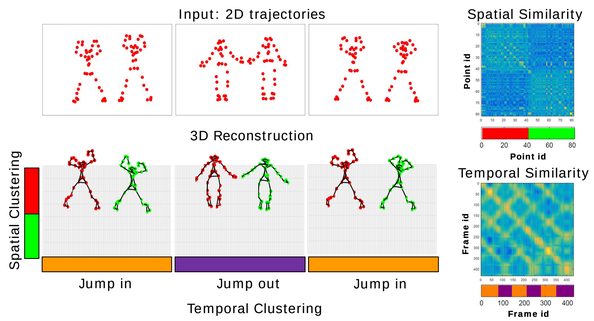
DUST: Dual Union of Spatio-Temporal Subspaces for Monocular Multiple Object 3D Reconstruction

Antonio Agudo and Francesc Moreno-Noguer

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

Problem Statement

- Given *2D point tracks* in a video where several objects are performing complex tasks, moving and even interacting between them.
- We want to jointly recover *spatial segmentation into objects*, *temporal segmentation into actions*, the *4D dynamic shape* and *camera pose*.
- We propose a **dual union of spatio-temporal subspaces** to encode deformation.
- The spatio-temporal segmentation is obtained by spectral clustering over *affinity matrices*.

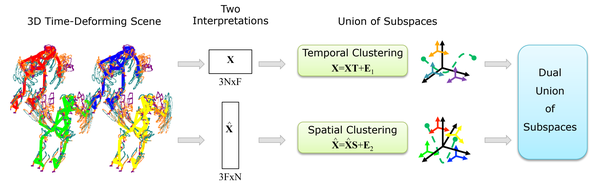


Dual Union of Spatio-Temporal Subspaces

- Under orthography, a time-deforming scenario $\hat{\mathbf{X}}$ can be projected as $\mathbf{P}$ such that:

$$\underbrace{\begin{bmatrix} \tilde{\mathbf{p}}_1^1 & \dots & \tilde{\mathbf{p}}_N^1 \\ \vdots & \ddots & \vdots \\ \tilde{\mathbf{p}}_1^F & \dots & \tilde{\mathbf{p}}_N^F \end{bmatrix}}_{\mathbf{P}} = \underbrace{\begin{bmatrix} \mathbf{R}^1 & & \\ & \ddots & \\ & & \mathbf{R}^F \end{bmatrix}}_{\mathbf{G}} \underbrace{\begin{bmatrix} \mathbf{x}_1^1 & \dots & \mathbf{x}_N^1 \\ \vdots & \ddots & \vdots \\ \mathbf{x}_1^F & \dots & \mathbf{x}_N^F \end{bmatrix}}_{\hat{\mathbf{X}}}$$

- We propose a **dual union of spatio-temporal subspaces**, where the 4D information can have *two different representations*: $\hat{\mathbf{X}}$ and $\mathbf{X}$.



- $\mathbf{T}$ and $\mathbf{S}$ are temporal $F \times F$ and spatial $N \times N$ *affinity matrices*, respectively.
- We re-arrange the elements of $\hat{\mathbf{X}}$ into a new $\mathbf{X}$:

$$\mathbf{X} = \begin{bmatrix} x_1^1 & \dots & x_N^1 & y_1^1 & \dots & y_N^1 & z_1^1 & \dots & z_N^1 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ x_1^F & \dots & x_N^F & y_1^F & \dots & y_N^F & z_1^F & \dots & z_N^F \end{bmatrix}^\top$$

- Qualitative comparison of our method with other NRSfM approaches. Our method is *the only one* that simultaneously provides 4D reconstruction, shape segmentation and temporal clustering:

	[3, 37]	[21, 22]	[16, 20]	[18, 24, 25]	[40]	Ours
Rank required	–	–	✓	✓	✓	✓
Occlusion handling	✓	✓	–	✓	–	✓
Multiple objects	–	✓	–	–	✓	✓
Temporal clustering	–	–	–	–	✓	✓
Shape clustering	–	–	–	–	–	✓

Augmented Lagrange Multiplier Optimization

- We propose to solve the optimization problem:

$$\underset{\mathbf{\Theta}}{\arg\min} \; \| (\mathbf{O} \otimes \mathbf{1}_2) \odot (\mathbf{P} - \bar{\mathbf{P}}) \|_F^2 + \beta \|\mathbf{P}\|_* + \phi \|\mathbf{T}\|_* + \phi \|\mathbf{S}\|_* + \gamma \|\mathbf{X}\|_* + \lambda_t \|\mathbf{E}_t\|_1 + \lambda_s \|\mathbf{E}_s\|_1$$

$$\text{subject to} \quad \begin{aligned} \mathbf{P} &= \mathbf{G}\hat{\mathbf{X}} && \text{3D-to-2D projection} \\ \mathbf{I}_{2F} &= \mathbf{G}\mathbf{G}^\top && \text{Rotation orthonormality} \\ \mathbf{X} &= \mathbf{X}\mathbf{T} + \mathbf{E}_t && \text{Temporal Clustering} \\ \hat{\mathbf{X}} &= \hat{\mathbf{X}}\mathbf{S} + \mathbf{E}_s && \text{Spatial Clustering} \end{aligned}$$

where $\mathbf{\Theta} \equiv \{\mathbf{P}, \mathbf{G}, \mathbf{T}, \mathbf{S}, \mathbf{X}, \mathbf{E}_t, \mathbf{E}_s\}$. $\bar{\mathbf{P}}$ be a possibly incomplete 2D measurement matrix, and $\mathbf{O}$ its corresponding $F \times N$ observation matrix.

- We present a **three-step strategy** to minimize the cost function: 1) completing missing entries (assuming low-rank), 2) estimating camera rotation and 3) joint clustering and 4D reconstruction (dual union of subspaces).
- For the third step, we present an Augmented Lagrange Multiplier (ALM) scheme, solving:

$$\underset{\mathbf{T},\mathbf{S},\mathbf{X},\mathbf{E}_t,\mathbf{E}_s}{\arg\min} \; \phi\|\mathbf{T}\|_* + \phi\|\mathbf{S}\|_* + \gamma\|\mathbf{X}\|_* + \lambda_t\|\mathbf{E}_t\|_1 + \lambda_s\|\mathbf{E}_s\|_1$$

$$\text{subject to} \quad \begin{aligned} \mathbf{P} &= \mathbf{G}\hat{\mathbf{X}} \\ \mathbf{X} &= \mathbf{X}\mathbf{T} + \mathbf{E}_t \\ \hat{\mathbf{X}} &= \hat{\mathbf{X}}\mathbf{S} + \mathbf{E}_s \end{aligned}$$

Spatio-Temporal Clustering and 4D Reconstruction Results

Evaluation on CMU videos with 2 and 4 subjects

Method / Data / Metric:	CSF e_X	KSTA e_X	BMM e_X	EM-PND e_X	TUS e_X	GBNR e_X	CNR e_X	Ours (DUST) 0% missing data e_X	Ours (DUST) 0% missing data $e_S[\%]$	Ours (DUST) 0% missing data $e_T[\%]$	Ours (DUST) sparse/structured e_X	Ours (DUST) sparse/structured e_X
Jump	0.053	0.071	0.078	0.065	0.054	0.070	0.074	**0.045**	0.0(2)	5.8(3)	0.047	0.062
Pull	0.123	0.128	0.146	**0.113**	0.116	0.138	0.183	0.118	0.0(2)	8.3(4)	0.120	0.121
Soldiers	0.104	0.106	0.080	0.342	0.073	0.076	0.091	**0.049**	1.2(2)	5.0(2)	0.050	0.067
Stares Down	0.036	0.022	0.050	**0.013**	0.032	0.048	0.038	0.016	0.0(2)	0.0(2)	0.018	0.024
Stumbles	**0.094**	0.102	0.124	0.099	0.112	0.119	0.119	0.096	0.0(2)	1.3(2)	0.098	0.111
Squats	0.047	0.041	0.040	0.055	0.016	0.036	0.023	**0.015**	4.8(2)	0.8(2)	0.018	0.021
Synchronized	0.141	0.145	0.152	0.145	0.091	0.147	0.112	**0.083**	0.0(2)	1.2(2)	0.085	0.086
Violence	0.072	0.073	0.090	0.150	0.081	0.085	0.135	**0.060**	0.0(2)	1.1(3)	0.062	0.076
Zombie	0.070	0.067	0.062	0.076	0.056	0.061	0.087	**0.043**	0.0(2)	9.3(3)	0.044	0.066
Average error:	0.082	0.084	0.091	0.117	0.070	0.087	0.096	**0.058**	0.6	3.6	0.060	0.070
Relative error:	1.41	1.44	1.56	2.01	1.21	1.50	1.65	**1.00**	-	-	1.04	1.21
Blind4	0.047	**0.040**	0.079	0.079	0.059	0.074	0.137	0.045	0.0(4)	0.3(2)	0.045	0.052
Chicken4	0.030	0.034	0.027	0.022	0.017	0.021	0.022	**0.015**	0.0(4)	0.2(3)	0.018	0.022
Greet4	0.048	**0.041**	0.078	0.069	0.072	0.077	0.085	0.051	0.0(4)	2.0(3)	0.052	0.053
Shelters4	0.055	0.053	0.087	0.053	0.037	0.085	0.069	**0.034**	0.0(3)	3.2(2)	0.034	0.045
Soda4	0.011	0.011	0.009	0.010	0.009	0.011	0.016	**0.007**	0.0(4)	1.0(2)	0.008	0.011
Synchronized4	0.093	0.077	0.056	0.042	0.046	0.049	0.078	**0.041**	0.0(4)	1.2(2)	0.043	0.045
Zombie4	0.055	0.067	0.047	0.051	0.043	0.046	0.061	**0.033**	0.0(4)	8.9(3)	0.034	0.034
Average error:	0.048	0.046	0.055	0.046	0.040	0.052	0.067	**0.032**	0.0	2.4	0.033	0.037
Relative error:	1.50	1.43	1.69	1.44	1.26	1.62	2.09	**1.00**	-	-	1.03	1.17

4D Reconstruction: 2 and 4 subjects