International Journal of Computer Vision manuscript No.
(will be inserted by the editor)

Temporally coherent general dynamic scene reconstruction

Armin Mustafa, Marco Volino, Hansung Kim, Jean-Yves Guillemaut, and Adrian Hilton

Received: date / Accepted: date

Abstract Existing techniques for dynamic scene reconstruction from multiple wide-baseline cameras primarily focus on reconstruction in controlled environments, with fixed calibrated cameras and strong prior constraints. This paper introduces a general approach to obtain a 4D representation of complex dynamic scenes from multi-view wide-baseline static or moving cameras without prior knowledge of the scene structure, appearance, or illumination. Contributions of the work are: An automatic method for initial coarse reconstruction to initialize joint estimation; Sparse-to-dense temporal correspondence integrated with joint multi-view segmentation and reconstruction to introduce temporal coherence; and a general robust approach for joint segmentation refinement and dense reconstruction of dynamic scenes by introducing shape constraint. Comparison with state-of-the-art approaches on a variety of complex indoor and outdoor scenes, demonstrates improved accuracy in both multi-view segmentation and dense reconstruction. This paper demonstrates unsupervised reconstruction of complete temporally coherent 4D scene models with improved non-rigid object segmentation and shape reconstruction and its application to free-viewpoint rendering and virtual reality.

Keywords Reconstruction, Temporal coherence, Dynamic, Segmentation

1 Introduction

Reconstruction of general dynamic scenes is of great importance in entertainment applications such as visual effects in film and broadcast production and for content production in virtual reality. The ultimate goal of modelling dynamic scenes from multiple cameras is automatic understanding of real-world scenes from distributed camera networks, for applications in robotics and other autonomous systems. Existing methods have applied multiple view dynamic scene reconstruction techniques in controlled environment with known background or chroma-key studio [23,20,56,60]. Other multiple view stereo techniques require a relatively dense static camera network resulting in a large number of cameras [19]. Extensions to more general outdoor scenes [5,32,60] use prior reconstruction of the static geometry from images of the empty environment. However these methods either require accurate segmentation of dynamic foreground objects, or prior knowledge of the scene struc-

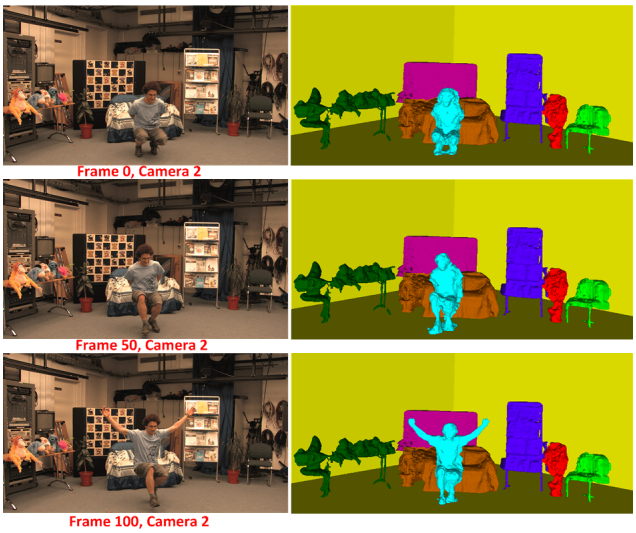

Fig. 1 Temporally consistent scene reconstruction for Odzemok dataset color-coded to show the scene object segmentation obtained.

All authors
Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, GU27XH, Guildford
E-mail: a.mustafa, m.volino, h.kim, j.guillemaut and a.hilton @surrey.ac.uk

ture and background, or are limited to static cameras and controlled environments. Scenes are reconstructed semi-automatically, requiring manual intervention for segmentation/rotoscoping, and result in temporally incoherent per-frame mesh geometries. Temporally coherent geometry with known surface correspondence across the sequence is essential for real-world applications and compact representation.

Our paper addresses the limitations of existing approaches by introducing a methodology for unsupervised temporally coherent dynamic scene reconstruction from multiple wide-baseline static or moving camera views without prior knowledge of the scene structure or background appearance. This temporally coherent dynamic scene reconstruction is demonstrated to work in applications for immersive content production such as free-viewpoint video (FVV) and virtual reality (VR). This work combines two previously published papers in general dynamic reconstruction [42] and temporally coherent reconstruction [43] into a single framework and demonstrates application of this novel unsupervised joint segmentation and reconstruction in immersive content production FVV and VR (Section 5).

The input is a sparse set of synchronised videos from multiple moving cameras of an unknown dynamic scene without prior scene segmentation or camera calibration. Our first contribution is automatic initilisation of camera calibration and sparse scene reconstruction from sparse feature correspondence using sparse feature detection and matching between pairs of frames. An initial coarse reconstruction and segmentation of all scene objects is obtained from sparse features matched across multiple views. This eliminates the requirement for prior knowledge of the background scene appearance or structure. Our second contribution is sparse-to-dense reconstruction and segmentation approach to introduce temporal coherence for every frame. We exploit temporal coherence of the scene to overcome visual ambiguities inherent in single frame reconstruction and multiple view segmentation methods for general scenes. Temporal coherence refers to the correspondence between the 3D surface of all objects observed over time. Our third contribution is spatio-temporal alignment to estimate dense surface correspondence for 4D reconstruction. A geodesic star convexity shape constraint is introduced for the shape segmentation to improve the quality of segmentation for non-rigid objects with complex appearance. The proposed approach overcomes the limitations of existing methods allowing an unsupervised temporally coherent 4D reconstruction of complete models for general dynamic scenes.

The scene is automatically decomposed into a set of spatio-temporally coherent objects as shown in Figure 1 where the resulting 4D scene reconstruction has temporally coherent labels and surface correspondence for each object. This can be used for free-viewpoint video rendering and imported to a game engine for VR experience production. The contributions explained above can be summarized as follows:

- Unsupervised temporally coherent dense reconstruction and segmentation of general complex dynamic scenes from multiple wide-baseline views.
- Automatic initialization of dynamic object segmentation and reconstruction from sparse features.
- A framework for space-time sparse-to-dense segmentation, reconstruction and temporal correspondence.
- Robust spatio-temporal refinement of dense reconstruction and segmentation integrating error tolerant photo-consistency and edge information using geodesic star convexity.
- Robust and computationally efficient reconstruction of dynamic scenes by exploiting temporal coherence.
- Real-world applications of 4D reconstruction to free-viewpoint video rendering and virtual reality.

This paper is structured as follows: First related work is reviewed. The methodology for general dynamic scene reconstruction is then introduced. Finally a thorough qualitative and quantitative evaluation and comparison to the state-of-the-art on challenging datasets is presented.

2 Related Work

Temporally coherent reconstruction is a challenging task for general dynamic scenes due to a number of factors such as motion blur, articulated, non-rigid and large motion of multiple people, resolution differences between camera views, occlusions, wide-baselines, errors in calibration and cluttered dynamic backgrounds. Segmentation of dynamic objects from such scenes is difficult because of foreground and background complexity and the likelihood of overlapping background and foreground color distributions. Reconstruction is also challenging due to limited visual cues and relatively large errors affecting both calibration and extraction of a globally consistent solution. This section reviews previous work on dynamic scene reconstruction and segmentation.

2.1 Dynamic Scene Reconstruction

Dense dynamic shape reconstruction is a fundamental problem and heavily studied area in the field of computer vision. Recovering accurate 3D models of a dynamically evolving, non-rigid scene observed by multiple synchronised cameras is a challenging task. Research on multiple view dense dynamic reconstruction has primarily focused on indoor scenes with controlled illumi-

nation and static backgrounds, extending methods for multiple view reconstruction of static scenes [53] to sequences [62]. Deep learning based approaches have been introduced to estimate shape of dynamic objects from minimal camera views in constrained environment [29, 68] and for rigid objects [58]. In the last decade, focus has shifted to more challenging outdoor scenes captured with both static and moving cameras. Reconstruction of non-rigid dynamic objects in uncontrolled natural environments is challenging due to the scene complexity, illumination changes, shadows, occlusion and dynamic backgrounds with clutter such as trees or people. Methods have been proposed for multi-view reconstruction [65,39,37] requiring a large number of closely spaced cameras for surface estimation of dynamic shape. Practical applications require relatively sparse moving cameras to acquire coverage over large areas such as outdoor. A number of approaches for mutli-view reconstruction of outdoor scenes require initial silhouette segmentation [67,32,22,23] to allow visual-hull reconstruction. Most of these approaches to general dynamic scene reconstruction fail in the case of complex (cluttered) scenes captured with moving cameras.

A recent work proposed reconstruction of dynamic fluids [50] for static cameras. Another work used RGB-D cameras to obtain reconstruction of non-rigid surfaces [55]. Pioneering research in general dynamic scene reconstruction from multiple handheld wide-baseline cameras [5,60] exploited prior reconstruction of the background scene to allow dynamic foreground segmentation and reconstruction. Recent work [46] estimates shape of dynamic objects from handheld cameras exploiting GANs. However these approaches either work for static/indoor scenes or exploit strong prior assumptions such as silhouette information, known background or scene structure. Also all these approaches give per frame reconstruction leading to temporally incoherent geometries. Our aim is to perform temporally coherent dense reconstruction of unknown dynamic non-rigid scenes automatically without strong priors or limitations on scene structure.

2.2 Joint Segmentation and Reconstruction

Many of the existing multi-view reconstruction approaches rely on a two-stage sequential pipeline where foreground or background segmentation is initially performed independently with respect to each camera, and then used as input to obtain visual hull for multi-view reconstruction. The problem with this approach is that the errors introduced at the segmentation stage cannot be recovered and are propagated to the reconstruction stage reducing the final reconstruction quality. Segmentation from multiple wide-baseline views has been proposed by exploiting appearance similarity [17,38,70]. These approaches assume static backgrounds and different colour distributions for the foreground and background [52,17] which limits applicability for general scenes.

Joint segmentation and reconstruction methods incorporate estimation of segmentation or matting with reconstruction to provide a combined solution. Joint refinement avoids the propagation of errors between the two stages thereby making the solution more robust. Also, cues from segmentation and reconstruction can be combined efficiently to achieve more accurate results. The first multi-view joint estimation system was proposed by Szeliski et al.[59] which used iterative gradient descent to perform an energy minimization. A number of approaches were introduced for joint formulation in static scenes and one recent work used training data to classify the segments [69]. The focus shifted to joint segmentation and reconstruction for rigid objects in indoor and outdoor environments. These approaches used a variety of techniques such as patch-based refinement [54, 48] and fixating cameras on the object of interest [11] for reconstructing rigid objects in the scene. However, these are either limited to static scenes [69,26] or process each frame independently thereby failing to enforce temporal consistency [11,23].

Joint reconstruction and segmentation on monocular video was proposed in [36,3,12] achieving semantic segmentation of scene limited to rigid objects in street scenes. Practical application of joint estimation requires these approaches to work on non-rigid objects such as humans with clothing. A multi-layer joint segmentation and reconstruction approach was proposed for multiple view video of sports and indoor scenes [23]. The algorithm used known background images of the scene without the dynamic foreground objects to obtain an initial segmentation. Visual-hull based reconstruction was performed with known prior foreground/background using a background image plate with fixed and calibrated cameras. This visual hull was used as a prior and was optimized by a combination of photo-consistency, silhouette, color and sparse feature information in an energy minimization framework to improve the segmentation and reconstruction quality. Although structurally similar to our approach, it requires the scene to be captured by fixed calibrated cameras and a priori known fixed background plate as a prior to estimate the initial visual hull by background subtraction. The proposed approach overcomes these limitations allowing moving cameras and unknown scene backgrounds.

An approach based on optical flow and graph cuts was shown to work well for non-rigid objects in indoor settings but requires known background segmentation to obtain silhouettes and is computationally expensive [24]. Practical application of temporally coherent joint

estimation requires approaches that work on non-rigid objects for general scenes in uncontrolled environments. A quantitative evaluation of techniques for multi-view reconstruction was presented in [53]. These methods are able to produce high quality results, but rely on good initializations and strong prior assumptions with known and controlled (static) scene backgrounds.

The proposed method exploits the advantages of joint segmentation and reconstruction and addresses the limitations of existing methods by introducing a novel approach to reconstruct general dynamic scenes automatically from wide-baseline cameras with no prior. To overcome the limitations of existing methods, the proposed approach automatically initialises the foreground object segmentation from wide-baseline correspondence without prior knowledge of the scene. This is followed by a joint spatio-temporal reconstruction and segmentation of general scenes. Temporal correspondence is exploited to overcome visual ambiguities giving improved reconstruction together with temporal coherence of surface correspondence to obtain 4D scene models.

2.3 Temporal coherent 4D Reconstruction

Temporally coherent 4D reconstruction refers to aligning the 3D surfaces of non-rigid objects over time for a dynamic sequence. This is achieved by estimating point-to-point correspondences for the 3D surfaces to obtain 4D temporally coherent reconstruction. 4D models allows to create efficient representation for practical applications in film, broadcast and immersive content production such as virtual, augmented and mixed reality. The majority of existing approaches for reconstruction of dynamic scenes from multi-view videos 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. Recent research has shown that exploiting temporal information can improve reconstruction accuracy as well as achieving temporal coherence [43].

3D scene flow estimates frame to frame correspondence whereas 4D temporal coherence estimates correspondence across the complete sequence to obtain a single surface model. Methods to estimate 3D scene flow have been reported in the literature [41] for autonomous vehicles. However this approach is limited to narrow baseline cameras. Other scene flow approaches are dependent on 2D optical flow [66,6] and they require an accurate estimate for most of the pixels which fails in the case of large motion. However, 3D scene flow methods align two frames independently and do not produce temporally coherent 4D models.

Research investigating spatio-temporal reconstruction across multiple frames was proposed by [20,37,24] exploiting the temporal information from the previous frames using optical flow. An approach for recovering space-time consistent depth maps from multiple video sequences captured by stationary, synchronized and calibrated cameras for depth based free viewpoint video rendering was proposed by [39]. However these methods require accurate initialisation, fixed and calibrated cameras and are limited to simple scenes. Other approaches to temporally coherent reconstruction [4] either requires a large number of closely spaced cameras or bi-layer segmentation [72,30] as a constraint for reconstruction. Recent approaches for spatio-temporal reconstruction of multi-view data either work on indoor studio data [47].

The framework proposed in this paper addresses limitations of existing approaches and gives 4D temporally coherent reconstruction for general dynamic indoor or outdoor scenes with large non-rigid motions, repetitive texture, uncontrolled illumination, and large capture volume. The scenes are captured with sparse static/moving cameras. The proposed approach gives 4D models of complete scenes with both static and dynamic objects for real-world applications (FVV and VR) with no prior knowledge of scene structure.

2.4 Multi-view Video Segmentation

In the field of image segmentation, approaches have been proposed to provide temporally consistent monocular video segmentation [21,49,45,71]. Hierarchical segmentation based on graphs was proposed in [21], directed acyclic graph were used to propose an object followed by segmentation [71]. Optical flow is used to identify and consistently segment objects [45,49]. Recently a number of approaches have been proposed for multi-view foreground object segmentation by exploiting appearance similarity spatially across views [16,35,38,70]. An approach for space-time multi-view segmentation was proposed by [17]. However, multi-view approaches assume a static background and different colour distributions for the foreground and background which limits applicability for general 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 [25] (convex and concave shapes), but require manual interaction. The proposed approach performs automatic multi-view video segmentation by initializing the foreground object model using spatio-temporal information from wide-baseline feature correspondence followed by a multi-

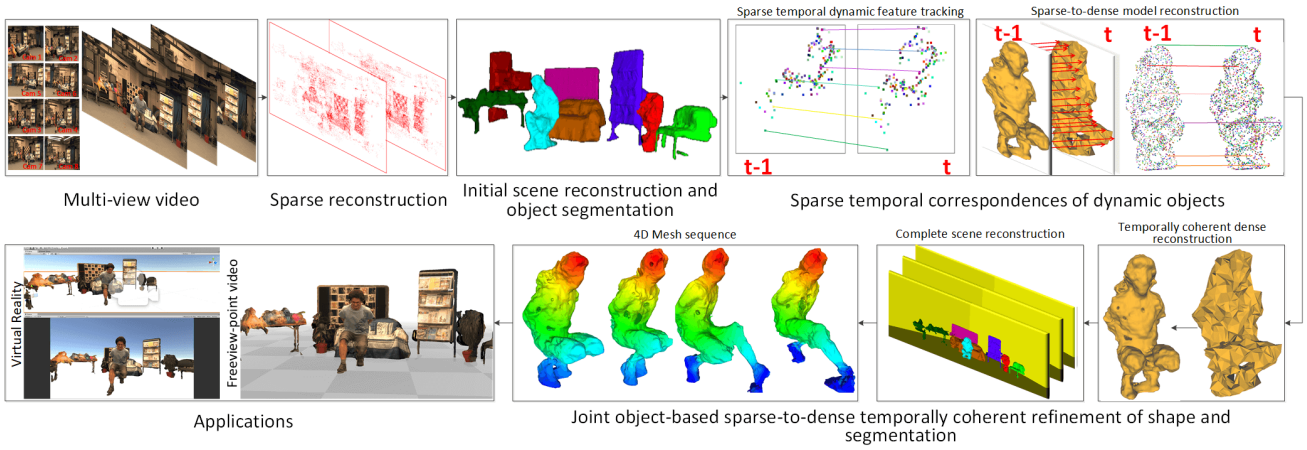

Fig. 2 Overview of temporally consistent scene reconstruction framework.

layer optimization framework. Geodesic star convexity previously used in single view segmentation [25] is applied to constraint the segmentation in each view. Our multi-view formulation naturally enforces coherent segmentation between views and also resolves ambiguities such as the similarity of background and foreground in isolated views.

2.5 Summary and Motivation

Image-based temporally coherent 4D dynamic scene reconstruction without a prior model or constraints on the scene structure is a key problem in computer vision. Existing dense reconstruction algorithms need some strong initial prior and constraints for the solution to converge such as background, structure, and segmentation, which limits their application for automatic reconstruction of general scenes. Current approaches are also commonly limited to independent per-frame reconstruction and do not exploit temporal information or produce a coherent model with known correspondence.

The approach proposed in this paper aims to overcome the limitations of existing approaches to enable robust temporally coherent wide-baseline multiple view reconstruction of general dynamic scenes without prior assumptions on scene appearance, structure or segmentation of the moving objects. Static and dynamic objects in the scene are identified for simultaneous segmentation and reconstruction using geometry and appearance cues in a sparse-to-dense optimization framework. Temporal coherence is introduced to improve the quality of the reconstruction and geodesic star convexity is used to improve the quality of segmentation. The static and dynamic elements are fused automatically in both the temporal and spatial domain to obtain the final 4D scene reconstruction.

This paper presents a unified framework, novel in combining multiple view joint reconstruction and segmentation with temporal coherence to improve per-frame reconstruction performance and produce a single framework from the initial work presented in [43,42]. In particular the approach gives 4D surface model with full correspondence over time. A comprehensive experimental evaluation with comparison to the state-of-the-art in segmentation, reconstruction and 4D modelling is also presented extending previous work. Application fo the resulting 4D models to free-viewpoint video rendering and content production for immersive virtual reality experiences is also presented.

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 [19,23], or require restrictive assumptions on scene complexity and structure and often assume prior camera calibration and foreground segmentation [60,24]. We address these issues by initializing the joint reconstruction and segmentation algorithm automatically, introducing temporal coherence in the reconstruction and geodesic star convexity in segmentation to reduce ambiguity and ensure consistent non-rigid structure initialization at successive frames. The proposed approach is demonstrated to achieve improved reconstruction and segmentation performance over state-of-the-art approaches and produce temporally coherent 4D models of complex dynamic scenes.

3.1 Overview

An overview of the proposed framework for temporally coherent multi-view reconstruction is presented in Figures 2 and consists of the following stages:
Multi-view video: The scenes are captured using multiple video cameras (static/moving) separated by wide-baseline ($> 15°$). The cameras can be synchronized

during the capture using time-code generator or later using the audio information. Camera extrinsic calibration and scene structure are assumed to be unknown.

Sparse reconstruction: The intrinsics are assumed to be known. Segmentation based feature detection (SFD) [44] is used to 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 camera extrinsics and a sparse 3D point-cloud for each time instant for the entire sequence [27].

Initial scene segmentation and reconstruction - Section 3.2: Automatic initialisation is performed without prior knowledge of the scene structure or appearance to obtain an initial approximation for each object. The sparse point cloud is clustered in 3D [51] with each cluster representing a unique foreground object. Object segmentation increases efficiency and improve robustness of 4D models. This reconstruction is refined using the framework explained in Section 3.4 to obtain segmentation and dense reconstruction of each object.

Sparse-to-dense temporal reconstruction with temporal coherence - Section 3.3 Temporal coherence is introduced in the framework to initialize the coarse reconstruction and obtain frame-to-frame dense correspondences for dynamic object. 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.

Joint object-based sparse-to-dense temporally coherent refinement of shape and segmentation - Section 3.4: The initial estimate is refined for each object per-view in the scene through joint optimisation of shape and segmentation using a robust cost function combining matching, color, contrast and smoothness information for wide-baseline matching with a geodesic star convexity constraint. A single 3D model for each dynamic object is obtained by fusion of the view-dependent depth maps using Poisson surface reconstruction [31]. Surface orientation is estimated based on neighbouring pixels.

Applications - Section : The 4D representation from the proposed joint segmentation and reconstruction framework has a number of applications in media production, including free-viewpoint video (FVV) rendering and virtual reality (VR).

The process above is repeated for the entire sequence for all objects in the first frame and for dynamic objects at each time-instant. The proposed approach enables automatic reconstruction of all objects in the scene as a 4D mesh sequence. Subsequent sections present the novel contributions of this work in initialisation and refinement to obtain a dense temporally coherent reconstruction. The approach is demonstrated to outperform previous approaches to dynamic scene reconstruction and does not require prior knowledge of the scene.

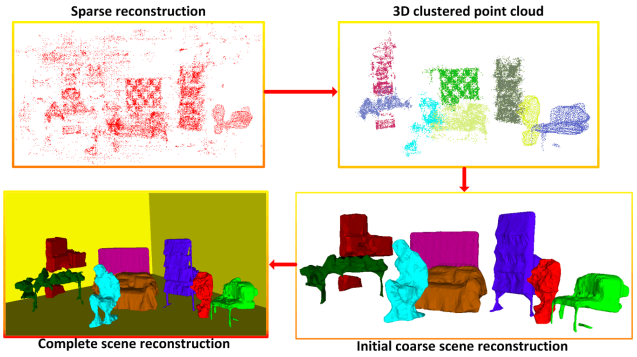

Fig. 3 Overview of stages for estimation of an initial dense scene reconstruction. For more details refer to Section 3.2.

3.2 Initial Scene Segmentation and Reconstruction

For general dynamic scene reconstruction, we need to reconstruct and segment the objects in the scene. This requires an initial coarse approximation for initialisation of a subsequent refinement step to optimise the segmentation and reconstruction with respect to each camera view. We introduce an approach based on sparse point cloud clustering, an overview is shown in Figure 3. Initialisation gives a complete coarse segmentation and reconstruction of each object in the first frame of the sequence for subsequent refinement. The dense reconstruction of the foreground objects and background are combined to obtain a full scene reconstruction at the first time instant. A rough geometric proxy of the background is created using the method. For consecutive time instants dynamic objects and newly appeared objects are identified and only these objects are reconstructed and segmented. The reconstruction of static objects is retained which reduces computational complexity. The optic flow and cluster information for each dynamic object ensures that we retain same labels for the entire sequence.

3.2.1 Background Reconstruction

Accurate reconstruction of the background is often challenging due to uniform appearance of large regions. A rough geometric proxy of the background is created by computing the minimum oriented bounding box for the sparse 3D point-cloud using principal component analysis (PCA) [15]. Different methods are used for background estimation for indoor and outdoor scenes. For

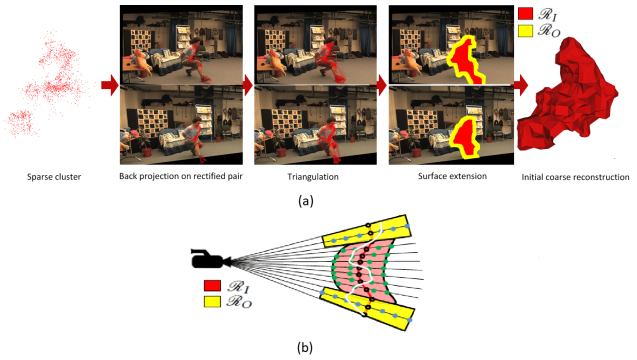

Fig. 4 Initial coarse reconstruction: (a) Sparse-to-dense initial coarse reconstruction of the dynamic object in Odzemok dataset; and (b) White line represents the actual surface, Depth labels are represented as circles; blue circles depict depth labels in $\mathscr{D}_O$, green circles depict depth labels in $\mathscr{D}_I$ and black circles depict the initial surface estimate.

outdoor scenes a plane is inserted at infinity perpendicular to the ground plane as there are no consistent constraints like room, walls, corridors etc. in such datasets. For indoor scenes the Manhattan world assumption [13] is applied and the process used for estimation of the background is described below:

- The centroid $\mathbf{A} = (a_0, a_1, a_2)$ and normalized covariance of the point-cloud are estimated to compute the eigenvectors $\vec{e} = (e_0, e_1, e_2)$ for the covariance matrix of the point-cloud (PCA). We define the reference system as $\mathbf{R}$= $(e_0, e_1, e_0 \times e_1)$ such that: $e_0 \times e_1 = +/-e_2$. The sparse points are mapped in reference frame using $\mathbf{R}$ as the rotation matrix and $\mathbf{A}$ as the translation.
- The rotation and translation are calculated using the eigenvectors to place a box in correct location. The minimum and maximum values of coordinates in the x, y and z directions for the transformed cloud are computed to determine the minimum oriented box width, height, and depth.
- Given a box centred at the origin with size defined above the rotation $\mathbf{R}$ and translation $\mathbf{R}\times\mathbf{C}+\mathbf{A}$ is applied, where $\mathbf{C}$ is the middle of the minimum and maximum points.

This background reconstruction is a rough geometric proxy estimate of the background of the scene but gives reasonable results to give complete scene reconstruction.

3.2.2 Sparse Point-cloud Clustering

The sparse representation of the scene is processed to remove outliers using the point neighbourhood statistics to filter outlier data [51]. We segment the objects in the sparse scene reconstruction, this allows only moving objects to be reconstructed at each frame for efficiency and this also allows object shape similarity to be propagated across frames to increase robustness of reconstruction.

We use data clustering approach based on the 3D grid subdivision of the space using an octree data structure in Euclidean space to segment objects at each frame. In a more general sense, nearest neighbor information is used to cluster, which is essentially similar to a flood fill algorithm. We choose this data clustering because of its computational efficiency and robustness. The approach allows segmentation of objects in the scene and is demonstrated to work well for cluttered and general outdoor scenes as shown in Section 4.

Objects with insufficient detected features are reconstructed as part of the scene background. Appearing, disappearing and reappearing objects are handled by sparse dynamic feature tracking, explained in Section 3.3. Clustering results are shown in Figure 3. This is followed by a sparse-to-dense coarse object based approach to segment and reconstruct general dynamic scenes.

3.2.3 Coarse Object Reconstruction

The process to obtain the coarse reconstruction for the first frame of the sequence is shown in Figure 4. The sparse representation of each element is back-projected on the rectified image pair for each view. Delaunay triangulation [18] is performed on the set of back projected points for each cluster on one image and is propagated to the second image using the sparse matched features. Triangles with edge length greater than the median length of edges of all triangles are removed. For each remaining triangle pair direct linear transform is used to estimate the affine homography. Displacement at each pixel within the triangle pair is estimated by interpolation to get an initial dense disparity map for each cluster in the 2D image pair labelled as $\mathscr{R}_I$ depicted in red in Figure 4. The initial coarse reconstruction for the observed objects in the scene is used to define the depth hypotheses at each pixel for the optimization.

The region $\mathscr{R}_I$ does not ensure complete coverage of the object, so we extrapolate this region to obtain a region $\mathscr{R}_O$ (shown in yellow) in 2D by 5% of the average distance between the boundary points($\mathscr{R}_I$) and the centroid of the object. To allow for errors in the initial approximate depth from sparse features we add volume in front and behind of the projected surface by an error tolerance, along the optical ray of the camera. This ensures that the object boundaries lie within the extrapolated initial coarse estimate and depth at each pixel for the combined regions may not be accurate. The tolerance for extrapolation may vary if a pixel belongs to $\mathscr{R}_I$ or $\mathscr{R}_O$ as the propagated pixels of the extrapolated regions ($\mathscr{R}_O$) may have a high level of errors compared to error at the points from sparse representation ($\mathscr{R}_I$) requiring a comparatively higher tolerance. The calculation of threshold depends on the capture

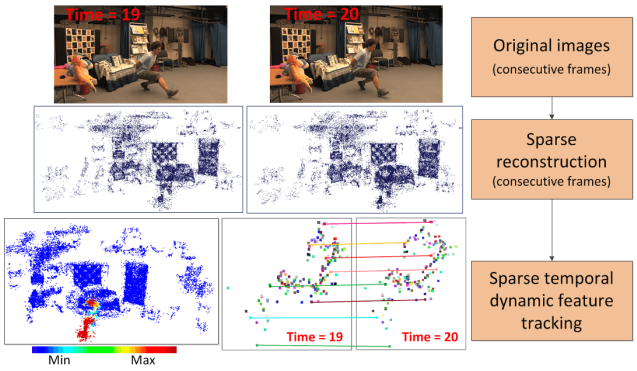

Fig. 5 Sparse temporal dynamic feature tracking algorithm: Results on Odzemok dataset; Min and Max is the minimum and maximum movement in the 3D points respectively.

volume of the datasets and is set to 1% of the capture volume for $\mathscr{R}_O$ and half the value for $\mathscr{R}_I$. This volume in 3D corresponds to our initial coarse reconstruction of each object and enables us to remove the dependency of the existing approaches on background plate and visual hull estimates. This process of cluster identification and initial coarse object reconstruction is performed for multiple objects in general environments. Initial object segmentation using point cloud clustering and coarse segmentation is insensitive to parameters. Throughout this work the same parameters are used for all datasets. The result of this process is a coarse initial object segmentation and reconstruction for each object.

3.3 Sparse-to-dense temporal reconstruction with temporal coherence

Once the static scene reconstruction is obtained for the first frame, we perform temporally coherent reconstruction for dynamic objects at successive time instants instead of whole scene reconstruction for computational efficiency and to avoid redundancy. 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 used to propagate an initial model of the moving object for refinement. Figure 5 presents the sparse reconstruction and temporal correspondence. New objects are identified per frame from the clustered sparse reconstruction and are labelled as dynamic objects.

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

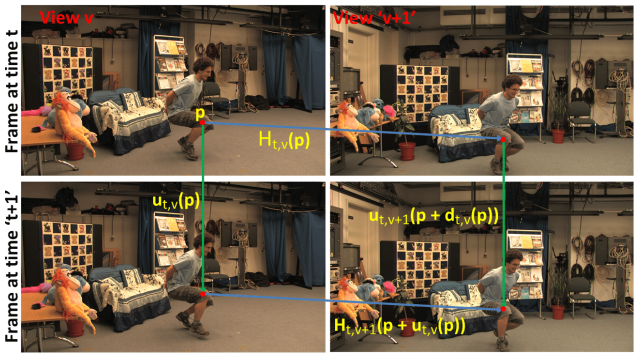

Fig. 6 Spatio-temporal consistency check for 3D tracking for Odzemok dataset.

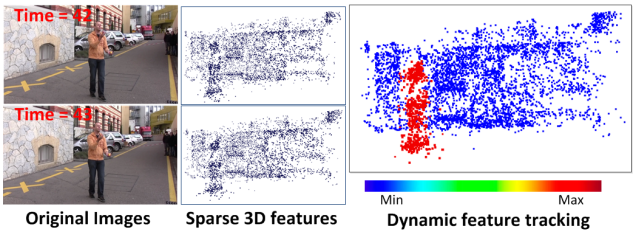

Fig. 7 Sparse temporal dynamic feature tracking for Juggler dataset captured with only moving cameras. Min and Max is the minimum and maximum movement in the 3D points respectively.

using SFD [44] 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 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 back-projected to each view. These features are matched temporally using 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 6. Each match must satisfy the constraint:

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

where p is the feature image point in view v at frame t, $H_{t,v}(p)$ is the disparity at frame t from views v and r, $u_{t,v}(p)$ 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 [71] 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

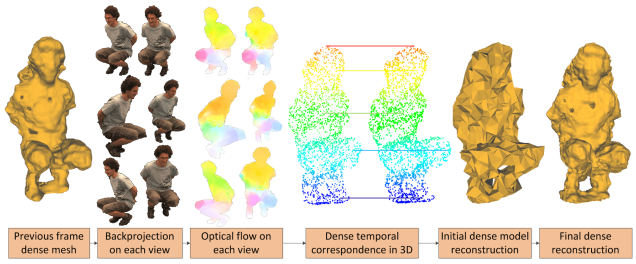

Fig. 8 Initial sparse-to-dense model reconstruction workflow

Fig. 9 Improvement in segmentation for the Odzemok dataset and reconstruction for the Juggler dataset with temporal coherence (highlighted in yellow)

top and bottom 5 percentile values are removed. Median filtering is then applied to identify the dynamic features. Figure 7 shows an example with moving cameras for Juggler [5].

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 [23]. Temporal coherence also provides a more accurate initialisation to overcome visual ambiguities at individual frames. Figure 8 illustrates the use of temporal coherence for reconstruction initialisation 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. Lucas Kanade Optical flow [8] 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. The sparse-to-dense initial coarse reconstruction improves the quality of segmentation and reconstruction after the refinement. Examples of the improvement in segmentation and reconstruction for Odzemok [1] and Juggler [5] datasets are shown in Figure 9. As observed limbs of the people is retained by using information from the previous frames in both the cases.

3.4 Joint object-based sparse-to-dense temporally coherent refinement of shape and segmentation

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 previously been shown to give improved performance in interactive image segmentation for structures with fine details (for example a person's fingers or hair)[25].

Shape is a powerful cue for object recognition and segmentation. Shape models represented as distance transforms from a template have been used for category specific segmentation [33]. Some works have introduced generic connectivity constraints for segmentation showing that obtaining a globally optimal solutions under the connectivity constraint is NP-hard [64]. Veksler et al. have used shape constraint in segmentation framework by enforcing star convexity prior on the segmentation, and globally optimal solutions are achieved subject to this constraint [63]. The star convexity constraint ensures connectivity to seed points, and is a stronger assumption than plain connectivity. An example of a star-convex object is shown in Figure 10 along with a failure case for a non-rigid articulate object. To handle more complex objects the idea of geodesic forests with multiple star centres was introduced to obtain a globally optimal solution for interactive 2D object segmentation [25]. The main focus was to introduce shape constraints in interactive segmentation, by means of a geodesic star convexity prior. The notion of connectivity was extended from Euclidean to geodesic so that paths can bend and adapt to image data as opposed to straight Euclidean rays, thus extending visibility and reducing the number of star centers required.

The geodesic star-convexity is integrated as a constraint on the energy minimisation for joint multi-view

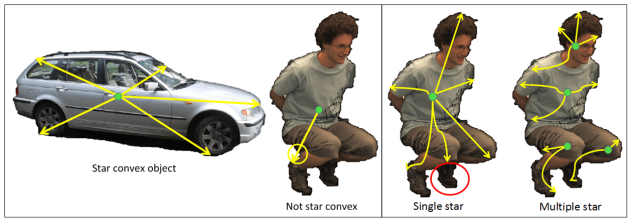

Fig. 10 (a) Representation of star convexity: The left object depicts example of star-convex objects, with a star center marked. The object on the right with a plausible star center shows deviations from star-convexity in the fine details, and (b) Multiple star semantics for joint refinement: Single star center based segmentation is depicted on the left and multiple star is shown on the right.

reconstruction and segmentation [23]. In this work the shape constraint is automatically initialised for each view from the initial segmentation. 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 Figure 12. 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 12.

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 3.4.1).

3.4.1 Optimization on initial coarse object reconstruction based on geodesic star convexity

The depth of the initial coarse reconstruction estimate is refined per view for each dynamic object at a per pixel level. View-dependent optimisation of depth is performed with respect to each camera which is robust to errors in camera calibration and initialisation. Calibration inaccuracies produce inconsistencies limiting the applicability of global reconstruction techniques which simultaneously consider all views; view-dependent techniques are more tolerant to such inaccuracies because they only use a subset of the views for reconstruction of depth from each camera view.

Our goal is to assign an accurate depth value from a set of depth values $\mathscr{D} = \{d_1, ..., d_{|\mathscr{D}|-1}, \mathscr{U}\}$ and assign a layer label from a set of label values $\mathscr{L} = \{l_1, ..., l_{|\mathscr{L}|}\}$ 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 [23] for label (segmentation) and depth (reconstruction):

$$E(l,d) = \lambda_{data}E_{data}(d) + \lambda_{contrast}E_{contrast}(l) + \lambda_{smooth}E_{smooth}(l,d) + \lambda_{color}E_{color}(l) \quad (2)$$

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 3.4.2. The equation consists of four terms: the data term is for the photo-consistency scores, the smoothness term is to avoid sudden peaks in depth and maintain the consistency and the color and contrast terms are to identify the object boundaries. Data and smoothness terms are common to solve reconstruction problems [7] and the color and contrast terms are used for segmentation [34]. 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, if every point $p \in l$ is visible to a star center c via l in the image x which can be expressed as an energy cost:

$$E^{\star}(l|x,c) = \sum_{p \in R} \sum_{q \in \Gamma_{c,p}} E^{\star}_{p,q}(l_p, l_q) \quad (3)$$

$$\forall q \in \Gamma_{c,p},\ E^{\star}_{p,q} = \begin{cases} \infty \text{ if } l_p \neq l_q \\ 0 \text{ otherwise} \end{cases} \quad (4)$$

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

$$\Gamma_{c,p} = \arg\min_{\Gamma \in P_{c,p}} L(\Gamma) \quad (5)$$

where $P_{c,p}$ denotes the set of all discrete paths between c and p and $L(\Gamma)$ is the length of discrete geodesic path as defined in [25]. In the case of image segmentation the gradients in the underlying image provide information to compute the discrete paths between each pixel and star centers and $L(\Gamma)$ is defined below:

$$L(\Gamma) = \sum_{i=1}^{N_D - 1} \sqrt{(1-\delta_g) j(\Gamma^i, \Gamma^{i+1})^2 + \delta_g \left\| \bigtriangledown I(\Gamma^i) \right\|^2} \quad (6)$$

where Γ is an arbitrary parametrized discrete path with N_D pixels given by $\{\Gamma^1, \Gamma^2, \cdots \Gamma^{N_D}\}$, $j(\Gamma^i, \Gamma^{i+1})$ is the Euclidean distance between successive pixels, and the

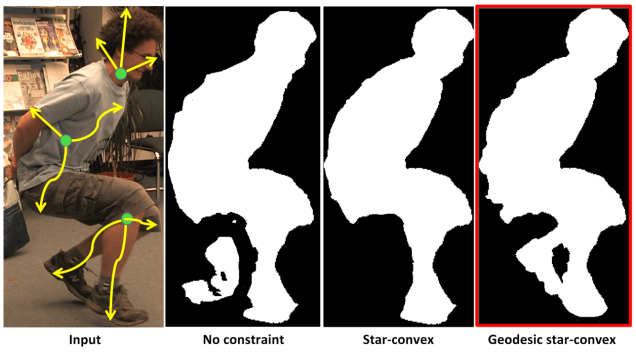

Fig. 11 Segmentation comparison results with no constraint, star convexity constraint and geodesic star convexity constraint for Odzemok dataset.

quantity $\left\|\bigtriangledown I(\Gamma^i)\right\|^2$ is a finite difference approximation of the image gradient between the points $\left(\Gamma^i, \Gamma^{i+1}\right)$. The parameter weights δ_g the Euclidean distance with the geodesic length. Using the above definition, one can define the geodesic distance as defined in Equation 5.

An extension of single star-convexity is to use multiple stars to define a more general class of shapes. Introduction of multiple star centers reduces the path lengths and increases the visibility of small parts of objects like small limbs as shown in Figure 10. Hence Equation 3 is extended to multiple stars. 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 $\mathscr{C} = \{c_1, ..., c_{N_T}\}$ via l in the image x, where N_T is the number of star centers [25]. This is expressed as an energy cost:

$$E^\star(l|x, \mathscr{C}) = \sum_{p \in R} \sum_{q \in \Gamma_{c,p}} E^\star_{p,q}(l_p, l_q) \tag{7}$$

In our case all the correct temporal sparse feature correspondences are used 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. Comparison of segmentation constraint with geodesic multi-star convexity against no constraints and Euclidean multi-star convexity constraint is shown in Figure 11. The figure demonstrates the usefulness of the proposed approach with an improvement in segmentation quality on non-rigid complex objects. The energy in the Equation 2 is minimized as follows:

$$min_{\substack{(l,d) \\ s.t.}} \; E(l,d) \underset{l \in S^\star(\mathscr{C})}{\Leftrightarrow} \min_{(l,d)} E(l,d) + E^\star(l|x, \mathscr{C}) \tag{8}$$

where $S^\star(\mathscr{C})$ is the set of all shapes which lie within the geodesic distances with respect to the centers in $\mathscr{C}$. Optimization of Equation 8, subject to each pixel p in the region $\mathscr{R}$ being at a geodesic distance $\Gamma_{c,p}$ from the star centers in the set $\mathscr{C}$, is performed using the α-expansion algorithm for a pixel p by iterating through

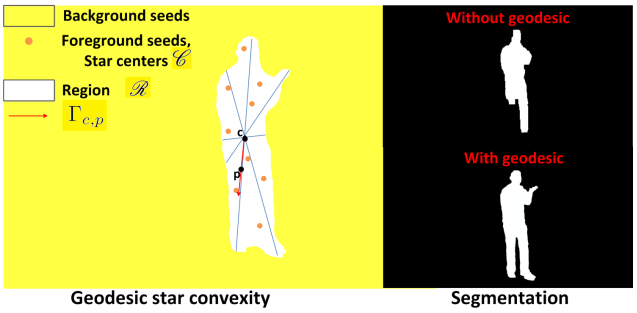

Fig. 12 Geodesic star convexity: A region $\mathscr{R}$ with star centers $\mathscr{C}$ connected with geodesic distance $\Gamma_{c,p}$. Segmentation results with and without geodesic star convexity based optimization are shown on the right for the Juggler dataset.

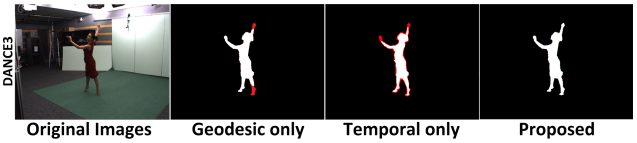

Fig. 13 Comparison of segmentation with introduction of temporal coherence, Geodesic star convexity(GSC) and proposed method (GSC and temporal coherence) for Dance2 dataset.

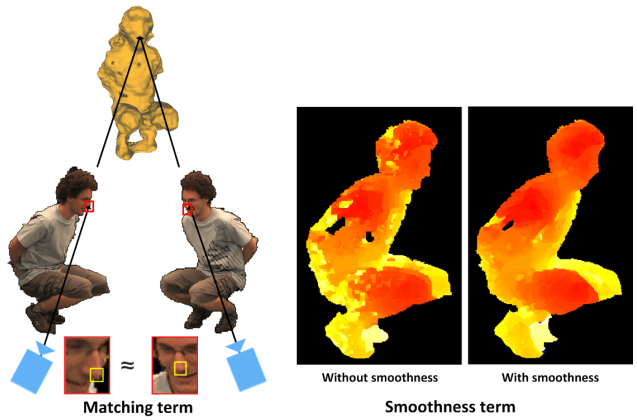

Fig. 14 Illustration of matching and smoothness term for the energy minimization

the set of labels in $\mathscr{L} \times \mathscr{D}$ [10]. Graph-cut is used to obtain a local optimum [9]. The improvements in the results using geodesic star convexity in the framework is shown in Figure 12 and by using temporal coherence is shown in Figure 9. Figure 13 shows improvements using geodesic shape constraint, temporal coherence and combined proposed approach for Dance2 [2] dataset.

3.4.2 Energy cost function for joint segmentation and reconstruction

For completeness in this section we define each of the terms in Equation 2, these are based on previous terms used for joint optimisation over depth for each pixel introduced in [42], 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 (Figure 14) as follows:

$E_{data}(d) = \sum_{p \in \mathscr{P}} e_{data}(p, d_p) =$

$$\begin{cases} M(p,q) = \sum_{i \in \mathscr{O}_k} m(p,q), & \text{if } d_p \neq \mathscr{U} \\ M_{\mathscr{U}}, & \text{if } d_p = \mathscr{U} \end{cases} \tag{9}$$

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.

For textured scenes Normalized Cross Correlation (NCC) over a squared window is a common choice [53]. The NCC values range from -1 to 1 which are then mapped to non-negative values by using the function $1 - NCC$. A maximum likelihood measure [40] is used in this function for confidence value calculation between the center pixel p and the other pixels q and is based on the survey on confidence measures for stereo [28]. The measure is defined as:

$$m(p,q) = \frac{exp\frac{c_{min}}{2\sigma_i^2}}{\sum_{(p,q) \in \mathscr{N}} exp\frac{-(1-NCC(p,q))}{2\sigma_i^2}} \tag{10}$$

where σ_i^2 is the noise variance for each auxiliary camera i; this parameter was fixed to 0.3. $\mathscr{N}$ denotes the set of interacting pixels in $\mathscr{P}$. c_{min} is the minimum cost for a pixel obtained by evaluating the function $(1-NCC(.,.))$ on a 15×15 window.

Contrast term: Segmentation boundaries in images tend to align with contours of high contrast and it is desirable to represent this as a constraint in stereo matching. A consistent interpretation of segmentation-prior and contrast-likelihood is used from [34]. We used a modified version of this interpretation in our formulation to preserve the edges by using Bilateral filtering [61] instead of Gaussian filtering. The contrast term is as follows:

$$E_{contrast}(l) = \sum_{p,q \in \mathscr{N}} e_{contrast}(p, q, l_p, l_q) \tag{11}$$

$$e_{contrast}(p,q,l_p,l_q) = \begin{cases} 0, & \text{if } (l_p = l_q) \\ \frac{1}{1+\epsilon}(\epsilon + exp^{-C(p,q)}), & \text{otherwise} \end{cases} \tag{12}$$

$\|\cdot\|$ is the L_2 norm and $\epsilon = 1$. The simplest choice for $C(p,q)$ would be the squared Euclidean color distance between intensities at pixel p and q as used in [23]. We propose a term for better segmentation as $C(p,q) = \frac{\|B(p)-B(q)\|^2}{2\sigma_{pq}^2 d_{pq}^2}$ where $B(.)$ represents the bilateral filter, d_{pq} is the Euclidean distance between p and q, and $\sigma_{pq} = \left\langle \frac{\|B(p)-B(p)\|^2}{d_{pq}^2} \right\rangle$ This term enables to remove the regions with low photo-consistency scores and weak edges and thereby helps in estimating the object boundaries.

Smoothness term: This term is inspired by [23] and it ensures the depth labels vary smoothly within the object reducing noise and peaks in the reconstructed surface. This is useful when the photo-consistency score is low and insufficient to assign depth to a pixel (Figure 14). It is defined as:

$$E_{smooth}(l,d) = \sum_{(p,q) \in \mathscr{N}} e_{smooth}(l_p, d_p, l_q, d_q) \tag{13}$$

$e_{smooth}(l_p, d_p, l_q, d_q) =$

$$\begin{cases} min(|d_p - d_q|, d_{max}), & \text{if } l_p = l_q \text{ and } d_p, d_q \neq \mathscr{U} \\ 0, & \text{if } l_p = l_q \text{ and } d_p, d_q = \mathscr{U} \\ d_{max}, & \text{otherwise} \end{cases} \tag{14}$$

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

Color term: This term is computed using the negative log likelihood [9] 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 Foreground/Background mixed with uniform color models [14] as the markers are sparse.

$$E_{color}(l) = \sum_{p \in \mathscr{P}} -logP(I_p|l_p) \tag{15}$$

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

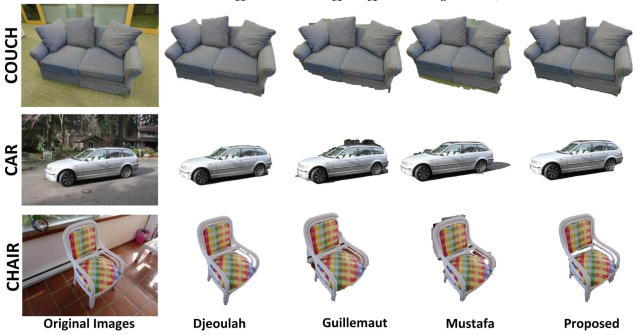

Fig. 15 Comparison of segmentation on benchmark static datasets using geodesic star-convexity.

Datasets	Resolution	No. of views	Baseline	Type
Office[1]	1920 × 1080	8(all S)	25°-35°	dynamic
Juggler[5]	960 × 544	6(all M)	15°-45°	dynamic
Dance1[1]	1920 × 1080	8(all S)	20°-30°	dynamic
Odzemok[1]	1920 × 1080	8(2 M)	15°-30°	dynamic
Dance2[2]	780 × 582	8(all S)	35°-45°	dynamic
Magician[5]	960 × 544	6(all M)	15°-45°	dynamic
Couch[35]	640 × 480	9(all S)	25°-30°	static
Chair[35]	640 × 480	17(all S)	5°-8°	static
Car[35]	640 × 480	16(all S)	5°-8°	static

Table 1 Properties of all datasets where **Type** represents whether the data is static or dynamic. In **No. of views** S stands for static cameras and M for moving cameras.

	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 $\lambda_{contrast}$

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

Table 3 Static segmentation completeness comparison with existing methods on benchmark datasets (%)

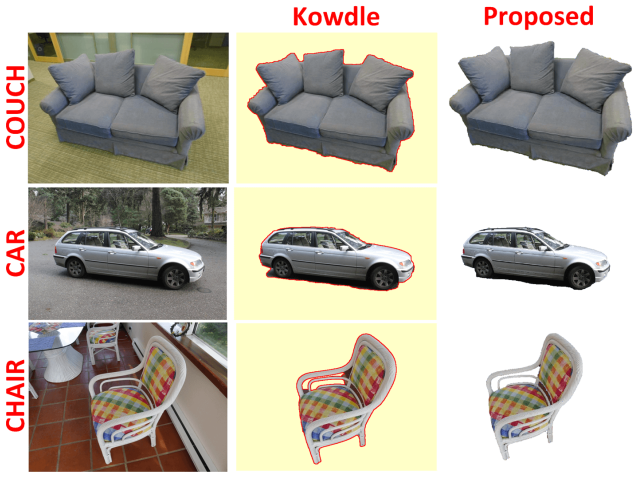

Fig. 16 Comparison of segmentation with Kowdle [35].

4 Results and Performance Evaluation

The proposed system is tested on publicly available multi-view research datasets of indoor and outdoor scenes, details of datasets explained in Table 1. The parameters used for all the datasets are defined in Table 2. More information is available on the website[1].

4.1 Multi-view segmentation evaluation

Segmentation is evaluated against the state-of-the-art methods for multi-view segmentation Kowdle [35] and Djelouah [16] for static scenes and joint segmentation reconstruction methods Mustafa [42] (per frame) and Guillemaut [24] (using temporal information) 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 3.4.1. 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 [35]. Comparisons are shown in Table 3 and Figure 15, 16 for static benchmark datasets. Comparison for dynamic scene segmentations are shown in Table 4 and Figure 17, 18. 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 17 and 18. This overcomes limitations of previous multi-view per-frame segmentation.

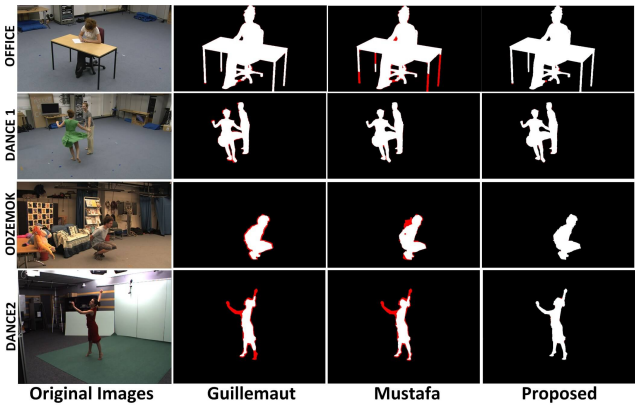

Fig. 17 Segmentation results for dynamic scenes (Error against ground-truth is highlighted in red).

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 4 Dynamic scene segmentation completeness (%)

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

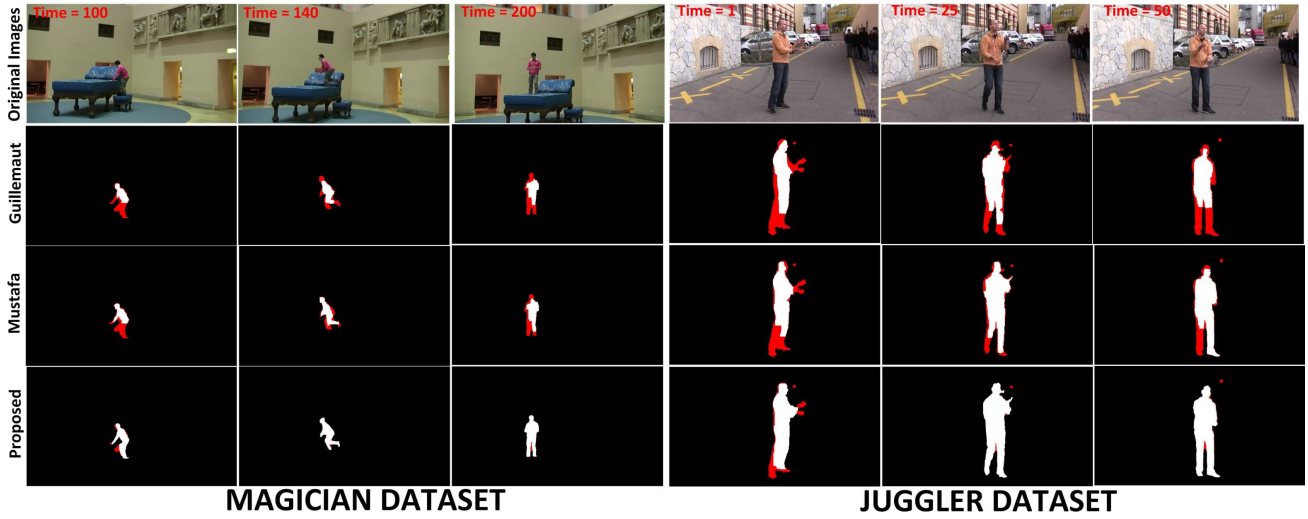

Fig. 18 Segmentation results for dynamic scenes on sequence of frames (Error against ground-truth is highlighted in red).

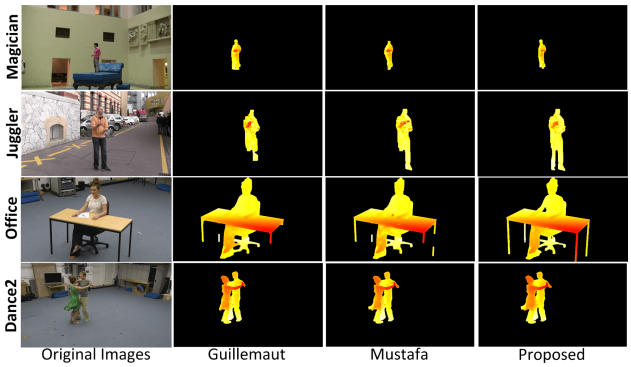

Fig. 19 Comparison of depth maps against existing methods for two indoor and two outdoor benchmark datasets.

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 5 Comparison of computational efficiency for dynamic datasets (time in seconds (s))

4.2 Reconstruction evaluation

Reconstruction results obtained using the proposed method are compared against Mustafa [42], Guillemaut [24], and Furukawa [19] for dynamic sequences. Furukawa [19] is a per-frame multi-view wide-baseline stereo approach which ranks highly on the middlebury benchmark [53] but does not refine the segmentation.

The depth maps obtained using the proposed approach are compared against Mustafa and Guillemaut in Figure 19. The depth map obtained using the proposed approach are smoother with low reconstruction noise compared to the state-of-the-art methods. Figure 20 and 21 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 reconstruction from real multi-view video. In Figure 21 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 20 resulting in more complete shape. Figure 22 shows three complete scene reconstructions with 4D models of multiple objects. The Juggler and Magician sequences are reconstructed from moving handheld cameras.

Computational Complexity: Computation times for the proposed approach vs other methods are presented in Table 5. 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.

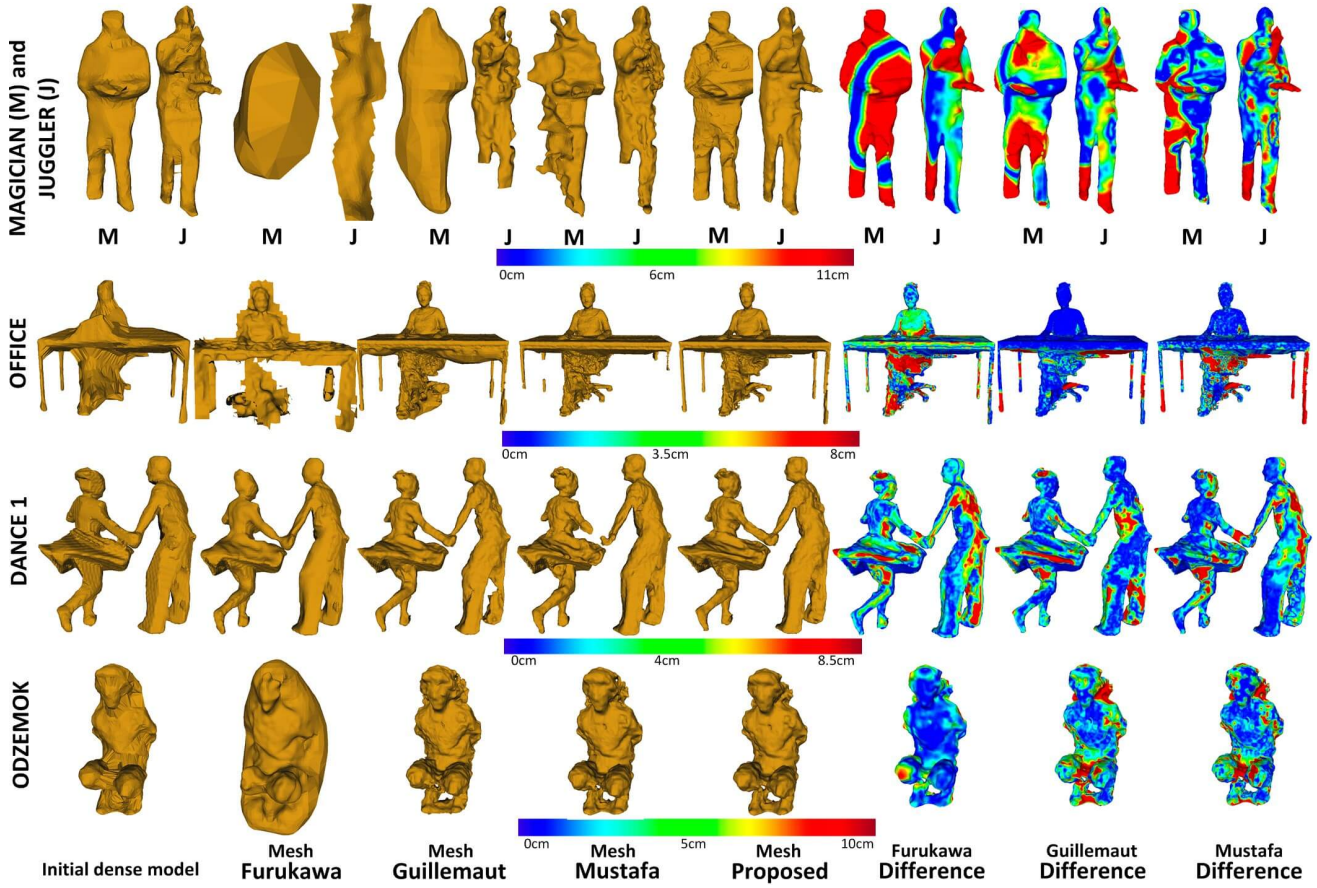

Fig. 20 Reconstruction result mesh comparison against state-of-the-art methods with errors shown in the last three columns

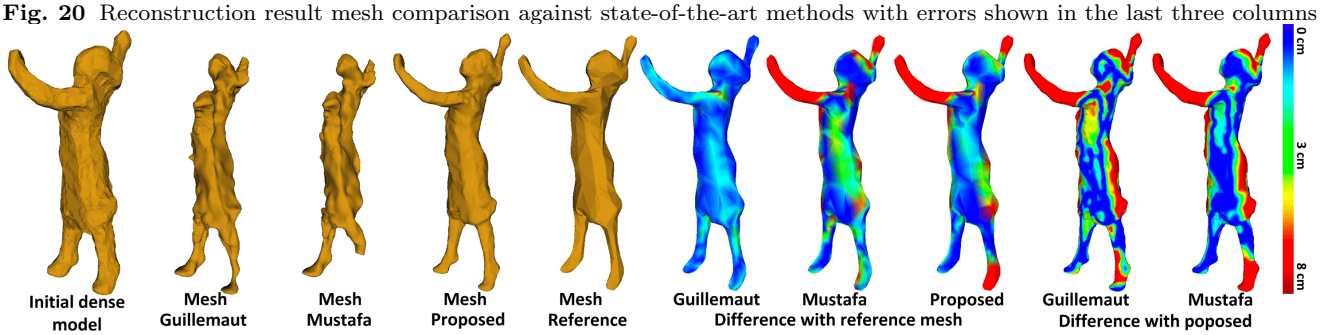

Fig. 21 Reconstruction result comparison with reference mesh and proposed for Dance2 benchmark dataset

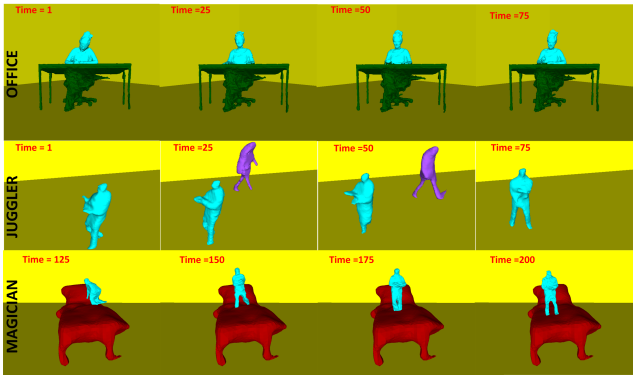

Fig. 22 Complete scene reconstruction with 4D mesh sequence.

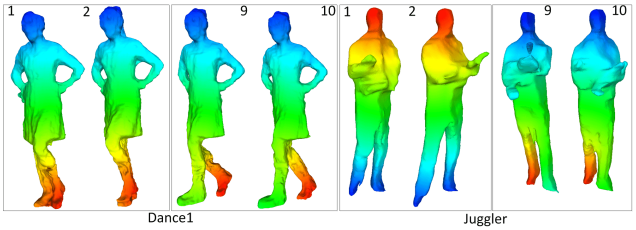

Fig. 23 Frame-to-frame temporal alignment for Dance1 and Juggler dataset

Temporal coherence: A frame-to-frame alignment is obtained using the proposed approach as shown in Figure 23 for Dance1 and Juggle dataset. The meshes of the dynamic object in Frame 1 and Frame 9 are color

coded in both the datasets and the color is propagated to the next frame using the dense temporal coherence information. The color in different parts of the object is retained to the next frame as seen from the figure. The proposed approach obtains sequential temporal alignment which drifts with large movement in the object, hence successive frames are shown in the figure.

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.

Fig. 24 Application of proposed method for freeview-point video for Dance2 dataset.

5 Applications to immersive content production

The 4D meshes generated from the proposed approach can be used for applications in immersive content production such as FVV rendering and VR. This section demonstrates the results of these applications.

5.1 Free-viewpoint rendering

In FVV, the virtual viewpoint is controlled interactively by the user. The appearance of the reconstruction is sampled and interpolated directly from the captured camera images using cameras located close to the virtual viewpoint [57].

The proposed joint segmentation and reconstruction framework generates per-view silhouettes and a temporally coherent 4D reconstruction at each time instant of the input video sequence. This representation of the dynamic sequence is used for FVV rendering. To create FVV, a view-dependent surface texture is computed based on the user selected virtual view. This virtual view is obtained by combining the information from camera views in close proximity to the virtual viewpoint [57]. FVV rendering gives user the freedom to interactively choose a novel viewpoint in space to observe the dynamic scene and reproduces fine scale temporal surface details, such as the movement of hair and clothing wrinkles, that may not be modelled geometrically. An example of a reconstructed scene and the camera configuration is shown in Figure 24.

A qualitative evaluation of images synthesised using FVV is shown in Figure 25 and 26. These demonstrate reconstruction results rendered from novel viewpoints from the proposed method against Mustafa [43] and Guillemaut [23] on publicly available datasets. This is particularly important for wide-baseline camera configurations where this technique can be used to synthesize intermediate viewpoints where it may not be practical or economical to physically locate real cameras.

5.2 Virtual reality rendering

There is a growing demand for photo-realistic content in the creation of immersive VR experiences. The 4D temporally coherent reconstructions of the dynamic scenes obtained using the proposed approach enables the creation of photo-realistic digital assets that can be incorporated into VR environments using game engines such as Unity and Unreal Engine, as shown in Figure 27 for single frame of four datasets and for a series of frames for Dance1 dataset.

In order to efficiently render the reconstructions in a game engine for applications in VR, a UV texture atlas is extracted using the 4D meshes from the proposed approach as a geometric proxy. The UV texture atlas at each frame are applied to the models at render time in unity for viewing in a VR headset. A UV texture atlas is constructed by projectively texturing and blending multiple view frames onto a 2D unwrapped UV texture atlas, see Figure Figure 28. This is performed once for each static object and at each time instance for dynamic objects allowing efficient storage and real-time playback of static and dynamic textured reconstructions within a VR headset.

6 Conclusion

This paper introduced a novel technique to automatically segment and reconstruct dynamic scenes captured from multiple moving cameras in general dynamic uncontrolled environments without any prior on background appearance or structure. The proposed automatic initialization was used to identify and initialize the segmentation and reconstruction of multiple objects. 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

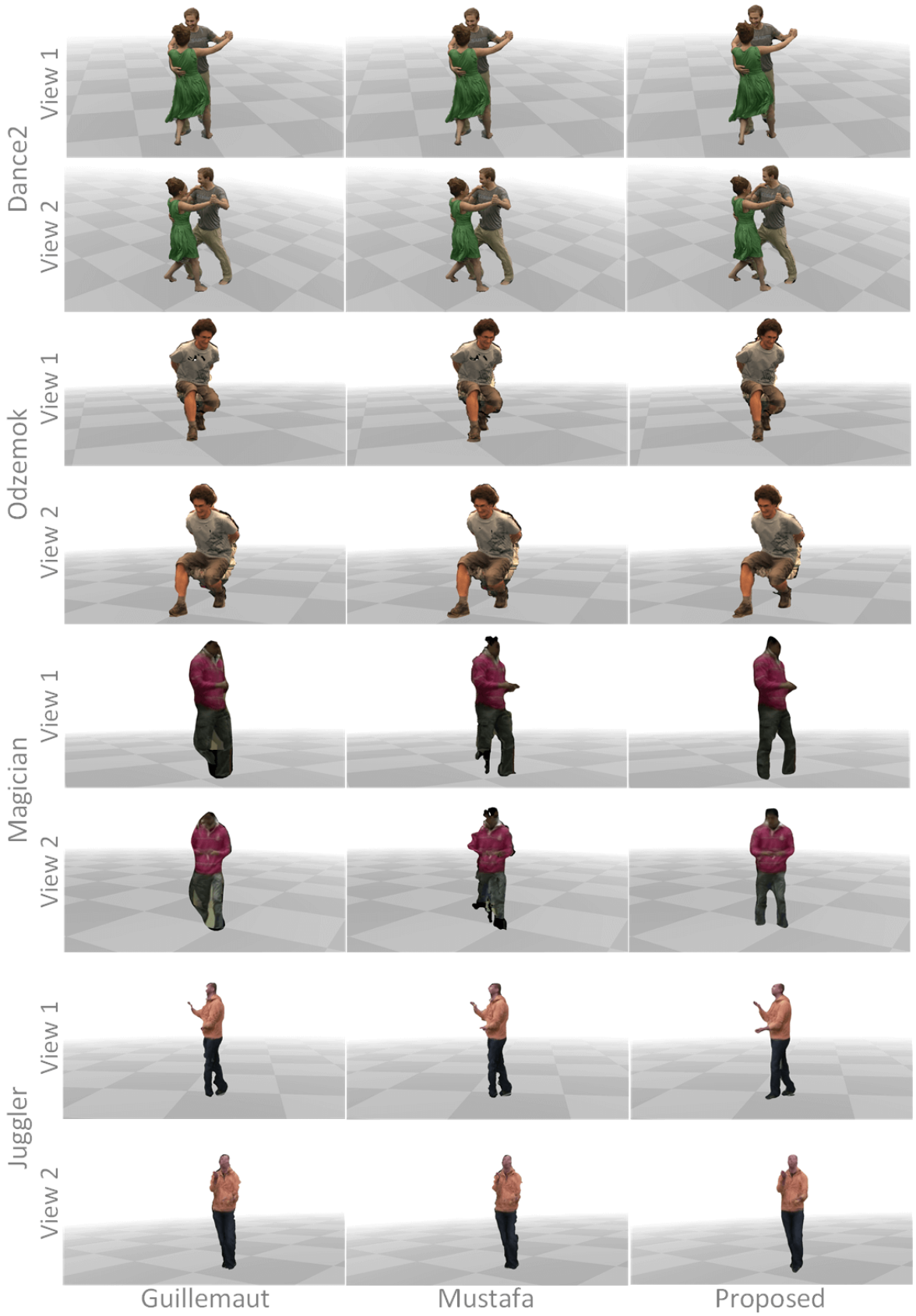

Fig. 25 Comparison of Free-viewpoint rendering of proposed method against Mustafa and Guillemaut for 4 datasets.

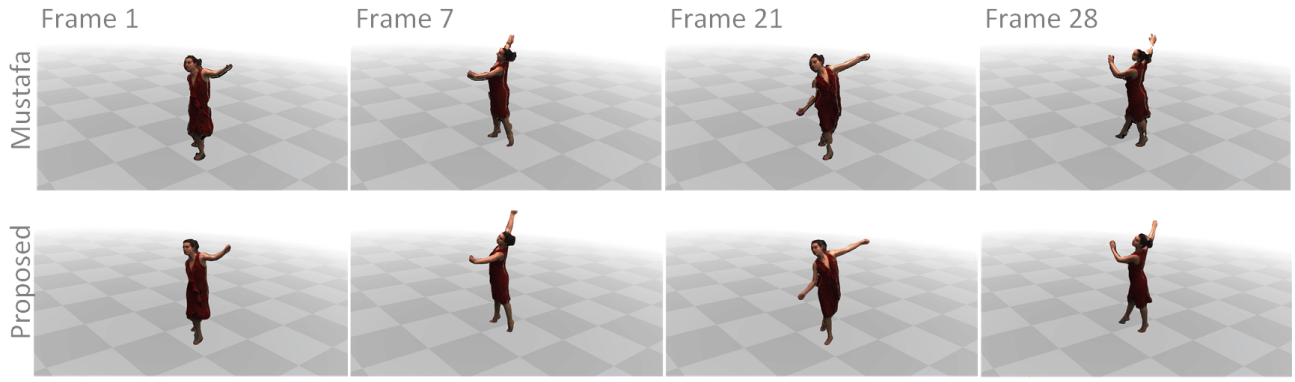

Fig. 26 Comparison of Free-viewpoint rendering of proposed method against Mustafa for Dance1 sequence.

Fig. 27 Application of proposed method for Virtual Reality. Renderings in Unity are shown for five datasets, including a sequence for Dance1.

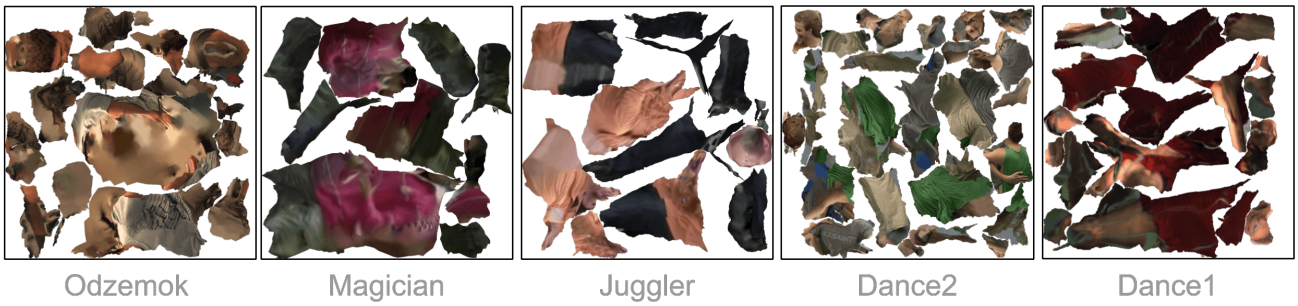

Fig. 28 UV texture atlases for different dynamic datasets to render in VR at a single time instance.

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.

Acknowledgements This research was supported by the Royal Academy of Engineering Research Fellowship RF-201718-17177 and the EPSRC Platform Grant on Audio-Visual Media Research EP/P022529.

References

1. 4d and multiview video repository. In: Centre for Vision Speech and Signal Processing, University of Surrey, UK
2. 4d repository, http://4drepository.inrialpes.fr/. In: Institut national de recherche en informatique et en automatique (INRIA) Rhone Alpes
3. Atapour-Abarghouei, A., Breckon, T.P.: Veritatem dies aperit - temporally consistent depth prediction enabled by a multi-task geometric and semantic scene understanding approach. In: CVPR (2019)
4. Bailer, C., Taetz, B., Stricker, D.: Flow fields: Dense correspondence fields for highly accurate large displacement optical flow estimation. In: ICCV (2015)
5. Ballan, L., Brostow, G.J., Puwein, J., Pollefeys, M.: Unstructured video-based rendering: Interactive exploration of casually captured videos. ACM Trans. on Graphics pp. 1–11 (2010)
6. Basha, T., Moses, Y., Kiryati, N.: Multi-view scene flow estimation: A view centered variational approach. In: CVPR, pp. 1506–1513 (2010)
7. Bleyer, M., Rhemann, C., Rother, C.: Patchmatch stereo - stereo matching with slanted support windows. In: BMVC (2011)
8. Bouguet, J.: Pyramidal implementation of the lucas kanade feature tracker. Intel Corporation, Microprocessor Research Labs (2000)
9. Boykov, Y., Kolmogorov, V.: An experimental comparison of min-cut/max- flow algorithms for energy minimization in vision. PAMI **26**, 1124–1137 (2004)
10. Boykov, Y., Veksler, O., Zabih, R.: Fast approximate energy minimization via graph cuts. PAMI **23**, 1222–1239 (2001)
11. Campbell, N., Vogiatzis, G., Hernndez, C., Cipolla, R.: Automatic 3d object segmentation in multiple views using volumetric graph-cuts. Image and Vision Computing **28**, 14 – 25 (2010)
12. Chen, P.Y., Liu, A.H., Liu, Y.C., Wang, Y.C.F.: Towards scene understanding: Unsupervised monocular depth estimation with semantic-aware representation. In: CVPR (2019)
13. Coughlan, J.M., Yuille, A.L.: The manhattan world assumption: Regularities in scene statistics which enable bayesian inference. In: NIPS, pp. 845–851 (2000)
14. Das, P., Veksler, O., Zavadsky, V., Boykov, Y.: Semiautomatic segmentation with compact shape prior. Image and Vision Computing **27**, 206–219 (2009)
15. Dimitrov, D., Knauer, C., Kriegel, K., Rote, G.: On the bounding boxes obtained by principal component analysis (2006)
16. Djelouah, A., Franco, J.S., Boyer, E., Le Clerc, F., Perez, P.: Multi-view object segmentation in space and time. In: ICCV, pp. 2640–2647 (2013)
17. Djelouah, A., Franco, J.S., Boyer, E., Le Clerc, F., Perez, P.: Sparse multi-view consistency for object segmentation. PAMI pp. 1–1 (2015)
18. Fortune, S.: Handbook of discrete and computational geometry. chap. Voronoi Diagrams and Delaunay Triangulations, pp. 377–388 (1997)
19. Furukawa, Y., Ponce, J.: Accurate, dense, and robust multiview stereopsis. PAMI **32**, 1362–1376 (2010)
20. Goldluecke, B., Magnor, M.: Space-time isosurface evolution for temporally coherent 3d reconstruction. In: CVPR, pp. 350–355 (2004)
21. Grundmann, M., Kwatra, V., Han, M., Essa, I.: Efficient hierarchical graph based video segmentation. CVPR (2010)
22. Guan, L., Franco, J.S., Pollefeys, M.: Multi-view occlusion reasoning for probabilistic silhouette-based dynamic scene reconstruction. IJCV **90**, 283–303 (2010)
23. Guillemaut, J.Y., Hilton, A.: Joint Multi-Layer Segmentation and Reconstruction for Free-Viewpoint Video Applications. IJCV **93**, 73–100 (2010)
24. Guillemaut, J.Y., Hilton, A.: Space-time joint multi-layer segmentation and depth estimation. In: 3DIMPVT, pp. 440–447 (2012)
25. Gulshan, V., Rother, C., Criminisi, A., Blake, A., Zisserman, A.: Geodesic star convexity for interactive image segmentation. In: CVPR, pp. 3129–3136 (2010)
26. Hane, C., Zach, C., Cohen, A., Angst, R., Pollefeys, M.: Joint 3d scene reconstruction and class segmentation. In: CVPR, pp. 97–104 (2013)
27. Hartley, R., Zisserman, A.: Multiple View Geometry in Computer Vision, 2 edn. Cambridge University Press (2003)
28. Hu, X., Mordohai, P.: A quantitative evaluation of confidence measures for stereo vision. PAMI pp. 2121–2133 (2012)
29. Huang, Z., Li, T., Chen, W., Zhao, Y., Xing, J., LeGendre, C., Luo, L., Ma, C., Li, H.: In: ECCV (2018)
30. Jiang, H., Liu, H., Tan, P., Zhang, G., Bao, H.: 3d reconstruction of dynamic scenes with multiple handheld cameras. In: ECCV, pp. 601–615 (2012)
31. Kazhdan, M., Bolitho, M., Hoppe, H.: Poisson surface reconstruction. In: Eurographics Symposium on Geometry Processing, pp. 61–70 (2006)
32. Kim, H., Guillemaut, J., Takai, T., Sarim, M., Hilton, A.: Outdoor Dynamic 3-D Scene Reconstruction. CSVT **22**, 1611–1622 (2012)
33. Kohli, P., Rihan, J., Bray, M., Torr, P.H.: Simultaneous segmentation and pose estimation of humans using dynamic graph cuts. Int. J. Comput. Vision **79**(3), 285–298 (2008)
34. Kolmogorov, V., Criminisi, A., Blake, A., Cross, G., Rother, C.: Probabilistic fusion of stereo with color and contrast for bilayer segmentation. PAMI **28**, 2006 (2006)
35. Kowdle, A., Sinha, S., Szeliski, R.: Multiple view object cosegmentation using appearance and stereo cues. In: ECCV, pp. 789–803 (2012)
36. Kundu, A., Li, Y., Dellaert, F., Li, F., Rehg, J.M.: Joint semantic segmentation and 3d reconstruction from monocular video. In: ECCV, vol. 8694, pp. 703–718 (2014)
37. Larsen, E., Mordohai, P., Pollefeys, M., Fuchs, H.: Temporally consistent reconstruction from multiple video streams using enhanced belief propagation. In: ICCV, pp. 1–8 (2007)
38. Lee, W., Woo, W., Boyer, E.: Silhouette segmentation in multiple views. PAMI pp. 1429–1441 (2011)
39. Lei, C., Chen, X.D., Yang, Y.H.: A new multiview spacetime-consistent depth recovery framework for free viewpoint video rendering. In: ICCV, pp. 1570–1577 (2009)
40. Matthies, L.: Stereo vision for planetary rovers: Stochastic modeling to near real-time implementation. IJCV **8**, 71–91 (1992)
41. Menze, M., Geiger, A.: Object scene flow for autonomous vehicles. In: CVPR (2015)
42. Mustafa, A., Kim, H., Guillemaut, J., Hilton, A.: General dynamic scene reconstruction from wide-baseline views. In: ICCV (2015)
43. Mustafa, A., Kim, H., Guillemaut, J.Y., Hilton, A.: Temporally coherent 4d reconstruction of complex dynamic scenes. In: CVPR, Oral (2016)

44. Mustafa, A., Kim, H., Imre, E., Hilton, A.: Segmentation based features for wide-baseline multi-view reconstruction. In: 3DV (2015)
45. Narayana, M., Hanson, A., Learned-Miller, E.: Coherent motion segmentation in moving camera videos using optical flow orientations. In: ICCV, pp. 1577–1584 (2013)
46. Ngo, T., Nagahara, H., Nishino, K., Taniguchi, R., Yagi, Y.: Reflectance and shape estimation with a light field camera under natural illumination. IJCV (2019)
47. Oswald, M., Sthmer, J., Cremers, D.: Generalized connectivity constraints for spatio-temporal 3d reconstruction. In: ECCV 2014, pp. 32–46 (2014)
48. Ozden, K., Schindler, K., Van Gool, L.: Simultaneous segmentation and 3d reconstruction of monocular image sequences. In: ICCV, pp. 1–8 (2007)
49. Papazoglou, A., Ferrari, V.: Fast object segmentation in unconstrained video. In: ICCV, pp. 1777–1784 (2013)
50. Qian, Y., Gong, M., Yang, Y.H.: Stereo-based 3d reconstruction of dynamic fluid surfaces by global optimization. In: CVPR (2017)
51. Rusu, R.B.: Semantic 3d object maps for everyday manipulation in human living environments. Ph.D. thesis, Computer Science department, Technische Universitaet Muenchen, Germany (2009)
52. Sarim, M., Hilton, A., Guillemaut, J.Y.: Temporal trimap propagation for video matting using inferential statistics. In: ICIP, pp. 1745–1748 (2011)
53. Seitz, S., Curless, B., Diebel, J., Scharstein, D., Szeliski, R.: A comparison and evaluation of multi-view stereo reconstruction algorithms. In: CVPR, pp. 519–528 (2006)
54. Shin, Y.M., Cho, M., Lee, K.M.: Multi-object reconstruction from dynamic scenes: An object-centered approach. CVIU **117**, 1575 – 1588 (2013)
55. Slavcheva, M., Baust, M., Cremers, D., Ilic, S.: Killingfusion: Non-rigid 3d reconstruction without correspondences. In: CVPR (2017)
56. Starck, J., Hilton, A.: Surface Capture for Performance-Based Animation. IEEE Computer Graph. and Appl. **27**, 21–31 (2007)
57. Starck, J., Kilner, J., Hilton, A.: A free-viewpoint video renderer. Journal of Graphics, GPU, and Game Tools **14**(3), 57–72 (2009)
58. Stutz, D., Geiger, A.: Learning 3d shape completion under weak supervision. IJCV (2018)
59. Szeliski, R., Golland, P.: Stereo matching with transparency and matting. In: ICCV, pp. 517–524 (1998)
60. Taneja, A., Ballan, L., Pollefeys, M.: Modeling dynamic scenes recorded with freely moving cameras. In: ACCV, pp. 613–626 (2011)
61. Tomasi, C., Manduchi, R.: Bilateral filtering for gray and color images. In: ICCV, pp. 839–846 (1998)
62. Tung, T., Nobuhara, S., Matsuyama, T.: Complete multi-view reconstruction of dynamic scenes from probabilistic fusion of narrow and wide baseline stereo. In: ICCV, pp. 1709–1716 (2009)
63. Veksler, O.: Star shape prior for graph-cut image segmentation. In: ECCV, pp. 454–467 (2008)
64. Vicente, S., Kolmogorov, V., Rother, C.: Graph cut based image segmentation with connectivity priors. In: CVPR, pp. 1–8 (2008)
65. Vo, M., Narasimhan, S.G., Sheikh, Y.: Spatiotemporal bundle adjustment for dynamic 3d reconstruction. In: CVPR (2016)
66. Wedel, A., Brox, T., Vaudrey, T., Rabe, C., Franke, U., Cremers, D.: Stereoscopic scene flow computation for 3d motion understanding. IJCV **95**, 29–51 (2011)
67. Wu, C.: Towards linear-time incremental structure from motion. In: 3DV, pp. 127–134 (2013)
68. Wu, S., Huang, H., Portenier, T., Sela, M., Cohen-Or, D., Kimmel, R., Zwicker, M.: Specular-to-diffuse translation for multi-view reconstruction. In: ECCV (2018)
69. Zach, C., Cohen, A., Pollefeys, M.: Joint 3d scene reconstruction and class segmentation. In: CVPR (2013)
70. Zeng, G., Quan, L.: Silhouette extraction from multiple images of an unknown background. In: ACCV (2004)
71. Zhang, D., Javed, O., Shah, M.: Video object segmentation through spatially accurate and temporally dense extraction of primary object regions. In: CVPR (2013)
72. Zhang, G., Jia, J., Hua, W., Bao, H.: Robust bilayer segmentation and motion/depth estimation with a handheld camera. PAMI (2011)