International Journal of Computer Vision (2021) 129:882–899
https://doi.org/10.1007/s11263-020-01406-y

A Benchmark and Evaluation of Non-Rigid Structure from Motion

Sebastian Hoppe Nesgaard Jensen[1] · Mads Emil Brix Doest[1] · Henrik Aanæs[1] · Alessio Del Bue[2]

Received: 28 April 2020 / Accepted: 10 November 2020 / Published online: 29 December 2020

Abstract

Non-rigid structure from motion (NRS*f*M), is a long standing and central problem in computer vision and its solution is necessary for obtaining 3D information from multiple images when the scene is dynamic. A main issue regarding the further development of this important computer vision topic, is the lack of high quality data sets. We here address this issue by presenting a data set created for this purpose, which is made publicly available, and considerably larger than the previous state of the art. To validate the applicability of this data set, and provide an investigation into the state of the art of NRS*f*M, including potential directions forward, we here present a benchmark and a scrupulous evaluation using this data set. This benchmark evaluates 18 different methods with available code that reasonably spans the state of the art in sparse NRS*f*M. This new public data set and evaluation protocol will provide benchmark tools for further development in this challenging field.

Keywords Non-rigid structure from motion · Dataset · Evaluation · Deformation modelling

1 Introduction

The estimation of structure from motion (SfM) using a monocular image sequence is one of the central problems in computer vision. This problem has received a lot of attention, and truly impressive advances have been made over the last ten to twenty years (Hartley and Zisserman 2000; Szeliski 2010; Özyeşil et al. 2017). It plays a central role in robot navigation, self-driving cars, and 3D reconstruction of the environment, to mention a few. A central part of maturing regular SfM is the availability of sizeable data sets with rigorous evaluations, e.g. Menze and Geiger (2015) and Aanæs et al. (2012).

The regular SfM problem, however, primarily deals with rigid objects, which is somewhat at odds with the world we see around us. That is, trees sway, faces express themselves in various expressions, and organic objects are generally non-rigid. The issue of making this obvious and necessary extension of the SfM problem is referred to as the non-rigid structure from motion problem (NRS*f*M). A problem that also has a central place in computer vision. The solution to this problem is, however, not as mature as the regular SfM problem. A reason for this is certainly the intrinsic difficulty of the problem and the scarcity of high quality data sets and accompanying evaluations. Such data and evaluations allow us to better understand the problem domain and better determine what works best and why.

To address this issue, we here introduce a high quality data set, with accompanying ground truth (or reference data to be more precise) aimed at evaluating non-rigid structure from motion. To the best of our knowledge, this data set is significantly larger and more diverse than what has previously been available—c.f. Sect. 3 for a comparison to previous evaluations of NRS*f*M. The presented data set better capture the variability of the problem and gives higher statistical strength of the conclusions reached via it. Accompanying this data set, we have conducted an evaluation of 18 state of the art methods, hereby validating the suitability of our

Communicated by Konrad Schindler.

H. Aanæs and A. Del Bue have equal contribution.

✉ Alessio Del Bue
alessio.delbue@iit.it

Sebastian Hoppe Nesgaard Jensen
snje@dtu.dk

Mads Emil Brix Doest
mebd@dtu.dk

Henrik Aanæs
aanes@dtu.dk

[1] DTU Compute, Kongens Lyngby, Denmark

[2] Pattern Analysis and Computer Vision (PAVIS), Visual Geometry and Modelling (VGM) Lab, Istituto Italiano di Tecnologia (IIT), 08028 Genova, Italy

data set, and providing insight into the state of the art within NRS*f*M. This evaluation was part of the competition we held at a CVPR 2017 workshop, and still ongoing. It is our hope and belief that this data set and evaluation will help in furthering the state of the art in NRS*f*M research, by providing insight and a benchmark. The data set is publicly available at http://nrsfm2017.compute.dtu.dk/dataset together with the description of the evaluation protocol.

This paper is structured by first giving an overview of the NRS*f*M problem, followed by a general description of related work, wrt. other data sets. This section is then followed by a presentation of our data set, including an overview of the design considerations, c.f. Sect. 3, which is followed by a presentation of our proposed protocol for evaluation, c.f. Sect. 4. This leads to the result of our benchmark evaluation in Sect. 5. The paper is rounded off by a discussion and conclusions in Sect. 6.

2 The NRS*f*M Problem

In this section, we will provide a brief introduction of the NRS*f*M problem, followed by a more detailed overview of the ways this problem has been addressed. The intention is to establish a taxonomy to base our experimental design and evaluation upon. In particular, we review sparse NRSfM methods as these approaches are the one evaluated in our benchmark.

The standard/rigid SfM problem, c.f. e.g. Hartley and Zisserman (2000), is an inverse problem aimed at finding the camera positions (and possibly internal parameters) as well as 3D structure—typically represented as a static 3D point set, Q—from a sequence of 2D images of a rigid body. The 2D images are typically reduced to a sparse set of tracked 2D point features, corresponding to the 3D point set, Q. The most often employed observation model, linking 2D image points to 3D points and camera motion is either the *perspective camera model*, or the *weak perspective* approximation hereof. The weak perspective camera model is derived from the full perspective model, by simplifying the projective effect of 3D point depth, i.e. the distance between the camera and 3D point.

The extension from rigid structure from motion to the non-rigid case is by allowing the 3D structure, here points $\mathbf{Q}_f$, to vary from frame to frame, i.e.

$$\mathbf{Q}_f = \left[\mathbf{Q}_{f,1}\ \mathbf{Q}_{f,2}\ \cdots\ \mathbf{Q}_{f,P}\right] \ , \tag{1}$$

where $\mathbf{Q}_{f,p}$ is the 3D position of point p at frame f. To make this NRS*f*M problem well-defined, a prior or regularization is often employed. Here most of the cases target the spatial and temporal variations of $\mathbf{Q}_f$. The fitness of the prior to deformation in question is a crucial element in successfully solving the NRS*f*M problem, and a main difference among NRS*f*M methods is this prior.

In this study, we denote NRS*f*M methods according to a three category taxonomy, i.e. the *deformable model* used (statistical or physical), the *camera model* (affine, weak or full perspective) and the ability to deal with *missing data*. The remainder of this section will elaborate this taxonomy by relating it with the current literature, leading up to a discussion of how the NRS*f*M methods we evaluate, c.f. Table 1, span the state of the art.

2.1 Deformable Models

The description of our taxonomy will start with the underlying structure deformation model category, divided into statistical and physical based models.

2.1.1 Statistical

This set of algorithms apply a statistical deformation model with no direct connection to the physical process of structure deformations. They are in general heuristically defined a priori to enforce constraints that can reduce the ill-posedness of the NRS*f*M problem. The most used low-rank model in the NRS*f*M literature falls into this category, utilizing the assumption that 3D deformations are well described by linear subspaces (also called basis shapes). The low-rank model was first introduced almost 20 years ago by Bregler et al. (2000) solving NRS*f*M through the formalisation of a factorization problem, as analogously proposed by Tomasi and Kanade for the rigid case (Tomasi and Kanade 1992). However, strong nonlinear deformations, such as the one appearing in articulated shapes, may drastically reduce the effectiveness of such models. Moreover, the first low-rank model presented in Bregler et al. (2000) acted mainly as a constraint over the spatial distribution of the deforming point cloud and it did not restrict the temporal variations of the deforming object.

Differently, Gotardo and Martinez (2011a) had the intuition to use the very same DCT bases to model camera and deformation motion instead, assuming those factors are smooth in a video sequence. This approach was later expanded on by explicitly modeling a set of complementary rank-3 spaces, and to constrain the magnitude of deformations in the basis shapes (Gotardo and Martinez 2011c). An extension of this framework, increased the generalization of the model to non-linear deformations, with a kernel transformation on the 3D shape space using radial basis functions (Gotardo and Martinez 2011b). This switch of perspective addressed the main issue of increasing the number of available DCT bases, allowing more diverse motions, while not restricting the complexity of deformations. Later, further extension and optimization have been made to low-rank and DCT based approaches. Valmadre and Lucey (2012) noticed that the trajectory should be a low-frequency signal, thus laying the ground for an automatic selection of DCT basis rank via penalizing the trajectory's response to one or more high-pass filters. Moreover, spatio-temporal constraints have been imposed both for temporal and spatial deformations (Akhter et al. 2012).

A related idea proposed by Li et al. (2018) attempts at grouping recurrent deformations in order to better describe deformations. At its core, the method has an additional clustering step that links together similar deformations. Recently a new prior model, related to the Kronecker–Markov structure of the covariance of time-varying 3D point, very well generalizes several priors introduced previously (Simon et al. 2017). Another recent improvement is given by Dawud Ansari et al. (2017) usage of DCT basis in conjunction with singular value thresholding for camera pose estimation.

Similar spatial and temporal priors have been introduced as regularization terms while optimizing a cost function solving for the NRS*f*M problem, mainly using a low-rank model only. Torresani et al. (2008) proposed a probabilistic PCA model for modelling deformations by marginalizing some of the variables, assuming Gaussian distributions for both noise and deformations. Moreover, in the same framework, a linear dynamical model was used to represent the deformation at the current frame as a linear function of the previous. Brand and Bhotika (2001) penalizes deformations over the mean shape of the object by introducing sensible parameters over the degree of flexibility of the shape. Del Bue et al. (2005a) instead compute a more robust non-rigid factorization, using a 3D mean shape as a prior for NRS*f*M (Del Bue 2013). In a non-linear optimization framework, Olsen and Bartoli (2008) include l_2 penalties both on the frame-by-frame deformations and on the closeness of the reconstructed points in 3D given their 2D projections. Of course, penalty costs introduce a new set of hyper-parameters that weights the terms, implying the need for further tuning, that can be impracticable when cross-validation is not an option. Regularization has also been introduced in formulations of Bundle Adjustment for NRS*f*M (Aanæs and Kahl 2002) by including smoothness deformations via l_2 penalties mainly (Del Bue et al. 2007) or constraints over the rigidity of pre-segmented points in the measurement (Del Bue et al. 2006).

Another important statistical principal is enforcing that low-rank bases are independent. In the coarse to fine approach of Bartoli et al. (2008), base shapes are computed sequentially by adding the basis, which explains most of the variance in respect to the previous ones. They also impose a stopping criteria, thus, achieving the automatic computation of the overall number of bases. The concept of basis independence clearly calls for a statistical model close to independent component analysis (ICA). To this end, Brandt et al. (2011) proposed a prior term to minimize the mutual information of each basis in the NRS*f*M model. Low-rank models are indeed compact but limited in the expressiveness of complex deformations, as noted in Zhu et al. (2014). To solve this problem, Zhu et al. (2014) use a temporal union of subspace that associate at each cluster of frames in time a specific subspace. Such association is solved by adopting a cost function promoting self-expressiveness (Elhamifar and Vidal 2013). Similarly, both spatial and temporal union of subspaces was used also to account for independently deforming multiple shapes (Agudo and Moreno-Noguer 2017a; Kumar et al. 2017). Interestingly, such union of subspaces strategy was previously adopted to solve for the multi-body 3D reconstruction of independently moving objects (Zappella et al. 2013). Another option is to use an over-complete representation of subspaces that can still be used by imposing sparsity over the selected bases (Kong and Lucey 2016). In this way, 3D shapes in time can have a compact representation, and they can be theoretically characterized as a block sparse dictionary learning problem. In a similar spirit, Hamsici et al. (2012) propose to use the input data for learning spatially smooth shape weights using rotation invariant kernels.

All these approaches for addressing NRS*f*M with a low-rank model have provided several non-linear optimization procedures, mainly using alternating least squares (ALS), Lagrange multipliers and alternating direction method of multipliers (ADMM). Torresani et al. first proposed to alternate between the solution of camera matrices, deformation parameters and basis shapes. This first initial solution was then extended by Wang et al. (2008) by constraining the camera matrices to be orthonormal at each iteration, while Paladini et al. (2012) strictly enforced the matrix manifold of the camera matrices to increase the chances to converge to the global optimum of the cost function. All these methods were not designed to be strictly convergent, for this reason, a bilinear augmented multiplier method (BALM) (Del Bue et al. 2012) was introduced to be convergent while implying all the problems constraints being satisfied. Furthermore, robustness in terms of outlying data was then included to improve results in a proximal method with theoretical guarantees of convergence to a stationary point (Wang et al. 2015).

Despite the non-linearity of the problem, it is possible to relax the rank constraint with the trace norm and solve the problem with convex programming. Following this strategy, Dai et al. (2014) provided one of the first effective closed form solutions to the low-rank problem. Although their convex solution, resulting from relaxation, did not provide the best performance, a following iterative optimization scheme gave improved results. In this respect, Kumar et al. (2017) proposed a further improvement on their previous approach, where deformations are represented as a spatio-temporal union of subspaces rather than a single subspace. Thus complex deformation can be represented as the union of several simple ones as already described in the previous paragraphs. To notice that evaluation is performed with synthetic generated data only.

Later Kumar (2020) proposed a set of improvements over Dai et al. approach 2014. Namely, metric rectification was performed using incomplete information by choosing arbitrarily a triplet of solutions among the one available. The solution in Kumar (2020) proposes a method to select the best among the available triplets using a rotation smoothness heuristic as a decision criteria. Then, a further improvement is algorithmic. Instead of using Dai et al. strategy with a matrix shrinkage operator that equally penalizes all the singular values, the method in Kumar (2020) introduces a weighted nuclear norm function during optimisation. More recently Ornhag and Olsson (2020) proposed a unified optimization framework for low-rank inducing penalties that can be readily applied to solve for NRS*f*M. The main advantage of the approach is the ability to combining bias reduction in the estimation and nonconvex low-rank inducing objectives in the form of a weighted nuclear norm.

On the one hand, the procrustean normal distribution (PND) model was proposed as an effective way to implicitly separate rigid and non-rigid deformations (Lee et al. 2017; Park et al. 2018). This separation provides a relevant regularization, since rigid motion can be used to obtain a more robust camera estimation, while deformations are still sampled as a normal distribution as done similarly previously (Torresani et al. 2008). Such a separation is obtained by enforcing an alignment between the reconstructed 3D shapes at every frame. This should in practice factor out the rigid transformations from the statistical distribution of deformations. The PND model has been then extended to deal with more complex deformations and longer sequences (Cho et al. 2016).

2.1.2 Physical

Physical models represent a less studied class wrt. NRSfM, which should ideally be the most accurate for modelling NRS*f*M. Of course, applying the right physical model requires a knowledge of the deformation type and object material, which is information not readily available a priori.

A first class of physical models assume that the non-rigid object is a piecewise partition into parts, i.e. a collection of pre-defined or estimated patches that are mostly rigid or slightly deformable. This observation is certainly true for objects with articulated deformations, as it naturally models natural and mechanical shapes connected into parts. One of the first approaches to use this strategy is given by Varol et al. (2009). By preselecting a set of overlapping patches from the 2D image points, and assuming each patch is rigid, homography constraints can be imposed at each patch, followed by global 3D consistency being enforced using the overlapping points. However, the rigidity of a patch, even if small, is a very hard constraint to impose and it does not generalise well for every non-rigid shape. Moreover, dense point-matches over the image sequence are required to ensure a set of overlapping points among all the patches. A relaxation to the piece-wise rigid constraint was given by Fayad et al. (2010), assuming each patch deforming with a quadratic physical model, thus, accounting for linear and bending deformations. These methods all require an initial patch segmentation and the number of overlapping points, to this end, Russell et al. (2011) optimize the number of patches and overlap by defining an energy based cost function. This approach was further extended and generalised to deal with general videos (Russell et al. 2014) and energy functional that includes temporal smoothing (Golyanik et al. 2019). The method of Lee et al. (2016) instead use 3D reconstructions of multiple combinations of patches and define a 3D consensus between a set of patches. This approach provides a fast way to bypass the segmentation problem and robust mechanism to prune out wrong local 3D reconstructions. The method was further improved to account for higher degrees of missing data in the chosen patches so to generalise better the capabilities of the approach in challenging NRS*f*M sequences (Cha et al. 2019).

Differently from these approaches, Taylor et al. (2010) constructs a triangular mesh, connecting all the points, and considering each triangle as being locally rigid. Global consistency is here imposed to ensure that the vertexes of each triangle coincide in 3D. Again, this approach is to a certain extent similar to Varol et al. (2009), which requires a dense set of points in order to comply with the local rigidity constraint.

A strong prior, which helps dramatically to mitigate the ill-posedness of the problem, is obtained by considering the deformation isometric, i.e. the metric length of curves does not change when the shape is subject to deformations (e.g. paper and metallic materials to some extent). A first solution considering a regularly sampled surface mesh model was presented in Salzmann et al. (2007). Using an assumption that a surface can be approximated as infinitesimally planar, Chhatkuli et al. (2014) proposed a local method that frame NRS*f*M as the solution of partial differential equations (PDE) being able to deal with missing data as well. As a further update (Parashar et al. 2017) formalizes the framework in the context of Riemannian geometry, which led to a practical method for solving the problem in linear time and scaling for a relevant number of views and points. Furthermore, a convex formulation for NRS*f*M with inextensible deformation constraints was implemented using second-order cone programming (SOCP), leading to a closed form solution to the problem (Chhatkuli et al. 2018). Vicente and Agapito (2012) implemented soft inextensibility constraints in an energy minimization framework, e.g. using recently introduced techniques for discrete optimization.

Another set of approaches try to directly estimate the deformation function using high order models. Del Bue and

Bartoli (2011) extended and applied 3D warps such as the thin plate spline, to the NRS*f*M problem. Starting from an approximate mean 3D reconstruction, the warping function can be constructed and the deformation at each frame can be solved by iterating between camera and 3D warp field estimation. Finally, Agudo et al. (2016) introduced the use of finite elements models (FEM) in NRS*f*M. As these models are highly parametrized, requiring the knowledge of the material properties of the object (e.g. the Young modulus), FEM needs to be approximated in order to be efficiently estimated, however, in ideal conditions it might achieve remarkable results, since FEM is a consolidated technique for modelling structural deformations. Lately, Agudo and Moreno-Noguer (2017b) presented a duality between standard statistical rank-constrained model and a new proposed force model inspired from the Hooke's law. However, in principle, their physical model can account for a wider range of deformations than rank-based statistical approaches.

2.2 Missing Data

The initial methods for NRS*f*M assumed complete 2D point matches among views when observing a deformable object. However, given self and standard occlusions, this is rarely the case. Most approaches for dealing with such missing data in NRS*f*M were framed as a matrix completion problem, i.e. estimate the missing entries of the matrix storing the 2D coordinates obtained by projecting each deforming 3D point.

Torresani et al. (2001) first proposed removing rows and lines of the matrix corresponding to missing entries in order to solve the NRS*f*M problem. However, this strategy suffers greatly from even small percentages of missing data, since the subset of completely known entries can be very small. Most of the iterative approaches indeed include an update step of the missing entries (Paladini et al. 2012; Del Bue et al. 2012) where the missing entries become an explicit unknown to estimate. Gotardo and Martinez (2011a) instead strongly reduce the number of parameters by estimating only the camera matrix explicitly under severe missing data. This variable reduction is known as VARPRO in the optimization literature. It has been recently revisited in relation to several structure from motion problems (Hyeong Hong et al. 2017).

2.3 Camera Model

Most NRS*f*M methods in the literature assume a weak perspective camera model. However, in cases where the object is close to the camera and undergoing strong changes in depth, time-varying perspective distortions can significantly affect the measured 2D trajectories.

As low-rank NRS*f*M is treated as a factorization problem, a straightforward extension is to follow best practices from rigid SfM for perspective camera. Xiao and Kanade (2005) have developed a two step factorization algorithm for reconstruction of 3D deformable shapes under the full perspective camera model. This is done using the assumption that a set of basis shapes are known to be independent. Vidal and Abretske (2006) have also proposed an algebraic solution to the non-rigid factorization problem. Their approach is, however, limited to the case of an object being modelled with two independent basis shapes and viewed in five different images. Wang et al. (2007) proposed a method able to deal with the perspective camera model, but under the assumption that its internal calibration is already known. They update the solutions from a weak perspective to a full perspective projection by refining the projective depths recursively, and then refine all the parameters in a final optimization stage. Finally, Hartley and Vidal (2008) have proposed a new closed form linear solution for the perspective camera case. This algorithm requires the initial estimation of a multifocal tensor, which the authors report is very sensitive to noise. Lladó et al. (2006) and Lladó et al. (2010) proposed a non-linear optimization procedure. It is based on the fact that it is possible to detect nearly rigid points in the deforming shape, which can provide the basis for a robust camera calibration.

2.4 Evaluated Methods

We have chosen a representative subset of the aforementioned methods, which are summarized according to our taxonomy in Table 1. This gives us a good representation of recent works, distributed according to our taxonomy with a decent span of deformation models (statistical/physical) and camera models (orthographic, weak perspective or perspective). This also takes into account in-group variations such as DCT basis for statistical deformation and isometry for physical deformation. Even lesser used priors, such as compressibility, are represented. While this is not a full factorial study, we think this reasonably spans the recent state of the art of NRS*f*M. Our choice has, of course, also been influence by method availability, as we want to test the author's original implementation, to avoid our own implementation bias/errors. All in all, we have included 18 methods in our evaluation.

Note that we have chosen not to include the method of Taylor et al. (2010), even if code is available, the approach failed approximately two thirds of the time when tested on our data set.

3 Dataset

As stated, in order to compare state of the art methods for NRS*f*M, we have compiled a larger data set for this purpose. Even though there is a lack of empirical evidence w.r.t. NRS*f*M, it does not imply, that no data sets for NRS*f*M exist.

Table 1 Methods included in our NRS*f*M evaluation with annotations of how they fit into our taxonomy

Method	Citation	Deformable model	Camera model	Missing data
BALM	Del Bue et al. (2012)	Statistical	Orthographic	Yes
Bundle	Del Bue et al. (2007)	Statistical	Weak perspective	Yes
Compressible	Kong and Lucey (2016)	Statistical	Weak perspective	–
Consensus	Lee et al. (2016)	Physical	Orthographic	–
CSF	Gotardo and Martinez (2011a)	Statistical	Weak perspective	Yes
CSF2	Gotardo and Martinez (2011c)	Statistical	Orthographic	Yes
EM PPCA	Torresani et al. (2008)	Statistical	Weak perspective	Yes
KSTA	Gotardo and Martinez (2011b)	Statistical	Orthographic	Yes
MDH	Chhatkuli et al. (2018)	Physical	Perspective	Yes
MetricProj	Paladini et al. (2012)	Statistical	Orthographic	Yes
MultiBody	Kumar et al. (2017)	Statistical	Orthographic	–
PTA	Akhter et al. (2011)	Statistical	Orthographic	–
RIKS	Hamsici et al. (2012)	Statistical	Orthographic	–
ScalableSurface	Dawud Ansari et al. (2017)	Statistical	Orthographic	Yes
SoftInext	Vicente and Agapito (2012)	Physical	Perspective	Yes
SPFM	Dai et al. (2014)	Statistical	Orthographic	–
CMDR	Golyanik et al. (2019)	Physical	Orthographic	–
F-consensus	Cha et al. (2019)	Physical	Orthographic	Yes

As an example in Lee et al. (2016), Gotardo and Martinez (2011a, b, c), Kumar et al. (2017), Akhter et al. (2011), Hamsici et al. (2012) and Dai et al. (2014), a combination of two data sets are used. Namely seven sequences of a human body from the CMU motion capture database (University 2002), two MoCap sequences of a deforming face (Torresani et al. 2004; Del Bue et al. 2005b), a computer animated shark (Torresani et al. 2004) and a challenging flag sequence (Fayad et al. 2010). To the best of our knowledge, this list in Table 2 represents the most used evaluation data sets for NRS*f*M with available ground truth.

The CMU data set (University 2002) captures the motion of humans. Since the other frequently used data sets are also related to animated faces (Torresani et al. 2004; Del Bue et al. 2005b), this implies that there is a high over representation of humans in this state of the art and that a higher variability in the deformed scenes viewed is deemed beneficial. In addition, the shark sequence (Torresani et al. 2004) is not based on real images and objects but on computer graphics and pure simulation. As such, there is a need for new data sets, with reliable ground truth or reference data,[1] and a higher variability in the objects and deformations used.

As such, we here present a data set consisting of five widely different objects/scenes and deformations. The physical object motions are generated mechanically using animatronics, therefore assuring experimental repeatability. Furthermore, we have defined six different camera motions using orthographic and full perspective camera models. This setup, all in all, gives 60 different sequences organized in a factorial experimental design, thus, enabling a more stringent statistical analysis. In addition to this, since we have tight 3D surface models of our objects or scenes, we are able to determine occlusions of all 2D feature points. This in turn gives a realistic handling of missing data, which is often due to object self occlusion. Given this procedure of generating occlusions, missing data always follow a more realistic structured pattern in contrast with the most common, and unrealistic, random process of removing 2D measurement entries used in previous evaluation dataset.

As indicated, these data sets are achieved by stop-motion using mechanical animatronics. These are recorded in our robotic setup, Fig. 1, previously used for generating high quality data sets c.f. e.g. Aanæs et al. (2016). We will here present details of our data capture pipeline, followed by a brief outline and discussion of design considerations.

The goal of the data capturing is to produce 3 types of related data:

Ground Truth: A series of 3D points that change over time.
Input Tracks: 2D tracks used as input.
Missing Data: Binary data indicating the tracks that are occluded at specific image frames.

We record the step-wise deformation of our animatronics from K static views, obtaining both image data and dense 3D

[1] With real measurements like ours the ’ground truth’ data also include noise, why ’reference data’ is a more correct term.

Table 2 A description of the previous data set sequences with available ground truth

Name	Citation	Frames × points	Type	Shape
shark	Torresani et al. (2008)	240 × 91	Synthetic	Animal motion
face1	Torresani et al. (2008)	74 × 37	Mocap	Face motion
face2	Torresani et al. (2008)	316 × 40	Mocap	Face motion
cubes	Xiao et al. (2006)	200 × 14	Synthetic	ToyProblem
face_occ	Paladini et al. (2012)	70 × 37	Mocap	Face motion
flag	Fayad et al. (2010)	540 × 50	Mocap	Cloth deformation
yoga	Akhter et al. (2011)	307 × 41	Mocap	Human motion
drink	Akhter et al. (2011)	1102 × 41	Mocap	Human motion
stretch	Akhter et al. (2011)	307 × 41	Mocap	Human motion
dance	Akhter et al. (2011)	264 × 41	Mocap	Human motion
pickup	Akhter et al. (2011)	357 × 41	Mocap	Human motion
walking	Akhter et al. (2011)	260 × 41	Mocap	Human motion
capoeira	Gotardo and Martinez (2011a)	250 × 41	Mocap	Human motion
jaws	Gotardo and Martinez (2011a)	321 × 49	Synthetic	Animal motion

The table shows the number of frames and points, the way to generate the sequence (mainly with motion capture data) and the type of shape used

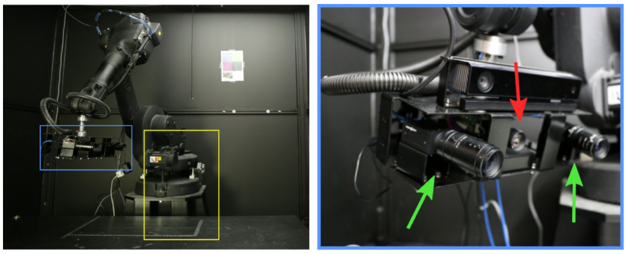


Fig. 1 Images of the robot cell for dataset acquisition. Left image shows the robot with the structured light scanner (blue box) and the area where the animatronic systems are positioned (yellow box). Right image shows the structured light scanner up close, green arrows show the position of the PointGrey Grasshopper3 cameras, and the red arrow marks the Lightcrafter 4500 projector (Color figure online)

surface geometry. We obtain 2D point features by applying standard optical flow tracking (Bouguet 2001) to the image sequence obtained from each of the K views, which is then reprojected onto the recorded surface geometry. The ground truth is then the union of these 3D tracks. By using optical flow for tracking instead of MoCap markers, we obtain a more realistic set of ground truth points. We create input 2D points by projecting the recorded ground truth using a virtual camera in a fully factorial design of camera paths and camera models.

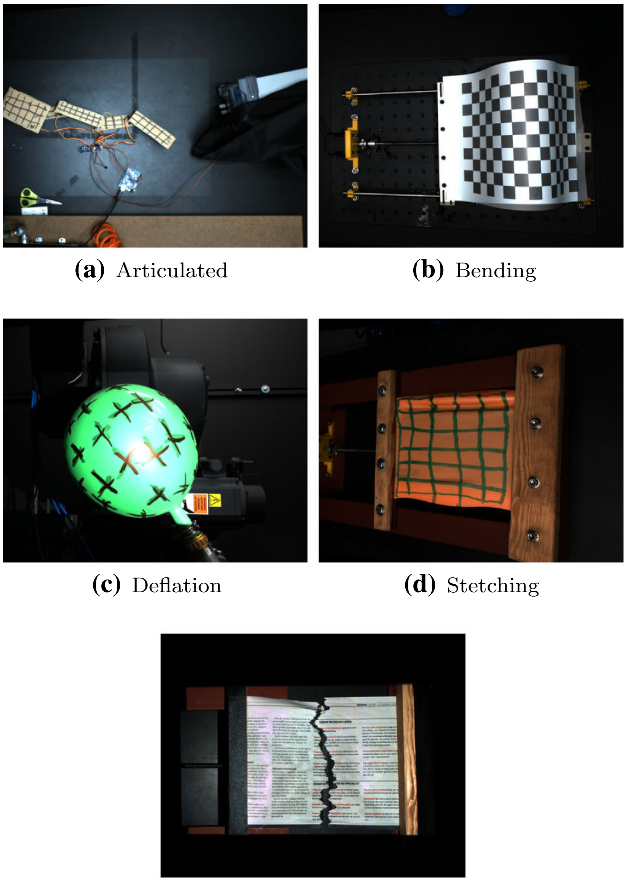
In the following, we will detail some of the central parts of the above procedure.

3.1 Animatronics & Recording Setup

Our stop-motion animatronics are five mechatronic devices capable of computer controlled gradual deformation. They are shown in Fig. 2, and they cover five types of deformations: Articulated Motion, Bending, Deflation, Stretching, and Tearing. We believe this covers a good range of interesting and archetypal deformations. It is noted, that NRS*f*M has previously been tested on bending and tearing (Taylor et al. 2010; Vicente and Agapito 2012; Chhatkuli et al. 2018; Lee et al. 2016), but without ground truth for quantitative comparison. Additionally, elastic deformations, like deflation and stretching, are quite commonplace but did not appear in any previous data sets, to the best of our knowledge.

The animatronics can hold a given deformation or pose for a large extent of time, thus, allowing us to record accurately the object's geometry. We, therefore, do not need a real-time 3D scanner or elaborate multi-scanner setup. Instead, our recording setup consists of an in-house built structured light scanner mounted on an industrial robot as shown in Fig. 1. This does not only provide us with accurate 3D scan data, but the robot's mobility also enables a full scan of the object at each deformation step.

The structured light scanner utilizes two PointGrey Grasshopper3 9.1MP CCD cameras and a projector WinTech Lightcrafter 4500 Pro projecting patterns onto the scene and acquiring images. Then, we use the Heterodyne Phase Shifting method (Reich et al. 1997) to compute the point clouds using 16 periods across the image and 9 shifts. We verified precision according to standard VDI 2634-2 (Deutsches Institut für Normung 2012), and found that the scanner has a form error of [0.01, 0.32 mm], a sphere distance error of [−0.33, 0.50 mm] and a flatness error of [0.29, 0.56 mm]. This is approximately 2 orders of magnitude better than the results we see in our evaluation of the NRS*f*M methods.



(a) Articulated **(b)** Bending

(c) Deflation **(d)** Stetching

(e) Tearing

Fig. 2 Animatronic systems used for generating specific types of non-rigid motion

3.2 Recording Procedure

The recording procedure acquires for each shape a series of image sequences and surface geometries of its deformation over F frames. We record each frame from K static views with our aforementioned structured light scanner. As such we obtain K image sequences with F images in each. We also obtain F dense surface reconstructions, one for each frame in the deformation. The procedure is summarized in pseudo code in Algorithm 1. Figure 3 illustrates sample images of three views obtained using the above process.

3.3 3D Ground Truth Data

The next step is to take acquired images $I_{f,k}$ and surfaces S_f, and extract the ground truth points. We do this by applying optical flow tracking (Bouguet 2001) as implemented in OpenCV 2.4 to obtain 2D tracks, which are then reprojected onto S_f. The union of these reprojected tracks gives us the ground truth, Q. This process is summarized in pseudo code in Algorithm 2.

Algorithm 1: Process for recording image data for tracking and dense surface geometry for an animatronic.

```
Let $F$ be the number of frames
Let $k$ be the number of static scan views $K$
for $f \in F$ do
    Deform animatronic to pose $f$
    for $k \in K$ do
        Move scanner to view $k$
        Acquire image $I_{f,k}$
        Acquire structured light scan $S_{f,k}$
    end
    Combine scans $S_{f,k}$ for full, dense surface $S_f$
end
```

Algorithm 2: Process for extracting the ground truth Q from recorded images and surface scans.

```
Let $F$ be the number of frames
Let $k$ be the number of static scan views $K$
Let $S_f$ be the surface at frame $f$
Let $I_{f,k}$ be the image from view $k$, frame $f$
$S = \{S_1 \dots S_F\}$
for $k \in K$ do
    $I_k = \{I_{1,k} \dots I_{F,k}\}$
    Apply optical flow (Bouguet 2001) to $I_k$ to get 2D tracks $T_k$
    Reproject $T_k$ onto $S$ to get 3D tracks $Q_k$
end
$Q = \{Q_1 \dots Q_K\}$
```

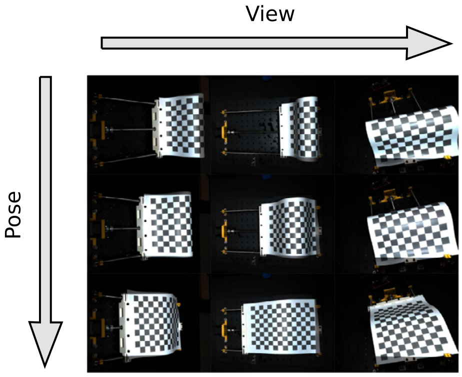


Fig. 3 Illustrative sample of our multi-view, stop-motion recording procedure. Animatronic pose evolves vertically and scanner view change horizontally

3.4 Projection using a Virtual Camera

To produce the desired input, we project the ground truth **Q** using a virtual camera, similar to what has been done in Lee et al. (2016), Gotardo and Martinez (2011a), Dai et al. (2014), and Del Bue et al. (2005b). This step has two factors related to the camera that we wish to control for: Path and camera model. To keep our design factorial, we define six different camera paths, which will all be used to create the 2D input. They are illustrated in Fig. 4. We believe these are a good representation of possible camera motion with both linear

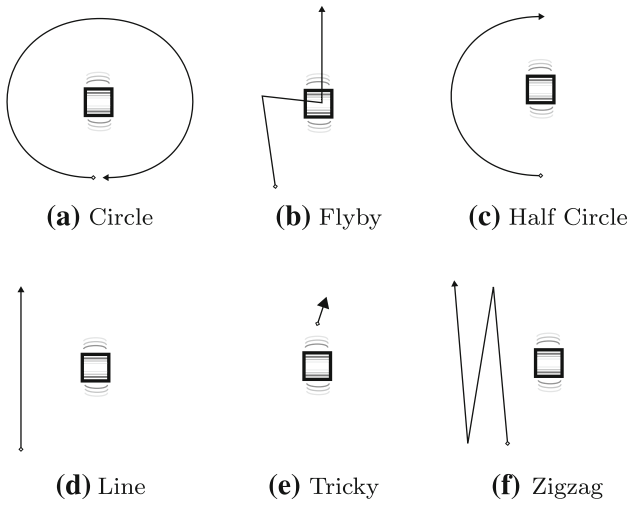


Fig. 4 Camera path taxonomy. The box represents the deforming scene and the wiggles illustrates the main direction of deformation, e.g. the direction of stretching

Algorithm 3: Creation of input tracks $W_{c,p}$ and missing data $D_{c,p}$ from ground truth Q for each combination of camera path p and model c.

```
Let $F$ be the number of frames
Let $P$ be the set of camera paths shown in Fig. 4
Let $C$ be either perspective or orthographic
Let $Q_f$ be the ground truth at frame $f$
Let $S_f$ be the surface at frame $f$
for $S_f \in \{S_1 \dots S_F\}$ do
    Estimate mesh $M_f$ from $S_f$
end
for $c \in C$ do
    for $p \in P$ do
        for $f \in F$ do
            Set camera pose to $p_f$
            Project $Q_f$ using model $c$ to get points $w_f$
            Do occlusion test $q_f$ against $M_f$ to get missing data
              $d_f$
        end
        $W_{c,p} = \{w_1 \dots w_F\}$
        $D_{c,p} = \{d_1 \dots d_F\}$
    end
end
```

motion and panoramic panning. The Circle and Half Circle paths correspond well to the way scans are performed in SfM and structured light methods: By moving around the target object we try to cover most of its shape. Line and Flyby are to simulate a scenario where instead the camera move linearly as in the automotive and drone-alike movements respectively. Zigzag and Tricky motions are about having depth variations in the camera movement, which is important for perspective camera, where each frame will have different projective distortions. Tricky camera path resembles more a critical motion in the direction of the optical ray of the camera as expected, for instance, in medical imaging. To conclude, as mentioned earlier, the camera model can be either orthographic or perspective.

The factorial combination of these elements yields to 12 input sequences for each ground truth. Additionally, as we have previously recorded the dense surface for each frame (see Sect. 3.2), we estimate missing data via self-occlusion. Specifically, we create a triangular mesh for each S_f and estimate occlusion via raycasting into the camera along the projection lines. Vertices whose ray intersects a triangle on the way to the camera are removed, from the input for the given frame, as those vertices would naturally be occluded. In this way, we ensure as realistic as possible structured missing data by modelling self-occlusion given the different camera paths. This process is summarized in pseudo code in Algorithm 3.

3.5 Discussion

While stop-motion does allow for diverse data creation, it is not without drawbacks. Natural acceleration is easily lost when objects deform in a step-wise manner and recordings are unnaturally free of noise like motion blur. However, without this technique, it would have been prohibitive to create data with the desired diversity and accurate 3D ground truth.

The same criticism could be levied against the use of a virtual camera, it lacks the shakiness and acceleration of a real world camera. On the other hand, it allows us to precisely vary both the camera path and camera model. This enables us to perform a factorial analysis, in which we can study the effects of different configurations on NRS*f*M. As we show in Sec. 5 some interesting conclusions are drawn from this analysis. Most NRS*f*M methods are designed with an orthographic camera in mind. As such investigating the difference between data under orthographic and perspective projection is of interest. Such an investigation is only practically possible using a virtual camera.

4 Evaluation Metric

In order to compare the methods of Table 1 w.r.t. our data set, a metric is needed. The purpose is to project the high dimensional 3D reconstruction error into (ideally) a one dimensional measure.

Several different metrics have been proposed for NRS*f*M evaluation in the past literature, e.g. the Frobenius norm (Paladini et al. 2009), mean (Hamsici et al. 2012), variance normalized mean (Gotardo and Martinez 2011c) and RMSE (Taylor et al. 2010).

All of the above mentioned evaluation metrics are based on the $L2$-norm in one form or another. A drawback of the $L2$-norm is its sensitive to large errors, often letting a few outliers dominating the evaluation. To address this, we incorporate robustness into our metric, by introducing truncation of the individual 3D point reconstruction errors. In particular, our metric is based on a RMSE measure similar used in Taylor et al. (2010).

Given the visualisation effectiveness and general adoption of box plots (Velleman and Hoaglin 1981), we propose to use their whisker function to identify and to model outliers in the error distribution. Such a strategy will enable the inclusion of outliers in the metric with the additional benefit of reducing their influence in the RMSE. Consider E being the set of point-wise errors ($||\mathbf{X}_{f,p} - \mathbf{Q}_{f,p}||$) and E_1, E_3 as the first and third quartile of that set. As described in Williamson et al. (1989), we define the whisker as $w = \frac{3}{2}(E_3 - E_1)$, then any point that is more than a whisker outside of the interquantile range ($IQR = E_3 - E_1$) is considered as an outlier. Those outliers are then truncated at $E_3 + w$ allowing them to be included in a RMSE without dominating the result. This strategy works well for approximately normally distributed data. With this in mind, our truncation function is defined as follows,

$$t(\mathbf{x}, \mathbf{q}) = \begin{cases} ||\mathbf{x} - \mathbf{q}||, & ||\mathbf{x} - \mathbf{q}|| < E_3 + w \\ E_3 + w, & \text{otherwise} \end{cases} \tag{2}$$

Thus the robust RMSE is defined as,

$$m(\mathbf{Q}, \mathbf{X}) = \sqrt{\frac{1}{FP} \sum_{f,p}^{F,P} t\left(\mathbf{X}_{f,p}, \mathbf{Q}_{f,p}\right)}. \tag{3}$$

A NRS*f*M reconstruction is given in an arbitrary coordinate system, thus we must align the reference and reconstruction before computing the error metric. This is typically done via Procrustes Analysis (Gower 1975), but as it minimizes the distance between two shapes in a $L2$-norm sense it is also sensitive to outliers. Therefore, we formulate our alignment process as an optimization problem based on the robust metric of Eq. 3. Thus the combined metric and alignment is given by,

$$\begin{aligned} m(\mathbf{X}, \mathbf{Q}) &= \min_{s,\mathbf{R},\mathbf{t}} \sqrt{\frac{1}{FP} \sum_{f,p} t\left(s\left[\mathbf{R}\mathbf{X}_{fp} + \mathbf{t}\right], \mathbf{Q}_{fp}\right)}, \\ \text{where } s &= \text{scale}, \\ \mathbf{R} &= \text{rotation and reflection}, \\ \mathbf{t} &= \text{translation}. \end{aligned} \tag{4}$$

An implication of using a robust, as opposed to a $L2$-norm, is that the minimization problem of (4) cannot be achieved by a standard Procrustes alignment, as done in Taylor et al. (2010). As such, we optimize (4) using the Levenberg-Marquardt method, where s, $\mathbf{R}$ and $\mathbf{t}$ have been initialized via Procrustes alignment (Gower and Dijksterhuis 2004). In summary, (4) defines the alignment and metric that has been used for the evaluation presented in Sect. 5.

Notice also that this registration procedure estimates a single rotation and translation for the entire sequence. In this way, we avoid the practise of registering the GT 3D shape at every frame of the reconstructed 3D sequence. Such frame-by-frame procedure does not account for the global temporal consistency of the reconstructed 3D sequence and in particular regarding possible sign flips of the 3D shape, scale variations, or reflections that might happen abruptly from one frame to the other during reconstruction. Registering the 3D ground truth frame-by-frame is also unrealistic, because in general, it is not feasible to do in a real operative reconstruction scenario where 3D GT is not available.

To conclude, the choice of an evaluation metric always has a streak of subjectivity and for this reason, we investigated the sensitivity of choosing a particular one. We did this by repeating our evaluation with another robust metric, where the minimum track-wise distance between the ground truth and reconstruction was used. By just using the n-th percentile, instead of our truncation, the magnitude of the RMSE significantly decreases, but the major findings and conclusions, as presented in Sect. 5, were the same. As such we conclude that our conclusions are not overly sensitive to the choice of metric.

5 Evaluation

With our data set and robust error metric, we have performed a thorough evaluation and analysis of the state-of-the-art in NRS*f*M, which is presented in the following. This is done in part as an explorative analysis and in part to answer some of what we see as most pressing, open questions in NRS*f*M. Specifically:

- Which algorithms perform the best?
- Which deformable models have the best performance or generalization?
- How well can the state-of-the-art handle data from a perspective camera?
- How well can the state-of-the-art handle occlusion-based missing data?

To answer these questions, we perform our analysis in a factorial manner, aligned with the factorial design of our data set. To do this, we view a NRS*f*M reconstruction as a function of the following factors:

Algorithm a_i: Which algorithm was used.

Camera Model m_j: Which camera model was used (perspective or orthographic).
Animatronics s_k: Which animatronics sequence was reconstructed.
Camera Path p_l: How the camera moved.
Missing Data d_n: Whether occlusion based missing data was used.

We design our evaluation to be almost fully crossed, meaning we obtain a reconstruction for every combination of the above factors.

The only missing part is that the authors of Multi-Body (Kumar et al. 2017) only submitted reconstructions for orthographic camera model.

Our factorial experimental design allows us to employ a classic statistical method known as ANalysis Of VAriance (ANOVA) (Seber and Lee 2012). The ANOVA not only allow us to deduce the precise influence of each factor on the reconstruction but also allows for testing their significance. To be specific, we model the reconstruction error in terms of the following bilinear model,

$$\begin{aligned} y =& \mu + a_i + m_j + s_k + p_l + d_n \\ &+ as_{ik} + ap_{il} + ad_{in} + ms_{jk} \\ &+ mp_{jl} + md_{jn} + sp_{kl} + sd_{kn} + pd_{ln}, \end{aligned} \tag{5}$$

where,

$$\begin{aligned} y &= \text{reconstruction error}, \\ \mu &= \text{overall average error}, \\ xy_{i,j} &= \text{interaction term between factor } x_i \text{ and } y_j. \end{aligned}$$

This model, Eq. (5), contains both linear and interaction terms, meaning the model reflects both factor influence as independent and as cross effects, e.g. as_{ik} is the interaction term for ’algorithm’ and ’animatronics’. For each term, we test for significance by choosing between two hypotheses:

$$\begin{aligned} &\mathcal{H}_0 : c_0 = c_1 = \ldots = c_N \\ &\mathcal{H}_1 : c_0 \neq c_1 \neq \ldots \neq c_N \end{aligned} \tag{6}$$

with c_n being a term from (5) e.g. a_i or md_{jn}. Typically, $\mathcal{H}_0$ is referred to as the null hypothesis, meaning the term c_n has no significant effect. ANOVA allows for estimating the probability of falsely rejecting the null hypothesis for each factor. This statistic is referred to as the p-value. A term is referred to as being statistically significant if its p-value is below a certain threshold. In this paper we consider a significance threshold of 0.0005 or approximately 3.5σ. As such, we clearly evaluated which factors are important for NRS*f*M and which are not.

Table 3 ANOVA table for NRS*f*M reconstruction error without missing data with sources as defined in (5)

Factor	Sum sq.	DoF	Mean sq.	F	p-value
a_i	3.6×10^5	15	2.4×10^4	204.8	5.5×10^{-242}
m_j	1.1×10^4	1	1.1×10^4	90.4	3.2×10^{-20}
s_k	1.0×10^5	4	2.6×10^4	219.0	3.6×10^{-121}
p_l	1.5×10^4	5	3.0×10^3	25.6	9.3×10^{-24}
as_{ik}	4.1×10^4	60	6.9×10^2	5.9	2.9×10^{-33}
ap_{il}	4.1×10^4	75	5.5×10^2	4.7	2.3×10^{-28}
ms_{jk}	1.3×10^3	4	3.2×10^2	2.7	0.03
mp_{jl}	1.8×10^3	5	3.6×10^2	3.1	0.0086
sp_{kl}	1.1×10^4	20	5.7×10^2	4.9	2.3×10^{-11}
Error	8×10^4	689	1.2×10^2		
Total	7×10^5	878			

All factors are statistically significant at a 0.0005 level except ms_{jk} and mp_{jl}

Another interesting property of the ANOVA is that all coefficients in a given factor sums to zero,

$$\sum_{i=0}^{N} c_i = 0. \tag{7}$$

So each factor can be seen as adjusting the predicted reconstruction error from the overall average. It should be noted that the “algorithm”/“camera model” interaction am_{ij} has been left out of (5) due to MultiBody (Kumar et al. 2017) only being tested with one camera model.

The error model of (5) is not directly applicable to the error of all algorithms as not all state-of-the-art methods from Table 1 can deal with missing data. As such we perform the evaluation in two parts. One where we disregard missing data and include all available methods from Table 1, and one where we use the subset of methods that handles missing data and utilize the full model of (5). The former is covered in Sect. 5.1 and the latter is covered in Sect. 5.2.

5.1 Evaluation without Missing Data

In the following, we discuss the results of the ANOVA without taking ’missing data’ into account, using the model as in Eq. (5) without terms related to d_n:

$$\begin{aligned} y =& \mu + a_i + m_j + s_k + p_l + as_{ik} \\ &+ ap_{il} + ms_{jk} + mp_{jl} + sp_{kl}. \end{aligned} \tag{8}$$

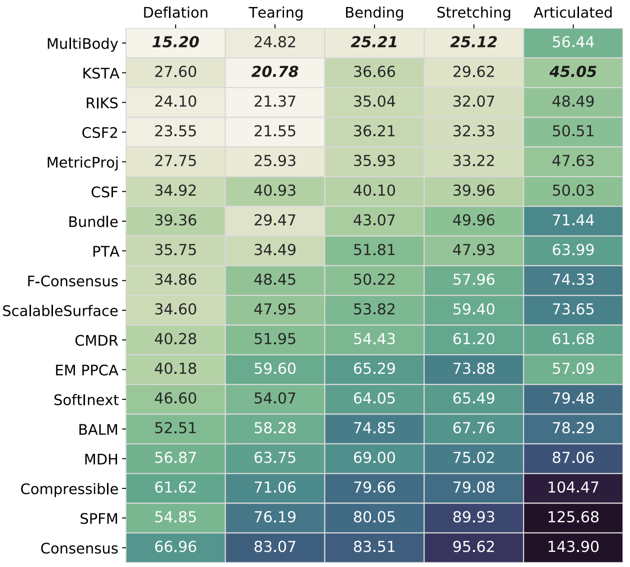
The results of the ANOVA using Eq. (8) is summarized in Table 3. All factors except ms_{jk} and mp_{jl} are statistically significant. As such, we can conclude that all the aforementioned factors have a significant influence on the reconstruction

Table 4 Linear term $\mu + a_i$ sorted in ascending numerical order, this is the average error for the given algorithm

MultiBody	KSTA	RIKS
29.36	31.94	32.21
CSF2	MetricProj	CSF
32.83	34.09	41.19
Bundle	PTA	F-Consensus
46.66	46.80	53.17
ScalableSurface	CMDR	EM PPCA
53.88	53.91	59.21
SoftInext	BALM	MDH
61.94	66.34	70.34
Compressible	SPFM	Consensus
79.18	85.34	94.61

Algorithms are referred to by their alias in Table 1. All numbers are given in millimeters

Table 5 Interaction term $\mu + a_i + s_k + as_{ik}$



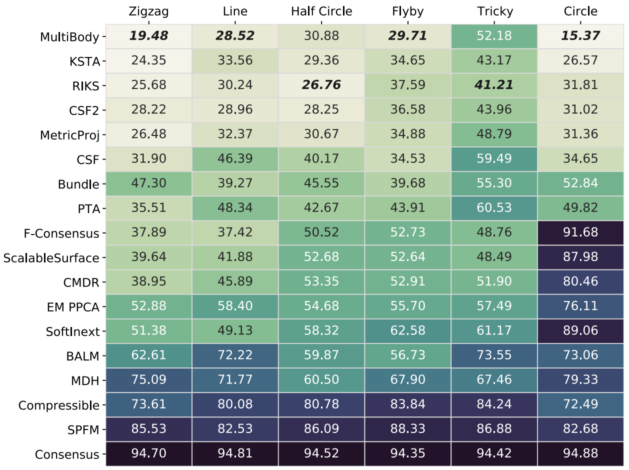
	Deflation	Tearing	Bending	Stretching	Articulated
MultiBody	***15.20***	24.82	***25.21***	***25.12***	56.44
KSTA	27.60	***20.78***	36.66	29.62	***45.05***
RIKS	24.10	21.37	35.04	32.07	48.49
CSF2	23.55	21.55	36.21	32.33	50.51
MetricProj	27.75	25.93	35.93	33.22	47.63
CSF	34.92	40.93	40.10	39.96	50.03
Bundle	39.36	29.47	43.07	49.96	71.44
PTA	35.75	34.49	51.81	47.93	63.99
F-Consensus	34.86	48.45	50.22	57.96	74.33
ScalableSurface	34.60	47.95	53.82	59.40	73.65
CMDR	40.28	51.95	54.43	61.20	61.68
EM PPCA	40.18	59.60	65.29	73.88	57.09
SoftInext	46.60	54.07	64.05	65.49	79.48
BALM	52.51	58.28	74.85	67.76	78.29
MDH	56.87	63.75	69.00	75.02	87.06
Compressible	61.62	71.06	79.66	79.08	104.47
SPFM	54.85	76.19	80.05	89.93	125.68
Consensus	66.96	83.07	83.51	95.62	143.90

This is equivalent to the algorithms average error on each animatronic. Lowest error for each animatronic is marked with bold text. Algorithms are referred to by their alias in Table 1. All numbers are given in millimeters

error. Therefore, we will explore the specifics of each factor in the following, starting with ’algorithm’.

Table 4 shows the average reconstruction error for each algorithm. The method MultiBody (Kumar et al. 2017) has the lowest average reconstruction error over all experiments followed by KSTA (Gotardo and Martinez 2011b) and RIKS (Hamsici et al. 2012). For more detailed insights refer to Table 5 showing the ’algorithm’ vs ’animatronic’ effect on the reconstruction error. As it can be seen, MultiBody (Kumar et al. 2017) does not have the lowest error for all animatronics, as e.g. KSTA (Gotardo and Martinez 2011b) has a significantly lower error on the Tearing and Articulated deformations. Both of these can roughly be described as rigid bodies moving relative to each other, and it would seem KSTA (Gotardo and Martinez 2011b) is the best at handling these deformations.

Table 6 Interaction term $\mu + a_i + p_l + ap_{il}$



	Zigzag	Line	Half Circle	Flyby	Tricky	Circle
MultiBody	***19.48***	***28.52***	30.88	***29.71***	52.18	***15.37***
KSTA	24.35	33.56	29.36	34.65	43.17	26.57
RIKS	25.68	30.24	***26.76***	37.59	***41.21***	31.81
CSF2	28.22	28.96	28.25	36.58	43.96	31.02
MetricProj	26.48	32.37	30.67	34.88	48.79	31.36
CSF	31.90	46.39	40.17	34.53	59.49	34.65
Bundle	47.30	39.27	45.55	39.68	55.30	52.84
PTA	35.51	48.34	42.67	43.91	60.53	49.82
F-Consensus	37.89	37.42	50.52	52.73	48.76	91.68
ScalableSurface	39.64	41.88	52.68	52.64	48.49	87.98
CMDR	38.95	45.89	53.35	52.91	51.90	80.46
EM PPCA	52.88	58.40	54.68	55.70	57.49	76.11
SoftInext	51.38	49.13	58.32	62.58	61.17	89.06
BALM	62.61	72.22	59.87	56.73	73.55	73.06
MDH	75.09	71.77	60.50	67.90	67.46	79.33
Compressible	73.61	80.08	80.78	83.84	84.24	72.49
SPFM	85.53	82.53	86.09	88.33	86.88	82.68
Consensus	94.70	94.81	94.52	94.35	94.42	94.88

Algorithms are referred to by their alias in Table 1. All numbers are given in millimeters

Table 7 Linear term $\mu + m_j$ sorted in ascending numerical order, this is the average error for the given camera model

Orthographic	Perspective
50.45	57.66

All numbers are given in millimeters

Methods with a physical prior, like MDH (Chhatkuli et al. 2018) and SoftInext (Vicente and Agapito 2012) have in general lower performance, as it is evident from Tables 1, 5 and 6. MDH (Chhatkuli et al. 2018) is designed with an isometry prior, therefore one would expect it to perform well in the bending deformation. Indeed, while its interaction term as_{ik} has its lowest value for the bending deformation, denoting the fitness of the chosen prior, the average reconstruction error is higher. On a more careful inspection of the reconstructed 3D sequences, it is evident that for a few frames MDH and SoftInext struggle to obtain an accurate 3D reconstruction and this affects the whole evaluation. Moreover, the 3D reconstruction shows intermittent sign flips of the 3D reconstructed shape. To this end, a stronger temporal consistency may help to reduce this negative effect and improve the method performance.

A similar trend can be observed in Table 6, which shows the ’algorithm’ vs ’camera path’ effect on the reconstruction error. While MultiBody (Kumar et al. 2017) has the lowest average error, it is surpassed in the Half Circle and Tricky ’camera path’ by RIKS (Hamsici et al. 2012). On the other

Table 8 Linear term $\mu + s_k$ sorted in ascending numerical order, this is the average error for the given animatronic

Deflation	Tearing	Bending
39.86	46.32	54.38
Stretching	Articulated	
56.42	73.29	

All numbers are given in millimeters

Table 9 Linear term $\mu + p_l$ sorted in ascending numerical order, this is the average error for the given camera path

Zigzag	Line	Half Circle
47.29	51.21	51.42
Flyby	Tricky	Circle
53.29	59.94	61.18

All numbers are given in millimeters

Table 10 ANOVA table for NRS*f*M reconstruction error with missing data. Factors are as defined in (5) and described at the beginning of this section

Factor	Sum sq.	DoF	Mean sq.	F	p-value
a_i	1.3×10^5	8	1.6×10^4	90.9	7.7×10^{-108}
m_j	1.4×10^4	1	1.4×10^4	81.6	1.2×10^{-18}
s_k	7.5×10^4	4	1.9×10^4	106.5	3.8×10^{-73}
p_l	4.1×10^4	5	8.2×10^3	47.0	8.8×10^{-43}
d_n	1.6×10^4	1	1.6×10^4	89.8	2.7×10^{-20}
as_{ik}	1.6×10^4	32	5.0×10^2	2.9	3.4×10^{-7}
ap_{il}	5.6×10^4	40	1.4×10^3	8.0	6.4×10^{-37}
ad_{in}	1.1×10^4	8	1.3×10^3	7.5	1.1×10^{-9}
ms_{jk}	2.6×10^3	4	6.5×10^2	3.7	0.0052
mp_{jl}	2.5×10^3	5	5.1×10^2	2.9	0.013
md_{jn}	2.9×10^2	1	2.9×10^2	1.6	0.2
sp_{kl}	2.7×10^4	20	1.4×10^3	7.8	6.7×10^{-21}
sd_{kn}	3.6×10^3	4	8.9×10^2	5.1	0.00048
pd_{ln}	8.1×10^3	5	1.6×10^3	9.3	1.4×10^{-8}
Error	1.4×10^5	824	1.8×10^2		
Total	5.7×10^5	962			

All factors are statistically significant at a 0.0005 level except ms_{jk}, mp_{jl} and md_{jn}

hand, MultiBody has the lowest error under the Circle path by quite a significant margin.

From this analysis we can conclude that MultiBody performs the best on average, but is surpassed w.r.t. to certain camera paths and animatronic deformations by algorithms such as RIKS (Hamsici et al. 2012) and KSTA (Gotardo and Martinez 2011b). This also clearly indicates that one needs to control for both deformation type and camera motion in future NRS*f*M comparisons, as the above conclusion could be changed by choosing the right combination of camera path and deformation. On the other hand, these findings show that NRS*f*M performance can be optimized by choosing the right camera path (e.g. Zigzag) and the right algorithm for the deformation in question (Tables 7 and 8).

The camera model and its path have a significant impact on reconstruction error, a trend that can be observed from Table 6.

Table 9 shows that there is a significant difference in average error w.r.t. 'camera path'. It is interesting to note, that the Circle path has one of the highest average errors, only surpassed by the Tricky camera path. The latter was specifically designed to be challenging, as such, it is surprising to find that the Circle and Tricky path's average error only differ by 3.08mm. In fact, MultiBody (Kumar et al. 2017) seems to be the only method that benefits from the circle type of camera path, as can be seen in Table 6. Table 7 shows the average error of reconstructions for an orthographic and a perspective camera model. As it can be seen, there is a difference of 7.20mm, which is significant but not as large as the difference w.r.t. 'algorithm' (Table 4) or 'camera path' (Table 9). This suggests that, while the error increases the state-of-the-art in NRS*f*M can still operate under a perspective camera model. This is quite interesting as most NRS*f*M approaches are not designed with a perspective camera in mind. It would seem that an orthographic or weak-perspective camera acts a reasonable approximation given the perspective distortions and the scale of the object deformation.

There is also a significant difference between the average reconstruction error of each animatronic which Table 8 shows. Articulated has by far the highest average reconstruction error, making it the most difficult to reconstruct for the current state-of-the-art in NRS*f*M. Since most approaches use low-rank methods, a highly structured motion such as an Articulated is difficult to handle with a low-rank prior, especially if points are densely sampled on all joints. On the other hand, Deflation seems to be quite easy to handle for most of the state-of-the-art methods.

5.2 Evaluation with Missing Data

As previously mentioned, we are interested in 'missing data' and its effect on NRS*f*M. We, thus, here use Eq. (5), which is used to evaluate the subset of methods capable of handling missing data, as shown in Table 1.

It should be noted that while MDH (Chhatkuli et al. 2018) is nominally capable of handling missing data, it has not been included in this part of the study. The reason being that the code provided only reconstructs frames with minimum ratio of visible data, thus our error metric cannot be applied. As such, we have 9 methods in total in this category.

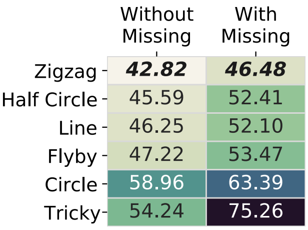
We treat 'missing data' as a categorical factor having two states: with or without missing data. This is because the missing percentage of our occlusion-based missing data is dependent on the 'animatronic', 'camera path' and 'camera model' factors. Additionally, there is a significant sampling bias in the occlusion-based missing data. For example, in-plane motion, like Articulated and Tearing, rarely get a missing percentage above 25% and more volumetric motion such as Deflation rarely go below 40% missing data. This would make it difficult to distinguish between the influence of the 'missing data' factor and the animatronic factor.

The results of the ANOVA is summarized in Table 10 and all factors except ms_{jk}, mp_{jl} and md_{jn} are statistically significant (Tables 11 and 12). This means that 'missing data' has a significant influence on the reconstruction error. Table 13 shows the interaction between 'algorithm' and 'missing data'. As expected, the mean error without missing data is very similar to the averages in Table 4 with KSTA (Gotardo and Martinez 2011b) having the lowest expected error. However, with missing data, MetricProj (Paladini et al. 2012) actually has a lower average reconstruction error. This is due to its low increase in error of 5.85mm when operating under occlusion-based missing data. In comparison, KSTA (Gotardo and Martinez 2011b), CSF2 (Gotardo and Martinez 2011c) and CSF (Gotardo and Martinez 2011a) are much more unstable with average increases in error of 9.65, 18.15 and 13.49 mm respectively. Common among the three methods is the fact that they assume a Discrete Cosine Transform (DCT) as their prior. Indeed, we see a similar increase for ScalableSurface of 16.52 mm and this method also uses a DCT basis.

These results suggest that while DCT-based approaches are quite accurate without missing data, they are not very robust when operating under occlusion-based missing data. Thus, they would likely not be very robust when applied to real-world deformations, where occlusion-based missing data is unavoidable. This indicates that future research should focus on making DCT basis methods more robust or to modify the DCT model to better generalize for 'missing data'. Finally, BALM (Del Bue et al. 2012) method exhibit some peculiar behavior as its average error actually decreases by 3.33mm, contrary to expectation. A likely cause is a different computational structure of the algorithm, since the full data case uses mainly SVD for factorisation while the missing data approach has a more elaborated algorithmic approach with manifold projections and matrix entries imputation.

Table 12 shows the average error as an interaction between 'animatronic' and 'missing data', i.e. the average reconstruction error of each animatronic with and without missing data. It is interesting to note that the in-plane deformations, i.e. Tearing, Stretching and Articulated, generally have a smaller increase in error with missing data compared to the more volumetric deformation, i.e. Deflation and Bending, compared to the error without missing data. The increase is respectively 3.96, 4.65 and 8.38 mm versus 12.27 and 13.47 mm. The main difference between the two groups is that the ratio of missing data is consistently low for the in-plane deformations. This would suggest that the ratio of missing data has an impact on the reconstruction error.

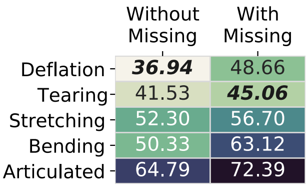
Table 11 Interaction between 'camera path'/'missing data'; $\mu + p_l + d_n + pd_{ln}$



	Without Missing	With Missing
Zigzag	***42.82***	***46.48***
Half Circle	45.59	52.41
Line	46.25	52.10
Flyby	47.22	53.47
Circle	58.96	63.39
Tricky	54.24	75.26

Numbers are given in milimeters

Table 12 Interaction between 'animatronic'/'missing data'; $\mu + s_k + d_n + sd_{kn}$



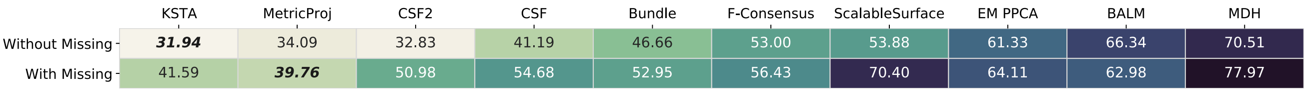
	Without Missing	With Missing
Deflation	***36.94***	48.66
Tearing	41.53	***45.06***
Stretching	52.30	56.70
Bending	50.33	63.12
Articulated	64.79	72.39

Numbers are given in milimeters

Table 11 shows the average error as interaction between 'camera path' and 'missing data'. The Tricky path has by far the highest average error. This is expected, as the small camera movement ensures that a portion of the tracked points is consistently hidden. As such, while Tricky and Circle were almost equally difficult without missing data, this is no longer the case with missing data as Circle's average error only increases by 4.9 mm. Indeed, all other camera paths have approximately the same increase in error with missing data. These paths also ensure that all observed points are equally visible. What differs consistently is the spatio-temporal distribution of missing data, which has a physical plausible structured pattern. the missing data distributions in our dataset are in contrast with previous evaluations where often missing entries were generated randomly, thus not reflecting a real 3D modelling scenario. These results also suggest that the distribution of missing data is as important as the ratio in affecting the reconstruction error. Indeed this is in line with the observations made by Paladini et al. (2012).

The aforementioned observations demonstrate the importance of testing against occlusion-based missing data as it contains a spatio-temporal structure of missing data that a randomly removed subset lacks. Many NRS*f*M methods treat missing data as a matrix fill-in problem, meaning recreating missing values from interpolation of spatio-temporally close observations. Thus, it is clear that conceptually it

Table 13 Interaction between 'algorithm'/'missing data'; $\mu + a_i + d_n + ad_{in}$



	KSTA	MetricProj	CSF2	CSF	Bundle	F-Consensus	ScalableSurface	EM PPCA	BALM	MDH
Without Missing	***31.94***	34.09	32.83	41.19	46.66	53.00	53.88	61.33	66.34	70.51
With Missing	41.59	***39.76***	50.98	54.68	52.95	56.43	70.40	64.11	62.98	77.97

This is the average error for each algorithm either with or without occlusion-based missing data

is much easier to interpolate random, evenly distributed missing data, compared to the spatio-temporally clustered structure of occlusion-based missing data. It is noted, that KSTA (Gotardo and Martinez 2011b) and CSF (Gotardo and Martinez 2011a) were both evaluated using random subset missing data in the original works, and was found to approximately have the same performance whether from 0 to 50% missing data. These results are obviously quite different from the conclusion of our study and we hypothesize, that the spatio-temporal structure of our occlusion-based missing is probably the primary cause for the drop in performance of many approaches.

6 Discussion and Conclusion

To summarize our findings, we would like to firstly mention that, the algorithm with the lowest error on average without missing data was found to be MultiBody (Kumar et al. 2017).

There is, however, a large variation between the different algorithms performance depending on the factors chosen. As such our study does not conclude that Multibody (Kumar et al. 2017) is definitively better than all other methods in general. As an example, for some camera paths RIKS (Hamsici et al. 2012) had lower average error than MultiBody (Kumar et al. 2017). Also, with missing data MetricProj (Paladini et al. 2009) has the lowest reconstruction error. Other observations include that methods with a DCT basis were found to have a great increase in error with occlusion-based missing data. In general, the evaluated methods stay about two orders of magnitude behind the accuracy of the ground truth, showing that there is a need of improving current approaches.

Our study also shows findings that support hypotheses of where NRS*f*M research could head in the future. Even though some of these hypotheses have been stated before in related work, the strength of our data set and evaluation is able to confirm these. Firstly, it is clear that methods using the weak perspective approximation to the perspective camera model only incur a small penalty for doing so on average. This camera model seems like a good approximation, although it should be noted, that our data set does not challenge the algorithms extremely in this regard, with only an average 1.6 fold change in the depth variations. In particular, NRS*f*M applied in the medical domain, e.g. endoscopic imaging, may better benefit from a perspective camera model as the deforming body can be imaged at different depths while approaching with the endoscope to the regions of interest. Providing an in vivo data set for this scenario is a complex task requiring medical staff support. Some initial and promising efforts have been done for evaluating deformable registration methods (Modrzejewski et al. 2019) that could lead to a related NRS*f*M evaluation.

Moreover, given continuously deforming shapes, global temporal consistency should be enforced in order to avoid frame-by-frame sign flips, reflections and other ambiguities given the stronger geometrical expressiveness of deformable models. This is truly necessary in an operative scenario where such a problem might drastically reduce the effectiveness of the NRS*f*M approaches.

Another main avenue of investigation was the effect of missing data. Here we found, that that this aspect has a large impact on the reconstruction error. This is somewhat at odds with previous findings, and we speculate that this has to do with our missing data having structure originating from object self occlusion, as opposed to generate missing data with random sampling. In particular, occlusion-based missing data increases the reconstruction error of all methods except BALM (Del Bue et al. 2012). Our study thus indicates this area to be a fruitful area of investigation for NRS*f*M research.

Another observation is that the physical based methods did quite poorly compared to the methods using a statistically based deformation model. This is in a sense counter intuitive, provided that the physical models capture the deformation physics well. This, in turn, leads us to the observation that stronger efforts could be beneficial as far as better physical based deformation models.

As stated, many of these observations, support hypothesis held in the NRS*f*M community, and it strengths them, that we have here provided empirical support for them. On the other hand, this study also helps to validate the suitability of our compiled data set. In regard to which, it should be noted, both deformation types and camera paths have a statistically significant impact on reconstruction error, regardless of the algorithm used. This indicated that our proposed taxonomy and the data set design has value.

All in all, we have here presented a state of the art data set for NRS*f*M evaluation. We have applied 18 different NRS*f*M method to this data set. Methods that span the state of the art of NRS*f*M. This evaluation validates the usability of our proposed, and publicly available data set, and gives several insights into the current state of the art of NRS*f*M, including directions for further research.

Funding Open access funding provided by Istituto Italiano di Tecnologia within the CRUI-CARE Agreement.

References

Aanæs, H., Dahl, A., & Steenstrup Pedersen, K. (2012). Interesting interest points. *International Journal of Computer Vision*, *97*, 18–35.

Aanæs, H., Jensen, R., Vogiatzis, G., Tola, E., & Dahl, A. (2016). Large-scale data for multiple-view stereopsis. *International Journal of Computer Vision*, *120*, 1–16.

Aanæs, H., & Kahl, F. (2002). Estimation of deformable structure and motion. In: *In workshop on vision and modelling of dynamic scenes, ECCV'02*.

Agudo, A., & Moreno-Noguer, F. (2017a) Dust: Dual union of spatio-temporal subspaces for monocular multiple object 3d reconstruction. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1513–1521).

Agudo, A.,& Moreno-Noguer, F. (2017b). Force-based representation for non-rigid shape and elastic model estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(9), 2137–2150.

Agudo, A., Moreno-Noguer, F., Calvo, B., & Montiel, J. M. M. (2016). Sequential non-rigid structure from motion using physical priors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *38*(5), 979–994.

Akhter, I., Sheikh, Y. S., & Kanade, T. (2011). Trajectory space: A dual representation for nonrigid structure from motion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *33*(7), 1442–1456.

Akhter, I., Simon, T., Khan, S., Matthews, I., & Sheikh, Y. (2012). Bilinear spatiotemporal basis models. *ACM Transactions on Graphics (TOG)*, *31*(2), 17.

Bartoli, A., Gay-Bellile, V., Castellani, U., Peyras, J., Olsen, S., & Sayd, P. (2008). Coarse-to-fine low-rank structure-from-motion. In: *International conference on computer vision and pattern recognition*.

Bouguet, J. Y. (2001). Pyramidal implementation of the affine Lucas Kanade feature tracker description of the algorithm. *Intel Corporation*, *5*(1–10), 4.

Brand, M., & Bhotika, R. (2001). Flexible flow for 3d nonrigid tracking and shape recovery. In: *International conference on computer vision and pattern recognition* (pp. 315–22).

Brandt, S., ad J. Kannala, P.K., & Heyden, A. (2011). Uncalibrated non-rigid factorisation with automatic shape basis selection. In: *Workshop on non-rigid shape analysis and deformable image alignment*.

Bregler, C., Hertzmann, A., & Biermann, H. (2000). Recovering non-rigid 3D shape from image streams. In: *International conference on computer vision and pattern recognition* (pp. 690–696).

Cha, G., Lee, M., Cho, J., & Oh, S. (2019). Reconstruct as far as you can: Consensus of non-rigid reconstruction from feasible regions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Chhatkuli, A., Pizarro, D., & Bartoli, A. (2014). Non-rigid shape-from-motion for isometric surfaces using infinitesimal planarity. In: *BMVC*.

Chhatkuli, A., Pizarro, D., Collins, T., & Bartoli, A. (2018). Inextensible non-rigid structure-from-motion by second-order cone programming. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(10), 2428–2441.

Cho, J., Lee, M., & Oh, S. (2016). Complex non-rigid 3D shape recovery using a procrustean normal distribution mixture model. *International Journal of Computer Vision*, *117*(3), 226–246.

Dai, Y., Li, H., & He, M. (2014). A simple prior-free method for non-rigid structure-from-motion factorization. *International Journal of Computer Vision*, *107*(2), 101–122.

Dawud Ansari, M., Golyanik, V., & Stricker, D.(2017). Scalable dense monocular surface reconstruction. In: *International conference on 3D vision*.

Del Bue, A. (2013). Adaptive non-rigid registration and structure from motion from image trajectories. *International Journal of Computer Vision*, *103*, 226–239. https://doi.org/10.1007/s11263-012-0577-9.

Del Bue, A., & Bartoli, A. (2011). Multiview 3D warps. In: *International conference on computer vision* (pp. 675–682).

Del Bue, A., Lladó, X., & Agapito, L. (2005a). Non-rigid face modelling using shape priors. In W. Zhao, S. Gong, & X. Tang (Eds.), *Analysis and Modelling of Faces and Gestures*. AMFG 2005. Lecture Notes in Computer Science (Vol. 3723). Berlin, Heidelberg: Springer. https://doi.org/10.1007/11564386_9

Del Bue, A., Lladó, X., & Agapito, L. (2005b). Non-rigid face modelling using shape priors. In: *AMFG* (pp. 97–108). Springer.

Del Bue, A., Llado, X., & Agapito, L. (2006). Non-rigid metric shape and motion recovery from uncalibrated images using priors. In: *International conference on computer vision and pattern recognition*.

Del Bue, A., Smeraldi, F., & Agapito, L. (2007). Non-rigid structure from motion using Ranklet-based tracking and non-linear optimization. *Image and Vision Computing*, *25*(3), 297–310.

Del Bue, A., Xavier, J., Agapito, L., & Paladini, M. (2012). Bilinear modeling via augmented Lagrange multipliers (BALM). *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *34*(8), 1496–1508. 10.1109/TPAMI.2011.238. http://users.isr.ist.utl.pt/~adb/publications/2012_PAMI_Del_Bue.pdf.

Deutsches Institut für Normung. (2012). VDI 2634: Optical 3-D measuring systems. Optical systems based on area scanning. Tech. rep., Deutsches Institut für Normung.

Elhamifar, E., & Vidal, R. (2013). Sparse subspace clustering: Algorithm, theory, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(11), 2765–2781.

Fayad, J., Agapito, L., & Del Bue, A.(2010). Piecewise quadratic reconstruction of non-rigid surfaces from monocular sequences. In: *European conference on computer vision*.

Golyanik, V., Jonas, A.,& Stricker, D.(2019). Consolidating segmentwise non-rigid structure from motion. In: *Machine vision applications (MVA)*.

Gotardo, P. F. U., & Martinez, A. M. (2011a). Computing smooth time-trajectories for camera and deformable shape in structure from motion with occlusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *33*(10), 2051–2065.

Gotardo, P.F.U., & Martinez, A.M. (2011b). Kernel non-rigid structure from motion. In: *IEEE international conference on computer vision*.

Gotardo, P.F.U., & Martinez, A.M. (2011). Non-rigid structure from motion with complementary rank-3 spaces. In: *IEEE conference on computer vision and pattern recognition*.

Gower, J. C. (1975). Generalized procrustes analysis. *Psychometrika*, *40*(1), 33–51.

Gower, J. C., & Dijksterhuis, G. B. (2004). *Procrustes problems* (Vol. 30). Oxford: Oxford University Press.

Hamsici, O. C., Gotardo, P. F., & Martinez, A. M. (2012). *Learning spatially-smooth mappings in non-rigid structure from motion* (pp. 260–273). New York: Springer.

Hartley, R., & Vidal, R.(2008). Perspective nonrigid shape and motion recovery. In: *European conference on computer vision* (pp. 276–289).

Hartley, R. I., & Zisserman, A. (2000). *Multiple view geometry in computer vision*. Cambridge: Cambridge University Press.

Hyeong Hong, J., Zach, C., & Fitzgibbon, A. (2017). Revisiting the variable projection method for separable nonlinear least squares problems. In: *The IEEE conference on computer vision and pattern recognition (CVPR)*.

Kong, C.,& Lucey, S. (2016). Prior-less compressible structure from motion. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 4123–4131). https://doi.org/10.1109/CVPR.2016.447.

Kumar, S. (2020). Non-rigid structure from motion: Prior-free factorization method revisited. In: *The IEEE winter conference on applications of computer vision* (pp. 51–60).

Kumar, S., Dai, Y., & Li, H. (2017). Spatio-temporal union of subspaces for multi-body non-rigid structure-from-motion. *Pattern Recognition*, *71*, 428–443.

Lee, M., Cho, J., & Oh, S. (2016). Consensus of non-rigid reconstructions. In: *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*) (pp. 4670–4678). https://doi.org/10.1109/CVPR.2016.505.

Lee, M., Cho, J., & Oh, S. (2017). Procrustean normal distribution for non-rigid structure from motion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(7), 1388–1400. https://doi.org/10.1109/TPAMI.2016.2596720.

Li, X., Li, H., Joo, H., Liu, Y., & Sheikh, Y.(2018). Structure from recurrent motion: From rigidity to recurrency. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, (pp. 3032–3040).

Lladó, X., Del Bue, A., & Agapito, L. (2006). Euclidean reconstruction of deformable structure using a perspective camera with varying intrinsic parameters. In: *Proc. international conference on pattern recognition, Hong Kong*.

Lladó, X., Del Bue, A., & Agapito, L. (2010). Non-rigid metric reconstruction from perspective cameras. *Image and Vision Computing*, *28*(9), 1339–1353.

Menze, M., & Geiger, A.(2015). Object scene flow for autonomous vehicles. In: *Conference on computer vision and pattern recognition (CVPR)*.

Modrzejewski, R., Collins, T., Seeliger, B., Bartoli, A., Hostettler, A., & Marescaux, J. (2019). An in vivo porcine dataset and evaluation methodology to measure soft-body laparoscopic liver registration accuracy with an extended algorithm that handles collisions. *International Journal of Computer Assisted Radiology and Surgery*, *14*(7), 1237–1245.

Olsen, S. I., & Bartoli, A. (2008). Implicit non-rigid structure-from-motion with priors. *Journal of Mathematical Imaging and Vision*, *31*(2), 233–244.

Ornhag, M.V., & Olsson, C. (2020). A unified optimization framework for low-rank inducing penalties. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 8474–8483).

Özyeşil, O., Voroninski, V., Basri, R., & Singer, A. (2017). A survey of structure from motion*. *Acta Numerica*, *26*, 305–364.

Paladini, M., Del Bue, A., Stosic, M., Dodig, M., Xavier, J., & Agapito, L. (2009). Factorization for non-rigid and articulated structure using metric projections. In: *International conference on computer vision and pattern recognition*. https://doi.org/10.1109/CVPRW.2009.5206602

Paladini, M., Del Bue, A., Stosic, M., Dodig, M., Xavier, J., & Agapito, L. (2012). Optimal metric projections for deformable and articulated structure-from-motion. *International Journal of Computer Vision (IJCV)*, *96*, 252–276. https://doi.org/10.1007/s11263-011-0468-5.

Parashar, S., Pizarro, D., & Bartoli, A. (2017). Isometric non-rigid shape-from-motion with Riemannian geometry solved in linear time. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(10), 2442–2454.

Park, S., Lee, M., & Kwak, N. (2018). Procrustean regression: A flexible alignment-based framework for nonrigid structure estimation. *IEEE Transactions on Image Processing*, *27*(1), 249–264. https://doi.org/10.1109/TIP.2017.2757280.

Reich, C., Ritter, R., & Thesing, J. (1997). White light heterodyne principle for 3D-measurement. In O. Loffeld (Ed.), *Sensors, sensor systems, and sensor data processing* (Vol. 3100, pp. 236–244). Washington: International Society for Optics and Photonics, SPIE. https://doi.org/10.1117/12.287750.

Russell, C., Fayad, J., & Agapito, L. (2011). Energy based multiple model fitting for non-rigid structure from motion. In: *IEEE conference on computer vision and pattern recognition*.

Russell, C., Yu, R., & Agapito, L. (2014). Video pop-up: Monocular 3d reconstruction of dynamic scenes. In: *European conference on computer vision* (pp. 583–598). Springer.

Salzmann, M., Pilet, J., Ilic, S., & Fua, P. (2007). Surface deformation models for nonrigid 3d shape recovery. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *29*(8), 1481–1487.

Seber, G. A., & Lee, A. J. (2012). *Linear regression analysis* (Vol. 936). New York: Wiley.

Simon, T., Valmadre, J., Matthews, I., & Sheikh, Y. (2017). Kronecker–Markov prior for dynamic 3D reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(11), 2201–2214.

Szeliski, R. (2010). *Computer vision: Algorithms and applications*. Berlin: Springer.

Taylor, J., Jepson, A.D.,& Kutulakos, K.N. (2010). Non-rigid structure from locally-rigid motion. In: *IEEE conference on computer vision and pattern recognition*.

Tomasi, C., & Kanade, T. (1992). Shape and motion from image streams under orthography: A factorization approach. *International Journal of Computer Vision*, *9*(2), 137–154.

Torresani, L., Hertzmann, A., & Bregler, C. (2004). Learning non-rigid 3D shape from 2D motion. In S. Thrun, L. Saul, & B. Schölkopf (Eds.), *Advances in neural Information processing systems* (Vol. 16). Cambridge: MIT Press.

Torresani, L., Hertzmann, A., & Bregler, C. (2008). Non-rigid structure-from-motion: Estimating shape and motion with hierarchical priors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *30*(5), 878–892.

Torresani, L., Yang, D., Alexander, E., & Bregler, C. (2001). Tracking and modeling non-rigid objects with rank constraints. In: *International conference on computer vision and pattern recognition*.

University, C.M.: Cmu graphics lab motion capture database (2002). http://mocap.cs.cmu.edu/. Accessed Nov 15 2019

Valmadre, J., Lucey, S. (2012). General trajectory prior for non-rigid reconstruction. In: *IEEE conference on computer vision and pattern recognition*.

Varol, A., Salzmann, M., Tola, E.,& Fua, P. (2009). Template-free monocular reconstruction of deformable surfaces. In: *International conference on computer vision* (pp. 1811–1818).

Velleman, P. F., & Hoaglin, D. C. (1981). *Applications, basics, and computing of exploratory data analysis*. Boston: Duxbury Press.

Vicente, S., & Agapito, L. (2012). Soft inextensibility constraints for template-free non-rigid reconstruction. In: *European conference on computer vision* (pp. 426–440).

Vidal, R.,& Abretske, D. (2006). Nonrigid shape and motion from multiple perspective views. In: *European conference on computer vision* (pp. 205–218). Springer.

Wang, G., Tsui, H., & Wu, Q. (2008). Rotation constrained power factorization for structure from motion of nonrigid objects. *Pattern Recognition Letters*, *29*(1), 72–80.

Wang, G., Tsui, H. T., & Hu, Z. (2007). Structure and motion of nonrigid object under perspective projection. *Pattern Recognition Letters*, *28*(4), 507–515.

Wang, Y. X., Lee, C. M., Cheong, L. F., & Toh, K. C. (2015). Practical matrix completion and corruption recovery using proximal alternating robust subspace minimization. *International Journal of Computer Vision*, *111*(3), 315–344.

Williamson, D. F., Parker, R. A., & Kendrick, J. S. (1989). The box plot: A simple visual method to interpret data. *Annals of Internal Medicine*, *110*(11), 916–921.

Xiao, J., Chai, J., & Kanade, T. (2006). A closed-form solution to non-rigid shape and motion recovery. *International Journal of Computer Vision*, *67*(2), 233–246.

Xiao, J., & Kanade, T.(2005). Uncalibrated perspective reconstruction of deformable structures. In: *IEEE international conference on computer vision* (pp. 1075–1082).

Zappella, L., Del Bue, A., Lladó, X., & Salvi, J. (2013). Joint estimation of segmentation and structure from motion. *Computer Vision and Image Understanding*, *117*(2), 113–129.

Zhu, Y., Huang, D., De La Torre, F., & Lucey, S. (2014). Complex non-rigid motion 3D reconstruction by union of subspaces. In: *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1542–1549).