Image Collection Pop-up: 3D Reconstruction and Clustering of Rigid and Non-Rigid Categories

Antonio Agudo Melcior Pijoan Francesc Moreno-Noguer
Institut de Robòtica i Informàtica Industrial, CSIC-UPC, 08028, Barcelona, Spain

Abstract

This paper introduces an approach to simultaneously estimate 3D shape, camera pose, and object and type of deformation clustering, from partial 2D annotations in a multi-instance collection of images. Furthermore, we can indistinctly process rigid and non-rigid categories. This advances existing work, which only addresses the problem for one single object or, if multiple objects are considered, they are assumed to be clustered a priori. To handle this broader version of the problem, we model object deformation using a formulation based on multiple unions of subspaces, able to span from small rigid motion to complex deformations. The parameters of this model are learned via Augmented Lagrange Multipliers, in a completely unsupervised manner that does not require any training data at all. Extensive validation is provided in a wide variety of synthetic and real scenarios, including rigid and non-rigid categories with small and large deformations. In all cases our approach outperforms state-of-the-art in terms of 3D reconstruction accuracy, while also providing clustering results that allow segmenting the images into object instances and their associated type of deformation (or action the object is performing).

1. Introduction

Simultaneously estimating 3D object shape and camera pose from a collection of RGB images either acquired from different viewpoints or by a single moving camera is one of the most active research areas in computer vision. Early works addressed this problem under the assumption of a rigid structure [1, 29, 33]. More recently, many efforts have focused on the non-rigid case, to retrieve deforming 3D shape and camera motion from only 2D measurements in a monocular video [2, 23, 26, 37, 39]. This problem is known to be inherently ambiguous and demands introducing sophisticated priors. Probably, the most standard priors include the use of different modalities of low-rank subspaces to constrain the solution space [3, 7, 9, 26, 35]. Moreover, these algorithms exploit the fact that input images smoothly change viewpoints. This allows introducing temporal smoothness on the shape deformations and obtain more accurate solutions [5, 28].

All these previous approaches, however, solve the problem for one single object instance. There exist works addressing scenarios with multiple objects within a category. For instance, if the observed category is rigid (e.g., cars or aeroplanes) and all objects in it have the same geometry, the problem can be addressed as a rigid Structure from Motion (SfM) one [31, 38]. When object instances within the category have distinct geometry, even if they are rigid (e.g., different model cars), the global problem of retrieving their shape can be formulated in a non-rigid manner [20]. This can be extended to inherently non-rigid classes (e.g., faces, animal poses), in which case, both inter- and intra-object deformations shall be considered [4]. However, all these works are only focused on the reconstruction problem, and assume the object clustering to be known a priori.

In this paper we move a step forward and tackle the problem in which the object clusters are not known a priori. That is, given an input collection of images of a specific category, we aim at simultaneously clustering them into different object instances and recovering their 3D shape regardless of whether the objects are *rigid* or *non-rigid*. Camera pose is also estimated. For instance, as shown in Fig. 1-Left, given a number of images of bicycles (5 models seen from different viewpoints) our approach clusters them into each of the models and reconstructs their 3D shape. Note that some observations of the bicycle instances are very similar and difficult to distinguish from only 2D annotations. Simultaneously reasoning about the clustering and 3D reconstruction helps improving both tasks. The proposed method generalizes to non-rigid categories as well. As shown in Fig. 1-Right, given a collection of face images of 5 humans under different viewpoints and facial expressions, our algorithm jointly splits the images into each of the individuals and their actions, and retrieves their 3D deformable shape.

In order to simultaneously tackle clustering and reconstruction from a collection of unordered images, we propose a novel optimization framework that builds upon recent Non-Rigid Structure from Motion approaches (NRSfM) [6,

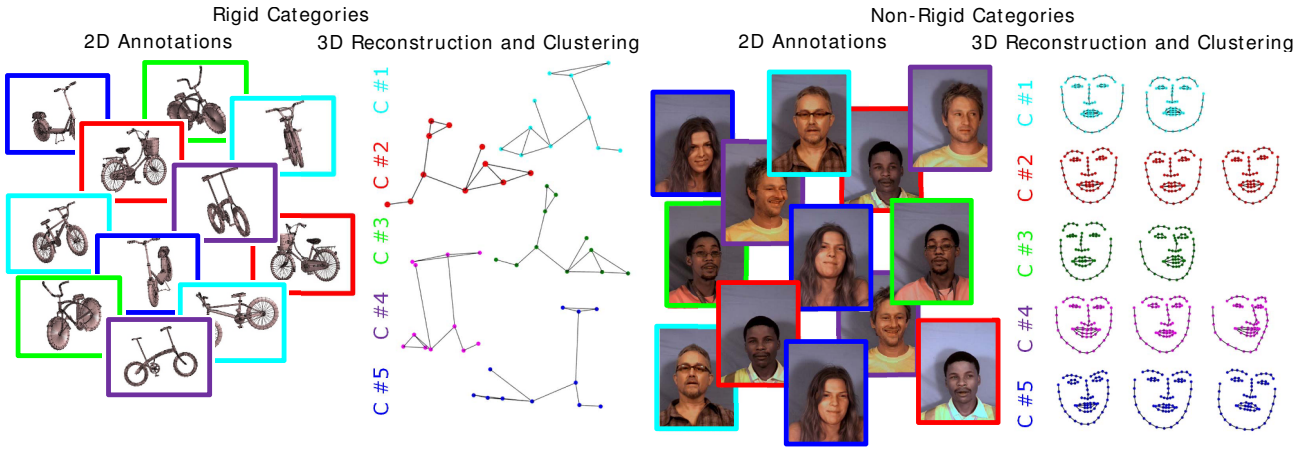

Figure 1. **Simultaneous 3D reconstruction and object clustering from partial 2D annotations of rigid and non-rigid categories.** In both cases, input data consists of a collection of RGB images with –possibly incomplete– 2D semantic point detections. The number of objects within the category is not known. Our goal is to jointly estimate the 3D object reconstruction in each image, the camera pose, and the instance cluster (we use a different color per each object instance). **Left:** A rigid *bicycle* category, in which each instance has a single 3D configuration. **Right:** A non-rigid *face* category, where every instance may potentially have as many 3D configurations as its number of images. This graph only shows instance clustering, but as we shall see in the results, our approach also permits segmenting every non-rigid instance into several types of deformation (or expressions in the case of the faces).

41]. More specifically, we model the 3D shape by multiple unions of unknown subspaces, accounting for rigid plus small and large non-rigid deformations. These subspaces, in conjunction with additional matrices encoding the similarities among the samples and among their deformations, are retrieved from partial 2D annotations using an efficient Augmented Lagrange Multiplier (ALM) scheme. A subsequent spectral clustering on the similarity matrices yields the results of the segmentation. The whole algorithm works in a fully unsupervised manner, without requiring to know a priori the number of object clusters nor any other information about the type of deformation (if any) undergone by the objects. We thoroughly evaluate this algorithm on synthetic and real images for rigid and non-rigid categories, and provide improved 3D reconstructions compared to state-of-the-art approaches for which ground truth clustering is given.

2. Related Work

Inferring the 3D shape while retrieving camera pose from only 2D point measurements in a collection of RGB images, is a mature problem when the observed object is rigid. In this case, the rigidity constraint is sufficient to make the problem well-posed, yielding impressively accurate solutions [1, 29, 38]. In contrast, handling non-rigid scenarios becomes an ill-posed problem that requires to exploit the denominated *art of priors* to constrain the solution space. The most standard prior used in NRSfM consists in constraining the deforming shape to lie in a low-rank subspace. To learn such a low-rank model, early approaches rely on factorization [10, 34, 40], or optimization-based strategies [9, 26, 39]. More recently, the low-rank constraint has been imposed by means of PCA-like formulations in which the rank of the shape matrix is optimized. These type of methods either assume the data lies in a single low dimensional shape space [16, 19, 21], or in a union of temporal [41] or spatio-temporal subspaces [6]. Low-rank models were also extended to the temporal domain, by exploiting pre-defined trajectory basis [7, 35], the combination of shape-trajectory domains [22, 23], and the force space that induces the deformations [3]. As most of the methods process video sequences, additional temporal smoothness priors have allowed to obtain more consistent solutions for rigid [33] and non-rigid domains [9, 21, 22, 27].

In any event, while achieving remarkable results, all previous approaches aim at modeling one single object in a category, typically observed from smoothly changing viewpoints. This means they are not directly applicable to the multi-object scenario we contemplate in this paper. However, there have been some attempts along this line. Recent solutions to reconstruct rigid categories from single images [24], resort to large amounts of training data to constrain the solution space. Our approach, instead, aims at learning the solution space on the fly from a collection of images, without requiring any training data at all. There exist very recent works implementing this idea on rigid object categories, either exploiting the concept of symmetry [20], or imposing a sparse shape-space model [25]. In [4], this was extended to non-rigid categories through a dual low-rank shape model which allowed handling small deformations. Nevertheless, these works are still limited by the fact that they assume the clustering of the image collection into objects needs to be known a priori.

Our Contributions. We overcome most of the limitations

Meth. \ Feat.	Automatic rank	Occlusion handing	Object / Type of deformation clustering	Rigid/Non-Rigid categories
[3, 9, 39]	–	✓	–/–	–/–
[22, 23]	–	✓	–/–	–/–
[16, 21]	✓	–	–/–	–/–
[19, 26, 27]	✓	✓	–/–	–/–
[41]	✓	–	–/✓	–/–
[4]	–	✓	–/–	–/✓
[20, 25]	✓	✓	–/–	✓/–
Ours	✓	✓	✓/✓	✓/✓

Table 1. **Comparison of our approach with state-of-the-art NRSfM methods.** Our approach is the only one that simultaneously provides 3D reconstruction of both rigid and non-rigid categories, and estimates clustering per object instance and type of deformation. Additionally, it can also handle incomplete 2D annotations, and does not need to adjust the rank of the basis.

of previous methods with an approach that jointly retrieves 3D shape, camera pose, object and deformation clustering, and the incomplete 2D annotations, for both rigid and non-rigid categories of objects. To this end, we encode object deformation by means of multiple unions of subspaces, without requiring any prior knowledge about the dimensionality of the subspaces nor which data points belong to which subspace. As a result, we obtain a unified and unsupervised framework which does not need training data. We are not aware of any other work jointly offering all these characteristics. Table 1 provides a qualitative comparison of the main features offered by our solution and the most relevant state of the art.

3. Revisiting Structure from Motion

We next review the SfM formulation that will be later used to describe our approach on rigid and non-rigid category reconstruction and clustering. Let us consider a set of P points detected on I images. Let $\mathbf{x}_p^i = [x_p^i, y_p^i, z_p^i]^\top$ be the 3D coordinates of the p-th point in image i, and $\mathbf{w}_p^i = [u_p^i, v_p^i]^\top$ its 2D position according to an orthographic projection. We can jointly write the 3D-to-2D mapping of all points as the following linear system:

$$\underbrace{\begin{bmatrix} \mathbf{w}_1^1 & \dots & \mathbf{w}_P^1 \\ \vdots & \ddots & \vdots \\ \mathbf{w}_1^I & \dots & \mathbf{w}_P^I \end{bmatrix}}_{\mathbf{W}} = \mathbf{G} \underbrace{\begin{bmatrix} \mathbf{x}_1^1 & \dots & \mathbf{x}_P^1 \\ \vdots & \ddots & \vdots \\ \mathbf{x}_1^I & \dots & \mathbf{x}_P^I \end{bmatrix}}_{\hat{\mathbf{X}}} + \mathbf{T}, \quad (1)$$

where $\mathbf{W}$ is a $2I \times P$ matrix with the 2D measurements arranged in columns, $\mathbf{G}$ is a $2I \times 3I$ block diagonal matrix made of I truncated 2×3 camera rotations $\mathbf{R}^i$, $\hat{\mathbf{X}}$ is a $3I \times P$ matrix with the 3D locations of the points for all the collection, also arranged in columns, and $\mathbf{T}$ is a $2I \times P$ matrix that stacks P copies of the I bi-dimensional translation vectors $\mathbf{t}^i$. The SfM problem consists in recovering the 3D shape $\hat{\mathbf{X}}$, along with the camera motion $\{\mathbf{R}^i, \mathbf{t}^i\}$ with $i = \{1, \dots, I\}$, from 2D point detections $\mathbf{W}$.

When a rigid object is observed, i.e., $\mathbf{x}_p^1 = \mathbf{x}_p^2 = \dots = \mathbf{x}_p^I$, the shape matrix can be simplified. In this case, the shape can be estimated by applying SVD-based factorization strategies, and enforcing a 3-rank constraint on $\mathbf{W}$ [31, 38] together with orthonormality constraints on $\mathbf{G}$. If, by contrast, the observed object were non-rigid, the I locations of every point can be potentially different. Then, shape and motion can be retrieved by enforcing a $3K$-rank decomposition over the measurement matrix $\mathbf{W}$ [10, 40], where K represents the rank of the linear subspace.

For later computations, we will also re-arrange the elements of $\hat{\mathbf{X}}$ into a new $3P \times I$ matrix $\mathbf{X}$ encoding the x, y and z coordinates in different rows. Both matrices can be related through a function $q(\cdot)$ such that $\hat{\mathbf{X}} = q(\mathbf{X})$ [6, 16, 19, 21]. This new interpretation has the advantage of allowing for a K-rank decomposition, rather than $3K$, avoiding the use of unnecessary degrees of freedom.

4. Shape as Multiple Unions of Subspaces

This section describes the deformation model we propose to represent the 3D shape of an unknown number of objects belonging to a specific family and their relation with the 2D measurements in a collection of images. In the following we shall consider three scenarios depending on the nature of the deformation: rigid objects, and non-rigid shapes with small and large deformations.

4.1. Case 1: Rigid Objects

Let us consider a collection of I images of a number of rigid objects that belong to the same category (e.g., *bus* in Fig. 2-Left). Each object is characterized by P semantic 3D points, which, for the moment, we will assume to be all visible in all images. The number of objects and images per object is not known a priori. Our goal is, given the 2D annotations, to reconstruct the 3D position of the P points in all images, and identify and group the images belonging to the same object. When only considering one single object instance, the problem becomes a standard rigid SfM [31, 38], which we will not tackle in this paper. When more than one type of object is considered, we can consider their P semantic points to be related by a geometric transformation that includes both a rigid and a non-rigid deformation. Reconstructing the P points can then be addressed in a NRSfM context, although without enforcing temporal consistency between consecutive images.

Assuming a single low-rank constraint could be sufficient to span the solution space of the 3D shape in this case, as was shown in [20]. However, this formulation is very sensitive to the chosen rank of the subspace, and its optimal value may be very difficult to discover when the number of object instances is unknown. Additionally, the maximum rank, and hence the expressiveness of the subspace, is limited by construction by the number of semantic

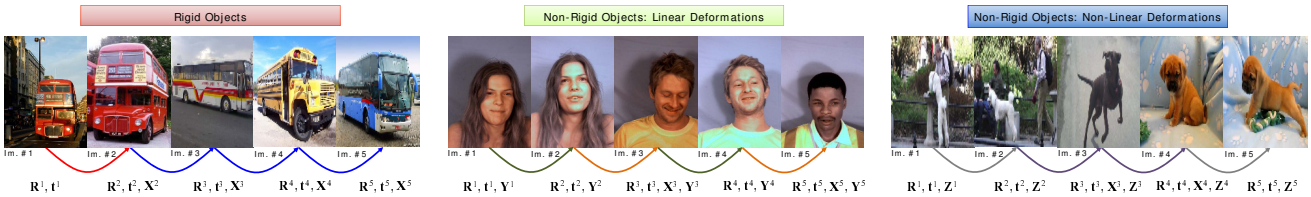

Figure 2. **Unified formulation to recover the 3D shape of rigid and non-rigid categories from a collection of RGB images.** Our deformation model considers several types of transformations. In all cases, between every pair of images, we define a rigid motion consisting of a rotation matrix $\mathbf{R}^i$ and a translation vector $\mathbf{t}^i$. **Left:** The geometric relation between pairs of objects in a rigid category (e.g., *bus*) can be defined in the context of a NRSfM problem using a global deformation $\mathbf{X}^i$. **Middle:** In some categories (e.g., *face*) each of the objects deform by themselves. In this case, besides the global deformation between objects, we define a linear deformation $\mathbf{Y}^i$ to encode the non-rigid motion that each object may undergo. **Right:** Other categories (e.g., *dog*), following more complex patterns. In this case we consider a non-linear deformation $\mathbf{Z}^i$. Our deformation model, simultaneously considers all types of deformations and automatically learns the contribution of each term to describe the geometry of the objects in a specific category. Images in this figure are taken from the PASCAL VOC [18], MUCT [32], and TigDog [17] datasets, respectively.

points P, which in most of our scenarios is rather small[1]. To overcome these difficulties, we introduce a formulation that models deformation using a union of subspaces, allowing to automatically represent a wide range of deformations, from simple low-rank solution spaces to highly expressive ones. We mathematically write this model as:

$$\mathbf{X} = \mathbf{X}\mathbf{Q} + \mathbf{E}_1 \,, \tag{2}$$

where $\mathbf{E}_1$ is a $3P \times I$ residual noise matrix, and $\mathbf{Q}$ is a $I \times I$ similarity matrix which should have higher entries for pairs of images of the same object. In essence, by doing this, we bring the standard scenario of the rigid SfM problem to the non-rigid domain, with the additional outcome of clustering the input images into different objects, with no a priori knowledge about the dimensionality of the subspaces nor which data points belong to which subspace. As we shall see later, once the similarity matrix $\mathbf{Q}$ is estimated, spectral clustering algorithms [13] can be applied on it to discover and match the different objects within the collection.

4.2. Case 2: Non-Rigid Objects with Small Deformations

We next consider the case in which the objects, besides a rigid motion, also undergo small deformations or a partial deformation of some of their points. Figure 2-Middle shows an example of such situation for faces, where most of the deformation is concentrated around the mouth and eyes areas. Existing solutions address this case by enforcing a single low-rank subspace [9, 16, 34], when only considering one object, or through a dual low-rank shape representation [4] when multiple objects appear in the set of images. Most these approaches, however, still require accurately adjusting a priori the dimensionality of the subspace.

In order to account for such small and sparse deformations we will introduce a matrix $\mathbf{Y} \in \mathbb{R}^{3P \times I}$ in our model. In contrast to the aforementioned approaches, no low-rank constraint will be enforced, but only a sparsity constraint that allows the deformation of just a few points.

[1]Note that defining the same semantic points in all objects of a category is a difficult task. In this paper they were manually annotated for some collections, and the exact position can be very subjective in certain cases.

4.3. Case 3: Non-Rigid Objects with Large Deformations

We finally consider the case in which the images correspond to a number of non-rigid objects of a given category, that can potentially undergo large deformations. The articulated motion of humans or animals (see Fig. 2-Right) are examples of this scenario. Again, we consider the number of objects in the category is not known.

In order to model this situation, we require a model with large expressibility. This is achieved by introducing into the model a matrix $\mathbf{Z} \in \mathbb{R}^{3P \times I}$ which is enforced to be formed by another union of subspaces:

$$\mathbf{Z} = \mathbf{Z}\mathbf{Q}\mathbf{H} + \mathbf{E}_2 \,, \tag{3}$$

where $\mathbf{H}$ is again a $I \times I$ similarity matrix, and $\mathbf{E}_2$ is a residual noise one. Note that in this case we are considering the total similarity to be defined by the product $\mathbf{QH}$, that is, we jointly consider similarity between objects and types of deformation. Like mentioned before for the matrix $\mathbf{Q}$, applying spectral clustering on the similarity $\mathbf{QH}$ will yield clusters of objects with similar deformation (e.g., person 'A' or 'B' smiling, person 'A' or 'B' with closed mouth).

5. 3D Shape and Clustering per Object and Deformation Type

Our goal is to jointly recover 3D shape, camera motion, and object and deformation type from partial 2D observations. In this section we formulate this problem by integrating the three deformation cases discussed above into the 3D-to-2D projection model defined in Eq. (1). We then describe the optimization scheme we propose to solve it.

5.1. Problem Formulation

Let $\bar{\mathbf{W}}$ be a possibly incomplete matrix of 2D detections (recall that I is the number of images of an object class and P the number of points defining the class), and $\mathbf{O}$ the corresponding $I \times P$ observation matrix with $\{1, 0\}$ entries indicating whether a specific point in an image is observed or not. Given $\bar{\mathbf{W}}$ and $\mathbf{O}$, we aim at recovering: 1) the 3D locations of all points in all images, encoded by the shape matrices $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$ defined in Section 4; 2) the object specific $\mathbf{Q}$ and deformation specific $\mathbf{QH}$ similarity matrices which we shall use later for clustering; 3) the camera pose parameters $(\mathbf{G}, \mathbf{T})$ in all images; and 4) the complete 2D detections matrix $\mathbf{W}$. We denote all these unknown parameters, plus the corresponding noise matrices by $\mathbf{\Psi} \equiv \{\mathbf{W}, \mathbf{G}, \mathbf{T}, \mathbf{Q}, \mathbf{H}, \mathbf{X}, \mathbf{Y}, \mathbf{Z}, \mathbf{E}_1, \mathbf{E}_2\}$.

In order to tackle this problem we propose optimizing a cost function that enforces the correct reprojection of the estimated 3D shape onto the image and incorporates the shape constraints we mentioned when describing the model in Section 4. In particular, the matrices $\mathbf{X}$ and $\mathbf{Z}$ are enforced to lie in low-rank subspaces. Since rank minimization is a non-convex NP-hard problem [36], the nuclear norm is used as a convex relaxation [12, 14]. Sparsity on the component $\mathbf{Y}$ is encouraged through l_1-norm minimization. Additionally, we consider the mixed $l_{2,1}$-norm over the matrices of residual noise $\mathbf{E}_1$ and $\mathbf{E}_2$, as this type of norm favors structured sparsity. Note that structured noise patterns may occur on the shape matrices $\mathbf{X}$ and $\mathbf{Z}$ when specific data points are missing or corrupted by noise. Taking all this into consideration we formulate the optimization problem as follows:

$$\begin{aligned}\arg\min_{\mathbf{\Psi}} \; & \| \left(\mathbf{O} \otimes \mathbf{1}_2\right) \odot \left(\mathbf{W} - \bar{\mathbf{W}}\right) \|_F^2 + \beta \|\mathbf{W}\|_* + \phi \|\mathbf{Q}\|_* \\ & + \gamma(\|\mathbf{X}\|_* + \|\mathbf{Y}\|_1 + \|\mathbf{Z}\|_*) + \phi \|\mathbf{H}\|_* \\ & + \lambda(\|\mathbf{E}_1\|_{2,1} + \|\mathbf{E}_2\|_{2,1}) \end{aligned} \tag{4}$$

$$\begin{aligned}\text{subject to} \quad & \mathbf{W} = \mathbf{G}\, q(\mathbf{X} + \mathbf{Y} + \mathbf{Z}) + \mathbf{T} \\ & \mathbf{G}\mathbf{G}^\top = \mathbf{I}_{2I} \\ & \mathbf{X} = \mathbf{X}\mathbf{Q} + \mathbf{E}_1 \\ & \mathbf{Z} = \mathbf{Z}\mathbf{Q}\mathbf{H} + \mathbf{E}_2 \end{aligned}$$

where $\otimes$ and $\odot$ represent the Kronecker and Hadamard products, respectively. $\mathbf{1}$ is a vector of ones, and $\mathbf{I}$ the identity matrix. $\|\cdot\|_F$ indicates the Frobenius norm, $\|\cdot\|_*$ denotes the nuclear norm, and $\|\cdot\|_1$, and $\|\cdot\|_{2,1}$ are the l_1-norm and $l_{2,1}$-norm, respectively. Finally, $\{\beta, \phi, \gamma, \lambda\}$ represent the set of penalty weights.

We approximately solve Eq. (4) in three stages: 1) complete missing entries ; 2) estimate camera pose parameters, and 3) recover the 3D shape reconstruction, and perform clustering per object and type of deformation. We next describe each of these stages.

5.2. Complete Missing Entries

To complete the unobserved 2D detections of $\bar{\mathbf{W}}$ (zeros in the observation matrix $\mathbf{O}$), we independently optimize $\mathbf{W}$ in the first two terms of Eq. (4) while enforcing this matrix to be low rank. As shown in [6, 8, 11], this optimization can be done by means of bilinear factorization, defining $\mathbf{W} = \mathbf{U}\mathbf{V}^\top$. We write the equivalent problem as:

$$\begin{aligned}\arg\min_{\mathbf{W},\mathbf{U},\mathbf{V}} \; & \| \left(\mathbf{O} \otimes \mathbf{1}_2\right) \odot \left(\mathbf{W} - \bar{\mathbf{W}}\right) \|_F^2 + \frac{\beta}{2} \left(\|\mathbf{U}\|_F^2 + \|\mathbf{V}\|_F^2\right) \\ \text{subject to} \quad & \mathbf{W} = \mathbf{U}\mathbf{V}^\top \end{aligned}$$

This can be efficiently solved via ALM. To improve convergence, the missing entries of $\bar{\mathbf{W}}$ are initialized in every image as the mean value of the observed data points.

5.3. Camera Pose Recovery

Once the missing observations are estimated, the camera translation $\mathbf{t}^i$ and rotation $\mathbf{R}^i$ in every image can be inferred from the rest of model parameters. For this purpose, we first estimate the translations in $\mathbf{T}$ as $\mathbf{t}^i = \frac{1}{P}\sum_{p=1}^{P} \mathbf{w}_p^i$. The rotations matrices in $\mathbf{G}$ can then be jointly estimated by solving the following non-convex problem:

$$\begin{aligned}\arg\min_{\mathbf{G}} \; & \frac{1}{2}\|\mathbf{W} - \mathbf{T} - \mathbf{G}\hat{\mathbf{X}}\|_F^2 \\ \text{subject to} \quad & \mathbf{G}\mathbf{G}^\top = \mathbf{I}_{2I} \end{aligned} \tag{5}$$

where the constraint enforces the camera rotation matrices to be orthonormal. This optimization is solved by factorization, using different values of rank and stopping automatically when there is no additional improvement in the average camera orthonormality.

5.4. Joint 3D Reconstruction and Clustering

We finally formulate the problem of simultaneously recovering 3D shape in all images as well as the type of object and deformation clustering. Assuming the matrices $\mathbf{W}$, $\mathbf{G}$ and $\mathbf{T}$ to be known, the optimization problem that needs to be solved becomes:

$$\begin{aligned}\arg\min_{\mathbf{\Psi}'} \; & \gamma(\|\mathbf{X}\|_* + \|\mathbf{Y}\|_1 + \|\mathbf{Z}\|_*) + \|\mathbf{Q}\|_* + \|\mathbf{H}\|_* \\ & + \lambda(\|\mathbf{E}_1\|_{2,1} + \|\mathbf{E}_2\|_{2,1}) \end{aligned} \tag{6}$$

$$\begin{aligned}\text{subject to} \quad & \mathbf{W} = \mathbf{G}\left(\mathbf{A} + \mathbf{B} + \mathbf{C}\right) + \mathbf{T} \\ & \mathbf{X} = \mathbf{X}\mathbf{Q} + \mathbf{E}_1 \\ & \mathbf{Z} = \mathbf{Z}\mathbf{F} + \mathbf{E}_2 \\ & q(\mathbf{X}) = \mathbf{A} \\ & q(\mathbf{Y}) = \mathbf{B} \\ & q(\mathbf{Z}) = \mathbf{C} \\ & \mathbf{F} = \mathbf{Q}\mathbf{H} \end{aligned}$$

where $\mathbf{\Psi}' \equiv \{\mathbf{Q}, \mathbf{H}, \mathbf{F}, \mathbf{X}, \mathbf{A}, \mathbf{Y}, \mathbf{B}, \mathbf{Z}, \mathbf{C}, \mathbf{E}_1, \mathbf{E}_2\}$. Note that compared to the original Eq. (4) we have included three

additional constraints, namely $q(\mathbf{X}) = \mathbf{A}$, $q(\mathbf{Y}) = \mathbf{B}$ and $q(\mathbf{Z}) = \mathbf{C}$, where $q(\cdot)$ simply rearranges the elements of a matrix as discussed in Section 3. Furthermore, to reduce the computational burden, we have included the constraint $\mathbf{F} = \mathbf{QH}$. Without loss of generality we have also reduced the number of weight parameters originally appearing in Eq. (4), by setting $\phi = 1$ and re-scaling the rest. In order to solve this optimization problem, we again resort to the ALM method.

6. Experimental Evaluation

We now present our experimental results for different types of scenarios, including synthetic and real image collections of rigid and non-rigid categories. We provide quantitative and qualitative evaluation and compare our approach against state-of-the-art solutions on several synthetic datasets with 3D ground truth. For quantitative evaluation, we provide the reconstruction error in terms of the normalized mean 3D error e_X used before in [7, 16, 22].

To evaluate the object clustering accuracy, we apply spectral clustering [13] over the estimated matrix $\mathbf{Q}$, and retrieve the $I-$dimensional vector $\mathcal{C}$, where each entry is an integer representing the cluster index. To this end, we define $a_C = 1 - \frac{1}{I}\sum_{i=1}^{I} \mathbb{I}(\mathcal{C}_i \neq \mathcal{C}_i^{GT})$, where $\mathbb{I}(v)$ is the indicator function, i.e., $\mathbb{I}(v) = 1$ if v is true, and 0 otherwise, and $\mathcal{C}_i^{GT}$ is the ground truth cluster index of the i-th image.

6.1. Synthetic Images

We first evaluate our approach on synthetic collections of images of rigid object categories, where the 3D ground truth is obtained from the CAD models of the PASCAL VOC dataset [18]. We choose the categories which are defined by at least eight points. Based on this, we evaluate our approach on eight categories which contain between seven and ten objects each (see Table 2). The penalty terms were tuned with the *Bicycle* collection, and then kept fixed for the rest of experiments. Specifically, we use $\lambda = 0.03$ and $\gamma = 10$.

We compare the 3D reconstruction accuracy of our approach, dubbed MUS (Multiple Union of Subspaces), with two SfM baselines: TK [38] and MC [31]; as well as with seven NRSfM solutions: the shape-trajectory methods CSF [22] and KSTA [23]; the block matrix approach BMM [16], the probabilistic-normal-distribution method EM-PND [26], the temporal union of subspaces TUS [41], the grouping-based NRSfM of GBNR [19] and the consensus NRSfM of CNR [27]. We also include the baseline LRR [30] to obtain the object clustering from 2D annotations. The parameters of these methods were set in accordance to their original papers. We manually set the rank of the subspace for the methods CSF [22] and KSTA [23], using the value that gave the best results. As the source code for TUS [41] is not publicly available, we used our own implementation. In this particular case, we also used our annotation completion and camera motion estimation, as the method did not address any strategy to solve these problems. We would like to recall that our approach does not need manually tuning any subspace rank parameter, neither assigning which images belong to which object class.

Table 2 summarizes the reconstruction errors for all methods and the object clustering accuracy of ours and LRR [30], considering both noise-free and noisy annotations. For the noisy case, we corrupt 2D detections with a zero mean Gaussian perturbation with standard deviation $\sigma_{noise} = 0.01 \max_{i,j,k}\{|d_{ijk}|\}$, where d_{ijk} represents the maximum distance of an image point to the centroid of all the points. Note that MUS consistently outperforms the rest of competing techniques in terms of 3D reconstruction accuracy for both cases, reducing, for instance, the 3D error of other methods by large margins between the 5% and 380% for the noise-free case. Note also that GBNR [19] and CNR [27] do not provide solutions for all collections, as the number of points is not sufficient for their formulation. In addition, our approach also estimates the object clustering, as seen in the right-most column, resulting in very accurate segmentations compared to the LRR [30] solution. Figure 3 shows a few sample images for the *Bicycle* and *Chair* categories, and the 3D reconstructions we obtain.

6.2. Real Images

We next evaluate our approach on several real image collections either deforming linearly (faces) or highly non-linearly (animal motion). Since no ground truth is available for these datasets we only provide qualitative evaluation.

The MUCT collection [32] is made of 72 images of faces of seven people, both men and women, of different ages and races, and under varying poses and expressions. The 2D annotations are obtained by using an off-the-shelf 2D active appearance model [15]. This model consists of 68 2D points, which are all visible in all frames. The results we provide in this dataset are shown in Fig. 4. Despite no quantitative estimates are available, the 3D reconstruction we obtain seems very realistic. We can, however, manually annotate the results of the object segmentation. Even though the 2D shapes are very similar (recall that object segmentation is computed based just on the 2D location of points) we obtain a segmentation accuracy $a_C = 0.68(7)$.

In order to validate our approach against missing annotations, we process the ASL collection [23], consisting of 229 images of a man and a woman. The number of 2D feature points is 77, but some of them are not visible due to structured occlusions (by the hands or face self-rotation). In total, 14.43% of the points are missing. The 3D reconstruction results are shown in Fig. 5. Note that the inferred shapes seem to be very accurate, even when hallucinating the occluded points. In this case, the object segmentation is computed with no error, i.e., $a_C = 1.0(2)$. For this experi-

Algorithm / Data / Metric:	TK [38] e_X	MC [31] e_X	CSF [22] e_X	KSTA [23] e_X	BMM [16] e_X	EM-PND [26] e_X	TUS [41] e_X	GBNR [19] e_X	CNR [27] e_X	LRR [30] a_C	Ours (MUS) e_X	Ours (MUS) a_C
Aeroplane	0.679	0.584	0.363	**0.145**	0.843	0.578	0.294	–	0.263	0.39(7)	0.261	0.95(7)
Bicycle	0.309	0.440	0.424	0.442	0.308	0.763	0.182	0.221	–	0.39(10)	**0.178**	0.95(10)
Bus	0.202	0.238	0.217	0.214	0.300	1.048	0.129	0.214	–	0.44(10)	**0.113**	0.75(10)
Car	0.239	0.256	0.195	0.159	0.266	0.496	0.084	0.217	0.099	0.36(10)	**0.078**	0.87(10)
Chair	0.356	0.447	0.398	0.399	0.357	0.687	0.211	–	–	0.39(10)	**0.210**	0.87(10)
Diningtable	0.386	0.512	0.406	0.372	0.422	0.670	0.265	0.351	–	0.41(10)	**0.264**	0.86(10)
Motorbike	0.339	0.346	0.278	0.270	0.336	0.740	0.228	0.268	–	0.41(10)	**0.222**	0.91(10)
Sofa	0.381	0.390	0.409	0.298	0.279	0.692	0.179	0.264	0.214	0.44(9)	**0.167**	0.85(9)
Average error:	0.361	0.402	0.336	0.287	0.388	0.709	0.196	0.256*	0.192*	0.40	**0.186**	0.88
Relative error:	1.93	2.15	1.80	1.54	2.08	3.80	1.05	1.37*	1.03*	–	**1.00**	–
Aeroplane	0.677	0.583	0.233	**0.183**	0.566	0.760	0.297	–	0.294	0.41(7)	0.271	0.87(7)
Bicycle	0.308	0.442	0.455	0.457	0.307	0.808	0.195	0.231	–	0.38(10)	**0.188**	0.93(10)
Bus	0.204	0.241	0.227	0.218	0.255	1.197	0.139	0.223	–	0.44(10)	**0.122**	0.80(10)
Car	0.241	0.259	0.169	0.164	0.161	0.624	0.100	0.222	0.122	0.36(10)	**0.093**	0.92(10)
Chair	0.358	0.447	0.398	0.396	0.258	0.818	0.221	–	–	0.41(10)	**0.220**	0.91(10)
Diningtable	0.392	0.522	0.414	0.383	0.358	0.807	0.268	0.370	–	0.38(10)	**0.267**	0.89(10)
Motorbike	0.342	0.348	0.295	0.290	0.299	0.748	0.237	0.277	–	0.41(10)	**0.233**	0.89(10)
Sofa	0.384	0.392	0.303	0.294	0.240	0.726	0.188	0.271	0.228	0.42(9)	**0.174**	0.91(9)
Average error:	0.363	0.404	0.312	0.298	0.305	0.811	0.206	0.266*	0.215*	0.40	**0.196**	0.89
Relative error:	1.95	2.17	1.67	1.60	1.64	4.35	1.10	1.42*	1.15*	–	**1.05**	–

Table 2. **Evaluation on synthetic collections for several object categories under noise-free and noisy annotations.** The table reports the 3D reconstruction error e_X for the following SfM baselines: TK [38] and MC [31]; and the NRSfM baselines: CSF [22], KSTA [23], SPM [16], EM-PND [26], TUS [41], GBNR [19] and CNR [27]; and ours (MUS). In all cases, we consider full and clean 2D annotations. The symbol "$-$" indicates the algorithm did not manage to process the sequence, and *, that the summary is obtained considering only the successful cases. Relative error is always computed with respect to MUS reconstruction, on average, the most accurate solution. In addition, for LRR [30] and our approach we also show the clustering accuracies a_C, and the number of object clusters in parentheses.

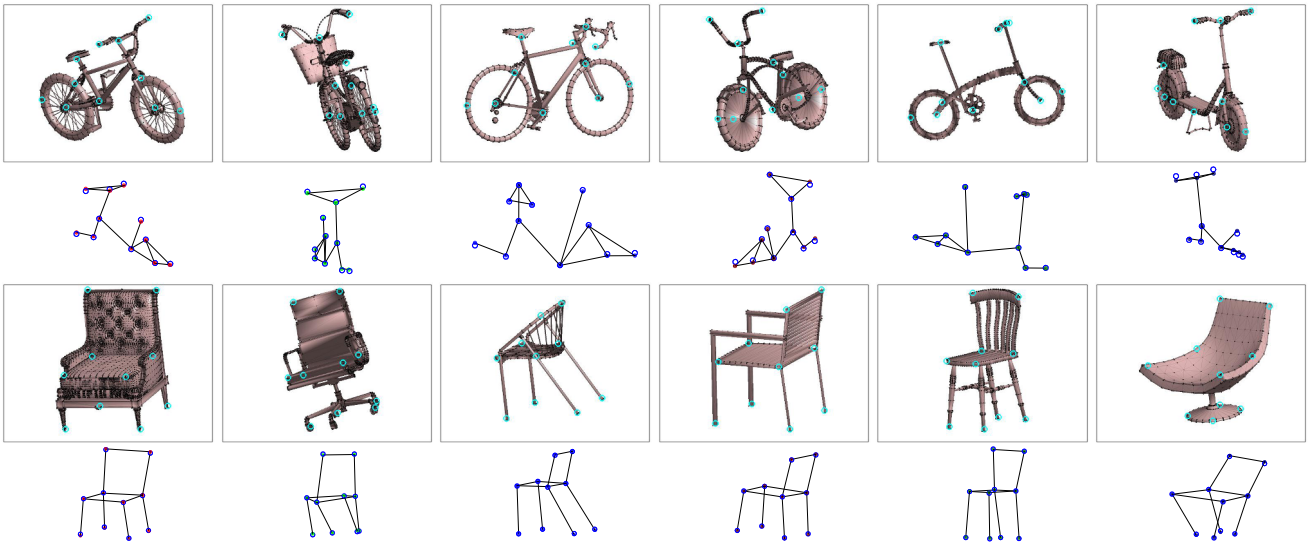

Figure 3. **Bicycle and Chair collections.** The same information is shown for the two experiments. **Top:** Images $\{\#2, \#31, \#53, \#70, \#83, \#148\}$ and $\{\#21, \#37, \#49, \#63, \#93, \#139\}$ for the bicycle and chair collections, respectively. The semantic 2D point measurements fed to our model are represented by cyan circles. **Bottom:** Color-coded dots correspond to our 3D estimation where every color represents a different object, and empty circles represent the 3D ground truth.

ment, we also display the clustering in terms of type of deformation (colored lines in the 3D reconstruction of Fig. 5). These clusters seem to have a clear physical meaning indicating face deformations with closed or open mouth.

We finally evaluate our approach on a challenging collection of dog images [17] with 33 dog instances. This collection is made of 52 images, and we define a model with 19 points, which was manually annotated. Not all points are visible in all images. Concretely, 11.34% of the points are missing. The 3D reconstruction and clustering results are shown in Fig. 6. Again, the 3D shapes we obtain seem very plausible, even for the points that are not observed.

7. Conclusion

In this paper we have extended NRSfM to a new scenario in which we can retrieve 3D shape of either rigid or non-rigid categories from collections of RGB images. Consider-

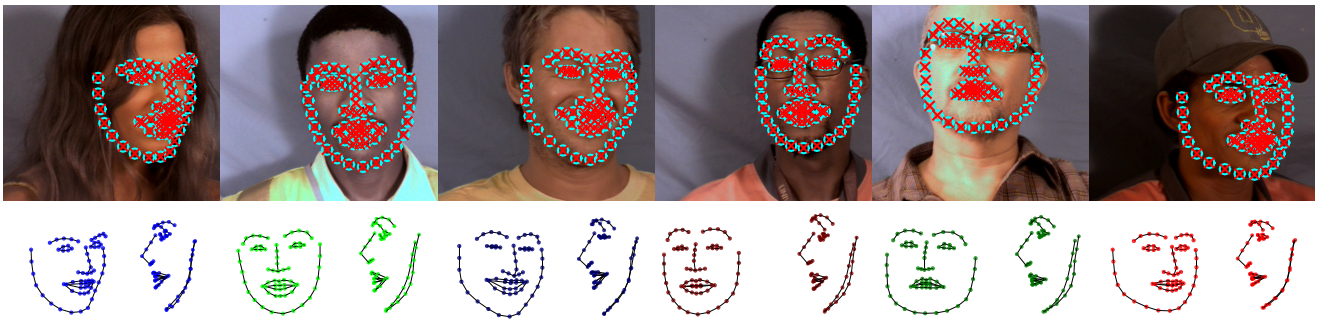

Figure 4. **MUCT collection**. **Top:** Images #3, #26, #32, #46, #65 and #70 of the dataset. Input 2D detections and reprojected 3D shape are shown as cyan circles and red squares, respectively. **Bottom:** Camera viewpoint and side views of the estimated 3D shape. The colored dots indicate the object cluster index estimated by our approach, i.e., a different person in the manifold of faces. Best viewed in color.

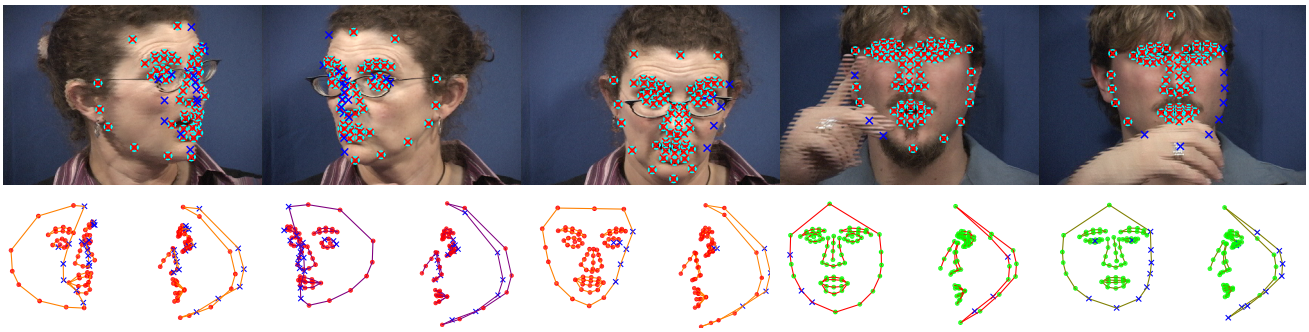

Figure 5. **ASL collection. Top:** Images #29, #47, #100, #142 and #228 of the dataset. Input 2D detections and reprojected 3D shape are shown as cyan circles and red crosses, respectively. Blue crosses correspond to reconstructed (hallucinated) missing points. **Bottom:** Camera viewpoint and side views of our 3D reconstruction, where colored dots (red and green) indicate every human in the collection. The colored lines indicate a specific deformation cluster that was recovered by our approach. These estimated clusters have a clear physical meaning and correspond to open/close mouth (shown in orange/magenta for the woman, and red/dark green for the man). In all cases, 3D reconstructed missing points are represented by blue crosses. Best viewed in color.

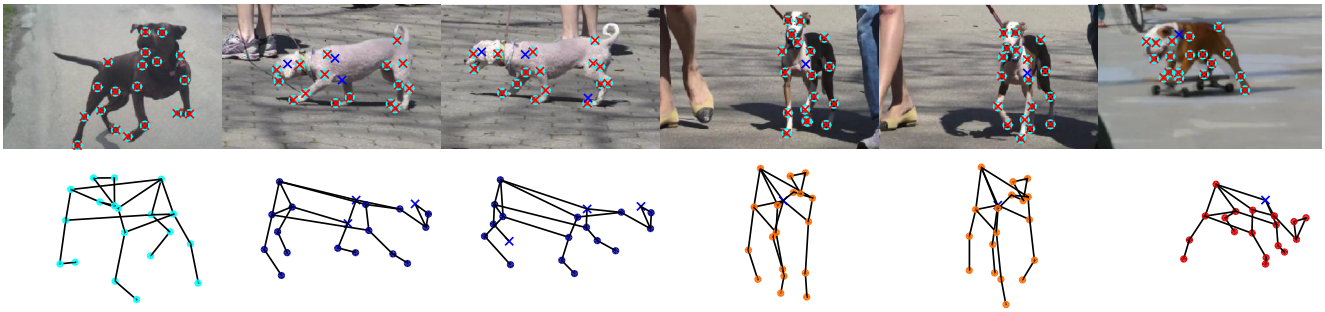

Figure 6. **Dog collection**. **Top:** Images #4, #14, #15, #24, #25 and #51 of the dataset. Input 2D detections and reprojected 3D shape are shown as cyan circles and red crosses, respectively. **Bottom:** 3D reconstruction from a novel point of view, where colored dots indicate the object cluster index estimated by our approach. In both cases, missing points are shown as blue crosses. Best viewed in color.

ing only partial 2D point annotations per image, we propose an approach that besides reconstructing 3D shape, it also estimates camera pose per image, as well as segments the collection of images into different objects and each object geometry, into several deformation primitives. For this purpose, we have introduced a unified formulation that models object shape using multiple unions of subspaces, able to render from rigid motion to highly non-rigid deformations. The model parameters are learned via an ALM scheme in a completely unsupervised manner. We have evaluated our approach on synthetic and real collections of images, of both rigid and non-rigid categories. 3D reconstruction results outperform existing state-of-the-art solutions by large margins. An interesting avenue for future research is to extend our formulation to collections of images of multiples categories, exploring the union of several solution spaces.

Acknowledgments: This work is supported in part by a Google Faculty Research Award, by the Spanish Ministry of Science and Innovation under projects HuMoUR TIN2017-90086-R, and María de Maeztu Seal of Excellence MDM-2016-0656.

References

[1] S. Agarwal, N. Snavely, I. Simon, S. M. Seitz, and R. Szeliski. Building Rome in a day. In *ICCV*, 2009.

[2] A. Agudo, J. M. M. Montiel, L. Agapito, and B. Calvo. Modal space: A physics-based model for sequential estimation of time-varying shape from monocular video. *JMIV*, 57(1):75–98, 2017.

[3] A. Agudo and F. Moreno-Noguer. Learning shape, motion and elastic models in force space. In *ICCV*, 2015.

[4] A. Agudo and F. Moreno-Noguer. Recovering pose and 3D deformable shape from multi-instance image ensembles. In *ACCV*, 2016.

[5] A. Agudo and F. Moreno-Noguer. Combining local-physical and global-statistical models for sequential deformable shape from motion. *IJCV*, 122(2):371–387, 2017.

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

[7] I. Akhter, Y. Sheikh, S. Khan, and T. Kanade. Non-rigid structure from motion in trajectory space. In *NIPS*, 2008.

[8] F. Bach, J. Mairal, and J. Ponce. Convex sparse matrix factorizations. *Technical report HAL-00345747*, 2008.

[9] 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.

[10] C. Bregler, A. Hertzmann, and H. Biermann. Recovering non-rigid 3D shape from image streams. In *CVPR*, 2000.

[11] R. Cabral, F. de la Torre, J. P. Costeira, and A. Bernardino. Unifying nuclear norm and bilinear factorization approaches for low-rank matrix decomposition. In *ICCV*, 2013.

[12] E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717, 2008.

[13] W. Y. Chen, Y. Song, H. Bai, C. Lin, and E. Chang. Parallel spectral clustering in distributed systems. *TPAMI*, 33(3):568–586, 2010.

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

[15] T. Cootes, G. Edwards, and C. Taylor. Active appearance models. In *ECCV*, 1998.

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

[17] L. Del Pero, S. Ricco, R. Sukthankar, and V. Ferrari. Articulated motion discovery using pairs of trajectories. In *CVPR*, 2015.

[18] M. Everingham, L. V. Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (VOC) challenge. *IJCV*, 88(2):303–338, 2010.

[19] K. Fragkiadaki, M. Salas, P. Arbeláez, and J. Malik. Grouping-based low-rank trajectory completion and 3D reconstruction. In *NIPS*, 2014.

[20] Y. Gao and A. L. Yuille. Symmetric non-rigid structure from motion for category-specific object structure estimation. In *ECCV*, 2016.

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

[22] P. F. U. Gotardo and A. M. Martinez. Computing smooth time-trajectories for camera and deformable shape in structure from motion with occlusion. *TPAMI*, 33(10):2051–2065, 2011.

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

[24] A. Kar, S. Tulsiani, L. Carreira, and J. Malik. Category-specific object reconstruction from a single image. In *CVPR*, 2015.

[25] C. Kong, R. Zhu, H. Kiani, and S. Lucey. Structure from category: A generic and prior-less approach. In *3DV*, 2016.

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

[27] M. Lee, J. Cho, and S. Oh. Consensus of non-rigid reconstructions. In *CVPR*, 2016.

[28] M. Lee, C. H. Choi, and S. Oh. A procrustean markov process for non-rigid structure recovery. In *CVPR*, 2014.

[29] J. Lim, J. Frahm, and M. Pollefeys. Online environment mapping. In *CVPR*, 2011.

[30] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma. Robust recovery of subspace structures by low-rank representation. *TPAMI*, 35(1):171–184, 2013.

[31] M. Marques and J. Costeira. Optimal shape from estimation with missing and degenerate data. In *WMVC*, 2008.

[32] S. Milborrow, J. Morkel, and F. Nicolls. The MUCT Landmarked Face Database. *Pattern Recognition Association of South Africa*, 2010.

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

[34] 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.

[35] H. S. Park, T. Shiratori, I. Matthews, and Y. Sheikh. 3D reconstruction of a moving point from a series of 2D projections. In *ECCV*, 2010.

[36] B. Recht, M. Fazel, and P. A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.

[37] C. Russell, R. Yu, and L. Agapito. Video pop-up: Monocular 3D reconstruction of dynamic scenes. In *ECCV*, 2014.

[38] C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: A factorization approach. *IJCV*, 9(2):137–154, 1992.

[39] L. Torresani, A. Hertzmann, and C. Bregler. Nonrigid structure-from-motion: estimating shape and motion with hierarchical priors. *TPAMI*, 30(5):878–892, 2008.

[40] J. Xiao, J. Chai, and T. Kanade. A closed-form solution to non-rigid shape and motion. *IJCV*, 67(2):233–246, 2006.

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