, to appear*, 2020.