General Dynamic Scene Reconstruction from Multiple View Video

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

Abstract

This paper introduces a general approach to dynamic scene reconstruction from multiple moving cameras without prior knowledge or limiting constraints on the scene structure, appearance, or illumination. Existing techniques for dynamic scene reconstruction from multiple wide-baseline camera views primarily focus on accurate reconstruction in controlled environments, where the cameras are fixed and calibrated and background is known. These approaches are not robust for general dynamic scenes captured with sparse moving cameras. Previous approaches for outdoor dynamic scene reconstruction assume prior knowledge of the static background appearance and structure. The primary contributions of this paper are twofold: an automatic method for initial coarse dynamic scene segmentation and reconstruction without prior knowledge of background appearance or structure; and a general robust approach for joint segmentation refinement and dense reconstruction of dynamic scenes from multiple wide-baseline static or moving cameras. Evaluation is performed on a variety of indoor and outdoor scenes with cluttered backgrounds and multiple dynamic non-rigid objects such as people. Comparison with state-of-the-art approaches demonstrates improved accuracy in both multiple view segmentation and dense reconstruction. The proposed approach also eliminates the requirement for prior knowledge of scene structure and appearance.

1. Introduction

Reconstruction of general dynamic scenes is motivated by potential applications in film and broadcast production together with the ultimate goal of automatic understanding of real-world scenes from distributed camera networks.

Over the past decades, effective approaches have been proposed to reconstruct dense dynamic shape from wide-baseline camera views in controlled environments with static backgrounds and illumination. A common assumption of widely used visual-hull based reconstruction approaches is prior foreground/background segmentation, which is commonly achieved using a uniform chroma-key color background or background image plate. Alternatively, multiple view stereo techniques have been developed which require a relatively dense camera network resulting in large numbers of cameras.

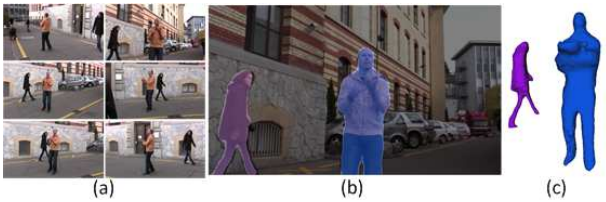

Figure 1. General dynamic scene reconstruction (a) Multi-view frames for Juggler dataset, (b) Segmentation of dynamic objects and (c) Reconstructed mesh

Recent research has applied multiple view dynamic scene reconstruction techniques to less controlled outdoor scenes. Initial research focused on reconstruction in sports [10] exploiting known background images or the pitch color to obtain an initial segmentation. Extension to more general outdoor scenes [1, 15, 31] uses prior reconstruction of the static geometry from images of the empty environment. Research has also exploited strong prior models of dynamic scene structure such as people or used active depth sensors to reconstruct dynamic scenes.

This paper presents an approach for unsupervised dynamic scene reconstruction from multiple wide-baseline static or moving camera views without prior knowledge of the scene structure or background appearance. The input is a sparse set of synchronised multiple view videos without segmentation. Camera extrinsics are automatically calibrated using scene features. An initial coarse reconstruction and segmentation of all dynamic 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. Joint segmentation and dense reconstruction refinement is then performed to estimate the non-rigid shape of dynamic objects at each frame. Robust methods are introduced to handle complex dynamic scene geometry in cluttered scenes from independently moving wide-baseline cameras views. The proposed approach overcomes constraints of existing approaches allowing the reconstruction of more general dynamic scenes. Results for a popular dataset, Juggler [1] captured with a network of moving handheld cameras are shown in Figure 1. The contributions are as follows:

- Unsupervised dense reconstruction and segmentation of general dynamic scenes from multiple wide-baseline views.
- Automatic initialization of dynamic object segmentation and reconstruction from sparse features.
- Robust spatio-temporal refinement of dense reconstruction and segmentation integrating error tolerant photo-consistency and edge information.

2. Related work

2.1. Dynamic scene reconstruction

Research on multiple view dense dynamic reconstruction has primarily focused on indoor scenes with controlled illumination and backgrounds extending methods for multiple view reconstruction of static scenes [28] to sequences [34]. 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. Initial research focused on narrow baseline stereo [20, 18] requiring a large number of closely spaced cameras for complete reconstruction of dynamic shape. Practical reconstruction requires relatively sparse moving cameras to acquire coverage over large outdoor areas. A number of approaches for reconstruction of outdoor scenes require initial silhouette segmentation [35, 15, 9, 10] to allow visual-hull reconstruction. Recent research has proposed reconstruction from a single handheld moving camera given a strong prior of bilayer segmentation [39]. Bi-layer segmentation is used for depth-map reconstruction with the DAISY descriptor for matching [13], results are presented for handheld cameras with a relatively narrow baseline.

Pioneering research in general dynamic scene reconstruction from multiple handheld wide-baseline cameras [1, 31] exploited prior reconstruction of the background scene to allow dynamic foreground segmentation and reconstruction. This requires images of the environment captured in the absence of dynamic elements to recover the background geometry and appearance.

Most of these approaches to general dynamic scene reconstruction fail in case of complex (cluttered) scenes captured with moving cameras. These approaches either work for static/indoor scenes or exploit strong prior assumptions like silhouette information, known background or scene structure. Our aim is to perform dense reconstruction of dynamic scene automatically without any prior knowledge of background or segmentation of dynamic object.

2.2. Joint segmentation and reconstruction

Segmentation from multiple wide-baseline views has been proposed by exploiting appearance similarity [6, 19, 38]. These approaches assume static backgrounds and different colour distributions for the foreground and background [27, 6] which limits applicability for general scenes. In contrast to overcome these limitations, the proposed approaches initialised the foreground object segmentation from wide-baseline feature correspondence followed by joint segmentation and reconstruction.

Joint segmentation and reconstruction methods incorporate estimation of segmentation or matting with reconstruction to provide a combined solution. The first multi-view joint estimation system was proposed by Szeliski et al.[30] 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 [36]. The focus shifted to joint segmentation and reconstruction for rigid objects in indoor and outdoor environment. Approaches used a variety of techniques like patch based refinement [29, 25] and fixation of cameras on the object of interest [5].

Practical application of joint estimation requires these approaches to work on non-rigid objects like humans with clothing. Recent work proposed joint reconstruction and segmentation on monocular video achieving semantic segmentation of scene but does not work with dynamic objects [17]. A multi-layer segmentation and reconstruction approach was proposed for sports data and indoor sequences [10] for multi-view videos. The algorithm used visual hull as a prior obtained from segmentation of the dynamic objects. The visual hull was optimized by 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 a background plate (assumed unknown in our case) as a prior to estimate the initial visual hull by background subtraction. The probabilistic color models of foreground and background are also used for optimization. A quantitative evaluation of state-of-the-art techniques for reconstruction from multiple views was presented by [28]. These methods are able to produce high quality results, but rely on good initializations and strong prior assumptions.

Image-based 3D dynamic scene reconstruction without a prior model is a key problem in computer vision. This research aims to overcome the limitations of the discussed approaches enabling robust wide-baseline multiple view reconstruction of general dynamic scenes without prior assumptions on scene appearance, structure or segmentation of the moving objects. The approach identifies and obtains an initial coarse reconstruction of dynamic objects automatically which is then refined using geometry and appearance cues in an optimization framework. The approach is a significant development over existing approaches as it works for the scenes captured only with moving cameras with unknown background and structure. Existing state-of-the-art techniques has not addressed this problem until now.

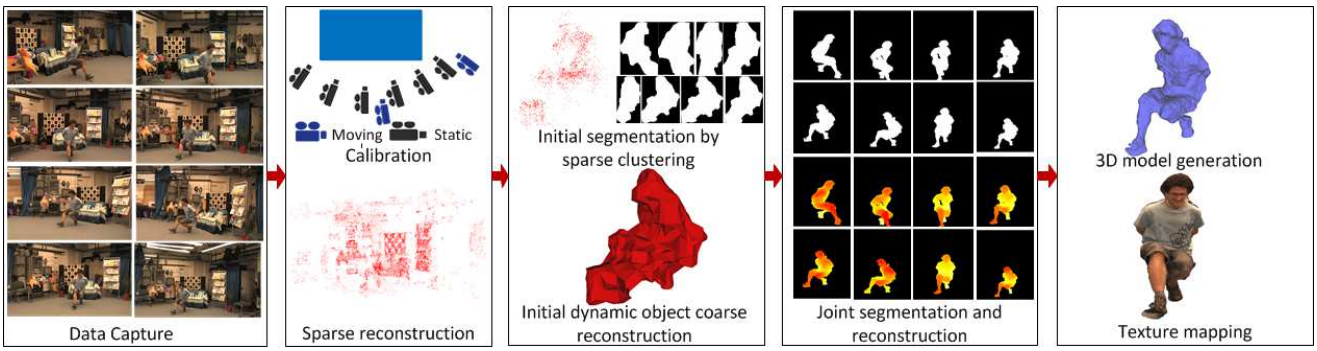

Figure 2. Dense dynamic reconstruction framework

3. Overview

The motivation of our work is to obtain automatic dense reconstruction and segmentation of complex dynamic scenes from multiple wide-baseline camera views without restrictive assumptions on scene structure or camera motion. The proposed approach estimates per-pixel dense depth with respect to each camera view of the observed moving non-rigid objects in the scene. View-dependent depth maps are then fused to obtain a reconstruction for each dynamic object. An overview of the approach is presented in Figure 2 and consists of the following stages:

Data Capture: The scene is captured using multiple synchronised video cameras separated by wide-baseline.

Calibration and sparse reconstruction: The intrinsics are assumed to be known for the static cameras and extrinsics are calibrated using Fundamental matrix estimation for pairs of images followed by bundle adjustment. Moving cameras are calibrated automatically using multi-camera calibration [12]. A sparse 3D point-cloud is then reconstructed from wide-baseline feature matches.

Initial dynamic object segmentation and reconstruction: Automatic initialisation is performed without prior knowledge of the scene structure or appearance to obtain an initial approximation for each dynamic object. Dynamic objects are segmented from the sparse 3D point cloud by combining optic flow with 3D clustering (section 4).

Joint segmentation and reconstruction for each dynamic object: The initial coarse reconstruction is refined for each dynamic object through joint optimisation of shape and segmentation using a robust cost function for wide-baseline matching. View-dependent optimisation of depth is performed with respect to each camera which is robust to errors in camera calibration and initialisation. This gives a set of dense depth maps for each dynamic object.

3D model generation and texture mapping: A single 3D model for each dynamic object is obtained by fusion of the view-dependent depth maps using Poisson surface reconstruction [14]. Surface orientation is estimated based on neighbouring pixels. Projective texture mapping is then performed for free-viewpoint video rendering.

Dense reconstruction of sequence: The process above is repeated for the entire sequence for all dynamic objects.

The proposed approach enables automatic reconstruction of all dynamic 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 reconstruction. The approach is demonstrated to outperform previous approaches to dynamic scene reconstruction and does not require prior knowledge of the scene structure.

4. Initial dynamic object reconstruction

For general dynamic scene reconstruction, we need to reconstruct and segment the dynamic objects in the scene at each frame instead of whole scene reconstruction for computational efficiency and to avoid redundancy. 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 and optical flow labelling. This approach is robust to scene clutter in the 3D point cloud segmentation and partial segmentation of the dynamic object using optic flow due to partial motion or correspondence failure. Initialisation gives a complete coarse segmentation and reconstruction of each dynamic object for subsequent refinement. The optic flow and cluster information for each dynamic object helps us to retain same labels for the entire sequence.

4.1. Sparse point cloud clustering

Feature detection is performed on all the multi-view images [24]. This is followed by SIFT descriptor based feature matching [21] to obtain sparse reconstruction of the scene using the calibration information [12] for each time instant. This representation of the scene is processed to remove outliers using the point neighbourhood statistics to filter outlier data [26]. To retrieve the sparse features corresponding to the dynamic objects from the sparse reconstruction of the scene, we classify this representation into clusters followed by optical flow labelling. Data clustering approach is applied based on the 3D grid subdivision of the space using an octree data structure in Eucildean space. In a more general sense, nearest neighbors information is used to cluster, that is essentially similar to a flood fill algorithm [26]. We choose this because of its computational efficiency and ro-

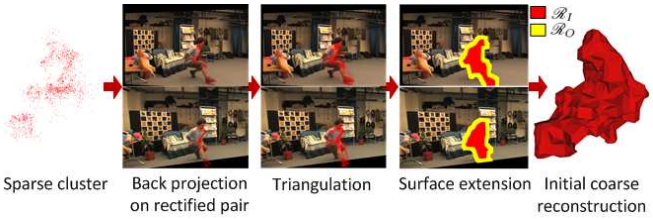

Figure 3. Initial coarse reconstruction of the dynamic object in Odzemok dataset

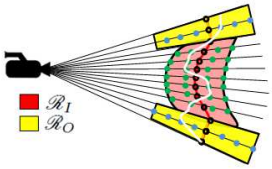

Figure 4. Initial coarse reconstruction: White line represents the actual surface, Depth labels are represented as circles; blue circles depict depth labels in $\mathscr{D}_O$, green circles depict depth labels in $\mathscr{D}_I$ and black circles depict the initial surface estimate.

bustness. The approach allows unsupervised segmentation of dynamic objects and is proved to work well for cluttered and general outdoor scenes as shown in Section 6.

4.2. Coarse scene reconstruction

Dynamic elements of the scene are identified by performing optical flow [37] on consecutive frames for a single view of each cluster. For each cluster the optimal camera view is dynamically selected to maximise visibility based on the sparse dynamic feature points at each frame. This allows efficient selection of the best view for optical flow. Optical flow is used to assign a unique label for each dynamic cluster throughout the sequence. If an object does not move between two consecutive time instants the reconstruction from this previous frame is retained. This limits the dynamic scene reconstruction to objects which have moved between frames reducing computational cost.

The process to obtain the coarse reconstruction is shown in Figure 3 and 4. The sparse representation of dynamic element is back-projected on the rectified image pair for each view. Delaunay triangulation [7] 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 [23]. 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 3 and 4. 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. We assume that the object boundaries lie within the initial coarse estimate and depth at each pixel for the combined regions may not be accurate. Hence, to handle these errors in depth we add volume in front and behind of the projected surface by an error tolerance (calculated experimentally), along the optical ray of the camera. This tolerance 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 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 the dynamic 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 coarse reconstruction can be performed for multiple dynamic objects in the complex general environments. Initial dynamic object segmentation using point cloud clustering and coarse segmentation is insensitive to parameters. Throughout this work the same parameters are used for all datasets.

5. Joint segmentation and reconstruction

5.1. Problem statement

In this section our aim is to refine the depth of the initial coarse reconstruction estimate of each dynamic object. We aim to assign an accurate depth value to each pixel p from a set of depth values $\mathscr{D} = \{d_1, ..., d_{|\mathscr{D}|-1}, \mathscr{U}\}$. Each d_i is obtained by sampling the optical ray from the camera and $\mathscr{U}$ is an unknown depth value to handle occlusions and to refine object segmentation. We assume that the depth of a particular pixel lies within the given threshold around the initial estimate as depicted in Figure 4 and varies depending upon the regions $\mathscr{R}_I$ or $\mathscr{R}_O$. Hence we divide our depth labels in two sets, one for the region $\mathscr{R}_I$ ($\mathscr{D}_I$) and other for $\mathscr{R}_O$ ($\mathscr{D}_O$) such that $|\mathscr{D}_I| < |\mathscr{D}_O|$.

5.2. Proposed approach

We formulate the computation of depth at each point as energy minimization of the cost function defined in Eq. (1). This equation is specifically designed to refine the reconstruction and segmentation and is used to estimate a view-dependent depth map for each dynamic object with respect to each camera.

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

where, d is the depth at each pixel for our dynamic object for the region $\mathscr{R}_I + \mathscr{R}_O$ and can be assigned $\mathscr{U}$ to refine object segmentation. The equation consist of three 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 contrast term is to identify the object boundaries. Data and smoothness terms are common to solve reconstruction problems [2] and the contrast term is used for segmentation [16].

5.2.1 Matching term

To measure photo-consistency, we use a data term measure based on NCC recently proposed in [11]. They suggests this to be the best photo-consistency measure for wide baseline multi-view datasets because of its ability to obtain a high number of correct matches and preserve boundaries.
$E_{data}(d) = \sum_{p\in\mathscr{P}} e_{data}(p, d_p) =$

$$\begin{cases} M(p,q) = \sum_{i\in\mathscr{C}_k} m(p,q), & \text{if } d_p \neq \mathscr{U} \\ M_{\mathscr{U}}, & \text{if } d_p = \mathscr{U} \end{cases} \quad (2)$$

where $\mathscr{P}$ is the 4-connected neighbourhood of pixel p, $M_{\mathscr{U}}$ is the fixed cost of labelling a pixel unknown and $q = \Pi(p, d_p)$ denotes the projection of the hypothesised point P in an auxiliary camera where P is the coordinates of $3D$ point along the optical ray passing through pixel p located at a distance d_p from the reference camera. $\mathscr{C}_k$ is the set of k most photo-consistent pairs with reference camera. For textured scenes NCC over a squared window is a common choice [28]. 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[22] 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 [11]. The measure is defined as:

$$m(p,q) = \frac{exp\frac{c_{min}}{2\sigma_i^2}}{\sum_{(p,q)\in\mathscr{N}} exp\frac{-(1-NCC(p,q))}{2\sigma_i^2}} \quad (3)$$

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.

5.2.2 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 [16]. We used a modified version of this interpretation in our formulation to preserve the edges by using Bilateral filtering [33] instead of Gaussian filtering.

$$E_{contrast} = \sum_{p,q\in\mathscr{N}} e_{contrast}(p,q) \quad (4)$$

$$e_{contrast}(p,q) = \begin{cases} 0, & \text{if } (d_p = d_q = \mathscr{U}) \text{ or} \\ & (d_p = d_q \neq \mathscr{U}) \\ \frac{1}{1+\epsilon}(\epsilon + exp^{-C(p,q)}), & \text{otherwise} \end{cases} \quad (5)$$

$\|\cdot\|$ 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 [10]. 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.

5.2.3 Smoothness term

This term is inspired by [10] 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.

$$E_{smooth}(d) = \sum_{(p,q)\in\mathscr{N}} e_{smooth}(d_p, d_q) \quad (6)$$

$$e_{smooth}(d_p,d_q) = \begin{cases} min(|d_p - d_q|, d_{max}), & \text{if } d_p, d_q \neq \mathscr{U} \\ 0, & \text{if } d_p, d_q = \mathscr{U} \\ d_{max}, & \text{otherwise} \end{cases} \quad (7)$$

d_{max} is set to 50 times the size of the depth sampling step defined in Section 5.1 for all datasets.

5.3. Optimization of Reconstruction and Segmentation

The energy minimization for Eq. (1) is performed by using the α-expansion move algorithm from [4]. We choose graph cuts because of its strong optimality properties over belief propagation [32]. Graph-cut using the min-cut/max-flow algorithm is used to obtain a local optimum [3]. The α-expansion for a pixel p is performed by iterating through the set of depth labels $\mathscr{D}_I$, if $p \in \mathscr{R}_I$ and $\mathscr{D}_O$, if $p \in \mathscr{R}_O$. Convergence is achieved after 4 or 5 iterations. A final model is obtained by merging the view-dependent depth representations through the Poisson surface reconstruction algorithm as explained in Section 3.

6. Results and Evaluation

Evaluation is performed from publicly available research datasets: Indoor and Outdoor dataset with simple background (Dance2 and Cathedral), Indoor datasets with cluttered background (Odzemok and Dance1) (cvssp.org/cvssp3d) and Indoor and Outdoor datasets captured with moving handheld cameras (Magician and Juggler) [1]. The detailed characteristics of these datasets and the parameter settings for Eq.(1) are summarised in Table 1. The framework explained in Section 3 is applied to all datasets, starting from sparse reconstruction followed by clustering and initial coarse reconstruction of dynamic objects which is then optimized using the proposed joint segmentation and reconstruction approach. Most existing methods do not perform simultaneous segmentation and reconstruction, therefore the method is compared to two state of the art approaches Furukawa and Ponce [8] for wide-baseline reconstruction and Guillemaut and Hilton [10] for joint reconstruction and segmentation. Both of these approaches are top performers on the Middlebury for multi-view reconstruction of wide-baseline views [28].

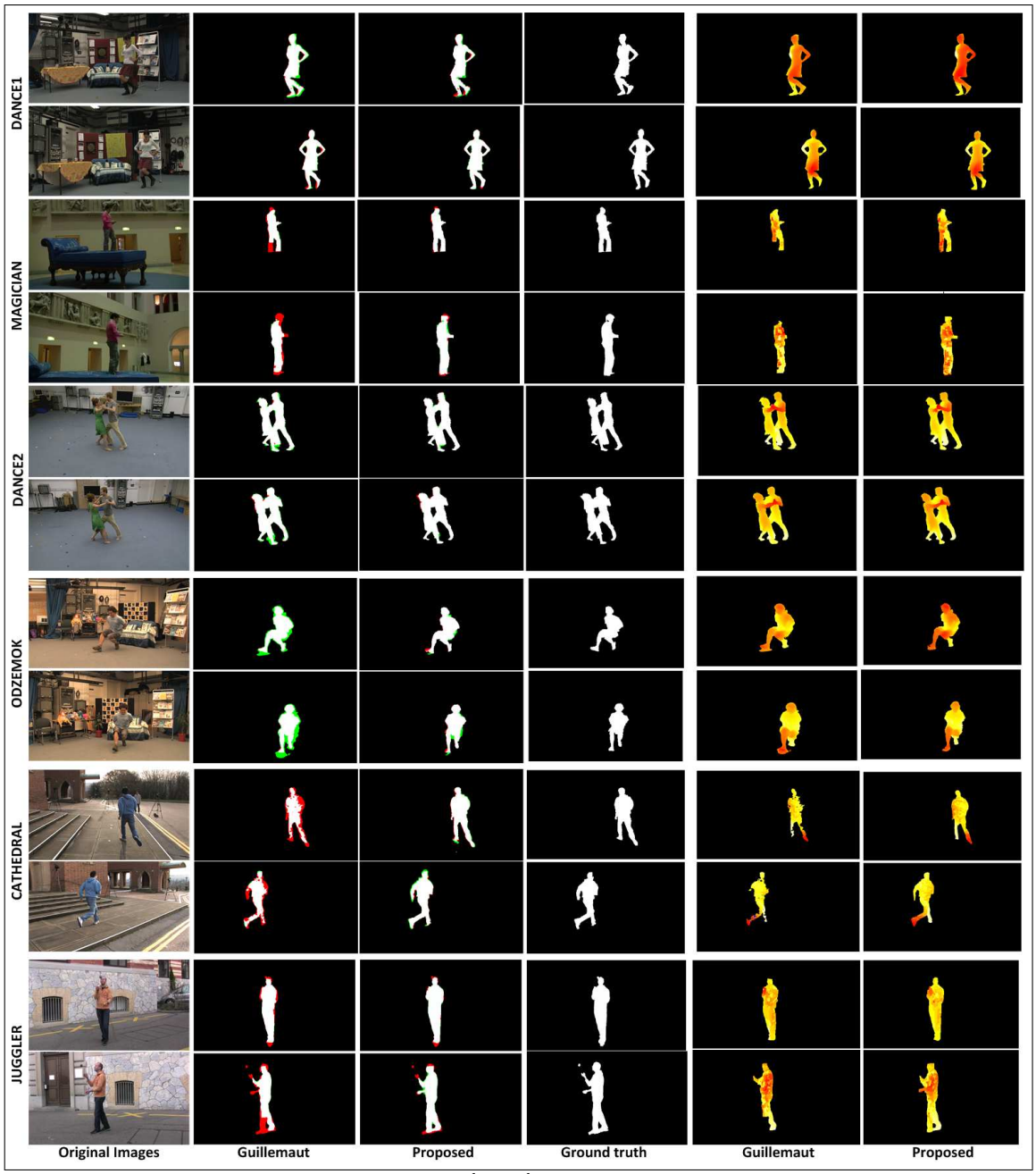

Figure 5. Results for a pair of images from each dataset: $2^{nd} - 4^{th}$ column: Segmentation (Red represents true negatives and green represents false positives compared to the ground truth) and $5^{th} - 6^{th}$ column: Depth

6.1. Segmentation results

The segmentation results from the proposed approach are compared against the segmentation from Guillemaut and Hilton [10] and the ground-truth. Ground truth is obtained by manually labelling the foreground for all datasets except Juggler and Magician where ground-truth is available online.

Guillemaut [10]: This approach requires an initial coarse foreground segmentation retrieved by differencing against a static background plate to obtain a visual hull required as a prior for reconstruction. In the proposed approach we do not assume a known background allowing the use of moving cameras. We modified the Guillemaut method by assigning the coefficient of the color term to be zero because we assume no prior knowledge of the background and we initialized this approach using our initial coarse reconstruction instead of the visual hull.

6.1.1 Qualitative results

The segmentation results for two frames from each dataset are shown in Figure 5. Guillemaut requires accurate vi-

Dataset	Number of Cameras	Number of frames	Image resolution	Baseline	λ_{data}	λ_{smooth}	$\lambda_{contrast}$
Dance1	8 (1 moving)	250	1920×1080	15 degrees	0.5	0.005	1.0
Magician	6 (all moving)	6900	960×544	40-55 degrees	0.6	0.01	3.0
Dance2	8 (all static)	125	1920×1080	45 degrees	0.5	0.005	1.0
Odzemok	8 (2 moving)	250	1920×1080	15 degrees	0.5	0.005	1.0
Cathedral	8 (all static)	143	1920×1080	45 degrees	0.6	0.01	5.0
Juggler	6 (all moving)	3500	960×544	25-35 degrees	0.6	0.01	5.0

Table 1. Characteristics and parameter settings for datasets

sual hull initialization, in this case the proposed coarse reconstruction is erroneous and far-away from the actual object boundaries as shown in Figure 3. This results in less accurate segmentation compared to the proposed approach which disambiguates the problem by improving the contrast and data terms in the energy formulation. The data term removes the regions with very low photo-consistency and the contrast term introduces affinity towards strong edges of foreground. The artefacts with respect to ground truth in the proposed approach are from shadow areas and occlusions.

6.1.2 Quantitative evaluation

To perform the quantitative evaluation of the segmentation we measured the $HitRatio$, $BkgRatio$ and $OverlapRatio$ as defined in [29] against the ground truth pixels. The three criterion are defined as follows:

$$HitRatio = |Result \bigcap GT| \,/\, |GT|$$
$$BkgRatio = |Result - GT| \,/\, |Result|$$
$$OverlapRatio = \left|Result \bigcap GT\right| / \left|Result \bigcup GT\right| \quad (8)$$

The results are shown in Table 2 for all the dataset. The comparison parameters are averaged over the entire sequence to ensure the accuracy of the result. Higher hit, overlap ratio and lower background ratio represents better segmentation. The $HitRatio$ is the ratio of true positive in the result with the ground truth. The $OverlapRatio$ is the ratio of true positives in the result with the sum of result and ground truth. The ratios for the proposed approach are higher than Guillemaut for all the datasets, generally much higher for more complex datasets like outdoor scenes or scenes captured with only handheld moving cameras. This demonstrates the robustness of the proposed approach to general dynamic scene segmentation compared to Guillemaut as seen in Figure 5. The $BkgRatio$ measures the proportion of result which actually belongs to background i.e. false positives in the segmentation. In case of Guillemaut this value is higher as compared to the proposed approach for most of the datasets. To conclude the segmentation obtained by the proposed approach vs. a state-of-the-art technique which assumes static cameras and a known background plate is better in quality with higher hit, overlap ratio and lower background ratio.

6.2. Reconstruction results

We have compared our results with Guillemaut (Section 6.1) and **Furukawa [8]**: This represents a state-of-the-art multi-view wide-baseline stereo approach. Furukawa [8] does not refine the segmentation but gives a 3D point cloud which is converted into a mesh using Poisson surface reconstruction. For fair comparisons all of the approaches are initialised with the same calibration and coarse reconstruction obtained using the method explained in Section 4.

6.2.1 Qualitative results

The depth maps for the proposed approach and Guillemaut are shown in Figure 5. The consistency of depth maps in

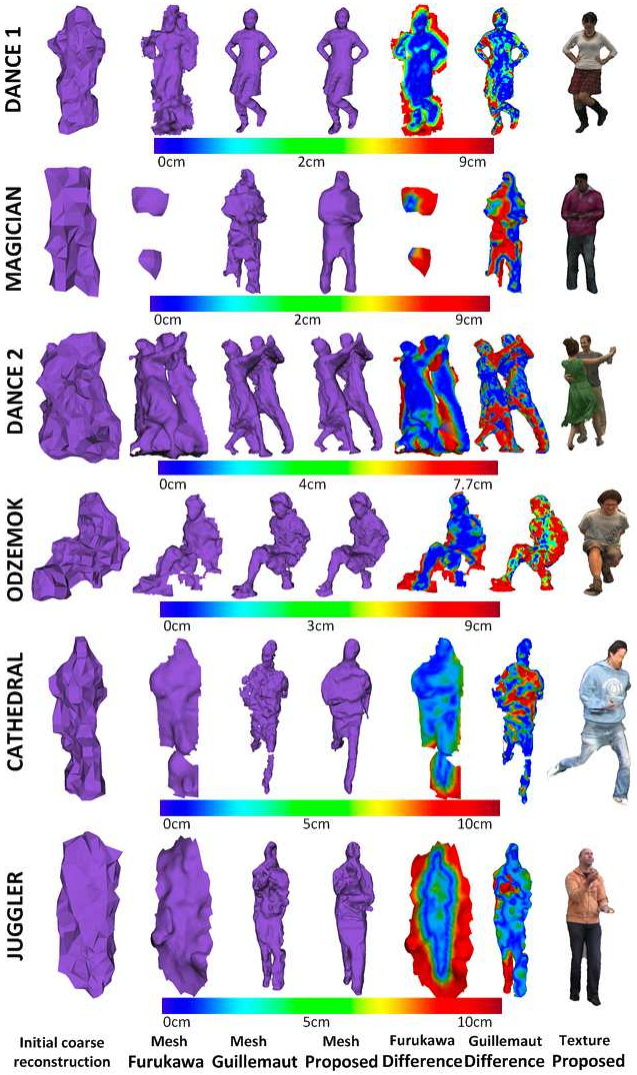

Figure 6. Results for each dataset: $1^{st} - 4^{th}$ column: Meshes and $5^{th} - 6^{th}$ column: Difference meshes against proposed approach with color coded error in cms and 7^{th} is textured mesh.

Criteria	Dance1		Magician		Dance2		Odzemok		Cathedral		Juggler	
	Ours	Guill.	Ours	Guill.	Ours	Guill.	Ours	Guill.	Ours	Guill.	Ours	Guill.
HitRatio	**0.995**	0.993	**0.887**	0.663	**0.994**	0.992	**0.899**	0.895	**0.891**	0.796	**0.879**	0.646
BkgRatio	**0.023**	0.042	0.022	**0.018**	**0.020**	0.031	**0.381**	0.507	0.021	**0.015**	**0.025**	0.038
Overlap	**0.947**	0.928	**0.855**	0.595	**0.963**	0.941	**0.611**	0.469	**0.849**	0.745	**0.841**	0.577

Table 2. Segmentation performance comparison for all datasets (best for each dataset is highlighted in bold) (Guill. depicts Guillemaut)

the case of the proposed approach are better because of the use of an improved data term for robustly matching between views and preserving edges.

The 3D models of the dynamic foreground obtained from the proposed approach are compared with Guillemaut and Furukawa in Figure 6 for all the datasets. For Magician dataset Furukawa gives very few points on a small part of the object in the reconstruction due to the complexity of the dataset. Results are compared closely with Guillemaut in Figure 7. In Figure 6 the meshes obtained by Furukawa do not have clear boundaries because it is not designed to refine the segmentation of the object. The meshes obtained from the proposed approach are visibly more accurate compared to the other techniques especially in the case of outdoor datasets. Some errors in the mesh reconstruction are present due to camera noise, uniform textures and similarity to the background. Results for Juggler sequence are shown in Figure 8 and more results are available in supplementary material and video.

6.2.2 Quantitative evaluation

Due to the absence of ground-truth 3D models for the datasets the accuracy evaluation is limited to the qualitative analysis. In this section we compare the computational efficiency of different approaches against the proposed method. The run-time per frame is shown in Table 3. The speed of the proposed approach is slightly lower than Furukawa (which does not perform segmentation) and the improvement in the speed for the proposed approach is approximately 25% as compared to Guillemaut.

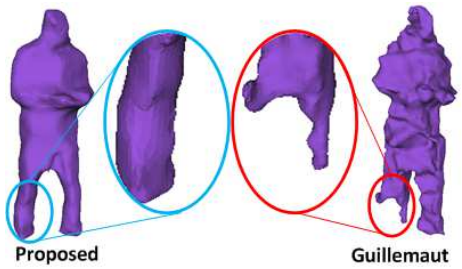

Figure 7. Result for magician dataset

Figure 8. Result for Juggler sequence: Original images from one view with frame numbers and mesh reconstructions alternatively

Dataset	Furukawa[8]	Guillemaut[10]	Proposed
Dance1	326 s	448 s	295 s
Magician	311 s	452 s	377 s
Dance2	502 s	655 s	471 s
Odzemok	381 s	498 s	364 s
Cathedral	525 s	679 s	501 s
Juggler	399 s	466 s	374 s

Table 3. Comparison of computational efficiency for all datasets (time in seconds (s))

6.3. Limitations and Future work

The proposed approach reconstructs and segments multiple close objects as a single dynamic object. This is not a failure case, but it increases the overall computational time of general scene reconstruction. Secondly, the proposed technique does not handle textureless scenes due to the sparcity of 3D points and crowded scenes due to the failure of the clustering algorithm used for initialisation. We aim to handle these scenes in future, by inclusion of full scene reconstruction from the sequence.

7. Conclusion

This paper introduced a novel technique to automatically segment and reconstruct dynamic objects 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 segment and reconstruction of multiple dynamic objects. The initial coarse approximation is refined using a a joint view-dependent optimisation of segmentation and reconstruction by a view-dependent graph-cut optimization using the photo-consistency and contrast cues from wide-baseline images.

Unlike previous method the proposed approach allows unsupervised reconstruction without prior information on scene appearance or structure. The segmentation and reconstruction accuracy are significantly improved over previous methods allows application to more general dynamic scenes. Tests on challenging datasets demonstrate improvements in quality of reconstruction and segmentation compared to state-of-the-art methods.

Acknowledgements

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

References

[1] 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. 1, 2, 5

[2] M. Bleyer, C. Rhemann, and C. Rother. Patchmatch stereo - stereo matching with slanted support windows. In *BMVC*, 2011. 4

[3] 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

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

[5] 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

[6] 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

[7] S. Fortune. Handbook of discrete and computational geometry. chapter Voronoi Diagrams and Delaunay Triangulations, pages 377–388. 1997. 4

[8] Y. Furukawa and J. Ponce. Accurate, dense, and robust multiview stereopsis. *PAMI*, 32:1362–1376, 2010. 5, 7, 8

[9] L. Guan, J. S. Franco, and M. Pollefeys. Multi-view occlusion reasoning for probabilistic silhouette-based dynamic scene reconstruction. *IJCV*, 90:283–303, 2010. 2

[10] J. Y. Guillemaut and A. Hilton. Joint Multi-Layer Segmentation and Reconstruction for Free-Viewpoint Video Applications. *IJCV*, 93:73–100, 2010. 1, 2, 5, 6, 8

[11] X. Hu and P. Mordohai. A quantitative evaluation of confidence measures for stereo vision. *PAMI*, 2012. 5

[12] 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. 3

[13] 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

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

[15] H. Kim, J. Guillemaut, T. Takai, M. Sarim, and A. Hilton. Outdoor Dynamic 3-D Scene Reconstruction. *CSVT*, 22:1611–1622, 2012. 1, 2

[16] V. Kolmogorov, A. Criminisi, A. Blake, G. Cross, and C. Rother. Probabilistic fusion of stereo with color and contrast for bilayer segmentation. *PAMI*, 28:2006, 2006. 4, 5

[17] A. Kundu, Y. Li, F. Dellaert, F. Li, and J. M. Rehg. Joint semantic segmentation and 3d reconstruction from monocular video. In *ECCV*, volume 8694, pages 703–718. 2014. 2

[18] 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

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

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

[21] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60:91–110, 2004. 3

[22] L. Matthies. Stereo vision for planetary rovers: Stochastic modeling to near real-time implementation. *IJCV*, 8:71–91, 1992. 5

[23] A. Mustafa, H. Kim, E. Imre, and A. Hilton. Initial disparity estimation using sparse matching for wide-baseline dense stereo. In *CVMP*, 2014. 4

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

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

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

[27] M. Sarim, A. Hilton, and J.-Y. Guillemaut. Temporal trimap propagation for video matting using inferential statistics. In *ICIP*, pages 1745–1748, 2011. 2

[28] S. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *CVPR*, 2006. 2, 5

[29] 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, 7

[30] R. Szeliski and P. Golland. Stereo matching with transparency and matting. In *ICCV*, pages 517–524, 1998. 2

[31] A. Taneja, L. Ballan, and M. Pollefeys. Modeling dynamic scenes recorded with freely moving cameras. In *ACCV*, pages 613–626. 2011. 1, 2

[32] M. F. Tappen and W. T. Freeman. Comparison of graph cuts with belief propagation for stereo, using identical mrf parameters. In *ICCV*, pages 900–908, 2003. 5

[33] C. Tomasi and R. Manduchi. Bilateral filtering for gray and color images. In *ICCV*, pages 839–846, 1998. 5

[34] T. Tung, S. Nobuhara, and T. Matsuyama. Complete multi-view reconstruction of dynamic scenes from probabilistic fusion of narrow and wide baseline stereo. In *ICCV*, pages 1709–1716, 2009. 2

[35] C. Wu. Towards linear-time incremental structure from motion. In *3DV*, pages 127–134, 2013. 2

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

[37] C. Zach, T. Pock, and H. Bischof. A duality based approach for realtime tv-l1 optical flow. In *In Ann. Symp. German Association Patt. Recogn*, pages 214–223, 2007. 4

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

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