Temporally coherent 4D reconstruction of complex dynamic scenes

Armin Mustafa Hansung Kim Jean-Yves Guillemaut Adrian Hilton
CVSSP, University of Surrey, Guildford, United Kingdom
a.mustafa@surrey.ac.uk

Abstract

This paper presents an approach for reconstruction of 4D temporally coherent models of complex dynamic scenes. No prior knowledge is required of scene structure or camera calibration allowing reconstruction from multiple moving cameras. Sparse-to-dense temporal correspondence is integrated with joint multi-view segmentation and reconstruction to obtain a complete 4D representation of static and dynamic objects. Temporal coherence is exploited to overcome visual ambiguities resulting in improved reconstruction of complex scenes. Robust joint segmentation and reconstruction of dynamic objects is achieved by introducing a geodesic star convexity constraint. Comparative evaluation is performed on a variety of unstructured indoor and outdoor dynamic scenes with hand-held cameras and multiple people. This demonstrates reconstruction of complete temporally coherent 4D scene models with improved non-rigid object segmentation and shape reconstruction.

1. Introduction

Existing reconstruction frameworks for general dynamic scenes commonly operate on a frame-by-frame basis [14, 32] or are limited to simple scenes [15]. Previous work on indoor and outdoor dynamic scene reconstruction has shown that joint segmentation and reconstruction across multiple views gives improved reconstruction [17]. In this work we build on this concept exploiting temporal coherence of the scene to overcome visual ambiguities inherent in single frame reconstruction and multiple view segmentation methods for general scenes. This is illustrated in Figure 1 where the resulting 4D scene reconstruction has temporally coherent labels and surface correspondence for each object.

We present a sparse-to-dense approach to estimate dense temporal correspondence and surface reconstruction for non-rigid objects. Initially sparse 3D feature points are robustly tracked from wide-baseline image correspondence using spatio-temporal information to obtain sparse temporal correspondence and reconstruction. Sparse 3D feature correspondences are used to constrain optical flow estimation to obtain an initial dense temporally consistent model of dynamic regions. The initial model is then refined using a novel optimisation framework using a geodesic star convexity constraint for simultaneous multi-view segmentation and reconstruction of non-rigid shape. The proposed approach overcomes limitations of existing methods allowing an unsupervised temporally coherent 4D reconstruction of complete models for general scenes. The scene is automatically decomposed into a set of spatio-temporally coherent objects as shown in Figure 1. The contributions are as follows:

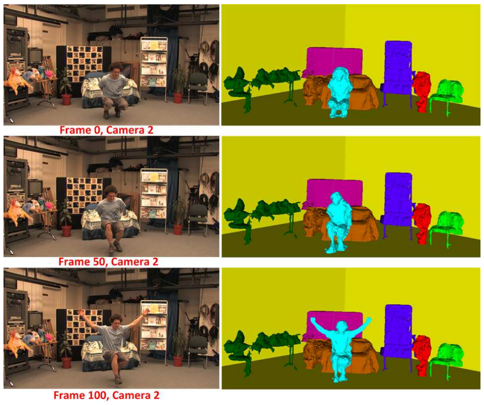

Figure 1. Temporally consistent scene reconstruction for Odzemok dataset colour-coded to show the obtained scene segmentation.

- Temporally coherent reconstruction of complex dynamic scenes.
- A framework for space-time sparse-to-dense segmentation and reconstruction.
- Optimisation of dense reconstruction and segmentation using geodesic star convexity.
- Robust and computationally efficient reconstruction of dynamic scenes by exploiting temporal coherence.

2. Related work

2.1. Temporal multi-view reconstruction

Extensive research has been performed in multi-view reconstruction of dynamic scenes. Most existing approaches process each time frame independently due to the difficulty of simultaneously estimating temporal correspondence for non-rigid objects. Independent per-frame reconstruction can result in errors due to the inherent visual ambiguity caused by occlusion and similar object appearance for general scenes. Quantitative evaluation of state-of-the-art techniques for static object reconstruction from multiple views was presented [39]. Research investigating spatio-temporal reconstruction across multiple frames [15, 18] requires accurate initialisation, is limited to simple scenes and does not produce temporally coherent 4D models. A number of approaches that use temporal information [2, 30, 28] either require a large number of closely spaced cameras or bi-layer segmentation [46, 25] as a constraint for complete reconstruction. Other approaches for reconstruction of general scenes from multiple handheld wide-baseline cameras [3, 41] exploit prior reconstruction of the background scene to allow dynamic foreground segmentation and reconstruction. Recent approaches for spatio-temporal reconstruction of multi-view data either work on indoor studio data [35] or for dynamic reconstruction of crowd sourced data [24].

Methods to estimate 3D scene flow have been reported in the literature [31]. However existing approaches are limited to narrow baseline correspondence for dynamic scenes. Scene flow approaches dependent on optical flow [42, 4] require an accurate estimate for most of the pixels which fails in the case of large motion. The approach presented in this paper is for general dynamic indoor or outdoor scenes with large non-rigid motions and no prior knowledge of scene structure. Temporal correspondence and reconstruction are simultaneously estimated to produce a 4D model of the complete scene with both static and dynamic objects.

2.2. Multi-view video segmentation

In the field of image segmentation, approaches have been proposed to provide impressive temporally consistent video segmentation [16, 37, 34, 45]. Hierarchical segmentation based on graphs was proposed in [16], directed acyclic graph were used to propose an object followed by segmentation in [45] and [37, 34] used optical flow. All of these methods work only for monocular videos. Recently a number of approaches have been proposed for multi-view foreground object segmentation by exploiting appearance similarity [12, 11, 27, 29, 44] . These approaches assume a static background and different colour distributions for the foreground and background which limits applicability for general complex scenes and non-rigid objects.

To address this issue we introduce a novel method for spatio-temporal multi-view segmentation of dynamic scenes using shape constraints. Single image segmentation techniques using shape constraints provide good results for complex scene segmentation [19](convex and concave shapes), but requires manual interaction. The proposed approach performs multi-view video segmentation by initializing the foreground object model using spatio-temporal information from wide-baseline feature correspondence followed by a multi-layer optimization framework using geodesic star convexity to constrain the segmentation. Our multi-view formulation naturally enforces coherent segmentation between views and also resolves ambiguities such as the similarity of background and foreground.

2.3. Joint segmentation and reconstruction

Joint segmentation and reconstruction methods simultaneously estimate multi-view segmentation or matting with reconstruction and have been shown to given improved performance for complex scenes. A number of approaches have been introduced for joint optimization. However, these are either limited to static scenes [43, 20] or process each frame independently thereby failing to enforce temporal consistency [8, 32, 17]. A joint formulation for multi-view video was proposed for sports data and indoor sequences in [17] and for challenging outdoor scenes in [32]. Recent work proposed joint reconstruction and segmentation on monocular video achieving semantic segmentation of static scenes. Other joint segmentation and reconstruction approaches that use temporal information based on patch refinement [40, 36] work only for rigid objects. An approach based on optical flow and graph cuts was shown to work well for non-rigid objects in indoor settings but requires silhouettes and is computationally expensive [18]. Practical application of temporally coherent joint estimation requires approaches that work on non-rigid objects for general scenes in uncontrolled environments.

The proposed approach overcomes the limitations of previous methods enabling robust wide-baseline spatio-temporal reconstruction and segmentation of general scenes. Temporal correspondence is exploited to overcome visual ambiguities giving improved reconstruction together with temporally coherent 4D scene models.

3. Methodology

This work is motivated by the limitations of existing multiple view reconstruction methods which either work independently at each frame resulting in errors due to visual ambiguity and occlusion [14, 17, 32], or commonly require restrictive assumptions on scene complexity and structure [41, 18]. We address these issues by introducing temporal coherence in the reconstruction to reduce ambiguity, ensure consistent non-rigid structure initialisation at successive frames and improve reconstruction quality.

3.1. Overview

A novel automatic multi-object dynamic segmentation and reconstruction method based on the geodesic star-convexity shape constraint is proposed to obtain a 4D model of the scene including both dynamic and static objects. An overview of the framework is presented in Figure 2 :

Sparse reconstruction: The input to the system is multiple view wide-baseline video with known camera intrinsics. Extrinsic parameters are calibrated automatically [21, 23] using sparse wide-baseline feature matching. Segmentation-based feature detection (SFD) [33] is used to

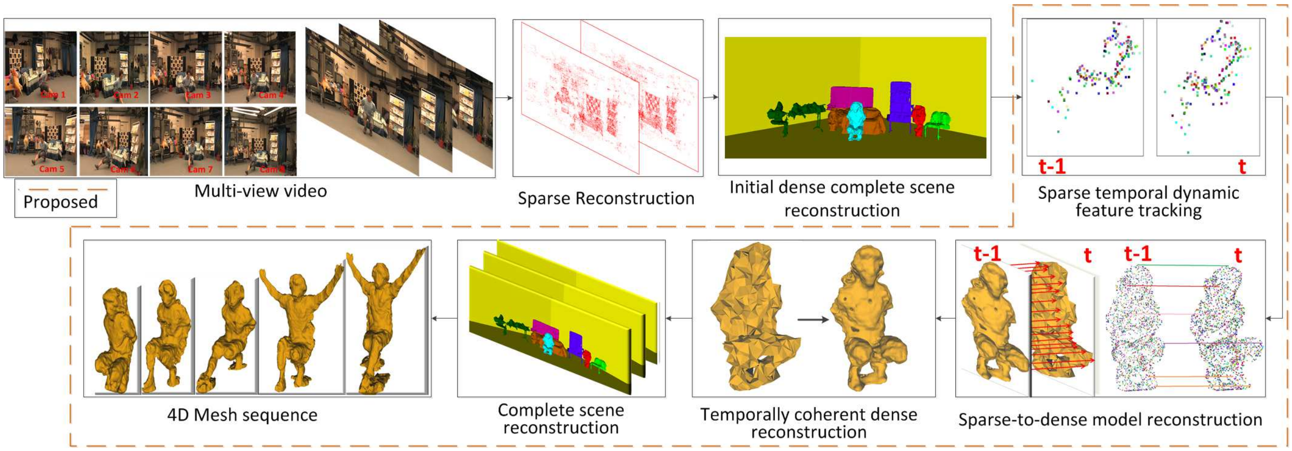
Cam 1
Cam 2
Cam 3
Cam 4
Cam 5
Cam 6
Cam 7
Cam 8
Multi-view video
Proposed
Sparse Reconstruction
Initial dense complete scene reconstruction
t-1
t
Sparse temporal dynamic feature tracking
t-1
t
t-1
t
Sparse-to-dense model reconstruction
Temporally coherent dense reconstruction
Complete scene reconstruction
4D Mesh sequence

Figure 2. Temporally consistent scene reconstruction framework

obtain a relatively large number of sparse features suitable for wide-baseline matching which are distributed throughout the scene including on dynamic objects such as people. SFD features are matched between views using a SIFT descriptor giving a sparse 3D point-cloud and camera extrinsics for each time instant. The sparse point cloud is clustered in 3D [38] with each cluster representing a unique foreground object. Objects with insufficient detected features are reconstructed as part of the scene background.

Initial dense complete scene reconstruction: Sparse reconstruction at each time instant is clustered in 3D[38] to obtain an initial coarse object segmentation. Delaunay triangulation [13] is performed on the set of back projected sparse features for each object in the camera image plane with best visibility. This is propagated to the other views using the sparse feature matching to obtain an initial object reconstruction. This reconstruction is refined using the framework explained in Section 3.3 to obtain segmentation and dense reconstruction of each object.

Accurate reconstruction of the background object is often challenging due to the lack of features, repetitive texture, occlusion, textureless regions and relatively narrow baseline for distant objects. Hence we create a rough geometric proxy of the background by computing the minimum oriented bounding box for the sparse 3D point cloud using principal component analysis (PCA) [10]. The dense reconstruction of the foreground objects and background are combined to obtain a full scene reconstruction at the first time instant. For consecutive time instants only dynamic objects are reconstructed with the segmentation and reconstruction of static objects retained which reduces computational complexity.

Temporally coherent reconstruction of dynamic objects: Dynamic object regions are detected at each time instant by sparse temporal correspondence of SFD features at successive frames. Sparse temporal feature correspondence allows propagation of the dense reconstruction for each dynamic object to obtain an initial approximation (Section 3.2). The initial estimate is refined using a joint optimisation of segmentation and reconstruction based on geodesic star convexity (Section 3.3). A single 3D model for each dynamic object is obtained by fusion of the view-dependent depth maps using Poisson surface reconstruction [26].

Subsequent sections present the novel contributions of this work in identifying the dynamic points, initialisation using space-time information and refinement using geodesic star convexity to obtain a dense reconstruction. The approach is demonstrated to outperform state-of-the-art dynamic scene reconstruction and gives a temporally coherent 4D model.

3.2. Initial temporally coherent reconstruction

Once the static scene reconstruction is obtained for the first frame, we perform temporally coherent dynamic scene reconstruction at successive time instants. Dynamic regions are identified using temporal correspondence of sparse 3D features. These points are used to obtain an initial dense model for the dynamic objects using optical flow. The initial coarse reconstruction for each dynamic region is refined in the subsequent optimization step with respect to each camera view. Dynamic scene objects are identified from the temporal correspondence of sparse feature points. Sparse correspondence is then used to propagate an initial model of the moving object for refinement. Figure 3 presents the sparse reconstruction and temporal correspondence.

Sparse temporal dynamic feature tracking: Numerous approaches have been proposed to track moving objects in 2D using either features or optical flow. However these methods may fail in the case of occlusion, movement parallel to the view direction, large motions and moving cameras. To overcome these limitations we match the sparse 3D feature points obtained using SFD from multiple wide-baseline views at each time instant. The use of sparse 3D features is robust to large non-rigid motion, occlusions and camera movement. SFD [33] detects sparse features which are stable across wide-baseline views and consecutive time instants for a moving camera and dynamic scene. Sparse 3D feature matches between consecutive time instants are

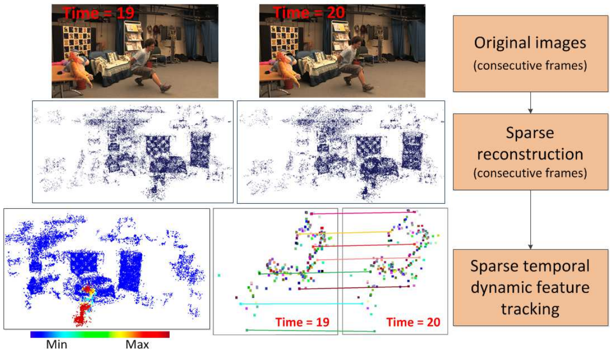

Figure 3. Sparse temporal dynamic feature tracking algorithm: Results on two datasets; Min and Max is the minimum and maximum movement in the 3D points respectively.

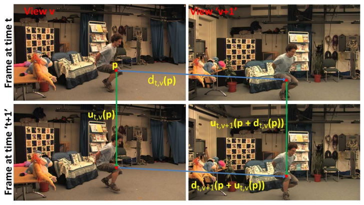

Figure 4. Spatio-temporal consistency check for 3D tracking

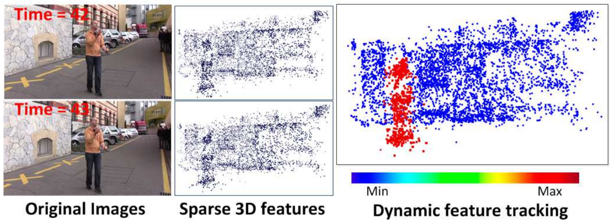

Figure 5. Sparse dynamic feature tracking for Juggler dataset.

back-projected to each view. These features are matched temporally using a SIFT descriptor to identify the moving points. Robust matching is achieved by enforcing multiple view consistency for the temporal feature correspondence in each view as illustrated in Figure 4. Each match must satisfy the constraint:

$$\|d_{t,v}(p) + u_{t,v+1}(p + d_{t,v}(p)) - u_{t,v}(p) - d_{t,v+1}(p + u_{t,v}(p))\| < \epsilon$$

where p is the feature image point in view v at frame t, $d_{t,v}(p)$ is the disparity at frame t from view v to $v + 1$, $u_{t,v}(p)$ is the temporal correspondence from frames t to $t + 1$ for view v. The multi-view consistency check ensures that correspondences between any two views remain temporally consistent for successive frames. Matches in the 2D domain are sensitive to camera movement and occlusion, hence we map the set of refined matches into 3D to make the system robust to camera motion. The Frobenius norm is applied on the 3D point gradients in all directions [45] to obtain the 'net' motion at each sparse point. The 'net' motion between pairs of 3D points for consecutive time instants are ranked, and the top and bottom 5 percentile values removed. Median filtering is then applied to identify the dynamic features. Figure 5 shows an example with moving cameras.

Sparse-to-dense model reconstruction: Dynamic 3D feature points are used to initialize the segmentation and reconstruction of the initial model. This avoids the assumption of static backgrounds and prior scene segmentation commonly used to initialise multiple view reconstruction with a coarse visual-hull approximation [17]. Temporal coherence also provides a more accurate initialisation to overcome visual ambiguities at individual frames. Figure 6 illustrates the use of temporal coherence for reconstruction initialization and refinement. Dynamic feature correspondence is used to identify the mesh for each dynamic object. This mesh is back projected on each view to obtain the region of interest. Optical flow [5] is performed on the projected mask for each view in the temporal domain using the dynamic feature correspondences over time as initialization. Dense multi-view wide-baseline correspondences from the previous frame are propagated to the current frame using the information from the flow vectors to obtain dense multi-view matches in the current frame. The matches are triangulated in 3D to obtain a refined 3D dense model of the dynamic object for the current frame.

For dynamic scenes, a new object may enter the scene or a new part may appear as the object moves. To allow the introduction of new objects and object parts we also use information from the cluster of sparse points for each dynamic object. The cluster corresponding to the dynamic features is identified and static points are removed. This ensures that the set of new points not only contain the dynamic features but also the unprocessed points which represent new parts of the object. These points are added to the refined sparse model of the dynamic object. To handle the new objects we detect new clusters at each time instant and consider them as dynamic regions.

Once we have a set of dense 3D points for each dynamic object, Poisson surface reconstruction is performed on the set of sparse points to obtain an initial coarse model of each dynamic region $\mathscr{R}$, which is subsequently refined using the optimization framework (Section 3.3).

3.3. Temporally coherent dense reconstruction

The initial reconstruction and segmentation from dense temporal feature correspondence is refined using a joint optimization framework. A novel shape constraint is introduced based on geodesic star convexity which has previ-

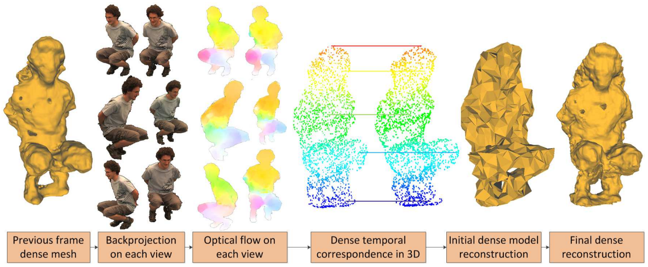
Previous frame dense mesh
Backprojection on each view
Optical flow on each view
Dense temporal correspondence in 3D
Initial dense model reconstruction
Final dense reconstruction

Figure 6. Initial sparse-to-dense model reconstruction workflow

ously been shown to give improved performance in interactive image segmentation for structures with fine details (for example a persons fingers or hair)[19]. In this work the shape constraint is automatically initialised for each view from the initial segmentation. The geodesic star-convexity is integrated as a constraint on the energy minimisation for joint multi-view reconstruction and segmentation [17]. The shape constraint is based on the geodesic distance with foreground object initialisation (seeds) as star centres to which the object shape is restricted. The union formed by multiple object seeds form a geodesic forest. This allows complex shapes to be segmented. In this work to automatically initialize the segmentation we use the sparse temporal feature correspondence as star centers (seeds) to build a geodesic forest automatically. The region outside the initial coarse reconstruction of all dynamic objects is initialized as the background seed for segmentation as shown in in Figure 7. The shape of the dynamic object is restricted by this geodesic distance constraint that depends on the image gradient. Comparison with existing methods for multi-view segmentation demonstrates improvements in recovery of fine detail structure as illustrated in Figure 7.

3.3.1 Optimization based on geodesic star convexity

The depth of the initial coarse reconstruction estimate is refined for each dynamic object at a per pixel level. Our goal is to assign an accurate depth value from a set of depth values $\mathscr{D} = \left\{d_1, ..., d_{|\mathscr{D}|-1}, \mathscr{U}\right\}$ and assign a layer label from a set of label values $\mathscr{L} = \left\{l_1, ..., l_{|\mathscr{L}|}\right\}$ to each pixel p for the region $\mathscr{R}$ of each dynamic object. Each d_i is obtained by sampling the optical ray from the camera and $\mathscr{U}$ is an unknown depth value to handle occlusions. This is achieved by optimisation of a joint cost function [17] for label (segmentation) and depth (reconstruction):

$$E(l,d) = \lambda_{data} E_{data}(d) + \lambda_{contrast} E_{contrast}(l) + \lambda_{smooth} E_{smooth}(l,d) + \lambda_{color} E_{color}(l) \quad (1)$$

where, d is the depth at each pixel, l is the layer label for multiple objects and the cost function terms are defined in section 3.3.2. This is solved subject to a geodesic star-convexity constraint on the labels l. A label l is star convex with respect to center c_i, if every point $p \in l$ is visible to a star center c_i in set $\mathscr{C} = \{c_1, ..., c_n\}$ via l in the image x, where n is the number of star centers[19]. This is expressed as an energy cost:

$$E^\star(l|x, \mathscr{C}) = \sum_{p \in \mathscr{R}} \sum_{q \in \Gamma_{c,p}} E^\star_{p,q}(l_p, l_q) \quad (2)$$

$$\forall q \in \Gamma_{c,p},\ E^\star_{p,q} = \begin{cases} \infty \text{ if } l_p \neq l_q \\ 0 \text{ otherwise} \end{cases} \quad (3)$$

where $\forall p \in \mathscr{R} : p \in l \Leftrightarrow l_p = 1$ and $\Gamma_{c,p}$ is the geodesic path joining p to any star center in set $\mathscr{C}$ given by:

$$\Gamma_{c,p} = \arg \min_{\Gamma \in \mathscr{P}_{c,p}} \mathscr{L}(\Gamma) \quad (4)$$

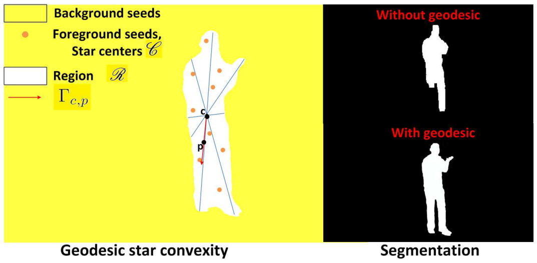

Figure 7. Geodesic star convexity: A region $\mathscr{R}$ with star centers $\mathscr{C}$ connected with geodesic distance $\Gamma_{c,p}$. Segmentation results with and without geodesic star convexity based optimization are shown on the right for the Juggler dataset.

where $\mathscr{P}_{c,p}$ denotes the set of all discrete paths between c and p and $\mathscr{L}(\Gamma)$ is the length of discrete geodesic path as defined in [19]. In our case we define the temporal sparse feature correspondences as star centers, hence the segmentation will include all the points which are visible to these sparse features via geodesic distances in the region $\mathscr{R}$, thereby employing the shape constraint. Since the star centers are selected automatically, the method is unsupervised. The energy in the Eq. 1 is minimized as follows:

$$\underset{s.t.}{min_{(l,d)}} \underset{l \epsilon S^\star(\mathscr{C})}{E(l,d)} \Leftrightarrow \min_{(l,d)} E(l,d) + E^\star(l|x, \mathscr{C}) \quad (5)$$

where $S^\star(\mathscr{C})$ is the set of all shapes which lie within the geodesic distances wrt to the centers in $\mathscr{C}$. Optimization of eq. 5, subject to each pixel p in the region $\mathscr{R}$ being at a geodesic distance from the star centers in the set $\mathscr{C}$, is performed using the α-expansion algorithm for a pixel p by iterating through the set of labels in $\mathscr{L} \times \mathscr{D}$ [7]. Graph-cut is used to obtain a local optimum [6].

3.3.2 Energy cost function

For completeness in this section we define each of the terms in eq. 1, these are based on previous terms used for joint optimisation over depth for each pixel introduced in [32], with modification of the color matching term to improve robustness and extension to multiple labels.

Matching term: The data term for matching between views is specified as a measure of photo-consistency as follows:

$$E_{data}(d) = \textstyle\sum_{p \in \mathscr{P}} e_{data}(p, d_p) = \begin{cases} M(p,q) = \sum_{i \in \mathscr{O}_k} m(p,q), & \text{if } d_p \neq \mathscr{U} \\ M_{\mathscr{U}}, & \text{if } d_p = \mathscr{U} \end{cases} \quad (6)$$

where $\mathscr{P}$ is the 4-connected neighbourhood of pixel p, $M_{\mathscr{U}}$ is the fixed cost of labelling a pixel unknown and q denotes the projection of the hypothesised point P in an auxiliary camera where P is a $3D$ point along the optical ray passing through pixel p located at a distance d_p from the reference camera. $\mathscr{O}_k$ is the set of k most photo-consistent pairs with reference camera and $m(p,q)$ is inspired from [22].

Contrast term: The contrast term is as follows:

$$E_{contrast}(l) = \sum_{p,q \in \mathscr{N}} e_{contrast}(p, q, l_p, l_q) \quad (7)$$

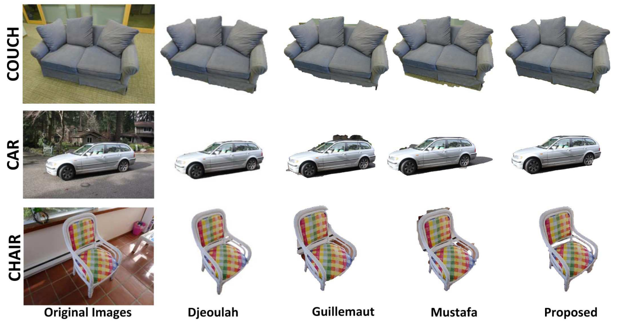

Figure 8. Comparison of segmentation on benchmark static datasets using geodesic star-convexity.

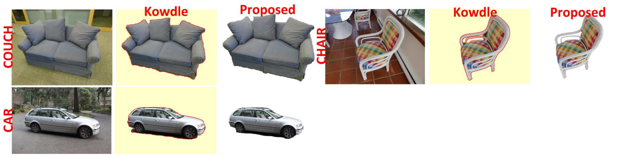

Figure 9. Comparison of segmentation with Kowdle.

$$e_{contrast}(p,q,l_p,l_q) = \begin{cases} 0, & \text{if } (l_p = l_q) \\ \frac{1}{1+\epsilon}(\epsilon + exp^{-C(p,q)}), & \text{otherwise} \end{cases} \tag{8}$$

Smoothness term: This term is defined as:

$$E_{smooth}(l,d) = \sum_{(p,q)\in\mathscr{N}} e_{smooth}(l_p,d_p,l_q,d_q) \tag{9}$$

$$e_{smooth}(l_p,d_p,l_q,d_q) =$$
$$\begin{cases} min(|d_p - d_q|, d_{max}), & \text{if } l_p = l_q \text{ and } d_p, d_q \neq \mathscr{U} \\ 0, & \text{if } l_p = l_q \text{ and } d_p, d_q = \mathscr{U} \\ d_{max}, & \text{otherwise} \end{cases} \tag{10}$$

d_{max} is set to 50 times the size of the depth sampling step defined in Section 3.3.1 for all datasets.

Color term: This term is computed using the negative log likelihood [6] of the color models learned from the foreground and background markers. The star centers obtained from the sparse 3D features are foreground markers and for background markers we consider the region outside the projected initial coarse reconstruction for each view. The color models use GMMs with 5 components each for FG/BG mixed with uniform color models [9] as the markers are sparse.

$$E_{color}(l) = \sum_{p\in\mathscr{P}} -logP(I_p|l_p) \tag{11}$$

where $P(I_p|l_p = l_i)$ denotes the probability at pixel p in the reference image belonging to layer l_i.

	λ_{data}	λ_c	λ_{smooth}	λ_{color}
Magician/Dance2	0.4	5.0	.0005	0.6
Juggler	0.5	5.0	.0005	0.4
Odzemok/Dance1/Office	0.4	3.0	.001	0.6

Table 2. Parameters used for all datasets: λ_c represents $\lambda_{contrast}$

4. Results and Performance Evaluation

The proposed system is tested on publicly available multi-view research datasets of indoor and outdoor scenes: static data for segmentation comparison Couch, Chair and Car[27]; and dynamic data for full evaluation Dance2[1], Office[1], Dance1[1], Odzemok[1], Magician and Juggler [3]. Dance1, Dance2 and Office are captured from 8 static cameras, Odzemok from 6 static and 2 moving cameras and Magician and Juggler from 6 moving handheld cameras. More information is available on the website[2].

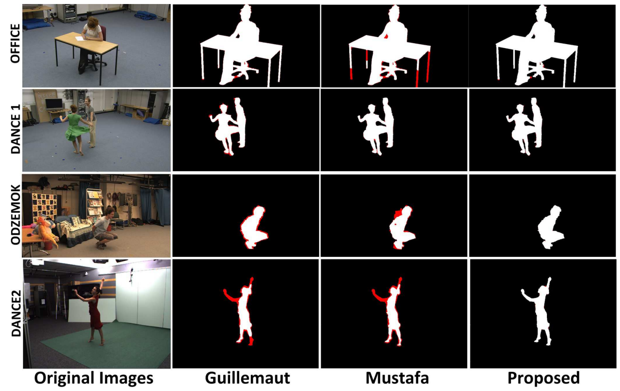

Figure 10. Segmentation results for dynamic scenes (Error against ground-truth is highlighted in red).

4.1. Multi-view segmentation evaluation

Segmentation is evaluated against the state-of-the-art methods for multi-view segmentation Kowdle[27] and Djelouah[11] for static scenes and joint segmentation reconstruction per frame Mustafa[32] and using temporal information Guillemaut[18] for both static and dynamic scenes. For static multi-view data the segmentation is initialised as detailed in Section 3.1 followed by refinement using the constrained optimisation Section 3.3. For dynamic scenes the full pipeline with temporal coherence is used as detailed in 3. Ground-truth is obtained by manually labelling the foreground for Office, Dance1 and Odzemok dataset, and for other datasets ground-truth is available online. We initialize all approaches by the same proposed initial coarse reconstruction for fair comparison.

To evaluate the segmentation we measure completeness as the ratio of intersection to union with ground-truth[27]. Comparisons are shown in Table 1 and Figure 8 and 9 for static benchmark datasets and in Table 3 and Figure 10 and 11 for dynamic scenes. Results for multi-view segmentation of static scenes are more accurate than Djelouah, Mustafa and Guillemaut and comparable to Kowdle with improved segmentation of some detail such as the back of the chair.

For dynamic scenes the geodesic star convexity based optimization together with temporal consistency gives improved segmentation of fine detail such as the legs of the table in the Office dataset and limbs of the person in the Juggler, Magician and Dance2 datasets in Figure 10 and 11. This overcomes limitations of previous multi-view per-frame segmentation.

[1] http://cvssp.org/data/

[2] http://cvssp.org/projects/4d/4DRecon/

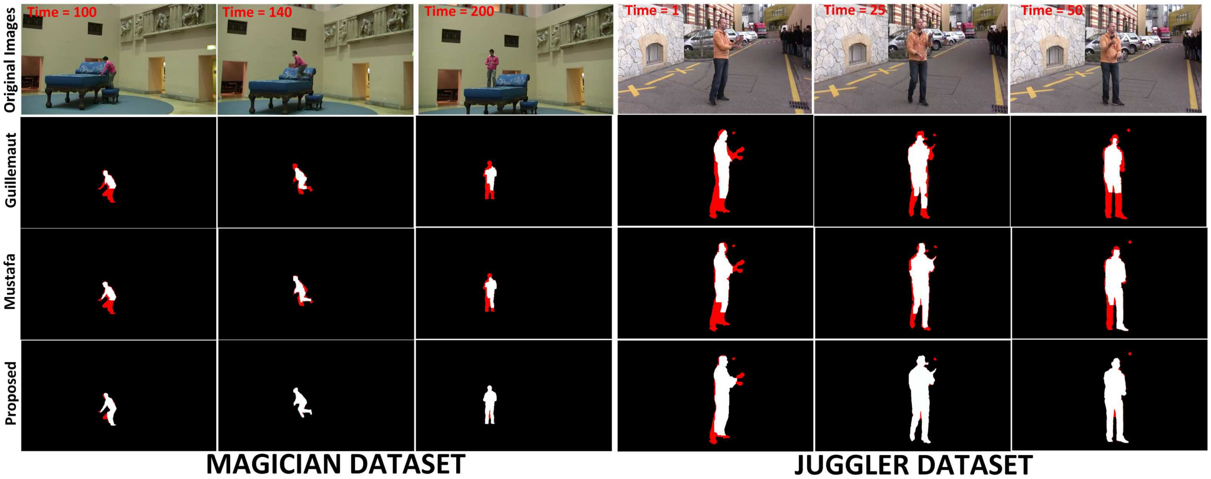

Figure 11. Segmentation results for dynamic scenes on sequence of frames (Error against ground-truth is highlighted in red).

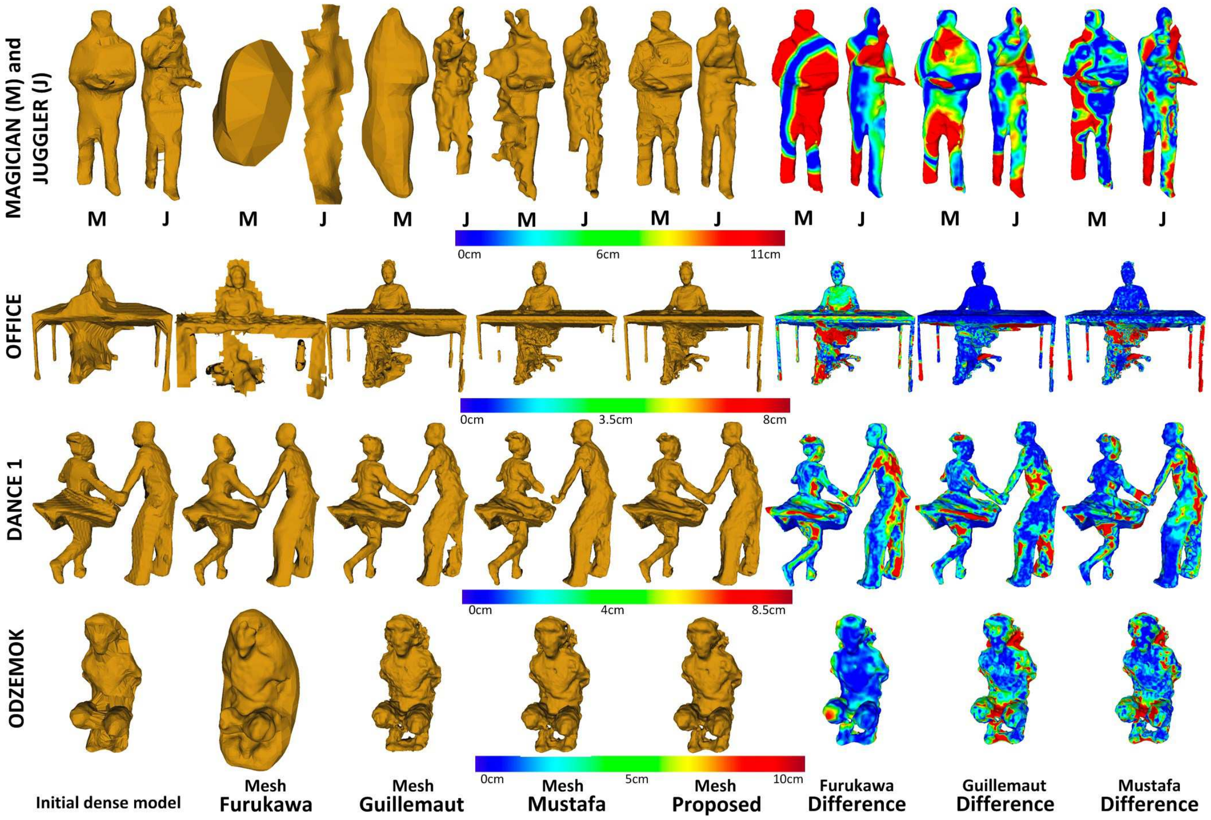

Figure 12. Reconstruction result mesh comparison

Dataset	Guillemaut	Mustafa	Proposed
Magician	68.0 ± 0.7	88.7 ± 0.5	**91.2 ± 0.2**
Juggler	84.6 ± 0.6	87.9 ± 0.6	**93.3 ± 0.2**
Odzemok	90.1 ± 0.3	89.9 ± 0.3	**91.8 ± 0.2**
Dance1	99.2 ± 0.5	99.4 ± 0.2	**99.5 ± 0.2**
Office	99.3 ± 0.4	99.0 ± 0.3	**99.4 ± 0.2**
Dance2	98.6 ± 0.3	99.0 ± 0.2	**99.0 ± 0.2**

Table 3. Dynamic scene segmentation completeness in %

4.2. Reconstruction evaluation

Reconstruction results obtained using the proposed method with parameters defined in Table 2 are compared against Mustafa[32], Guillemaut[18], and Furukawa [14] for dynamic sequences. Furukawa [14] is a per-frame multi-view wide-baseline stereo approach which ranks highly on the middlebury benchmark [39] but does not refine the segmentation.Figure 12 and 13 present qualitative and quantitative comparison of our method with the state-of-the-art approaches. Comparison of reconstructions demonstrates that the proposed method gives consistently more complete and accurate models. The colour maps highlight the quantitative differences in reconstruction. As far as we are aware no ground-truth data exist for dynamic scene reconstruc-

Dataset	Number of Views	Kowdle	Djelouah	Guillemaut	Mustafa	Proposed
Couch	11	99.6 ± 0.1	99.0 ± 0.2	97.0 ± 0.3	98.5 ± 0.2	**99.7 ± 0.3**
Chair	18	**99.2 ± 0.4**	98.6 ± 0.3	97.9 ± 0.5	98.0 ± 0.5	99.1 ± 0.3
Car	44	98.0 ± 0.7	97.0 ± 0.8	95.0 ± 0.7	97.6 ± 0.3	**98.6 ± 0.4**

Table 1. Static segmentation completeness comparison with existing methods on benchmark datasets

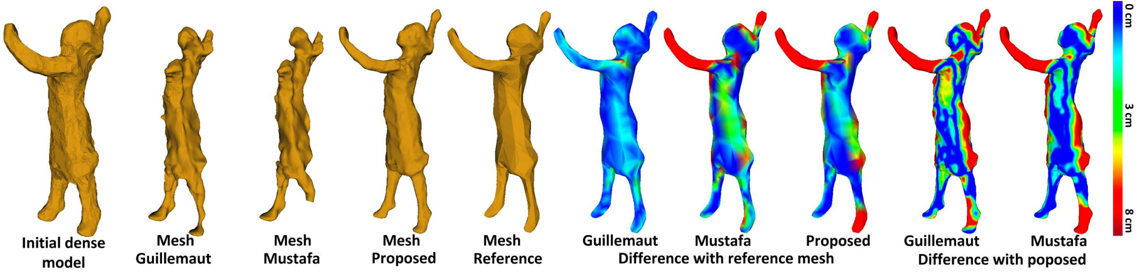

Figure 13. Reconstruction result comparison with reference mesh and proposed for Dance2 benchmark dataset

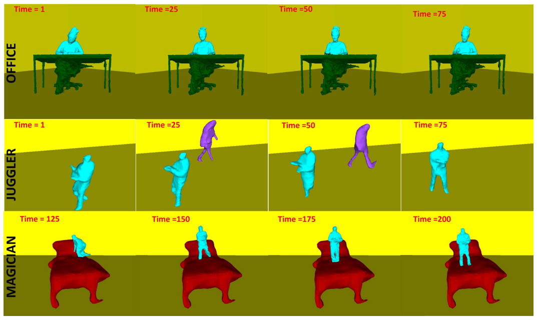

Figure 14. Complete scene reconstruction with 4D mesh sequence.

Dataset	Furukawa	Guillemaut	Mustafa	Ours
Dance1	326 s	493 s	295 s	**254 s**
Magician	**311 s**	608 s	377 s	325 s
Odzemok	381 s	598 s	394 s	**363 s**
Office	339 s	533 s	347 s	**291 s**
Juggler	394 s	634 s	411 s	**378 s**
Dance2	312 s	432 s	323 s	**278 s**

Table 4. Comparison of computational efficiency for dynamic datasets (time in seconds (s))

tion from real multi-view video. In Figure 13 we present a comparison with the reference mesh available with the Dance2 dataset reconstructed using a visual-hull approach. This comparison demonstrates improved reconstruction of fine detail with the proposed technique.

In contrast to all previous approaches the proposed method gives temporally coherent 4D model reconstructions with dense surface correspondence over time. The introduction of temporal coherence constrains the reconstruction in regions which are ambiguous on a particular frame such as the right leg of the juggler in Figure 12 resulting in more complete shape. Figure 14 shows three complete scene reconstructions with 4D models of multiple objects. The Juggler and Magician sequences are reconstructed from moving hand-held cameras.

Computation times for the proposed approach vs other methods are presented in Table 4. The proposed approach to reconstruct temporally coherent 4D models is comparable in computation time to per-frame multiple view reconstruction and gives a ∼50% reduction in computation cost compared to previous joint segmentation and reconstruction approaches using a known background. This efficiency is achieved through improved per-frame initialisation based on temporal propagation and the introduction of the geodesic star constraint in joint optimisation. Further results can be found in the supplementary material.

5. Conclusion

This paper present a framework for temporally coherent 4D model reconstruction of dynamic scenes from a set of wide-baseline moving cameras. The approach gives a complete model of all static and dynamic non-rigid objects in the scene. Temporal coherence for dynamic objects addresses limitations of previous per-frame reconstruction giving improved reconstruction and segmentation together with dense temporal surface correspondence for dynamic objects. A sparse-to-dense approach is introduced to establish temporal correspondence for non-rigid objects using robust sparse feature matching to initialise dense optical flow providing an initial segmentation and reconstruction. Joint refinement of object reconstruction and segmentation is then performed using a multiple view optimisation with a novel geodesic star convexity constraint that gives improved shape estimation and is computationally efficient. Comparison against state-of-the-art techniques for multiple view segmentation and reconstruction demonstrates significant improvement in performance for complex scenes. The approach enables reconstruction of 4D models for complex scenes which has not been demonstrated previously.

Limitations: As with previous dynamic scene reconstruction methods the proposed approach has a number of limitations: persistent ambiguities in appearance between objects will degrade the improvement achieved with temporal coherence; scenes with a large number of inter-occluding dynamic objects will degrade performance; the approach requires sufficient wide-baseline views to cover the scene.

Acknowledgements:
This research was supported by the European Commission, FP7 Intelligent Management Platform for Advanced Real-time Media Processes project (grant 316564).

References

[1] 4d repository, http://4drepository.inrialpes.fr/. In *Institut national de recherche en informatique et en automatique (INRIA) Rhone Alpes*. 6

[2] C. Bailer, B. Taetz, and D. Stricker. Flow fields: Dense correspondence fields for highly accurate large displacement optical flow estimation. In *ICCV*, 2015. 2

[3] L. Ballan, G. J. Brostow, J. Puwein, and M. Pollefeys. Unstructured video-based rendering: Interactive exploration of casually captured videos. *ACM Trans. on Graph.*, pages 1–11, 2010. 2, 6

[4] T. Basha, Y. Moses, and N. Kiryati. Multi-view scene flow estimation: A view centered variational approach. In *CVPR*, pages 1506–1513, 2010. 2

[5] J. Bouguet. Pyramidal implementation of the lucas kanade feature tracker. *Intel Corporation, Microprocessor Research Labs*, 2000. 4

[6] Y. Boykov and V. Kolmogorov. An experimental comparison of min-cut/max- flow algorithms for energy minimization in vision. *PAMI*, 26:1124–1137, 2004. 5, 6

[7] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *PAMI*, 23:1222–1239, 2001. 5

[8] N. Campbell, G. Vogiatzis, C. Hernndez, and R. Cipolla. Automatic 3d object segmentation in multiple views using volumetric graph-cuts. *Image and Vision Computing*, 28:14 – 25, 2010. 2

[9] P. Das, O. Veksler, V. Zavadsky, and Y. Boykov. Semiautomatic segmentation with compact shape prior. *Image and Vision Computing*, 27:206–219, 2009. 6

[10] D. Dimitrov, C. Knauer, K. Kriegel, and G. Rote. On the bounding boxes obtained by principal component analysis, 2006. 3

[11] A. Djelouah, J.-S. Franco, E. Boyer, F. Le Clerc, and P. Perez. Multi-view object segmentation in space and time. In *ICCV*, pages 2640–2647, 2013. 2, 6

[12] A. Djelouah, J.-S. Franco, E. Boyer, F. Le Clerc, and P. Perez. Sparse multi-view consistency for object segmentation. *PAMI*, pages 1–1, 2015. 2

[13] S. Fortune. Handbook of discrete and computational geometry. chapter Voronoi Diagrams and Delaunay Triangulations, pages 377–388. 1997. 3

[14] Y. Furukawa and J. Ponce. Accurate, dense, and robust multiview stereopsis. *PAMI*, 32:1362–1376, 2010. 1, 2, 7

[15] B. Goldluecke and M. Magnor. Space-time isosurface evolution for temporally coherent 3d reconstruction. In *CVPR*, pages 350–355, 2004. 1, 2

[16] M. Grundmann, V. Kwatra, M. Han, and I. Essa. Efficient hierarchical graph based video segmentation. *CVPR*, 2010. 2

[17] J. Y. Guillemaut and A. Hilton. Joint Multi-Layer Segmentation and Reconstruction for Free-Viewpoint Video Applications. *IJCV*, 93:73–100, 2010. 1, 2, 4, 5

[18] J.-Y. Guillemaut and A. Hilton. Space-time joint multi-layer segmentation and depth estimation. In *3DIMPVT*, pages 440–447, 2012. 2, 6, 7

[19] V. Gulshan, C. Rother, A. Criminisi, A. Blake, and A. Zisserman. Geodesic star convexity for interactive image segmentation. In *CVPR*, pages 3129–3136, 2010. 2, 5

[20] C. Hane, C. Zach, A. Cohen, R. Angst, and M. Pollefeys. Joint 3d scene reconstruction and class segmentation. In *CVPR*, pages 97–104, 2013. 2

[21] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. 2 edition, 2003. 2

[22] X. Hu and P. Mordohai. A quantitative evaluation of confidence measures for stereo vision. *PAMI*, pages 2121–2133, 2012. 5

[23] E. Imre, J. Y. Guillemaut, and A. Hilton. Calibration of nodal and free-moving cameras in dynamic scenes for post-production. In *3DIMPVT*, pages 260–267, 2011. 2

[24] D. Ji, E. Dunn, and J. M. Frahm. 3d reconstruction of dynamic textures in crowd sourced data. In *ECCV*, volume 8689, pages 143–158. 2014. 2

[25] H. Jiang, H. Liu, P. Tan, G. Zhang, and H. Bao. 3d reconstruction of dynamic scenes with multiple handheld cameras. In *ECCV*, pages 601–615. 2012. 2

[26] M. Kazhdan, M. Bolitho, and H. Hoppe. Poisson surface reconstruction. In *Eurographics Symposium on Geometry Processing*, pages 61–70, 2006. 3

[27] A. Kowdle, S. Sinha, and R. Szeliski. Multiple view object cosegmentation using appearance and stereo cues. In *ECCV*, pages 789–803. 2012. 2, 6

[28] E. Larsen, P. Mordohai, M. Pollefeys, and H. Fuchs. Temporally consistent reconstruction from multiple video streams using enhanced belief propagation. In *ICCV*, pages 1–8, 2007. 2

[29] W. Lee, W. Woo, and E. Boyer. Silhouette segmentation in multiple views. *PAMI*, pages 1429–1441, 2011. 2

[30] C. Lei, X. D. Chen, and Y. H. Yang. A new multiview spacetime-consistent depth recovery framework for free viewpoint video rendering. In *ICCV*, pages 1570–1577, 2009. 2

[31] M. Menze and A. Geiger. Object scene flow for autonomous vehicles. In *CVPR*, 2015. 2

[32] A. Mustafa, H. Kim, J. Guillemaut, and A. Hilton. General dynamic scene reconstruction from wide-baseline views. In *ICCV*, 2015. 1, 2, 5, 6, 7

[33] A. Mustafa, H. Kim, E. Imre, and A. Hilton. Segmentation based features for wide-baseline multi-view reconstruction. In *3DV*, 2015. 2, 3

[34] M. Narayana, A. Hanson, and E. Learned-Miller. Coherent motion segmentation in moving camera videos using optical flow orientations. In *ICCV*, pages 1577–1584, 2013. 2

[35] M. Oswald, J. Sthmer, and D. Cremers. Generalized connectivity constraints for spatio-temporal 3d reconstruction. In *ECCV 2014*, pages 32–46. 2014. 2

[36] K. Ozden, K. Schindler, and L. Van Gool. Simultaneous segmentation and 3d reconstruction of monocular image sequences. In *ICCV*, pages 1–8, 2007. 2

[37] A. Papazoglou and V. Ferrari. Fast object segmentation in unconstrained video. In *ICCV*, pages 1777–1784, 2013. 2

[38] R. B. Rusu. *Semantic 3D Object Maps for Everyday Manipulation in Human Living Environments*. PhD thesis, Computer Science department, Technische Universitaet Muenchen, Germany, 2009. 3

[39] S. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *CVPR*, pages 519–528, 2006. 2, 7

[40] Y. M. Shin, M. Cho, and K. M. Lee. Multi-object reconstruction from dynamic scenes: An object-centered approach. *CVIU*, 117:1575 – 1588, 2013. 2

[41] A. Taneja, L. Ballan, and M. Pollefeys. Modeling dynamic scenes recorded with freely moving cameras. In *ACCV*, pages 613–626. 2011. 2

[42] A. Wedel, T. Brox, T. Vaudrey, C. Rabe, U. Franke, and D. Cremers. Stereoscopic scene flow computation for 3d motion understanding. *IJCV*, 95:29–51, 2011. 2

[43] C. Zach, A. Cohen, and M. Pollefeys. Joint 3d scene reconstruction and class segmentation. In *CVPR*, 2013. 2

[44] G. Zeng and L. Quan. Silhouette extraction from multiple images of an unknown background. In *ACCV*, 2004. 2

[45] D. Zhang, O. Javed, and M. Shah. Video object segmentation through spatially accurate and temporally dense extraction of primary object regions. In *CVPR*, 2013. 2, 4

[46] G. Zhang, J. Jia, W. Hua, and H. Bao. Robust bilayer segmentation and motion/depth estimation with a handheld camera. *PAMI*, 2011. 2