

Real-Time Sequential Model-based Non-rigid SFM

S. Bronte¹, M. Paladini², L. M. Bergasa¹, L. Agapito³ and R. Arroyo¹

Abstract—Tracking non-rigid objects from video is useful in robotic systems such as HMIs or robotic manipulator arms which interact with deformable objects. This paper proposes a method for sequential model-based 3D reconstruction of deformable objects and camera localization in real time. Non-rigid SFM methods commonly process a video sequence offline in a batch way. While there are real-time methods for rigid models, reconstruction of deformable 3D shapes for real-time applications is still unsolved. Dense approaches offer promising results, but processing all frames in batch, offline. We propose a real-time non-rigid reconstruction method based on a known deformable model. Object shape and pose is tracked by real-time estimation of camera pose and deformation coefficients. An extensive evaluation of the algorithm on several data sets, and comparison with state-of-the-art techniques is performed. The tests include different outlier rates, noise levels and occlusions handling.

I. INTRODUCTION

Understanding the deformations of non-rigid objects is vital for some robotic systems. For instance, Human-Machine Interfaces (HMIs) must represent 3D shape and viewpoint correctly when giving feedback to the user. Manipulators need the 3D position of suitable contact points to grasp deformable objects, and robotic surgery requires responsive and accurate 3D reconstruction. A robust and real-time reconstruction process is needed in all these cases, where current non-rigid structure from motion methods cannot be applied, as they work in batch mode. Hence, the development of a real-time non-rigid 3D reconstruction technique by using a single camera is justified. In this context, the real-time constraint is defined by the interval between incoming frames from a camera, which is between 33 and 66 ms.

Currently, most of the Non-Rigid Structure from Motion (NR-SFM) approaches work in an offline mode [1], [2], [3], [4]. The goal of these algorithms is to retrieve the camera pose and the shape of a non-rigid object. The reconstruction is cast as the problem of finding camera and shape parameters that minimize the reprojection error over the whole sequence. On the other hand, successful real-time 3D reconstruction has been achieved for rigid scenes, where the 3D structure captured by the camera does not change over time. Klein and Murray [5] demonstrated a system for online reconstruction which used no prior information about the scene structure or camera motion. Nowadays, the most complex cases of rigid

dense reconstruction from monocular video can be solved in real-time [6], [7]. For the dense non-rigid case there are some proposals in the literature such as [8]. The drawback of this method is that, even using GPU technology, the authors report a run-time of 8 seconds per frame. Thus, this approach is not suitable for online applications.

For the feature-based non-rigid approach, Paladini *et al.* [9] proposed a method to reconstruct each frame independently, while updating the basis shapes as more frames are processed. Although the method does not require a batch process of the whole frames, each frame reconstruction is computationally expensive, making it unsuitable for real-time constraints.

Some non-rigid model-based approaches are able to simultaneously track feature points from images and reconstruct the 3D shape [10], [11], by using a large training set of labeled image data to train the model. Such training images are synthetically created or often manually labeled, a time-consuming task prone to errors. Model-free approaches to non-rigid structure from motion do not rely on any prior knowledge about the scene or the kinds of deformations present. However, those methods require accurate feature correspondences between frames, computed by a feature tracking algorithm. This condition is difficult to be fulfilled in real applications.

This paper proposes a sequential model-based method for real-time 3D reconstruction of deformable objects from unreliable feature correspondences by using monocular vision. Accurate tracking for non-rigid objects is a difficult problem, as the visual appearance of feature points changes due to shadows, reflections and self-occlusions caused by the deformations.

II. RELATED WORK

Simultaneous reconstruction and tracking for rigid objects is a mature area in which real-time methods have been developed using sparse features [5] and dense maps [6], [7].

Tracking techniques for a known rigid object using SLAM techniques are also developed. In the work by Pisacariu *et al.* [12], the authors develop a method for model-based segmentation and tracking based on the assumption that, given an accurate 3D model of an object, its segmentation from any given image is fully defined by its pose.

Non-rigid factorization was introduced by the seminar work by Bregler *et al.* [13], modelling deformations as a linear combination of *basis shapes*, resulting in specific metric constraints on the factorization problem. Available non-rigid factorization methods like Torresani *et al.* [4], [14] use alternation and expectation-maximization (EM) as the

¹Department of Electronics, University of Alcalá, Madrid, Spain
sebastian.bronte, bergasa, roberto.arroyo at depeca.uah.es

²Robotics Research, Ocado Technology, Ocado Ltd, Hatfield, United Kingdom marco.paladini at ocado.com

³Department of Computer Science, University College London, London, United Kingdom l.agapito at cs.ucl.ac.uk

core of the 3D shape estimation. Paladini *et al.* [1] also proposed an alternation method, with an additional projection step to enforce the metric constraints. Other approaches like Gotardo *et al.* [3], reformulate the problem, mixing multiple trajectory and shape modelling, combining them to extract the 3D shape deformation as a single smooth time-trajectory within a linear space spanned by 3D basis shapes.

All the previous non-rigid approaches approximate the camera with orthographic projection, which is not valid when dealing with typical robotic applications. None of them tackle the problem of outliers in the input 2D trajectories. Recently, an approach to handle self occlusions and robust reconstruction is taken by Pizarro *et al.* [15] using an orthographic model.

In the work from [2] MLE+EM is replaced by a MAP formulation. This algorithm is designed for perspective camera model, which makes the algorithm easily applicable to robotics, although it was tested with a small set of points per model. This algorithm addresses the problem of noise and outlier rejection, without the need of handling missing data, since tracking points can change between frames.

In [10], a model-based 3D tracking algorithm is performed by combining texture information for each projection, and the 3D deformation models, in a similar way to object tracking using Active Appearance Models (AAM) [11]. This approach is similar to the one used in this paper, although the point tracks do not directly correspond to the model points.

An incremental solution for the problem of generating a deformation model and tracking was introduced in [9]. The method constructs a model based on several initial frames, and builds deformation modes as the reprojection error reaches a threshold, adding more basis shapes as needed. Another model-building strategy [16] consists of forming several frame clusters and 3D reconstructions for each cluster, later optimizing the whole model as a tree.

Dense methods such as the work of Garg *et al.* [8], can build depth maps of the non-rigid shape by computing dense optical flow. The principal drawback of this method is its low frame rate, even using parallel processing on the GPU.

NR-SFM methods are computationally expensive, in terms of memory and time, especially when the number of points is high. These reasons make them unsuitable to be implemented for real-time estimation of deformable objects. In this paper we propose a real-time non-rigid reconstruction method inspired by the real-time reconstruction algorithm in PTAM (Parallel Tracking and Mapping) [5]. In brief, we extend PTAM to include deformations in the 3D model inferred from the camera motion.

III. ALGORITHM DESCRIPTION

Given a video sequence, a set of deformation bases and an initial estimation of the pose and rigid shape, our algorithm estimates the camera pose and 3D model deformations of an object in the sequence for each frame, tracking some features over the images.

We use the linear basis shapes model introduced by Bregler *et al.* [13] for modelling deformations. The shape

is expressed as a linear combination of a fixed set of *basis shapes* multiplied by time-varying coefficients:

$$S_{NR}(f) = S_R + \sum_{k=1}^K L_k(f) B_k \quad (1)$$

where $S_{NR} \in \mathbb{R}^{3 \times p}$ is the (time-varying) shape for frame f , where p is the number of points. $S_R \in \mathbb{R}^{3 \times p}$ is the average rigid shape, $L(f) \in \mathbb{R}^K$ are the K deformation weights for frame f , and K is the number of basis shapes. The basis shapes $B \in \mathbb{R}^{3K \times p}$ are fixed for the whole sequence and are assumed to be known, normally learnt in an initial set-up from a set of 3D training data.

Each point p of the non-rigid shape S_{NR} is transformed from world to camera coordinates as

$$X_p = (x \ y \ z)^T = [R|T] * S_{NR_p} \quad (2)$$

The camera extrinsic parameters are represented by the 4x3 transformation matrix $[R|T] \in \mathbb{SE}(3)$, where $R \in \mathbb{R}^{3 \times 3}$ specifies camera rotation, and $T \in \mathbb{R}^3$ is the camera translation vector. Points transformed in the camera coordinates are projected with perspective projection as:

$$U_p^S = \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} u_0 \\ v_0 \end{pmatrix} + \alpha \begin{pmatrix} f_u & 0 \\ 0 & f_v \end{pmatrix} \begin{pmatrix} x/z \\ y/z \end{pmatrix} \quad (3)$$

$U^S \in \mathbb{R}^{2 \times p}$ represents the set of shape points, $(u_0, v_0)^T$ are the camera center coordinates, (f_u, f_v) the focal length, known from the prior calibration process, and α the radial distortion function, described in [5].

A. Measurement model and matching

The measurement model is based on the detection of sparse features on the image by using the FAST [17] detector. The most challenging task in the tracking is the association between detected features and projected shape points. In order to address this task, the initial non-rigid shape is used as a template.

The points of the model (S_{NR}) are projected onto the image plane, and a Delaunay triangulation is computed with these projections, resulting in a set of connected vertices v_i . Feature points will not match the vertices of the model, but can be expressed using barycentric coordinates. A tracked point $U'_p = (u'v')^T$ has barycentric coordinates (a, b, c) computed as in [2]:

$$\begin{pmatrix} a \\ b \\ c \end{pmatrix} = \text{pinv} \begin{pmatrix} v_{1x} & v_{2x} & v_{3x} \\ v_{1y} & v_{2y} & v_{3y} \end{pmatrix} * \begin{pmatrix} u' \\ v' \end{pmatrix} \quad (4)$$

Where $\text{pinv}(\cdot)$ is the matrix pseudo-inverse. Using the barycentric coordinates we can define a set of interpolated basis B' such that the 3D position of a feature point X'_p relates to the model basis shapes as:

$$\begin{aligned} X'_p &= aX_{v1} + bX_{v2} + cX_{v3} \\ &= \sum_k L_k (aB_{v1k} + bB_{v2k} + cB_{v3k}) = \sum_k L_k B'_{pk} \end{aligned} \quad (5)$$

Feature points detected in the image correspond to the vertices of B' and are easier to track than the points given by the known basis shapes model. This soft constraint applied to the detected points provides flexibility to the tracking and facilitates the matching of the points among frames, as there is no need to strictly look for features near the initial model points, but we use directly the detected points.

The detected features points on a frame are matched with those detected in the previous frame by using an algorithm similar to the one used in PTAM [5]. However, some modifications are needed to deal with non-rigid features. For each point, the following matching is followed:

- First, an affine warping is applied (based on point, patch and pose), like in [5]. Most of the nearly rigid points will be found using this approach.
- If this matching fails, it is usually due to deformations in the search area. In this case, a multilevel correlation-based approach is performed. The previous frame feature is looked up and matched in the top level of the current frame pyramid, and refined by matching it in the lower pyramid levels. Matching is discarded if correlation is too low or if displacement is too large. Married matching (current to previous frame) is applied to discard false positives.

This matching approach fails for points with high deformed texture. The effect of these points is mitigated by using all available matches on the model estimation in the calculation of the deformation coefficients.

B. Motion model

In order to improve convergence in the tracking, a linear motion model of the camera is computed. Then the pose is updated as in (6).

$$\begin{aligned} vel_t &= \beta/2 (vel_{t-1} + \mu) \\ \widehat{[R|T]}^t &= exp(vel_t \Delta t) [R|T]^{t-1} \end{aligned} \quad (6)$$

Where vel_t is the camera speed at frame t calculated using the ESM homography between frames [18], β is the factor that modulates the influence of vel on the update, and μ is the camera motion vector. The choice of the selected algorithm comes from the original tracking in [5].

C. E-M optimization

Several estimation methods implement an Expectation Maximization approach to compute deformation weights, keeping the rest of the parameters fixed [4], [14], [19]. The likelihood function is defined as,

$$f_{MLE}(U', \mu, L, B, S_R) \propto \sum_p \left| U'_p - proj(X'_p) \right| \quad (7)$$

The parameters we want to estimate are the camera pose vector $\mu = (\phi_x \phi_y \phi_z t_x t_y t_z)$ composed by the angles and the translations in the 3 axes, and the coefficient vector L for each frame.

We need to minimize (7) w.r.t. $\theta = [\mu \ L]$, the state vector, which is formed by $6 + K$ components for pose and deformation weights. We assume, B , S_R and K fixed.

We run a maximum of 10 EM iterations per frame, alternating between camera pose and deformation coefficients optimization. The current number of iterations depends on the RMS reprojection error. If the error is not significantly reduced between frames the process is stopped. If the error increases, the whole optimization algorithm is applied and, if the error continues increasing, the estimation is stopped. The best solution for the state vector is kept for the next frame. Each estimation is performed by weighted least squares minimization of (7).

1) *E-Step. Coefficient estimation:* In this step, the goal is to estimate the set of K deformation coefficients to improve the posterior pose estimation. A preliminary set of coefficients are computed with the current estimation of μ^{t-1} , S_R and B' .

The E-step computes the lower bound for the observed data likelihood f_{MLE} given the previous $(t-1)$ state vector. The M-step will find the best camera pose given L . The bound is obtained by minimizing the reprojection error over the current set of coefficients L , and taking into account the motion model:

$$e_p^L = \left(\begin{array}{c} u' \\ v' \end{array} \right) - proj \left(\widehat{[R|T]}^t \sum_k L_k^{t-1} B'_k \right) \quad (8)$$

Where L^{t-1} is the current value for the coefficients, e_p^L the reprojection error for each point p in the current frame. We weight the observed points using the Tuckey bi-weight M-estimator function of the reprojection error, with a median-based estimation of the standard deviation, in the same way that the proposed in [5]. The M-estimator reduces outlier and noise influence on the results, weighting each measurement depending on its reprojection error. Only the detected points from the matching are taken into account.

To compute the deformation weights, starting from (3), we undistort each 2D measurement, and project the 3D shape on the image plane. For the orthographic case, the 2D camera coordinates are approximated to the first 2 transformed coordinates, and the following expression is valid

$$proj \left(\sum_k L_k B'_k \right) = \sum_k L_k proj(B'_k) \quad (9)$$

However, in the case of perspective projection, (9) is not valid because projected points depend also on the depth. To compute the coefficients, we start from equations (2) and (3) and expanding the projection function we reach:

$$\left(\begin{array}{c} \lambda u \\ \lambda v \\ \lambda \end{array} \right) = \left(\begin{array}{c} f_u \left(\sum_k L_k \vec{r}_x B'_k + t_x \right) - u_0 \lambda \\ f_v \left(\sum_k L_k \vec{r}_y B'_k + t_y \right) - v_0 \lambda \\ \sum_k L_k \vec{r}_z B'_k + t_z \end{array} \right) \quad (10)$$

Where $R = (\vec{r}_x \ \vec{r}_y \ \vec{r}_z)^T$, $T = (t_x \ t_y \ t_z)^T$ are the camera pose and λ the projection scale.

By grouping the terms L_k and writing the system in a least squares form $Ax = C$, where $x = L$, $\Delta u_p = u_p - u_0$, and $\Delta v_p = v_p - v_0$, we obtain:

$$A = \begin{pmatrix} (\Delta u_1 \vec{r}_z - f_u \vec{r}_x) B'_{11} & \cdots & (\Delta u_1 \vec{r}_z - f_u \vec{r}_x) B'_{1k} \\ (\Delta v_1 \vec{r}_z - f_v \vec{r}_y) B'_{11} & \cdots & (\Delta v_1 \vec{r}_z - f_v \vec{r}_y) B'_{1k} \\ \cdots & \cdots & \cdots \\ (\Delta u_p \vec{r}_z - f_u \vec{r}_x) B'_{p1} & \cdots & (\Delta u_p \vec{r}_z - f_u \vec{r}_x) B'_{pk} \\ (\Delta v_p \vec{r}_z - f_v \vec{r}_y) B'_{p1} & \cdots & (\Delta v_p \vec{r}_z - f_v \vec{r}_y) B'_{pk} \end{pmatrix} \quad (11)$$

$$C = \begin{pmatrix} f_u t_x - t_z \Delta u_1 \\ f_v t_y - t_z \Delta v_1 \\ \vdots \\ f_u t_x - t_z \Delta u_p \\ f_v t_y - t_z \Delta v_p \end{pmatrix} \quad (12)$$

In the case that an average shape is given for the first frame (S_R), this will help to improve the algorithm convergence, describing it as a function of the original basis shapes:

$$S_R = \sum_k L_{Rk} B_k \quad (13)$$

Where $L_R \in \mathbb{R}^K$ is the vector of coefficients that describe the average rigid shape. Computing the transformation $L_{NR} = L - L_R$, we remove the rigid shape influence.

2) *M-Step. Pose estimation:* In this step, we compute the camera pose by maximizing the likelihood (7) of the observed data.

The maximum likelihood pose is estimated while keeping the 3D shape fixed, again minimizing the reprojection error:

$$e_p^\mu = \begin{pmatrix} u' \\ v' \end{pmatrix} - \text{proj} \left(\exp(\mu^t) [\widehat{R|T}]^t \sum_k L_k^t B_k' \right) \quad (14)$$

Equation (14) is similar to (8), but in this case we search for the minima w.r.t. μ instead of L .

We compute the μ update in the same way as done in [5]. Similarly to the E-step, the points not found are taken out from the pose estimation on the current frame.

IV. EXPERIMENTS

Our algorithm has been tested on several commonly employed sequences of the state of the art. A summary of the main features of these sequences are shown in Table I.

Seq. - Ref.	#Points	#Frames	σ	outl
CMUface - [9]	40	315	yes	yes
Flag - [20]	540	450	yes	yes
Rendered flag - [21]	-	60	-	-

TABLE I: Main features of the tested sequences

For data sets we set $K=15$. In order to perform a thorough comparison from the motion captured data, an increasing noise strength, $\sigma = [0, 1, 2, 3, 4]$, and outlier percentage, $outl = [0, 5, 10, 20, 30, 40]\%$ is added to the points.

The CMUface sequence comes from [9], the point wise flag sequence from [20] and the rendered flag sequence is

presented in [21]. The last sequence consist of images and dense 3D ground truth to evaluate the performance of optical flow methods.

All the experiments were computed using a 4-core processor at 2.4 GHz, with 4GB RAM. The algorithm is implemented in c++, runs on Ubuntu Linux, using CGAL, OpenMP and PTAM derived libraries. The experiments for [9] and [3] are run with the default parameters in the examples. For the experiments, we set the motion movement constant β to 0.9 and the feature search radius to 15.

It must be noted that although the proposed method can reconstruct non-rigid shapes in real-time given a deformation model, current state-of-the art methods cannot because they are not model-based, hence, a direct comparison is not possible. Results from other non-rigid SFM methods are only used to show qualitative differences.

In Table II a summary of results is shown, depicting processing time (whole sequence without data loading and rendering), 3D accuracy and 2D error. Our proposal gets the best 3D accuracy and reprojection error for both sequences. Also, note that for CMUface and Flag sequences, the data association is directly given. Therefore, no transformed bases are needed.

Sequence	Method	proc. T / Tpf	3D err (%)	2D err (px)	model
CMUface	ours	4,5 s / 0.7 ms	0.1	0.28	yes
CMUface	[9]	14' / 2.66 s	3.18	1.06	no
CMUface	[3]	43 s / 136 ms	3.19	0.6	no
Flag	ours	7.5 s / 2ms	0.6	1.99	yes
Flag	[9]	20' / 2.66 s	91.65	6.14	no
Flag	[3]	36' / 4.8 s	66.65	29.79	no

TABLE II: Summary of results for the sequences.

Sequence	Method	proc. T / Tpf / Tpp	2D err (pix)	model
Rend. flag	ours	2.5s / 30ms / 54 μ s	5.4	yes
Rend. flag	[21]	8' ¹ / 8s / 890 μ s	≈ 1	no
Rend. flag	[15]	-	≈ 1.25	no

TABLE III: Summary of results for the sequences.

In Table III the results obtained for the Rendered flag are shown, depicting processing time (total/per frame/per point), and 2D reprojection error from the features or the flow.

A. CMUface sequence

This motion captured sequence represents a moving head turning and talking. We add random noise and outliers to the sequence for the different tests. Reconstruction results with varying noise and outlier levels are shown in Fig. 1.

Our method obtains low 3D and reprojection errors for all rates of noise and outliers. These results are due to our method can handle outliers, whereas [9] minimizes reprojection error as it assumes the tracks do not contain errors. This continuous adaptation leads to an increment of the error compared to the ground truth.

¹as reported by the authors using GPU and server clusters

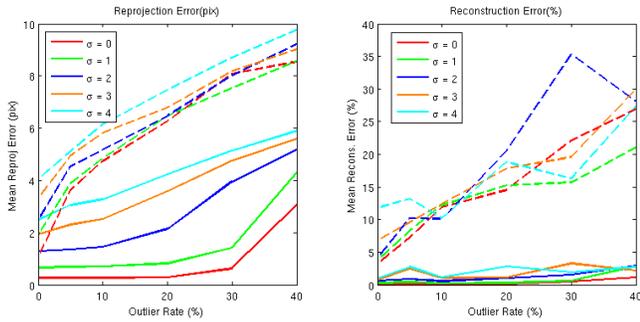


Fig. 1: Results for *CMUface* sequence. Continuous line represents our results, dashed line represents results from [9].

B. Flag sequence

In this sequence a motion captured flag is bending with the wind. The huge deformation of the flag makes this sequence one of the most challenging. Moreover, the amount of points in the sequence is significantly higher than other data sets.

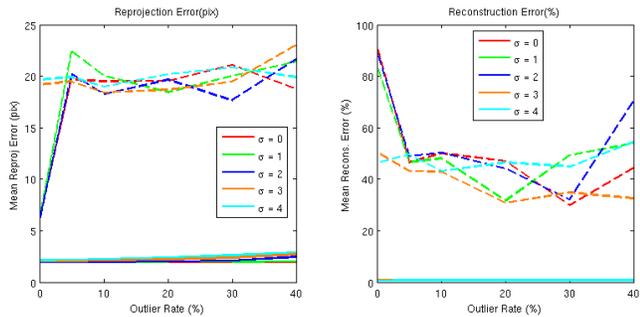


Fig. 2: Results for *Flag* sequence. Results from our approach are in continuous line, results from [9] are in dashed line

Looking at the processing times of state of the art approaches (Table II), it gives an idea of the data set complexity. This table shows that current solutions to model-free NR-SFM cannot recover a correct 3D shape for the flag sequence. Moreover, the speed reached by our method is quite faster than the other approaches.

It can be seen on Fig. 2 that the performance difference between methods is high and even higher than in Fig. 1 due to the sequence complexity and the same reason pointed out in the previous point. Comparison with [3] is not shown as the results are worse than for [9].

C. Rendered flag sequence

The flag sequence is interpolated and rendered using isometric projection, in Garg *et al.* [21]. We re-rendered this sequence with perspective projection.

Initialization of the model and pose is provided to the tracking algorithm. As shown on Fig. 3.a the tracking is focused on the detected (in magenta) and matched points (in red). The shape composed of the current coefficient estimation and the prior basis is also represented, to have an idea of how the estimation fits on the image (cyan grid).

On Fig 3.b the obtained 2D error is shown for frames 20, 30 and 60 and warped to the first (reference) frame (0-10 px). The 3D error is shown on Fig. 3.c (0 - 15%). As we can see on Fig. 3, the unmatched areas of the image are prone to have a high reprojection error, as there are no correct measurements to estimate coefficients correctly.

The results given after processing this sequence are depicted in Table. III. Our reprojection error is about 5 times higher than the other methods, being the 3D reconstruction error of the sequence below 3.5%. The processing time of this method is more than 200 times lower than [21]. It must be noted that we are using just one core, except for the matching, where OpenMP library is utilized. The authors of [21] are using clusters and GPU technology.

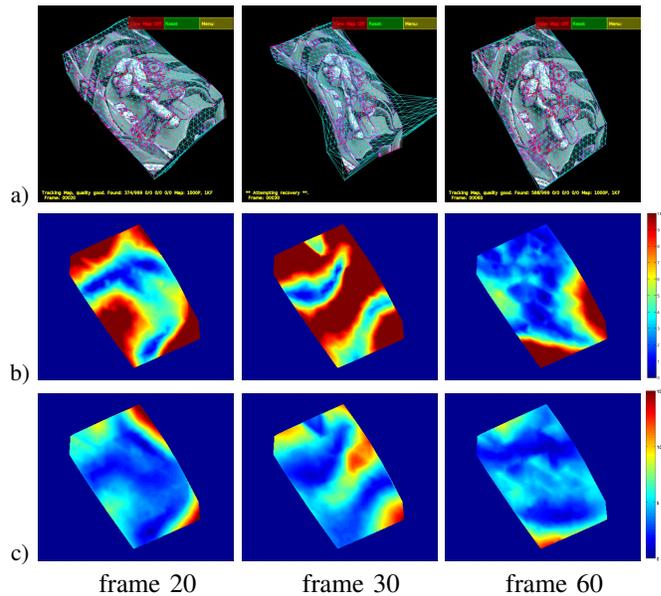


Fig. 3: Results from *Rendered flag* sequences. a): processed sequence output. b): 2D error map (px). c) 3D error map (%).

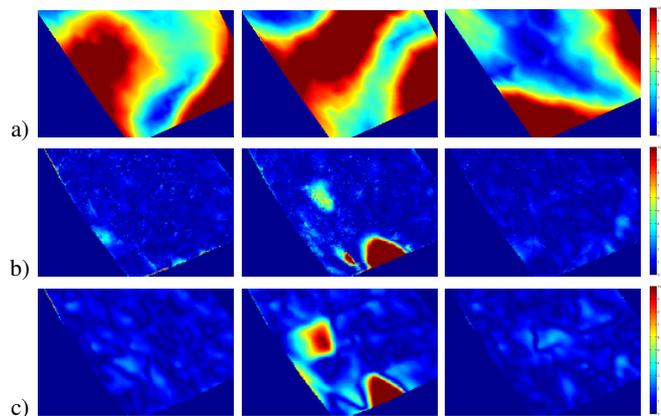


Fig. 4: 2D error comparison for *Rendered flag* sequence (0-10 px). a) current approach. b) Garg *et al.* [21] approach. c) Pizarro *et al.* [15] approach.

On Fig. 4 a comparative study between our approach, [21]

and [15] is shown in terms of reprojection error. Even though our error is higher, the tracking is not lost in any frame. On the other hand, our method can be run in real time. Thus, the balance between processing time and precision is clearly favorable to our proposal

V. CONCLUSIONS AND FURTHER WORK

This paper has presented a sequential approach to solve the structure and motion estimation for non rigid objects. It takes the best from the PTAM algorithm [5], and extends it to include deformations in the 3D model inferred from the camera, showing that it is able to deal with strong deformations.

Current state-of-the-art non rigid methods are not focused on real time implementation, since they are not sequential and require the whole data set.

The proposed method is able to deal with noise and outliers, which is important when facing real applications.

Even though the reprojection error is worse than some state-of-the-art approaches for real sequences, the 3D error is below 3.5% and the processing time is more than 200 times faster without using GPU computation. Our model-based approach provides a balance between performance and accuracy, which makes it suitable to be included on a robotic tracking system when it has to deal with deformable surfaces and elastic objects.

When the pose is highly variant, the balance between the pose and the deformation estimation is not accurate enough, resulting in a tendency to model the deformations better than the pose. Shape and time priors could be introduced to penalize strong deformations between frames.

When the 2D-3D correspondences are not given beforehand, an alternative data association procedure is needed, like the ones in [2], [15]. A matching that handles difficult and deformed texture points in a fast way is required, as the better the points are matched, the better the estimations will be.

Descriptor-based matching will be included in the near future to improve tracking, as the non-detected points of the most deformed areas are critical to get a proper reconstruction and to minimize the reprojection error.

A good initialization is crucial to get accurate results, which is a drawback for several real applications. Further works will be focused on this line, trying to automatically initialize a model, basis and pose after several frames.

VI. ACKNOWLEDGMENTS

This work has been partially financed with funds from the MINECO through the project Smart Driving Applications (TEC2012-37104) and with funds from the CAM under project Robocity2030 (S2009/DPI-1559). This work was supported by the European Research Council under the ERC Starting Grant agreement 204871-HUMANIS. Additionally, we would like to thank the authors of [21] for the provided data sets and generation scripts and J. Yebes for the useful comments of this paper.

REFERENCES

- [1] 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 *IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 2898–2905.
- [2] F. Moreno-Noguer and J. M. Porta, "Probabilistic simultaneous pose and non-rigid shape recovery," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 1289–1296.
- [3] P. F. Gotardo and A. M. Martinez, "Non-rigid structure from motion with complementary rank-3 spaces," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 3065–3072.
- [4] L. Torresani, A. Hertzmann, and C. Bregler, "Learning non-rigid 3d shape from 2d motion," in *Advances in Neural Information Processing Systems 16*, 2003, pp. 1555–1562.
- [5] G. Klein and D. Murray, "Parallel tracking and mapping for small AR workspaces," in *Proceedings of the 2007 6th IEEE and ACM International Symposium on Mixed and Augmented Reality*, 2007, pp. 1–10.
- [6] R. A. Newcombe, S. J. Lovegrove, and A. J. Davison, "DTAM: Dense tracking and mapping in real-time," in *International Conference on Computer Vision*, 2011, pp. 2320–2327.
- [7] J. Stühmer, S. Gumhold, and D. Cremers, "Real-time dense geometry from a handheld camera," in *DAGM-Symposium*, 2010, pp. 11–20.
- [8] R. Garg, A. Roussos, and L. Agapito, "Dense variational reconstruction of non-rigid surfaces from monocular video," in *The IEEE Conference on Computer Vision and Pattern Recognition*, June 2013, pp. 1272–1279.
- [9] M. Paladini, A. Bartoli, and L. Agapito, "Sequential non-rigid structure-from-motion with the 3D-implicit low-rank shape model," in *Proceedings of the 11th European Conference on Computer vision*, 2010, pp. 15–28.
- [10] E. Muñoz, J. M. Buenaposada, and L. Baumela, "A direct approach for efficiently tracking with 3D morphable models," in *International Conference on Computer Vision*, 2009, pp. 1615–1622.
- [11] T. Cootes, G. Edwards, and C. Taylor, "Active appearance models," in *IEEE European Conference on Computer Vision*, 1998, pp. 484–498.
- [12] V. Prisacariu and I. Reid, "PWP3D: Real-time segmentation and tracking of 3d objects," *International Journal of Computer Vision*, pp. 1–20, 2012.
- [13] C. Bregler, A. Hertzmann, and H. Biermann, "Recovering non-rigid 3D shape from image streams," in *IEEE Conference on Computer Vision and Pattern Recognition*, 1999, pp. 690–696.
- [14] L. Torresani, A. Hertzmann, and C. Bregler, "Nonrigid structure-from-motion: Estimating shape and motion with hierarchical priors," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 5, 2008, pp. 878–892.
- [15] D. Pizarro and A. Bartoli, "Feature-based deformable surface detection with self-occlusion reasoning," *International Journal of Computer Vision*, vol. 97, no. 1, pp. 54–70, 2012.
- [16] S. Zhu, L. Zhang, and B. M. Smith, "Model evolution: An incremental approach to non-rigid structure from motion," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 1165–1172.
- [17] E. Rosten and T. Drummond, "Machine learning for high-speed corner detection," in *European Conference on Computer Vision*, vol. 1, May 2006, pp. 430–443.
- [18] S. Benhimane and E. Malis, "Homography-based 2d visual tracking and servoing," *International Journal of Robotic Research*, vol. 26, no. 7, pp. 661–676, July 2007.
- [19] F. Dellaert, "The expectation maximization algorithm," College of Computing, Georgia Institute of Technology, Tech. Rep., 2002.
- [20] R. White, K. Crane, and D. Forsyth, "Capturing and animating occluded cloth," in *ACM Transactions on Graphics (SIGGRAPH)*, 2007, p. 34.
- [21] R. Garg, A. Roussos, and L. de Agapito, "Robust trajectory-space tv-II optical flow for non-rigid sequences," in *EMMVCVPR*, vol. 6819, 2011, pp. 300–314.