SAFARI FROM VISUAL SIGNALS: RECOVERING VOLUMETRIC 3D SHAPES

Antonio Agudo

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

ABSTRACT

In this paper we propose a convex approach for recovering a detailed 3D volumetric geometry of several objects from visual signals. To this end, we first present a minimal detailed surface energy that is optimized together with a volume constraint by considering some geometrical priors, and without requiring neither additional training data nor templates in order to constrain the solution. Our problem can be efficiently solved by means of a gradient descent, and be applied for single RGB images or monocular videos even with very small rigid motions. Temporal-aware solutions and driven by point correspondences are incorporated without assuming any 2D tracking data over time. Thanks to this formulation, both rigid and non-rigid objects can be considered. We have extensively validated our approach in a wide variety of scenarios in the wild, recovering challenging type of shapes that have not been previously attempted without assuming any training data.

Index Terms— Visual Signals, Mininal Surfaces, Image and Video Understanding.

1. INTRODUCTION

Visual signals are nowadays present in everyone's life and they can be easily accessed through the Internet, mainly thanks to the rapid development of acquiring devices. In the last decades, many efforts have been made in obtaining systems able to perceive in three dimensions from still visual information. Unfortunately, building algorithms that can emulate the human 3D perception has proven to be a much harder task than initially anticipated. There is a large body of literature on shape estimation from a single RGB image, especially considering human motion [1]. These approaches normally assume a relatively large set of training data to solve the inverse problem, making them a very specific problem since the learned knowledge cannot be applied to another type of objects. To solve this limitation, other approaches relied on the use of planar templates to establish correspondences [2], RGB-D sensors, or image silhouettes [3, 4, 5] to infer the geometry. Unfortunately, the previous approaches cannot recover some parts of very thin objects as well as retrieve a finer level of details.

This work has been partially supported by the Spanish Ministry of Science and Innovation under project MoHuCo PID2020-120049RB.

The alternative strand of methods known as structure-from-motion [6] relied on 2D point trajectories throughout the video to recover the 3D geometry. Later, these formulations were extended to the non-rigid domain, coining the term non-rigid structure from motion [7]. In these cases, the problem is inherently ill-posed and requires to exploit the denominated art of priors, in shape and motion, to constrain the solution space. Most of techniques represent the time-varying shape as a linear combination of shape [8, 9], trajectory [10] or force [11] vectors. Sadly, these approaches have a few important restrictions in practice: 1) they can only recover the part of the shape that is fully observed in the image, and 2) a considerable amount of out of plane rotation is needed to solve the problem. Nonetheless, there exist some works addressing the volumetric 3D reconstruction problem for rigid scenes [12], and non-rigid ones such as humans [13, 14] and animals [15, 16], where the use of training data is mandatory. Other approaches have relied on knowing a volumetric 3D initial mean shape [17, 18] that is the main component of a deformation model. Unfortunately, acquiring volumetric 3D shapes for some objects, such as animals, is a hard task because the standard 3D scanning techniques used to capture human motion are not applicable to those scenarios in the wild. The large diversity of animals and 3D configurations they can take [19], makes the problem extremely complex, particularly in unconstrained environments with uncalibrated cameras. Recovering the corresponding volumetric 3D closed shape from pictures is a longstanding challenge with potentially many real-world applications in biology, agriculture, animal conservation, animal-inspired design, the movie industry, motion capture or neuroscience, to name just a few.

We overcome most of the limitations of current methods with a convex approach that can retrieve a closed and detailed 3D shape from a single image without needing any training data. Moreover, our approach can be also applied for rigid and non-rigid objects from a monocular video where the camera motion is almost null, and without assuming any 2D tracking data. To the best of our knowledge, no previous approach has reconstructed animals in the wild without considering one or multiple shape models as it is assumed in template-based and learning-based approaches, respectively. In contrast, our approach relies on an almost fronto-parallel image to infer the 3D volumetric shape, that represents a drastic reduction of priors in comparison with previous techniques [15, 16, 18].

2. MINIMAL DETAILED SURFACES FROM SPATIO AND SPATIO-TEMPORAL SIGNALS

Let $\mathcal{I} \subset \mathbb{R}^2$ be an image plane where it appears an object we want to reconstruct. For that object, we also define $S \subset \mathcal{I}$ as its shape segmentation, and by means of $B \subset S$ its silhouette boundary, i.e., S contains the pixels inside the silhouette of the object shape and B only its boundary. Our goal is to recover the object 3D geometry from the single image $\mathcal{I}$ by determining silhouette consistent surfaces of minimal area together with a volumetric constraint. In other words, our problem consists in estimating a height map function $z : S \rightarrow \mathbb{R}$, assigning a depth value $z(x, y)$ for every point $(x, y) \in S$. To this end, we propose to minimize an energy function composed of a data term to obtain the minimal surface [3], and a shape prior to regularize it as:

$$\mathcal{A}(z) = \int_S \sqrt{1 + |\nabla z|^2} + \lambda (z - w)^2 \; dxdy \qquad (1)$$
$$\text{subject to} \int_S z \; dxdy = V$$

where $\nabla(\cdot)$ denotes a gradient operator, λ is a weight coefficient, V indicates the volume of the object, and w is a function to regularize the solution. It is worth noting that this formulation never exploits any depth value known in advance at any specific point to constrain the solution. As the estimation is performed from a single view, depth information will be up to scale, and it can be fixed by imposing a volume value V, as the product of the area of the silhouette S and the estimated average depth value of the object.

To define the shape prior, we first assume that the thickness of the object increases as we move inward from its silhouette boundary B. This assumption is especially relevant in nature, where the shape normally evolves from the boundary to the interior following a smooth and harmonious way, as we can see in many natural objects, such as animals, and a few human-made ones. To become effective, the distance $d(p, \partial B)$ to the boundary B for any interior point $p \in S$ can be computed as $d(p, \partial B) = \min_{b \in \partial B} \|p - b\|$.

While this assumption by itself is good for many points [4, 5], it still represents a non-realistic 3D constraint in others, since an ample variety of details are not retrieved. To solve this limitation, we define a detail map $e(\mathcal{I})$ by exploiting image information as $e(\mathcal{I}) = \gamma \frac{(|\nabla \mathcal{I}| - \min(|\nabla \mathcal{I}|))}{\max(|\nabla \mathcal{I}|) - \min(|\nabla \mathcal{I}|)}$, where γ is a weight coefficient, and $e(\mathcal{I}(m)) = 0$ for any point $m \notin S$. The previous terms are now combined to define the function w, that for the pixel location (x, y) can be written as:

$$w(x, y) = \min\{\phi, \mu + \kappa \, d((x, y), \partial B) + e(\mathcal{I}(x, y))\}, \quad (2)$$

where $\{\phi, \mu, \kappa\}$ are parameters to code the type of prior. Particularly, ϕ is to limit the level of extrusion of the object and it can be set as $\phi = \alpha \max(d((x, y), \partial B))$, with $\alpha \in [0, 1]$. μ is to guarantee a minimum of extrusion in those points close to the boundary. Note that the influence of this term can be attenuated with decreasing the shape prior by modifying the coefficient λ in Eq. (1). Function w in Eq. (2) can be seen as a shape constraint to encode that the object gets thicker the further away the point is from the boundary, while considering spatial details and encouraging that thin areas are still extruded. As the optimization problem in Eq. (1) is convex, we can obtain a global optimal solution by means of an iterative gradient descent method with a projection step to enforce the volume constraint. Finally, we also impose boundary conditions to guarantee silhouette boundary consistency. To this end, Dirichlet's boundary conditions are considered as $z(x, y) = 0, \quad \forall (x, y) \in B$. Once the minimal detailed surface z with fixed volume is achieved, a closed surface (including that unobservable part) can be computed by applying a reflection from the image plane and adding an internal mesh. This makes our algorithm more accurate as the plane of symmetry of the object coincides with the image plane observed.

While our formulation in Eq. (1) could be directly used to process video sequences frame by frame, the formulation can also be extended to be temporally consistent, and provide more realistic and faster solutions. To this end, we just need some matching points between two consecutive images in the video. To recover the 3D geometry of an object at frame $f+1$ considering the estimation in frame f, our energy problem can be written as:

$$\mathcal{T}(z^{f+1}) = \int_S \sqrt{1 + |\nabla z^{f+1}|^2} + \lambda (z^{f+1} - w^{f+1})^2 + \zeta (o^{f+1} \odot (z^{f+1} - r^f(z^f)))^2 \; dxdy$$
$$\text{subject to} \int_S z^{f+1} \; dxdy = V \qquad (3)$$

where $\odot$ represents a Hadamard product and ζ is a weight coefficient. o^{f+1} includes $\{0, 1\}$ entries indicating whether the coordinates of a point in the $f + 1$-th frame were matched or not in the f-th one. The z^f values for matching points are included in $r^f(z^f)$ in a correct location according to o^{f+1}, and they are used to constrain the new estimation z^{f+1}. To solve the matching problem, we exploit the segmentations S^f/S^{f+1}, reducing the amount of pixels to be considered. Additionally, and to avoid bad matches in objects with similar local texture, we also include a guided search of correspondences by means of a 25×25 window where SIFT [20] points between consecutive images are considered. Therefore, our approach does not need 2D tracking data to sort out the spatio-temporal problem.

3. EXPERIMENTAL EVALUATION

We now present our experimental evaluation for different scenarios, by considering single images as well as monocular videos of both rigid and non-rigid animal objects.

Single RGB Image. We first evaluate our approach on real RGB images in the wild taken from the DAVIS

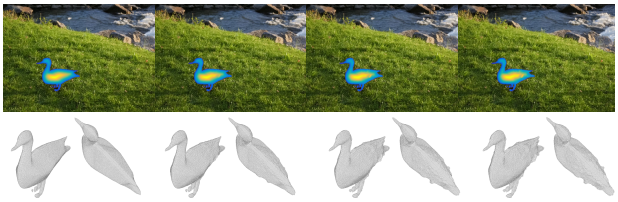

Fig. 1. **Ablation study on detailed surfaces. Top:** From left to right it is represented the minimal surface estimation z by modifying the parameter $\gamma = \{0, 10, 20, 25\}$ from 0 (without considering any details) to 25 (probably, a type of over-estimation). **Bottom:** Two novel view points of the 3D reconstruction.

	cow	flamingo	camel	mallard	rhino	black
[3]	$1.0 \cdot 10^{-4}$	$3.1 \cdot 10^{-4}$	$1.3 \cdot 10^{-4}$	$7.6 \cdot 10^{-5}$	$7.1 \cdot 10^{-5}$	$1.3 \cdot 10^{-4}$
[5]	$1.4 \cdot 10^{-4}$	$7.9 \cdot 10^{-4}$	$7.2 \cdot 10^{-4}$	$4.2 \cdot 10^{-4}$	$2.5 \cdot 10^{-4}$	$2.1 \cdot 10^{-4}$
Ours	$1.9 \cdot 10^{-9}$	$1.2 \cdot 10^{-7}$	$3.9 \cdot 10^{-9}$	$5.1 \cdot 10^{-9}$	$2.3 \cdot 10^{-9}$	$5.5 \cdot 10^{-9}$

Table 1. **Quantitative error comparison.** Error for some images considered in Fig. 2.

dataset [21]. As the object segmentation is out of the scope of this paper, we directly use the original object segmentations provided in the dataset, since similar solutions can be obtained by means of semi-supervised [22, 23] and unsupervised [24] approaches. Particularly, we consider the following RGB images: *cow*, *flamingo*, *camel*, *mallard*, *rhino*, *black swan*, *bicycle* and *break dance*. In all cases, image resolution is 480×854, producing 3D reconstructions from 10,249 (*mallard*) to 77,088 (*rhino*) points (this number represents the size of z, i.e., the final value for the closed 3D shape will contain many more).

We first use the *mallard* image to evaluate how the detailed map in Eq. (2) acts. To this end, we directly modify the γ coefficient from 0 to 25. The results are displayed in Fig. 1, including our z estimation along with the corresponding volumetric 3D shape. As it can be seen, thanks to this term, our algorithm provides more physically plausible estimations.

Considering $\gamma = 10$, we use the rest of the images for evaluation. An outline of our results are summarized in Fig. 2. As it can be seen, the minimal detailed surface z we obtain produces solutions with spatial details while thin areas are still extruded and the variation from the boundary is consistent (see third column in the figure). Without loss of generality, as the distance from the camera to the object shape is within reasonable bounds, the relation between shape area and shape volume is always similar, simplifying the search for a volume value V. Finally, the solution z is employed to generate physically-aware closed 3D surfaces (see the right part of the figure). It is worth noting that those representations are only included for visualization purposes, since our estimation is z, as it can be seen in the third column of the cited figure.

A qualitative comparison is also provided with respect to the approaches [3, 5] (as they use the same input signal) in Fig. 3 for the *flamingo* and *mallard* images. To make a fair comparison, we use for all methods the same volume V, and apply the same strategy to infer the closed surface. As it is

Fig. 2. **Qualitative evaluation on real images in the wild**. From top to bottom are considered the images: *cow*, *flamingo*, *camel*, *mallard*, *rhino*, *black swan*, *bicycle* and *break dance*. **First column:** Input RGB image. **Second column:** Object image segmentation. **Third column:** Minimal detailed surface with volume z. Yellowish areas mean bigger z values. **Fourth column:** 3D reconstruction from a novel point of view. **Fifth column:** 3D reconstruction from a novel point of view with original texture. Best viewed in color.

shown, thin (see for instance *mallard* and *flamingo* faces, as well as the *flamingo* legs) and detailed areas (see main body parts and feathered areas) cannot be recovered properly by competing techniques [3, 5], as we can do. After performing a qualitative comparison, we can see that [5] handles better thin areas than [3], but it is not enough to obtain physically-aware surfaces as those recovered by our approach. Thanks to our novel energy, besides producing more accurate solutions, our algorithm is more stable and converges faster, giving an speed up of $22.92\times$ and $20.70\times$ in comparison with [3] and [5].

Figure 4 shows the evolution of the error for the eight cases considered in Fig. 2, comparing these values with competing techniques in similar conditions (see table 1). Note that after around 50 iterations our approach drastically reduces the error, being a few extra iterations needed to guarantee the exact satisfaction and, obtaining lower errors than other methods. As it can be also seen in the figure, the volume constraint is perfectly satisfied throughout the iterations. On average, the median computation time with non-optimized Matlab code was from 105 to 152 seconds, on a laptop with an Intel Core i7 processor at 2.4GHz.

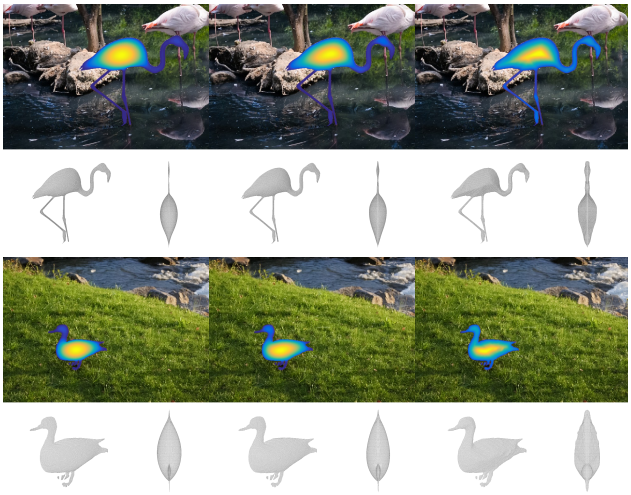

Fig. 3. **Qualitative comparison w.r.t. competing approaches**. In all cases, the displayed data are the same, including the images *flamingo*, and *mallard*. **Top:** From left to right it is represented the minimal surface estimation z by considering the approaches [3], [5] and ours, respectively. **Bottom:** Two novel 3D views.

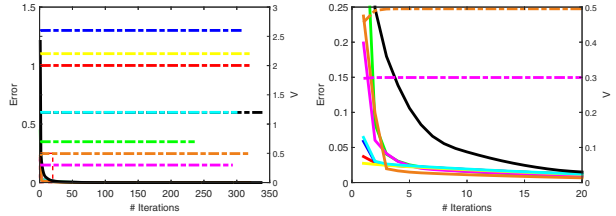

Fig. 4. **Convergence analysis and volume enforcement vs. number of iterations. Left:** Error evolution for the eight images considered in Fig. 2, and the volume enforcement V as a function of the iterations until convergence. Note that two different scales (left and right vertical axes) are used to represent the errors and volume evolution by using non-dashed and dashed lines, respectively. The correspondence between colors and pictures is: *cow* in red, *flamingo* in green, *camel* in blue, *mallard* in magenta, *rhino* in yellow, *black swan* in cyan, *bicycle* in orange, and *break dance* in black. **Right:** Zoom of the area within the red dashed rectangle in the left plot.

Monocular Videos. We now evaluate our approach on real monocular videos. It is worth noting that we only consider videos with fronto-parallel views of the object to be reconstructed, i.e., our approach does not need large camera motions to solve the problem. While this could be considered a limitation of our approach against rigid-structure-from-motion [6, 25], multi-view [26] and non-rigid-structure-from-motion approaches [9, 11, 27, 28], in the same way, this is also our great strength against those methods since in the absence of rigid motion they cannot be used. In other words, the previous frameworks rely on motion parallax to achieve a solution, while our approach can do it without that assumption.

Again, we rely on the DAVIS dataset [21] to get RGB videos of rigid and non-rigid objects. Particularly, we consider *boat* and *flamingo* classes for rigid and non-rigid shapes, respectively. Our results are summarized in Fig. 5. Despite not assuming any camera motion, our approach can accurately estimate these challenging shapes over time, a couple of estimations that have not been previously attempted, especially without assuming any training data at all. Some of the 2D correspondences we use to infer the solution are represented in the figure. As it can be seen, just a few correspondences –without assuming any tracking– are needed to apply our energy in Eq. (3). In these sequences, we can compute the error $\epsilon = \frac{\|z - z_{(1)}\|_{\mathcal{F}}}{\|z_{(1)}\|_{\mathcal{F}}}$ with z and $z_{(1)}$ the minimal solutions after applying Eqs. (3) and (1), respectively, denoting $\mathcal{F}$ a Frobenius norm. We obtain $\epsilon = 2.85 \cdot 10^{-4}$ and $\epsilon = 9.15 \cdot 10^{-5}$ for *boat* and *flamingo* classes, that means both estimations are quite similar. A visualization can be seen in the last rows of Fig. 5. Apart from that, our new energy in Eq. (3) exploits the previous frame estimation and reduces a 40% the computational cost w.r.t. the direct frame-by-frame application of Eq. (1), producing faster solutions.

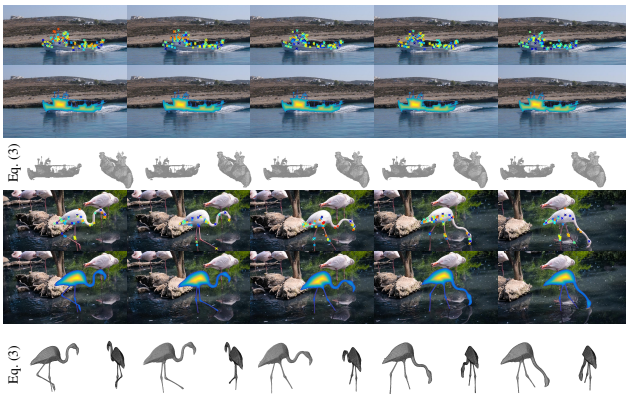

Fig. 5. **Qualitative evaluation for rigid and non-rigid objects on real videos in the wild**. In both cases, the same information is displayed for the categories *boat* and *flamingo*, respectively. **First row:** Images of the categories *boat* and *flamingo*. In both cases 2D correspondences with the previous frame are displayed in crosses. No tracking is assumed. **Second row:** Minimal detailed surfaces z with volume. **Third row:** Two novel views of our detailed, closed and volumetric 3D reconstruction. Best viewed in color.

4. CONCLUSION

We have proposed a convex method to retrieve the 3D geometry of an object from visual signals and without assuming any training data at all. To this end, an energy is optimized to be consistent with a pre-defined volume, while enforcing some geometrical priors to acquire fine details of the object. Our approach can be applied for single RGB images as well as for video sequences, where a temporal-aware solution is automatically enforced. Our solution is also fast in a commodity laptop even for dense estimations, and it obtains convergence within reasonable bounds, while satisfying perfectly the constraints. We have experimentally evaluated our approach on a wide variety of real scenarios, by using images in the wild where obtaining 3D training data could be a very hard task.

5. REFERENCES

[1] J. Martinez, R. Hossain, J. Romero, and J. J. Little, "A simple yet effective baseline for 3D human pose estimation," in *ICCV*, 2017.

[2] J. Ostlund, A. Varol, and P. Fua, "Laplacian meshes for monocular 3D shape recovery," in *ECCV*, 2012.

[3] M. R. Oswald, E. Toeppe, and D. Cremers, "Fast and globally optimal single view reconstruction of curved objects," in *CVPR*, 2012.

[4] E. Toeppe, M. R. Oswald, D. Cremers, and C. Rother, "Silhouette-based variational methods for single view reconstruction," in *ICVPCV*, 2011.

[5] S. Vicente and L. Agapito, "Balloon shapes: reconstructing and deforming objects with volume from images," in *3DV*, 2013.

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

[7] A. Agudo and F. Moreno-Noguer, "Combining local-physical and global-statistical models for sequential deformable shape from motion," *IJCV*, vol. 122, no. 2, pp. 371–387, 2017.

[8] A. Agudo and F. Moreno-Noguer, "Shape basis interpretation for monocular deformable 3D reconstruction," *TMM*, vol. 21, no. 4, pp. 821–834, 2019.

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

[10] I. Akhter, Y. Sheikh, S. Khan, and T. Kanade, "Trajectory space: A dual representation for nonrigid structure from motion," *TPAMI*, vol. 33, no. 7, pp. 1442–1456, 2011.

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

[12] A. O. Ulusoy, M. J. Black, and A. Geiger, "Patches, planes and probabilities: A non-local prior for volumetric 3D reconstruction," in *CVPR*, 2016.

[13] E. Grant, P. Kohli, and M. van Gerven, "Deep disentangled representations for volumetric reconstruction," in *ECCVW*, 2016.

[14] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "SMPL: A skinned multi-person linear model," *TOG*, vol. 34, no. 6, pp. 1–16, 2015.

[15] S. Zuffi, A. Kanazawa, and M. J. Black, "Lions and tigers and bears: Capturing non-rigid, 3D, articulated shape from images," in *CVPR*, 2018.

[16] S. Zuffi, A. Kanazawa, D. Jacobs, and M. J. Black, "3D menagerie: Modeling the 3D shape and pose of animals," in *CVPR*, 2017.

[17] T. J. Cashman and A. W. Fitzgibbon, "What shape are dolphins? building 3D morphable models from 2D images," *TPAMI*, vol. 35, no. 1, pp. 232–244, 2012.

[18] S. Parashar, D. Pizarro, A. Bartoli, and T. Collins, "As-rigid-as-possible volumetric shape-from-template," in *ICCV*, 2015.

[19] A. Agudo, "Unsupervised 3D reconstruction and grouping of rigid and non-rigid categories," *TPAMI*, vol. 44, no. 1, pp. 519–532, 2022.

[20] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, vol. 60, no. 2, pp. 91–110, 2004.

[21] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung, "A benchmark dataset and evaluation methodology for video object segmentation," in *CVPR*, 2016.

[22] J. Luiten, P. Voigtlaender, and B. Leibe, "PReMVOS: Proposal-generation, refinement and merging for video object segmentation," in *ACCV*, 2018.

[23] Z. Wang, J. Xu, L. Liu, F. Zhu, and L. Shao, "RANet: Ranking attention network for fast video object segmentation," in *ICCV*, 2019.

[24] T. Zhou, S. Wang, Y. Zhou, Y. Yao, J. Li, and L. Shao, "MATNet: Motion-attentive transition for zero-shot video object segmentation," in *AAAI*, 2020.

[25] C. Tomasi and T. Kanade, "Shape and motion from image streams under orthography: A factorization approach," *IJCV*, vol. 9, no. 2, pp. 137–154, 1992.

[26] S. M. Seitz, B. Curless ad J. Diebel, D. Scharstein, and R. Szeliski, "A comparison and evaluation of multi-view stereo reconstruction algorithms," in *CVPR*, 2006.

[27] A. Agudo, J. M. M. Montiel, L. Agapito, and B. Calvo, "Modal space: A physics-based model for sequential estimation of time-varying shape from monocular video," *JMIV*, vol. 57, no. 1, pp. 75–98, 2017.

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