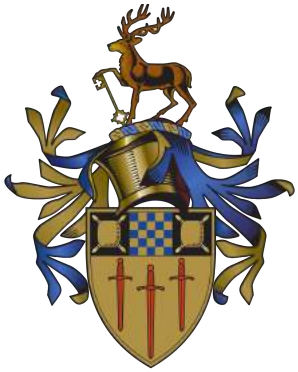
General 4D Dynamic Scene Reconstruction from Multiple View Video



Armin MUSTAFA

Centre for Vision, Speech and Signal Processing,
Department of Electronic Engineering

University of Surrey

This dissertation is submitted for the degree of

Doctor of Philosophy

December 2016

I would like to dedicate this thesis to my loving parents for their love and support …

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 65,000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 100 figures.

Armin MUSTAFA
December 2016

Acknowledgements

I would like to thank my principal supervisor Prof. Adrian Hilton for giving me an opportunity to work in 3D Reconstruction. His advice and support motivated me throughout my research and this work would not have been possible without his guidance. I would also like to thank my co-supervisor Dr. Hansung Kim for his inputs and encouragement throughout this study.

I express my gratitude to past and present members of CVSSP who have been a part of this journey. I would like to thank Dr. Jean-Yves Guillemaut and Dr. Evren Imre for their valuable ideas. I am most grateful to Volkan Kilic for being a great friend and providing encouragement and support when I needed the most. I am thankful to Marco Volino, Alexandros Neophytou, Premkumar Elangovan, Sheaka Alobaidli, Charles Malleson, Liz James and Anna Korzeniowska for their support and friendship over the years. My heartfelt thanks goes to my parents and my family for giving me strength and encouraging me at every step. Finally a special thanks to my husband Gowhar for all his endless love and motivation. This study would not have been possible without his support.

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

Abstract

This thesis addresses the problem of reconstructing complex real-world dynamic scenes without prior knowledge of the scene structure, dynamic objects or background. Previous approaches to 3D reconstruction of dynamic scenes either require a controlled studio set-up with chroma-key backgrounds or prior knowledge such as static background appearance or segmentation of the dynamic objects. This thesis presents a new approach which enables general dynamic scene reconstruction. This is achieved by initializing the reconstruction with sparse wide-baseline feature matches between views which avoids the requirement for prior knowledge of the background appearance or assumptions that the background is static. To achieve sparse reconstruction of dynamic objects a novel segmentation based feature detector SFD is introduced. SFD is shown to give an order of magnitude increase in the number and reliability of features detected. A coarse-to-fine approach is introduced for reconstruction of dense 3D models of dynamic scenes. This uses joint segmentation and shape refinement to achieve robust reconstruction of dynamic object such as people. The approach is evaluated across a wide-range of indoor and outdoor scenes.

The second major contribution of this research is to introduce temporal coherence into the reconstruction process. The dynamic scene is segmented into objects based on the initial sparse 3D feature reconstruction of the scene. Dense reconstruction is then performed for each object. For dynamic objects the reconstruction is propagated over time to provide a prior for the reconstruction at successive frames in the sequence. This is combined with the introduction of a geodesic star convexity constraint in the segmentation refinement to improve the segmentation of complex objects. Evaluation on general dynamic scene demonstrates significant improvement in both segmentation and reconstruction with temporal coherence reducing the ambiguity in the reconstruction of complex shape.

The final significant contribution of this research is the introduction of a complete framework for 4D temporally coherent shape reconstruction from one or more camera views. The 4D match tree is introduced as an intermediate representation for robust alignment of partial surface reconstructions across a complete sequence. SFD is used to achieve wide-timeframe matching of partial surface reconstructions between any pair of frames in the sequence. This allows the evaluation of a frame-to-frame shape similarity metric. A 4D match tree

is then reconstructed as the minimum spanning tree which represents the shortest path in shape similarity space for alignment across all frames in the sequence. The 4D match tree is applied to achieve robust 4D shape reconstruction of complex dynamic scenes. This is the first approach to demonstrate 4D reconstruction of general real-world dynamic scenes with non-rigid shape from video.

Email: a.mustafa@surrey.ac.uk

Table of contents

List of figures

List of tables

List of Notations and Symbols

Abbreviations

2D	Two-Dimensional
3D	Three-Dimensional
4D	Four-Dimensional
AKAZE	Accelerated KAZE
BRIEF	Binary Robust Independent Elementary Features
CV	Computer Vision
FAST	Features from Accelerated Segment Test
FVVR	Free Viewpoint Video Renderer
GSC	Geodesic star convexity
GMM	Gaussian mixture model
ICR	Initial coarse reconstruction
MSER	Maximally Stable Extremal Regions
MRF	Markov Random Field
MST	Minimum Spanning Tree
MVS	Multi-view Stereo
NRSFM	Non-rigid structure from motion
NSA	Non-sequential alignment
ORB	Oriented FAST and Rotated BRIEF
SFD	Segmentation based Feature Detector
SFM	Structure from Motion
SIFT	Scale Invariant Feature Transform
SOE	Silhouette overlap error
SURF	Speeded Up Robust Features
VH	Visual Hull

Typesetting

Scalar Values	Lower-case letters in italic (e.g. *x,y*)
Vector Values	Lower-case letters with arrow above elements (e.g. $\vec{n}$, $\vec{v}$)
Matrices	Upper-case letters in bold (e.g. **K,R**)
Functions	Upper-case letters with brackets for variable inputs (e.g. A(x), F(a,b))
Sequences	Lower-case letter enclosed with curly brackets over a defined range (e.g. $\{c(i)\}N_i = 1$)
Sets	Upper-case letters in italic (e.g. *F,M*)
Total number of elements	Upper-case N with subscript identifier (e.g. N_T, N_C)
Elements from the graph theory	Upper-case in calligraphic style (e.g. $\mathscr{G}$, $\mathscr{V}$, $\mathscr{E}$)

Symbols

Chapter 3

N	Neighbourhood of a feature
F	Set of initial features on an image
G	Image gradient for each feature
F^*	Set of features detected on an image
N_F	Total number of features
$T(.)$	Cost function for sub-pixel refinement

Chapter 4

R_I	Interior region for initial coarse reconstruction
R_O	Outer region for initial coarse reconstruction
N_C	Number of photo-consistent camera pairs with reference camera
N_P	Neighbourhood of a pixel for reconstruction and segmentation
N_I	Number of interacting pixels in N_P
$\mathscr{D}_I$	Set of depth labels for inner region
$\mathscr{D}_O$	Set of depth labels for outer region
d_i	Different depth values as a part of $\mathscr{D}$
$E(.)$	Cost function for joint refinement
$M(.,.)$	Data cost assignment
$NCC(.,.)$	Normalized correlation coefficient
$B(.)$	Bilateral filter
$J(.,.)$	Squared Euclidean color distance

Chapter 5

$\mathbf{A}$	Centroid of sparse point cloud
$\vec{ev}$	Eigen vectors of point cloud
$\mathbf{R}$	Rotation matrix for PCA
$H_{.,.}(.)$	Disprity between two pixels for a single view
$u_{.,.}(.)$	Temporal correspondence between two frames for a single view
R	Region for initial coarse reconstruction
O	Frobenius norm
$\mathscr{D}$	Set of depth labels for initial coarse reconstruction
$\mathscr{L}$	Set of labels for joint refinement
$\mathscr{G}$	Graph cut for joint refinement
$\mathscr{V}$	Vertices for graph structure
$\mathscr{E}$	Edges for graph structure
C	Set of star centers
N_T	Number of star centers
N_D	Number of pixels for discrete paths
N_C	Number of photo-consistent camera pairs with reference camera
N_P	Neighbourhood of a pixel for reconstruction and segmentation
$E(.,.)$	Cost function for joint refinement
$M(.,.)$	Data cost assignment
$P(.,.)$	Probability at a pixel to compute color term
$N(.,.)$	Normal distribution to compute color term

Chapter 6

$q(.)$	Frames of a sequence
$v(.)$	Views of a frame
N_Q	Number of frames in a sequence
N_V	Number of views per frame
X^C	Initial set of keypoints for all frames at view C
$S^C_{.,.}$	Feature correspondence at view C
$\mathbf{D}$	Dissimilarity matrix for non-sequential alignment
$K^C_{.,.}$	Feature match metric for view C
$R^C_{.,.}$	Total number of initial matches at view C
$I^C_{.,.}$	Silhouette match metric for view C
$A^C_{.,.}$	Area under the silhouette at view C
$h^C_{.,.}$	Aligned silhouette intersection at view C

$\mathscr{T}$	Traversal tree for Non-sequential alignment
$\mathscr{N}$	Set of nodes in traversal tree
$\mathscr{P}$	Set of edges in traversal tree
$\mathscr{P}$	Optimal tree (MST)
Ω	Graph of all possible frame-to-frame alignment pairs

Chapter 1

Introduction

1.1 Background and Motivation

We live in a 3D world and directly perceive 3D information. Inspired by this, a fundamental problem of computer vision is to obtain 3D geometric information with the help of computers and digital sensing devices. This process is called '3D reconstruction'. Over the past two decades there has been extensive research in visual geometry and static 3D scene reconstruction. Due to the rapid advances in technology as well as decreasing cost of computing and sensing hardware, the focus of 3D reconstruction has gradually shifted from 3D static structure reconstruction to dynamic scene modeling. A 'dynamic scene' is a scene containing one or more moving objects to be modeled, possibly with shape deformation over time. Existing systems for dynamic 3D reconstruction from multiple view video use controlled indoor environments with uniform illumination and backgrounds to allow accurate segmentation of the dynamic foreground objects. The general case of outdoor dynamic 3D scene reconstruction with uncontrolled illumination and backgrounds remains an open challenge. Reconstruction of general dynamic scenes is motivated by potential applications in film and broadcast production, assisted living, surveillance and smart cities/workplaces.

Recent decades have seen increasing use of visual effects in film, television and video game production. 3D static scene reconstruction from images and video is widely used in media production to allow addition of graphical elements to generate a photo-realistic experience for the audience. The line between real images and computer graphics has been blurred by the introduction of hardware and robust algorithms for static scene reconstruction. Reconstruction of dynamic scenes such as people in controlled studio environments has been shown to allow photo-realistic rendering from the reconstructed 3D models. However, content production requires the ability to exit and manipulate the content as part of the creative process. The goal is to reconstruct dynamic scene representations which can be

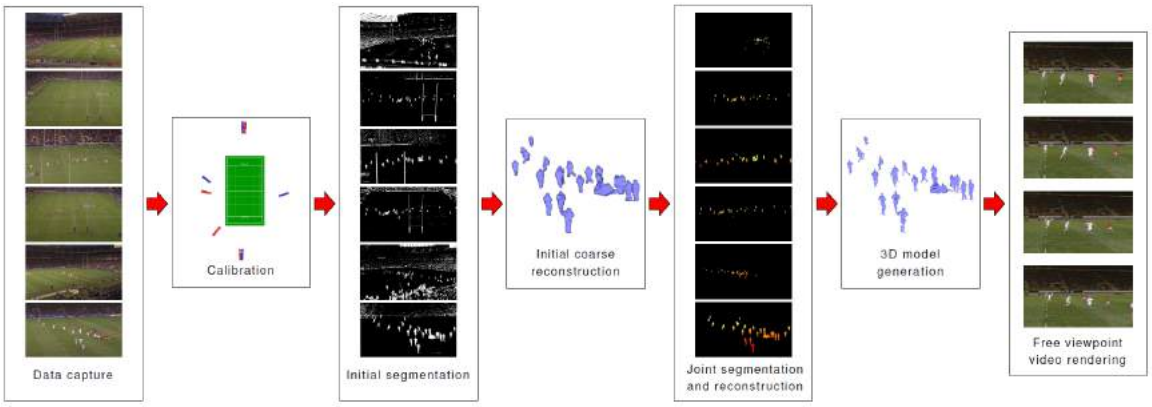


Fig. 1.1 Pipeline for dynamic scene reconstruction in [62]

rendered with the photo-realism of video and manipulated or edited with the flexibility of computer generated models. To achieve this requires temporally coherent 4D reconstruction. Methods for temporally coherent 4D reconstruction have recently been introduced but are restricted to controlled scenes with known backgrounds or they on work on water-tight meshes of dynamic objects with full visibility. This thesis addresses the general problem of reconstruction of temporally coherent (4D) representations of general dynamic scenes from multiple camera views without prior knowledge of the scene structure. 4D reconstruction is defined as a temporally consistent representation of the 3D points of entire scene for the whole sequence over time. The per frame 3D points of the dynamic parts of the object are connected in time for 4D scene reconstruction.

Previous work in dynamic outdoor scene reconstruction from multi-view images for sports data was proposed by [62]. A joint multi-layer reconstruction and segmentation approach was introduced to obtain per frame reconstruction of the scene using the framework shown in Figure 1.1. Initial coarse reconstruction was retrieved using visual hull from the segmentation of dynamic objects which was refined using a joint depth and segmentation optimization framework. Another recent work in dynamic outdoor scene reconstruction which targets film and broadcast applications for outdoor dynamic datasets was proposed by [82]. The algorithm is divided in three main stages: Environment modelling; Dynamic foreground scene reconstruction; and Model composition and rendering. The pipeline is shown in Figure 1.2. These existing dynamic 3D reconstruction techniques suffer from the following problems:

- These approaches require accurate segmentation of dynamic foreground objects. Automatic segmentation techniques fail for complex scenes with dynamic backgrounds, changing illumination and similar foreground/background appearance and manual segmentation is a tedious task.

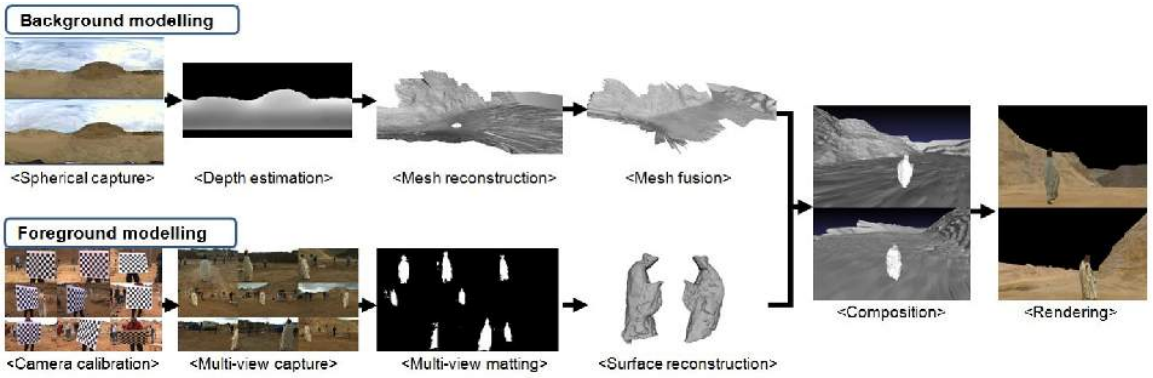


Fig. 1.2 Pipeline for dynamic scene reconstruction in [82]

- The assumptions of prior knowledge of the scene structure, background and segmentation limit previous approaches to static cameras and controlled environments with static background and these approaches do not work with multi-view datasets capture with a network of moving cameras.
- Static background and dynamic foreground are processed separately requiring different capture technologies and offline processing to estimate the background model. The method in [82] reconstructs dynamic scene elements from multi-camera views along with spherical capture for static scene/background to obtain full scene reconstruction.
- Semi-automatic operation: Tasks like segmentation and registering reconstructions from multiple methods may require manual intervention making the whole process of scene reconstruction slow.
- These methods give temporally incoherent mesh geometries as an output which makes it difficult to apply these methods in real-world applications.

The main goal of the research is to is to overcome the limitations of previous approaches and enable 4D reconstruction of general real-world dynamic scenes.

1.2 Problem Statement

Over the past decades, effective approaches have been proposed to recover 3D dynamic shapes in indoor laboratory environments with controlled lighting, uniform-colored static backgrounds and empty viewing space without any visual occluders [19]. However, such systems cannot be directly extended to an uncontrolled natural environment. A natural scene is far more complicated with environmental variations, including illumination changes,

soft shadows cast by clouds, dark shadows cast by trees or dynamic shapes themselves, visual occluders that are common in outdoor scenes, varying dynamic backgrounds with clutter such as trees or people, and specular reflections on glassy or metallic surfaces. Existing approaches in multi-view reconstruction require strong initial prior like knowledge of background, structure information, assumption of static cameras or foreground segmentation to obtain reconstruction of dynamic objects for these natural scenes which limits their applicability. The proposed approach to be developed in this research aims to overcome the limitations of existing approaches through unsupervised complete dense scene reconstruction of complex dynamic scenes captured from a network of static or moving cameras. Also the existing systems for dynamic reconstruction from general outdoor scenes either produce per frame reconstruction of the dynamic objects or give sequential temporally consistent geometries which fail in case of large motion. Our aim is to obtain temporally coherent (4D) dynamic scene reconstruction automatically for partial non-rigid geometries with large and articulated motion of general scenes from a single or multiple camera acquisition system. The requirements of the proposed approach are:

- The input to our system is synchronized multi-view videos captured using a network of multiple static or moving cameras.
- Currently methods for dense scene reconstruction of dynamic scenes requires extra knowledge like segmentation, depth data or background knowledge. We assume no prior knowledge of the foreground, background, scene structure or appearance.
- The methods proposed in the literature for 4D scene reconstruction assume rigid objects or limited motion. Our goal is to obtain temporally coherent reconstruction for articulated objects with large non-rigid motion and partial surface geometries.

1.3 Methodology

Previously some approaches to dynamic scene reconstruction have achieved the solution through joint segmentation and reconstruction of the scene. It has been shown that segmentation and reconstruction are related problems which require simultaneous solution and produce reliable results [62]. However, existing joint segmentation and reconstruction approaches use the foreground segmentation to obtain visual hull and background information as prior to obtain the final results [62, 63]. The visual hull is used as a prior for the reconstruction of the foreground which is refined using an joint optimization framework. Our aim is to perform segmentation and reconstruction of the general dynamic scene without any prior on

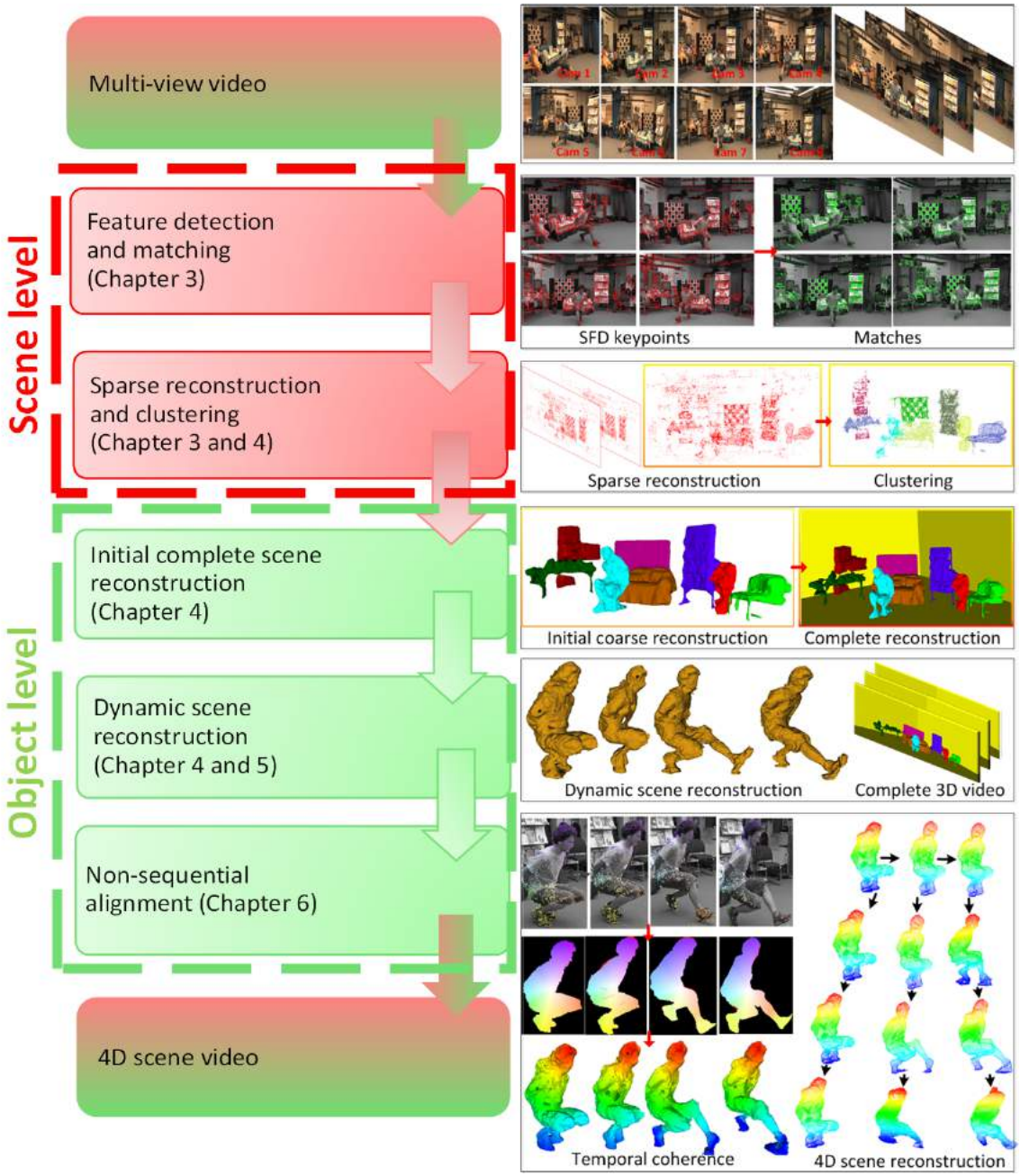


Fig. 1.3 Proposed pipeline for 4D general scene reconstruction

the scene structure, background or segmentation and to use this concept in our framework we need to initialize the geometry of objects in the scene automatically. To initialize the joint optimization of the general dynamic scenes automatically without any prior unlike existing approaches we use the information from the sparse reconstruction of the scene. The proposed framework is shown in Figure 1.3 with two stages of processing: Scene level which is performed on the entire scene depicted in dashed red line and Object level is performed on each object of the scene depicted in dashed green line. The separation of the processing at scene and object level is to improve the efficiency of the system. Reconstructing the entire scene in a single minimization is computationally expensive and may not lead to an efficient solution. Separating the scenes in different objects allows us to retain the reconstruction of static parts of the scene over-time and reconstruct only dynamic parts of the scene for consecutive time instants. The stages of the proposed system, illustrated in the figure are explained below:

- **Multi-view videos:** We consider synchronized multi-view videos as input to our system captured from a network of static or moving hand-held cameras. Camera intrinsic calibration is assumed to be known.

- **Feature detection and matching:** Sparse 2D feature correspondences are obtained for the entire scene between multi-view image pairs at each frame.

- **Sparse reconstruction and clustering:** Camera extrinsics and sparse 3D points are obtained from the feature correspondences. The sparse reconstruction of the scene represents different objects in foreground and some information of the background. To identify various objects in the scene we cluster the sparse 3D points, with each cluster representing an foreground object. A reliable and uniform initial estimate of sparse 3D structure is required for salient object identification through clustering which depends upon the accuracy and number of correspondences obtained from multi-view image pairs. Existing feature detection approaches result in a highly sparse non-uniform distribution of scene features and the resulting feature set often results in poor scene coverage. We introduced a feature detection suitable for wide-baseline matching to obtain a good estimate with uniform scene coverage in the initial sparse reconstruction of the scene. This process is performed at scene level for the entire sequence as shown in the Figure 1.3.

- **Initial coarse reconstruction and refinement:** Once we have clusters of objects, for each cluster an initial coarse reconstruction is obtained for all the objects in the scene for first frame of the sequence. The initialization of segmentation and reconstruction is

a rough estimate prone to errors. Application of existing joint refinement techniques on such inaccurate estimate produces unreliable results. Hence the existing state-of-the-art methods cannot be used directly and there is a need to improve the optimization framework to obtain an accurate estimate of the objects in the scene and make it robust to errors in the initialization. The initial reconstruction of the scene is refined using a joint segmentation and reconstruction refinement framework to obtain complete scene reconstruction for the first frame of the sequence and a rough proxy of the background is added.

- **Dynamic scene reconstruction and refinement:** For consecutive time instants, only dynamic objects in the scene are identified and reconstructed. The initial coarse reconstruction is obtained by integrating the information from the mesh in the previous frame and the sparse cluster of 3D points at current frame. The initial coarse reconstruction of the dynamic scene is refined using a joint optimization of object segmentation and reconstruction and is combined with the static scene reconstruction at that frame to obtain complete scene reconstruction for the entire sequence. We also introduce temporal coherence and shape information in the framework to achieve reliable results. The static parts of the scene are retained over time and are fused with dynamic elements to obtain the final 4D scene reconstruction for the entire dynamic sequence.

- **Non-sequential alignment for 4D scene reconstruction:** Applications of reconstruction to film, broadcast and gaming requires temporally consistent geometries. State-of-the-art approaches in dynamic scene reconstruction either produce per frame reconstruction, frame-to-frame or sequential alignment (errors due to drift and large motion) or work for water-tight dynamic geometries. To overcome the limitations of existing methods, temporal coherence is introduced in our framework by using a global non-rigid non-sequential alignment framework based on the movement and shape of the dynamic objects in the scene.

1.3.1 Contributions of Research

Our contributions include:

- A novel segmentation based feature detector SFD is introduced in Chapter 3 for accurate wide-baseline matching which gives an increased number of good and repeatable feature detection for different viewpoints, accurate feature localization and improved coverage for natural scenes.

  - A comprehensive performance evaluation for wide-baseline matching on benchmark datasets of existing feature detectors (Harris, SIFT, SURF, FAST, MSER, ORB, AKAZE) and descriptors(SIFT, BRIEF, ORB, SURF) is presented showing improved performance of the SFD detector in terms of both number of features and matching accuracy.
  - Application to scene reconstruction demonstrates an order of magnitude increase in the number of reconstructed points with improved scene coverage and reduced error compared to previous detectors against ground-truth.

- An unsupervised dense reconstruction and segmentation of general dynamic scenes from multiple wide-baseline views is proposed in Chapter 4.
  - Automatic initialization of dynamic object initial segmentation and reconstruction from sparse features is introduced.
  - Robust joint refinement of dense reconstruction and segmentation integrating error tolerant photo-consistency and edge information is proposed.

- Robust spatio-temporal reconstruction and segmentation of dynamic scenes by exploiting temporal coherence and shape information is shown to improve the quality of the results in Chapter 5.
  - A framework for space-time sparse-to-dense reconstruction is introduced to create a good estimate of initial coarse reconstruction. The initial dense reconstruction is jointly optimized using geodesic star convexity constraint along with addition of color information in the minimization framework.

- Robust global 4D alignment of partial reconstructions of non-rigid shape from multi-view sequences with moving cameras is introduced in Chapter 6.
  - 4D Match Trees are introduced to represent the optimal non-sequential alignment path which minimizes change in the observed shape.
  - Sparse matching between wide-timeframe image pairs of non-rigid shape using SFD is proposed. A novel method to obtain dense 4D surface correspondence using optical flow guided by sparse SFD temporal matching is introduced.

1.3.2 Thesis Outline

Chapter 1:Introduction
This chapter presents the background of the research problem, motivation, problem statement,

methodology and contributions of the research. A list of publications resulting from this research is also given.

Chapter 2: Literature Survey and Background
Chapter 2 presents a review of reconstruction techniques in static and dynamic scene. General 4D scene reconstruction framework is introduced and an overview of state-of-the-art techniques at each stage is presented: Data capture, Feature detection and matching, Dense reconstruction and Temporal coherence. A description of datasets used in this thesis for evaluation purposes is also included.

Chapter 3: SFD: Segmentation based Features for Wide-baseline Reconstruction
In this chapter we introduce a novel feature detector SFD that produces an increased number of 'good' features for accurate wide-baseline reconstruction. This work was published at 3DV, 2015 [126]. Each image is segmented into regions by over-segmentation and feature points are detected at the intersection of the boundaries for three or more regions. Segmentation-based feature detection locates features at local maxima giving a relatively large number of feature points which are consistently detected across wide-baseline views and accurately localized. A comprehensive comparative performance evaluation with previous feature detection approaches demonstrates that: SFD produces a large number of features with increased scene coverage; detected features are consistent across wide-baseline views for images of a variety of indoor and outdoor scenes; and the number of wide-baseline matches is increased by an order of magnitude compared to alternative detector-descriptor combinations. Sparse scene reconstruction from multiple wide-baseline stereo views using the SFD feature detector demonstrates at least a factor six increase in the number of reconstructed points with reduced error distribution compared to SIFT when evaluated against ground-truth and similar computational cost to SURF/FAST.

Chapter 4: Dense Reconstruction of Real-world Dynamic Scenes
This chapter 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 using SFD features introduced in the previous chapter. The primary contributions of this chapter 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 multi-view segmentation and dense reconstruction. The proposed approach also eliminates the requirement for prior knowledge of scene structure and appearance. This work was presented at ICCV, 2015 [122] and initial work in CVMP, 2014 [125].

Chapter 5: Temporally Coherent Dynamic Scene Reconstruction

Chapter 5 presents an approach for reconstruction of temporally coherent models of complex dynamic scenes. The method was published at CVPR, 2016 [123] as oral presentation. Sparse-to-dense temporal correspondence is integrated with joint multi-view segmentation and reconstruction to obtain a complete 4D representation of static and dynamic objects. Temporal coherence is exploited to overcome visual ambiguities resulting in improved reconstruction of complex scenes. Robust joint segmentation and reconstruction of dynamic objects is achieved by introducing a geodesic star convexity constraint. Comparative evaluation is performed on a variety of unstructured indoor and outdoor dynamic scenes with hand-held cameras and multiple people. This demonstrates reconstruction of complete temporally coherent scene models with improved non-rigid object segmentation and shape reconstruction.

Chapter 6: 4D Match Trees for Non-rigid Surface Alignment

This chapter presents a method for dense 4D temporal alignment of partial reconstructions of non-rigid surfaces observed from single or multiple moving cameras of complex scenes. *4D Match Trees* are introduced for robust global alignment of non-rigid shape based on the similarity between images across sequences and views. Wide-timeframe sparse correspondence between arbitrary pairs of images is established using SFD. Sparse SFD correspondence allows the similarity between any pair of image frames to be estimated for moving cameras and multiple views. This enables the 4D Match Tree to be constructed which minimizes the observed change in non-rigid shape for global alignment across all images. Dense 4D temporal correspondence across all frames is then estimated by traversing the 4D Match tree using optical flow initialized from the sparse feature matches. The approach is evaluated on single and multi-view images sequences for alignment of partial surface reconstructions of dynamic objects in complex indoor and outdoor scenes to obtain a temporally consistent 4D representation. Comparison to previous 2D and 3D scene flow demonstrates that 4D Match Trees achieve reduced errors due to drift and improved robustness to large non-rigid deformations. This work is to be presented at ECCV, 2016 [124].

Chapter 7: Conclusion and Future Work
Chapter 7 concludes the work presented throughout this thesis and suggests potential future research directions.

1.3.3 Publications

The following publications are a result of the work described in this thesis:

- A. Mustafa, H. Kim and A. Hilton; **4D Match Trees for Non-rigid Surface Alignment**; *European Conference on Computer Vision (ECCV) 2016*; Poster Presentation.
- A. Mustafa, H. Kim, J.-Y. Guillemaut and A. Hilton; **Temporally coherent 4D reconstruction of complex dynamic scenes**; *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) 2016*; Oral presentation.
- A. Mustafa, H. Kim, J.-Y. Guillemaut and A. Hilton; **General Dynamic Scene Reconstruction from Multiple View Video**; *International Conference on Computer Vision (ICCV) 2015*; Poster Presentation.
- A. Mustafa, H. Kim, E. Imre and A. Hilton; **Segmentation based Features for Wide-baseline Multi-view Reconstruction**; *3D Vision (3DV) 2015*; Oral Presentation.
- A. Mustafa, H. Kim, E. Imre and A. Hilton; **Initial Disparity Estimation using Sparse Matching for Wide-baseline Dense Stereo**; *European Conference on Visual Media Production (CVMP) 2014*; Short Paper.
- A. Mustafa, H. Kim, J.-Y. Guillemaut and A. Hilton; **Temporally coherent general dynamic scene reconstruction**; *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 2016*; To be submitted.
- A. Mustafa, H. Kim and A. Hilton; **Segmentation based Features for Wide-baseline Matching**; *Pattern Recognition (PR) 2016*; To be submitted.

Chapter 2

Literature Survey and Background

The recovery of 3D structure information from image sequences is a fundamental problem in computer vision. Reconstruction of general dynamic scenes is motivated by potential applications in gaming, film, broadcast production, surveillance etc. The ultimate goal is to estimate 3D geometry of real-world scenes captured from distributed static or moving camera networks. Over the past three decades numerous approaches have been proposed and evaluated for multi-view reconstruction of static scenes [156]. State of the art 3D reconstruction systems use different methods like structure-from-motion (SFM), multi-view stereo, depth maps etc. to generate 3D models of scenes. The estimated shape is commonly associated with static structures where the scene is rigid. Our research focuses on the more difficult problem of 3D recovery when the object in the video is non-rigid, that is, its shape can change through time by deforming or articulating as many objects in the real world, including people, animals, elastic objects and clothing. However, reconstructing the shape of such objects from imagery remains an open problem. Approaches have been introduced to obtain 3D models from single or multiple cameras in the scene based on some prior knowledge of the object structure or deformation. In this work we focus on the videos captured with multiple static or moving cameras and we seek 3D models to express the time-varying shape of the objects in the image along with an approximate background reconstruction with no prior on the background, object structure or appearance. This chapter discusses the literature in the field of 4D multi-view reconstruction from image sequences focusing on the case of dynamic scene reconstruction.

2.1 Multi-view Scene Reconstruction

There is an increasing demand of 3D reconstruction in movie industry, gaming and virtual environments. Existing methods for these application use multi-view cameras to obtain

3D photo-realistic models of the moving objects in the scenes. Numerous approaches have been proposed depending on various factors like camera projection model, the type of correspondences between the images (lines, corners, multi-scale gradient histograms) and how they are matched (windowed search, tracking, optical flow). These methods are also dependent on the unknowns in the motion model (calibrated, semi-calibrated, uncalibrated internal parameters, restricted motions, wide or short baseline) and the unknowns in the structure (completely free, planar, piecewise planar scenes). Also, the type of the optimization algorithms (linear, non-linear), timing constraints (batch, online, real-time), and the strategy to handle robustness and degeneracy issues (M-estimators, RANSAC, planar degeneracies, probabilistic tracking) define the decision of approaches. In addition to all of the above, a 3D modelling phase must be implemented where photo-consistency, smoothness factors, occlusions, and specularities are simultaneously taken into account in the ideal case. All the aforementioned concepts have been extensively studied in the case of static scenes where multiple views can be captured with a single moving camera referred to as 'structure-from-motion' or multi-camera set-ups using 'multi-view stereo'. Not surprisingly, such issues are also relevant for dynamic scenes and a survey was presented in [156]. The input data for these reconstruction systems is captured using static or moving multi-view cameras, which may or may not be synchronized or calibrated, an example is shown in Figure 2.1. The deformable objects in the scene are observed at different angles and reconstructed independently using information from all video streams. A general solution to multi-view stereo was explained in [69]. Some approaches perform surface tracking on the per frame reconstructions to establish temporal correspondences in cases where a complete view of the objects is visible. Often, these multi-view and temporal correspondence methods are data-driven and makes use of additional strong prior to obtain a solution. General review on existing techniques in static and dynamic reconstruction is presented in [106]. The multi-view reconstruction techniques for dynamic scenes can be broadly classified into four approaches:

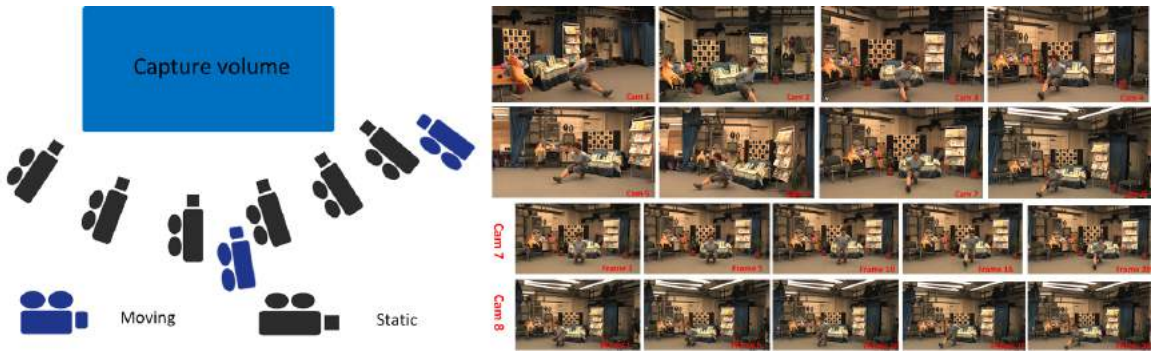


Fig. 2.1 Multi-view capture volume and a frame for Odzemok dataset

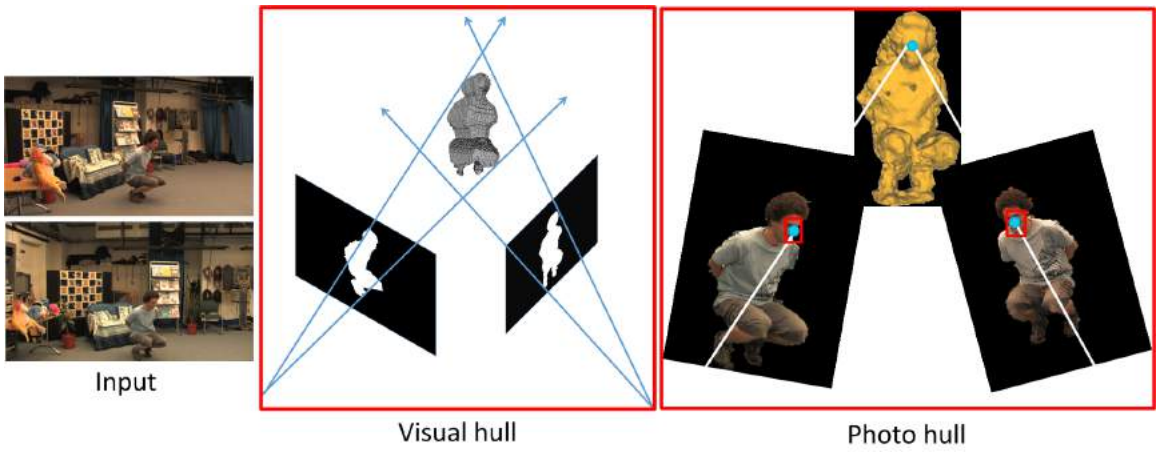


Fig. 2.2 Visual hull and Photo-hull from pair of input images of Odzemok dataset

1. Visual-hull: These require a foreground segmentation of the dynamic object of interest from the background. Reconstruction can be performed using a discrete voxel grid [59] or exact polyhedral visual hull[49]. An example of reconstruction obtained using visual hull is shown in Figure 2.2.

2. Photo-hull: Reconstructs the complete scene based on pixel photo-consistency between views [93, 186]. In practice such approaches require distinct color contrast between foreground and background limiting their use on natural scenes. An example of reconstruction obtained using photo-consistency is shown in Figure 2.2.

3. Multiple-view stereo: Stereo matching requires either narrow-baseline between camera views [156] or prior initialization which can be based on visual hull [165] which limits practical application to natural scenes. A sub-class of multi-view stereo based on oriented patch matching approach was introduced [51] which requires a large number of cameras to produce reliable reconstruction of the entire scene.

4. Multiple view depth maps: Active depth sensors can be used to acquire 2.5D depth maps from multiple viewpoints which are then fused into a single surface model [50, 114], but these methods focus on architectural scenes.

Research on multi-view dynamic scene reconstruction over the past two decades has primarily focused on indoor scenes with controlled illumination and backgrounds. Recent research has extended this work to outdoor scenes including sports and general environments [63, 77, 82, 170]. Current approaches to outdoor dynamic scene reconstruction exploit strong prior assumptions to enable segmentation of the foreground objects such as pitch color in sports or background scene geometry and appearance for general scenes. This research aims to

overcome the limitations of these assumptions enabling robust multi-view reconstruction of general dynamic scenes without prior assumptions on scene geometry or appearance and an example for dynamic dataset is shown in Figure 2.3. The research aims to develop a solution which addresses the following problems:

1. Uncalibrated multi-view scene acquisition from static or moving camera networks.
2. Challenging outdoor scenes with non-uniform dynamic and repetitive backgrounds.
3. Uncontrolled changing scene illumination.
4. Automatic dense reconstruction of the foreground and background.
5. Reconstruction to output a model with temporally consistent structure and known correspondences.

This will enable practical reconstruction of general scenes with commodity hardware. Potential applications include entertainment production, surveillance, biomechanics, animal welfare monitoring and dynamic scene understanding. It is useful to divide scene components into two categories: dynamic objects such as people, bikes, cars, birds or street vendors that move about and likely to have the same appearance a single time instant taken at a particular time and static backgrounds such as buildings, streets, landscaping, or benches that are visible in many images taken in the same general location and only change slowly in appearance due to scene illumination, and non-lambertian reflectance.

The remainder of this section reviews the development of multi-view reconstruction methods for static and dynamic scenes. This review focuses on recent advances moving towards reconstruction of dynamic scenes in uncontrolled environments. Finally a typical 4D dynamic scene reconstruction pipeline is outlined.

2.1.1 Static Scene Reconstruction

Multiple views of a scene can be used to obtain 3D reconstructions with appropriate assumptions like camera calibration information. Structure can be estimated from a moving camera or from a multi-camera system at a single time instant. Correspondences obtained from multi-views are used to recover a geometric model as well as a model of the camera locations. Different application have various requirements on the reconstruction. A few applications like navigation and robot localization require estimating sparse 3D points in the scene while other may require a per-pixel geometry estimate along with a dense surface reconstruction. For static (rigid) backgrounds, a classic approach to scene understanding is

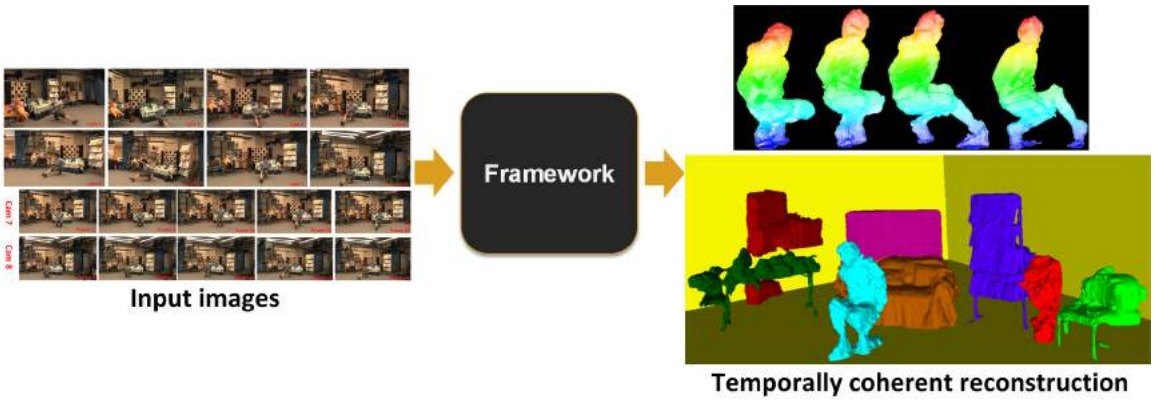


Fig. 2.3 Illustration of the required 4D scene reconstruction from dynamic scene Odzemok dataset as input.

to use structure-from-motion(SFM) and multi-view stereo (MVS) techniques to build up an explicit model of the scene geometry and appearance. There has been extensive research into accurate outdoor scene reconstruction from multi-view images [7, 35, 139] a quantitative evaluation against ground-truth is presented in [167].

Methods in static scene reconstruction are now well developed and work robustly on large unstructured photo collections ranging from point based reconstructions [51, 103] to line based SFM [133]. Patch-based MVS [51] generated point based reconstruction from a range of images by using standard SFM. Recently an alternative faster approach of PMVS has been introduced in [103]. A full 3D model of streets was reconstructed from 150,000 photos on Internet using grid computing with 500 cores for 24 hours in [7]. For accurate calibration of multiple cameras, it is always difficult to accurately calibrate cameras in advance. To solve this problem bundle adjustment introduced in [26, 177] was used for self calibration system in [162] and is widely adopted in state-of-the art 3D estimation techniques for calibration.

An accurate reconstruction algorithm for urban or archaeological sites was introduced in [9], including both geometry and texture, in order to obtain models useful for visualization, quantitative analysis in the form of measurements at large or small scales and potentially for studying their evolution through time. Instead of using fixed multiple cameras systems, SFM can also use video sequences from a moving camera [35, 139, 140]. The basic idea of the SFM is similar to multi-view reconstruction but it reconstructs 3D positions and the cameras motions simultaneously by feature correspondence. Uncalibrated 3D structure recovery is formulated as the joint estimation of the 3D position of the points in space and the pose and internal parameters of the camera. A variety of SFM strategies have been proposed including incremental [7, 190], hierarchical [56], and global approaches [94]. These systems still suffer from robustness, accuracy, completeness, and scalability problems which were

handled by state-of-the-art improved version of SFM in [155]. They propose an incremental SFM method which follows a sequential processing pipeline with an iterative reconstruction component. It starts with feature extraction and matching, followed by geometric verification to obtain the sparse 3D points.

In [130] a single hand-held RGB camera captures a static scene performing reconstruction in real-time using GPU hardware and does not rely on feature tracking instead on dense correspondence. The algorithm creates dense 3D surface model and uses it for dense camera tracking with whole image registration. This approach gives a high performance, quality and robustness but it requires a large number of narrow-baseline images [4].

2.1.2 Dynamic Scene Reconstruction

Given the practical importance of dealing with independent motions, there has been an increased interest in the detection and analysis of such cases. In general the analysis of dynamic scenes is of course not new in computer vision. The scenes which are subject to typical vision applications, such as tracking, background modeling, motion segmentation, contain dynamic elements by definition.

Dynamic shape reconstruction problem is a fundamental and heavily studied area in the field of computer vision. Methods have been proposed for dynamic reconstruction from monocular camera sequence [54, 92, 151, 192]. Garg et al. [54] introduced first variational approach to the problem of dense 3D reconstruction of non-rigid surfaces from a monocular video sequence. In [151] the whole scene is classified into background and foreground objects and the motion is modeled as a set of overlapping rigid parts. An approach for joint inference of 3D scene structure and semantic labeling for monocular video was proposed in [92]. A dense 3D template of the shape of the object is computed in [192] using a short rigid sequence and online reconstruction is performed subsequently of the non-rigid mesh as it evolves over time. However, this work focusses on recovering accurate 3D models of a dynamically evolving, non-rigid scene observed by multiple synchronized 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. Research on dense dynamic scene reconstruction from multiple views has primarily focused on indoor scenes with controlled illumination and backgrounds extending methods for multi-view reconstruction of static scenes [156] to sequences [178]. In the last decade, focus has shifted to more challenging outdoor scenes captured with both static and moving cameras.

The two main cues used frequently are the silhouettes of the object of interest and color consistency information. The former is used to form an approximate 3D shape estimate of

the original object, the visual hull. The latter is used to reconstruct the scene surfaces via multi-view stereo triangulation and both the cues are represented in Figure 2.2. Feature points are detected and matched for multi-view stereo. Many feature detection techniques have been introduced in the literature to find robust salient features. One of the most commonly used is the scale-invariant features is the SIFT feature detector [105] due to its performance on wide-baseline views with large appearance variation. In practice, for dynamic scenes of people most SIFT feature detections occur on the background or at the boundary of the silhouette. The latter unfortunately consists of partial foreground and background pixel information, and thus does not correspond to consistent features from a different view. Therefore, most of the recovered feature points cannot be used to perform 3D triangulation. Additional geometric constraints such as surface smoothness are often used to constrain the reconstruction where image data is inconclusive. A quantitative evaluation of state-of-the-art techniques for reconstruction from multiple camera views was presented by [156]. Existing methods generally focus on reconstruction in controlled environments, where the lighting is constrained and the viewing conditions used to obtain images of objects are optimal. They, however, face substantial difficulties when brought outdoors or in generally unconstrained environments. The prominent papers in the literature have been analyzed below:

Silhouette-Based Methods: A common approach to multi-view reconstruction uses the silhouettes of objects as sources of shape information. A 2D silhouette is the set of close contours that outline the projection of the object onto the image plane, as well as the regions inside the contours. Performing foreground/background segmentation of the object of interest is used to initialize or fit a 3D model as the silhouette provides a strong cue for shape understanding [59, 185]. The back-projection of the silhouettes from the camera optical center form generalized viewing cone's, and intersection of the viewing cones yields an approximation of the real object. The visual hull [97] refers to the intersected volume with an infinite number of viewpoints surrounding the object, in analogy to the convex hull of a set of points. The volume is maximal with respect to the visual silhouettes and the surface elements are tangent to the viewing rays (lines) along the silhouette boundaries [49]. Visual hull has been used as a constraint to obtain a high quality reconstruction of the dynamic object.

The 'visual hull' has a few characteristic properties. First, it is the maximal object that is consistent with all the silhouettes from the given viewpoints. This property is sometimes called 'conservativeness' of the visual hull, because the real shape, which is also consistent with all the silhouettes, is guaranteed to be contained in the visual hull. Secondly, every viewing cone can exclusively eliminate volumes outside of the cone. However, no matter how many cameras are used, surface concavities are not recovered, due to self-occlusion.

There is difference between 'concavity' and 'non-convexity': a visual hull is able to recover a 'saddle region', which is non-convex. Although the original visual hull concept is based on an infinite number of views, the above properties still hold for the shape recovered from finite views and this property is commonly used in practice for reconstruction of a volume which encloses the object of interest from a finite set of views.

Shape from silhouettes is a particularly good approach if only an approximate model of the real-world is required and the foreground can be segmented from the background. The methodology is intuitive and easy to implement. With the advances in computing powers, systems capturing real-time 3D digital video for dynamic reconstruction in studio environments are already on the market [3, 48, 110]. Nevertheless, such systems are restricted to an indoor environment with strictly controlled conditions such as uniform background of a distinct color allowing foreground segmentation. All such systems demand accurate silhouette extraction. If any part of a single silhouette were corrupted, due to the exclusiveness of the viewing cone carving, the corrupted parts would result in incomplete visual hull (which contradicts the visual hull 'conservativeness' property) and would never be recovered even if the silhouettes from all other views are correctly computed. These methods typically require minimum of 8 cameras covering the scene but typically $30-60$ cameras to get good approximation of object shape using visual hull.

Unfortunately, so far, there is no automatic solution that reliably produces accurate silhouettes for natural scenes. Varieties of visual hulls were introduced: [49] presented a technique to recover the exact representation of the visual hull corresponding to a polyhedral approximation of the silhouette contour; image-based visual hulls [112], an approximate view-dependent visual hull to efficiently render novel views without explicit reconstruction; classic work on 'Virtualized Reality' [80]; probabilistic visual hull [61] and safe-hull [118], the first visual hull reconstruction technique to produce a surface containing only foreground parts. Probabilistic visual hull methods use generative probabilistic sensor models, derived by analyzing the dependencies between the sensor observations and object labels which are assigned to each voxel location in the scene represented by a 3D volume. Bayesian reasoning is then applied to achieve robust reconstruction against real-world environment. A Bayesian approach [59] to silhouette estimation is used to compensate for modelling errors from false segmentation. They model prior density using probabilistic principal components analysis and estimated a maximum-a-posteriori reconstruction of multi-view contours. This reconstructs good error-compensated models from erroneous silhouette information, but needs prior knowledge about the objects and ground-truth training data. Another method based on multi-viewpoint silhouette images is presented in [95], but it is limited to controlled lighting and indoor datasets. A fully automated approach [165] is proposed for capturing a

human's shape, appearance and motion from multiple video cameras to create highly realistic animated content(time varying shape and appearance) from an actor's performance in full wardrobe based on visual hull and practical issues in designing studio capture systems were described in [166] including the studio backdrop, illumination, camera configuration and camera calibration but both focus on studio environment. Approach based on visual hull for complete outdoor dynamic scene reconstruction was introduced [62, 82].

In conclusion, shape-from-silhouette approaches can generate full 3D reconstruction of dynamic scenes, but these approaches assume foreground segmentation of the dynamic objects. They lack in reconstruction fidelity and are sensitive to errors in silhouette extraction unless large numbers of cameras are used, greater than 50. Therefore more efficient techniques, in terms of reconstruction quality, are employed to generate 3D models with an increased level of detail.

Multi-view Stereo: Image-based 3D scene reconstruction without a prior model is a key problem in computer vision. In the literature, there are several types of methods that are used to reconstruct 3D points of objects from a set of images. Conventional stereo-based techniques reconstruct a 2.5D depth image representation from two or more cameras through a regularized search for image correspondence using a geometric limitation such as the epipolar constraint. The epipolar constraint simplifies the correspondence search problem, as it limits the correspondence search region for a given point in one image to a single line in the other one. Due to these simplifications, stereo vision is a widely used technique in close range 3D reconstruction. One of the common method to achieve such dense reconstruction is to employ stereo-matching algorithms, where two images are matched pixel by pixel by exploiting intensity similarities, ordering constraints and smoothness constraints. A preprocessing step where both of the images are rectified with a projective transform is usually performed in order to come up with proper horizontal scan lines for both images. The stereo reconstruction problem is a relatively old and a well studied field but interestingly it is still an active research area where the interested reader can find a good review conducted in [154]. However, these work well for stereo image pair or narrow-baseline images where pixel by pixel feature matches can be easily retrieved.

Some multi-view stereo algorithms recover 3D surface point locations by triangulating corresponding visual features seen from different viewing angles [51, 103] and an initial survey was presented in [139]. Once these sparse critical points are recovered, smoothness constraints can be applied to regularize the recovery of the whole object surface. Previous research has focused on problems such as the recovery of the epipolar geometry between two stereo images [44, 69], and the calibration of multi-camera views [202]. These research

studies lay the foundation for current work in dense stereo reconstruction and volumetric modelling. These feature based approaches work well for narrow-baseline images where there is a large overlap between images and reliable correspondences can be retrieved. However, robust determination of the corresponding points and, consequently, of disparity is the central problem to be solved by stereo vision algorithms as matching suffers from ambiguities for images with uniform surface appearance, depth discontinuities and unknown surface visibility.

Other methods for multi-view stereo were proposed to handle these errors and wide-baseline scenes by using volumetric reconstruction techniques for calibrated cameras. These techniques derive the 3D volume that is consistent with multiple images. A volume representation allows inference of visibility and integration of appearance across multiple widely spaced camera views. A space carving framework to compute the 3D shape of an unknown scene from multiple images taken at known arbitrarily distributed viewpoints is introduced in [93]. It estimates the feature correspondences between views using the photo-consistency measures, that is the resemblance between the pixels/patches in one image to those in the other image is evaluated to see how well the two correlate. This is different from visual hull which provides an upper bound on the volume of the scene, and concavities that are occluded in silhouettes are not reconstructed. Space-carving techniques provide the photo-hull, which is the maximal volume that is photo-consistent across all visible camera images[93]. A space carving probabilistic framework for analyzing the 3D occupancy computation problem from multiple images is introduced in [22]. This framework enables a complete analysis of the complex probabilistic dependencies inherent in occupancy calculations and provides an expression for the tightest occupancy bound recoverable from noisy images referred to as the photo-hull. In [191], two major extensions to the space carving framework are presented. The first one is a progressive scheme for better reconstruction of surfaces lacking sufficient textures. The second one is a novel photo-consistency measure that is valid for both specular and diffuse (Lambertian) surfaces, without the need of illumination calibration. This method, unlike [22, 93], can deal with surfaces lacking sufficient textures.

In conclusion, the multi-view stereo framework suffers from several important limitations. The original space carving approach [93] makes hard decisions for the solution to converge. This limitation is partially overcome in [22]. The choice of the global threshold on the color variance is often problematic [22, 93]. An attempt to alleviate these photometric constraints in presented in [191]. Global optimization framework for such approaches was proposed in [86], thereby addressing some of the shortcomings of classical methodologies. A thorough survey is given in [156]. Unlike the visual hulls, concave regions can be recovered as long as feature correspondences in those regions can be triangulated from different camera views.

Another major advantage of multi-view stereo is that the recovered surfaces are a better approximation of the shape if triangulation is established for every point on the surface.

However, multi-view stereo algorithms generally work well on highly textured surfaces, where salient feature points are easy to find and match. In dynamic scene reconstruction applications, especially human modeling, uniformly colored clothes are common, thus, it is difficult to find salient features on the person. In the absence of sufficient robust features, stereo methods, such as [159, 186], have to 'reduce' to silhouette constraints with varieties of smoothness regularization such that the silhouette/visual hull is used as an initial approximation to constrain the stereo correspondence. In [176], a graph-cut algorithm is used to recover the 3D shape of an object using both silhouette and foreground color information. However, the solutions that incorporate silhouette constraints [159, 176] are only viable when exact silhouettes are available.

The second concern is the computational complexity. Unlike the straightforward visual hull algorithms which can reach real-time performance, multi-view stereo algorithms usually involve much slower optimization processes to obtain the detailed surfaces. From the Middlebury dataset evaluation [4], the fastest GPU accelerated algorithms take tens of seconds to output a final shape. This is not generally an issue for static scene modeling, where the reconstruction quality is the main concern. However, for dynamic scene modeling, these methods are not feasible for real-time applications.

The third concern is the self-occlusion problem. Since the 3D point triangulation requires at least two views of the same surface point from different perspectives, in order to obtain as many corresponding features as possible, neighbouring cameras should not be too far apart. In practice, existing successful multi-view stereo techniques either exploit tens or hundreds of views of a single static object, such as in [156, 191] which mainly works for static scenes, for multi-camera systems data volumes and color consistency across views can be an additional problem for large numbers of views or other approaches constraints the stereo correspondence using visual hull which requires foreground segmentation [166]. Besides pixel intensity, pixel color consistency is often used as a multi-view stereo correspondence measure, such as in [93]. The various color channels contain more information than the pixel intensity and thus produce more accurate 3D point correspondence. Although many advanced algorithms have been proposed for camera network photometric calibration as in [82], it is in general a tedious manual task.

Many recent approaches use a fusion of different reconstruction techniques to accomplish better reconstruction results. A high quality scene representation via graph-cut optimization of an energy function combining multiple image cues with strong prior was done in [62]. Optimization to solve labelling problem was done on color, contrast, sparse matching, dense

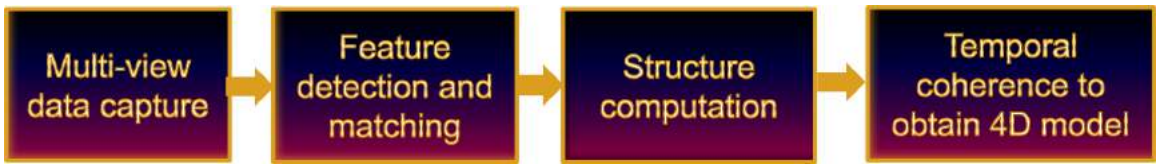


Fig. 2.4 4D reconstruction framework

matching and smoothness. A set of oriented points (or patches) covering the surface of an object or a scene of interests was used for reconstruction and named as Patch-based MVS algorithm [51]. This includes initial feature matching, patch expansion, patch filter, conversion of the patches into a polygonal mesh model and polygonal-mesh based multi-view stereo algorithm that refines the mesh. A dense 3D scene reconstruction from multiple images and simultaneous image segmentation was proposed by [193]. They utilize class-specific geometry priors for surface orientation and smoothness assumptions in order to improve the quality of obtained reconstruction and segment a volume of interest into a multi-label volumetric segmentation framework assigning object classes or free-space labels to voxels. There are many such methods like [57, 62, 82] to reconstruct a dynamic scene from multiple fixed cameras. They typically applied stereo algorithms to synchronized video frames of different cameras and smoothed the estimated disparities of corresponding pixels in the temporal domain. All these methods raise another concern which is the manual intervention to get good quality results, making it difficult for real-time applications.

Multi-view Depth Maps: Multi-view depth map approaches acquire multiple depth maps from 2.5D sensors and fuses them into a single reconstruction like DTAM [130], KinectFusion [75] and other reconstruction methods [131, 197]. However these methods work in case of static scenes. Recently RGBD based reconstruction methods were extended to dynamic scenes in [129]. First system to reconstruct non-rigidly deforming scenes in real-time by fusing together RGBD scans captured from commodity sensors was proposed. The method reconstructs scene geometry whilst simultaneously estimating a dense volumetric 6D motion field that warps the estimated geometry into a live frame to obtain real-time reconstruction of dyanmic scenes.

2.2 4D Reconstruction Pipeline

There are numerous applications of 3D dynamic scene modeling in our multimedia dominated modern world. In the movie industry, current motion capture systems have to attach trackable markers or sensors to the actors to recreate 3D structures and motions into a 4D marker or

model track i.e. 3D points on the objects are tracked over-time to create a 4D reconstruction sequence of dynamic scene. This is different to a 4D model where the entire surface is tracked over time. They are often inconvenient to put on and inhibit performance. The ability to obtain 3D shape and motion over time (4D) using images alone is therefore an attractive and desirable alternative. The demand for immersive 3D interactive games would be boosted by real-time modelling of full 3D human pose. The technology also has applications in the medical field. It can be used to create models of deforming organs, as well as to replay disease development over time. Other application areas include sports broadcasting and commentary, teleconferencing, robot navigation, object recognition, visual surveillance, digital historic archiving, assisted living and surveillance. In this section a general 4D reconstruction pipeline is presented to illustrate the steps required and review approaches for each step. Typical 4D reconstruction pipeline is illustrated in Figure 2.4 and it comprises of:

1. Image Acquisition: The reconstruction process starts with a data capturing step, where the image of the scene or the object is acquired by a fully calibrated and temporally synchronized camera or a regular hand-held camera.

2. Feature detection and matching: The images are processed to obtain the correspondence between the images either using sparse features or dense correspondences. Note that typically this step is visual hull not feature matching in existing techniques to general scene reconstruction.

3. Structure Computation: The correspondences are utilized to estimate the 3D dynamic scene structure including calibration of cameras and 3D points.

4. Temporal Coherence: The raw 3D reconstruction at each multi-view video frame is aligned over time to estimate a temporally coherent 4D representation using either a model-based prior or model-free approach.

This section reviews all the stages of general scene reconstruction from multi-view image sequences.

2.2.1 Data Capture

The most common set-up for a dynamic scene reconstruction consists of multiple cameras mounted at different fixed locations, looking at a common viewing region, as shown in Figure 2.5. The number of cameras placed in scene might vary typically vary from 5 to 16. Depending upon the applications cameras can be placed at narrow-baselines(within 10 degrees) or wide-baselines(more than 10 degrees and up to 50 degrees). With the development

Fig. 2.5 Data capture for IMPART project

in the algorithms and camera hardware, data has been captured using a single moving camera or a network of moving camera for any dynamic indoor and outdoor scenes. For example, a camera might be carried around a scene to build a geometric model from the resulting video or multiple hand-held cameras can be used to shoot the scene. Capturing data indoor is relatively easy as the capture volume is limited and lighting conditions can be controlled. In contrast, outdoor scene capture for film and broadcast production requires the following considerations.

- Large capture volume: Outdoor action capture commonly requires a relatively large area compared to indoor scenes.
- Natural scene backgrounds: Indoor captures have been performed with controlled chroma-key backgrounds and relatively static objects in the background, but outdoor capture will have natural image backgrounds that may be dynamic.
- Uncontrolled illumination: The illumination in the outdoor capture is subject to change based on the time-of-day and weather resulting in problems like shadows, shading, and specularities.
- Portable capture equipment: Set-up and reconfiguration of multiple camera equipment needs to be rapid to accommodate production requirements for principal photography.
- Fast scene motion: Rapid movement in the background like trees, cars etc.

The datasets used for our work is captured with a portable multiple HD camera system. The camera system comprises high-definition video (HDV) camcorders, Canon XH G1s, which have $f = 4.5$ – 90 mm, $F/1.6$ – 3.5 lens, and three $1/3$ in-charge-coupled devices providing uncompressed HD-serial digital interface (SDI) at 1920×1080 resolution. The

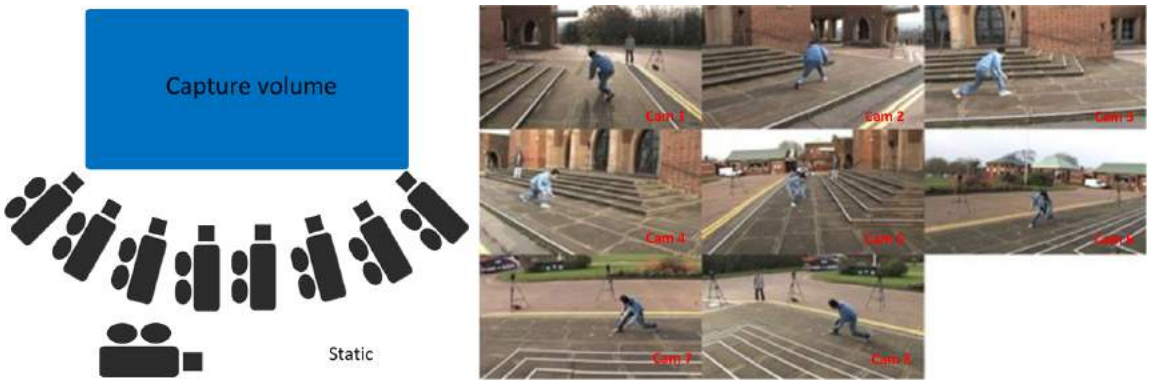


Fig. 2.6 Illustration of capture volume and a frame for Cathedral dataset

datasets are available online at [cvs]. The cameras can be synchronized during the capture using time-code generator or later using the audio information. The importance of synchronization is the genlock so that shutter opening is synchronized across views, without this it is not possible to reconstruct moving objects. In our work timecodes are synchronized to a master camera in advance. Cameras are then operated in a free running mode using their internal clocks for synchronization and video is captured to tape. This allows multiple cameras to be used without cables while maintaining synchronization enabling rapid set-up and reconfiguration with minimal impact onset. Synchronized multi-view image sequences can be extracted from the tapes with stored timecode for processing offline. Frame drift induced by using internal timecode is approximately one frame per hour in our experiments when cameras are free running, which is acceptable in most applications and can be detected using [74]. Since the cameras and the genlock synchronization signal generator can be driven by batteries, this system is fully portable and can produce synchronized multi-view image sequences of actions in any outdoor environment. In our work we have used various variety of datasets and all our datasets are publically available at [cvs]. The details are explained in this section:

- Odzemok: This is an indoor dataset with 8 cameras, 6 static and 2 moving. The background is cluttered the it is wide-baseline with stable lighting condition. An actor performs in the centre. Capture layout and example frames are illustrated in Figure 2.1.

- Dance1: This is a wide-baseline indoor dataset with 8 static cameras, cluttered background, stable lighting and actor performing in the centre illustrated in Figure 2.7.

- Dance2: This is a wide-baseline indoor dataset with 8 static cameras, plane background, stable lighting and two actors performing in the centre illustrated in Figure 2.7.

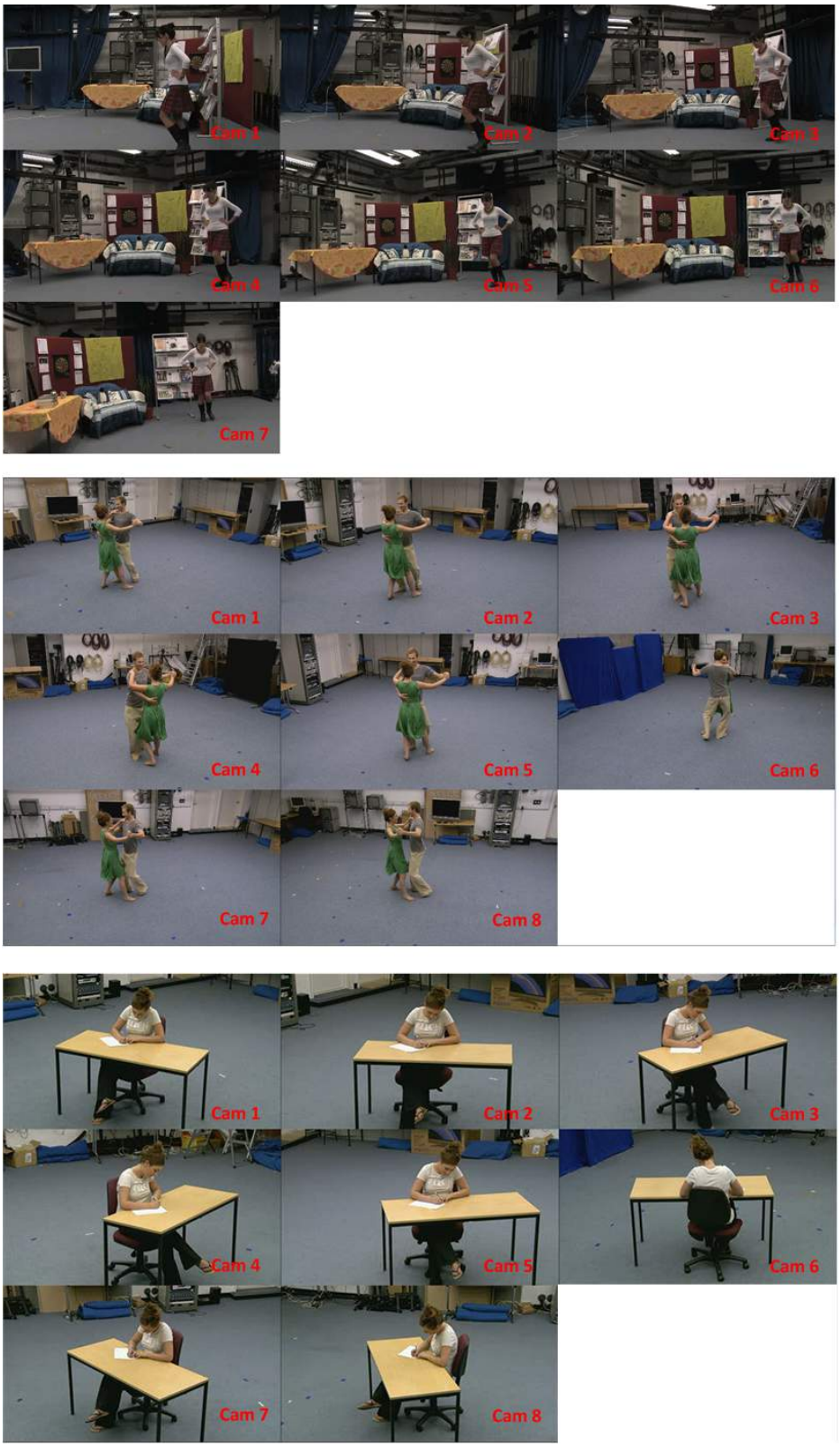


Fig. 2.7 Illustration of a frame with multi-views for Dance1, Dance2 and Office dataset

- Office: This is a wide-baseline indoor dataset with 8 static cameras, plane background, stable lighting and one actors performing with a prop table in the centre illustrated in Figure 2.7.

- Cathedral: This is a wide-baseline outdoor dataset with 8 static cameras, building as background, repetitive structure, varying illumination and actor in centre shown in Figure 2.6.

2.2.2 Feature Detection and Matching

Finding reliable correspondences between images is a fundamental problem in computer vision applications such as object recognition, camera tracking and automated $3D$ reconstruction. General scene reconstruction relies on obtaining reliable correspondences. Hence the presence of stable and representative features in the image is important, thus detecting, extracting and matching the image features is a vital step. The quality of the feature detection and matching directly affects the quality of reconstruction directly. The basic idea is to first detect interest regions (keypoints) that are covariant to a class of transformations. Two types of image features can be extracted form image content representation; namely global features and local features. Global features (e.g., color and texture) aim to describe an image as a whole and can be interpreted as a particular property of the image involving all pixels. While, local features aim to detect keypoints or interest regions in an image and describe them. In 3D reconstruction we use local keypoint features to obtain correspondences since they provide a limited set of well localized and individually identifiable anchor points. Local invariant features not only allow to find correspondences in spite of large changes in viewing conditions, occlusions, and image clutter (wide-baseline matching), but also yield an interesting description of the image content for image retrieval and object or scene recognition tasks (both for specific objects as well as categories) [179]. Various local feature detectors have been proposed in literature, such as Harris corner detector [68]. Depending on the datasets, requirements and speed efficiency various feature detectors were proposed consecutively like SIFT [105], SURF [18], MSER [109], KAZE [10] etc. for various applications.

Then, for each detected regions, an invariant feature vector representation (i.e., descriptor) for image data around the detected keypoint is built. Feature descriptors extracted from the image can be based on second-order statistics, parametric models, coefficients obtained from an image transform, or even a combination of these measures. A large variety of feature extraction methods have been proposed to compute reliable descriptors. Among these descriptors, the scale invariant feature transform (SIFT) descriptor [105] utilizing local extrema in a series of difference of Gaussian (DoG) functions for extracting robust features

and the speeded-up robust features (SURF) descriptor [18] partly inspired by the SIFT descriptor for computing distinctive invariant local features quickly are the most popular and widely used in several applications. These descriptors represent salient image regions by using a set of hand-crafted filters and non-linear operations.

Once the descriptors are computed, they can be compared to find a relationship between images for performing matching tasks. Generally, the performance of matching methods depends on the underlying keypoint detectors and choice of associated image descriptors. The datasets in our work are wide-baseline hence our focus is on wide-baseline matching. Several techniques have been proposed in the literature [105] and are reviewed in detail in Chapter 3.

2.2.3 Structure Computation

Once we have the correspondences the next step is to obtain the calibration of the scene. The calibration is used to derive the sparse and dense structure of the scene and a review of camera calibration and dense reconstruction is presented in this section.

Camera Calibration:

Camera calibration is a necessary step in 3D computer vision in order to extract metric information from 2D images. It has been studied extensively in computer vision and photogrammetry, and even recently new techniques have been proposed [202]. Primarily it means, finding the quantities internal to the camera that affect the imaging process such as, position of camera centre, focal length, different scaling factors for row pixels and column pixels, skew factor and lens distortion. Good calibration is important to reconstruct a world model with high accuracy and a benchmark is presented in [167].

Camera calibration is a well studied field and one crucial factor in the calibration is the availability of different camera projection models [32]. Indeed, a camera is just a device which projects a 3D scene to a 2D image plane and the choice of a proper mathematical function to model such a transformation depends on the type of application. Camera models can be broadly categorized as linear and non-linear. In the linear case, the simplest camera model is the orthographic one, where the projection of a 3D point is independent of its depth, i.e. perspectivity does not exist. Although its applicability is limited to scenes where perspective effects are negligible, the existence of powerful and relatively simple factorization algorithms such as the pioneering work of Tomasi and Kanade [173], makes it attractive. Kanade's work is based on the observation that when the feature tracks over a sequence are listed in a matrix, the rank of that matrix is at most three assuming orthographic projection. Consequently singular value decomposition factorization methods can be used to extract structure parameters from that matrix. This method has been extended to more realistic

linear camera models, for example para-perspective and perspective cameras [138]. However initialization of the feature point depths is required as a preprocessing step in the latter work.

When the linear models are not adequate due to strong perspective distortion, perspective camera models must be used. The pinhole model representing a perspective projection without lens distortion is widely used. Typically initialization is performed using two or three view geometric relationships, the fundamental matrix and tri-focal tensor respectively. Typically for a multi-camera network the relationship is initially estimated between a pair of cameras and new cameras are added sequentially to achieve full calibration. Such algorithms are reported to be quite successful [19]. The underlying multi-view geometric concepts are discussed in [43, 69].

Calibration is usually performed by: (1) estimating pairwise relations as outlined above i.e. fundamental matrix/epipolar geometry and, (2) performing bundle adjustment across all views to refine the cameras and estimate sparse 3D structure.

A practical technique which is proved to be effective is to compute the internal camera parameters in a separate step before processing the target images. Such a preliminary computation decreases the number of parameters to be estimated during the reconstruction procedure and consequently results in a better conditioned problem. This can be accomplished with a typical calibration scheme such as [201] or with the aforementioned uncalibrated techniques. A-priori knowledge on the internal matrices significantly improves the system's performance in the case of scene degeneracies such as dominantly planar regions or when the viewed object is relatively small compared to the image size. As a successful example, a solution for the 5-point pose estimation problem was reported in [132] and a real-time system exploiting it which could run for quite long sequences without drifting.

The calibration can be performed using reference objects or self-calibration techniques which requires only image point correspondences. Camera parameters are estimated from internet photo collection in [162]. Although no calibration objects are necessary, a larger number of parameters need to be estimated. The accuracy of such methods is low compared to other methods. Another work which targets review of camera calibration and dense multi-view stereo of outdoor scene is [167]. Ground-truth is obtained from laser scans and the tested aspects for 3D model building are pose estimation and multi view stereo with known internal parameters, camera calibration and multi-view stereo with raw images. The effect of any camera can be considered to be a projection of the 3-D real-world scene on to an image plane [45]. Camera model hierarchy was studied by [12]. Details of the theory on projective camera was investigated by [43]. Various camera models, their taxonomy and properties are given in [69].

For our internal datasets the intrinsic parameters are calibrated using a checker board. Extrinsic parameters are estimated by wand-based calibration using bundle adjustment between observed locations of colored markers [119]. For moving cameras, through-the-lens calibration is employed to register the moving camera with the multiple static witness cameras and estimate the extrinsic parameters [74].

Dense reconstruction:

The next step in the 3D reconstruction process is the structure recovery, given the feature matches and calibration as the input to the system. The structure information is recovered as the 3-D coordinates of the features obtained from the images. Geometrical constraints can be established between the feature correspondences. Finally, 3-D coordinates of the feature points can be recovered via triangulation. Although obtaining 3D points as output maybe sufficient for some applications, such as robot navigation, it is desirable to generate 3D depths for each pixel in the input images for realistic 3D graphics rendering and the review is presented in Section 2.1.2.

2.2.4 Temporal Coherence

A critical step for editing and reuse reconstructed 3D data from multi-camera capture is the temporal alignment of captured mesh sequences to obtain a temporally coherent mesh with surface correspondence over time. Such temporally coherent representation helps in efficient representation, scene manipulation and analysis of motion while using a conventional computer graphics pipeline.

Model-based approaches for 4D dynamic reconstruction inherently provide a 4D mesh representation of the motion by utilizing a geometric template which is iteratively deformed to match the captured silhouettes [53, 144]. However, the use of a geometric template limits the range of surface deformation which can be represented. Model-free approaches [28, 73] align a series of 3D mesh per frame with varying topology into a 4D model representation which allows the representation of more general shape deformation. The model free temporal alignment approaches can be classified in sequential and non-sequential. A sequential approach by Cagniart et al. [28] divided the meshes into overlapping patches which are tracked independently. The tracked patches are then mapped to a reference mesh to obtain the temporal coherence. Sequential coherence refers to estimating frame-to-frame temporal correspondences for each time frame. Another approach by Franco and Boyer [49] use Expectation Maximization to jointly track and piece-wise rigidly segment a dynamic mesh sequence reconstructed from a multi-view camera system. Both of these approaches assume that the full geometry of object is available to obtain temporal coherence for the entire sequence. However, sequential methods suffer from drift due to accumulation of errors

and fail in case of large motion. Temporal coherence is introduced frame-to-frame which leads to drift in the alignment due to accumulation of errors and rapid motion between frames introduce large errors. The non-sequential methods overcome these limitations and have been used to generate consistent geometric representation for model based multi-view and RGBD data. Non-sequential non-rigid global alignment [27] using Prim's minimum spanning tree was introduced due to errors in sequential approach. Huang et al. [73] produced non-sequential alignment over multiple sequences rather than a single sequence. Their proposed alignment algorithm uses shape similarity (via shape histograms) to select similar poses between sequences and Minimum Spanning Tree (MST) graph optimization to minimize the total non-rigid deformation required to bring all frames into registration with a common structure, thus reducing drift and increasing robustness against large non-rigid motion. However, these approaches require water-tight meshes on the dynamic objects and in real-world it is difficult to capture the full geometry of the dynamic objects and it is difficult to apply these full geometry based techniques to complex datasets and partial surfaces.

Sequential methods were introduced to align geometries for partial surfaces. The concept of optical flow was used to estimate scene flow between pixels of two 2D images to obtain frame to frame alignment [113, 180]. A piecewise rigid scene flow approach is proposed by Vogel et al. Other methods for 3D tracking jointly estimate the shape and motion of the dynamic scenes directly in the 3D space [52]. Wei et al. [188] proposed a sequential approach to align partial surfaces. A deep learning approach for finding dense correspondences between 3D partial geometries was introduced. A feature descriptor was trained on depth map pixels to solve a body region classification problem. As discussed before these sequential approaches suffer from errors due to drift. A non-sequential approach to align partial surfaces from Kinect RGBD data was proposed in [108]. Feature tracks were obtained in 2D for the entire sequence and non-rigid global alignment was performed to achieve temporal coherence. However, there are no approaches in literature for non-sequential alignment of partial surfaces obtained from multi-view RGB images.

2.3 Summary

This chapter reviewed existing multi-view static and dynamic scene reconstruction methods. A review of approaches at each step of a typical $4D$ scene reconstruction pipeline from wide-baseline multi-view video is presented. The review highlighted the following problems in the literature of general dynamic scene reconstruction from multi-view videos:

- Existing feature detection algorithms for wide-baseline multi-view reconstruction produce non-uniform distribution of features with poor scene coverage in the sparse scene reconstruction.
- The state-of-the-art multi-view scene reconstruction methods require strong assumptions of prior knowledge of the scene background, structure, segmentation or appearance information to obtain the solution.
- Reconstruction is typically performed independently at each frame resulting in reconstruction of an incoherent mesh sequence without temporal correspondence or consistent connectivity at successive frames.
- The methods used for temporal coherence across the entire sequence suffer from various limitations: sequential coherence (frame-to-frame) which is prone to errors; assume rigid objects with rigid motion in the scene; and some approaches only work for RGBD scenes.

In this thesis we address these limitations with previous approaches by introducing a novel approach to general 4D dynamic scene reconstruction from multi-view videos in the following chapters.

Chapter 3

SFD: Segmentation based Features for Wide-baseline Reconstruction

Feature detection and matching is the first step to obtain dynamic scene reconstruction. The feature correspondences between pair of multi-view images are used to obtain the sparse reconstruction of the scene. The quantity and quality of these feature matches determine the quality of sparse 3D reconstruction of the scene. Hence the feature matches must be distributed uniformly throughout the scene to get a reliable and complete representation of the scene structure.

The correspondence obtained using existing feature detectors tend to be clustered on a relatively few regions in the scene. These approaches result in a sparse non-uniform distribution of scene features. Whilst this may be sufficient for camera estimation using bundle adjustment the resulting feature set often results in poor scene coverage in the sparse reconstruction. To handle this issue, we introduce a new feature detector for wide-baseline scene reconstruction which gives uniform distribution of the sparse 3D points throughout the scene and produces consistent feature detection with change in viewpoint (wide-baseline views).

3.1 Introduction

Feature detection and matching plays a crucial role in multi-view stereo to obtain sparse reconstruction of the scene. Finding reliable correspondences is a well-studied field in computer vision and various feature detectors have been proposed in the literature. Existing feature detection approaches work well for narrow-baseline multi-view data giving a reliable sparse scene reconstruction, but the performance of these detectors is limited for wide-

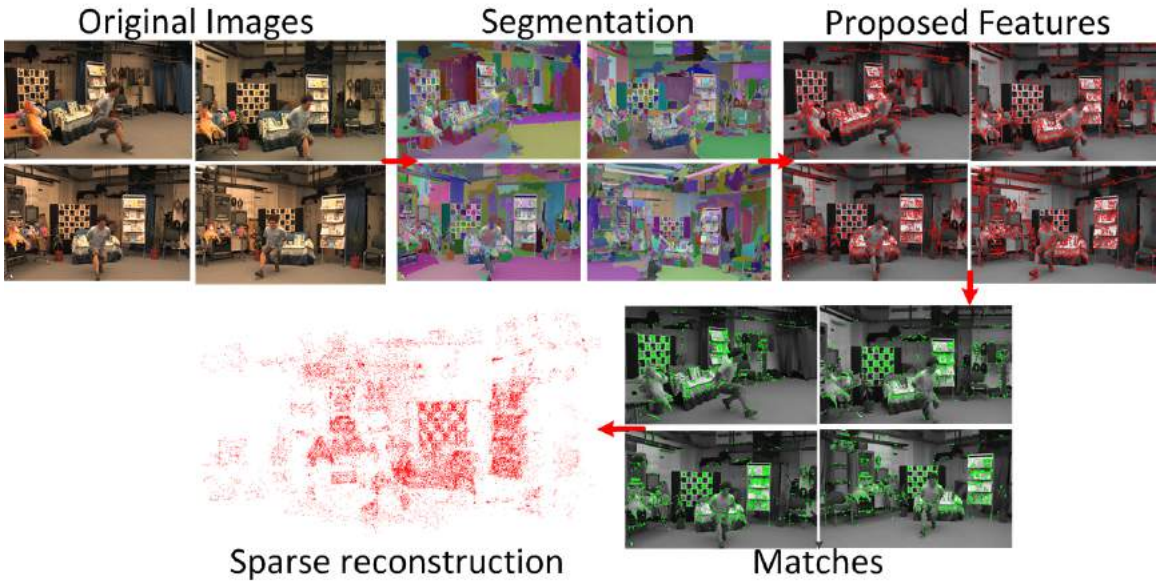


Fig. 3.1 SFD for wide-baseline matching and sparse reconstruction for Odzemok dataset.

baseline multi-view images. A common problem in wide-baseline sparse matching is the sparse and non-uniform distribution of correspondences when using conventional detectors such as SIFT, SURF, FAST, KAZE and MSER as seen in Section 3.5. Established feature detectors such as Harris [68], SIFT [105], SURF [18], FAST [147] and MSER [109] often yield sparse and non-uniformly distributed feature sets for wide-baseline matching.

Gradient based detectors (Harris, SIFT, SURF, STAR [8]) locate features at points of high-image gradient in multiple directions and scales to identify salient features which are suitable for affine-invariant matching across multiple scales resulting in non-uniform sparse correspondences. Alternatively, Watershed segmentation based detectors (MSER) identify salient regions which are stable across multiple scales which can be reliably matched across wide-baseline views also resulting in relatively few feature matches.

In this chapter we propose a new segmentation based feature detector SFD which uses the segmentation boundary (local maximal ridge lines of the image) rather than the segmentation regions for wide-baseline scene reconstruction as shown in Figure 3.1. We use the segmentation boundaries and SFD feature point detections are located at the intersection points of three or more region boundaries. The intersection points represent local maxima of the image function in multiple directions giving stable localization. Evaluation of SFD feature point detections across wide-baseline views demonstrates that although the segmentation changes with viewpoint the region intersection points are stable and accurately localized. SFD feature points are also demonstrated to give improved scene coverage with computational cost similar

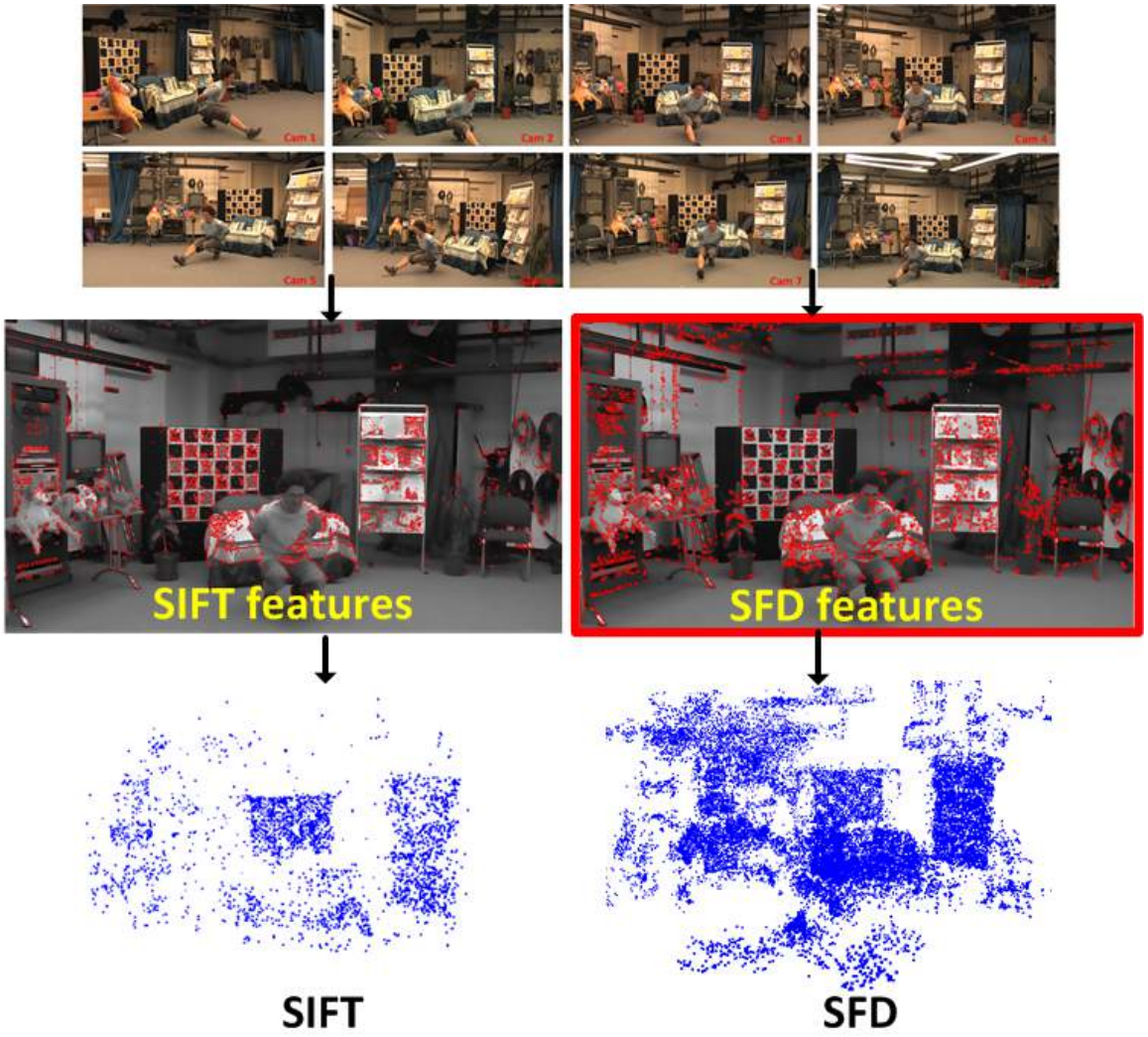


Fig. 3.2 Comparison of feature matching and sparse reconstruction using SIFT and SFD for Odzemok dataset.

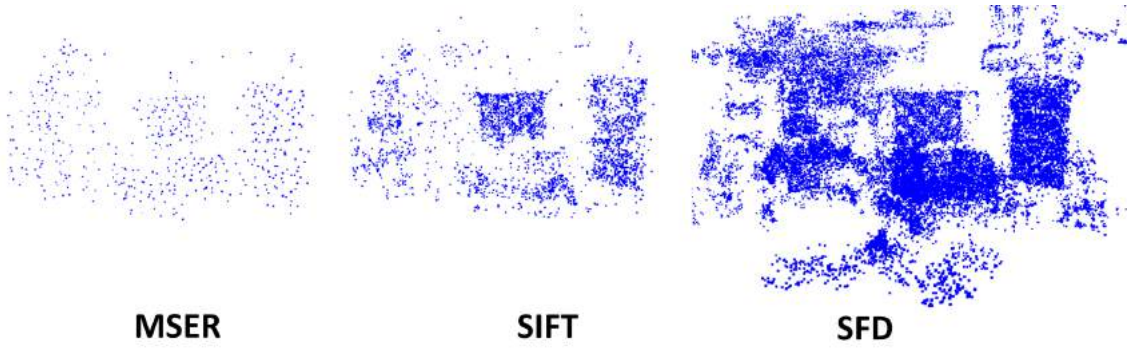


Fig. 3.3 Comparison of sparse reconstruction using MSER, SIFT and SFD for Odzemok dataset.

to existing efficient wide-baseline feature detectors (SURF/FAST). Figure 3.2 presents an example of comparison of wide-baseline correspondences and sparse 3D points for SFD and SIFT. Figure 3.3 illustrates comparison of sparse reconstruction obtained using SFD, MSER and SIFT. The sparse features based on SFD has better coverage compared to SIFT and MSER. Few features are detected in the regions around the corners of the images for SIFT and MSER, but these features are lost in matching and reconstruction stage as compared to SFD that detects increased number of features around the corners of the image, retained in the sparse reconstruction. Contributions in this chapter are:

- A novel segmentation based feature detector SFD for accurate wide-baseline matching; which gives an increased number of good and repeatable features for different viewpoints; accurate feature localization; and improved coverage for natural scenes. The detected features are dynamically adaptable by a intuitive threshold.
- Feature detection can be applied to multiple segmentation techniques like Watershed, Mean-shift, SLIC etc.
- A comprehensive performance evaluation for wide-baseline matching on benchmark datasets against existing feature detectors (Harris, SIFT, SURF, FAST, MSER, ORB, AKAZE) and descriptors(SIFT, BRIEF, ORB, SURF) showing improved performance of the SFD detector in terms of both number of features and matching accuracy;
- Application to scene reconstruction demonstrates an order of magnitude increase in the number of reconstructed points with improved scene coverage and reduced error compared to previous detectors against ground-truth.

3.2 Related Work

Features are defined as interesting image points with the following characteristic properties [179]:

- *Repeatability:* A feature should be invariant to noise and distortions.
- *Distinctiveness*: The feature detected should be unique in its neighbourhood.
- *Quantity*: The number of detected features should be sufficiently large, such that a reasonable number of features are detected even on small objects. However, the optimal number of features depends on the application. Ideally, the number of detected features should be adaptable over a large range by a simple and intuitive threshold.

- *Accuracy*: The detected features should be accurately localized.
- *Efficiency*: Preferably, the detection of features in a new image should allow for time-critical applications.

Existing feature detection techniques are proposed to satisfy these constraints. Decades of research has developed a lot of feature detection techniques and a review into interest-point detection reveals three main approaches [121, 149]: image gradient analysis, intensity templates, and contour analysis.

3.2.1 Image Gradient based Features

Early image gradient based approaches, such as Forstner corner detector [47], define an optimal point based on the distances from the local gradient lines and Harris corner detector [68], define an interest-point as the maximum of a function of the Hessian of the image. They used the local autocorrelation function of a signal to measure the local changes of the signal with patches shifted by a small amount in different directions. The Harris points are translation and rotation invariant. Moreover, they are stable under varying lighting conditions.

A multi-scale invariant extension was achieved by successive application of Gaussian kernels on scale-space representation of image and detecting interest-point as a local maximum both spatially, and across the scale-space [117] to deal with significant affine transformations. Mikolajcyzk and Schmid seek these maxima via the Laplacian-of-Gaussian (LoG) filter, which is a combination of the Gaussian smoothing and the differentiation operation [116]. The idea is to select the characteristic scale of a local structure, for which a given function attains an extremum over scales. The selected scale is characteristic in the quantitative sense, since it measures the scale at which there is maximum similarity between the feature detection operator and the local image structures. The size of the region is therefore selected independently of the image resolution for each point.

SIFT implements difference-of-Gaussians [105] to improve on earlier approaches by transforming an image into a large collection of local feature vectors which are invariant to changes in scale, illumination and local affine distortions. Lowe exploited locations that are maxima or minima of a Difference-of-Gaussian (DoG) function applied in scale space to generate local feature vector that represent an image as a one parameter. The key locations are computed by building an image pyramid with re-sampling between each level. A combination of gradient space with local symmetry was used in [71].

Gradient based techniques offer accurate localization [5], and are robust to many image transformations [117]. However, computation of the image gradients are sensitive to image noise and are computationally expensive. SURF mitigates this via the use of integral images

and 2D Haar wavelets [18]. Their approach is based on Hessian matrix (Fast Hessian) and sums of 2D Haar wavelet response. They rely on the determinant of the Hessian for selecting both location and scale. A box filter is used to approximate second order Gaussian derivative.

CenSurE achieves even faster operation by approximating the LoG operator with a bi-level filter [8]. These approaches suffer from drawbacks since Gaussian blurring does not preserve object boundaries and smooths to the same extent details and noise at all scales, spoiling localization accuracy and distinctiveness. To overcome this problem KAZE features have been presented that aim to detect and describe features in non-linear scale spaces [10]. The scale-space representation is computed by non-linear diffusion filtering (instead of Gaussian smoothing), yielding an improvement in the localization accuracy in [10](A-KAZE). This makes blurring locally adaptive to the image data, blurring small details but preserving object boundaries. This method increases repeatability and distinctiveness with respect to SIFT and SURF because of the use of non-linear diffusion filtering. The main drawback of KAZE is that it is computationally intense. However, A-KAZE claims superiority over all major gradient based methods in terms of computational complexity, by using efficient diffusion filtering [11].

3.2.2 Intensity based Features

Intensity template approaches seek patterns that are common manifestations of interest-points [149]. SUSAN [161] design a non-linear cornerness function, which evaluates the dissimilarity of a pixel to a disc surrounding it. FAST replaces the non-linear response functions by a simple, but effective heuristic: it first computes the intensity differences between the central pixel and a circle surrounding it, and then counts the contiguous pixels with a difference above a threshold [147]. A rotation-invariant implementation is proposed in [150], and a multi-scale extension, in [101]. An extension to a multi-scale detector by scale selection with the Laplacian function was proposed in [100]. The Laplacian is estimated using grey-level differences between pixels and location with largest estimate are retained. This produces a large number of keypoint candidates which can be refined in the matching process.

MSER or Maximally Stable Extremal Regions can be considered as a region detector responding to areas conforming to a “basin” template [109]. The regions are connected components of a thresholded image. The word “extremal” refers to the property that all pixels inside the MSER have either higher (bright extremal regions) or lower (dark extremal regions) intensity than all the pixels on its outer boundary. The “maximally stable” in MSER describes the property optimized in the threshold selection process since every extremal region is a connected component of a thresholded image. Intensity template methods are usually fast,

compared to their gradient based counterparts [149]. However, with the exception of MSER, they are not affine-invariant, which limits their ability to cope with viewpoint variations and a recent comparative evaluation is presented in [5].

3.2.3 Contour based Features

Originally, contour based methods were popular in applications like line drawings, piecewise constant regions, and cad–cam images rather than natural scenes and the focus was on the accuracy of point localization. Contour intersections and junctions often result in bi-directional signal changes. Therefore, a good strategy to detect features consists of extracting points along the contour with high curvature. Curvature of an analog curve is defined as the rate at which the unit tangent vector changes with respect to arc length. Contours are often encoded in chains of points or represented in a parametric form using splines. Hence, image contours give rise to two interest-point definitions: local maxima of the curvature along a contour, and intersections. Mokhtarian and Suomla [120] implement the former by building a scale-space representation of the contour map for the image, and detecting the local maxima of the curvature. The robustness was improved by using gradient correlation based detector [200]. Intersection of contour elements provides an alternative interest-point definition. T-junctions constitute a straightforward example [120][21] which inspires the proposed feature detector. Performance of curvature based techniques are dependent on the quality of the extracted edges [13]. Although they are generally fast, the scale-space approach introduces a compromise between robustness and accuracy. On the other hand, contours, especially intersections are distinctive. Therefore, they are more robust to viewpoint variation [13, 107].

3.2.4 Learning based features

Although work on feature detectors were mainly focused on handcrafted methods, learning based methods have been proposed recently [70, 145, 182, 183]. A classifier was learned to detect matchable keypoints for Structure-from-Motion (SFM) applications in [70]. They collect matchable keypoints by observing which keypoints are retained throughout the SFM pipeline and learn these keypoints. Although their method shows significant speed-up, they remain limited by the quality of the initial keypoint detector. [145] learns convolutional filters through random sampling and looking for the filter that gives the smallest pose estimation error when applied to stereo visual odometry. Approaches based on machine learning were proposed with FAST, [148] with aim of speeding up the detection process and improved

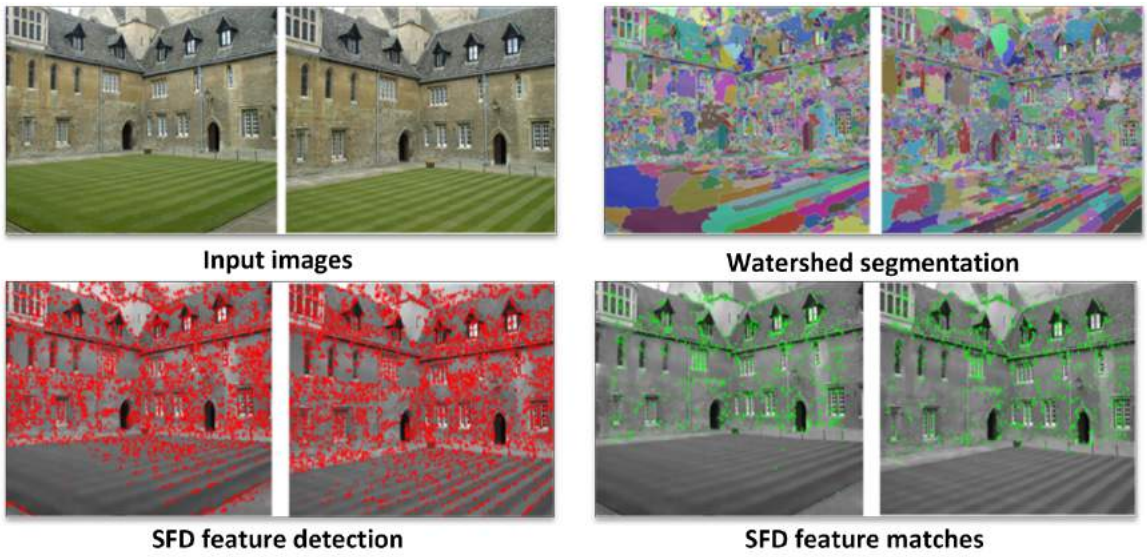


Fig. 3.4 SFD feature detection and matching using watershed segmentation on Merton dataset.

repeatability in FAST-ER [149]. But none of these features have been designed for the purpose of wide-baseline stereo.

3.2.5 Summary and Motivation

To overcome the limitations of existing feature detectors in terms of scene coverage and matching across wide-baseline views a segmentation based feature detector is introduced. It is based on the property of intersections of contours which is robust to changes in viewpoint as in [13, 107]. The number of features detected by curvature based techniques is quite small [13] and none of them have been evaluated on wide-baseline image pairs. They are based on only edge detection and vulnerable to the well-known difficulties in producing stable, connected, one-pixel wide contours [67]. To avoid this an over-segmentation based method for stable feature detection is proposed. The idea of using regions for salient feature matching is well known and is exploited in [17, 83, 175] for applications other than wide-baseline stereo. A survey on interest points based on Watershed, Mean-shift and Graph-cut segmentation was presented by [88]. A method is proposed [88] that uses boundaries and centres of gravity of the segments for extracting features. This demonstrates that Watershed is superior to the alternatives in terms of repeatability and Mean-shift segmentation performs best for natural scenes. Watershed detects the local maxima of the gradient magnitude intensities as the region boundaries and proposed feature detection is based on the the

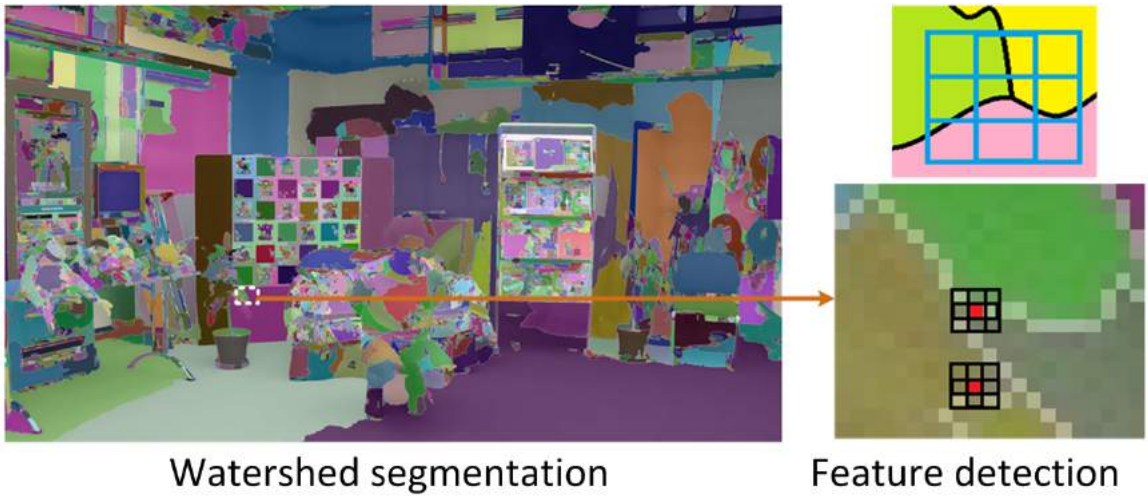


Fig. 3.5 Illustration of SFD feature detection on the watershed segmentation for Odzemok dataset.

detection of features as the intersection of local maxima, therefore Watershed is chosen as base segmentation technique. An example of watershed segmentation followed by feature detection and matching is shown on the Merton dataset in Figure 3.4. The detected features and the correspondences are consistent across changes in viewpoint and are uniformly spread across the scene.

3.3 SFD-Segmentation based Feature Detector

In this section new segmentation based feature detector is described. The main motivation for this approach is to increase the quantity and distribution of distinct features detected throughout the scene which are suitable for accurate wide-baseline matching and reconstruction, as shown in Figure 3.1. The approach is based on over-segmentation of the image into regions which ensures that detected features are distributed across the entire image as the region boundaries are located along contours of local maxima in the image which are consistent with respect to viewpoint change [88]. The use of local maximal contours overcomes the common problem of setting arbitrary thresholds or scales for feature detection, which is common to most existing feature detectors. SFD feature detection is based on the segmentation of the image such that the features are detected at the boundaries of the segmented regions hence the name ‘Segmentation based features’.

Fig. 3.6 SFD feature detection on Odzemok dataset for 4 views illustrating the stability of SFD with changes in viewpoint.

3.3.1 Feature Detection

Segmentation of an image results in a large number of small regions with uniform appearance. The region boundaries represent ridge lines corresponding to local maxima of the image function or maxima in gradient if the segmentation is performed on a gradient image. The boundary intersection points where three or more region boundaries meet are local maxima in the image function in multiple directions. Consequently these points are accurately localized, distinctive and stable under changes in viewpoint giving good features for matching across wide-baseline views. This observation forms the basis of our proposed region based feature detector, resulting in an increased number of salient features which are suitable for matching across wide-baseline views.

Over-segmentation is performed on the image using existing segmentation techniques such that the regions in the image are separated by a 1 pixel wide boundary. The intersection points of three or more region boundaries in the over-segmented image are detected as

features and a unique intersection can only be obtained with regions of 1 pixel wide boundary. The region intersection points are identified by traversing through the boundary points in the image such that for each point on the contour 3×3 pixel neighbourhood is tested for the number of region labels. If three or more region labels are present the point is detected as a feature as illustrated in Figure 3.5. These points are detected for the whole image on the region boundary contours. Locating features where multiple region boundaries (3 or more) intersect followed by sub-pixel refinement gives good localization, therefore SFD achieves good localization which is consistent with-respect-to changes in viewpoint, as illustrated in Figure 3.6 for Odzemok dataset. Segmentation is obtained for different viewpoints with a baseline varying between 0 to 90 degrees followed by SFD feature detection. Feature points consistent across views highlighted in the figure to show the stability of localization with viewpoint for SFD feature detection.

3.3.2 Sub-pixel Refinement

Let us denote the set of features detected for an image as $F = \{f_1, f_2, ..., f_{N_F}\}$, where N_F is the total number of features. These features are integer values of the pixels where intersections of regions are detected. We perform a local sub-pixel refinement to optimize the feature location f_i at a local gradient maxima using the Levenberg-Marquardt [102] method. This refinement is based on the observation that every vector from the feature f_i to a point p_j located within a neighbourhood N of $f_i = \{x,y\}^T$ is orthogonal to the image gradient $G_j = \{g_x, g_y\}^T$ at $p_j = \{x+\Delta x, y+\Delta y\}^T$, where $\Delta x, \Delta y$ is the shift at the point f_i. In our case a window size of $W \times W$ is chosen for the neighbourhood N, such that $W = \frac{min(W,H)}{100}$ which is the optimum window size for good localization of the features [55]. The cost function is defined as:

$$T(f_i) = \sum_{j \varepsilon \mathcal{N}} t_j(f_i), \text{ where, } t_j(f_i) = (G_j^T(f_i - p_j)(1 - e^{-\frac{\Delta x_i^2 + \Delta y_i^2}{2}}))^2 \tag{3.1}$$

Since the vectors G_j and $f_i - p_j$ are orthogonal, $t_j(f_i)$ is 0 if f_i is at a local maxima, thereby making $T(f_i)$ to be 0. The sub-pixel position of the feature point is the minima of $T(f_i)$. The process is repeated for the entire feature set F to obtain a new solution F^* and the speed is optimized by parallelization.

$$F^* = \underset{f_i}{\operatorname{argmin}} \{T(f_i)\} \tag{3.2}$$

Feature descriptors are then applied to the local image regions of F^* to perform matching. In Section 3.5 various feature detectors with descriptors based on SIFT and BRIEF for matching are evaluated.

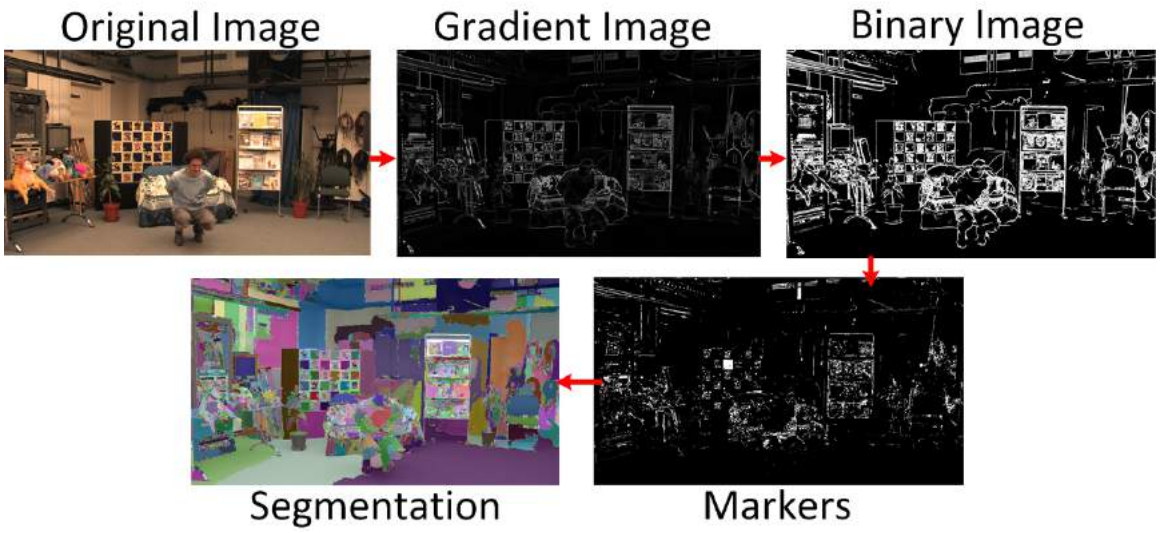


Fig. 3.7 Modified watershed algorithm used for SFD detection.

3.3.3 Segmentation

SFD can use different segmentation techniques, in this section we review possible segmentation methods. The performance of SFD for different segmentation methods is evaluated in Section 3.5.

Segmentation of an image is defined as the process of partitioning an image into multiple segments. Pixels in each region share similar properties and are distinct from the pixels in adjacent regions. The boundary of the segments define contours of local maxima in the image. Our focus is on finding fast, automatic and stable over-segmentation techniques suitable for wide-baseline matching in general indoor or outdoor scenes. The SFD features defined in Section 3.3.1 are evaluated on three different segmentation techniques:

Watershed (WA) [146]: The first segmentation technique is based on morphology. Readers are referred to [115] for detailed information on morphological segmentation techniques; the watershed transform [146] is used in this approach because of speed and efficiency. The watershed transformation considers the gradient magnitude of an image as a topographic surface. Pixels having the highest gradient magnitude correspond to watershed lines which represent the region boundaries. Water placed on any pixel enclosed by a common watershed line flows downhill to a common local intensity minimum. Pixels draining to a common minimum form a basin, which represents a segment partitioning the image into two different sets: the catchment basins and the watershed lines.

Implementing the transformation on the image gradient, the catchment basins correspond to homogeneous grey level regions of this image. In practice, this transform produces an

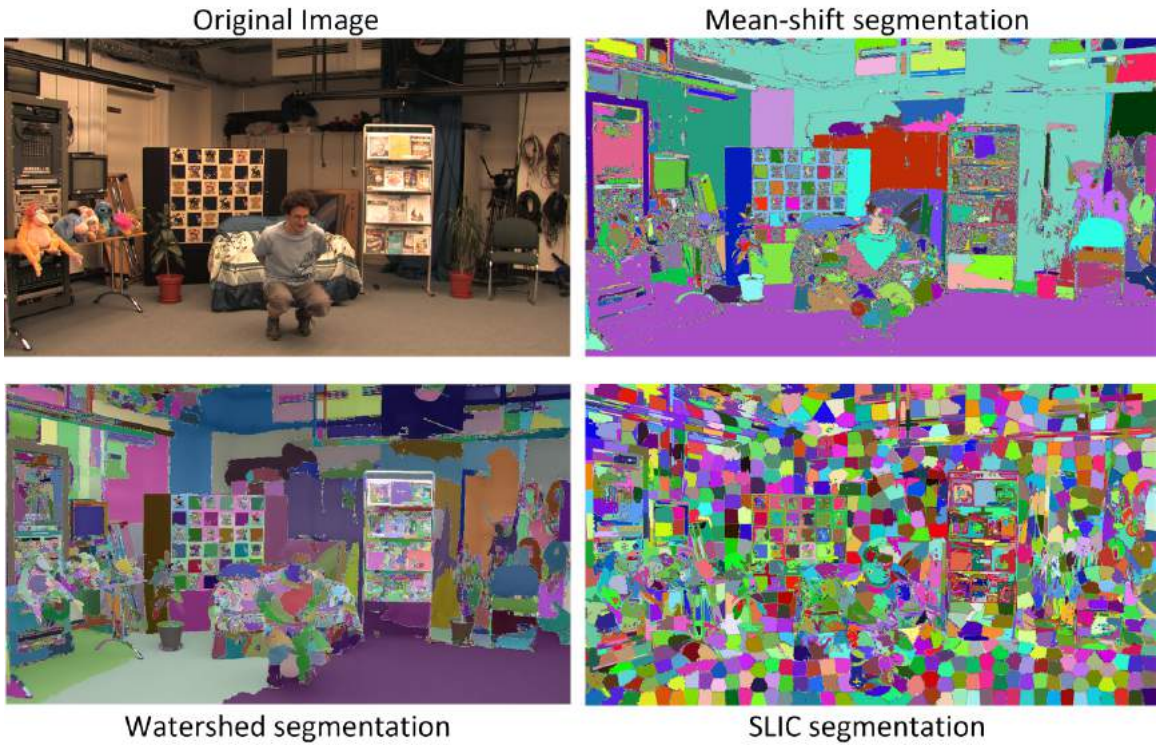


Fig. 3.8 Different segmentation algorithms for SFD feature detection.

over-segmentation due to scene structure, local appearance variation and image noise. We use the modified version of the watershed algorithm defined in [137]. A single marker for each region can be defined for watershed segmentation. These markers (or seeds) initiate the flooding, indicating the sector that gives rise to the basins. The characteristics of the markers determine the success of the watershed transform. For a repeatable over-segmentation the gradient of the image is used to retrieve markers. A step-by-step algorithm is shown in Figure 3.7 and is described as follows:

Step 1: To remove the presence of high-frequency image noise the image is pre-processed using a Bilateral filter [174]. The Bilateral filter is applied recursively with kernel length increasing from 1 to 31. This has been found empirically to give a stable noise reduction whilst preserving scene structure and edges. The same parameters are used for all the results on indoor and outdoor scenes presented in this work.

Step 2: The next step is to identify gradients in the image, a Sobel filter [163] with a kernel size of 3 is used to compute the horizontal and vertical image gradient components to obtain the gradient magnitude.

Step 3: To obtain markers for Watershed segmentation the gradient magnitude image is thresholded to obtain a binary image. The threshold is set as the lower quartile of the gradient

magnitude for all image pixels. This is obtained by trial and error and it removes the weak edges from the image which may introduce error in segmentation.
Step 4: The binary image is used to generate markers by detecting the hierarchical contours on the image. The contours at the bottom of the hierarchy are chosen as markers for watershed segmentation. These contours are innermost contours in the image and allows stability in region growing process of watershed. If the contours at the top of hierarchy are chosen then over-segmentation will not be obtained.
Step 5: These markers are used to perform watershed segmentation on the original image to obtain a stable over-segmentation of the image with respect to scene structure and low-frequency local appearance variation.
An example on Odzemok dataset is shown in Figure 3.8.

Mean-shift (MS) [34]: Mean-shift considers feature space as a empirical probability density function. If the input is a set of points then Mean-shift considers them as sampled from the underlying probability density function. If dense regions (or clusters) are present in the feature space, then they correspond to the mode (or local maxima) of the probability density function. For each data point, Mean-shift associates it with the nearby peak of the dataset's probability density function. For each data point, Mean-shift defines a window around it and computes the mean of the data point. Then it shifts the center of the window to the mean and repeats the algorithm till it converges. After each iteration, the window shifts to a more denser region of the dataset. There are three main parameters considered in this segmentation: Spatial resolution parameter(*SRP*) which affects the smoothing and connectivity of segments, it is chosen depending on the size of the image, Range resolution parameter (*RRP*) which affects the number of segments and the third parameter is Size of smallest segment (*S*3). The parameters are initialized automatically and assignments to each of these parameters are given below:

$$SRP = \frac{W \times H}{7.776 \times 10^4},\ RRP = \frac{W \times H}{7.776 \times 10^4},\ S3 = w_{min} * h_{min} \tag{3.3}$$

,where W and H are the width and height of input image and w_{min} and h_{min} are the minimum width and height of segmented regions which is set to approx 60×30 respectively.

The mean-shift segmentation method is based on connectedness criterion and is proved to give stable and repeatable segments for natural scenes [88]. All pixels of an image are considered as vectors in 5D consisting of spatial and color coordinates. Centroid based mode detection is employed and coordinates are ascribed modes. Recursive fusion of basins of attraction merges the modes located within a certain radius. This is an unsupervised

segmentation technique and over-segmentation is performed on image pre-processed using Bilateral filter to remove noise shown in Figure 3.8, followed by feature detection.

Simple Linear Iterative Clustering superpixels (SLIC) [6]: This segmentation technique is based on Superpixel methods and it clusters pixels in the combined five-dimensional color and image plane space to efficiently generate compact, nearly uniform superpixels with a low computational overhead. This approach generates superpixels by clustering pixels based on their color similarity and proximity in the image plane. SLIC is demonstrated to achieve good quality segmentation at a lower computational cost over state-of-the-art superpixel methods and to increase performance over pixel based methods.

The cluster centers are initialized by sampling pixels at regular grid steps. The cluster centers are perturbed in a neighbourhood, to the lowest gradient position for each cluster center. The best matching pixels are assigned from neighbourhood around the cluster center according to a distance measure. The residual error is reduced iteratively and connectivity is enforced. The segmentation requires the number of regions (S) as input and this is calculated it using the following equation in this work:

$$S = \frac{W \times H}{w_{min} \times h_{min}} \tag{3.4}$$

,where W and H are the width and height of input image and w_{min} and h_{min} are the minimum width and height of segmented regions which is set to approx 60×30 respectively to avoid very small segments as shown in Figure 3.8.

In this section SFD feature detection introduced for uniform scene coverage in the wide-baseline scene reconstruction was described in detail. Various segmentation techniques like Watershed, Mean-Shift and SLIC used for proposed feature detection were explained. Feature detection is performed on pair of multi-view images which is followed by feature matching and camera parameter estimation. The correspondences and camera parameters are used for wide-baseline sparse scene reconstruction which is explained in the following section.

3.4 Wide-baseline Scene Reconstruction

Wide-baseline correspondences are obtained for all pairs of images using SFD. These correspondences are used to reconstruct a sparse 3D representation of the scene. Figure 3.1 presents an overview of the algorithm for sparse reconstruction.

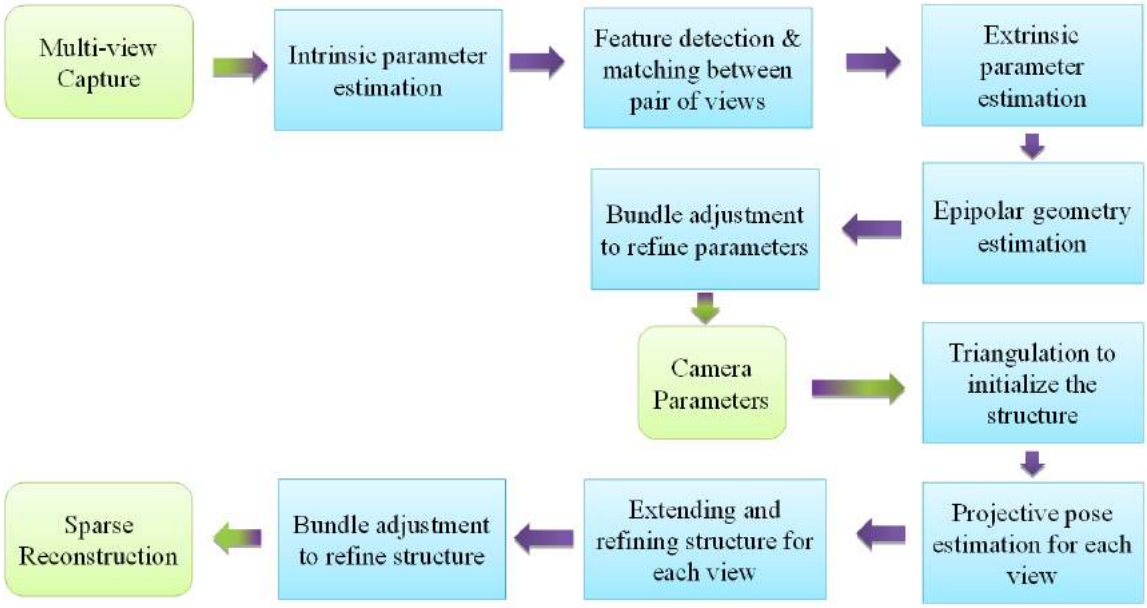


Fig. 3.9 Sparse scene reconstruction pipeline.

3.4.1 Sparse Scene Reconstruction

Sparse reconstruction is an important step for wide-baseline multi-view 3D reconstruction. It includes estimation of the camera parameters for multi-view system and 3D points for the structure using these parameters. The quality of sparse reconstruction is determined by the number and location of 3D points [69]. Existing feature detection techniques give few features leading to small number of correspondences and less 3D points. The proposed detector solves issue of limited reconstructed points and non-uniformity. The algorithm is shown in Figure 3.9. The data is captured and processed to obtain calibration and 3D structure is retrieved using the algorithm described. Both the tasks are executed simultaneously to get the final structure and parameters, but are depicted separately in the Figure 3.9 for more clear understanding.

An exhaustive matching is performed between all pairs of the multi-view images using the proposed SFD feature detector with a SIFT descriptor to get the correspondences using the feature matching technique described above. The pairs are sorted according to the highest number of matches and the first camera pair is selected from the list. Once correspondences between the two images are retrieved, the position and orientation of camera is estimated. This is described in the 3×4 Projection matrix P, which is combination of two elements: $P = [R|t]$, which are the Rotational element R and Translational element t. The P matrices can be recovered in multiple ways [69]. In this work we estimate the Fundamental matrix.

This 3×3 matrix encodes the epipolar constraint between the images: for each point x in image 1 and corresponding point x' in image 2 the following equation holds: $x'Fx = 0$.

It is proved in [69] that F can be used to infer the two projection P matrices given sufficient point matches to estimate F. F has 9 entries (but only 8 degrees of freedom), so if we have enough point pairs, we can solve for F in a least squares sense. The fundamental matrix estimation procedure employs RANSAC and the normalized 8-point algorithm [69], to find the epipolar geometry. However, [69] also point to a projective ambiguity problem with using F. This means that the recovered camera matrices may not be the "real" ones, but instead have gone through some 3D projective transformation. To cope with this, the Essential Matrix can be used instead, which holds epipolar constraint over points but for calibrated cameras. Using the Essential matrix removes the projective ambiguity and provides a Metric (or Singular) Reconstruction, which means the 3D points are true up to scaling alone, and not up to a projective transformation. The first camera is chosen as the world reference frame to obtain the camera matrix for the second camera from the fundamental matrix [69].

Once two camera matrices, P and P' are obtained, the 3D structure of the scene can be recovered. The 'optimal' triangulation method suggested by [69], optimizes the reconstruction solution based on the error from re-projection of the points back to the image plane, which is adopted by the current system. The re-projection error over the calibration and the structure parameters is minimized using bundle adjustment [162]. Bundle adjustment can be defined as the problem of simultaneously refining the 3D coordinates describing the scene geometry as well as the parameters of the relative motion and the optical characteristics of the camera(s) employed to acquire the images, according to an optimality criterion involving the corresponding image projections of all points. Bundle adjustment minimizes the re-projection error between the image locations of observed and predicted image points, which is expressed as the sum of squares of a large number of non-linear, real-valued functions. Thus, the minimization is achieved using a non-linear least-squares algorithms. Levenberg–Marquardt has proven to be one of the most successful due to its use of an effective damping strategy that lends it the ability to converge quickly from a wide range of initial guesses. By iteratively linearizing the function to be minimized in the neighbourhood of the current estimate, the Levenberg–Marquardt algorithm involves the solution of linear systems known as the normal equations.

Sparse pair-wise reconstruction results are shown for one camera pair for different datasets. The results compare the pairwise reconstruction obtained from two different feature detection approaches using SIFT and SFD.

We assume that the camera intrinsics are known to obtain metric reconstruction of the scene and camera extrinsics together with 3D point locations are estimated using the correspondences. Initialization is performed using the fundamental matrix estimation which employs RANSAC and the normalized 8-point algorithm [69], to find the epipolar geometry using the intrinsics. The first camera is chosen as the world reference frame to obtain the camera matrix for the second camera from the fundamental matrix. Then, for each image correspondence, the triangulation algorithm [69] seeks the $3D$ point that minimizes the re-projection error. After the initial pairwise sparse reconstruction is obtained for all camera pairs, a new camera is registered to the structure by finding the 2D and 3D correspondences between views and the 3D structure [69]. The view with highest correspondences is selected and pose is estimated for the view from 3D-2D point correspondences using the RANSAC algorithm. The estimated pose is used to augment the scene by triangulating the correspondences. The process is repeated for all the views until the camera pairs are exhausted. The algorithm employs global bundle adjustment [162] to minimize the re-projection error over the calibration and the structure parameters to get the optimal sparse reconstruction.

Dataset	Resolution	Number of views	Baseline	Type
Odzemok	1920×1080	8(2 moving)	15 degrees	Indoor, Dynamic
Dance1	1920×1080	7(1 moving)	15 degrees	Indoor, Dynamic
Office	1920×1080	8(all static)	15 degrees	Indoor, Dynamic
Magician	960×544	5(all moving)	40-55 degrees	Indoor, Dynamic
Rossendale	1920×1080	8(all static)	25 degrees	Outdoor, Dynamic
Cathedral	1920×1080	8(all static)	45 degrees	Outdoor, Dynamic
Patio	1920×1080	12(all static)	15 degrees	Outdoor, Dynamic
Juggler	960×544	6(all moving)	25-30 degrees	Outdoor, Dynamic
Building	800×600	119(all static)	15-30 degrees	Indoor, Static
Books	800×600	119(all static)	15-30 degrees	Indoor, Static
Cloth	800×600	119(all static)	15-30 degrees	Indoor, Static
Architecture	800×600	119(all static)	15-30 degrees	Indoor, Static
Merton	1024×768	3(all static)	10-15 degrees	Outdoor, Static
Valbonne	512×768	15(all static)	15-30 degrees	Outdoor, Static
Castle	1024×768	19(all static)	10-15 degrees	Outdoor, Static
Car	512×768	7(all static)	15-30 degrees	Outdoor, Static

Table 3.1 The characteristic properties of datasets used for evaluation.

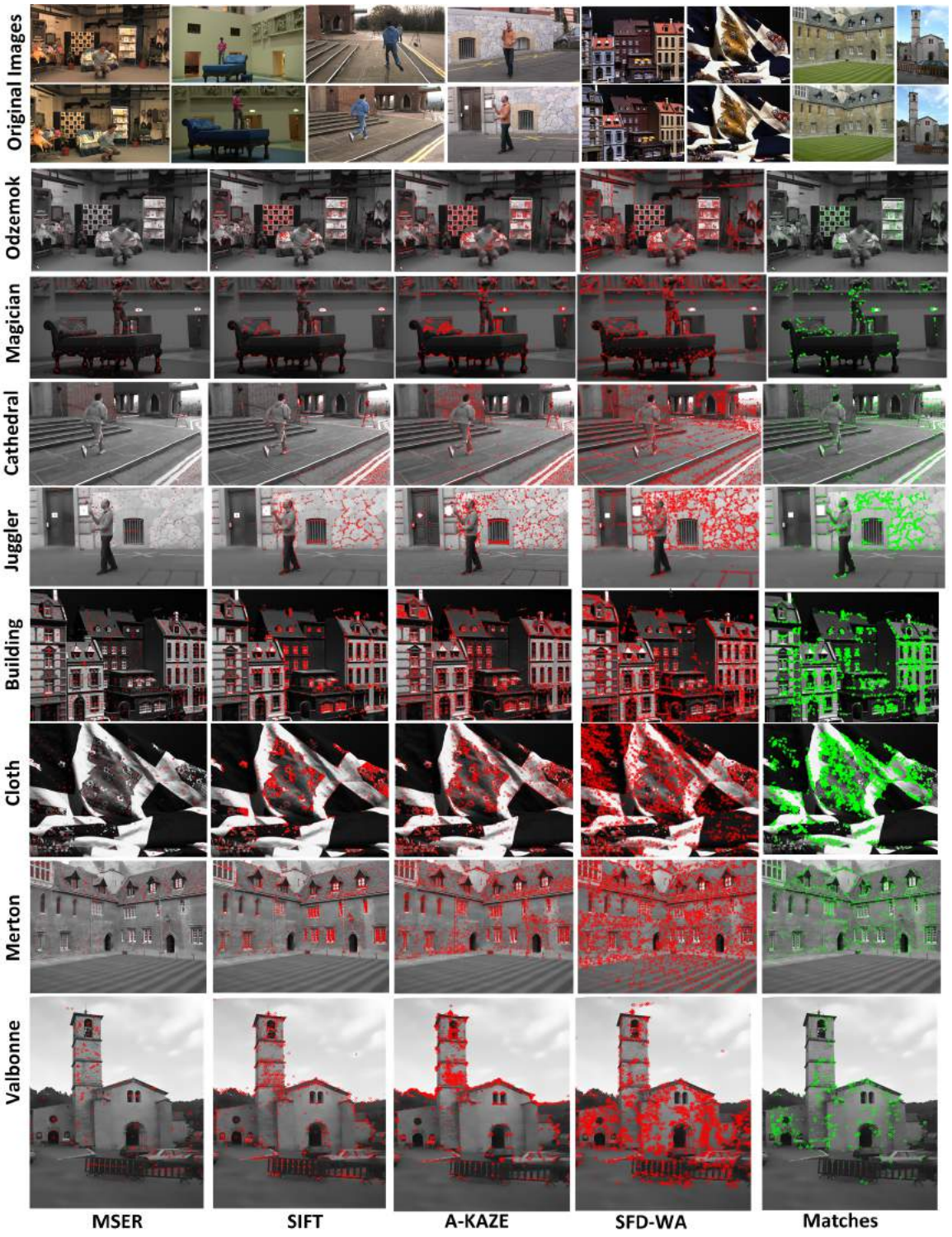


Fig. 3.10 Results for all datasets: Top two rows: Pair of images from each dataset, Bottom 8 rows: Column $1^{st} - 3^{rd}$ - Features detected on one image from each pair using MSER, SIFT and A-KAZE respectively, Column 4^{th} - Features detected by proposed SFD approach using watershed segmentation and Column 5^{th} - Features matched between pair of images using SFD features.

3.5 Results and Evaluation

This section presents evaluation of proposed SFD feature detector for different segmentation techniques Watershed, Mean-shift and SLIC against existing feature detection techniques. Extensive experimental results obtained on the standard evaluation set of [117] and on a practical wide-baseline image matching application are presented in this section.

Evaluation is performed on variety of datasets: static and dynamic; indoor and outdoor scenes. State-of-the-art feature detection techniques have used the static indoor and outdoor datasets in their evaluation, hence these datasets are included in our evaluation for fair comparison. Various benchmark dynamic datasets have been included to emphasize the importance of SFD in wide-baseline dynamic scene reconstruction. Wide-baseline image/video datasets (15-45 degree angle between adjacent cameras) of natural indoor and outdoor scenes under variable lighting are:
Static indoor datasets [5]: Building, Books, Cloth, Architecture.Challenges: Variable lighting and viewpoints.

Static outdoor datasets: Merton CollegeI [1], Valbonne[4], Castle[4], Car[4]. Challenges: Repetitive background, varying lighting condition.

Dynamic indoor datasets: Odzemok[2] , Dance1[2], Office[2], , Magician[3]. Challenges: Both scattered and uniform background, stable lighting condition, single and multiple objects. Magician is captured with only hand-held cameras.

Dynamic outdoor datasets: Rossendale[2], Cathedral[2], Patio[2], Juggler[3]. Challenges: Both scattered and uniform background, repetitive background, variation in illumination. Juggler is captured with only hand-held cameras.

The characteristics of datasets are presented in Table 3.1. Feature detections using default parameters for MSER, SIFT and A-KAZE feature detectors are shown in Figure 3.10 against SFD based on Watershed segmentation.

[1] http://www.robots.ox.ac.uk/ vgg/data/
[2] http://cvssp.org/data/cvssp3d/
[3] http://www.inf.ethz.ch/personal/lballan/datasets.html

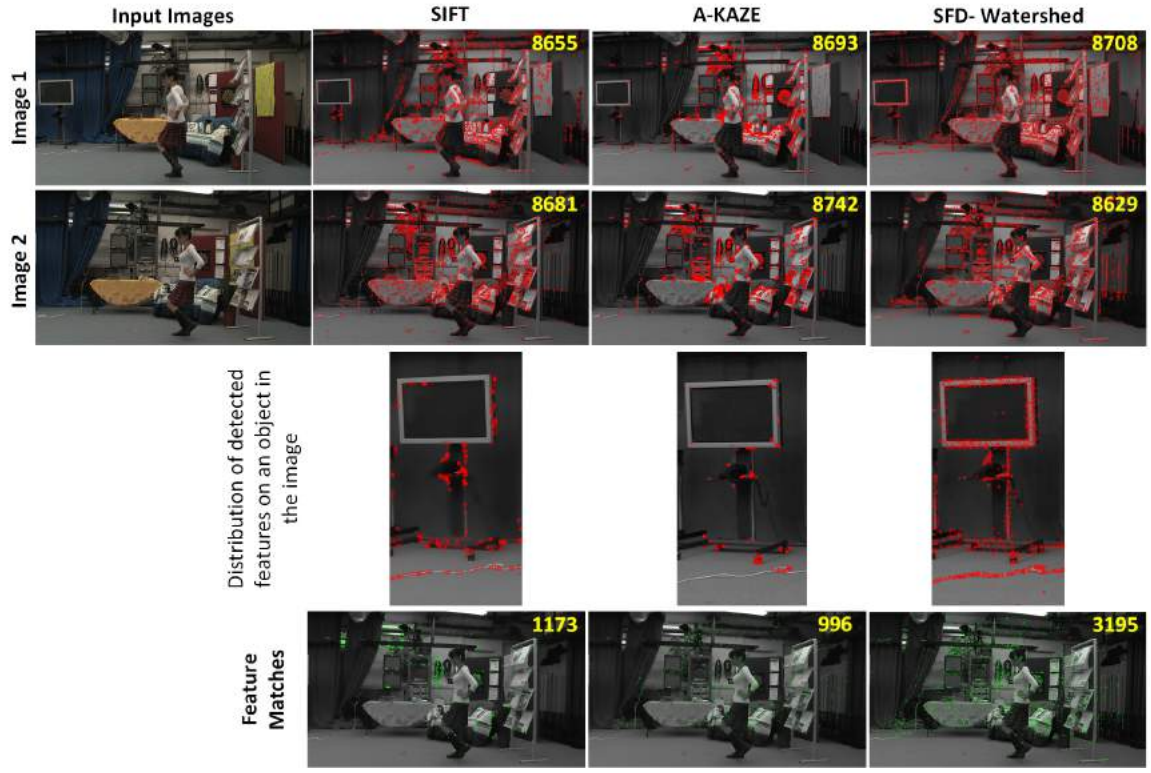


Fig. 3.11 Results for Dance1 dataset: Features detected on pair of images using SIFT, A-KAZE and SFD approach using watershed segmentation.

3.5.1 Evaluation Criteria

The SFD feature detector is evaluated based on the properties of good features described in [179]: quantity; efficiency; accuracy; repeatability; and coverage of the proposed SFD against the state-of-the-art detectors (FAST, BRIEF, Harris, SIFT, GFTT, MSER, ORB, STAR, SURF, KAZE, A-KAZE). The accuracy and repeatability is evaluated in Section 3.5.2 and quantity, coverage and efficiency is evaluated in Section 3.5.3. For SIFT, SURF, STAR, ORB, FAST, MSER, Harris and GFTT we use the OpenCV based implementation. For KAZE and A-KAZE the implementation available from the paper [10, 11] is used. The feature detection thresholds of the different methods are set to proper values to detect approximately the same number of features per image. Comparison of SFD features against SIFT and A-KAZE feature detectors is shown in Figure 3.11. A region is highlighted in the image where SFD gives uniform coverage of features as compared to existing state-of-the-art methods.

3.5.2 Feature Detection and Matching Accuracy Test

This section evaluates performance of different segmentation techniques for SFD. Adjacent pairs of images are taken from each dataset and segmentation is performed using Watershed, Mean-Shift and SLIC giving three variants SFD-WA, SFD-MS and SFD-SLIC respectively. The proposed SFD detection is performed on each pair of images for each segmentation method followed by feature matching using a SIFT descriptor. An exact nearest-neighbour matching algorithm followed by a ratio test as explained in [105] is used to evaluate the feature detector. All of the matches whose distance ratio is greater than 0.85 are rejected, which eliminates 90% of false matches and 5% of the correct matches [105]. After obtaining a set of refined matches, a left-right symmetry test is used to further remove inconsistent matches due to repeated patterns. This is followed by RANSAC based refinement [143] of matches without prior knowledge of camera parameters. The fundamental matrix is estimated using RANSAC and the inliers are chosen as the set of matches.

Segmentation	Watershed			Mean-Shift			SLIC		
Dataset	$\|F^*\|$	TC	RC	$\|F^*\|$	TC	RC	$\|F^*\|$	TC	RC
Odzemok	8169	6543	3717	7908	5913	3547	8093	7812	**4921**
Dance1	8242	6372	3394	7956	5652	3087	8305	7499	**4459**
Office	8074	6501	3508	7822	5908	3321	7998	7667	**4768**
Magician	7921	5057	2844	7878	4524	2698	7909	6629	**3066**
Rossendale	6576	4528	2332	6349	4075	2118	6542	4786	**2972**
Cathedral	7806	6324	3452	7747	6450	3601	7983	6161	**3882**
Patio	7207	5215	3270	7156	5309	3431	7176	5642	**3986**
Juggler	5231	4478	2342	5196	4563	2267	5435	4657	**2878**
Building	4981	3531	1983	4809	3467	2067	4943	3791	**2240**
Books	4877	3348	2019	4796	3259	1984	4814	3429	**2319**
Cloth	4532	3424	1732	4321	3349	1689	4559	3568	**1981**
Architecture	4790	3213	1654	4683	3563	1780	4897	3664	**2091**
Merton	9947	7644	4533	9817	6899	4485	10336	8899	**5920**
Valbonne	3251	2715	1135	2939	2217	1252	3065	2952	**1981**
Castle	5674	4043	2351	5547	4146	2474	5848	4420	**2559**
Car	7764	4915	3435	7939	5017	3552	8065	5152	**3975**

Table 3.2 Evaluation of feature detection and matching of SFD for three different segmentation approaches (best highlighted in bold): F^* shows the number of features detected, Total count (TC) is the number of matches obtained with brute force matching using a SIFT descriptor and RANSAC count (RC) is the number of correspondences that are consistent with the RANSAC based refinement.

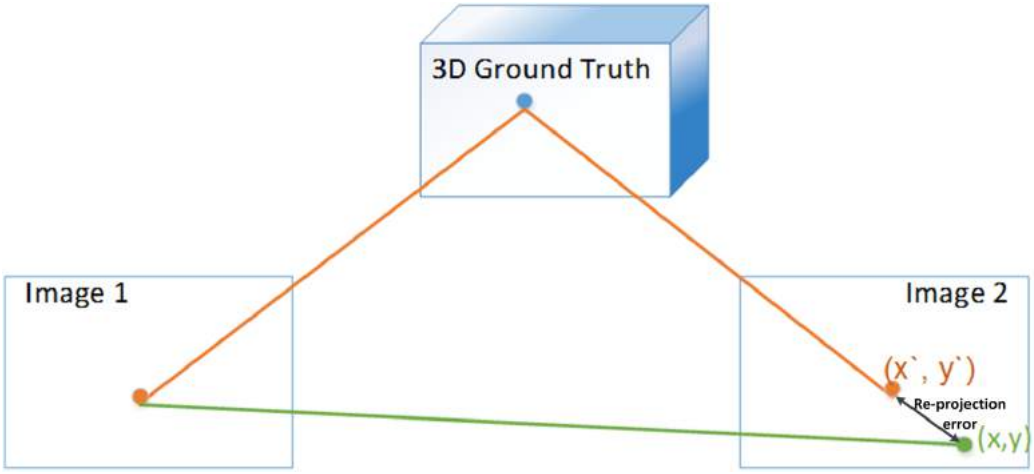


Fig. 3.12 Re-projection error illustration

Experimental results for a pair of image for each dataset and all segmentation methods (WA, MS and SLIC) are summarized in Table 3.2. The column headed '$|F^*|$' shows the number of features detected in one of the images. Total count (TC) is the number of matches obtained with brute force matching using a SIFT descriptor and RANSAC count (RC) is the number of correspondences that are consistent with the RANSAC based refinement performed after the ratio and symmetry tests. The number of features detected by all segmentation techniques are similar. The numbers of matches reduces by $30-40\%$ after refinement using the symmetry and RANSAC tests (RC).

Inlier ratio of the feature matches is defined as the ratio of the number of correct matches (RC) total number of matches (TC). It is a measure of the reliability of a feature detector and has been used in literature for evaluation extensively [117]. We calculate the inlier ratio from table 3.2 for all the datasets. As noticed the inlier ratio is highest for SFD-SLIC segmentation compared to MS and WA.

Matching accuracy evaluation: For further evaluation this section compares the feature matching accuracy of SFD against MSER, SIFT and A-KAZE. We choose only these detectors because they outperform the other feature detection methods for wide-baseline matching. SFD-WA is chosen as our base segmentation technique because of its computational efficiency. The aim is to evaluate the accuracy of each feature correspondence. The ground-truth camera calibration and reconstruction obtained using this camera calibration is used for evaluation of the accuracy of the feature matches for all datasets. Ground-truth camera calibration with both intrinsic and extrinsic parameters is known for all the datasets. The ground-truth reconstruction is available for static indoor datasets and for other datasets the reconstruction is computed using existing reconstruction algorithms using the ground-truth

	MSER		SIFT		A-KAZE		SFD-WA	
Dataset	RC	MRE	RC	MRE	RC	MRE	RC	MRE
Odzemok	119	1.390	1269	**1.175**	1209	1.181	3717	1.351
Dance1	102	1.362	1109	1.231	1081	**1.214**	3394	1.251
Office	111	1.431	1203	1.403	1143	1.361	3508	**1.354**
Magician	72	1.255	786	1.104	764	**1.045**	2844	1.195
Rossen.	26	1.411	237	1.323	516	1.318	2332	**1.315**
Cathedral	24	1.386	969	**1.152**	1158	1.154	3452	1.179
Patio	24	1.396	832	1.223	1207	**1.212**	3270	1.256
Juggler	28	1.298	208	1.155	686	**1.110**	2342	1.237
Building	57	1.240	629	**1.103**	689	1.120	1983	1.221
Books	72	1.314	783	1.210	832	1.223	2019	**1.207**
Cloth	40	1.211	636	**1.098**	845	1.201	1732	1.159
Archi.	49	1.273	719	**1.192**	905	1.206	1654	1.211
Merton	81	1.255	1272	1.177	1509	1.196	4533	**1.175**
Valbonne	41	1.258	636	1.181	755	1.187	1135	**1.159**
Castle	67	1.318	695	1.171	835	**1.152**	2351	1.208
Car	65	1.330	1018	1.213	1207	1.200	3435	**1.183**

Table 3.3 Evaluation of matching accuracy of SFD against MSER, SIFT and A-KAZE using the ground-truth reconstruction and camera calibration.

calibration information [62]. The accuracy is evaluated using the projection of a 3D point which gives the ground-truth match for a point in pair of images. The error between the ground-truth match and the match obtained using the different feature detection approaches gives the measure of accuracy.

Ground-truth correspondences are obtained by back-projecting the $3D$ location of the feature points detected in one image to the other image and evaluating the distance to the estimated feature match. Mean re-projection error (MRE) given in Equation 3.5 is used for accuracy evaluation of the estimated SFD feature matches again the ground-truth and the re-projection error is illustrated in Figure 3.12.

$$MRE = \frac{1}{(K+1)} \sum_{0}^{K} \sqrt{(x-x')^2 + (y-y')^2} \tag{3.5}$$

where (x,y) is the estimated SFD feature match, (x',y') is the re-projected point, and K is the number of feature matches, here $K = RC$, where RC is the RANSAC count depicting the final number of matches. Table 3.3 presents the results of the ground-truth correspondence for the proposed SFD using Watershed segmentation with a SIFT descriptor for matching and three other detector-descriptor combinations representing state-of-the-art detectors (MSER,

SIFT and A-KAZE). *RC* shows the number of correspondences obtained with each approach after symmetry and RANSAC consistency tests. The number of matches obtained with the proposed SFD feature detector is greater by an order of magnitude than MSER, and by a factor three greater than SIFT and A-KAZE. The *MRE* for SFD is lower compared to MSER and comparable with SIFT and A-KAZE feature detectors within approx. ± 0.2 pixels.

MRE gives an overall comparison of the accuracy of the feature matches, however the distribution of the matches at different pixel errors is not clear. Although the *MRE* of SFD is comparable with existing feature detectors, it is noted that SFD gives large number of matches and it would be interesting to see the comparison of number of matches at each pixel error against existing detectors. The more the number of matches at lower pixel error the better the accuracy of the feature detector. Hence to evaluate this re-projection error is calculated using the ground-truth reconstruction for each feature match. The errors are ranked from low to high and a graph is plotted for the number of feature matches at each pixel error.

The comparative evaluation of the re-projection errors for all the correspondences obtained by SIFT, A-KAZE and SFD is plotted and results for 2 datasets from each category are shown in Figure 3.13. Figure 3.13 shows that the number of wide-baseline matches for a given maximum re-projection error are consistently greater for SFD detection than for SIFT and A-KAZE. Approximately three times more points have less than 1 pixel error for SFD compared to SIFT and A-KAZE depicting the relatively high accuracy of the proposed method. This implies that taking the best N features from SFD will give higher accuracy calibration/reconstruction than for SIFT feature detection. Therefore SFD gives more accurate geometry estimation from wide-baseline views due to the improved accuracy of feature localization demonstrating the suitability of SFD for sparse 3D scene reconstruction.

Repeatability: Repeatability of the features is an important property which indicates whether or not the same feature will be detected in two or more different images of the same scene and it is generally related to Invariance and Robustness. Existing methods model the invariance mathematically and the detected features were designed to be unaffected by these transformations. The feature detection methods are made less sensitive to small deformations to increase the robustness. Typical variations that are tackled using robustness are image noise, discretization effects, compression artefacts, blur, etc.

In the case of the proposed SFD detection repeatability is ensured by the initial segmentation. When images undergo transformations or deformations the segmentation approaches are able to retrieve consistent edges in the images. The change is viewpoint or lighting causes a slight variance in the strength of the edges but the points of intersection remain robust to these deformations and transformations.

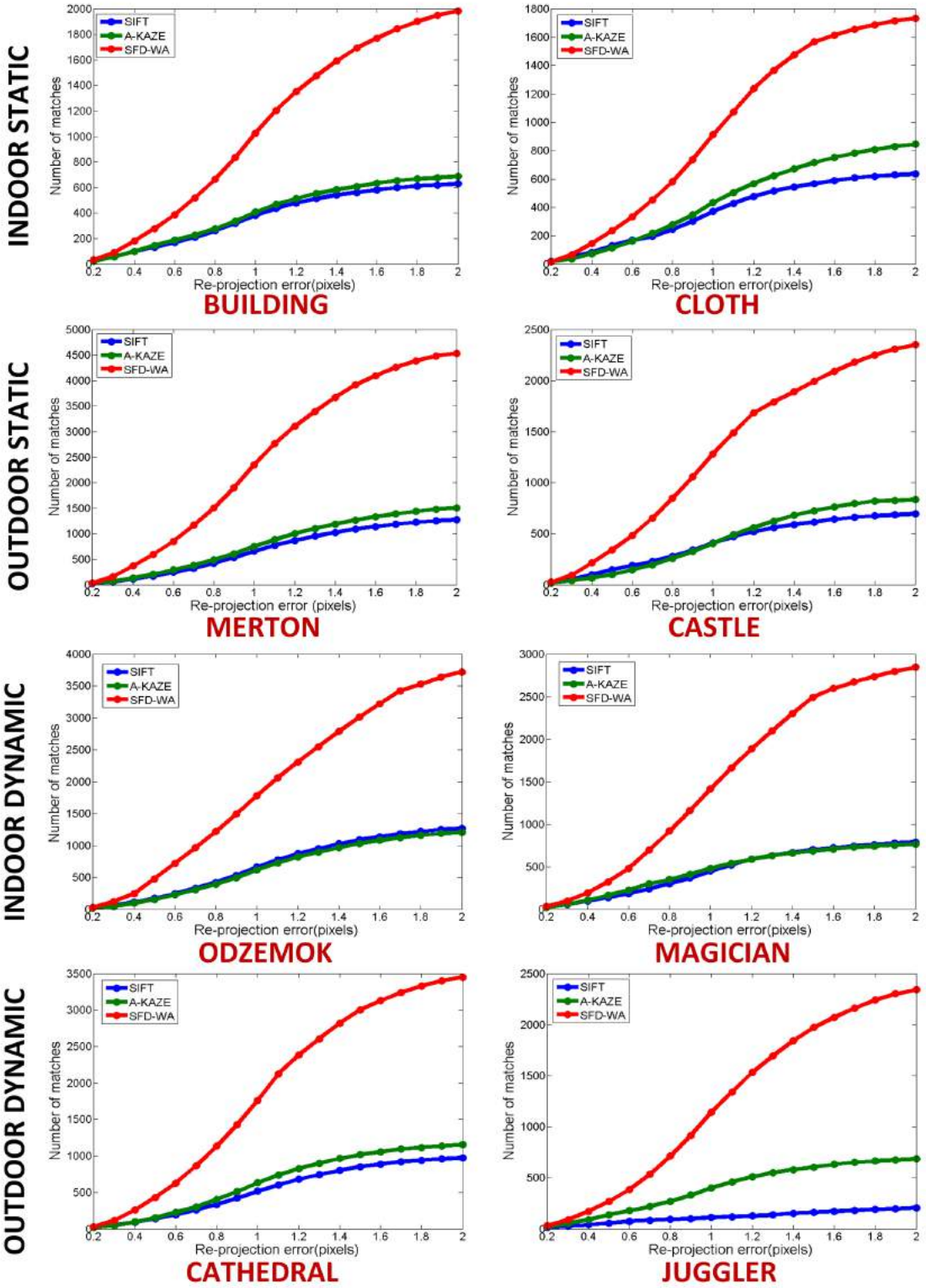


Fig. 3.13 Accuracy results for dynamic datasets: Re-projection error cumulative distribution of SIFT, A-KAZE and SFD-WA

We measure the repeatability (R) of SFD, defined as $R = \frac{\text{Correct Matches}}{\text{RC}}$ using the ground-truth information for Odzemok dataset. We eliminate the matches from RC with MRE greater that 2.5 pixels to obtain the 'Correct Matches', which is a standard setting to allow noise variance [69]. The comparisons with FAST, MSER, ORB, SIFT and A-KAZE are shown in Figure 3.14 for dynamic and in Figure 3.15 for static indoor and outdoor datasets. The assignments in the figure are: A: FAST-BRIEF, B: Harris-SIFT, C: GFTT-SIFT, D: MSER-SIFT, E: ORB-ORB, F: STAR-BRIEF, G: BRIEF-BRIEF, H: SIFT-SIFT, I: SURF-SURF, J: KAZE, K: A-KAZE, L: SFD-WA-BRIEF, M: SFD-WA-SIFT, N: SFD-MS-BRIEF, O: SFD-MS-SIFT, P: SFD-SLIC-BRIEF and Q: SFD-SLIC-SIFT. On left repeatability is shown between testing images 1-2, 1-3, ..., 1-7 with baseline 15-120 degrees and on the right for adjacent image pairs with baseline 15-30 degrees.

The repeatability of SIFT, A-KAZE and SFD detector is comparable and greater than other detectors like FAST, ORB and MSER. Watershed segmentation performed consistently better than other segmentation methods. As the baseline between the image pairs increases, the overlap between the images reduce which results in decrease in the number of matches (RC). It is noted that the repeatability for each feature detector reduces with the increase in baseline which indicates a drop in the percentage of correct matches from the set of matches (RC). The drop in the repeatability is similar for SFD, SIFT, A-KAZE and MSER. However, the percentage of correct matches for SFD is slightly higher than existing approaches, as seen from the Figure. The FAST and ORB detectors does not perform well for wide-baseline images.

Evaluation of SFD vs. Harris/High Frequency/Uniform Sampling: The proposed SFD feature detector results in an increased number of features against previous detectors designed for wide-baseline matching applications. Alternative approaches to increase the number and coverage of feature detections could be use of corner detectors such as Harris or uniform grid sampling. Evaluation of the performance of SFD vs. Harris/High Frequency/Uniform Sampling is presented in Table 3.4. For this comparison the threshold for the Harris detector, threshold for high frequency detections and the resolution for uniform grid resolution are set to give a similar number of features to SFD. Uniform grid sampling is performed by locating features at points of maximum gradient magnitude with a 13×13 grid resolution. The SIFT descriptor is used for all feature matching. Results presented in Table 3.4 show that the proposed SFD approach significantly outperforms the Uniform and Harris feature detectors after similarity and RANSAC tests are applied. This shows that the SFD approach detects stable features across wide-baseline views.

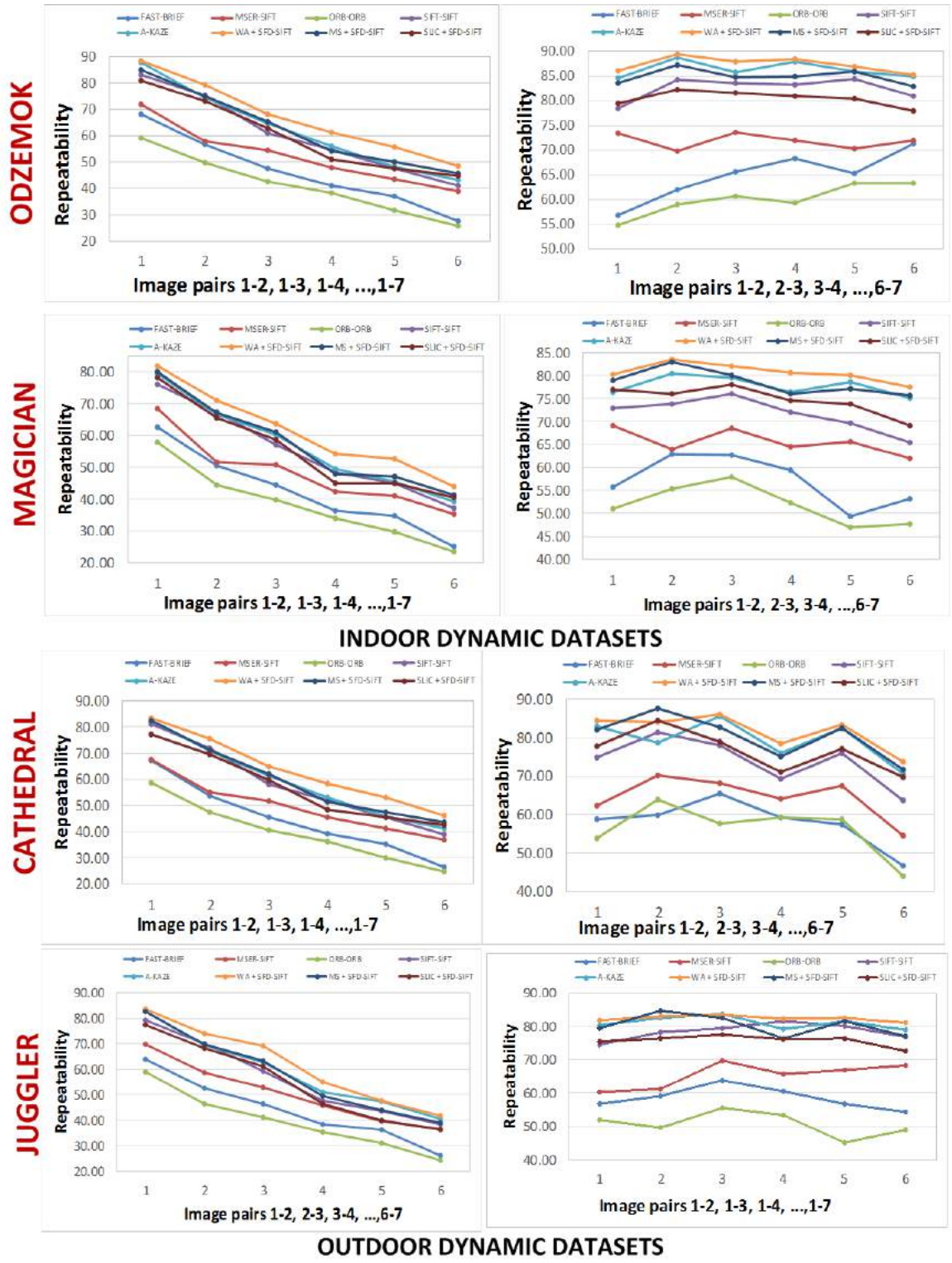


Fig. 3.14 Repeatability results for dynamic datasets: Left: Repeatability comparison for matching of camera 1 to all other views (15-120 degree baseline); and Right: Repeatability comparison for matching between adjacent views (15-30 degree baseline).

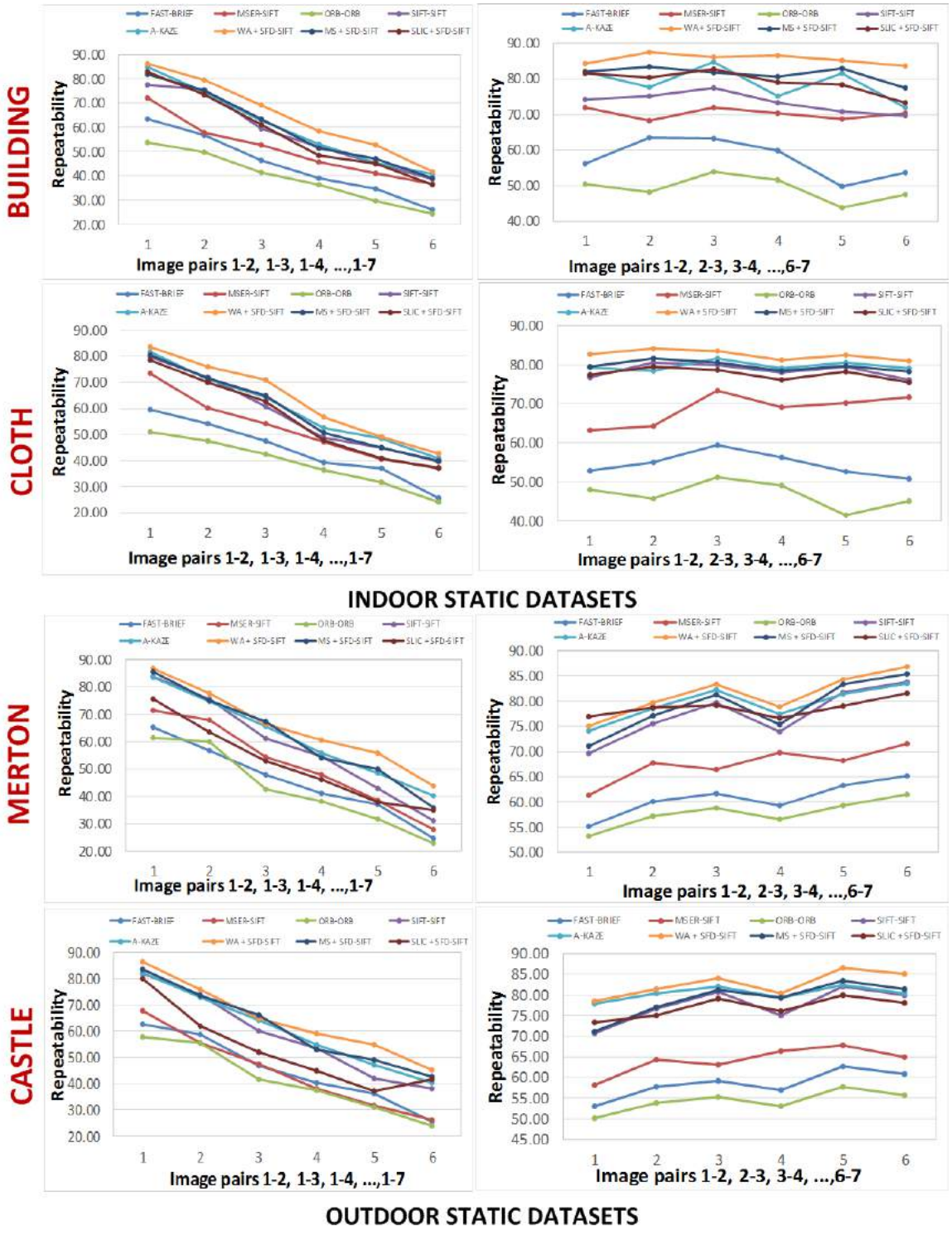


Fig. 3.15 Repeatability results for static datasets: Left: Repeatability comparison for matching of camera 1 to all other views (15-120 degree baseline); and Right: Repeatability comparison for matching between adjacent views (15-30 degree baseline).

Feature Detector	Descriptor	Features	RC
Uniform Sampling	SIFT	5508	78
High Frequency	SIFT	5582	143
Harris	SIFT	5542	155
SFD	SIFT	5533	2342

Table 3.4 Evaluation of feature matching performance of SFD vs. high frequency and dense feature sampling on Juggler dataset.

3.5.3 Benchmark Evaluation of Detector-Descriptor

To evaluate the performance of the proposed segmentation based feature detection approach for wide-baseline matching we present a comprehensive comparison with existing state-of-the-art feature detector and descriptor combinations. Comparison is performed with binary (FAST [11], ORB [150]), BRIEF [29] and floating point (Harris [68], GFTT [157], SIFT [105], SURF [18], STAR [8], MSER [109]), KAZE [10], A-KAZE [11] detectors. These detectors are combined with feature descriptors (BRIEF [29], ORB [150], SIFT [105], SURF [18]). Detectors and descriptors are used with default parameters. Figure 3.16 presents the evaluation results for each detector-descriptor combination for wide-baseline matching on all datasets. The left column presents the number of correct matches (*RC*) obtained after similarity and RANSAC tests.

Performance of the proposed SFD detector combined with WA, MS and SLIC segmentation techniques with BRIEF and SIFT descriptors is shown in bars labelled L - Q, respectively demonstrating that the approach consistently achieves a factor $3-10$ increase in the number of correct matches compared to previous detector-descriptor combinations. The right column of Figure 3.17 presents the average computation time/frame showing that the computational time is less than floating point detectors and similar to binary detectors. SFD-WA is the fastest detector compared to MS and SLIC, but the number of correct matches are highest for SLIC. MS gives lower number of correct matches compared to both WA and SLIC. The evaluation shows a trade-off between the performance and the number of correct matches for various segmentation techniques.

Scene Coverage: The distribution of the features across the scene is shown in Figure 3.10 for different detectors: Proposed SFD with WA, SIFT, MSER. Both the quantity and distribution of features for SFD give improved scene coverage for all the datasets.

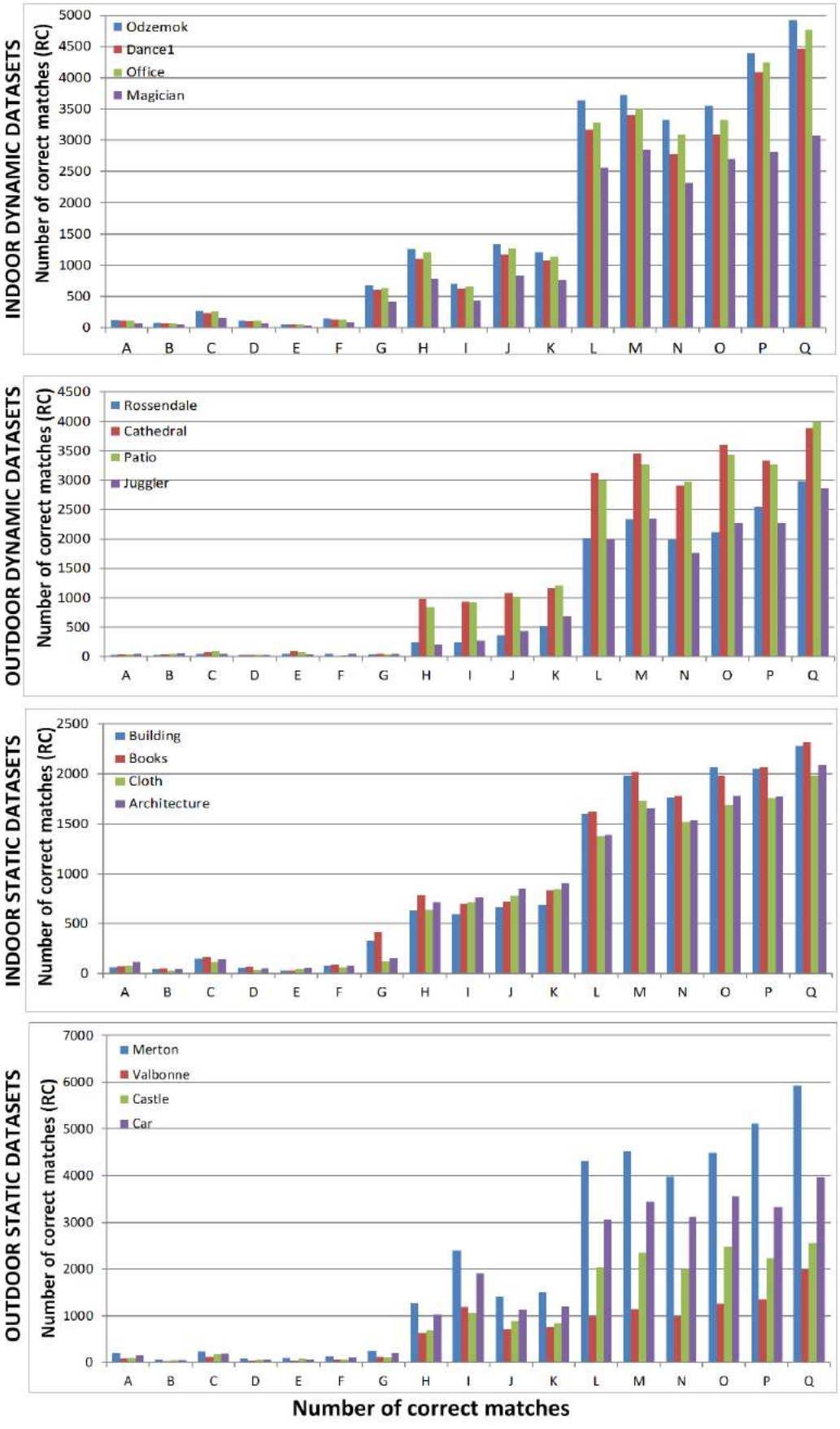


Fig. 3.16 Evaluation of number of correct matches on all datasets.

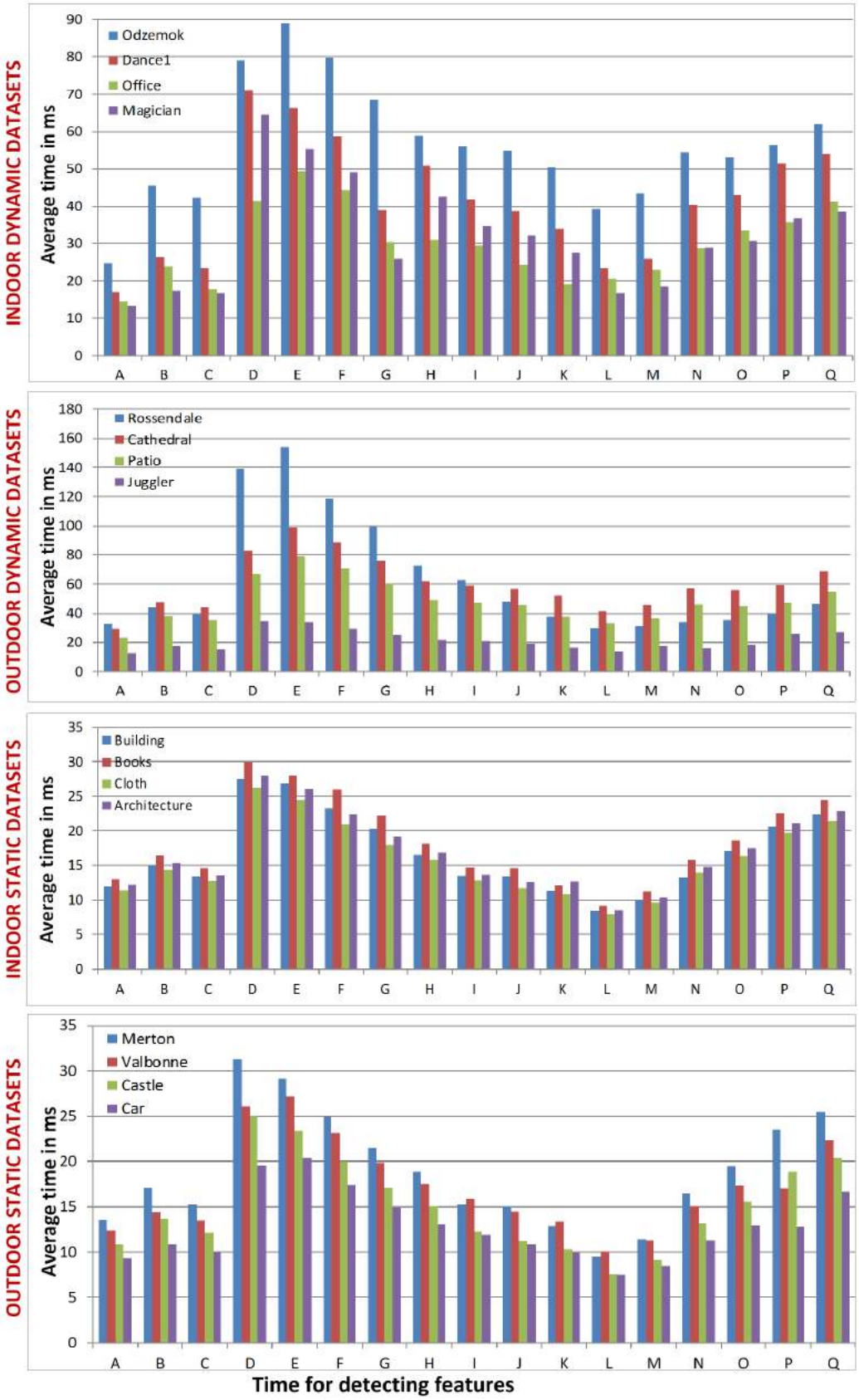


Fig. 3.17 Evaluation of time for detecting features on a wide-baseline stereo pair for each sequence in *ms* for all datasets.

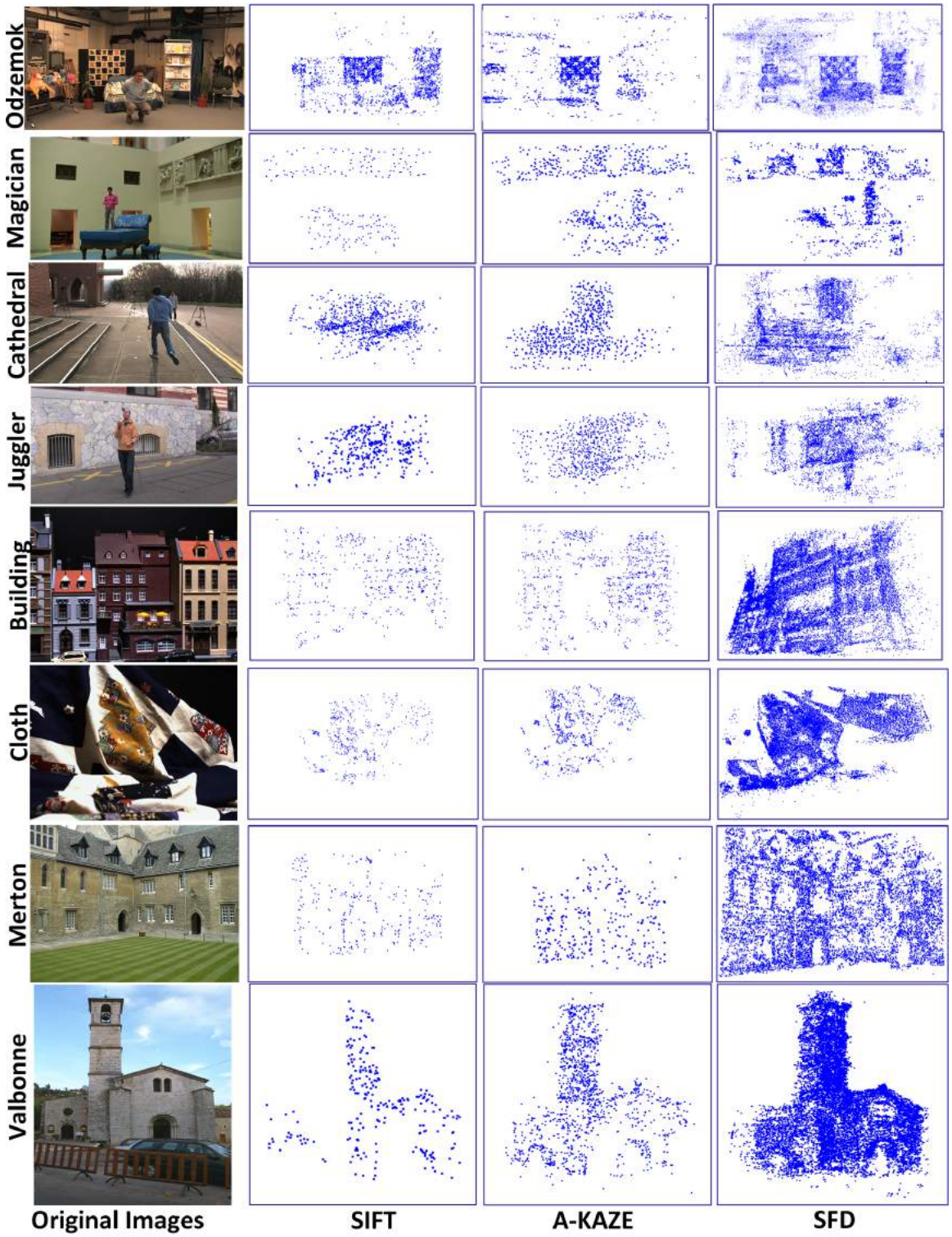


Fig. 3.18 Results of multi-view sparse reconstruction for all datasets for SIFT, A-KAZE and SFD-WA .

Dataset	MSER	SIFT	A-KAZE	SFD-WA	SFD-MS	SFD-SLIC
Odzemok	171	1884	4025	12385	9087	**14515**
Dance1	153	1652	3599	**13603**	8026	11302
Office	165	1792	3806	11681	8034	**14109**
Magician	128	1171	2544	**9470**	7014	9360
Rossen.	58	526	1145	2213	1017	**3983**
Cathedral	72	2153	2570	10840	9733	**12895**
Patio	61	1847	2679	**8845**	7261	7259
Juggler	67	461	1522	7211	6501	**8102**
Building	249	2788	2984	**8606**	6939	7660
Books	312	3398	3610	**8762**	7009	8610
Cloth	175	2760	3667	7516	5958	**7634**
Archi.	210	3107	3929	**7725**	6019	7178
Merton	316	2760	3274	9619	8118	**10965**
Valbonne	258	1380	1637	4084	3369	**5121**
Castle	261	1508	1811	**5368**	4333	4854
Car	252	2208	2619	**7705**	6755	7184

Table 3.5 Evaluation of the number of sparse 3D points from pair-wise reconstruction.

3.5.4 Application to Wide-baseline Reconstruction

Wide-baseline sparse scene reconstructions are presented for all the datasets in Figure 3.18. Reconstructions obtained using the proposed SFD features are compared with those obtained using the SIFT and A-KAZE detectors, in all cases the SIFT descriptor is used for matching. As expected from the evaluation of wide-baseline matching presented above the number of reconstructed points is much higher with the proposed approach as shown in Table 3.5 with WA, MS and SLIC. From Figure 3.18 it can be observed that sparse wide-baseline reconstruction based on SFD gives a significantly more complete representation of the scene (evaluation of the accuracy against ground-truth reconstruction for all datasets was presented in Table 3.3). The feature detection thresholds of the different methods are set to detect approximately the same number of features per image initially. However, the number of feature matches and sparse 3D points is much lower for SIFT and A-KAZE compared to SFD, showing the stability of SFD feature points. Hence, the SFD based dense reconstruction gives more complete coverage of scene compared to other detectors.

3.6 Limitations

Evaluation has been performed across a wide-variety of indoor and outdoor scenes to identify the limitations of SFD feature detection in the context of wide-baseline matching. As with other feature detection approaches the method is dependent on variation in surface appearance and consequently will produce fewer and less reliable features in areas of uniform appearance, or repetitive background texture like trees, sky etc. However, as demonstrated in the evaluation SFD increases the number of features and scene coverage for wide-baseline matching compared to previous feature detection approaches.

3.7 Conclusion

In this chapter we have proposed a novel feature detector for wide-baseline matching to support 3D scene reconstruction. The approach is based on over-segmentation of the scene and detecting features at intersections of three or more region boundaries. This approach is demonstrated to give stable feature detection across wide-baseline views with an increased number of features and more complete scene coverage than popular feature detectors used in wide-baseline applications. SFD is shown to give consistent performance for different segmentation approaches (Watershed, Mean-shift, SLIC), with SFD-SLIC giving a marginally higher number of features. The speed of SFD feature detection is comparable to other methods for wide-baseline matching.

A comprehensive performance evaluation against previous feature detectors (Harris, SIFT, SURF, FAST, ORB, MSER) in combination with widely used feature descriptors (SIFT, BRIEF, ORB, SURF) demonstrates that the proposed segmentation based feature detector SFD achieves a factor $3-10$ times more wide-baseline feature matches for a variety of indoor and outdoor scenes. Quantitative evaluation of SFD vs. SIFT feature detection shows that for a given error level SFD gives a significantly larger number of features. Improved accuracy in feature localization with SFD results in more accurate camera calibration and reconstruction of sparse scene geometry.

Application to stereo reconstruction from wide-baseline camera views demonstrates that the SFD feature detector combined with a SIFT descriptor achieves a significant increase in the number or reconstructed points and more complete scene coverage than existing detectors. The sparse scene reconstruction obtained from SFD keypoints is used to initialize our framework for general dynamic scene reconstruction. Large number of reliable sparse features in the scene gives us information about various objects in the scene, also large number of points provide detailed information on each object in the scene including dynamic

objects at each time instant. The increase in the scene coverage and the number of features in the wide-baseline sparse reconstruction using SFD compared to existing feature detectors provides for a better initialization for dense dynamic scene reconstruction presented in Chapter 4.

Chapter 4

Dense Reconstruction of Real-world Dynamic Scenes

4.1 Introduction

In this chapter we address the problem of dense surface reconstruction for real-world dynamic scenes without prior knowledge or assumptions on scene structure or background. The segmentation based feature detector SFD provides an initial sparse scene reconstruction with good feature coverage for both static and dynamic elements in the scene. This avoids the requirement for prior knowledge of the background and static backgrounds commonly assumed in previous work. The problem is then to obtain a dense surface reconstruction with accurate object segmentation given the sparse initialization. A large number of uniformly distributed sparse features in the 3D reconstructions leads to better initialization for dense reconstruction. Uniformly distributed large number of points are required on the moving objects in the scene to obtain initial coarse reconstruction from the correspondences of detected points. Salient object identification is performed by clustering the sparse features in the 3D space to obtain the initial coarse reconstruction. This is then refined for each object through joint optimization of shape and segmentation using a robust cost function for multi-view dynamic dense scene reconstruction explained in this chapter.

Recent research in multi-view dynamic scene reconstruction techniques has been applied to less controlled outdoor scenes. Initial research focused on reconstruction in sports [62] exploiting known background images or the pitch color to obtain an initial segmentation. Visual-hull based reconstruction was performed with known prior foreground/background, which is achieved using a uniform chroma-key color background or background image plate. Alternatively, multi-view stereo techniques have been developed which require a relatively

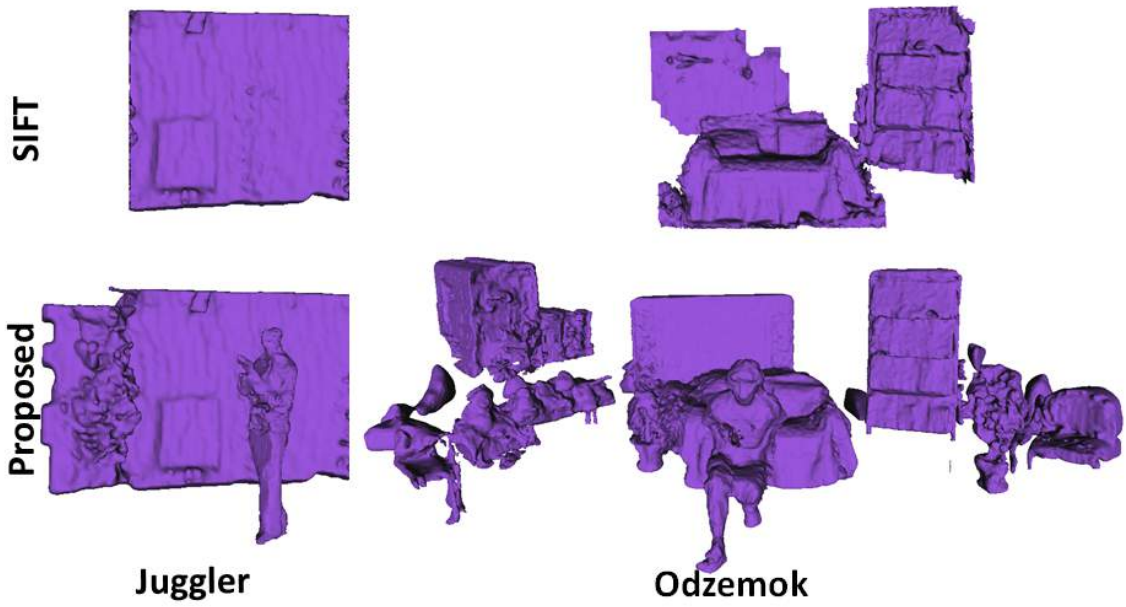


Fig. 4.1 Dense reconstruction results for Odzemok and Juggler datasets using SIFT and SFD feature detectors.

dense camera network resulting in large numbers of cameras. Extension to more general outdoor scenes [15, 82, 170] 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. These approaches to general dynamic scene reconstruction fail in case of complex (cluttered) scenes captured with moving cameras in absence of any priors.

This chapter 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. 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, calibrated and the 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 input is a sparse set of synchronized multi-view videos without any foreground segmentation. SFD features presented in Chapter 3 for wide-baseline matching are detected between pairs of views at each frame across the sequence for more complete coverage of the scene as compared to SIFT as shown in Figure 4.1 for Juggler and Odzemok dataset.

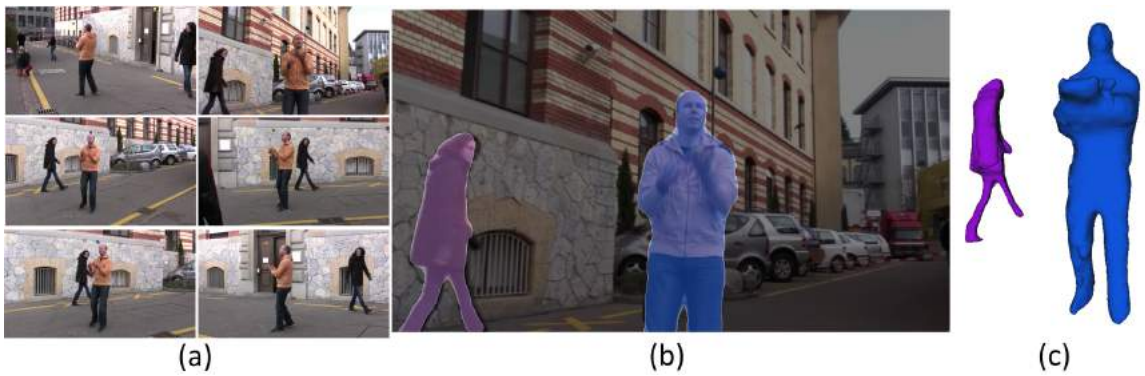


Fig. 4.2 General dynamic scene reconstruction (a) Multi-view frames for Juggler dataset, (b) Segmentation of dynamic objects and (c) Reconstructed mesh

The number of objects obtained from the sparse features in the final mesh reconstruction of SFD is higher than SIFT. Hence, the SFD based dense reconstruction gives more complete coverage of scene compared to SIFT. The scene structure is estimated automatically using the SFD correspondences. An initial coarse reconstruction and segmentation of 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. View-dependent optimization of depth is performed with respect to each camera which is robust to errors in camera calibration and initialization to obtain dense reconstruction. 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 [15] captured with a network of moving hand-held cameras are shown in Figure 4.2. 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 refinement of dense reconstruction and segmentation integrating error tolerant photo-consistency and edge information.

4.2 Related Work

Dense dynamic scene reconstruction is a challenging task as there are number of factors which need to be handled like errors in calibration, motion blur, cluttered background, articulated and non-rigid motion of multiple people, resolution differences between camera views, wide-baselines and non-uniform dynamic backgrounds. Segmentation of dynamic objects from such scenes is difficult because of background complexity and the likelihood of overlapping background and foreground appearance distributions. Reconstruction is also challenging due to limited visual cues and relatively large errors affecting both calibration and extraction of a globally consistent surface reconstruction.

Initial research focused on narrow-baseline stereo [96, 99] 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 [61, 62, 82, 190] to allow visual hull reconstruction. The unavailability of the visual hull discards many of the top-performing multi-view stereo algorithms. Other techniques employ optical flow based techniques which compute the displacement of each pixel between two images however such techniques usually make no assumptions on the existence of global geometric constraints, such as the epipolar geometry, and consequently the search region becomes 2D not 1D which makes the problem more ill-conditioned. There are also relatively new methods to come up with dense reconstructions, such as [43] describe a variational method where a surface is evolved with level set based PDE's to come up with the best 3D representation. Another type of method which also discretized the 3D space is called space or voxel carving, which typically assumes a hypothetical 3D grid in the bounding box of the viewed object and conducts a 3D search for the best possible surface representation. A survey of such volumetric methods has been made by [160]. Another interesting thread of research is from [167] where probabilistic methods are deployed to generate 3D depth from multiple wide-baseline views. Imaging, occlusions, and outliers are modeled with generative models and the most likely model is inferred in an Expectation-Maximization framework. But these approaches do not work for cluttered challenging outdoor scenes. Variational methods get stuck into local minima, unless they provide a way of estimating a close and reliable initial guess that takes visibility into account. In contrast to these methods, Furukawa et al [51] proposed a very accurate dense reconstruction of cluttered challenging scenes by using the 3D points as the initialization which is obtained using SIFT features for static scene.

Recent research has proposed reconstruction from a single hand-held moving camera given a strong prior for bilayer segmentation [199]. Bi-layer segmentation is used for depth-

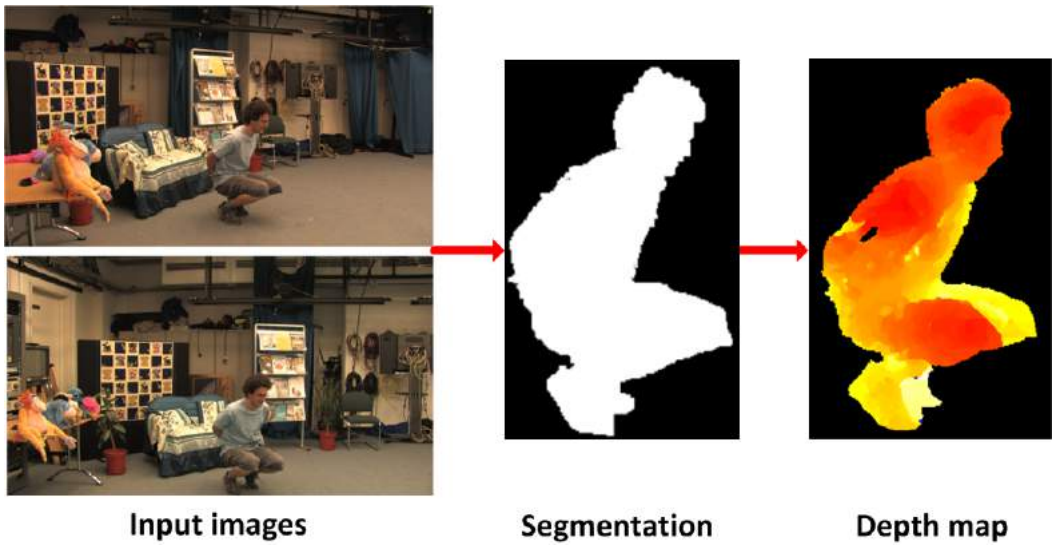


Fig. 4.3 Initialization of existing method with segmentation to obtain depth map [62].

map reconstruction with the DAISY descriptor for matching [77], results are presented for hand-held cameras with a relatively narrow-baseline. Pioneering research in general dynamic scene reconstruction from multiple hand-held wide-baseline cameras [15, 170] 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. 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.

4.2.1 Joint Segmentation and Reconstruction

Many of the existing multi-view reconstruction approaches rely on a two-stage sequential pipeline where foreground/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 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 [39, 98, 196]. These approaches assume static backgrounds and different color distributions for the foreground and background [39, 153] which limits applicability for general scenes.

Joint segmentation and reconstruction methods incorporate estimation of segmentation or matting with reconstruction to provide a combined solution and have been shown to given improved performance for complex scenes. 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.[169] 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 [193]. 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 [135, 158] and fixation of cameras on the object of interest [30].

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 scene segmentation but these approaches do not work with dynamic objects [92]. A multi-layer segmentation and reconstruction approach was proposed for sports data and indoor sequences [62] for multi-view videos.This work on indoor and outdoor dynamic scene reconstruction has shown that joint segmentation and reconstruction across multiple views gives improved reconstruction [62]. The algorithm used visual hull as a prior shape estimate obtained from segmentation of the dynamic objects as shown in Figure 4.3. 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 [156]. These methods are able to produce high quality results, but rely on good initializations and strong prior assumptions. To overcome these limitations, the proposed approaches initialized the foreground object segmentation automatically from wide-baseline feature correspondence followed by joint segmentation and reconstruction.

4.2.2 Graph-cuts in Computer Vision

Graph-cuts have been used extensively in CV applications like segmentation [25], stereo [24] and 3D reconstruction [62]. The problem of segmentation using graph-cuts can be defined as a partitioning of a graph structure in two disjoint sets. A graph consists of vertices and edges connecting those vertices with a non-negative edge weight. A graph structure is built

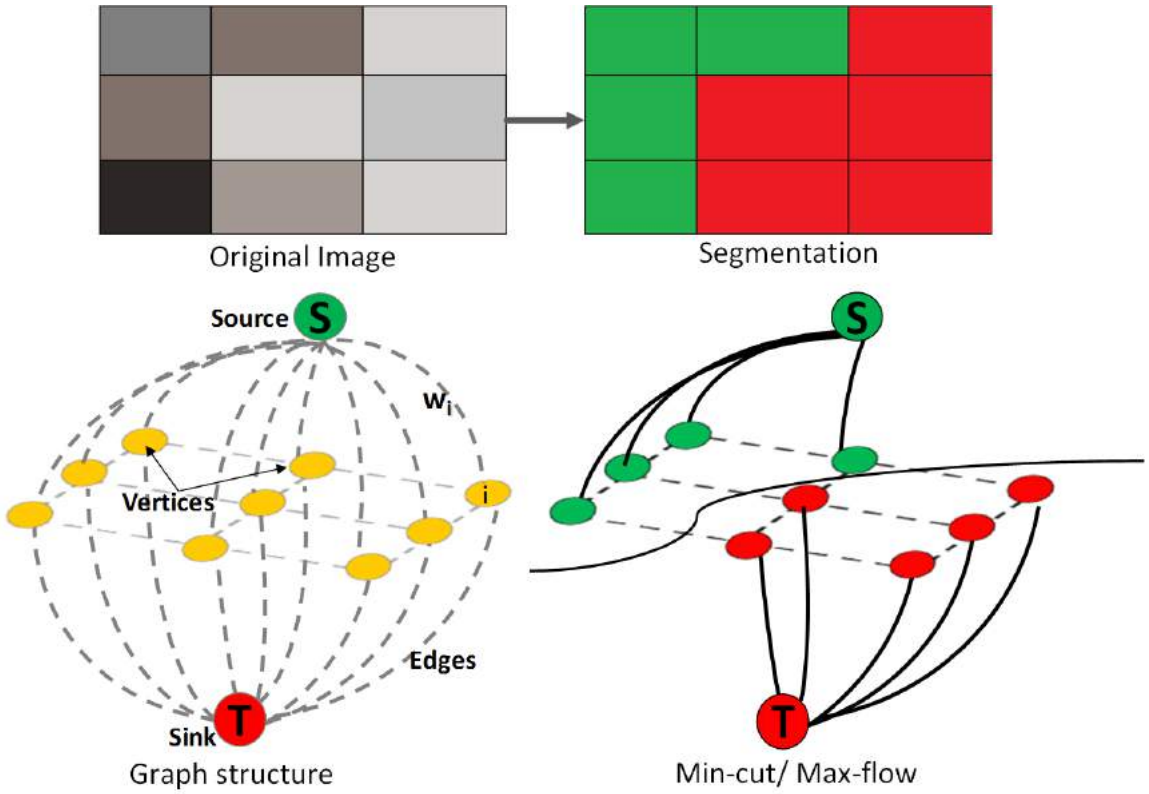


Fig. 4.4 Min-cut max-flow graph-cut example for segmentation

$\mathscr{G} = \{\mathscr{V}, \mathscr{E}\}$ where $\mathscr{V}$ is the set of the vertices and $\mathscr{E}$ is the set of edges with weights w_i assigned to each edge. Vertices are used to represent any entity in the problem domain and edges represent the dependencies between them. Additional nodes or vertices called as Source and Sink are introduced in the graph. A min-cut/max-flow algorithm is used to separate the graph in two sets, each starting from the Source and Sink nodes respectively; such that the cut represents minimum capacity (sum of edge weight) thereby ensuring maximum flow through the graph. An example is shown in Figure 4.4 depicting min-cut for segmentation. The min-cut/max-flow algorithm forms the basis of many other graph based algorithms. Graph-cut is a method of solving problems formulated as an Markov Random Field.

This two-way division of the graph is useful in cases when a binary decision needs to be made. For more common applications graph-cut is extended to multiple discrete labels by employing Markov Random Fields. A labelling algorithm is applied so as to minimize an energy function which consists of cost of assigning label to each vertex along with pairwise potentials in the graph as in [25]. Problems of this form are known to be NP-hard and therefore, a globally optimal solution cannot be obtained. However, a strong local minima can be obtained through graph-cut based optimization algorithms such as

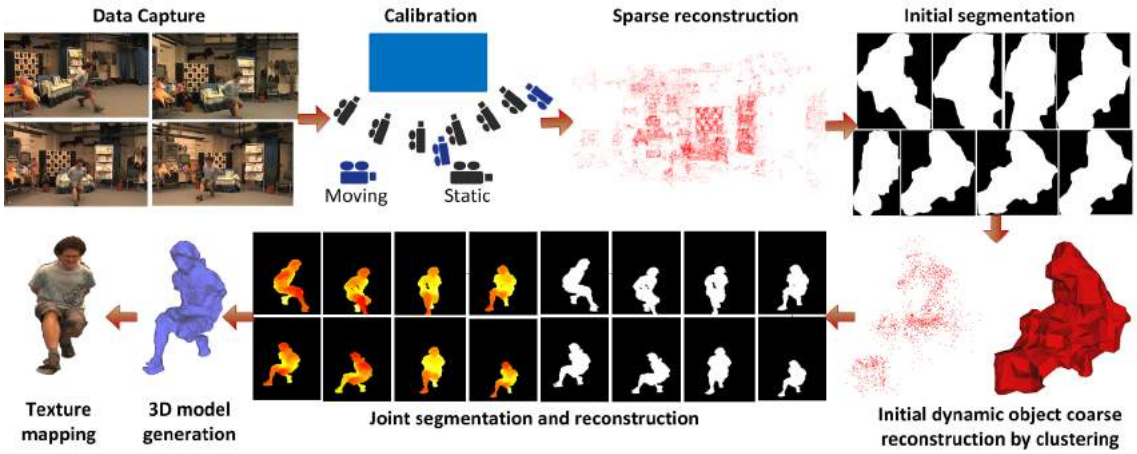


Fig. 4.5 Overview of dense dynamic scene reconstruction framework

α-expansion. In computer vision applications the main idea of the alpha-expansion algorithm is to successively segment all α and non-α pixels with graph-cuts and the algorithm will change the value of α at each iteration if it lowers the total energy. The algorithm will iterate through each possible label for α until it converges.

4.3 Overview

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 previous approaches enabling robust wide-baseline multi-view reconstruction of general dynamic scenes without prior assumptions on scene appearance, structure or segmentation of the moving objects (Figure 4.5). 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 scenes captured only with moving cameras with unknown background and structure. Existing state-of-the-art techniques have not addressed this problem until now.

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 4.5 and consists of the following stages:

Data Capture: The scene is captured using multiple synchronized video cameras (static or moving) separated by a wide-baseline. The cameras are synchronized during the capture using genlock and time-code generator or later using the audio information.
Calibration and sparse reconstruction: The procedure explained in Chapter 3 is followed for extrinsic calibration. Moving cameras are calibrated automatically on-the-fly using through-the-lens multi-camera calibration techniques [74]. A sparse 3D point-cloud is reconstructed from wide-baseline SFD feature matches for each time instant for the entire sequence.
Initial dynamic object segmentation and reconstruction: Automatic initialization 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 optical flow with 3D clustering (section 4.4). The initial coarse reconstruction for the observed moving objects in the scene is used to define the depth hypotheses at each pixel for the optimization.
Joint segmentation and reconstruction for each dynamic object: The initial coarse reconstruction is refined for each dynamic object through joint optimization of shape and segmentation using a robust cost function combining matching and smoothness information for wide-baseline matching. View-dependent optimization of depth is performed with respect to each camera which is robust to errors in camera calibration and initialization. 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 [81]. 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 per frame (each time-instant).

We assume multiple foreground objects corresponding to people/objects located at different depths and a single background layer. Depth is estimated only for foreground objects (background is assigned an unknown depth label). There are several motivations for not explicitly reconstructing the background. Firstly, this cannot be done reliably from the small number of cameras considered here (due to severe occlusions and framing limitations which prevent large portions of the background from being visible in more than a single view). Secondly, background reconstruction is often not required, as the post-production pipelines often only require foreground objects for composition with a virtual background set or a separately captured photo-realistic model.

The proposed approach enables automatic reconstruction of all dynamic objects in the scene as a 3D mesh sequence. The joint segmentation and reconstruction in a view-dependent manner helps in handling calibration errors unlike global approaches which require accurate calibration of the scene. 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.

Subsequent sections present the novel contributions of this work in initialization 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.

4.4 Initial Dynamic Object Reconstruction

For general dynamic scene reconstruction, there is a 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 initialization of a subsequent refinement step to optimize 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 optical flow due to partial motion or correspondence failure. Initialization gives a complete coarse segmentation and reconstruction of each dynamic object for subsequent refinement. The optical flow and cluster information for each dynamic object helps us to retain consistent labels for the entire sequence. The process is as follows:

- Sparse SFD features are matched in time for consecutive time instants for each dynamic object.

- Dense optical flow is performed on the initial coarse segmentation of each dynamic object using the sparse dynamic features as initialization to remove or reduce the errors in sparse correspondence.

- The dense flow is sampled at the sparse feature matches which were used for initialization previously. This is used to identify the dynamic points in the scene.

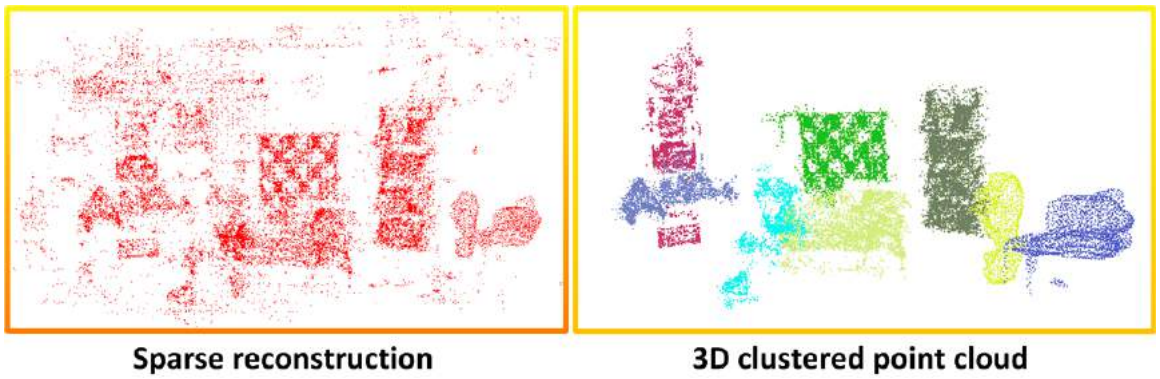


Fig. 4.6 Clustering of sparse point-cloud for Odzemok dataset

4.4.1 Sparse Point-cloud Clustering

Sparse reconstruction based on wide-baseline SFD feature matching, Chapter 3, of the scene consists of outliers due to the calibration errors and false feature matches. This complicates the estimation of local point-cloud characteristics for further processing, leading to erroneous values. These irregularities can be handled by using a sparse outlier removal [152]. The method is based on the computation of the distribution of point to neighbours distances in the input dataset. The method performs a statistical analysis on each point's neighbourhood assuming a Gaussian distribution, and points not satisfying the criteria are trimmed. To retrieve the sparse features corresponding to the dynamic objects from the sparse reconstruction of the scene, this representation is segmented into clusters followed by optical flow labelling.

Clustering: A data clustering approach is applied to identify the dynamic objects in the scene, an example is shown in Figure 4.6. Most clustering methods rely on spatial decomposition techniques that find subdivisions and boundaries to allow the data to be grouped together based on a given measure of "proximity". This measure is usually represented with the Manhattan (L1) and Euclidean (L2) distance metrics. For example a data clustering approach based on Euclidean distance uses 3D grid subdivision of the space with fixed width boxes. This particular representation is very fast to build and is useful for situations where either a volumetric representation of the occupied space is needed, or the data in each resultant 3D box (or octree leaf) can be approximated with a different structure. However this method can only be used for applications requiring equal spatial subdivisions. For situations where clusters can have different sizes a more complex algorithm is needed.

Hence for our framework an approach proposed by [152] is used. A cluster is defined such that every point in a cluster is at a threshold distance from points in different clusters.

The minimal distance between any two clusters is estimated by making use of approximate nearest neighbours queries via a kd-tree representation of sparse points. In a more general sense, nearest neighbours information is used to obtain the cluster, that is essentially similar to a flood fill algorithm. The algorithm is described in Algorithm 4.1. We choose this because of its computational efficiency and robustness. The approach allows unsupervised segmentation of dynamic objects and is shown to work well for cluttered and general outdoor scenes in Section 4.6.

```
Create a kd-tree representation for the input point-cloud;
Set up an empty list of clusters, and a queue that needs to be checked;
for every point in point-cloud do
    Add point to current queue;
    for every point in the current queue do
        Search for the set of point neighbours of the point in a sphere with radius less
          than the minimal distance;
        For every neighbour of the point above, check if the point has already been
          processed, and if not add it to the current queue;
    end
    When the list of all points in queue has been processed, add this set to the list of
      clusters, and reset queue to an empty list;
end
The algorithm terminates when all points in the point-cloud have been processed and
  are now part of the list of point clusters.
```

Algorithm 4.1: Sparse point-cloud clustering(object identification) algorithm[152]

4.4.2 Coarse Scene Reconstruction

Clustering helps in identification of various objects in the scene. For dynamic scene reconstruction, moving objects of the scene are identified by performing optical flow on consecutive frames for a single view of each cluster. For each cluster the optimal camera view is dynamically selected to maximize visibility based on the sparse dynamic feature points at each frame. This allows efficient selection of the best view for optical flow. The objects in close proximity and with similar motion may be clustered together but will be separated once these constraints are changed. 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 the previous frame is retained. This limits the dynamic scene reconstruction to objects which have moved between frames reducing computational cost.

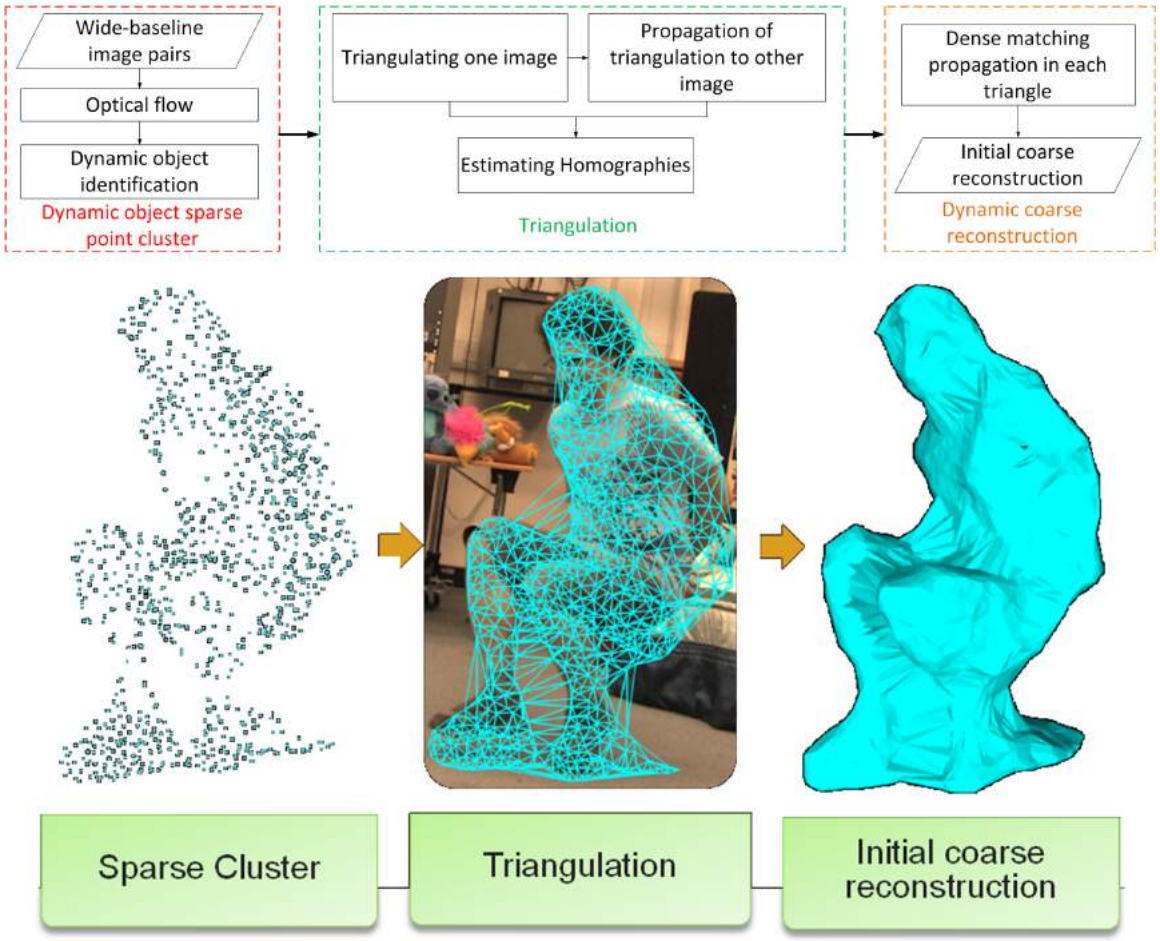


Fig. 4.7 Initial coarse reconstruction algorithm using SFD features

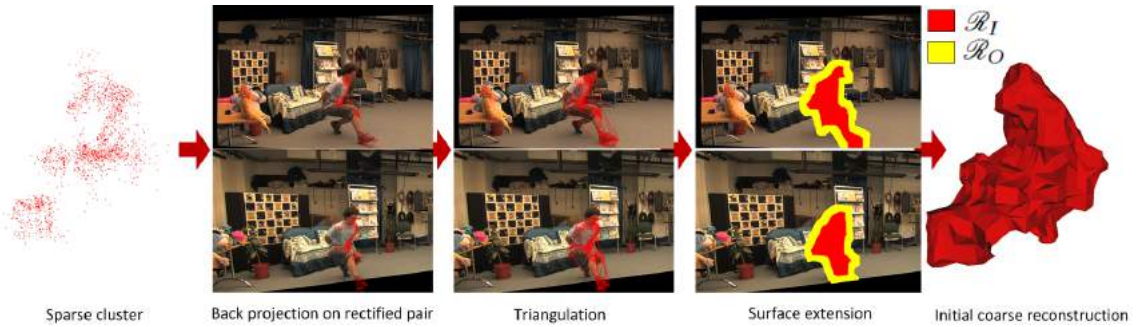


Fig. 4.8 Example of initial coarse reconstruction of the dynamic object in the Odzemok dataset

The process to obtain the coarse reconstruction is shown in Figure 4.7 and an example is shown in Figure 4.8. The sparse representation of dynamic object is back-projected on the rectified image pair for each view. Delaunay triangulation [46] is performed on the set of back projected points for each cluster on one image and is propagated to the other multi-view images at the same time instant using the sparse matched features. For each corresponding triangle pair direct linear transform is used to estimate the affine homography. Displacement at each pixel within the triangle is estimated by interpolation to get the initial dense disparity map for the image pair. Triangles with edge length greater than the median length of edges of all triangles are removed to avoid triangulation across step discontinuities. For each remaining triangle pair direct linear transform is used to estimate the affine homography [125]. 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 R_I depicted in red in Figure 4.8.

The region R_I does not ensure complete coverage of the object, so this region is extrapolated to obtain a region R_O (shown in yellow) in 2D by 5% of the average distance between the boundary points(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 a volume is added 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 R_I or R_O as the propagated pixels of the extrapolated regions (R_O) may have a high level of errors compared to error at the points from sparse representation (R_I) requiring a comparatively higher tolerance. The calculation of the threshold depends on the capture volume of the datasets and is set to 1% of the capture volume for R_O and half the value for 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 a 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.

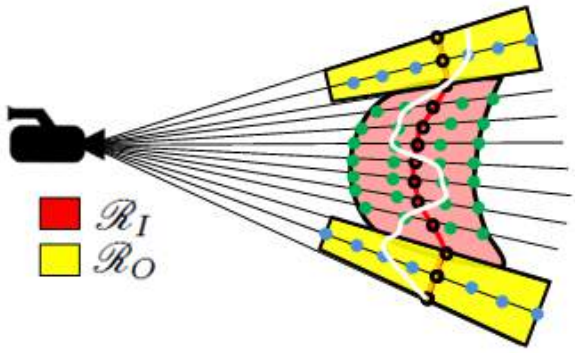


Fig. 4.9 Initial coarse reconstruction: White line represents the actual surface, Depth labels are represented as circles; blue circles depict depth labels in $\mathscr{D}_O$, green circles depict depth labels in $\mathscr{D}_I$ and black circles depict the initial surface estimate.

4.5 Joint Segmentation and Reconstruction

4.5.1 Problem Statement

In this section our aim is to refine the depth of the initial coarse reconstruction estimate of each dynamic object. The problem is to estimate the depth of all moving objects incorporating the initial segmentation and shape information at each time instance for all the views. We do not require a narrow-baseline set-up as is often the case in stereo reconstruction; in fact only a small number of synchronized cameras is assumed, these can be separated by a relatively large baseline. For simplicity, let us consider a single input camera, referred to thereafter as the reference camera, and its N_C neighbouring cameras indexed from 1 to N_C and referred to thereafter as auxiliary cameras; depth reconstruction for all cameras is obtained by considering each input camera in turn as the reference camera. Our aim is to obtain a reliable foreground segmentation of each moving object along with it's depth representation. Mathematically, the aim to assign an accurate depth value to each pixel p from a set of depth values $\mathscr{D}_\mathscr{I} = \{d_1,...,d_{|\mathscr{D}_\mathscr{I}|-1},U\}$ and $\mathscr{D}_\mathscr{O} = \{d_1,...,d_{|\mathscr{D}_\mathscr{O}|-1},U\}$ and assign a layer label from a set of label values $\mathscr{L} = \{l_1,...,l_{|\mathscr{L}|}\}$. Each d_i is obtained by sampling the optical ray from the camera and U is an unknown depth value to handle occlusions and to refine the object segmentation. If a pixel in the image is assigned a depth label value U the pixel is removed from the segmentation of the dynamic object. We assume that the depth of a particular pixel lies within the given threshold around the initial estimate as depicted in Figure 4.9 and varies depending upon the regions R_I or R_O. Hence the depth labels are divided in two sets, one for the region R_I ($\mathscr{D}_I$) and other for R_O ($\mathscr{D}_O$) such that $|\mathscr{D}_I| < |\mathscr{D}_O|$.

4.5.2 Proposed Approach

We formulate the computation of depth at each point as an energy minimization of the cost function defined in Eq. (4.1). 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.

$$E(l,d) = \lambda_{data}E_{data}(d) + \lambda_{contrast}E_{contrast}(l) + \lambda_{smooth}E_{smooth}(l,d) \tag{4.1}$$

where, d is the depth at each pixel for our dynamic object for the region R_I + R_O and can be assigned $U$and l is the layer label for multiple objects 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 [23] and the contrast term is used for segmentation [87] to encourage depth/layer discontinuities coincide with discontinuities in appearance. This is a view-dependent or per-view optimization of depth and the depth estimates per view will be combined to obtain an optimal surface reconstruction using Poisson surface reconstruction[81]. By performing the layered-depth estimation in a view-dependent manner, an accurate segmentation and reconstruction can be obtained in spite of calibration errors.

Matching term

The matching term encourages the reconstructed surface to be photo-consistent across multiple views. This is based on the idea that correct depth hypotheses yield similar appearances across the images in which they are visible, while incorrect depth hypotheses usually result in inconsistent projected appearances.

Several confidence measures for stereo matching have been proposed in the literature. An evaluation of existing confidence measures against ground-truth was performed in [72]. Robust photo-consistency scores were introduced to handle problems like uniform regions, repetitive appearance in the state-of-the-art measures for various applications in computer vision. A maximum likelihood measure was initially introduced in [111] using sum-of-squared-differences to measure the photo-consistency between images. This measure was generalized to the normalized-cross correlation (NCC) function to obtain a probability density function for disparity given cost by assuming that the cost follows a normal distribution and that the disparity prior is uniform as proposed in [72]. This gives a good photo-consistency measure for wide-baseline multi-view datasets because of its ability to obtain a high number of correct matches and preserve boundaries. Hence to measure the photo-consistency, a data

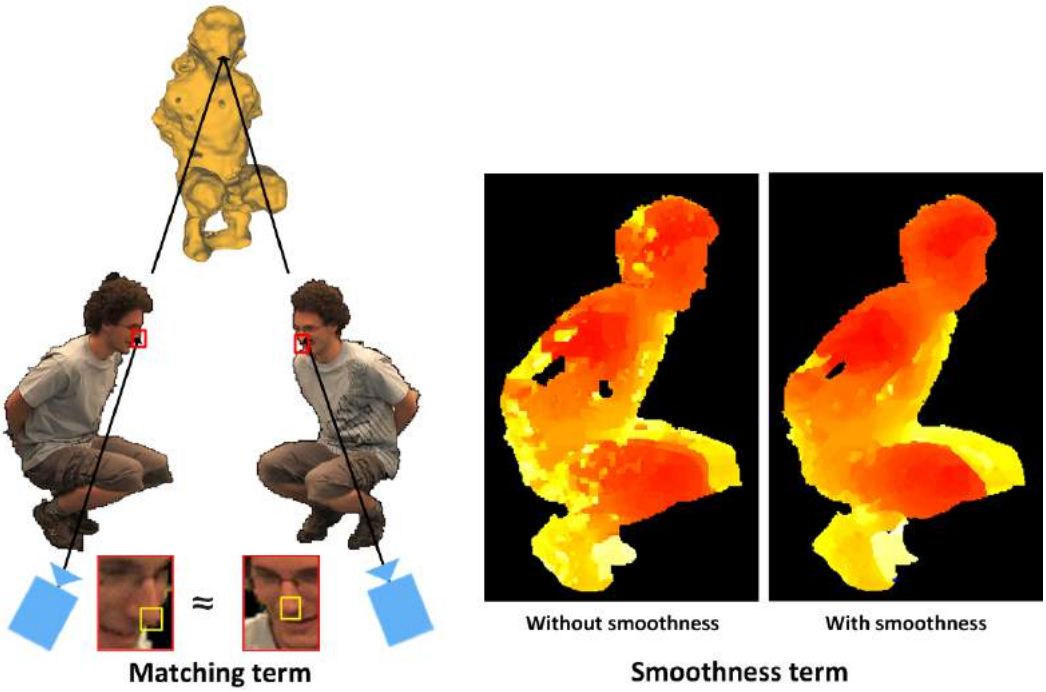


Fig. 4.10 Illustration of matching and smoothness term for the energy minimization

term score based on this measure is used, as illustrated in Figure 4.10, given by:

$$E_{data}(d) = \sum_{p \in N_P} e_{data}(p, d_p) = \begin{cases} M(p,q) = \sum_{i \in N_C} m(p,q), & \text{if } d_p \neq U \\ M_U, & \text{if } d_p = U \end{cases} \tag{4.2}$$

where N_P is the set of all pixels p, $m()$ is the confidence measure defined in equation 4.3, M_U is the fixed cost of labelling a pixel unknown and $q = \Pi(p, d_p)$ denotes the projection of the hypothesized 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. N_C is the set of most photo-consistent pairs with reference camera decided by the number of correct feature matches. The image pairs above a certain threshold are selected as most photo-consistent pairs with reference camera and the threshold is set to half of the maximum number of correspondences between an image pair for a particular time instant.

For textured scenes NCC over a squared window with different sizes is a common choice [156]. The NCC values range from -1 to 1 which are then mapped to non-negative values by using the function $1 - NCC$. The maximum likelihood measure[111] generalized to NCC to

obtain a probability density function is defined as:

$$m(p,q) = \frac{exp\frac{c_{min}}{2\sigma_i^2}}{\sum_{(p,q)\in N_I} exp\frac{-(1-NCC(p,q))}{2\sigma_i^2}} \tag{4.3}$$

where σ_i^2 is the noise variance for each auxiliary camera i; this parameter was fixed to 0.3. N_I denotes the set of interacting pixels in N_P. c_{min} is the minimum cost for a pixel obtained by evaluating the function $(1-NCC(.,.))$ on a $W \times W$ window, with $W = 15$. $NCC()$ is defined as:

$$NCC(p,q) = \frac{\sum_{k\in W}(I_L(p_k,q_k)-\mu_L)(I_R(p_k-d,q_k)-\mu_R)}{\sigma_L\sigma_R} \tag{4.4}$$

where, I_L and I_R are the image pair, μ_L and μ_R are the means and σ_L and σ_R are the standard deviations of all pixels in the square window for the image pair.

Contrast term

The contrast term encourages layer discontinuities to occur at high contrast locations. This naturally encourages low contrast regions to coalesce and favours discontinuities to follow strong edges. Object boundaries in images tend to align with contours of high contrast and it is desirable to represent this as a constraint in surface reconstruction. The simplest choice for the contrast term would be a squared Euclidean color distance, but this would yield a contrast term encouraging layer discontinuities to follow any edge in the image regardless of whether it is weak or strong. Better performance can be obtained by using combining the edge information with contrast-likelihood in the image [87]. A modified version of this interpretation is used in our formulation to preserve the edges through Bilateral filtering [174] instead of Gaussian filtering.

$$E_{contrast}(l) = \sum_{p,q\in N_I} e_{contrast}(p,q,l_p,l_q) \tag{4.5}$$

$$e_{contrast}(p,q,l_p,l_q) = \begin{cases} 0, & \text{if } (l_p = l_q) \\ \frac{1}{1+\varepsilon}(\varepsilon + exp^{-J(p,q)}), & \text{otherwise} \end{cases} \tag{4.6}$$

$\|\cdot\|$ is the L_2 norm and $\varepsilon = 1$. To improve the segmentation quality $J(p,q)$ is defined as:

$$J(p,q) = \frac{\|B(p)-B(q)\|^2}{2\sigma_{pq}^2 o_{pq}^2} \tag{4.7}$$

where $B(.)$ represents the bilateral filter, o_{pq} is the Euclidean distance between p and q, and σ_{pq} is defined below:

$$\sigma_{pq} = \left\langle \frac{\|B(p) - B(p)\|^2}{o_{pq}^2} \right\rangle \tag{4.8}$$

This term enables removal of regions with low photo-consistency scores and weak edges, this helps in estimating the object boundaries.

Smoothness term

Smoothness is useful in situations where matching constraints are weak (low photo-consistency scores) and insufficient to produce an accurate reconstruction without the support from neighbouring pixels. This term is same as [62] and to ensure that 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.

$$E_{smooth}(l,d) = \sum_{(p,q)\in\mathcal{N}} e_{smooth}(l_p, d_p, l_q, d_q) \tag{4.9}$$

$$e_{smooth}(l_p, d_p, l_q, d_q) = \begin{cases} min(\left|d_p - d_q\right|, d_{max}), & \text{if } l_p = l_q \text{ and } d_p, d_q \neq \mathscr{U} \\ 0, & \text{if } l_p = l_q \text{ and } d_p, d_q = \mathscr{U} \\ d_{max}, & \text{otherwise} \end{cases} \tag{4.10}$$

d_{max} is set to 50 times the size of the depth sampling step defined in Section 4.5.1 for all datasets. Such a distance is discontinuity preserving as it does not over-penalize large discontinuities; this is known to be superior to simpler non-discontinuity functions[25]. This term encourages unknown features to coalesce within each object and regularizes the reconstruction of regions with weak image support. The smoothness term reduces the noise in the reconstructed surface. An example of surface reconstruction with and without smoothness term for Odzemok dataset is shown in Figure 4.10. Depth reconstruction with smoothness has reduced noise in the object, making the shape estimate more realistic.

4.5.3 Optimization of Reconstruction and Segmentation

Optimization of the energy defined by Eq. (4.1) is known to be NP-hard. However, The energy minimization for Eq. (4.1) is performed by using the α-expansion move algorithm from [25] based on graph-cuts. We choose graph cuts because of its strong optimality properties over belief propagation [171]. Graph-cut using the min-cut/max-flow algorithm is

used to obtain a local optimum [24]. The α-expansion for a pixel p is performed by iterating through the set of labels in $\mathscr{L} \times \mathscr{D}_I$, if $p \in R_I$ and $\mathscr{L} \times \mathscr{D}_O$, if $p \in R_O$ until the energy cannot be decreased. Each α-expansion iteration can be solved exactly by performing a single graph-cut using the min-cut/max-flow algorithm [24]. Convergence is achieved after 4 or 5 iterations. After convergence of the algorithm, the result obtained is guaranteed to be a strong local optimum [25]. For all experiments reported in this chapter, the α-expansion algorithm was initialized with the initial coarse reconstruction estimate; convergence has been found to be insensitive to the choice of initialization. A final model is obtained by merging the view-dependent depth representations through the Poisson surface reconstruction algorithm as explained in next Section.

4.5.4 3D Model Generation

Final model generation is performed by merging the view-dependent depth estimates for each camera view into a single global shape representation using Poisson surface reconstruction [81]. The algorithm produces a water-tight reconstruction and eliminates the redundancy contained in the view-dependent meshes. The algorithm cannot be directly applied because of the noise and missing data in the view dependant meshes which tends to produce large protrusions. Vlasic et al. compensated for missing data in the final mesh by using the points from visual hull [185]. In our case the initial coarse reconstruction is not accurate enough to be used directly for view-dependent reconstructions. Instead a two-step Poisson surface reconstruction is performed. In the first step, the points from view dependent depth maps are refined to remove noise and a reconstruction is obtained with protrusions at occluded regions. In the second step, protrusions are removed by adding initial coarse reconstruction sample points located inside the reconstruction obtained at the first pass to the view-dependent sample points. This effectively reduces the addition of points to badly reconstructed areas, thus removing protrusions without deteriorating correctly reconstructed regions. Vertex connectivity is decided based on the layer segmentation and thresholding of the angle separating the line segment connecting 3D surface points defined by pairs of neighbouring pixels and the optical ray passing through the midpoint of the pixel pair (a threshold of 80° is used). This allows pixel belonging to different layers or located at a depth discontinuity to be correctly converted into separate mesh components. It should be noted that the Poisson surface reconstruction takes as input an oriented set of points while the depth recovered is not orientated. Orientation is estimated based on neighbouring pixels.

Dataset	Number of Cameras	Number of frames	Image resolution	Baseline
Dance1	7 (all static)	250	1920×1080	15 degrees
Magician	6 (all moving)	6900	960×544	40-55 degrees
Dance2	8 (all static)	125	1920×1080	45 degrees
Odzemok	8 (2 moving)	250	1920×1080	15 degrees
Cathedral	8 (all static)	143	1920×1080	45 degrees
Juggler	6 (all moving)	3500	960×544	25-35 degrees

Table 4.1 Characteristic properties of all the datasets used for evaluation.

4.6 Results and Evaluation

Evaluation is performed using publicly available research datasets: Indoor and Outdoor dataset with simple background (Dance2 and Cathedral), Indoor datasets with cluttered background (Odzemok and Dance1) and Indoor and Outdoor datasets captured with moving hand-held cameras (Magician and Juggler) [15]. The detailed characteristics of these datasets are given in Table 4.1 and the parameter settings for Eq.(4.1) are summarized in Table 4.2.

The framework explained in Section 4.3 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. The parameters for the experiments are shown in Table 4.2. 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 [51] for wide-baseline reconstruction and Guillemaut and Hilton [62] for joint reconstruction and segmentation. Both of these approaches are top performers on the Middlebury for multi-view reconstruction of wide-baseline views [156].

Quantitative evaluation of the 3D dynamic scene reconstruction against ground-truth would be valuable to validate the method. However ground-truth data in reconstruction is currently not available and as with other work, currently the only mechanism available for this is through simulated data, which does not represent the complexity of real scenes. The comparison is performed against two state-of-the-art methods explained below:

Guillemaut [62]: 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 no prior knowledge of the background is assumed and this approach is initialized using our initial coarse reconstruction instead of the visual hull.

Dataset	λ_{data}	λ_{smooth}	$\lambda_{contrast}$
Dance1	1.0	0.01	2.0
Magician	1.0	0.02	6.0
Dance2	1.0	0.01	2.0
Odzemok	1.0	0.01	2.0
Cathedral	1.0	0.02	8.0
Juggler	1.0	0.02	8.0

Table 4.2 Parameter settings used in Equation 4.1 for view-dependent reconstruction of all the datasets. The parameters for outdoor and moving hand-held camera sequences are same and parameters for indoor sequences captured with static and moving cameras remain consistent.

Furukawa [51]: This represents a state-of-the-art multi-view wide-baseline stereo approach. It takes multi-view images as input and generates dense reconstruction of the scene. Furukawa [51] does not refine the segmentation but gives a 3D point-cloud which is converted into a mesh using Poisson surface reconstruction.

4.6.1 Segmentation Results

The segmentation results from the proposed approach are compared against the segmentation from Guillemaut and Hilton [62] 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.

Qualitative Results

The segmentation results for two frames from each dataset are shown in Figure 4.11 for indoor datasets and Figure 4.12 for outdoor datasets. Guillemaut requires visual hull initialization, in this case the proposed initial coarse reconstruction is obtained from the sparse feature matches and deviates from the actual object boundaries as shown in Figure 4.8. 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 artefacts with respect to ground-truth in the proposed approach are from shadow areas and occlusions.

Quantitative Evaluation

To perform the quantitative evaluation of the segmentation the $HitRatio$, $BkgRatio$ and $OverlapRatio$ as defined in [158] is measured against the ground-truth pixels. Three criterion

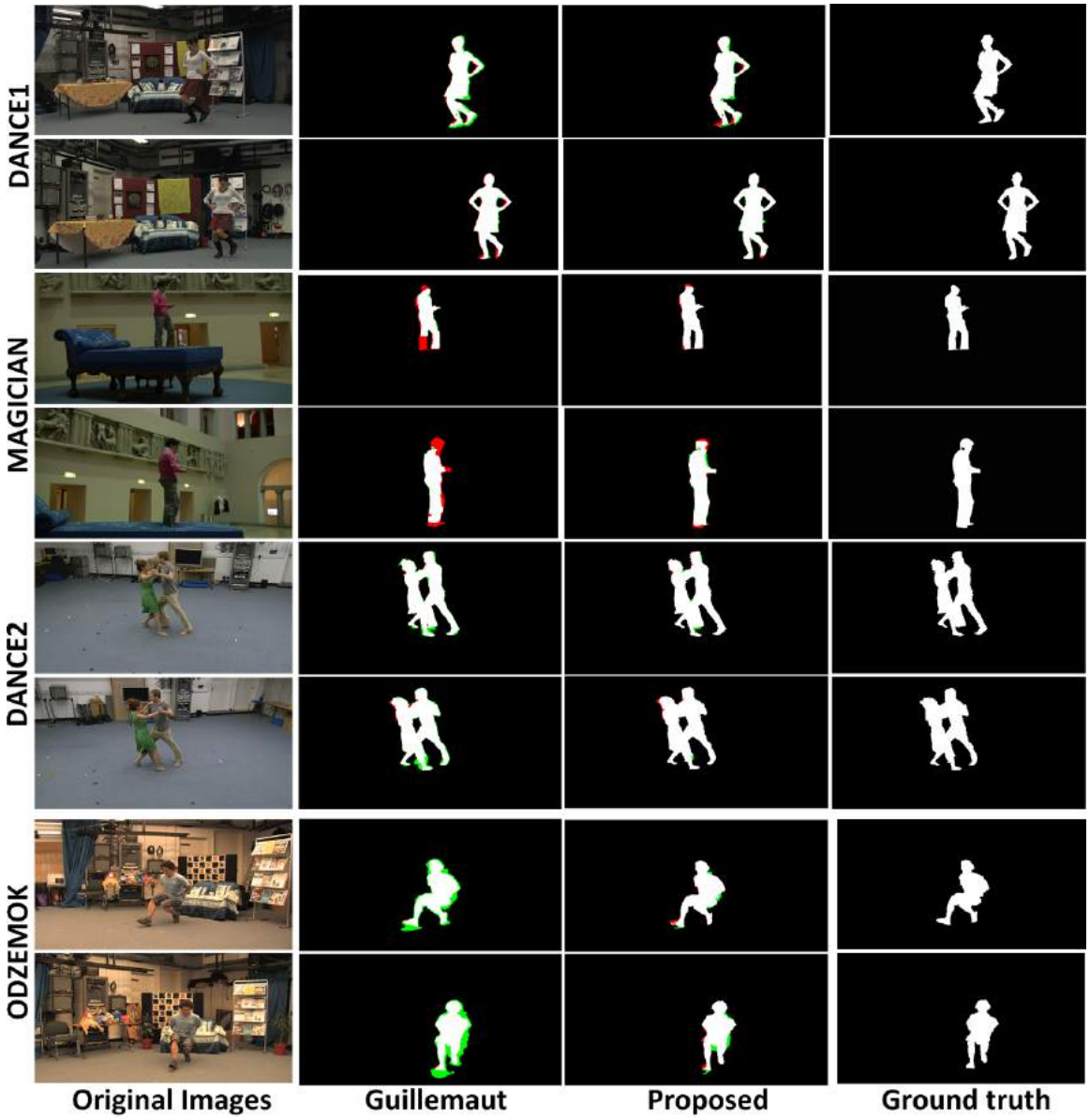


Fig. 4.11 Results for a pair of images from indoor datasets: $2^{nd} - 4^{th}$ column: Segmentation (Red represents true negatives and green represents false positives compared to the ground-truth)

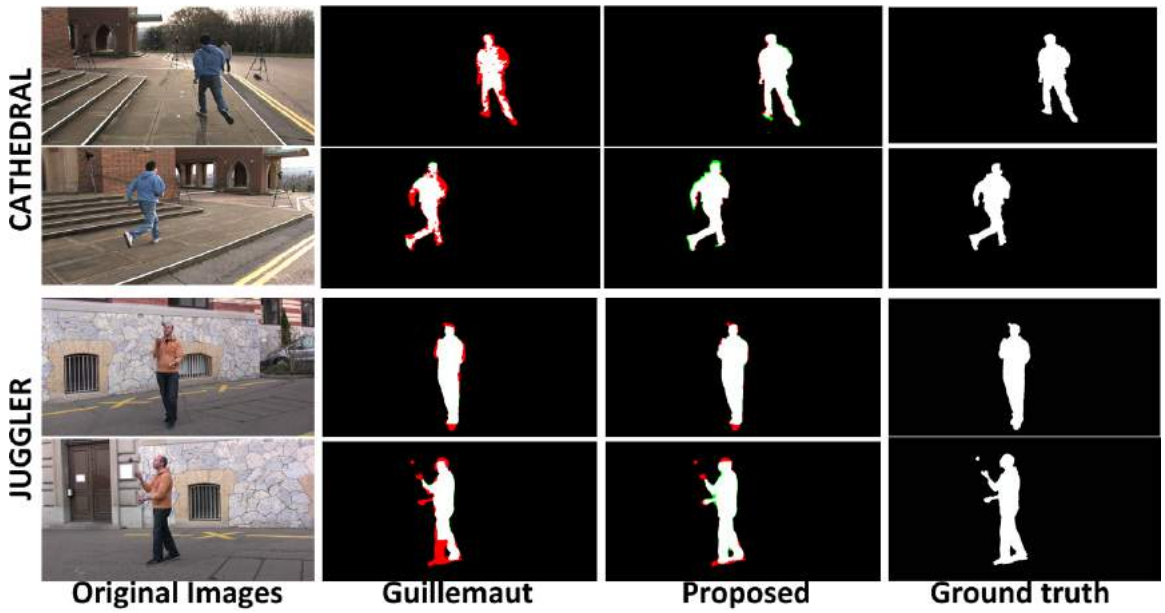


Fig. 4.12 Results for a pair of images from outdoor datasets: $2^{nd} - 4^{th}$ column: Segmentation (Red represents true negatives and green represents false positives compared to the ground-truth)

are defined as follows:

$$\begin{aligned} HitRatio &= |Result \bigcap GT| \,/\, |GT| \\ BkgRatio &= |Result - GT| \,/\, |Result| \\ OverlapRatio &= |Result \bigcap GT| \,/\, |Result \bigcup GT| \end{aligned} \tag{4.11}$$

where GT is the ground-truth segmentation image and $Result$ is the final segmentation image obtained using various approaches. The ratio is calculated by counting the number of pixels in the numerator and denominator of the criterion defined above. The results are shown in Table 4.3 for all the datasets. 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 datasets, generally significantly higher for more complex datasets like outdoor scenes or scenes captured with only hand-held moving cameras. This demonstrates the robustness of the proposed approach to general dynamic scene segmentation compared to Guillemaut as seen in Figure 4.11 and 4.12. The $BkgRatio$ measures the proportion of result which actually belongs to background i.e. false positives in the segmentation. In

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

Table 4.3 Segmentation performance comparison for all datasets (best for each dataset is highlighted in bold)

the case of Guillemaut this value is higher 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.

4.6.2 Reconstruction Results

We have compared our results with Guillemaut and Furukawa (explained in Section 4.6). For fair comparisons all of the approaches are initialized with the same calibration and coarse reconstruction obtained using the method explained in Section 4.4.

Qualitative Results

The depth maps for the proposed approach and Guillemaut are shown in Figure 4.13 for indoor datasets and 4.14 for outdoor datasets. The colors in the depth map ranges from yellow to red with red depicting closer to the camera. The consistency of depth maps in the case of the proposed approach is 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 4.15 for indoor and in Figure 4.16 for outdoor datasets. For the Magician dataset Furukawa gives very few points on a small part of

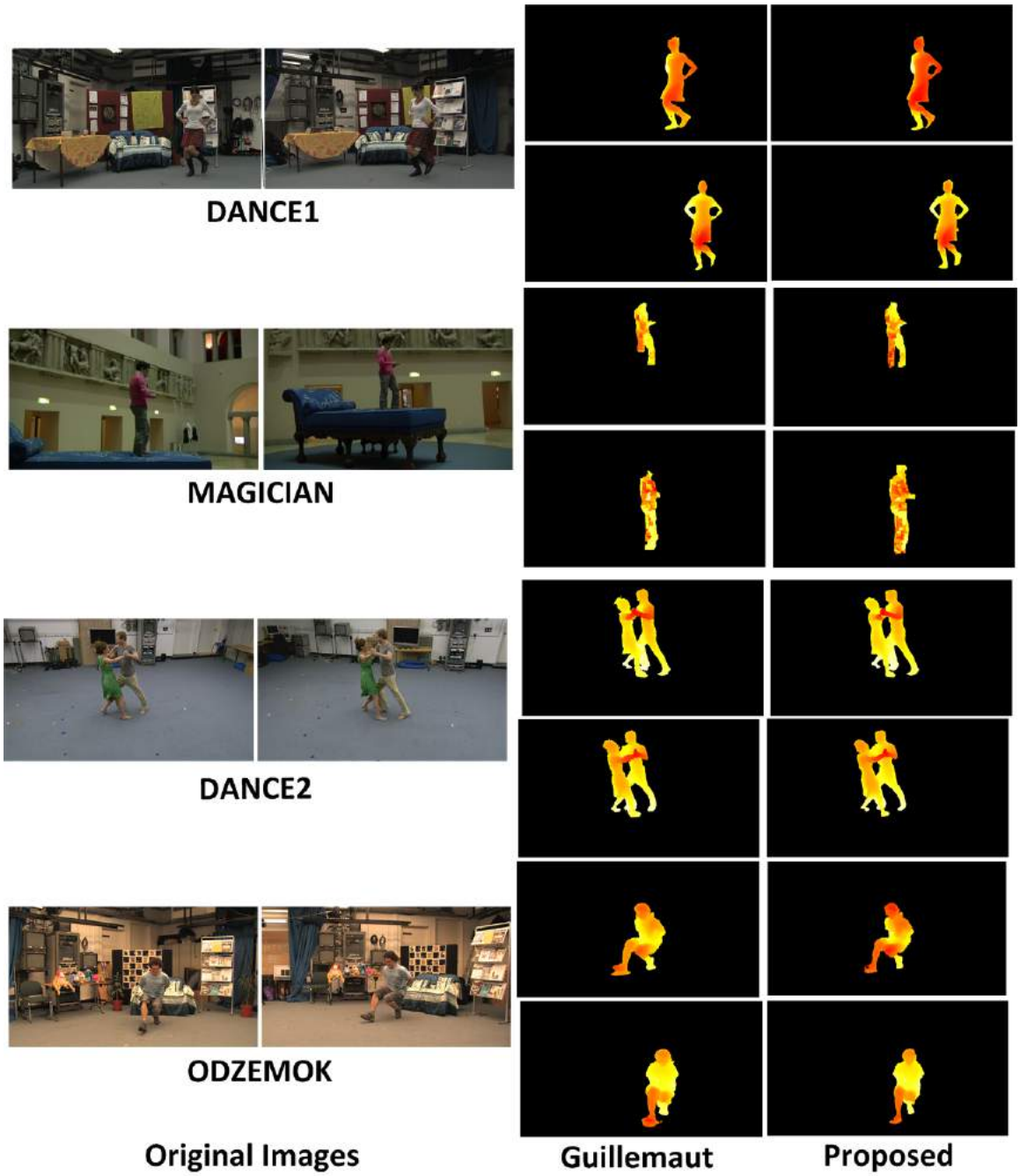


Fig. 4.13 Depth results for a pair of images from Indoor datasets: $2^{nd} - 3^{rd}$ column: Depth

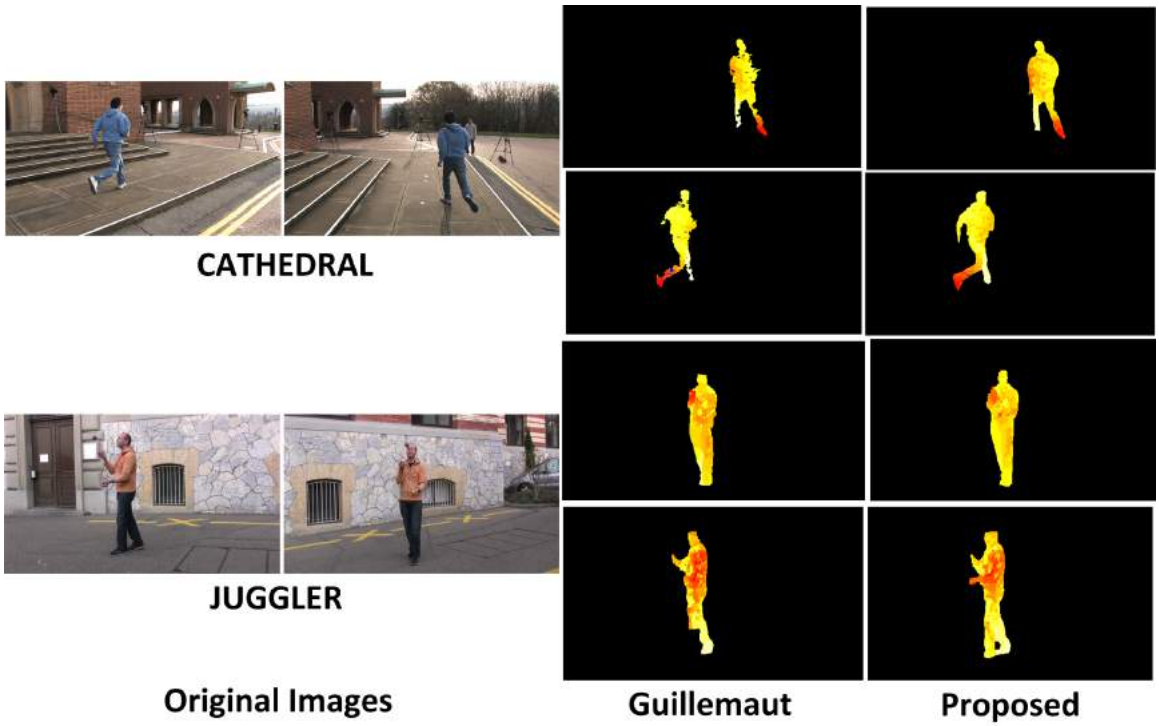


Fig. 4.14 Depth results for a pair of images from Outdoor datasets: $2^{nd} - 3^{rd}$ column: Depth

the object in the reconstruction due to the complexity of the dataset. Results are compared closely with Guillemaut in Figure 4.17. Reduced noise is observed in the reconstructed surface and the limb of the object is successfully recovered for proposed approach compared to Guillemaut. In Figure 4.15 and 4.16 the meshes obtained by Furukawa do not have clear boundaries because the method 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. The mislabellings in the initial coarse reconstruction in both the figures are recovered in the final reconstruction using the proposed method. Some errors in the mesh reconstruction are present due to camera noise, uniform textures and similarity to the background. Comparison of reconstruction obtained using initial coarse reconstruction is shown against visual hull based initialization in Figure 4.18. Visual hull is obtained using manual segmentation in each view. The initial coarse reconstruction is obtained using the sparse features. In case of absence of features on the leg of object we obtain incomplete reconstruction. To illustrate the usefulness of joint optimization we intialize it using visual hull at that frame to obtain a more complete reconstruciton. Reconstruction results for the Juggler sequence captured using only moving cameras are shown in Figure 4.19 for frames ranging from 100 to 150.

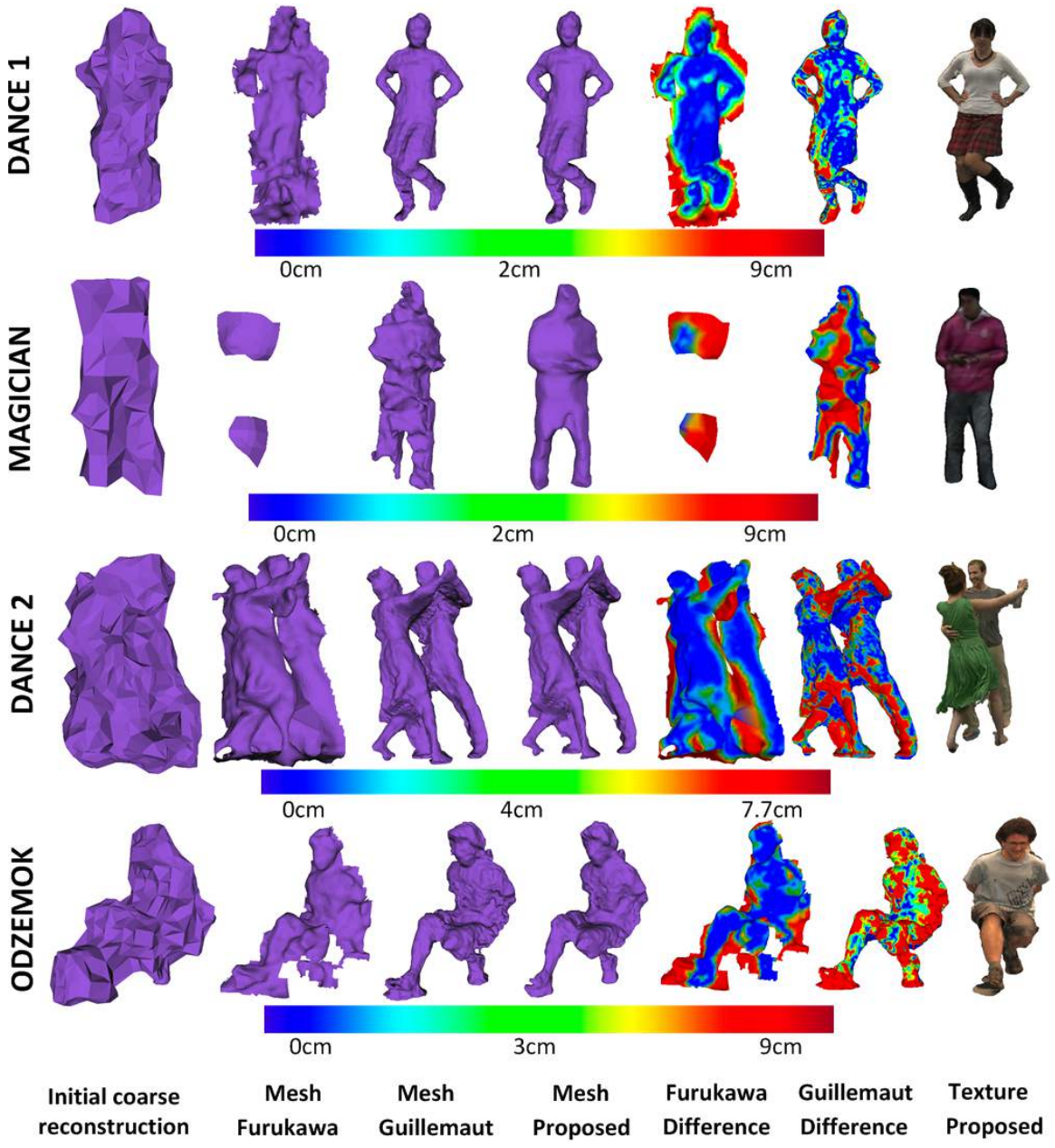


Fig. 4.15 Reconstruction results for indoor datasets: $1^{st} - 4^{th}$ column: Meshes and $5^{th} - 6^{th}$ column: Difference meshes against proposed approach with color coded error in cms and 7^{th} is the textured mesh.

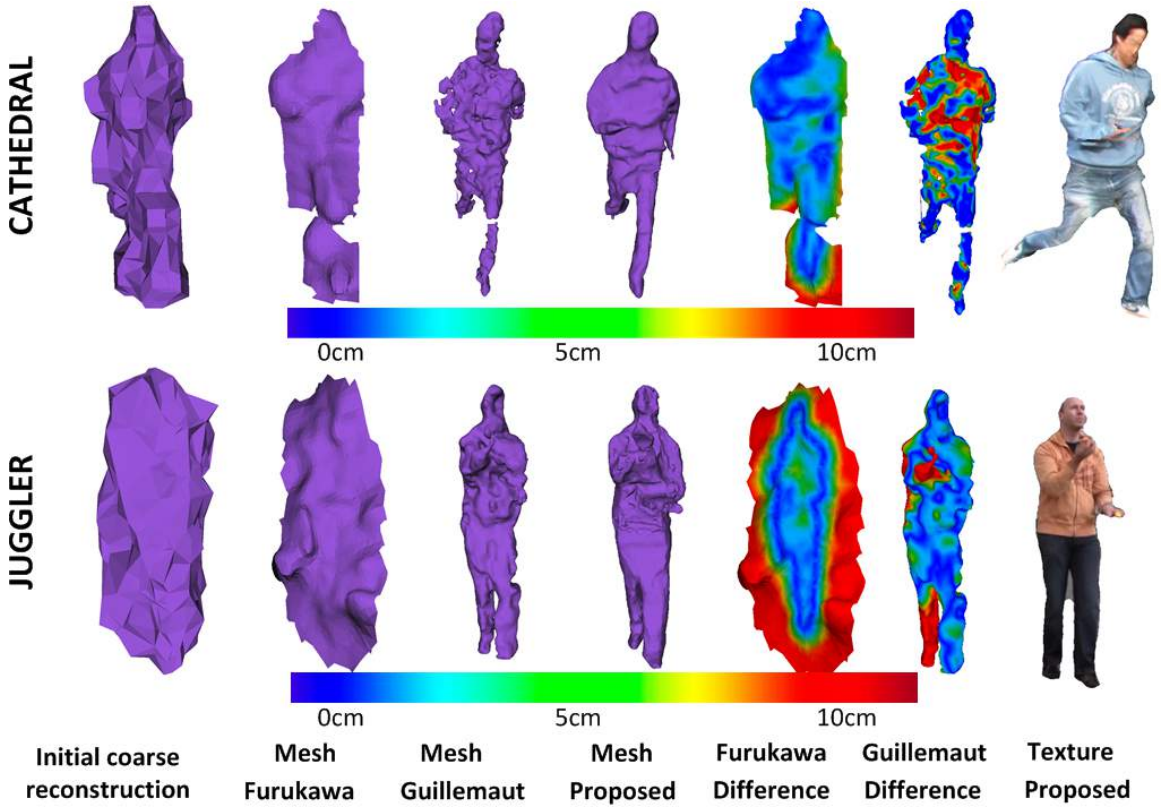


Fig. 4.16 Results for outdoor datasets: $1^{st} - 4^{th}$ column: Meshes and $5^{th} - 6^{th}$ column: Difference meshes against proposed approach with color coded error in cms and 7^{th} is textured mesh.

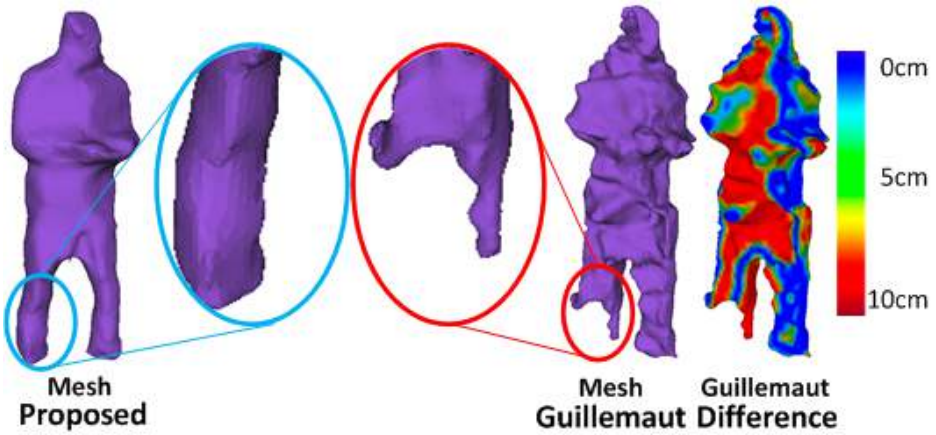


Fig. 4.17 Reconstruction result comparison against Guillemaut for Magician dataset

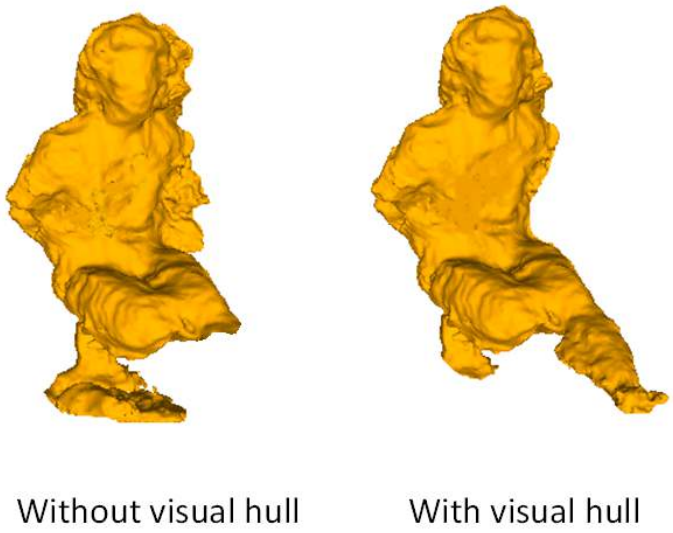


Fig. 4.18 Reconstruction results using proposed method initialized with initial coarse reconstruction against visual hull based initializatio for Odzemok dataset

Fig. 4.19 Result for Juggler sequence: Original images from one view with frame numbers and mesh reconstructions alternatively

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 4.4. 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% compared to Guillemaut.

The computation times can be improved using an alternative faster discrete optimization approach proposed recently. The method described in [90] can be used as an efficient solver to improve the timings further.

Dataset	Furukawa[51]	Guillemaut[62]	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 4.4 Comparison of computational efficiency for all datasets (time in seconds (s))

4.7 Limitations

The proposed method gives per frame reconstruction of the sequence, giving temporally incoherent shape and segmentation. This inconsistency also introduces errors in frames like missing limbs etc. for dynamic objects as shown in Figure 4.20. To handle such errors there is a need to introduce temporal alignment in the framework.

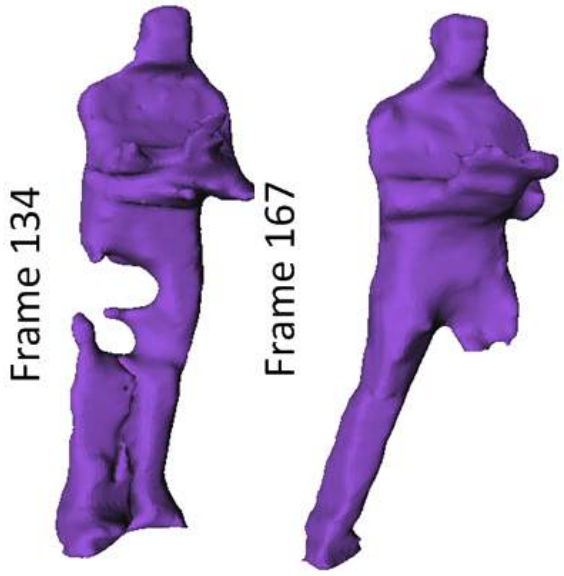


Fig. 4.20 Limitations of proposed method: Missing data in reconstruction of Juggler dataset for two different frames.

Another problem is the quality of the reconstruction and segmentation. Although the proposed approach performs better than the state-of-the-art techniques, the quality of results is far from perfect. The joint refinement framework needs to be improved to obtain higher quality output. These problems will be handled in the next chapter by introduction of temporal

coherence and shape constraint in the system to improve the quality of reconstruction and segmentation. Also, the reconstruction is limited to dynamic objects in the scene in this work and the aim is to combine the dynamic scene reconstruction with shape estimate of the static parts of the scenes to obtain full scene reconstruction from the sequence.

4.8 Conclusion

This chapter 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. This overcomes the common limitations of previous dynamic scene reconstruction methods which assume static backgrounds and known background appearance to allow prior segmentation of the scene. An automatic approach to initialize the dynamic scene reconstruction from the sparse 3D points obtained using SFD features (introduced in Chapter 3) to give good coverage of the dynamic objects in the scene is introduced. 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 joint view-dependent optimization of segmentation and reconstruction by a view-dependent graph-cut optimization using the photo-consistency and contrast cues from wide-baseline images.

This overcomes the limitation of previous reconstruction approaches which assume prior knowledge of either the empty background scene appearance or prior background scene reconstruction. 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. The improvement in quality is because of introduction of error-tolerant photo-consistency scores along with contrast information which introduces affinity towards strong foreground edges. Tests on challenging datasets demonstrate improvements in quality of reconstruction and segmentation compared to state-of-the-art methods.

The subsequent chapters in this thesis are build on the approach introduced in this chapter to address the limitations identified. As explained in Section 4.7, limitations of proposed framework of per frame reconstruction, temporal incoherence and limited quality of reconstruction and segmentation will be handled in Chapter 5.

Chapter 5

Temporally Coherent Scene Reconstruction

5.1 Introduction

General dynamic scene reconstruction introduced in Chapter 4 operates on a frame-by-frame basis and generates temporally incoherent results resulting in limited quality of the reconstruction and segmentation. In this work we build on the concept introduced in Chapter 4 by exploiting temporal coherence of the scene to overcome visual ambiguities inherent in single frame reconstruction and multi-view segmentation methods for general scenes. This is illustrated in Figure 5.1 where the resulting 4D scene reconstruction has temporally coherent labels and surface correspondence for each object. Temporal coherence refers to the correlation between 3D points of all objects observed at various time instants. Temporal information and shape constraints are introduced to extend the optimization framework proposed in the previous chapter. This improves the quality of the reconstruction and produce temporally coherent results. A complete scene reconstruction is produced by exploiting the static redundancy in the scene.

A sparse-to-dense approach is used to estimate dense temporal correspondence and surface reconstruction for non-rigid objects with no prior knowledge of scene structure or camera calibration allowing reconstruction from multiple moving cameras. Initially sparse 3D feature points are robustly tracked from wide-baseline image correspondence using spatio-temporal information to obtain sparse temporal correspondence and reconstruction. Sparse 3D feature correspondences are used to constrain optical flow estimation to obtain an initial dense temporally consistent model of dynamic regions. The initial model is then refined using a novel optimization framework using shape constraints for simultaneous multi-view

segmentation and reconstruction of non-rigid objects. Shape constraints are introduced by limiting the initial segmentation of the objects with geodesic star convexity. This enforces the segmentation to be connected to seed points via geodesic paths to handle complex object shapes [64]. The proposed approach overcomes constraints of existing methods allowing an unsupervised temporally coherent 4D reconstruction of complete models for general scenes. The scene is automatically decomposed into a set of spatio-temporally coherent objects as shown in Figure 5.1. The contributions are as follows:

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

5.2 Related Work

5.2.1 Temporal Multi-view Reconstruction

Extensive research has been performed in multi-view reconstruction of dynamic scenes. Reconstruction frameworks for general dynamic scenes commonly operate on a frame-by-frame basis [51, 122] or are limited to simple scenes [58]. Most existing approaches process each time frame independently due to the difficulty of simultaneously estimating temporal correspondence for non-rigid objects. Independent per frame reconstruction can result in errors due to the inherent visual ambiguity caused by occlusion and similar object appearance for general scenes. Quantitative evaluation of state-of-the-art techniques for static object reconstruction from multiple views was presented [156]. Research investigating spatio-temporal reconstruction across multiple frames [58, 63] requires accurate initialization, is limited to simple scenes and does not produce temporally coherent 4D models. A number of approaches that use temporal information [14, 96, 99] either require a large number of closely spaced cameras or bi-layer segmentation [77, 199] as a constraint for complete reconstruction. Other approaches for reconstruction of general scenes from multiple hand-held wide-baseline cameras [15, 170] exploit prior reconstruction of the background scene to allow dynamic foreground segmentation and reconstruction. Recent approaches for spatio-temporal reconstruction of multi-view data either work on indoor studio data [134] or for dynamic reconstruction of crowd sourced data [76].

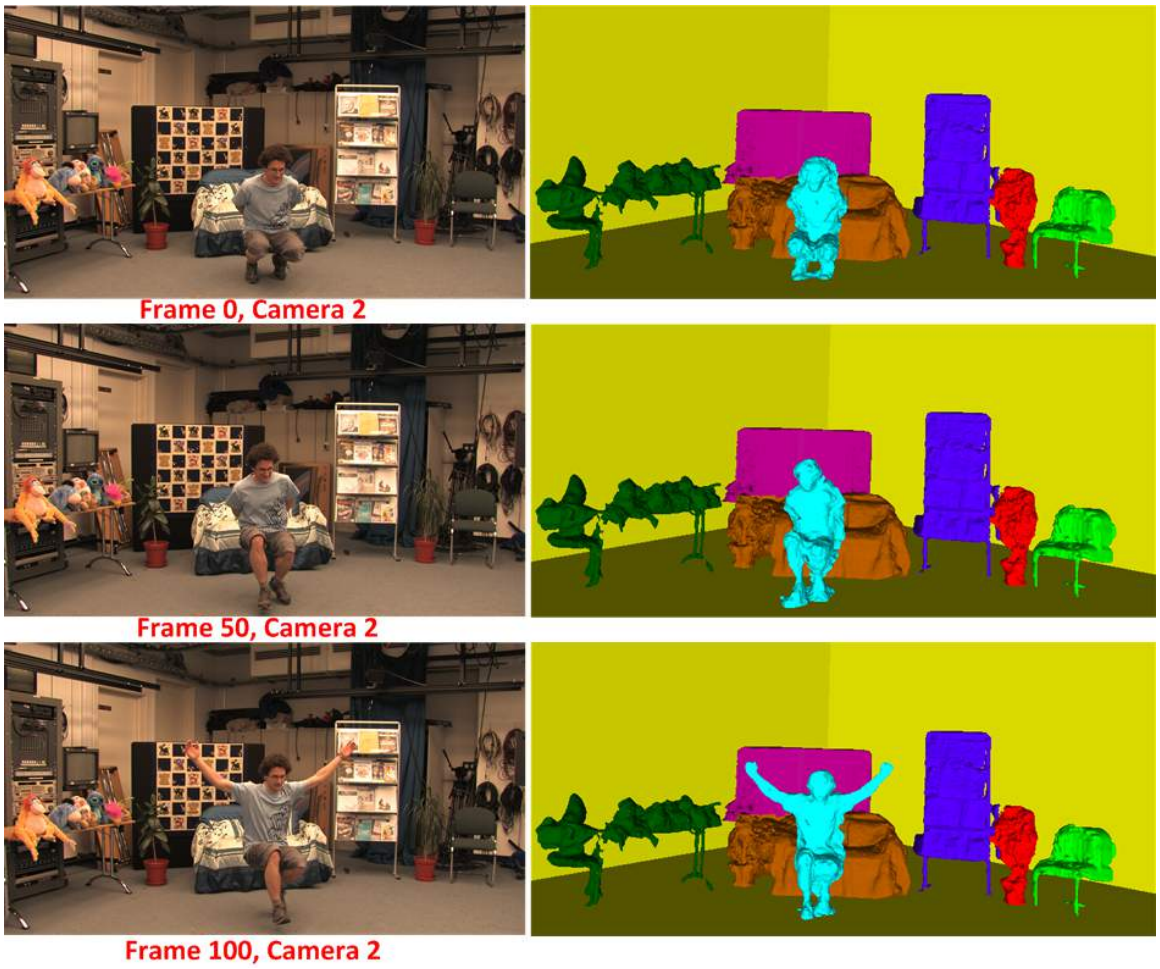


Fig. 5.1 Temporally consistent scene reconstruction for Odzemok dataset color-coded to show the scene object segmentation obtained.

A number of approaches have been introduced for joint optimization. However, these are either limited to static scenes [66, 193] or process each frame independently thereby failing to enforce temporal consistency [30, 62, 122]. A joint formulation for multi-view video was proposed for sports data and indoor sequences in [62] and for challenging outdoor scenes in [122] . Recent work proposed joint reconstruction and segmentation on monocular video achieving semantic segmentation of static scenes [92]. Other joint segmentation and reconstruction approaches that use temporal information based on patch refinement [135, 158] work only for rigid objects. An approach based on optical flow and graph-cuts was shown to work well for non-rigid objects in indoor settings but requires silhouettes and is computationally expensive [63]. Practical application of temporally coherent joint estimation requires approaches that work on non-rigid objects for general scenes in uncontrolled environments.

Methods to estimate 3D scene flow have been reported in the literature [113]. However existing approaches are limited to narrow-baseline correspondence for dynamic scenes. Scene flow approaches dependent on optical flow [16, 187] require an accurate motion estimate for most of the pixels which fails in the case of large motion.

The approach presented in this chapter is for general dynamic indoor or outdoor scenes with large non-rigid motions and no prior knowledge of scene structure. The target scenes are challenging natural outdoor scenes captured with only hand-held cameras, repetitive texture, uncontrolled illumination, fast scene motion, and large capture volume. Temporal correspondence and reconstruction are simultaneously estimated to produce a 4D model of the complete scene with both static and dynamic objects. The proposed approach overcomes the limitations of previous methods enabling robust wide-baseline spatio-temporal reconstruction and segmentation of general scenes. Temporal correspondence is exploited to overcome visual ambiguities giving improved reconstruction together with temporally coherent 4D scene models.

5.2.2 Temporally Consistent Multi-view Video Segmentation

In the field of image segmentation, approaches have been proposed to provide impressive temporally consistent video segmentation [60, 127, 136, 198]. Hierarchical segmentation based on graphs was proposed in [60]. Directed acyclic graphs were used for object proposal followed by segmentation in [198] and [127, 136] used optical flow. All of these methods work only for monocular videos. Recently a number of approaches have been proposed for multi-view foreground object segmentation exploiting appearance similarity between views [38, 89, 98, 196] . An approach that propagates segmentation coherence information in both space and time, allowing evidences in one image to be shared over the complete set was proposed by [39]. These approaches assume a static background and different color distributions for the foreground and background which limits applicability for general complex scenes and non-rigid objects.

To address this issue a novel method for spatio-temporal multi-view segmentation of dynamic scenes using shape constraints is introduced. The proposed approach performs multi-view video segmentation by initializing the foreground object model using spatio-temporal information from wide-baseline feature correspondence followed by a multi-layer optimization framework using geodesic star convexity to constrain the segmentation. Geodesic star convexity has previously been shown to give good results for single image segmentation of complex shapes [64] with manual interaction. The proposed multi-view formulation naturally enforces coherent segmentation between views and also resolves ambiguities such as the similarity of background and foreground in isolated views.

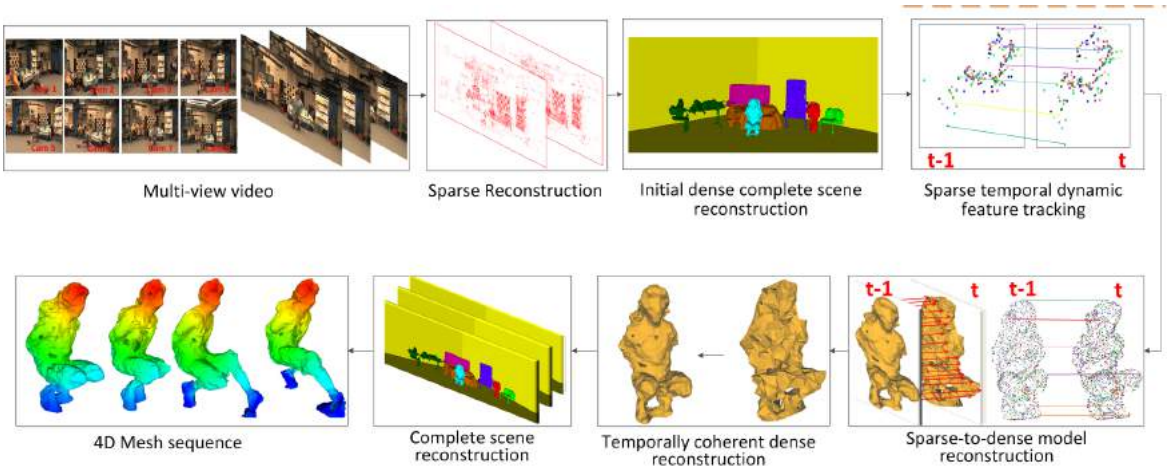


Fig. 5.2 Overview of temporally consistent scene reconstruction framework

5.2.3 Summary of Previous Work

Existing dense reconstruction algorithms need a strong initial prior for the solution to converge. Priors used in previous work include background appearance, scene structure and initial segmentation. These methods give per frame reconstruction of the scene, which is temporally incoherent. In this chapter we propose to obtain temporally coherent segmentation and reconstruction of dynamic scenes captured using uncalibrated multi-view static or moving cameras automatically. The static and dynamic objects in the scene are identified for simultaneous segmentation and reconstruction. Temporal coherence is introduced to improve the quality of the reconstruction and a geodesic star convexity constraint is used to improve the quality of segmentation. The static and dynamic elements are fused automatically in both the spatial and temporal domain to obtain the final 4D scene reconstruction. Figure 5.2 presents an overview of the temporally coherent reconstruction framework.

5.3 Methodology

This work is motivated by the limitations of existing multi-view reconstruction methods which either work independently at each frame resulting in errors due to visual ambiguity and occlusion [51, 62, 122], or commonly require restrictive assumptions on scene complexity and structure [63, 170], or gives reconstruction from single view video [151]. We address these issues by introducing temporal coherence in the reconstruction and geodesic star convexity in segmentation to reduce ambiguity, ensure consistent non-rigid structure initialization at successive frames and improve shape estimate quality.

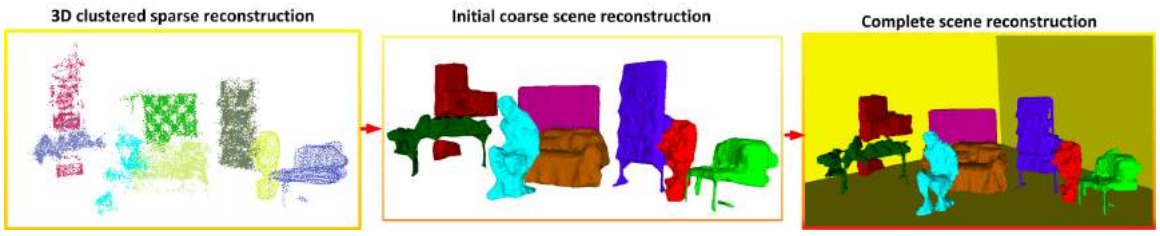


Fig. 5.3 Overview of stages for estimation of an initial dense scene reconstruction.

5.3.1 Overview

A novel automatic multi-object dynamic segmentation and reconstruction method based on the geodesic star convexity shape constraint is proposed to obtain a 4D model of the scene including both dynamic and static objects. An overview of the framework is presented in Figure 5.2 :

Sparse reconstruction: The sparse point-cloud based on SFD features is clustered in 3D [152] with each cluster representing a unique foreground object as explained in Chapter 4. Objects with insufficient detected features are reconstructed as part of the scene background.

Initial dense reconstruction of complete scene: Sparse reconstructions at each time instant are clustered in 3D[38] to obtain an initial coarse object segmentation. This reconstruction is refined using the framework explained in Section 5.3.3 to obtain segmentation and dense reconstruction of each object.

Accurate reconstruction of the background is often challenging due to the lack of features, repetitive texture, occlusion, texture-less regions and relatively narrow-baseline for distant objects. 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) [37]. Principal component analysis (PCA) is applied to the sparse 3D point-cloud obtained in Chapter 3 to retrieve the background. Different methods are used for background estimation for indoor and outdoor scenes. For 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 [36] is applied and the process used for estimation of the background is described below:

- The centroid $\mathbf{A} = (a0, a1, a2)$ and normalized covariance of the point-cloud are estimated to compute the eigenvectors $\vec{ev} = (ev0, ev1, ev2))$ for the covariance matrix of the point-cloud (PCA). We define the reference system as $\mathbf{R}$= $(ev0, ev1, ev0 \times ev1)$ such that: $ev0 \times ev1 = +/-ev2$. The sparse points are mapped in reference frame using $\mathbf{R}$ as the rotation matrix and $\mathbf{A}$ as the translation.

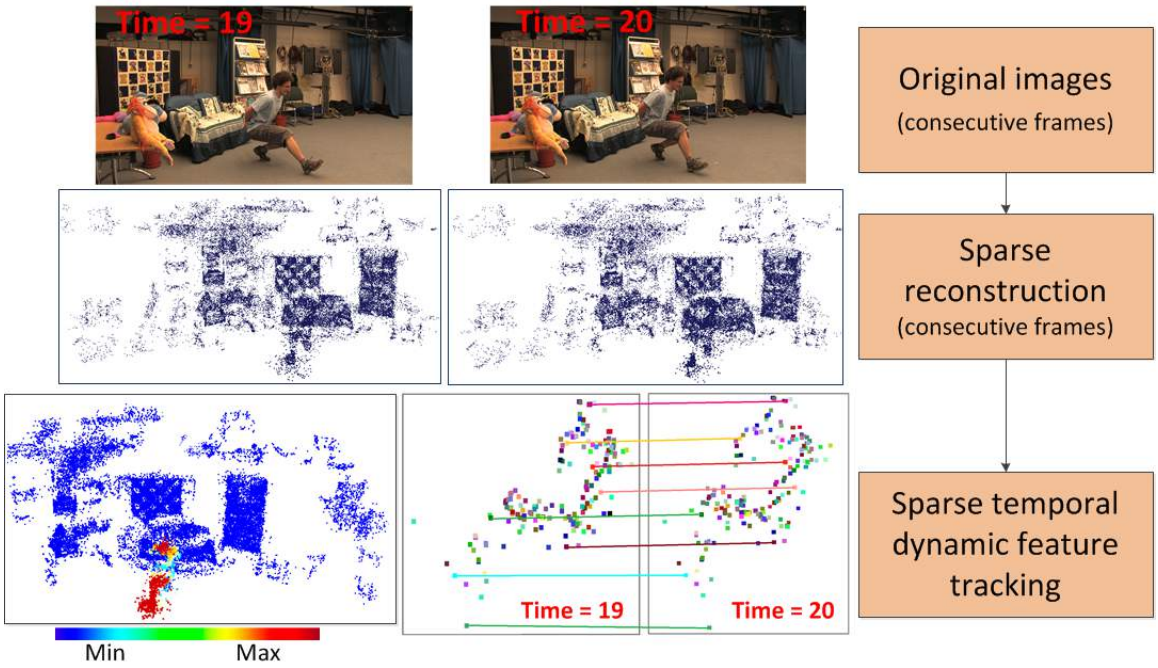


Fig. 5.4 Sparse temporal dynamic feature tracking algorithm: Results on Odzemok dataset; Min and Max is the minimum and maximum movement in the 3D points respectively.

- The rotation and translation is calculated using the eigenvectors to place a box in correct location. The minimum and maximum values of coordinates in x, y and z direction 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 **R** and translation $\mathbf{R} \times \mathbf{CP} + \mathbf{A}$ is applied, where **CP** 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 provide complete scene reconstruction.

The dense reconstruction of the foreground objects and background are combined to obtain a full scene reconstruction at the first time instant. The algorithm is shown in Figure 5.3. For consecutive time instants only dynamic objects are reconstructed and segmentation and reconstruction of static objects is retained which reduces computational complexity.

Temporally coherent reconstruction of dynamic objects: Temporal coherence is introduced in the framework to initialize the coarse reconstruction and obtain frame-to-frame dense correspondences. 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

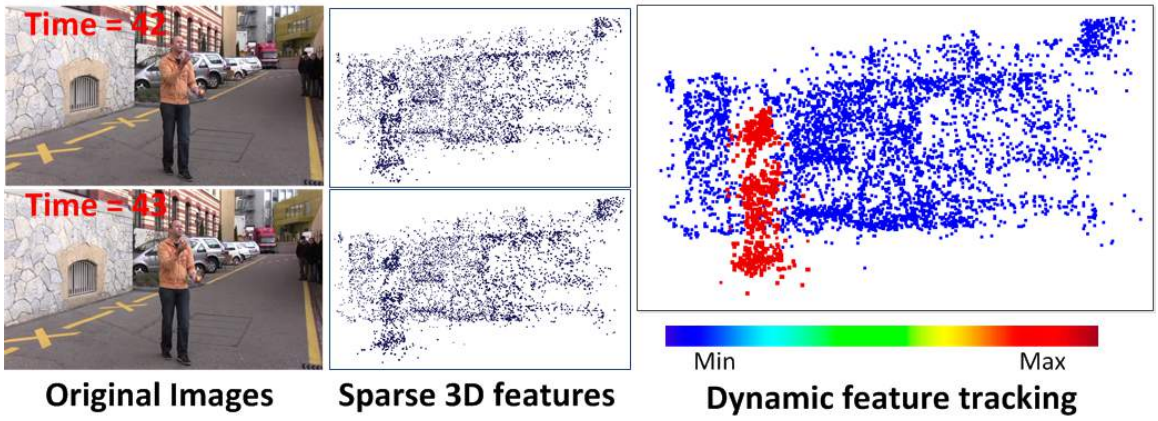


Fig. 5.5 Sparse temporal dynamic feature tracking for Juggler dataset captured with only moving cameras. Min and Max is the minimum and maximum movement in the 3D points respectively.

to obtain an initial approximation (Section 5.3.2). The initial estimate is refined using a joint optimization of segmentation and reconstruction based on geodesic star convexity (Section 5.3.3). A single 3D model for each dynamic object is obtained by fusion of the view-dependent depth maps using Poisson surface reconstruction [81].

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

5.3.2 Initial Temporally Coherent Reconstruction

Once the static scene reconstruction is obtained for the first frame, temporally coherent dynamic scene reconstruction is performed at successive time instants. Dynamic regions are identified using temporal correspondence of sparse 3D features. These points are used to obtain an initial dense model for the dynamic objects based on optical flow correspondence estimation. The initial coarse reconstruction for each dynamic region is refined in the subsequent optimization step with respect to each camera view. Dynamic scene objects are identified from the temporal correspondence of sparse feature points. Sparse correspondence is then used to propagate an initial model of the moving object for refinement. Figure 5.4 presents the sparse reconstruction and temporal correspondence.

Sparse temporal dynamic feature tracking: Numerous approaches have been proposed to track moving objects in 2D using either features or optical flow. However these methods

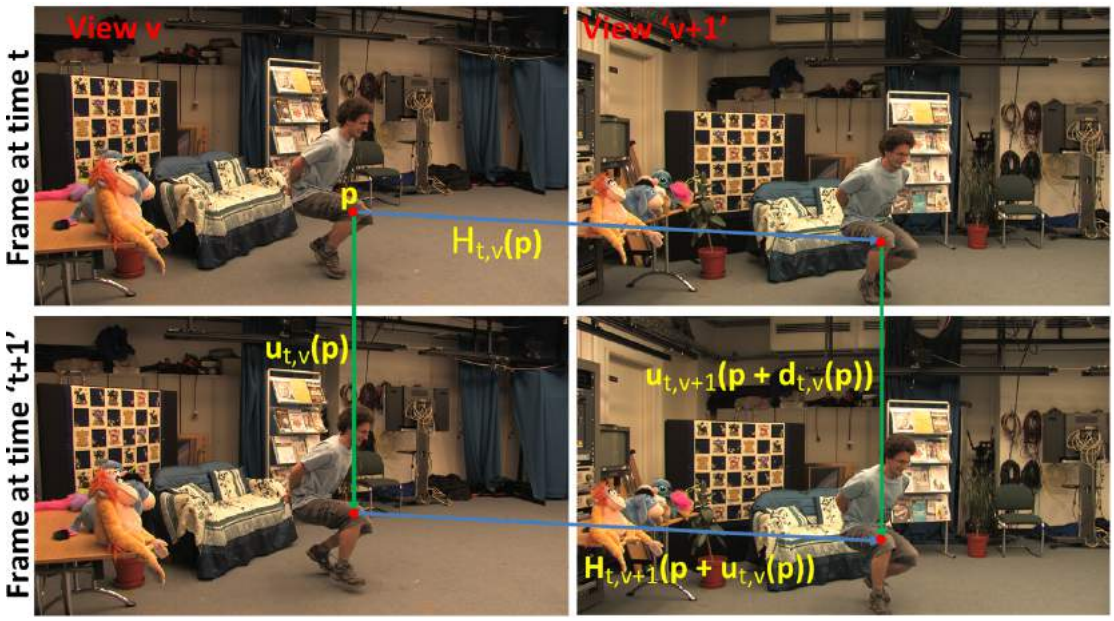


Fig. 5.6 Spatio-temporal consistency check for 3D tracking for Odzemok dataset.

may fail in the case of occlusion, movement parallel to the view direction, large motions and moving cameras. To overcome these limitations the sparse 3D feature points obtained using SFD from multiple wide-baseline views are matched at each time instant. The use of sparse 3D features is robust to large non-rigid motion, occlusions and camera movement. Sparse 3D feature matches between consecutive time instants are back-projected to each view. These features are matched temporally using a SIFT descriptor to identify the moving points. Robust matching is achieved by enforcing multi-view consistency for the temporal feature correspondence in each view as illustrated in Figure 5.6. Each match must satisfy the constraint:

$$\left\| H_{t,v}(p) + u_{t,v+1}(p + H_{t,v}(p)) - u_{t,v}(p) - H_{t,v+1}(p + u_{t,v}(p)) \right\| < \varepsilon$$

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 $v+1$, $u_{t,v}(p)$ is the temporal correspondence from frames t to $t+1$ for view v. The multi-view consistency check ensures that correspondences between any two views remain temporally consistent for successive frames. Matches in the 2D domain are sensitive to camera movement and occlusion, hence the set of refined matches are mapped in 3D to make the system robust to camera motion. The Frobenius norm is applied on the 3D

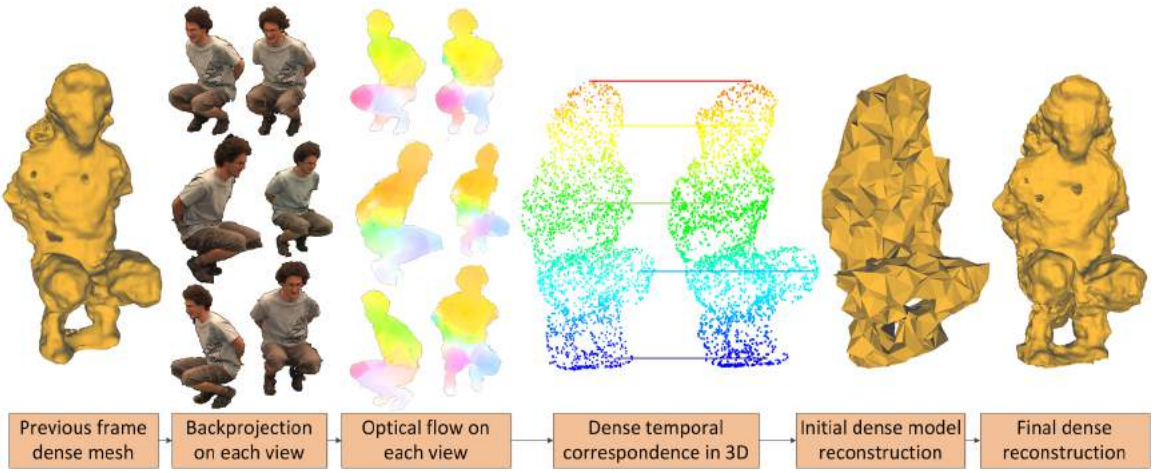


Fig. 5.7 Overview of initial sparse-to-dense model reconstruction for the Odzemok dataset.

point gradients in all directions to obtain the 'net' motion at each sparse point.

$$\|O\|_F = \left\| \begin{bmatrix} u_x & u_y \\ v_x & v_y \\ w_x & w_y \end{bmatrix} \right\|_F = \sqrt{u_x^2 + u_y^2 + v_x^2 + v_y^2 + w_x^2 + w_y^2} \tag{5.1}$$

A 3D extension of the 2D frobenius norm [198] is used to calculate the motion. The 'net' motion between two 3D points for consecutive time instants are ranked, and top and bottom 5 percentile values removed. Median filtering is then applied to identify the dynamic features. Figure 5.5 shows an example with moving camera.

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 initialize multi-view reconstruction with a coarse visual hull approximation [62]. Temporal coherence also provides a more accurate initialization to overcome visual ambiguities at individual frames. Figure 5.7 illustrates the use of temporal coherence for reconstruction initialization and refinement. Dynamic feature correspondence is used to identify the mesh for each dynamic object. This mesh is back projected on each view to obtain the region of interest(silhouette) at each time instant. Dense optical flow [194] is performed on the projected mask(silhouette) for each view in the temporal domain using the dynamic feature correspondences over time as initialization. The dense multi-view wide-baseline correspondences from the previous frame are propagated to the current frame using the information from the flow vectors to

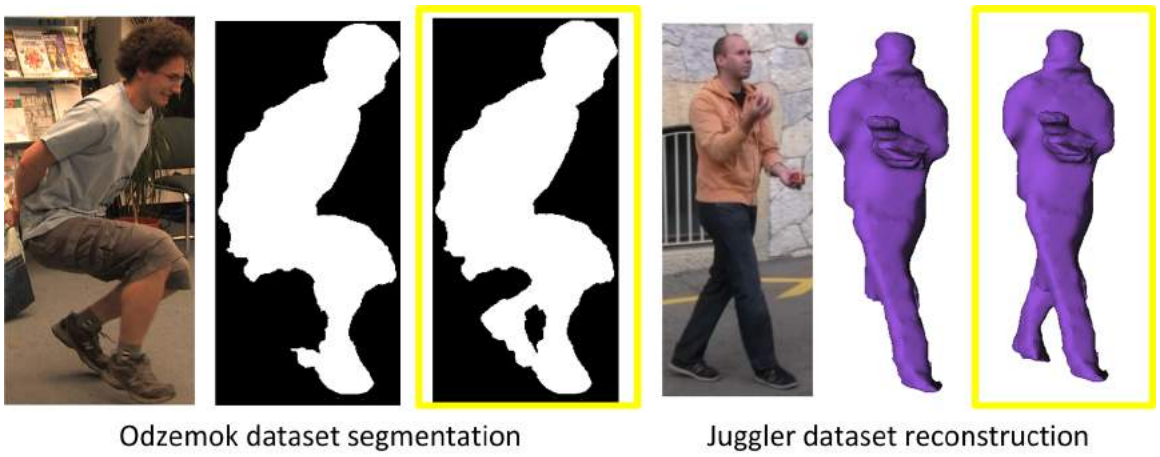


Fig. 5.8 Improvement in segmentation for the Odzemok dataset and reconstruction for the Juggler dataset with temporal coherence (highlighted in yellow)

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, information from the cluster of sparse points is used 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 new clusters are detected at each time instant and consider them as dynamic regions.

Once a set of dense 3D points is obtained 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 R, which is subsequently refined using the optimization framework (Section 5.3.3). Introduction of sparse-to-dense initial coarse reconstruction followed by the optimization framework improves the quality of segmentation and reconstruction. Examples of the improvement in segmentation and reconstruction for Odzemok and Juggler dataset are shown in Figure 5.8. As observed limbs of the object can be retained by using information from the previous frames in both the cases.

Fig. 5.9 Representation of star convexity: The left object depicts example of star convex object, with a star center marked in green. The object on the right with a plausible star center shows deviations from star convexity in the fine details.

5.3.3 Geodesic Star Convexity for Joint Refinement

Shape is a powerful cue for object recognition and segmentation. Shape models represented as some kind of distance transform from a template have been used for category specific segmentation [85]. Some works have introduced generic connectivity constraints for segmentation showing that obtaining globally optimal solutions under the connectivity constraint is NP-hard [184]. 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 [181]. The star convexity constraint ensures connectivity to seed points, and is a stronger assumption than plain connectivity. An example of star convex object is shown in Figure 5.9 along with a failure case for non-rigid articulate object. To handle more complex objects the idea of geodesic forests was introduced to obtain globally optimal solutions in [64]. 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 as geodesic 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.

In this work this concept is automatically applied in our joint segmentation and reconstruction refinement 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 peoples fingers or hair)[64]. The shape constraint is automatically initialized for each view from the initial segmentation. The geodesic star convexity is integrated as a constraint on the energy minimization for joint multi-view refinement. The shape constraint is based on the geodesic distance with foreground object initialization (seeds) as star centers 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 the sparse temporal feature correspondence are used as star centers (seeds) to build a geodesic forest automatically. The region outside the initial coarse reconstruction of all dynamic objects is initialized as the background seed for segmentation as shown in in Figure 5.11. 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 5.15.

Optimization based on Geodesic Star Convexity

Graph-cuts is used for joint refinement of reconstruction and segmentation. A graph structure is built $\mathscr{G} = \{\mathscr{V}, \mathscr{E}\}$ where $v_i \in \mathscr{V}$ are the vertices and $\{v_i, v_j\} \in \mathscr{E}$. Each vertex v_i represents a pixel and the depth of the initial coarse reconstruction estimate is refined for each dynamic object at a per pixel level. Our goal is to assign an accurate depth value from a set of depth values $\mathscr{D} = \{d_1, ..., d_{|\mathscr{D}|-1}, \mathscr{U}\}$ 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. Each edge e_i is assigned a non-negative weight followed by optimization of a joint cost function [62] for label (segmentation) and depth (reconstruction):

$$E(l,d) = \lambda_{data}E_{data}(d) + \lambda_{contrast}E_{contrast}(l) + \lambda_{smooth}E_{smooth}(l,d) + \lambda_{color}E_{color}(l) \quad (5.2)$$

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 5.3.3. 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 $p \in l$

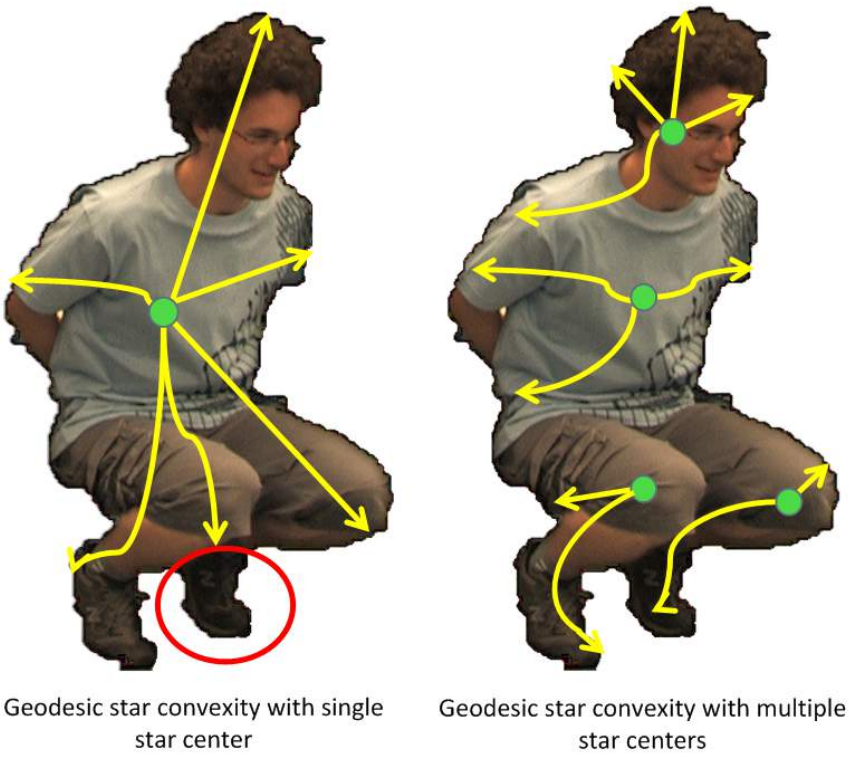


Fig. 5.10 Geodesic star convexity based segmentation: Left: Single star center and Right: Multiple star centers. The error with single star center based segmentation is highlighted in red.

is visible to a star center c via l in the image x which can be expressed as an energy cost:

$$E^{\star}(l|x,c) = \sum_{p \in R} \sum_{q \in \Gamma_{c,p}} E^{\star}_{p,q}(l_p, l_q) \tag{5.3}$$

$$\forall q \in \Gamma_{c,p},\ E^{\star}_{p,q} = \begin{cases} \infty \text{ if } l_p \neq l_q \\ 0 \text{ otherwise} \end{cases} \tag{5.4}$$

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:

$$\Gamma_{c,p} = \arg\min_{\Gamma \in P_{c,p}} L(\Gamma) \tag{5.5}$$

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 [64]. In the case of image segmentation the gradients in the underlying image provide information to compute the discrete paths between each pixel

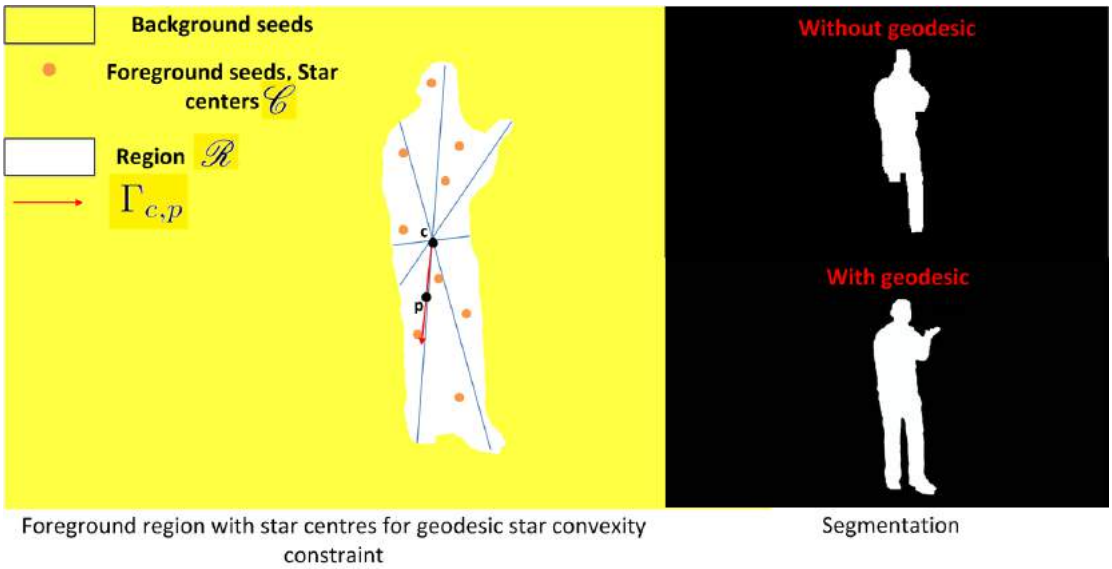


Fig. 5.11 Geodesic star convexity: A region $\mathscr{R}$ with star centers $\mathscr{C}$ connected with geodesic distance $\Gamma_{c,p}$. Segmentation results with and without geodesic star convexity based optimization are shown on the right for the Juggler dataset.

and star centers and $L(\Gamma)$ is defined below:

$$L(\Gamma) = \sum_{i=1}^{N_D-1} \sqrt{(1-\delta_g) j(\Gamma^i, \Gamma^{i+1})^2 + \delta_g \left\| \bigtriangledown I(\Gamma^i) \right\|^2} \tag{5.6}$$

where Γ is an arbitrary parametrized discrete path with N_D pixels given by $\left\{\Gamma^1, \Gamma^2, \cdots \Gamma^{N_D}\right\}$, $j(\Gamma^i, \Gamma^{i+1})$ is the Euclidean distance between successive pixels, and the 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 5.5.

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 for objects with concavities like small limbs as shown in Figure 5.10. Hence Equation 5.3 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 [64]. This is expressed as an energy cost:

$$E^\star(l|x, \mathscr{C}) = \sum_{p \in R} \sum_{q \in \Gamma_{c,p}} E^\star_{p,q}(l_p, l_q) \tag{5.7}$$

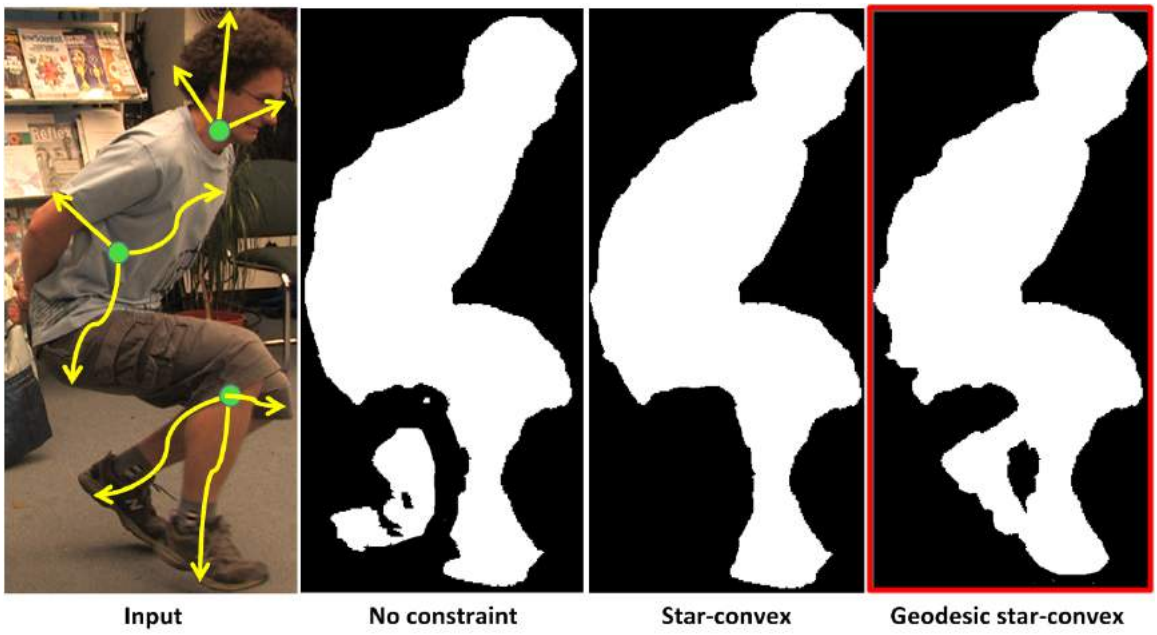


Fig. 5.12 Segmentation comparison results with no constraint, star convexity constraint and geodesic star convexity constraint for Odzemok dataset.

In our case the 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 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 constrains and Euclidean multi-star convexity constraint is shown in Figure 5.12. 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 5.2 is minimized as follows:

$$min_{\substack{(l,d) \\ s.t.}} \underset{l \varepsilon S^{\star}(\mathscr{C})}{E(l,d)} \Leftrightarrow \min_{(l,d)} E(l,d) + E^{\star}(l|x,\mathscr{C}) \tag{5.8}$$

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 eq. 5.8, 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 the set of labels in $\mathscr{L} \times \mathscr{D}$ [25]. Graph-cut is used to obtain a local optimum [24].

Energy Cost Function

For completeness in this section each of the terms in Equation 5.2 are defined, these are based on previous terms used for joint optimization over depth for each pixel introduced in [122](Chapter 4), with modification of the color matching term to improve robustness and extension to multiple labels.

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

$$E_{data}(d) = \sum_{p \in N_P} e_{data}(p, d_p) = \begin{cases} M(p,q) = \sum_{i \in N_C} m(p,q), & \text{if } d_p \neq \mathscr{U} \\ M_{\mathscr{U}}, & \text{if } d_p = \mathscr{U} \end{cases} \tag{5.9}$$

where N_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 hypothesized point P in an auxiliary camera where P is $3D$ point along the optical ray passing through pixel p located at a distance d_p from the reference camera. N_C is the set of most photo-consistent pairs with reference camera and $m(p,q)$ is the inspired from [72].

Contrast term: The contrast term naturally encourages low contrast regions to coalesce into layers and favours discontinuities to follow strong edges. It is defined as follows:

$$E_{contrast}(l) = \sum_{p,q \in \mathscr{N}} e_{contrast}(p, q, l_p, l_q) \tag{5.10}$$

$$e_{contrast}(p, q, l_p, l_q) = \begin{cases} 0, & \text{if } (l_p = l_q) \\ \frac{1}{1+\varepsilon}(\varepsilon + exp^{-J(p,q)}), & \text{otherwise} \end{cases} \tag{5.11}$$

$\|\cdot\|$ is the L_2 norm, $\varepsilon = 1$. $S(p,q)$ is defined in Chapter 4, Equation 4.7.

Smoothness term: This term is defined as:

$$E_{smooth}(l, d) = \sum_{(p,q) \in \mathscr{N}} e_{smooth}(l_p, d_p, l_q, d_q) \tag{5.12}$$

$$e_{smooth}(l_p, d_p, l_q, d_q) = \begin{cases} min(\left|d_p - d_q\right|, d_{max}), & \text{if } l_p = l_q \text{ and } d_p, d_q \neq \mathscr{U} \\ 0, & \text{if } l_p = l_q \text{ and } d_p, d_q = \mathscr{U} \\ d_{max}, & \text{otherwise} \end{cases} \tag{5.13}$$

d_{max} is set to 50 times the size of the depth sampling step defined in Section 5.3.3 for all datasets and the penalty is defined as a truncated linear distance within each layer. Such a distance is discontinuity preserving as it does not over-penalize large discontinuities within a

layer; this is known to be superior to simpler non-discontinuity functions [25]. This term also encourages unknown features to coalesce within each layer.

Color term: The color term uses learnt Gaussian mixture models (GMM) for each layer or group of layers following a similar distribution in order to assign the most likely layer label at each pixel in the reference image. This term is computed using the negative log likelihood [24] of the color models learned from the initially segmented foreground and background regions. The star centers obtained from the sparse 3D features are foreground markers and for background markers the region outside the projected initial coarse reconstruction is considered for each view. The color term is defined as:

$$E_{color}(l) = \sum_{p \in \mathscr{P}} -logP(I_p|l_p) \tag{5.14}$$

where $P(I_p|l_p = l_i)$ denotes the probability at pixel p in the reference image belonging to layer l_i. Existing methods have used a combination of global and local color models to compute the probability [62], but the local color model is applicable to static layers only like the background. In our case we handle moving hand-held cameras where the background pixel changes. Hence a global color model is used for computing the probabilities. The global color model for a layer l_i is defined as:

$$P(I_p|l_p = l_i) = \sum_{k=1}^{K_i} w_{ik} N(I_p|\mu_{ik}\varphi_{ik}) \tag{5.15}$$

where $N()$ is the normal distribution and the parameters w_{ik}, μ_{ik} and φ_{ik} represent the weight, the mean and the covariance matrix of the k^{th} component for layer l_i. K_i is the number of components of the mixture model for layer l_i. The color models use Gaussian mixture models with $K_i = 10$ components each for foreground/background mixed with uniform color models in RGB color space. The models are not updated over time as our algorithm is only applied to relatively short sequences (500 frames). For the processing of longer sequences, more sophisticated algorithms could be used to model temporal variations in illuminations.

The improvements in the results using geodesic star convexity in the framework is shown in Figure 5.12 and by using temporal coherence is shown in Figure 5.8. Figure 5.13 shows improvements using geodesic shape constraint, temporal coherence and combined proposed approach for Dance2 dataset. The segmentation is compared against ground-truth (publically available), red pixels denote error in the segmentation. The quality of the segmentation results obtained using the proposed approach by combining both GSC and temporal coherence is improved.

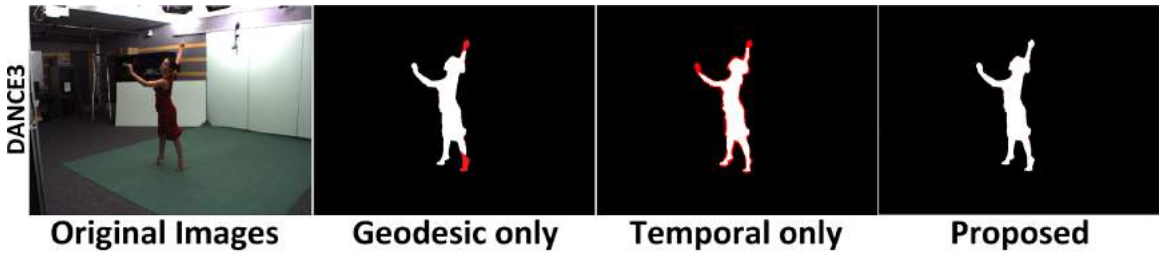


Fig. 5.13 Comparison of segmentation with introduction of temporal coherence, Geodesic star convexity(GSC) and proposed method (GSC and temporal coherence) for Dance2 dataset.

The energy terms in the Equation 5.2 can be constrained temporally to improve the performance further. Constraint can be applied on the pairwise terms as shown in [63] to include temporal contrast and smoothness information in the optimization.

5.4 Results and Evaluation

The proposed system is tested on publicly available multi-view research datasets of indoor and outdoor scenes: static data for segmentation comparison Couch, Chair and Car[89]; and dynamic data for full evaluation Dance3[4DI], Office, Dance2, Odzemok, Magician and Juggler [15]. Magician and Juggler from 6 moving hand-held cameras. The parameters used for experiments for all datasets are listed in Table 5.1. The parameters are set empirically.

Datasets	λ_{data}	$\lambda_{contrast}$	λ_{smooth}	λ_{color}
Magician/Dance3	1.0	12.5	.00125	1.5
Juggler	1.0	10.0	.001	0.8
Odzemok/Dance2/Office	1.0	7.5	.0025	1.5

Table 5.1 Parameter settings used in Equation 5.2 for reconstruction of all the datasets.

The evaluation performed using these datasets covers segmentation and reconstruction comparison against state-of-the-art methods in multi-view segmentation [38, 89], reconstruction [51] and joint segmentation and reconstruction [63, 122].

5.4.1 Multi-view Segmentation Evaluation

Segmentation is evaluated against state-of-the-art methods for multi-view segmentation for both static and dynamic data. Comparison is performed against ground-truth data for all datasets. Ground-truth is obtained by manually labelling the foreground for Office, Dance2

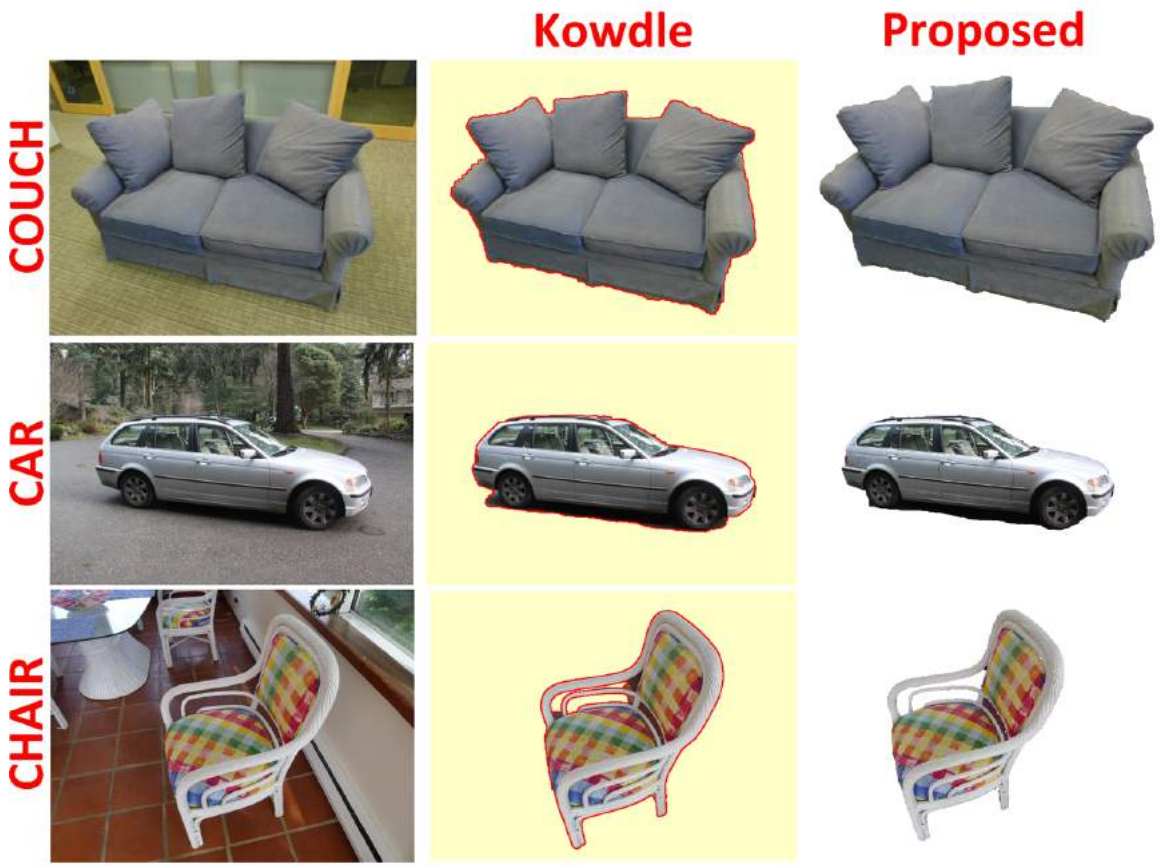


Fig. 5.14 Comparison of segmentation with Kowdle on benchmark static datasets using geodesic star convexity.

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.

Static scene segmentation methods: Comparison is performed against static multi-view segmentation methods Kowdle [89] and Djelouah [38] for static scenes. The segmentation is initialized as detailed in Section 5.3.1 followed by refinement using the constrained optimization Section 5.3.3. Qualitative results are shown in Figure 5.14 and 5.15 for static benchmark datasets. For quantitative evaluation the ratio of intersection to union with ground-truth is measured as proposed in [89] and the comparison is shown in Table 5.2. The segmentation quality and accuracy are comparable to the multi-view segmentation technique Kowdle [89] and more accurate than Djelouah [38].

Dynamic scene segmentation methods: Comparison is performed against joint dynamic scene segmentation and reconstruction per frame methods Mustafa [122] and Guillemaut [63] for both static and dynamic scenes. Mustafa is the method from Chapter 4 with no

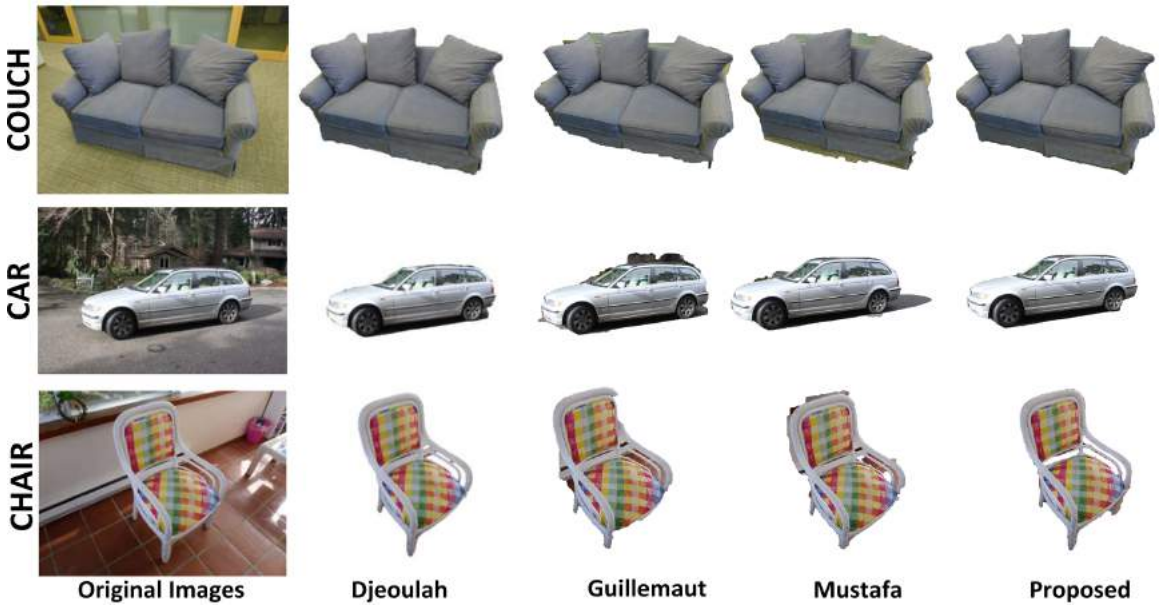


Fig. 5.15 Comparison of segmentation on benchmark static datasets using geodesic star convexity.

Dataset	No of Views	Kowdle	Djelouah	Guillemaut	Mustafa	Proposed
Couch	11	99.6 ± 0.1	99.0 ± 0.2	97.0 ± 0.3	98.5 ± 0.2	$\mathbf{99.7 \pm 0.3}$
Chair	18	$\mathbf{99.2 \pm 0.4}$	98.6 ± 0.3	97.9 ± 0.5	98.0 ± 0.5	99.1 ± 0.3
Car	44	98.0 ± 0.7	97.0 ± 0.8	95.0 ± 0.7	97.6 ± 0.3	$\mathbf{98.6 \pm 0.4}$

Table 5.2 Static segmentation comparison with existing methods on benchmark datasets

temporal information and Guillemaut[63] uses temporal information to improve the results. The results for the proposed approach are obtained by applying the full pipeline with temporal coherence as detailed in Section 5.3.

Qualitative results are shown in Figure 5.15 for static benchmark datasets and Figure 5.17 and 5.16 for dynamic scenes. Quantitative evaluation is performed same as for static segmentation methods and the comparison for static datasets is shown in Table 5.2 and in Table 5.3 for dynamic scenes. Results for multi-view segmentation of static scenes are more accurate than Djelouah, Mustafa and Guillemaut and are 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 5.17 and 5.16. This overcomes limitations of previous multi-view per frame segmentation.

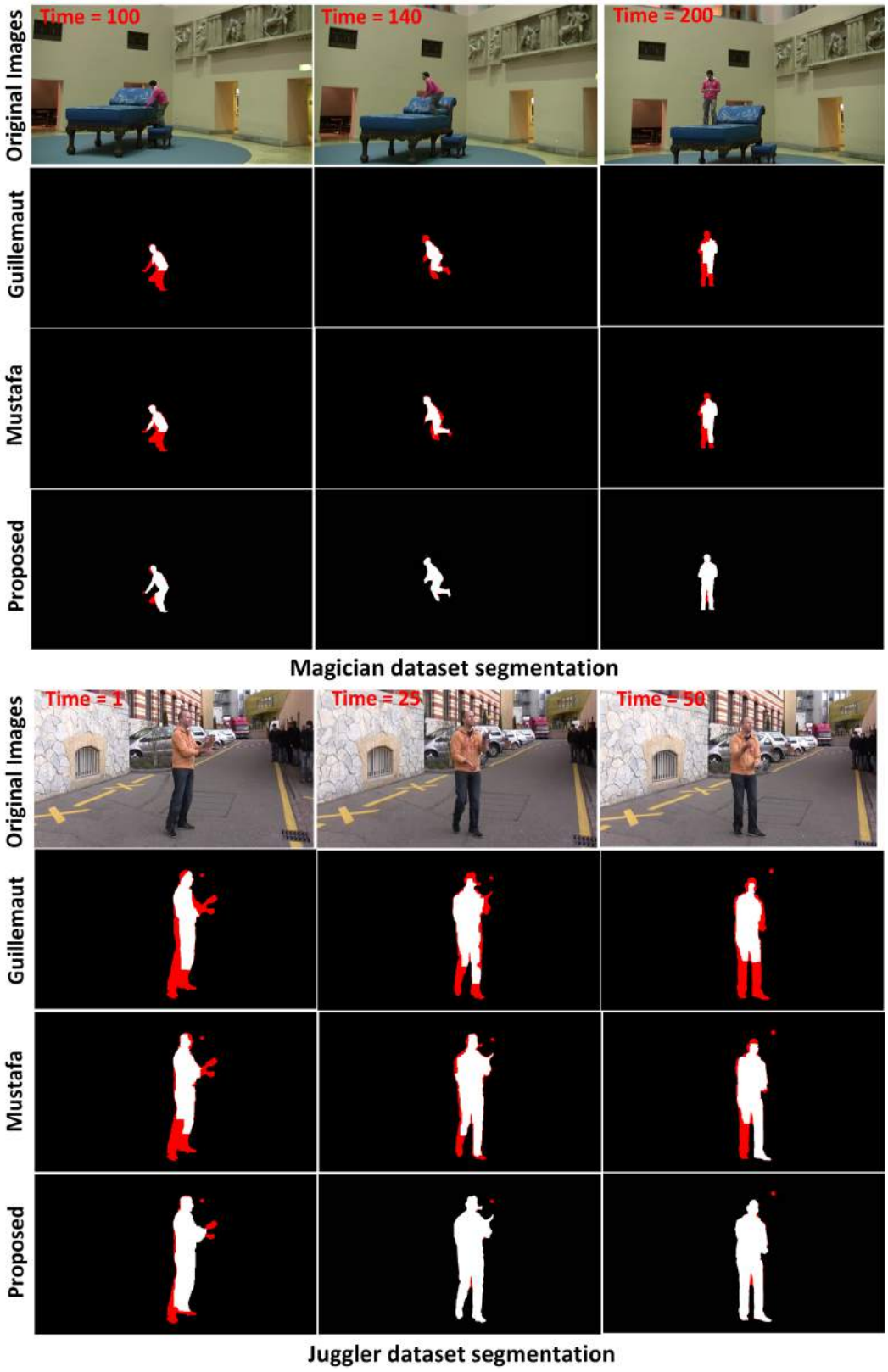


Fig. 5.16 Segmentation results for dynamic scenes on sequence of frames (Error against ground-truth is highlighted in red).

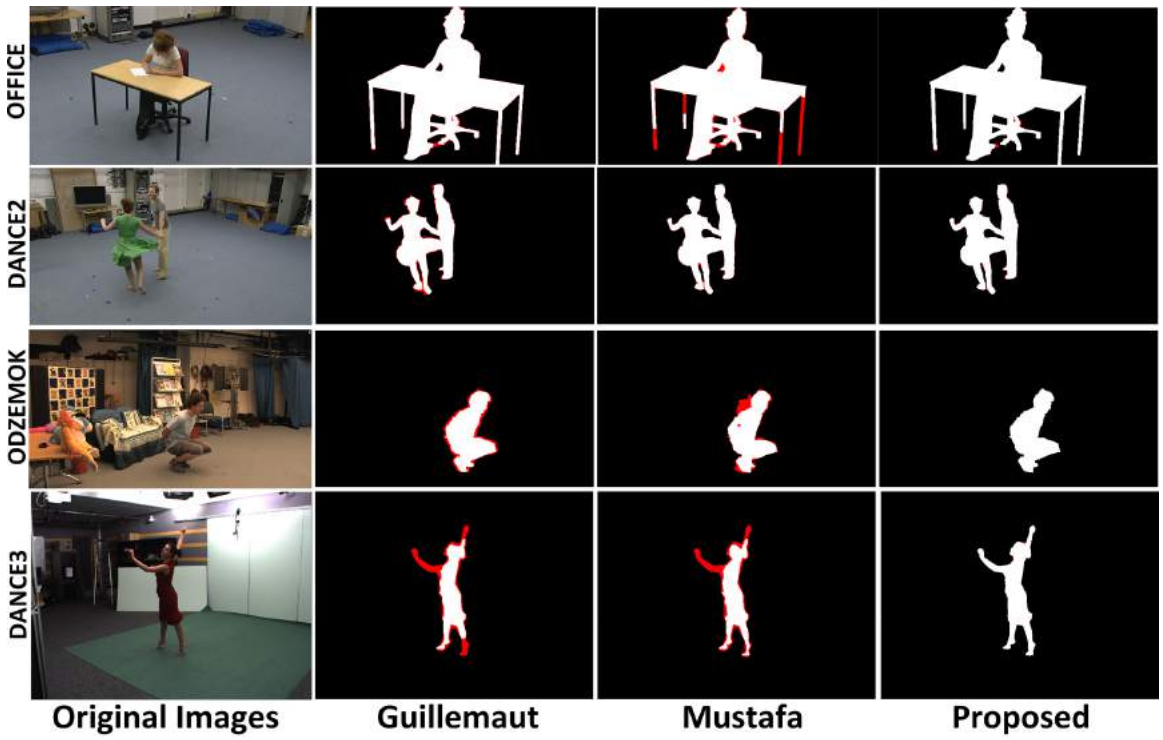


Fig. 5.17 Segmentation results for dynamic scenes (Error against ground-truth is highlighted in red).

Method	No Priors	Temporal coherence	Joint refinement (Segmentation)
Furukawa PAMI 2010	✓	X	X
Guillemaut 3DV 2012	X	✓	✓
Mustafa ICCV 2015	✓	X	✓
Proposed	✓	✓	✓

Fig. 5.18 State-of-the-art methods evaluated against the proposed method. MustafaICCV15 is the method from Chapter 4.

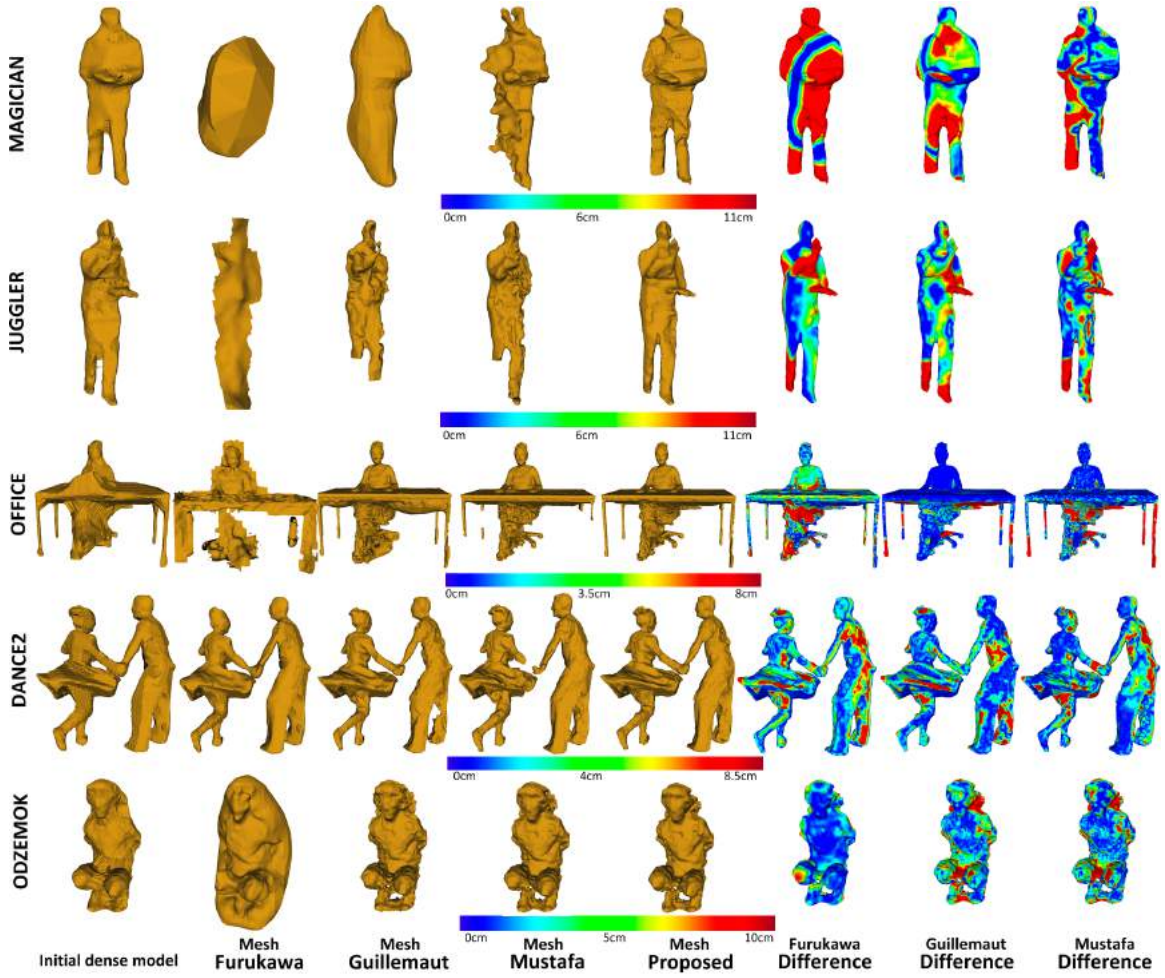


Fig. 5.19 Reconstruction result mesh comparison against state-of-the-art methods. Column 1^{st} represents the initial dense model, Column $2^{nd} - 5^{th}$: Meshes and Column $6^{th} - 8^{th}$: Difference meshes against proposed approach with color coded error in cms.

Dataset	Guillemaut	Mustafa	Proposed
Magician	68.0 ± 0.7	88.7 ± 0.5	$\mathbf{91.2 \pm 0.2}$
Juggler	84.6 ± 0.6	87.9 ± 0.6	$\mathbf{93.3 \pm 0.2}$
Odzemok	90.1 ± 0.3	89.9 ± 0.3	$\mathbf{91.8 \pm 0.2}$
Dance2	99.2 ± 0.5	99.4 ± 0.2	$\mathbf{99.5 \pm 0.2}$
Office	99.3 ± 0.4	99.0 ± 0.3	$\mathbf{99.4 \pm 0.2}$
Dance3	98.6 ± 0.3	99.0 ± 0.2	$\mathbf{99.0 \pm 0.2}$

Table 5.3 Comparison of the segmentation accuracy against ground-truth for dynamic scenes in %. Ground-truth is obtained by manually labelling the foreground for Office, Dance2 and Odzemok dataset, and for other datasets ground-truth is available online.

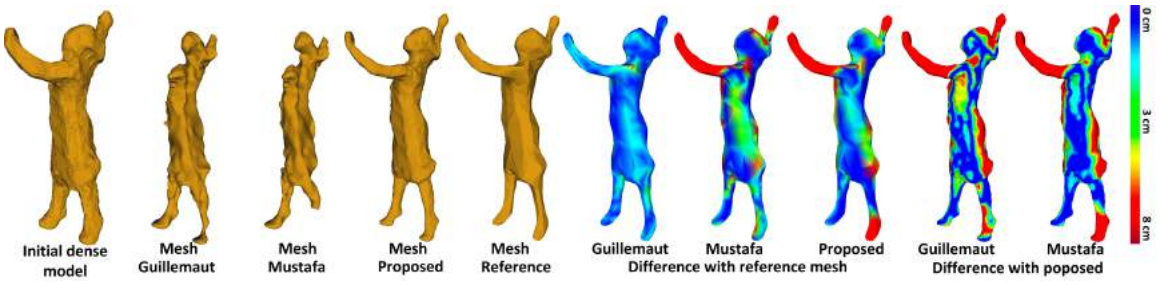


Fig. 5.20 Reconstruction result comparison with reference mesh and proposed for Dance3 benchmark dataset. Column 1^{st} represents the initial dense model, Column $2^{nd} - 4^{th}$: Meshes, Column 5^{th}:Reference mesh available online, Column $6^{th} - 8^{th}$: Difference meshes against reference approach with color coded error in cms and Column $9^{th} - 10^{th}$: Difference meshes against proposed approach.

5.4.2 Reconstruction Evaluation

Reconstruction results obtained using the proposed method with parameters defined in Table 5.1 are compared against Mustafa[122], Guillemaut[63], and Furukawa [51] for dynamic sequences. The state-of-the-art methods are illustrated in Figure 5.18. Furukawa [51] is a per frame multi-view wide-baseline stereo approach which ranks highly on the Middlebury benchmark [156] but does not refine the segmentation. Guillemaut is a joint refinement method which requires background structure and calibration to obtain the solution and uses temporal information in the framework. Mustafa is an automatic joint refinement method and gives per frame reconstruction.

Figure 5.19 and 5.20 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 color 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 5.20 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 5.19 resulting in more complete shape. Figure 5.21 and 5.22 shows complete scene reconstructions with 4D models of multiple objects for various datasets. The Juggler and Magician sequences are

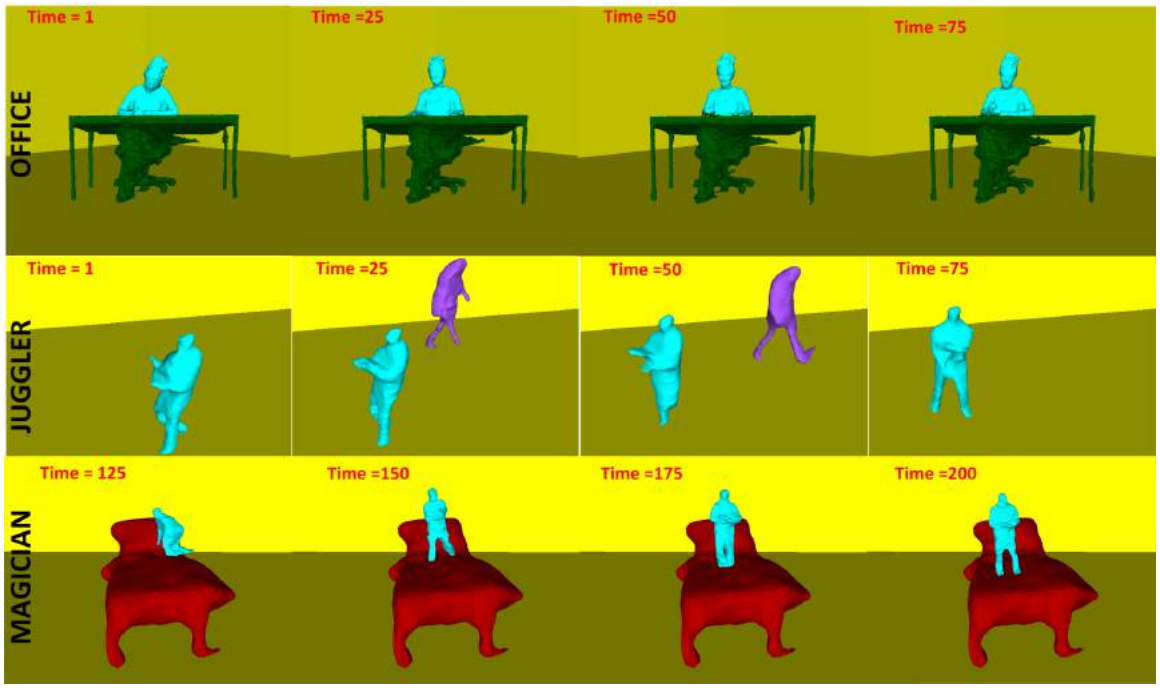


Fig. 5.21 Complete scene reconstruction with 4D mesh sequence.

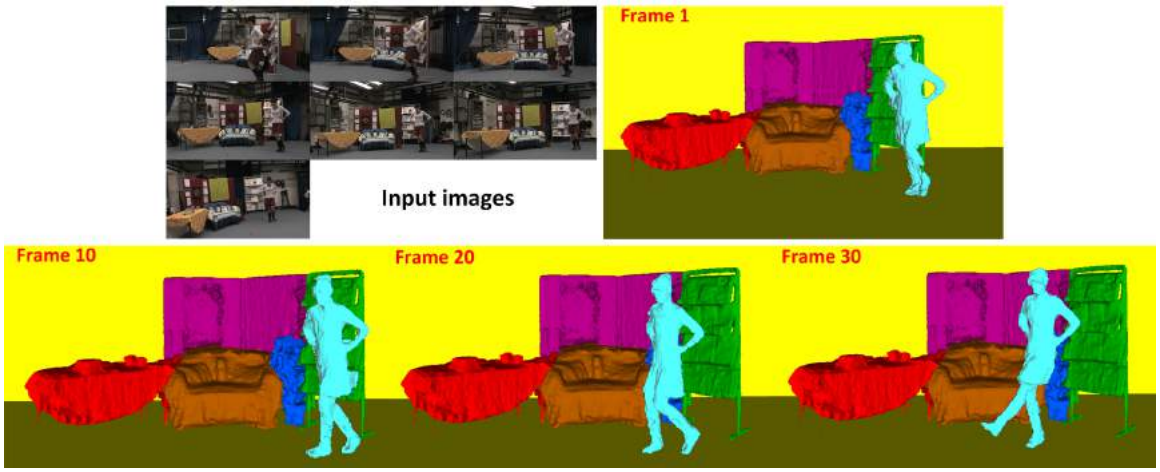


Fig. 5.22 Complete scene reconstruction for Dance1 dataset.

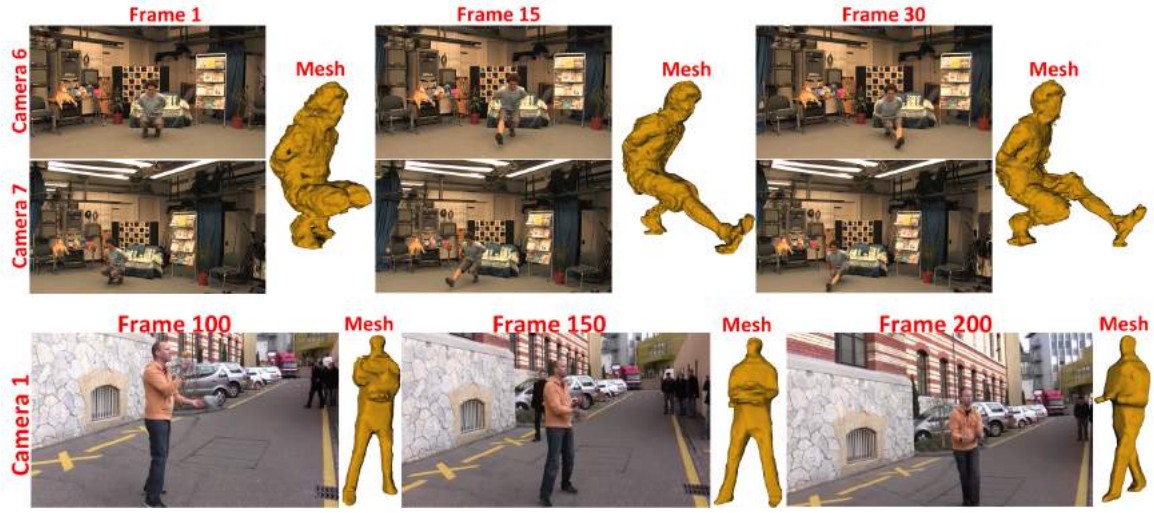


Fig. 5.23 Reconstruction for moving cameras for the Odzemok and Juggler datasets.

reconstructed from moving hand-held cameras. Odzemok and Juggler are scenes with significant movement in cameras. An example is shown in Figure 5.23 where camera moves by approx 0.25 meters and 30 degrees for Juggler dataset and by 15 degrees for Odzemok dataset.

The depth maps obtained using the proposed approach are compared against Mustafa and Guillemaut in Figure 5.24. The depth map obtained using the proposed approach are smoother with low reconstruction noise compared to the state-of-the-art methods.

Computational Complexity: Computation times for the proposed approach against other

Dataset	Furukawa	Guillemaut	Mustafa	Ours
Dance2	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**
Dance3	312 s	432 s	323 s	**278 s**

Table 5.4 Comparison of computational efficiency for dynamic datasets (time in seconds (s))

methods are presented in Table 5.4 for all the datasets. The proposed approach to reconstruct temporally coherent 4D models is comparable in computation time to per frame multi-view reconstruction and gives a $\sim$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 initialization based on temporal propagation and the

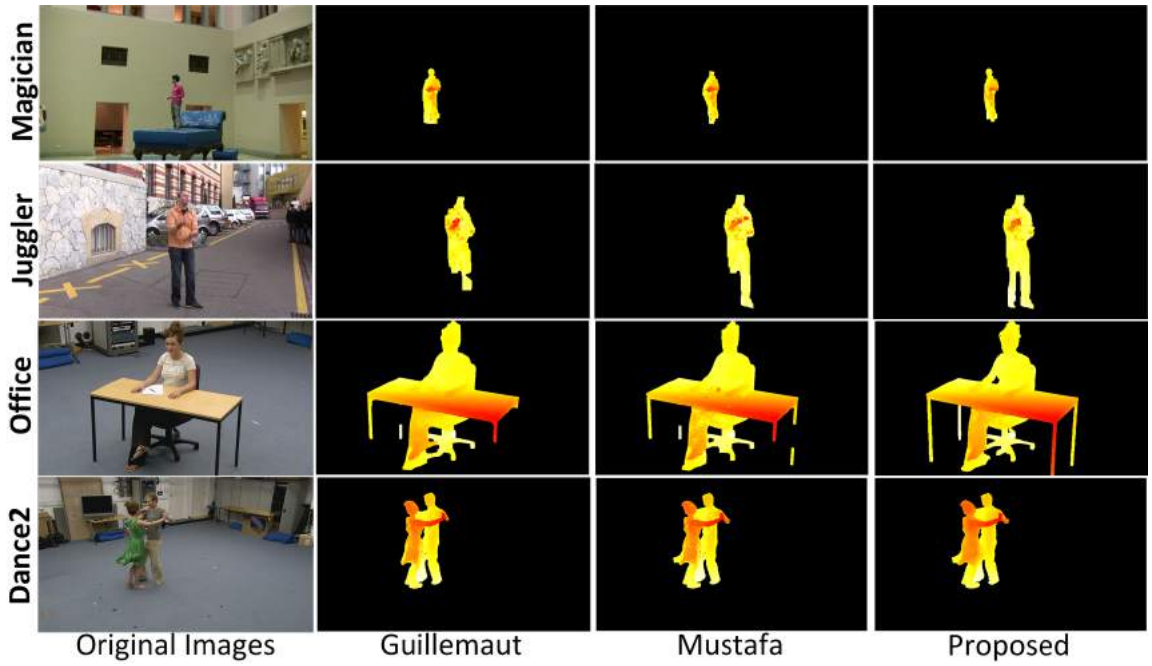


Fig. 5.24 Comparison of depth maps against existing methods for two indoor and two outdoor benchmark datasets.

introduction of the geodesic star constraint in joint optimization.

Temporal coherence: A frame-to-frame alignment is obtained using the proposed approach as shown in Figure 5.25 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. The limitations of sequential alignment will be addressed in Chapter 6.

5.5 Limitations

As with previous dynamic scene reconstruction methods the proposed approach has a number of limitations. The proposed approach may segment and reconstruct multiple objects which are in close proximity as a single dynamic object. This is not a failure case, but it increases the overall computational time for scene reconstruction. Secondly, 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;

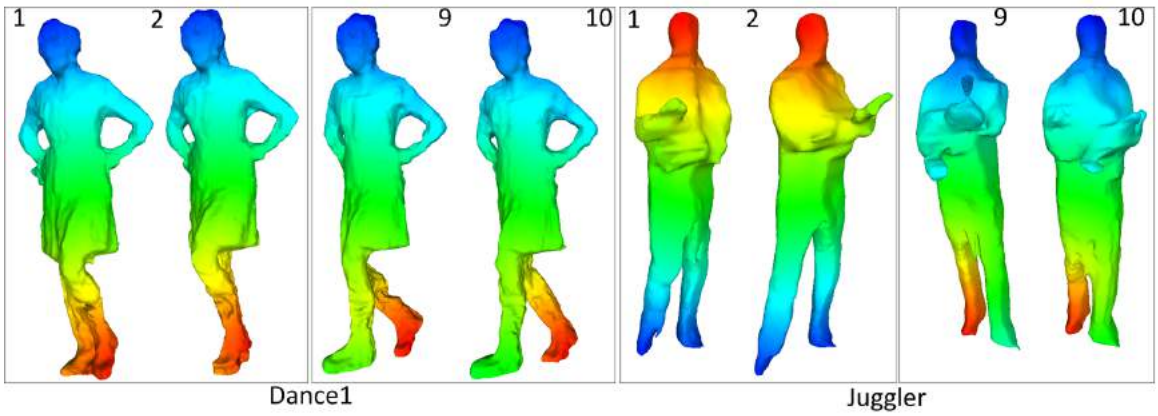


Fig. 5.25 Frame-to-frame temporal alignment for Dance1 and Juggler dataset

the approach requires sufficient wide-baseline views to cover the scene. Thirdly, the proposed technique does not handle textureless scenes due to the sparsity of 3D points and crowded scenes due to the failure of the clustering algorithm used for initialization. Most of the state-of-the-art methods in dynamic scene reconstruction suffer from this problem. Currently, the colour models (GMMs) are not updated due to short length of the sequences and datasets used, but this might suffer from failure due rapid rotations of dynamic objects. This problem can be resolved by updating the color models in case of long sequences and complex motions.

5.6 Conclusion

This chapter presented a framework for temporally coherent 4D model reconstruction of dynamic scenes from a set of wide-baseline moving cameras. The approach gives a complete model of all static and dynamic non-rigid objects in the scene. Temporal coherence for dynamic objects addresses limitations of previous per frame reconstruction giving improved reconstruction and segmentation together with dense temporal surface correspondence for dynamic objects. A sparse-to-dense approach is introduced to establish temporal correspondence for non-rigid objects using robust sparse feature matching to initialize dense optical flow. This provides an initial segmentation and reconstruction. Joint refinement of object reconstruction and segmentation is then performed using a multi-view optimization with a novel geodesic star convexity constraint that gives improved shape estimation and is computationally efficient. Comparison against state-of-the-art techniques for multi-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.

The approach proposed in this chapter introduces frame to frame temporal coherence and as an output we obtain a sequence of 3D shapes and sequential temporal coherence for the moving objects. The method suffers from drift due to accumulation of errors in alignment between successive frames and failure is observed due to large non-rigid motion over the entire sequence. This problem will be handled in the subsequent Chapter 6.

Chapter 6

4D Match Trees for Non-rigid Surface Alignment

6.1 Introduction

Temporally coherent reconstruction introduced in Chapter 5 introduces a sequential approach to obtain temporally coherent to obtain a 4D video. Sequential alignment suffers from drift due to accumulation of errors in alignment between successive frames and failure may occur due to fast motion resulting in large non-rigid deformation between successive frames.

Recent advances in computer vision have demonstrated reconstruction of complex dynamic real-world scenes from multiple view video or single view depth acquisition. These approaches typically produce an independent 3D scene model at each time instant with partial and erroneous surface reconstruction for moving objects due to occlusion and inherent visual ambiguity [77, 122, 170, 199]. For non-rigid objects, such as people with loose clothing or animals, producing a temporally coherent 4D representation from partial surface reconstructions remains a challenging problem. In this chapter an approach to address the problems of sequential alignment is presented by introducing a non-sequential global alignment across the sequence to estimate the dense surface correspondence across all observations from multiple view acquisition. A framework is proposed for global alignment of non-rigid shape observed in one or more views with a moving camera assuming that a partial surface reconstruction is available at each frame. The objective is to estimate the dense surface correspondence across all observations from multiple view acquisition. An overview of the approach is presented in Figure 6.1. Similarity evaluation between arbitrary pairs of frames is performed using robust sparse feature matches. This allows a *4D Match Tree* to be constructed which represents the optimal alignment path for all observations across multiple sequences and

views that minimizes the total dissimilarity between frames or non-rigid shape deformation. 4D alignment is then performed by traversing the 4D match tree using dense optical flow initialized from the sparse inter-frame non-rigid shape correspondence. This approach allows global alignment of partial surface reconstructions for complex dynamic scenes with multiple interacting people and loose clothing.

Previous work on 4D modelling of complex dynamic objects has primarily focused on acquisition under controlled conditions such as a multiple camera studio environment to reliably reconstruct the complete object surface at each frame using shape-from-silhouette and multi-view stereo[49, 80, 164]. Robust techniques have been introduced for temporal alignment of the reconstructed non-rigid shape to obtain a 4D model based on tracking the complete surface shape or volume [27, 28, 73] with impressive results for complex motion. However, these approaches assume reconstruction of the full non-rigid object surface at each time frame and do not easily extend to 4D alignment of partial surface reconstructions or depth maps.

The wide-spread availability of low-cost depth sensors has motivated the development of methods for temporal correspondence or alignment and 4D modelling from partial dynamic surface observations [108, 129, 172, 188]. Recently multi-view performance capture generating temporally coherent reconstructions in real-time is proposed for multiple RGBD cameras for indoor scenes [40]. Scene flow techniques [16, 187] typically estimate the pairwise surface or volume correspondence between reconstructions at successive frames but do not extend to 4D alignment or correspondence across complete sequences due to drift and failure for rapid and complex motion. Existing feature matching techniques either work in 2D[168] or 3D[113], or for sparse [78, 203] or dense[195] points. However these methods fail in the case of occlusion, large motions, background clutter, deformation, moving cameras and appearance of new parts of objects. Recent work has introduced approaches, such as DynamicFusion [129], for 4D modelling from depth image sequences integrating temporal observations of non-rigid shape to resolve fine detail. Approaches to 4D modelling from partial surface observations are currently limited to relatively simple isolated objects such as the human face or upper-body and do not handle large non-rigid deformations such as loose clothing.

In this chapter the *4D Match Tree* is introduced for robust global alignment of partial reconstructions of complex dynamic scenes. This enables the estimation of temporal surface correspondence for non-rigid shape across all frames and views from moving cameras to obtain a temporally coherent 4D representation of the scene. Contributions of this work include:

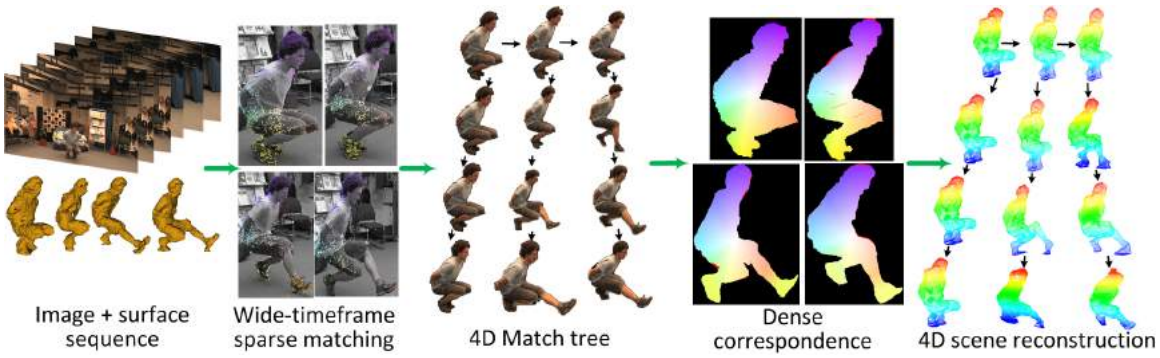


Fig. 6.1 4D Match Tree framework for global alignment of partial surface reconstructions

- Robust global 4D alignment of partial reconstructions of non-rigid shape from single or multi-view sequences with moving cameras
- Sparse matching between wide-timeframe image pairs of non-rigid shape using a segmentation based feature descriptor
- 4D Match Trees to represent the optimal non-sequential alignment path which minimizes change in the observed shape
- Dense 4D surface correspondence using optical flow guided by sparse matching

6.2 Related Work

Temporal alignment of dynamic scene reconstruction is an area of extensive research in computer vision. Temporally coherent mesh sequences finds application in 3D performance capture and animation. Research on temporal coherent reconstruction can be broadly classified into two main categories: Sequential and Non-sequential. This section will review different methods in these categories.

Sequential Alignment: Common approaches to surface reconstruction using stereo and shape-from-silhouette for surface reconstruction [99, 123] do not produce temporally coherent models for an entire sequence. These approaches align pairs of frames using existing techniques like sparse or dense optical flow and scene flow. Alignment between individual frames of sequence sequence can be established using correspondence information between frames. Methods have been proposed to obtain sparse [78, 168, 203] and dense [16, 113, 195] correspondence between consecutive frames for entire sequence. Existing sparse correspondence methods works sequentially on a frame-to-frame basis for single view [168] or multi-view [78] and require a strong prior initialization [203]. Existing dense

matching or scene flow methods [16, 187] require a strong prior which fails in the case of large motion and moving cameras. Other methods are limited to RGBD data [195] or narrow-timeframe [14, 113] for dynamic scenes.

Partial surface tracking methods for single view [151], from 3D scanner data [172] and RGBD data [65, 129, 188] perform sequential alignment of the reconstructions using frame-to-frame tracking. As an output a sequence of 3D shapes and instantaneous scene flows is obtained for the moving surface. But there is no explicit temporal consistency in this data because the shape and motion are typically sampled at each frame over a regular grid in the 3D space or image domain.

Other sequential methods proposed for 4D alignment of surface reconstructions assume that a complete mesh of the dynamic object is available for the entire sequence [28, 178, 185]. The model is deformed over time by surface tracking according to multi-view image sequences which results in temporal consistency into the resulting sequence of model instances. Sequential methods suffer from drift due to accumulation of errors in alignment between successive frames and failure is observed due to large non-rigid motion.

Non-Sequential Alignment: Recently, surface alignment methods have tackled the drift problem by non-sequential traversal of the input sequence. The template mesh is tracked from the root of the tree along tree structure of paths leading to the leaf nodes. Each node in the tree represents a frame in a sequence. Shorter chains of frame-to-frame alignments compared to sequential traversal reduce the amount of drift and the impact of a complete failure.

Non-sequential alignment has been proposed in [56] using structure-from-motion to reduce drift in reconstruction [56]. The reconstruction is performed over sub-sequences of the video and fused into a single scene reconstruction using a hierarchical tree structure. Beeler et al. [20] presented an approach based on anchor frames to reduce drift for alignment of reconstructed non-rigid face sequences. A patch based non-sequential surface alignment algorithm is proposed for face sequences in [84]. Shape similarity tree is used for non-sequential alignment across databases of multiple unstructured mesh sequences from non-rigid surface capture [27]. Recently non-sequential approaches have been proposed to temporally align unstructured meshes using motion graphs [141].

Although these methods provide a good alignment of the surfaces compared to sequential approaches, they require complete surface reconstruction of every frame of the sequence. Partial surfaces and incomplete geometries cannot be handled by these state-of-the-art approaches. The problem of structured representation of dynamic surfaces from partial surfaces from RGBD data was addressed by [108], but was limited to simple scenes. In this chapter a non-sequential method to align partial surface reconstruction of dynamic

objects is proposed for general dynamic outdoor and indoor scenes with large non-rigid motions across sequences and views. Robust sparse wide-timeframe correspondence are established to construct 4D Match Trees. Dense matching is performed on the 4D Match Tree non-sequentially using the sparse matches as an initialization for optical flow to handle large non-rigid motion and deformation across the sequence.

6.2.1 Summary of Previous Work

Existing methods have assumed complete surfaces [27, 28, 73] and/or sequential frame-by-frame alignment [108, 129, 172, 188] resulting in drift and fast motion failure. Motivated by robust non-sequential alignment of 'shape similarity trees'[27] a non-sequential alignment of partial surface reconstructions is proposed from one or more cameras. This requires a new measure of similarity between partial surface reconstructions for any pair of frames which is much more challenging than previous global shape descriptors[27]. To solve this we demonstrate for the first time that sparse temporal feature matching between widely spaced frames with large non-rigid deformations can be achieved using segmentation based feature detection SFD previously introduced for wide-baseline spatial matching [126]. This enables computation of surface overlap and shape similarity between any pair of frames used in '4D match trees' to initialize temporal alignment of partial surface reconstructions.

6.3 Methodology

The aim of this work is to obtain 4D temporally coherent models from partial surface reconstructions of dynamic scenes. Our approach is motivated by previous non-sequential approaches to surface alignment [27, 31, 33] which have been shown to achieve robust 4D alignment of complete surface reconstructions over multiple sequences with large non-rigid deformations. These approaches use an intermediate tree structure to represent the unaligned data based on a measure of shape similarity. This defines an optimal alignment path which minimizes the total shape deformation. In this chapter we introduce the 4D Match Tree to represent the similarity between unaligned partial surface reconstructions. In contrast to previous work the similarity between any pair of frames is estimated from wide-timeframe sparse feature matching between the images of the non-rigid shape and there are no prior assumptions on the reconstructed surface accuracy or mesh structure. Sparse correspondence gives a similarity measure which approximates the overlap and amount of non-rigid deformation between images of the partial surface reconstructions at different

time instants. This enables robust non-sequential alignment and initialization of dense 4D correspondence across all frames.

6.3.1 Overview

An overview of the 4D Match Tree framework is presented in Figure 6.1. The input is a partial surface reconstruction of a general dynamic scenes at each frame together with multiple view images. Cameras may be static or moving and camera calibration is assumed to be known or estimated together with the scene reconstruction [76, 123, 134, 170]. The first step is to estimate sparse wide-timeframe SFD feature correspondence. The 4D Match Tree is constructed as the minimum spanning tree based on the surface overlap and non-rigid shape similarity between pairs of frames estimated from the sparse feature correspondence. This tree defines an optimal path for alignment across all frames which minimizes the total dissimilarity or shape deformation. Traversal of the 4D Match Tree from the root to leaf nodes is performed to estimate dense 4D surface correspondence and obtain a temporally coherent representation. Dense surface correspondence is estimated by performing optical flow between each image pair initialized by the sparse feature correspondence. The 2D optical flow correspondence is back-projected to the 3D partial surface reconstruction to obtain a 4D temporally coherent representation. The approach is evaluated on publicly available benchmark datasets for partial reconstructions of indoor and outdoor dynamic scenes from static and moving cameras.

6.3.2 Robust Wide-timeframe Sparse Feature Correspondence

Sparse feature matching is performed between any pair of frames to obtain an initial estimate of the surface correspondence. This is used to estimate the similarity between partial observations of the non-rigid shape at different frames for construction of the 4D Match Tree and subsequently to initialize dense correspondence between adjacent pairs of frames on the tree branches. For partial reconstruction of non-rigid shape in general scenes we require feature matching which is robust to both large shape deformation, change in viewpoint, occlusion and errors in the reconstruction due to visual ambiguity. To overcome these challenges sparse feature matching is performed in the 2D domain between image pairs and projected onto the reconstructed 3D surface to obtain 3D matches. In the case of multiple view images consistency is enforced across views at each time frame. The input is the sequence of frames $\{q(i)\}_{i=1}^{N_Q}$ where N_Q is the number of frames. Each frame $q(i)$ consists of a set of images from multiple viewpoints $\{v(c)\}_{c=1}^{N_V}$, where N_V is the number of viewpoints for each time instant ($N_V \geq 1$).

Segmentation based Feature Detection for Wide-timeframe Matching: SFD features are detected on the segmented dynamic object for each view c and the set of initial keypoints are defined as: $X^c = \{x^c_{q_0}, x^c_{q_1}, ..., x^c_{q_N}\}$. The SIFT descriptor[105] for each detected SFD keypoint is used for wide-timeframe feature matching.

Wide-timeframe Matching: Once the features are extracted along with their descriptors from two or more images, the next step is to establish some preliminary feature matches between these images. There are potentially large non-rigid deformation and changes in viewpoint/visibility between the initial frame and the current frame, hence robust matching technique are used to establish correspondences. A match $S^{c,c}_{q_i,q_j}$ is a feature correspondence $S^c_{q_i,q_j} = (x^c_{q_i}, x^c_{q_j})$, between $x^c_{q_i}$ and $x^c_{q_j}$ in view c at frames i and in view c at frame j respectively. Nearest neighbour matching is used to establish matches between keypoints $x^c_{q_i}$ from the i^{th} frame to candidate interest points $x^c_{q_j}$ in the j^{th} frame. The ratio of the first to second nearest neighbour descriptor matching score is used to eliminate ambiguous matches ($ratio < 0.85$). This is followed by a symmetry test which employs the principal of forward and backward match consistency to remove the erroneous correspondences. Two-way matching is performed and inconsistent correspondences are eliminated. To further refine the sparse matching and eliminate outliers a local spatial coherence in the matching is enforced.

We enforce multi-view consistency on the feature matches to ensures that correspondences between any two views remain consistent for successive frames and views. Each match must satisfying the constraint:

$$\left\| S^{c,c}_{q_i,q_j} - (S^{c,k}_{q_j,q_j} + S^{k,k}_{q_i,q_j} + S^{c,k}_{q_i,q_i}) \right\| < \varepsilon \tag{6.1}$$

($\varepsilon = 0.25$). where c and k are two different viewpoints and $S^{c,c}_{q_i,q_j}$ is the match at view c at frame q_i and at view c at frame q_j. This gives a final set of sparse matches of the non-rigid shape between frames which is used to calculate the similarity metric for the non-sequential alignment of frames and initialize dense correspondence.

Several feature detection and matching approaches have been previously evaluated for wide-baseline spatial matching in Chapter 3. In this work existing detectors are evaluated against SFD for wide-timeframe matching between images of non-rigid shape on benchmark indoor and outdoor datasets with large deformation in the dynamic object: Dance1[4DI]; Dance2, Cathedral, Odzemok,[cvs]; Magician and Juggler [15]. Figure 6.2 and Table 6.1 present results for SIFT[105], FAST[149] and SFD[126] feature detection. SFD, SIFT and FAST feature detection is performed on all the views at each time instant in the sequence (approximate length of 400). Wide-timeframe matching is performed using the algorithm described above between all the frames such that frame 1 is matched to frame 2 to q_{N_Q}. The

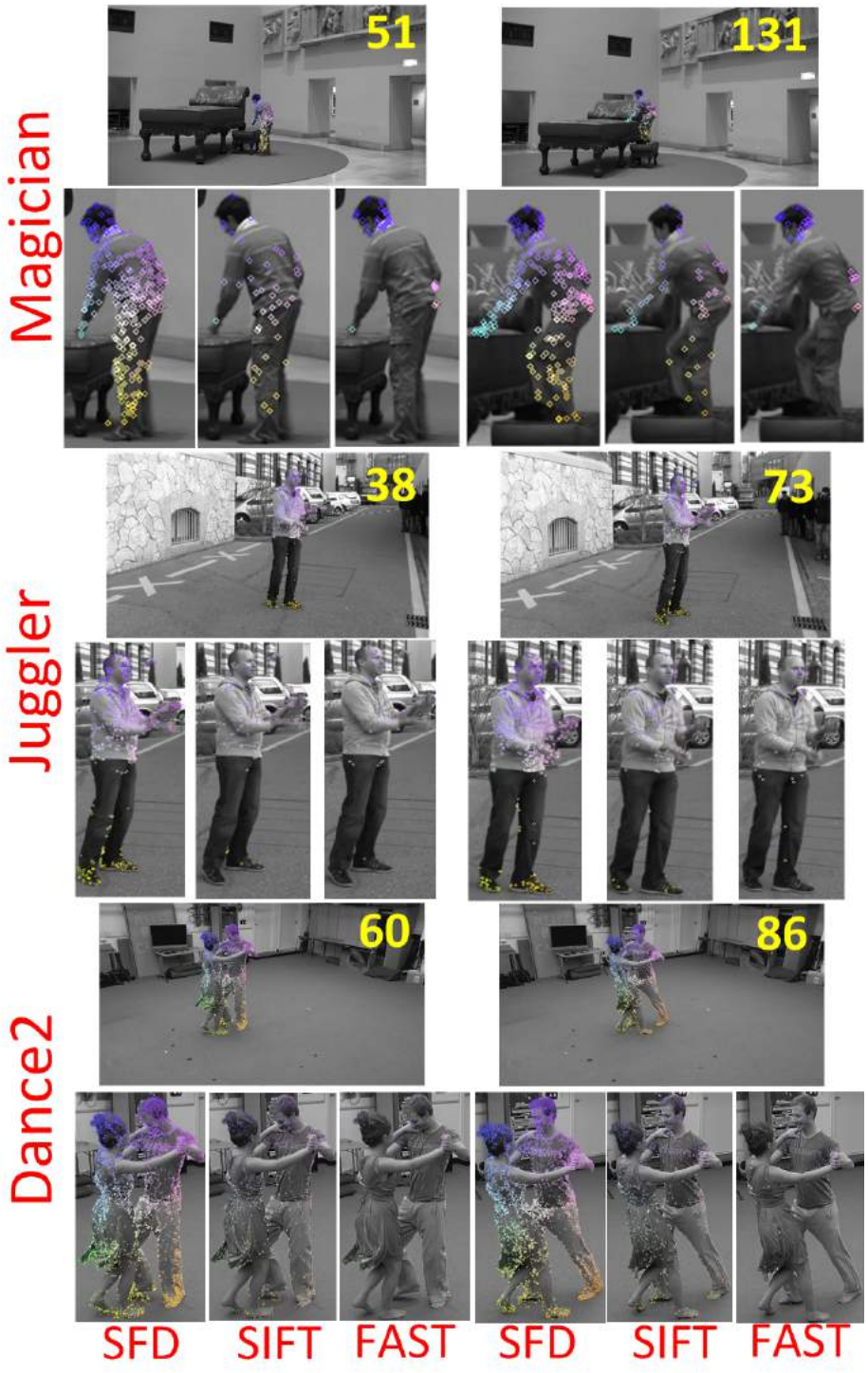


Fig. 6.2 Comparison of feature detectors for wide-timeframe matching on 3 datasets.

	Matches			
Datasets	Proposed	Nebehay	SIFT	FAST
Dance3	**416**	249	124	57
Dance2	**1233**	863	493	96
Odzemok	**916**	687	366	82
Cathedral	**665**	465	301	77
Magician	**392**	293	141	53
Juggler	**547**	437	273	68

Table 6.1 Comparison of number of sparse wide-timeframe correspondences for all datasets averaged over the entire sequence.

comparison is performed based on the following metric:

$$\text{Match Metric } = \frac{1}{N_Q^2 \times N_V} \sum_{c=1}^{N_V} \sum_{\substack{i=1 \\ \text{such that } i \neq j}}^{N_Q} \sum_{j=1}^{N_Q} \left| s^c_{q_i,q_j} \right| \tag{6.2}$$

where $\left| s^c_{q_i,q_j} \right|$ is the number of matches between view c of frame i and j. Table 6.1 shows that a greater number of temporal feature matches are achieved using SFD than SIFT or FAST detectors (rather than greater correct spatial matches [126]). This is significant because it allows sparse matching between non-rigid surfaces where other detectors fail to produce enough matches for initialization. SFD matches allow initialization of the dense correspondence, any remaining false sparse matches are eliminated in dense matching. This evaluation shows that SFD gives a relatively high number of correct matches. Results indicate that SFD can successfully establish sparse correspondence for large non-rigid deformations as well as changes in viewpoint with improved coverage and number of features.

6.3.3 4D Match Trees for Non-sequential Alignment

Our aim is to estimate dense correspondence for partial non-rigid surface reconstructions across complete sequences to obtain a temporally coherent 4D representation. Previous research has employed a tree structure to represent non-rigid shape of complete surfaces for robust non-sequential alignment of sequences with large non-rigid deformations [27, 31, 33]. These approaches use a similarity metric taking into account both shape and motion of the moving objects. However, previous similarity metrics are only applicable to objects with full visibility in the scene and require complete geometry as a prior. In this work we have partial reconstructions of complex dynamic scenes with partially visible objects with incomplete

geometries, hence existing shape similarity measures and matrices cannot be used in our work.

Inspired by the success of these approaches we propose the *4D Match Tree* as an intermediate representation for alignment of partial non-rigid surface reconstructions. An important difference of this approach is the use of an image based metric to estimate the similarity in non-rigid shape between frames. 4D match trees handle partial surfaces and use a sparse appearance similarity metric rather than global shape to build a tree, neither the representation or alignment of [27] can apply to partial surfaces. Similarity between any pair of frames is estimated from the sparse wide-timeframe feature matching. The 4D Match Tree represents the optimal traversal path for global alignment of all frames as a minimum spanning tree according to the similarity metric.

The space of all possible pairwise transitions between frames of the sequence is represented by a dissimilarity matrix $\mathbf{D}$ of size $N \times N$ where both rows and columns correspond to individual frames. The elements $\mathbf{D}(i,j) = d(q_i, q_j)$ are proportional to the cost of dissimilarity between frames i and j. The matrix is symmetrical ($d(q_i,q_j) = d(q_j,q_i)$) and has zero diagonal ($d(q_i,q_i) = 0$). For each dynamic object in a scene a graph Ω of possible frame-to-frame matching is constructed with nodes for all frames q_i. $d(q_i,q_j)$ is the similarity metric between two nodes and is computed using information from sparse correspondences and intersection of silhouettes obtained from the back-projection of the surface reconstructions in each view.

Feature match metric: SFD features detected for each view at each frame are matched between frames using all views. The feature match metric for non-sequential alignment $K^c_{i,j}$ between frame i and j for each view c is defined as the inlier ratio:

$$K^c_{i,j} = \frac{|s^c_{q_i,q_j}|}{R^c_{i,j}} \tag{6.3}$$

where $R^c_{i,j}$ is the total number of preliminary feature matches between frames i and j for view c before constraining, and $|s^c_{q_i,q_j}|$ is the number of matches between view c of frame i and frame j obtained using the method explained in Section 6.3.2. The preliminary feature matches and the refined correspondences after various tests. $K^c_{i,j}$ is a measure of the overlap between the visible surface for view c at frames i and j. The visible surface overlap is a measure of their suitability for pairwise dense alignment.

Silhouette match metric: The partial surface reconstruction at each frame is back-projected in all views to obtain silhouettes of the dynamic object. Silhouettes between two frames for the same camera view c are aligned by an affine warp [41] as shown in Figure 6.3. The aligned silhouette intersection area $h^c_{i,j}$ between frames i and j for view c is evaluated. The

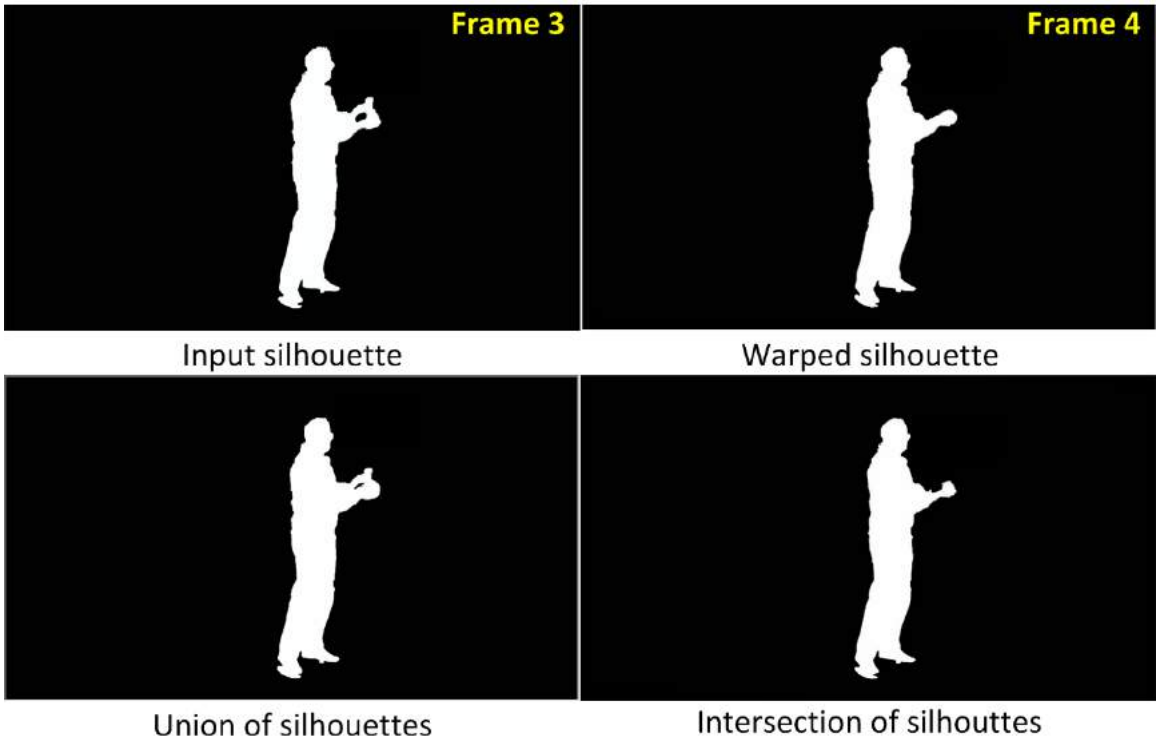


Fig. 6.3 Illustration of silhouette match metric: The input silhouette is the back-projection of mesh at Frame 3 and the warped silhouette shows the affine warp of back-projected mesh at Frame 4 w.r.t Frame 3. Union represents the addition of the two silhouettes from the top row ($A^c_{3,4}$) and Intersection represents the common area of the two silhouettes ($h^c_{3,4}$).

silhouette match metric $I^c_{i,j}$ is defined as:

$$I^c_{i,j} = \frac{h^c_{i,j}}{A^c_{i,j}} \tag{6.4}$$

where $A^c_{i,j}$ is the union of the area under the silhouette at frame i and j for view c. This gives a measure of the shape similarity between observations of the non-rigid shape between pairs of frames.

Similarity metric: The two metrics $I^c_{i,j}$ and $K^c_{i,j}$ are combined to calculate the dissimilarity between frames used as graph edge-weights. The edge-weight $d(q_i, q_j)$ for Ω is defined as:

$$d(q_i,q_j) = \begin{cases} \infty & \text{, if } |s^c_{q_i,q_j}| < 0.006 * max(\mathrm{W},\mathrm{H}) \\ \frac{1}{\sum_{c=1}^{N_V} K^c_{i,j} \times I^c_{i,j}} & \text{, otherwise} \end{cases} \tag{6.5}$$

where W and H are the width and height of the input image. Note small values of $d()$ indicates a high similarity in feature matches between frames. If the number of matches between any

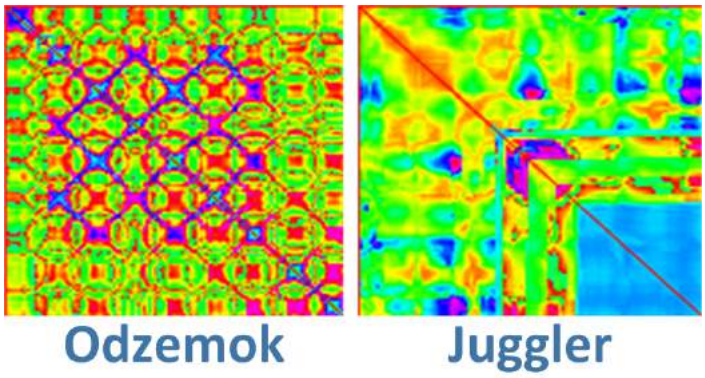


Fig. 6.4 The similarity matrix for Odzemok and Juggler datasets

two frames is lower than $0.006 * max(\mathrm{W},\mathrm{H})$ the edge weight is set to infinity. Figure 6.4 and 6.10 present the dissimilarity matrix D between all pairs of frames for all sequences (red indicates similar frames, blue dissimilar). The matrix off diagonal red areas indicate frames with similar views of the non-rigid shape suitable for non-sequential alignment.

A minimum spanning tree(MST) is constructed over this graph to obtain the 4D Match Tree. The MST defines the optimal path which minimizes the total non-rigid alignment costs (based on similarity) to optimize across all frames. Hence to define the optimal shape similarity MST is used rather than the shortest path tree (SPT) as it minimizes the total non-rigid deformation. SPT will favour short paths with large inter-frame differences in shape where disproportionally large errors in pairwise alignment may occur due to the non-linear relationship between error and dissimilarity. In contrast the MST identifies paths which favour small inter-frame differences in shape. Thus alignment based on the MST minimizes the accumulation of errors in alignment. MST also orders similar frames closer to the root with larger inter-frame changes towards the leaves of the tree, this limits the propagation of errors in alignment between meshes with relatively large differences in shape to the ends of the branches.

4D Match Tree: A fully connected graph is constructed using the dissimilarity metric as edge-weights and the minimum spanning tree is evaluated [91, 142]. Optimal paths through the sequence to every frame can be jointly optimized based on $d()$. The paths are represented by a traversal tree $\mathscr{T} = (\mathscr{N}; \mathscr{P})$ with the nodes $\mathscr{N} = \{Q_i\}_{i=1}^{N_Q}$. The edges $\mathscr{P}$ are undirected and weighted by the dissimilarity $p_{i,j} = d(q_i, q_j)$ for $p_{i,j} \in \mathscr{P}$. The optimal tree $\mathscr{O}$ is defined as the MST which minimizes the total cost of pairwise matching given by d:

$$\mathscr{O} = \arg\min_{\forall \mathscr{T} \in \Omega} \left(\sum_{\forall i,j \in \mathscr{T}} d(q_i, q_j) \right) \tag{6.6}$$

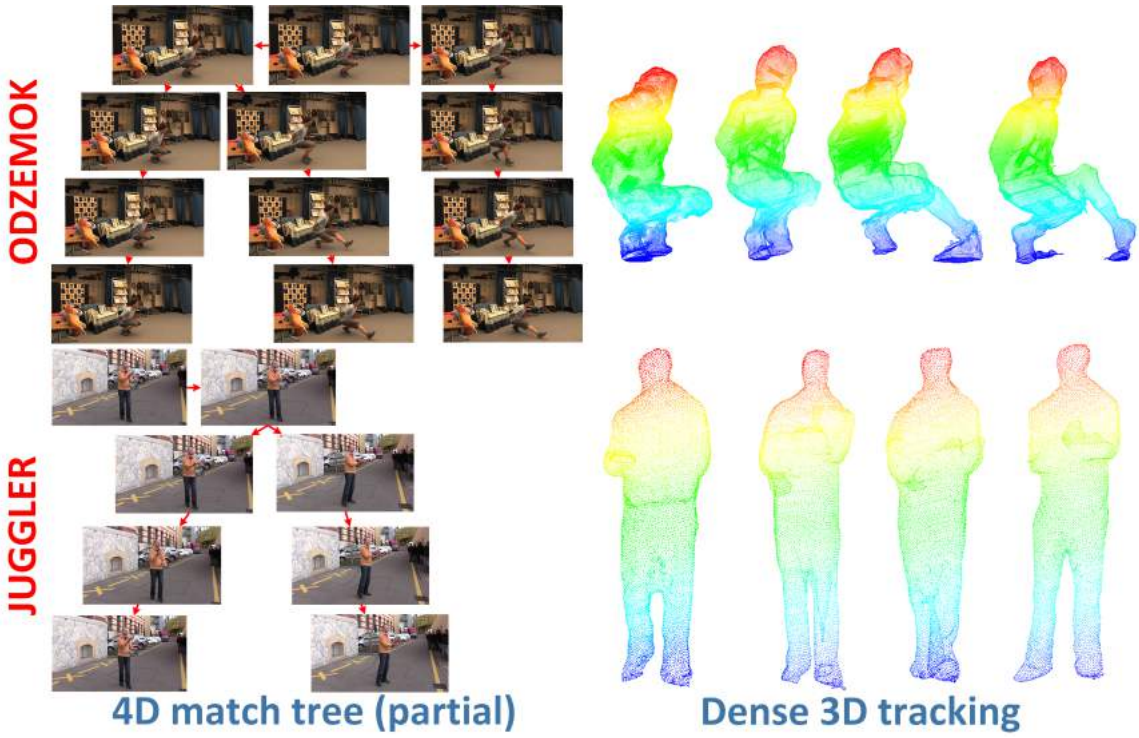


Fig. 6.5 The partial 4D Match Tree and 4D alignment for Odzemok and Juggler datasets

This results in the 4D Match Tree $\mathscr{O}$ which minimizes the total dissimilarity between frames due to non-rigid deformation and changes in surface visibility. Given $\mathscr{O}$ for a dynamic object the dense correspondence is estimated for the entire sequence to obtain a temporally coherent 4D representation of the surface shape. The tree root node M_{root} is defined as the node with minimum path length to all nodes in $\mathscr{O}$. The minimum spanning tree can be efficiently evaluated using established algorithms with order $O(n \log n)$ complexity where n is the number of nodes in the graph Ω. The mesh at the root node is subsequently tracked to other frames by traversing through the branches of the tree $\mathscr{T}$ towards the leaves. Partial 4D Match Trees for Juggler and Odzemok datasets are shown in Figure 6.5 on the left and aligned 3D dense points for different frames are shown on the right with the corresponding similarity matrix shown in Figure 6.4. To visualize the correspondence the tree root node M_{root} is color coded and the colors are propagated on various branches of the tree. The color in various parts of the object remain consistent over time indicating the success of the proposed method in estimating the surface correspondence.

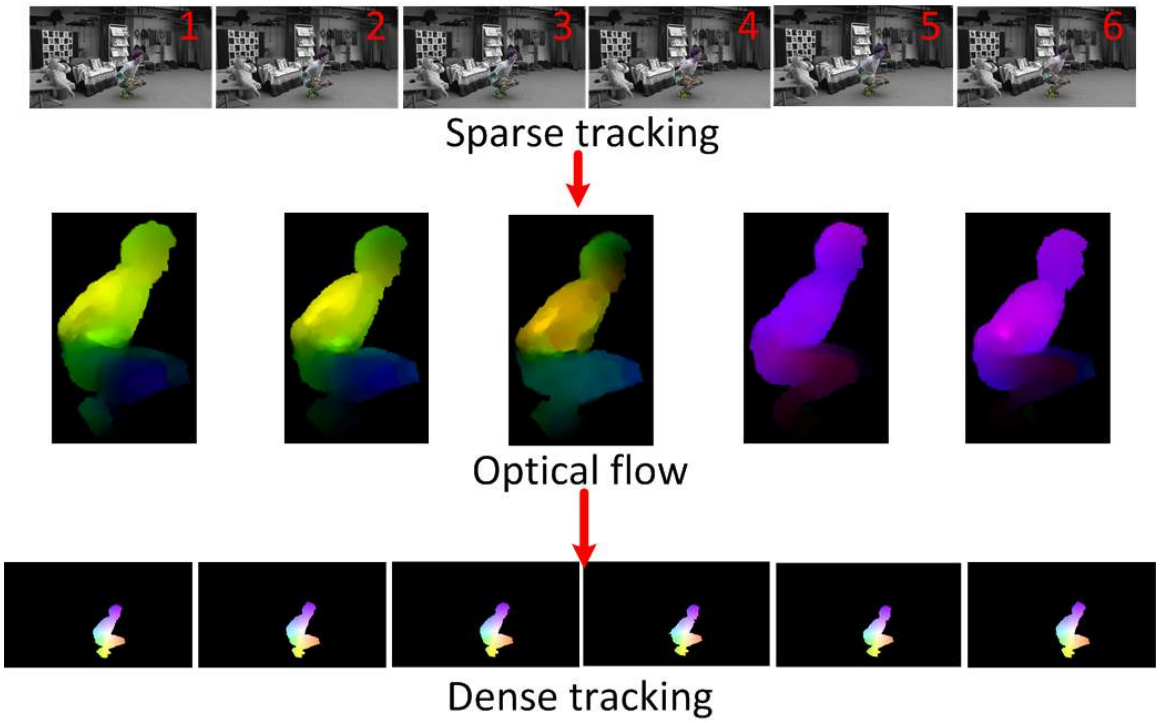


Fig. 6.6 Sparse to dense tracking using optical flow on series of frames for the Odzemok dataset

6.3.4 Dense Non-rigid Alignment

Given the 4D Match Tree global alignment is performed by traversing the tree to estimate dense correspondence between each pair of frames connected by an edge. Sparse SFD feature matches are used to initialize the pairwise dense correspondence which is estimated using optical flow [42] and the framework is shown in Figure 6.6. As illustrated in Figure 6.12 and 6.13 sparse features are non-uniformly distributed over the non-rigid surface. Hence, optical flow alignment rather than interpolation is required to achieve dense correspondence and eliminate any errors/inaccuracy in sparse matching. The sparse feature correspondences provides a robust initialization of the optical flow for large non-rigid shape deformation.

To illustrate the dense correspondence over time a coloring scheme is used as shown in Figure 6.7. A color wheel is used for the coding scheme such that the centroid of the back-projected mask is mapped to the the center of the color wheel and each pixel represents a color ID from the RGB gamut. The silhouette of the frame representing the tree root node M_{root} is color coded and the colors are propagated using the temporal dense matching information towards the leaves of the 4D match tree.

The estimated dense correspondence is back projected to the 3D visible surface to establish dense 4D correspondence between frames. In the case of multiple views dense

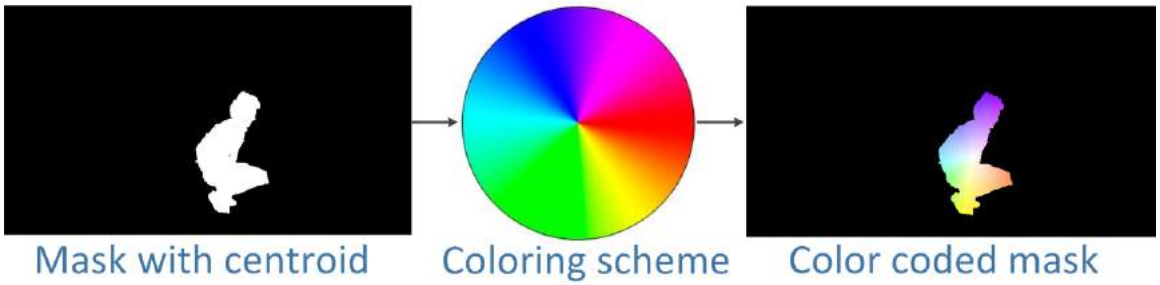


Fig. 6.7 Color coding scheme of dense correspondence for the Odzemok dataset

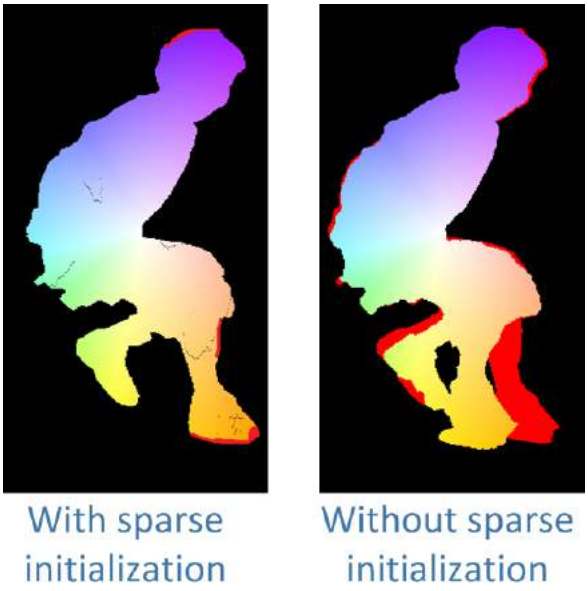


Fig. 6.8 Dense matching using optical flow with and without the sparse match initialization for the Odzemok dataset

4D correspondence is combined across views to obtain a consistent estimate and increase surface coverage. Dense temporal correspondence is propagated to new surface regions through the temporal propagation. As new surface regions appear they are incorporated into the temporal matching using the sparse feature matching and dense optical flow. An example of the propagated mask with and without sparse initialization for a single view is shown in Figure 6.8. The large motion in the leg of the actor is correctly estimated with sparse match initialization but fails without (shown by the red region indicating no correspondence).

Pairwise 4D dense surface correspondences are combined across the tree to obtain a temporally coherent 4D alignment across all frames. An example is shown for the Odzemok dataset in Figure 6.9 with sparse correspondence information in each frame. For visualization features are color coded in one frame according to the color map as illustrated in Figure 6.7

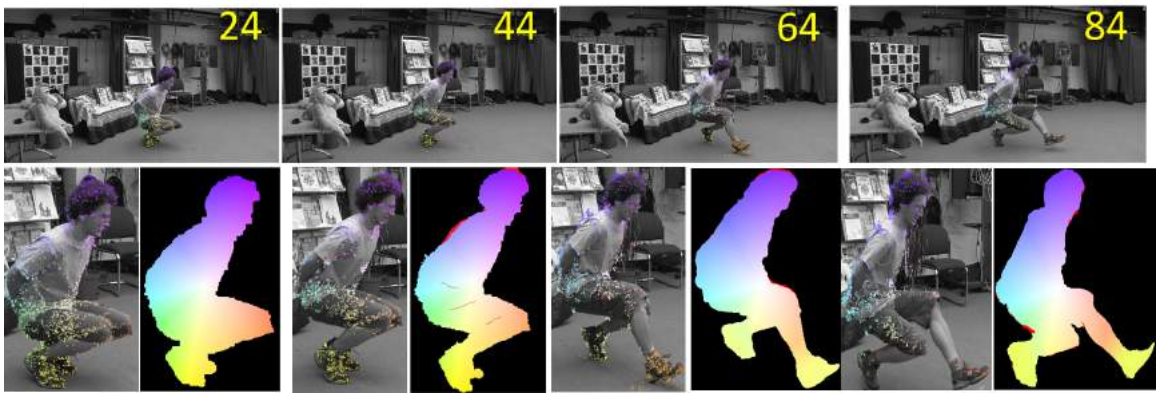


Fig. 6.9 Sparse feature matching and dense correspondence for the Odzemok dataset

Datasets	Number of views	Sequence length(frames)	Resolution	Tree depth (frames)	Tree depth (%)
Dance3	8 static	200	780×582	65	33
Dance2	8 static	244	1920×1080	73	29
Odzemok	6 static, 2 moving	232	1920×1080	82	35
Cathedral	8 static	217	1920×1080	92	42
Magician	6 moving	400	960×544	127	32
Juggler	6 moving	400	960×544	104	26

Table 6.2 Characteristic properties of all datasets and their 4D Match Trees.

and this color is propagated to feature matches in the other frames. As observed the color ID in various parts of the object remain consistent over time. Figure 6.5 presents two examples of 4D aligned meshes resulting from the global alignment with the 4D match tree for various frames in the sequence.

6.4 Results and Evaluation

The proposed approach is tested on a wide variety of indoor/outdoor scenes introduced in section 6.3.1 including fast/challenging motion and no/rigid motion and the properties of datasets are described in Table 6.2. Tree depth is defined as the maximum tree branch length in frames as a percentage of the total number of frames in the length of the sequence. 4D temporal alignment of partial reconstructions is achieved in all cases.

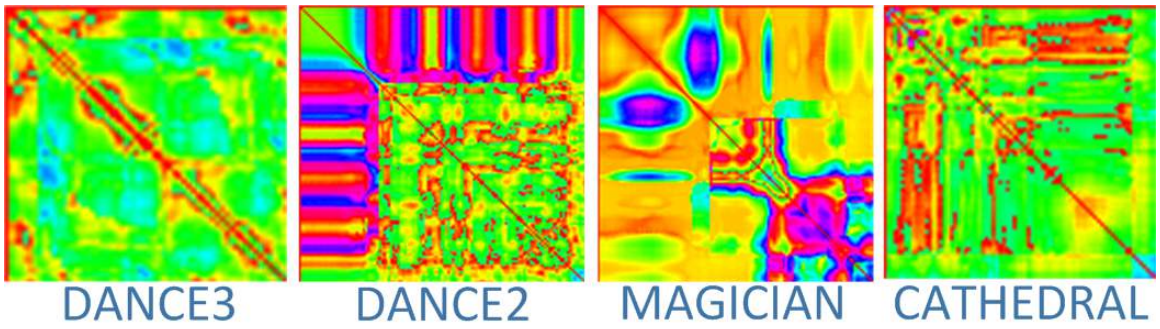


Fig. 6.10 Similarity matrix for non-sequential alignment of various datasets

6.4.1 Sequential vs. Non-sequential alignment

4D Match Trees are constructed for all datasets using the method described in Section 6.3.3 and the similarity matrix for the datasets are shown in Figure 6.10. The maximum length of branches in the 4D Match Tree for global alignment of each dataset is described in Table 6.2. The longest alignment path for all sequences is $< 50\%$ of the total sequence length leading to a significant reduction in the accumulation of errors due to drift in the sequential alignment process. Non-rigid alignment is performed over the branches of the tree to obtain a temporally consistent 4D representation for all datasets. Comparison of 4D aligned surfaces obtained from the proposed non-sequential approach against sequential tracking without the 4D Match tree is shown in Figure 6.11. Sequential tracking fails to estimate the correct 4D alignment (Odzemok-64, Dance2-66, Cathedral-55) whereas the non-sequential approach obtains consistent correspondence for all frames for sequences with large non-rigid deformations. To illustrate the surface alignment a color map is applied to the root mesh of the 4D Match tree and propagated to all frames based on the estimated dense correspondence. The color map is consistently aligned across all frames for large non-rigid motions of dynamic shapes in each dataset demonstrating qualitatively that the global alignment achieves reliable correspondence compared to sequential tracking.

6.4.2 Sparse Wide-timeframe Correspondence

Sparse correspondences are obtained for the entire sequence using the traversal path in the 4D Match tree from the root node towards the leaves. Results of the sparse and dense 4D correspondence are shown in 6.12 and 6.13. Sparse matches obtained using SFD are evaluated against a state-of-the-art method for sparse correspondence Nebehay[128] in Figure 6.14. For fair comparison Nebehay is initialized with SFD features instead of FAST (which produces a low number of matches). Table 6.1 shows that a greater number of temporal

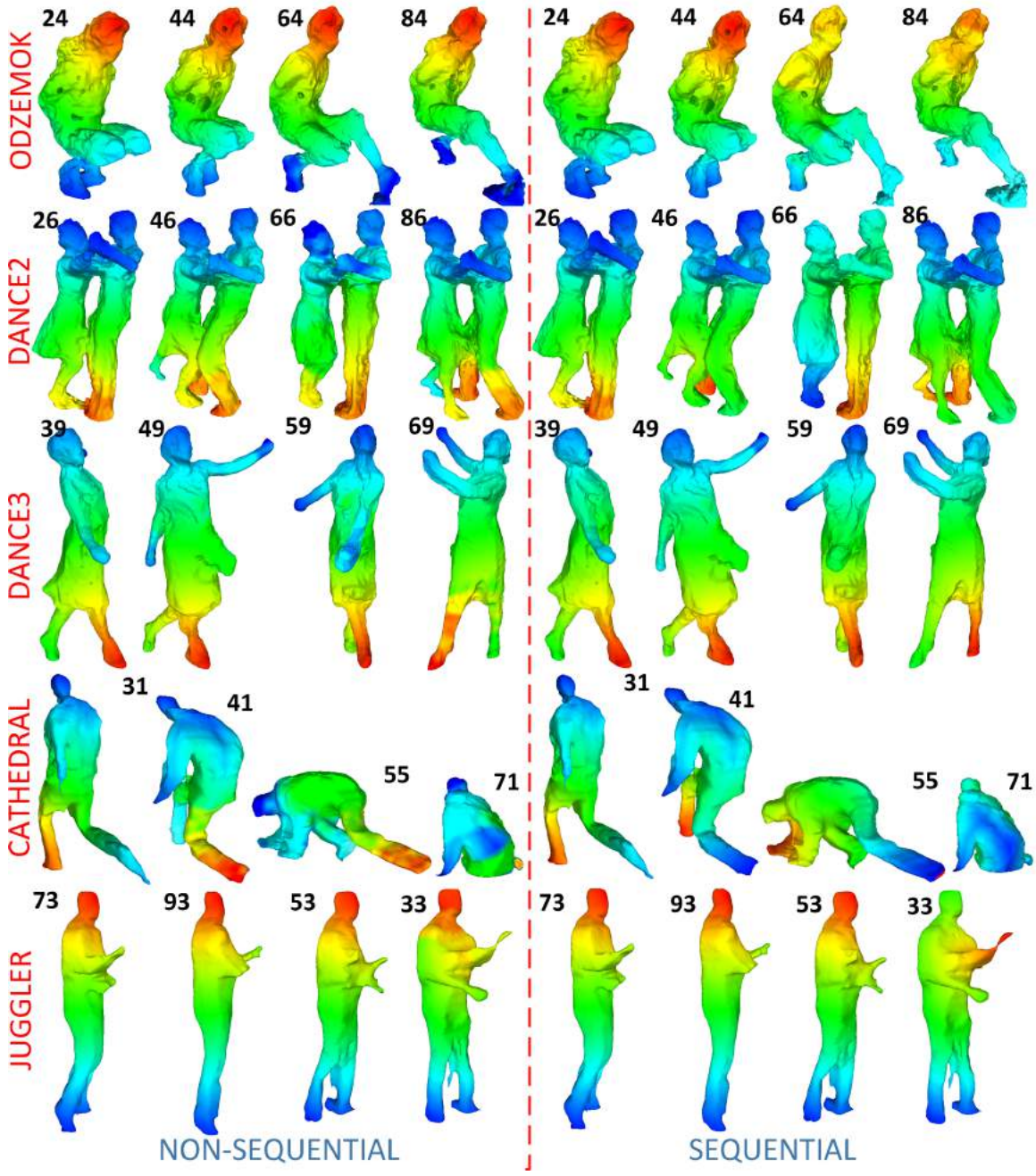


Fig. 6.11 Comparison of sequential and non-sequential alignment of all datasets for a sequence of frames.

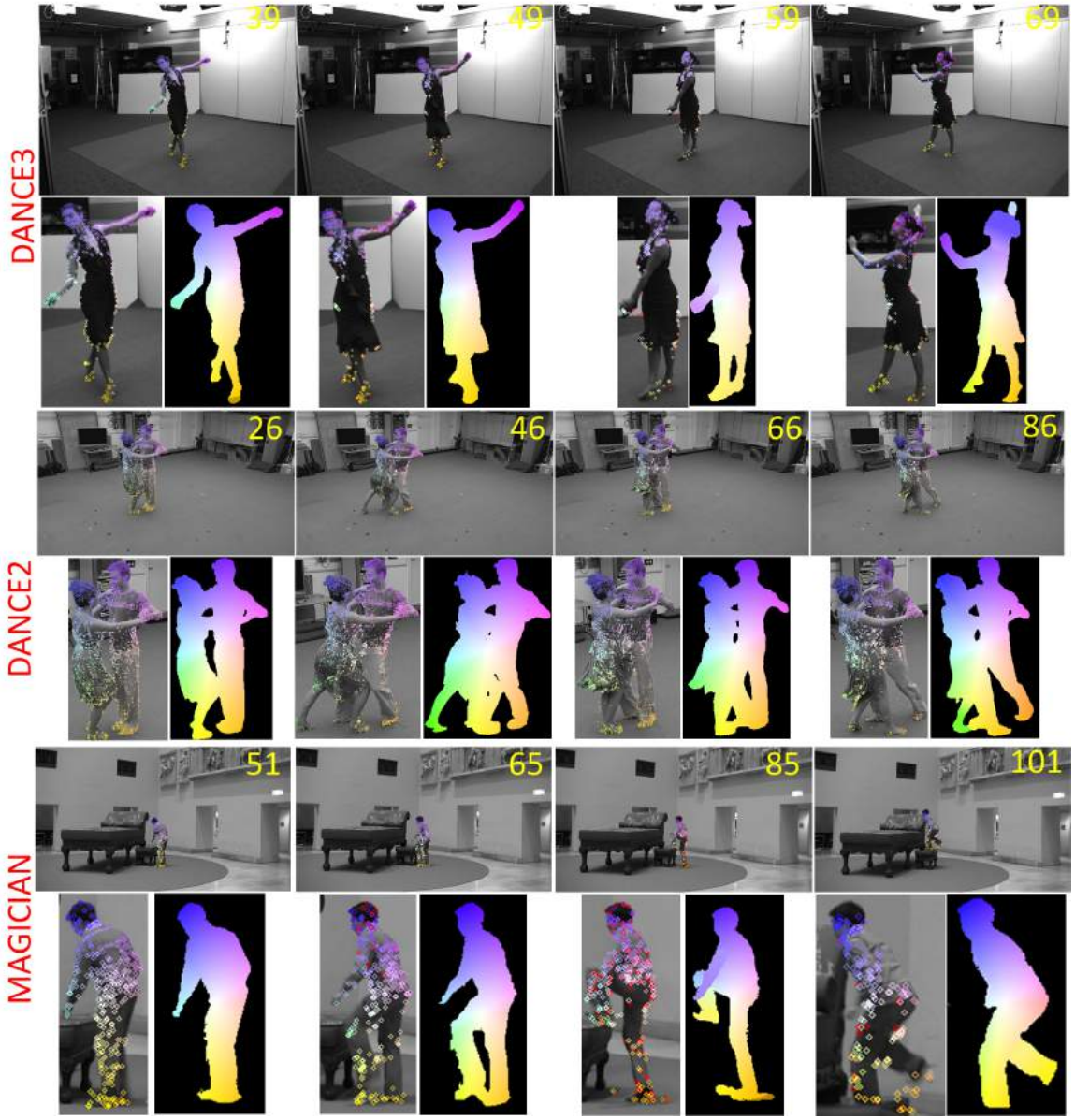


Fig. 6.12 Sparse and dense 2D tracking color coded for all datasets.

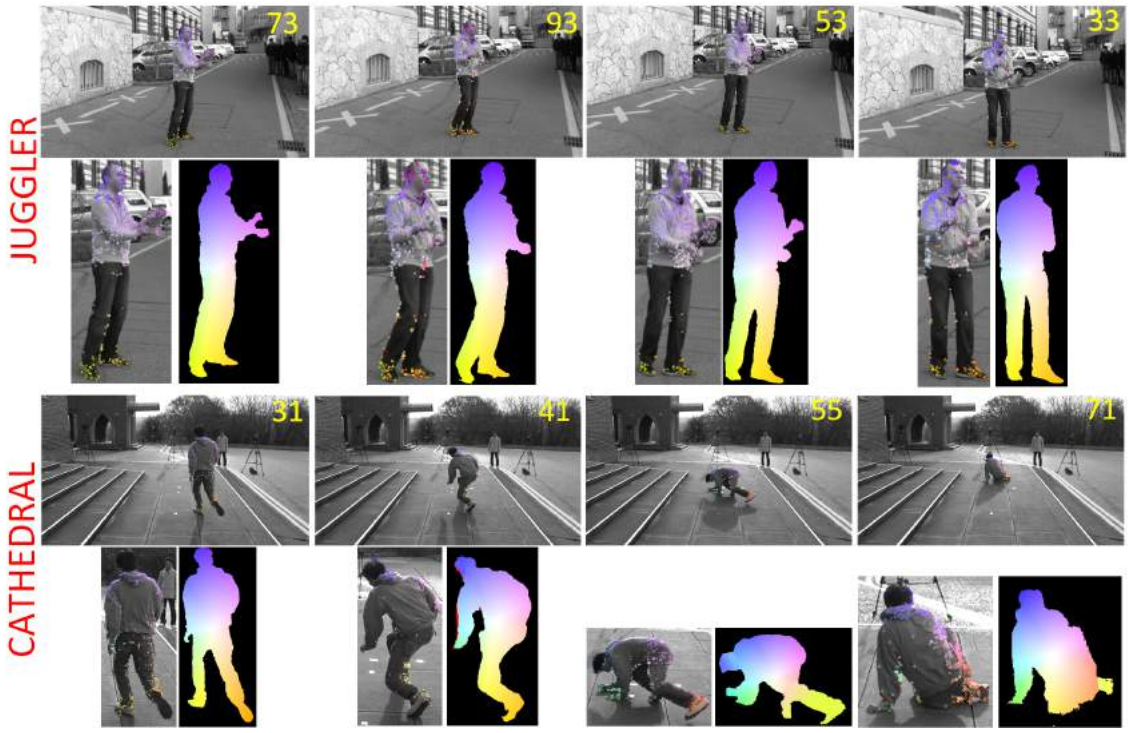


Fig. 6.13 Sparse and dense 2D tracking color coded for all datasets

feature matches are achieved using SFD than SIFT or FAST detectors. SFD matches are of sufficient accuracy to initialize dense correspondence, false sparse matches are eliminated in dense matching. Qualitative results are shown in Figure 6.14 and quantitative results are shown in Table 6.3. Matches obtained using the proposed approach are approximately 50% higher and consistent across frames compared to Nebehay[128] demonstrating the robustness of the proposed wide-timeframe matching using SFD features.

	Silhouette overlap error							
Datasets	Seq	Prop.	Deepflow	SIFT	Nebehay	1 view	2 views	4 views
Dance3	0.42	**0.35**	0.97	0.92	0.96	1.53	1.30	0.99
Dance2	0.83	**0.63**	1.36	1.43	1.38	2.13	1.78	1.47
Odzemok	0.98	**0.89**	2.82	2.59	2.69	4.35	3.66	2.76
Cathedral	0.83	**0.69**	1.14	1.10	1.29	1.92	1.65	1.09
Magician	1.07	**0.86**	3.43	3.22	3.77	5.46	4.67	3.18
Juggler	0.78	**0.65**	1.24	1.19	1.31	2.12	1.76	1.44

Table 6.3 Quantitative evaluation for sparse and dense correspondence for all the datasets; Prop. represents proposed non-sequential approach.

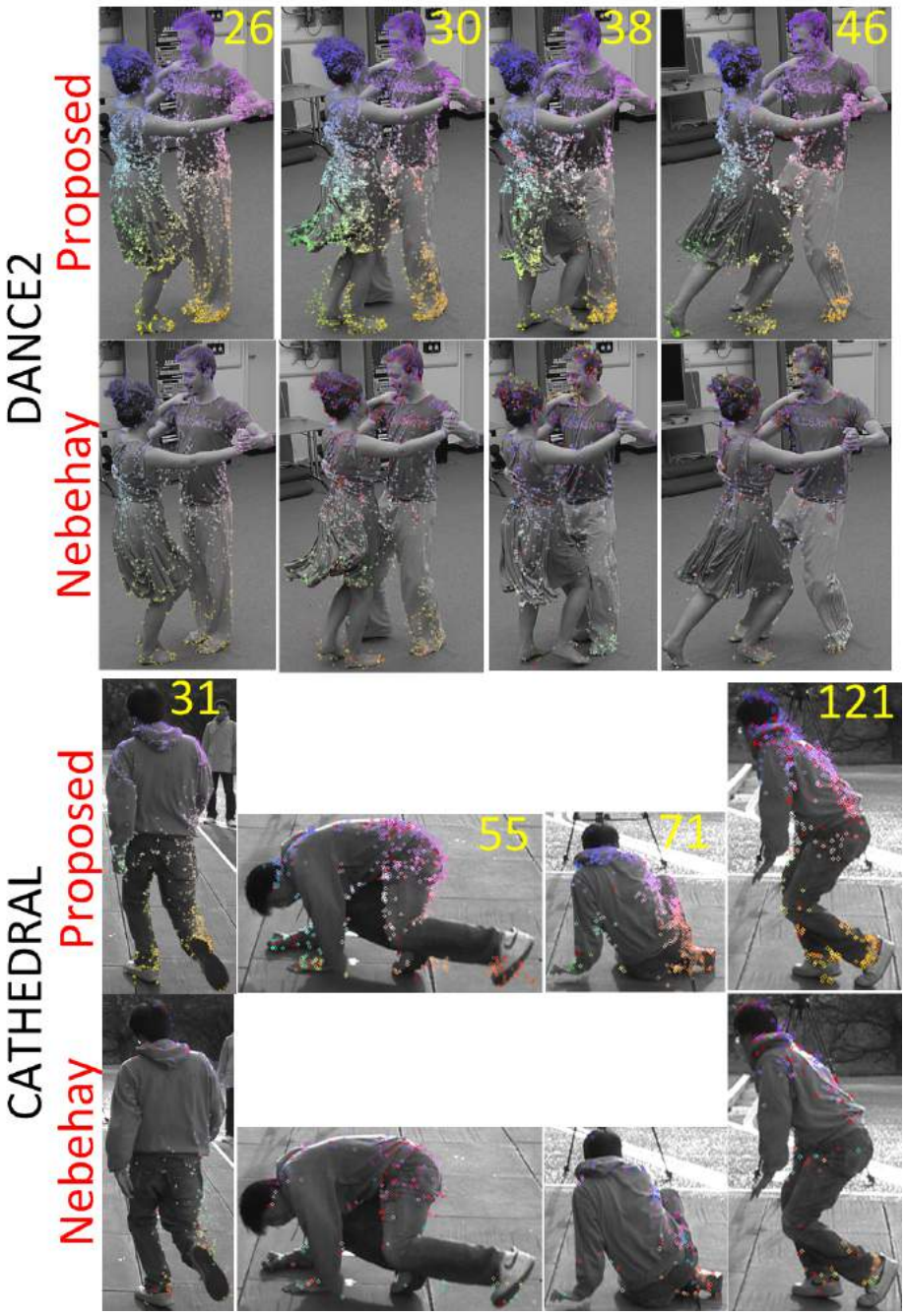


Fig. 6.14 Sparse tracking comparison for one indoor and one outdoor dataset.

6.4.3 Dense 4D Correspondence

Dense correspondences are obtained on the 4D match tree and the color coded results are shown in Figure 6.12 and 6.13 for all datasets. To illustrate the dense alignment the color coding scheme shown in Figure 6.9 is applied to the silhouette of the dense mesh on the root node for each view and propagated using the 4D Match Tree. The proposed approach is qualitatively shown to propagate the correspondences reliably over the entire sequence for complex dynamic scenes.

For comparative evaluation of dense matching we use:(a) SIFT features with the proposed method in section 6.3 to obtain dense correspondence; (b) Sparse correspondence obtained using Nebehay [128] with the proposed dense matching; and (c) a state-of-the-art dense flow algorithm Deepflow [189] over the 4D Match Tree for each dataset. Qualitative results against SIFT and Deepflow are shown in Figure 6.15 and 6.16. The propagated color map using deep flow and SIFT based alignment does not remain consistent across the sequence as compared to proposed method (red regions indicate correspondence failure).

For quantitative evaluation the silhouette overlap error(SOE) defined in equation 6.7 is compared. Dense correspondence is propagated over time to create a mask for each image. The propagated mask is overlapped with the silhouette of the projected partial surface at each frame to evaluate the accuracy of the dense propagation. The error is defined as:

$$SOE = \frac{1}{N_V * N_Q} \sum_{i=1}^{N_Q} \sum_{c=1}^{N_V} \frac{\text{Area of intersection}}{\text{Area of back-projected mask}} \tag{6.7}$$

Evaluation against sequential and non-sequential Deepflow, SIFT and Nebehay are shown in Table 6.3 for all datasets. As observed the silhouette overlap error is lowest for the proposed SFD based non-sequential approach showing relatively high accuracy. The completeness of the 3D points is also evaluated at each time instant as observed in Table 6.5 and 6.4:

$$completeness = \frac{100}{N_V * N_Q} \sum_{i=1}^{N_Q} \sum_{c=1}^{N_V} \frac{\text{Number of 3D points propagated}}{\text{Number of surface points visible from 'c' [79]}} \tag{6.8}$$

The proposed approach outperforms Deepflow, SIFT and Nebehay. As illustrated in Figure 6.12 and 6.13 sparse features are non-uniformly distributed. Optic flow alignment rather than interpolation is required to achieve dense correspondence. Results illustrate consistent matching over the entire sequences. The approach is tested on a wide variety of indoor and outdoor scenes which include fast and challenging motion including no motion, rigid or non-rigid motion. 4D temporal alignment of partial reconstructions is achieved in all cases. Ideally evaluation would be made against ground-truth however currently no ground-truth is available

Complete -ness(%)	Deepflow	SIFT	Nebehay	Proposed (all views)
Dance3	81.56	83.28	82.55	**98.22**
Dance2	83.26	85.80	83.96	**99.36**
Odzemok	81.46	79.83	80.91	**98.19**
Cathedral	79.54	81.53	81.78	**97.40**
Magician	82.58	82.92	80.65	**97.53**
Juggler	79.09	80.11	81.33	**97.89**

Table 6.4 Evaluation of completeness of dense 3D correspondence averaged over the entire sequence in %.

	Sequential	Proposed (Non-sequential)			
	All views	1 view	2 views	4 views	All views
Dance3	91.52	60.78	71.65	81.30	**98.22**
Dance2	92.76	61.98	72.30	82.87	**99.36**
Odzemok	90.51	62.73	70.87	77.64	**98.19**
Cathedral	89.21	59.77	69.05	76.98	**97.40**
Magician	89.58	61.29	71.23	75.56	**97.53**
Juggler	91.89	59.54	68.40	78.81	**97.89**

Table 6.5 Evaluation of completeness of dense 3D correspondence averaged over the entire sequence for different number of views in %.

for reconstructions of real non-rigid surfaces. As in previous work temporal alignment is evaluated against other methods using quantitative silhouette overlap and completeness measures Table 6.3, 6.4 and 6.5. This shows significant improvement in the resulting dense 4D correspondence.

6.4.4 Computational Complexity

The computational complexity of the proposed non-sequential alignment is evaluated against sequential method and other state-of-the-art methods and the results are shown in Table 6.6 for all the datasets. The experiments are performed on a quad-core processor with a speed of 1.8 GHz. The time required to obtain dense correspondence for the whole sequence is calculated and is averaged over number of frames to obtain computation time per frame. As observed the time complexity for the proposed method is comparable to other methods.

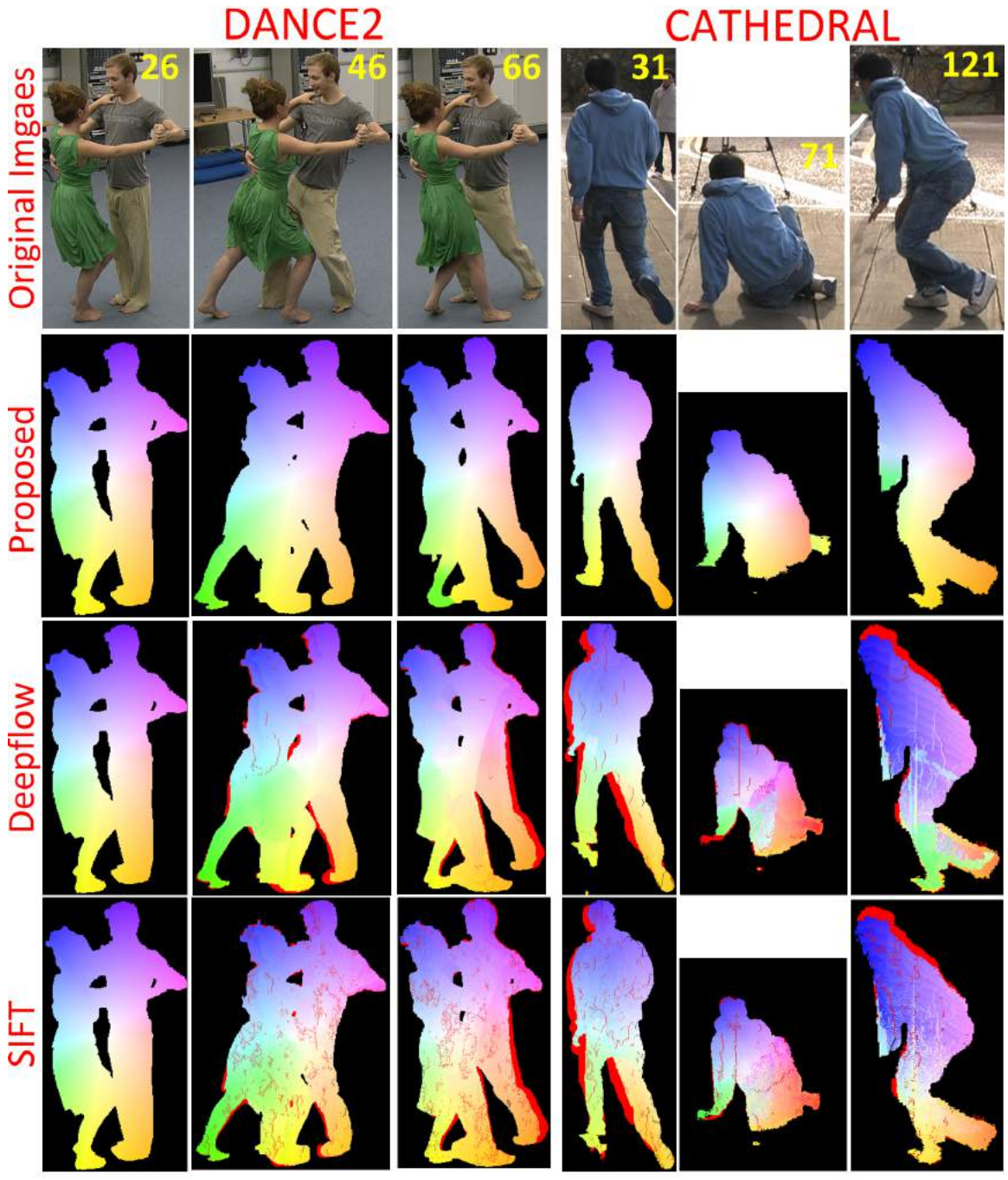


Fig. 6.15 Dense tracking comparison for one indoor and one outdoor datasets captured with static cameras.

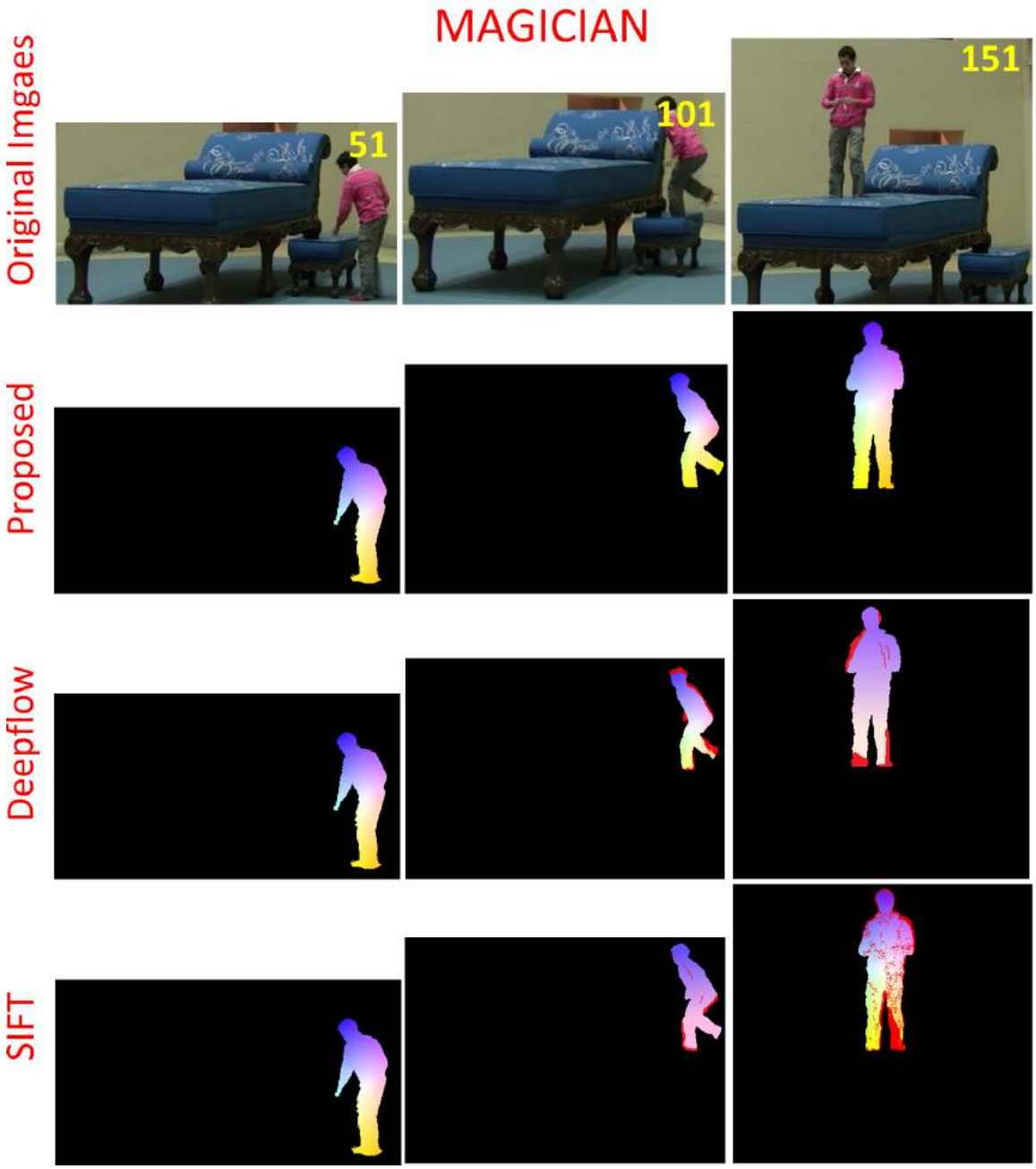


Fig. 6.16 Dense tracking comparison for one indoor dataset captured with only moving hand-held cameras.

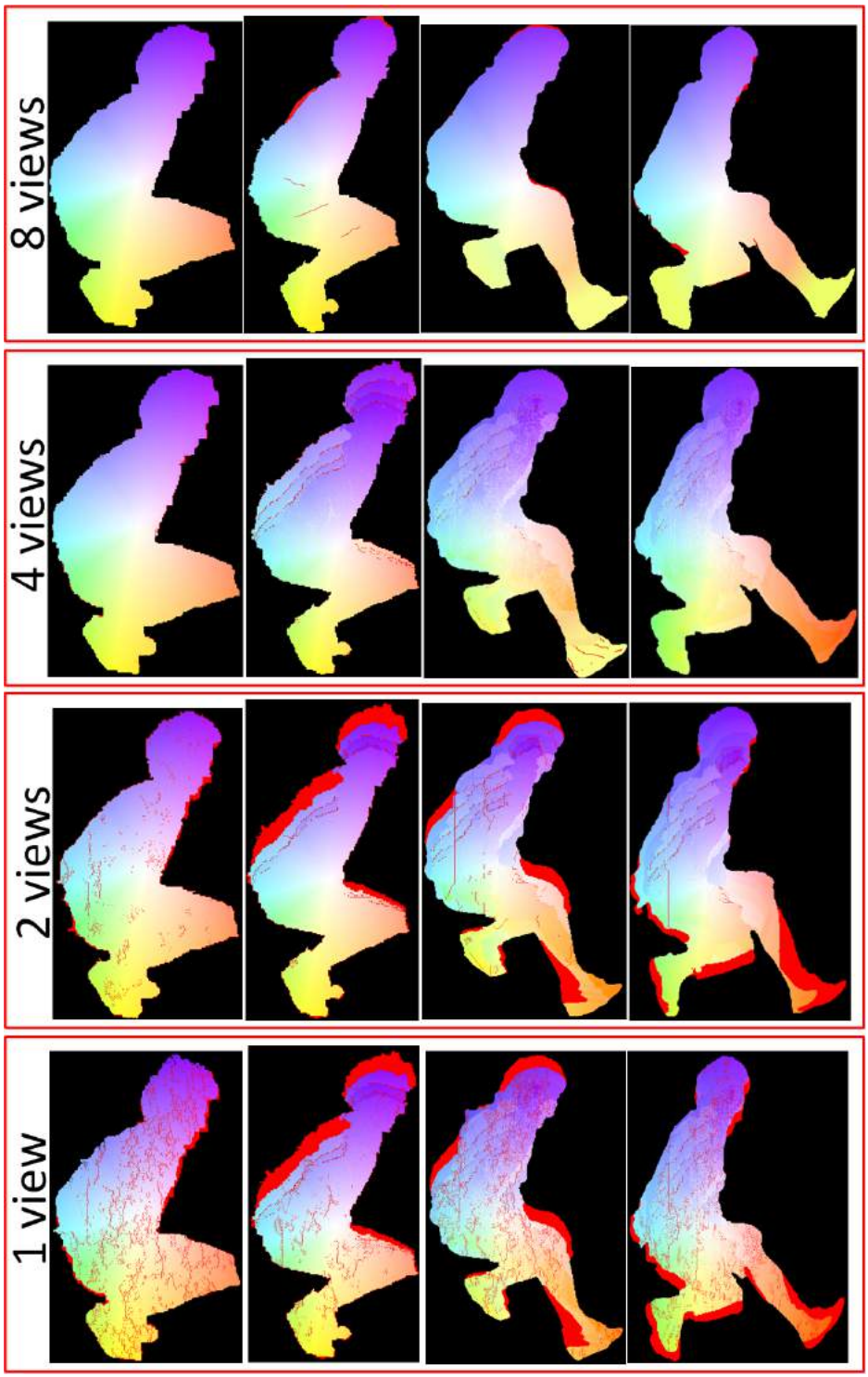


Fig. 6.17 Single and Multi-view alignment comparison results for Odzemok dataset on 2D images

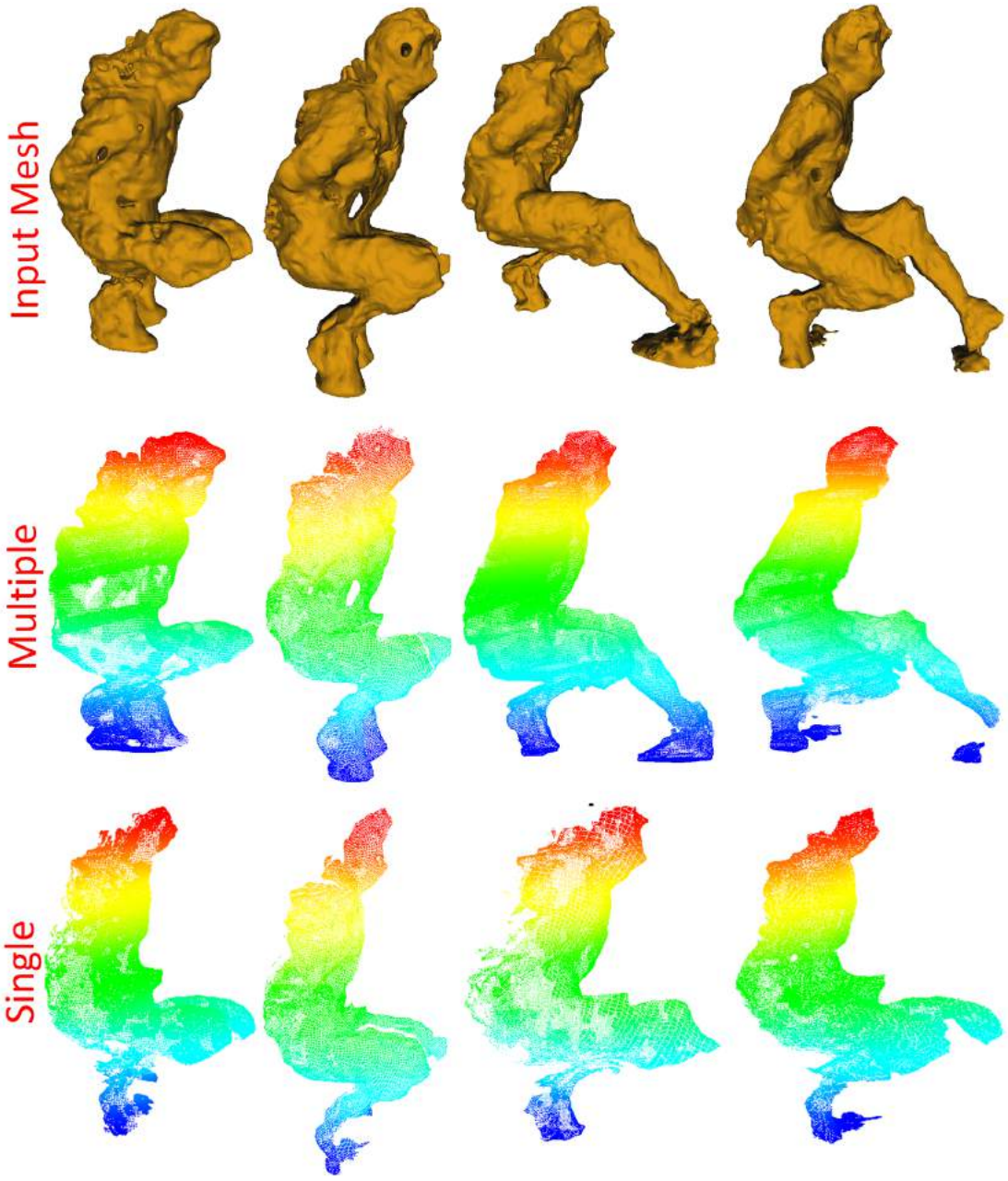


Fig. 6.18 Single and Multi-view alignment comparison results for Odzemok dataset in 3D

Datasets	Deepflow	SIFT	Nebehay	Sequential	Proposed Non-seq
Dance3	14.67	11.18	9.73	11.73	**9.59**
Dance2	15.85	11.50	10.08	12.67	**9.84**
Odzemok	17.12	12.54	10.98	13.83	**10.77**
Cathedral	16.76	12.78	11.12	13.41	**10.97**
Magician	10.48	7.99	7.006	8.38	**6.85**
Juggler	12.75	9.58	8.40	10.05	**8.22**

Table 6.6 Computational complexity per frame evaluation in seconds

6.4.5 Single vs Multi-view

The proposed 4D Match Tree global alignment method can be applied to single or multi-view image sequence with partial surface reconstruction. Dense correspondence for the Odzemok dataset using different numbers of views are compared in Figure 6.17. Dense 4D points are also evaluated for different number of views and is shown in Figure 6.18. Quantitative evaluation using *SOE* (Equation 6.7) and *completeness* (Equation 6.8) obtained from single, 2, 4 and all views for all datasets are presented in Table 6.3, 6.4 and 6.5. This shows that even with a single view the 4D Match Tree achieves 60% completeness in surface matching. As the number of views increase the restricted surface visibility improves. Completeness increases with the number of views to $> 97\%$ for all views which is significantly higher than other approaches.

6.5 Limitations:

The proposed method may fail in case of large deformation where there is high ambiguity in the surface or texture of the non-rigid object. Fast spinning objects with loose clothing in the scene lead to ambiguities in the optical flow due to the appearing/disappearing regions with large deformation. The failure of optical flow results in the failure of dense correspondence. Sparse temporal correspondence may fail in case of highly specular surfaces and objects with uniform appearance which leads to failure of temporal alignment in the sequence. Currently, the non-rigid alignment is not able to handle highly crowded dynamic environments like existing methods where no reliable sparse matches can be obtained.

6.6 Conclusion

A framework has been presented for dense 4D global alignment of partial surface reconstructions from one or more camera views of complex dynamic scenes using 4D Match trees.

4D Match Trees represent the similarity in the observed non-rigid surface shape across the sequence. This enables non-sequential alignment of partial surface reconstructions across a complete sequence to obtain dense surface correspondence across all frames. Robust wide-timeframe correspondence between pairs of frames is estimated using SFD. This sparse correspondence is used to estimate a frame-to-frame similarity in non-rigid shape and overlap between frames. The information from the sparse temporal correspondence is combined with the silhouette information of the non-rigid shape to estimate the similarity. A 4D match tree is then reconstructed as the minimum spanning tree which represents the shortest path in shape similarity space for alignment across all frames in the sequence. Dense 4D temporal correspondence is estimated from the 4D Match tree across all frames using guided optical flow. This is shown to provide improved robustness to large non-rigid deformation compared to sequential and other state-of-the-art sparse and dense correspondence methods. The proposed approach is evaluated on single and multi-view sequences of complex dynamic scenes with large non-rigid deformations to obtain a temporally consistent 4D representation. Results demonstrate completeness and accuracy of the resulting global 4D alignment over existing methods. It would also be interesting in future to add camera ego motion and object motion in the framework to improve the detection and alignment of dynamic parts in the scene.

Chapter 7

Conclusion and Future work

7.1 Conclusion

The primary goal of this research is to develop a general solution to the problem of dynamic outdoor scene reconstruction from multiple moving camera videos without any prior assumptions on scene structure or appearance. This is a challenging problem for state-of-the-art computer vision techniques due to scene complexity, relatively large volume, variable scene illumination and appearance. Existing methods require strong prior for general scene reconstruction of such complex scenes.

In this thesis, first the literature in static and dynamic scene reconstruction is reviewed, followed by a survey on each stage of the framework for general 4D scene reconstruction from multi-view data captured using static and moving cameras. This is followed by introduction of novel algorithms in consecutive chapters to obtain general 4D scene reconstruction from wide-baseline multi-view videos captured using a network of static or moving cameras automatically. An improved feature detection method to increase the distribution and coverage of features in the scene over time is proposed. This is an essential pre-requisite to dynamic scene reconstruction without prior scene segmentation or known static backgrounds.

General dynamic scene reconstruction without pre-segmentation or background information from multiple moving cameras is introduced by performing joint dense refinement on an initial coarse geometric proxy to constrain the search. The proposed approach exploits sparse feature matching for camera calibration and reconstruction of the static scene structure, together with dense stereo for foreground reconstruction. To improve reconstruction accuracy an approach to use temporal coherence and shape constraint in the optimization is proposed. A novel formulation is developed integrating sparse 3D feature constraints with dense stereo matching to obtain a complete reconstruction of both static and dynamic scene elements for general scenes. The redundancy of static structures is exploited over time to increase

computational efficiency by avoiding repeated reconstruction and improve resolution of detail by integrating reconstruction over time.

4D match trees were proposed for non-sequential non-rigid alignment of partial surface reconstructions from complex scenes to obtain 4D multi-layer segmentation and reconstruction. The approach is tested on a wide-range of general scenes with both hand-held and static cameras with the aim of demonstrating wide applicability of the approach in potential applications, such as film and broadcast production, which demand both temporal coherence and high accuracy in the reconstruction. The following sections discuss the main novel contributions of this research and suggestions for further investigation.

7.1.1 Segmentation based Feature Detection

Existing feature detectors give sparse non-uniform matches for complex dynamic scenes. The sparsity of the correspondences leads to an uneven distribution of 3D points which is often not sufficient to initialize wide-baseline stereo reconstruction. To address these problems a novel feature detector for wide-baseline matching is proposed to obtain relatively large number of uniformly distributed feature matches across the scene. This enables sparse reconstruction of 3D scene structure from wide-baseline views [126]. The approach is based on over-segmentation of the scene using the existing segmentation techniques and detecting features at three or more region intersections. This approach is demonstrated to give stable feature detection across wide-baseline views with an increased number of features and more complete scene coverage than previous feature detectors used in wide-baseline applications. A comprehensive performance evaluation against previous feature detectors (Harris, SIFT, SURF, FAST, ORB, MSER, KAZE) in combination with widely used feature descriptors (SIFT, BRIEF, ORB, SURF) demonstrates that the proposed segmentation based feature detector SFD achieves a factor $3-10$ times more wide-baseline features matches for a variety of indoor and outdoor scenes. Quantitative evaluation of feature correspondence accuracy demonstrates that SFD achieves a similar accuracy to previous wide-baseline methods with a significantly larger number of matches. Application to stereo reconstruction from wide-baseline camera views demonstrates that the SFD feature detector combined with SIFT achieves a significant increase in the number or reconstructed points and more complete scene coverage than SIFT detection and description.

More recently, there has been a trend towards feature detection with scale invariance. Although SFD keypoint detection can handle slight variations in scale, for example wide-baseline images with varying viewpoints, currently it is not scale invariant. A possible direction for future work is to extend SFD to a multi-scale approach using bilateral decomposition and performing hierarchical keypoint detection and multi-scale segmentations.

Multi-scale SFD will further improve the quality of feature detection and to obtain robust and reliable correspondences. Also it would be interesting to apply and evaluate multi-scale feature detection to various applications like Object Recognition, Registration and Tracking.

7.1.2 Dense Reconstruction of Dynamic Scenes

State-of-the-art wide-baseline multi-view dynamic reconstruction algorithms in the literature require prior knowledge to retrieve shape of dynamic objects in the scene. Chapter 4 introduces a novel method to allow general dynamic scene reconstruction without any prior knowledge of scene structure, background or appearance. An unsupervised method to initialize dynamic objects in the scene followed by joint reconstruction and segmentation of the dynamic parts of the scene is proposed. The matches obtained from the SFD features are used to obtain camera calibration and sparse reconstruction of the scene for each frame in the sequence. The sparse reconstruction is clustered in 3D and dynamic objects are identified to obtain an initial coarse reconstruction and segment the dynamic scene elements. The initial coarse reconstruction is refined using a joint segmentation and shape refinement framework automatically using the edge and photo-consistency information in the framework based on graph-cut. Evaluation of the segmentation against other state-of-the-art methods and ground-truth shows improved accuracy. Comparison of reconstruction against existing techniques demonstrates improvement in quality and accuracy of the shape estimate of dynamic objects. Computational complexity is comparable to the existing methods which require prior to obtain the results.

The proposed approach enables joint segmentation and reconstruction of general dynamic scenes without prior knowledge of scene structure of appearance. However, the approach has two principal limitations which were addressed in subsequent work. Firstly errors occur in reconstruction from frame-to-frame as each frame is reconstructed independently. Secondly the per frame reconstruction results in a temporally incoherent output. To overcome these limitations Chapter 5 introduces a general scene reconstruction approach which exploits temporal coherence.

7.1.3 Temporally Coherent Scene Reconstruction

Temporal coherence is introduced in Chapter 5 to improve the segmentation and reconstruction quality of general dynamic scenes. A novel method is introduced to exploit temporal coherence between consecutive frames to improve the quality of initial coarse reconstruction which is refined using joint optimization of reconstruction and segmentation. The sparse and dense temporal matching cues are integrated taking into account matching reliability

to introduce temporal coherence. The inherent redundancy of static scene elements over time is utilized to efficiently obtain multi-layer full scene reconstruction of complex dynamic scenes to reduce the computational cost by avoiding repeated reconstruction. Complete scene reconstruction is obtained for the first frame of the sequence and dynamic parts of the scene were updated at each frame. Static scene elements were integrated over time based on the previous reconstruction.

A novel shape constraint based on geodesic star convexity is introduced in the energy minimization framework to obtain robust joint multi-view segmentation and reconstruction. The shape of the dynamic object is restricted by the geodesic constraint which allows complex shapes to be segmented. The geodesic shape constraint is automatically initialized for each view from the initial segmentation and sparse temporal feature correspondences. Comparison with existing methods for multi-view segmentation demonstrates improvements in recovery of fine detail structure.

Evaluation shows that the quality of segmentation is comparable to the static state-of-the-art multi-view segmentation techniques. Comparison against previous joint segmentation and reconstruction refinement techniques demonstrates significant improvement in the quality and accuracy of reconstruction and segmentation for complex scenes. Reconstruction evaluation against existing methods on challenging indoor and outdoor datasets indicates a large improvement in the shape and resolution estimate of articulated objects. Complex non-rigid shapes are successfully recovered using the proposed approach, which state-of-the-art approaches are unable to retrieve. The computation time is comparable to the state-of-the-art methods which require prior information and manual interaction.

However, a few issues still remain. A frame-to-frame temporal coherence is introduced in the framework which gives rise to two problems: First, the alignment is sequential and hence leads to drift due to accumulation of errors and second, large errors in alignment may occur in the case of large motion deformation between consecutive frames which is propagated until the end of the sequence. This limitation is handled in Chapter 6 by introduction of non-sequential alignment of partial surface reconstructions from multiple cameras.

7.1.4 Robust 4D Scene Reconstruction

The ‘4D match trees’ is introduced in Chapter 6 to perform non-rigid non-sequential temporal alignment between partial surfaces for any pair of frames in the sequence. The alignment is based on a new measure of similarity exploiting temporal matching information. Sparse temporal feature matching between widely spaced frames with large non-rigid deformations is achieved using SFD previously introduced for wide-baseline spatial matching. This enables computation of surface overlap and similarity between any pair of frames used in ‘4D

match trees' to initialize temporal alignment of partial surface reconstructions. 4D temporal alignment is achieved for a wide variety of challenging indoor and outdoor scenes including fast non-rigid and no or rigid motion cases. As in previous work temporal alignment is evaluated against other methods using quantitative silhouette overlap and completeness measures. This shows significant improvement in the resulting dense 4D correspondence. The correspondence is consistent across the sequence and a significant improvement is achieved over sequential matching.

7.1.5 Contributions

The overall goal of the research is to obtain robust 4D scene reconstruction of general complex dynamic scenes from multiple moving cameras without any prior on scene structure, background or appearance. The contributions are concluded below:

- A novel segmentation based feature detector is proposed for uniform, stable wide-baseline and wide-timeframe matching.
- An automatic framework to initialize, reconstruct and segment general dynamic scenes without any prior is introduced.
- Temporal coherence is exploited in the framework to improve the segmentation and reconstruction quality and accuracy. Static redundancy is exploited over time to obtain complete scene reconstruction and reduce complexity.
- Shape constraint based on geodesic star convexity is integrated in the joint reconstruction and segmentation refinement framework to improve the quality of the results.
- '4D Match Trees' are introduced to perform non-sequential alignment of partial surfaces for robust 4D reconstruction of general dynamic scenes.

7.2 Future work

For future work, the proposed framework for general 4D scene reconstruction can be extended in multiple directions:

- The system currently works for multi-view videos, but it is possible to extend the system to a single moving camera. In the literature methods have been introduced for dynamic scene reconstruction from a single video using non-rigid structure from motion [151]. The same principle can be potentially applied to obtain the initial coarse

reconstruction which can be tracked and optimized over time to obtain dense 4D scene reconstruction from a single view. Initial coarse reconstruction can be obtained from the sparse 3D points at each frame on the object. The joint segmentation and reconstruction framework proposed in this work optimizes each view independently which makes it easier to apply this framework for a single camera. Temporal coherence can be introduced by using the trajectory of 3D points which can be estimated using existing methods [168] followed by non-sequential alignment of these per frame partial surfaces for the entire sequence.

- Joint semantic segmentation and reconstruction combines recognition with reconstruction to simultaneously identify the objects in the scene. State-of-the-art methods for joint semantic segmentation and reconstruction only work for static scenes [66] and it would be interesting to extend this concept to dynamic scenes by integrating recognition with the proposed framework. Semantic labels can be estimated for each pixel in the multi-view images at each frame using existing fully convolution network based semantic segmentation techniques[104]. The probabilities of each class at each pixel can be integrated in the joint segmentation and reconstruction refinement framework to obtain semantic segmentation and reconstruction of general dynamic scenes.

- Currently the system cannot handle crowded dynamic scenes with a large motion in background and foreground. It would be interesting in future to handle more complex and crowded dynamic scenes. In the literature methods have been proposed to build 3D models for dynamic scenes containing structures[76]. It would be interesting to extend such methods to more general complex dynamic scenes with a large number of people. Robust and efficient methods needs to be developed to identify and reconstruct general dynamic scenes. Instead of joint reconstruction and segmentation proposed in this work, only reconstruction can be optimized and temporal information can be exploited to offer a more robust solution.

- Research can be extended to online video-rate reconstruction of general outdoor dynamic scenes for onset reconstruction in the film industry to support directorial decision. Methods have been proposed to obtain real-time reconstruction from single camera and kinect data [130]. Online implementation for video-rate reconstruction is possible by exploiting GPU implementation of key processing bottlenecks such as stereo and feature matching in this work. Computational bottlenecks can be identified and efficient and faster solutions need to be developed. Novel general representations for dynamic scene structure can be considered for efficient storage and visualization.

References

[cvs] 4d and multiview video repository, http://cvssp.org/data/cvssp3d/. In *Centre for Vision Speech and Signal Processing, University of Surrey, UK*.

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

[3] (2007). Stagetm,organic motion., http://www.organicmotion.com/technology.

[4] (2008). Middlebury, http://vision.middlebury.edu/stereo/eval/.

[5] Aanæs, H., Dahl, A. L., and Pedersen, K. S. (2012). Interesting Interest Points - A Comparative Study of Interest Point Performance on a Unique Data Set. *International Journal of Computer Vision (IJCV)*, 97(1):18–35.

[6] Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., and Susstrunk, S. (2012). Slic superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 34(11):2274–2282.

[7] Agarwal, S., Furukawa, Y., Snavely, N., Simon, I., Curless, B., Seitz, S. M., and Szeliski, R. (2011). Building rome in a day. *Communication ACM*, 54(10):105–112.

[8] Agrawal, M., Konolige, K., and Blas, M. (2008). Censure: Center surround extremas for realtime feature detection and matching. In *European Conference on Computer Vision (ECCV)*, pages 102–115.

[9] Akbarzadeh, A., Frahm, J.-M., Mordohai, P., Clipp, B., Engels, C., Gallup, D., Merrell, P., Phelps, M., Sinha, S., Talton, B., Wang, L., Yang, Q., Stewenius, H., Yang, R., Welch, G., Towles, H., Nister, D., and Pollefeys, M. (2006). Towards Urban 3D Reconstruction from Video. In *3D Data Processing, Visualization, and Transmission, Third International Symposium on*, pages 1–8.

[10] Alcantarilla, P. F., Bartoli, A., and Davison, A. J. (2012). KAZE Features. In *European Conference on Computer Vision (ECCV)*, pages 214–227.

[11] Alcantarilla, P. F., Nuevo, J., and Bartoli, A. (2013). Fast Explicit Diffusion for Accelerated Features in Nonlinear Scale Spaces. In *The British Machine Vision Conference (BMVC)*.

[12] Aloimonos, J. Y. (1990). Perspective approximations. *Image Vision Computing*, 8(3):179–192.

[13] Awrangjeb, M., Lu, G., and Fraser, C. S. (2012). Performance Comparisons of Contour-Based Corner Detectors. *IEEE Transactions on Image Processing*, 21(4):4167–4179.

[14] Bailer, C., Taetz, B., and Stricker, D. (2015). Flow fields: Dense correspondence fields for highly accurate large displacement optical flow estimation. In *The IEEE International Conference on Computer Vision (ICCV)*.

[15] Ballan, L., Brostow, G. J., Puwein, J., and Pollefeys, M. (2010). Unstructured video-based rendering: Interactive exploration of casually captured videos. *ACM Transactions on Graphics*, 29(4):1–11.

[16] Basha, T., Moses, Y., and Kiryati, N. (2010). Multi-view scene flow estimation: A view centered variational approach. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1506–1513.

[17] Basri, R. and Jacobs, D. (1995). Recognition using region correspondences. *International Journal of Computer Vision (IJCV)*, 25(9):8–13.

[18] Bay, H., Tuytelaars, T., and Gool, L. (2006). Surf: Speeded up robust features. In *European Conference on Computer Vision (ECCV)*, pages 404–417.

[19] Beardsley, P., Torr, P., and Zisserman, A. (1995). 3d model acquisition from extended image sequences.

[20] Beeler, T., Hahn, F., Bradley, D., Bickel, B., Beardsley, P., Gotsman, C., Sumner, R. W., and Gross, M. (2011). High-quality passive facial performance capture using anchor frames. *ACM Transaction in Graphics*, 30(4):75:1–75:10.

[21] Beymer, D. J. (1991). Finding junctions using the image gradient. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 720–721.

[22] Bhotika, R., Fleet, D. J., and Kutulakos, K. N. (2002). A probabilistic theory of occupancy and emptiness. In *European Conference on Computer Vision (ECCV)*, volume 2352, pages 112–130.

[23] Bleyer, M., Rhemann, C., and Rother, C. (2011). Patchmatch stereo - stereo matching with slanted support windows. In *The British Machine Vision Conference (BMVC)*.

[24] Boykov, Y. and Kolmogorov, V. (2004). An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 26(11):1124–1137.

[25] Boykov, Y., Veksler, O., and Zabih, R. (2001). Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 23(11):1222–1239.

[26] Brown, D. C. (1976). The bundle adjustment - progress and prospects. *Congress of the International Society for Photogrammetry*.

[27] Budd, C., Huang, P., Klaudiny, M., and Hilton, A. (2013). Global non-rigid alignment of surface sequences. *International Journal of Computer Vision (IJCV)*, 102(1-3):256–270.

[28] Cagniart, C., Boyer, E., and Ilic, S. (2010). Probabilistic deformable surface tracking from multiple videos. In *European Conference on Computer Vision (ECCV)*, pages 326–339.

[29] Calonder, M., Lepetit, V., Strecha, C., and Fua, P. (2010). Brief: Binary robust independent elementary features. In *European Conference on Computer Vision (ECCV)*, pages 778–792.

[30] Campbell, N. D. F., Vogiatzis, G., Hernandez, C., and Cipolla, R. (2007). Automatic 3d object segmentation in multiple views using volumetric graph-cuts. In *The British Machine Vision Conference (BMVC)*, pages 58.1–58.10.

[31] Cao, X., Wei, Y., Wen, F., and Sun, J. (2012). Face alignment by explicit shape regression. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[32] Clarke, T. A. and Fryer, J. G. (1998). The development of camera calibration methods and models. *The Photogrammetric Record*, 16(91):51–66.

[33] Collet, A., Chuang, M., Sweeney, P., Gillett, D., Evseev, D., Calabrese, D., Hoppe, H., Kirk, A., and Sullivan, S. (2015). High-quality streamable free-viewpoint video. *ACM Transactions on Graphics*, 34(4):69:1–69:13.

[34] Comaniciu, D. and Meer, P. (2002). Mean shift: a robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 24(4):603–619.

[35] Cornelis, N., Leibe, B., Cornelis, K., and Gool, L. (2008). 3d urban scene modeling integrating recognition and reconstruction. *International Journal of Computer Vision (IJCV)*, 78(2):121–141.

[36] Coughlan, J. M. and Yuille, A. L. (2000). The manhattan world assumption: Regularities in scene statistics which enable bayesian inference. In *Neural Information Processing Systems (NIPS)*, pages 845–851.

[37] Dimitrov, D., Knauer, C., Kriegel, K., and Rote, G. (2006). On the bounding boxes obtained by principal component analysis. In *Proc. 22nd European Workshop on Computational Geometry*, pages 193–196.

[38] Djelouah, A., Franco, J.-S., Boyer, E., Le Clerc, F., and Perez, P. (2013). Multi-view object segmentation in space and time. In *The IEEE International Conference on Computer Vision (ICCV)*, pages 2640–2647.

[39] Djelouah, A., Franco, J.-S., Boyer, E., Le Clerc, F., and Perez, P. (2015). Sparse multi-view consistency for object segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 37(9):1890–1903.

[40] Dou, M., Khamis, S., Degtyarev, Y., Davidson, P., Fanello, S. R., Kowdle, A., Escolano, S. O., Rhemann, C., Kim, D., Taylor, J., Kohli, P., Tankovich, V., and Izadi, S. (2016). Fusion4d: Real-time performance capture of challenging scenes. *ACM Transaction on Graphics*, 35(4):114:1–114:13.

[41] Evangelidis, G. D. and Psarakis, E. Z. (2008). Parametric image alignment using enhanced correlation coefficient maximization. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 30(10):1858–1865.

[42] Farnebäck, G. (2003). Two-frame motion estimation based on polynomial expansion. In *SCIA*, pages 363–370.

[43] Faugeras, O. (1993). *Three-dimensional Computer Vision: A Geometric Viewpoint*. MIT Press.

[44] Fitzgibbon, A., Wexler, Y., and Zisserman, A. (2003). Image-based rendering using image-based priors. In *The IEEE International Conference on Computer Vision (ICCV)*, volume 2, pages 1176–1183.

[45] Forsyth, D. A. and Ponce, J. (2002). *Computer Vision: A Modern Approach*. Prentice Hall Professional Technical Reference.

[46] Fortune, S. (1997). Handbook of discrete and computational geometry. pages 377–388.

[47] Föstner, M. A. and Gülch, E. (1987). A Fast Operator for Detection and Precise Location of Distinct Points, Corners and Centers of Circular Features. In *ISPRS Intercommission Workshop*.

[48] Franco, J. and Boyer, E. (2005). Fusion of multiview silhouette cues using a space occupancy grid. In *The IEEE International Conference on Computer Vision (ICCV)*, volume 2, pages 1747–1753 Vol. 2.

[49] Franco, J.-S. and Boyer, E. (2003). Exact polyhedral visual hulls. In *The British Machine Vision Conference (BMVC)*, pages 32.1–32.10.

[50] Fuhrmann, S. and Goesele, M. (2011). Fusion of depth maps with multiple scales. *ACM Transactions on Graphics*, 30(6):148:1–148:8.

[51] Furukawa, Y. and Ponce, J. (2010a). Accurate, dense, and robust multiview stereopsis. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 32(8):1362–1376.

[52] Furukawa, Y. and Ponce, J. (2010b). *Dense 3D Motion Capture from Synchronized Video Streams*, pages 193–211. Springer Berlin Heidelberg.

[53] Gall, J., Stoll, C., Aguiar, E. D., Theobalt, C., Rosenhahn, B., and Seidel, H.-P. (2009). Motion capture using joint skeleton tracking and surface estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[54] Garg, R., Roussos, A., and Agapito, L. (2013). Dense variational reconstruction of non-rigid surfaces from monocular video. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1272–1279.

[55] Gauglitz, S., Höllerer, T., and Turk, M. (2011). Evaluation of interest point detectors and feature descriptors for visual tracking. *International Journal of Computer Vision (IJCV)*, 94(3):335–360.

[56] Gherardi, R., Farenzena, M., and Fusiello, A. (2010). Improving the efficiency of hierarchical structure-and-motion. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1594–1600.

[57] Goesele, M., Snavely, N., Curless, B., Hoppe, H., and Seitz, S. (2007). Multi-view stereo for community photo collections. In *The IEEE International Conference on Computer Vision (ICCV)*, pages 1–8.

[58] Goldluecke, B. and Magnor, M. (2004). Space-time isosurface evolution for temporally coherent 3d reconstruction. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 350–355.

[59] Grauman, K., Shakhnarovich, G., and Darrell, T. (2003). A bayesian approach to image-based visual hull reconstruction. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 187–194.

[60] Grundmann, M., Kwatra, V., Han, M., and Essa, I. (2010). Efficient hierarchical graph-based video segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[61] Guan, L., Franco, J. S., and Pollefeys, M. (2010). Multi-view occlusion reasoning for probabilistic silhouette-based dynamic scene reconstruction. *International Journal of Computer Vision (IJCV)*, 90(3):283–303.

[62] Guillemaut, J. Y. and Hilton, A. (2010). Joint Multi-Layer Segmentation and Reconstruction for Free-Viewpoint Video Applications. *International Journal of Computer Vision (IJCV)*, 93(1):73–100.

[63] Guillemaut, J.-Y. and Hilton, A. (2012). Space-time joint multi-layer segmentation and depth estimation. In *International Conference on 3D Imaging, Modeling, Processing, Visualization and Transmission (3DIMPVT)*, pages 440–447.

[64] Gulshan, V., Rother, C., Criminisi, A., Blake, A., and Zisserman, A. (2010). Geodesic star convexity for interactive image segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3129–3136.

[65] Guo, K., Xu, F., Wang, Y., Liu, Y., and Dai, Q. (2015). Robust non-rigid motion tracking and surface reconstruction using l0 regularization. In *The IEEE International Conference on Computer Vision (ICCV)*.

[66] Hane, C., Zach, C., Cohen, A., Angst, R., and Pollefeys, M. (2013). Joint 3d scene reconstruction and class segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 97–104.

[67] Haris, K., Efstratiadis, S. N., Maglaveras, N., and Katsaggelos, A. K. (1998). Hybrid image segmentation using watersheds and fast region merging. *IEEE Transactions on Image Processing*, 7(6):1684–1699.

[68] Harris, C. and Stephens, M. (1988). A combined corner and edge detector. In *Fourth Alvey Vision Conference*, pages 147–151.

[69] Hartley, R. and Zisserman, A. (2003). *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2 edition.

[70] Hartmann, W., Havlena, M., and Schindler, K. (2014). Predicting matchability. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9–16.

[71] Hauagge, D. C. and Snavely, N. (2012). Image matching using local symmetry features. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 206–213.

[72] Hu, X. and Mordohai, P. (2012). A quantitative evaluation of confidence measures for stereo vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 34(8):2121–2133.

[73] Huang, C., Cagniart, C., Boyer, E., and Ilic, S. (2016). A bayesian approach to multi-view 4d modeling. *International Journal of Computer Vision (IJCV)*, 116(2):115–135.

[74] Imre, E., Guillemaut, J. Y., and Hilton, A. (2011). Calibration of nodal and free-moving cameras in dynamic scenes for post-production. In *International Conference on 3D Imaging, Modeling, Processing, Visualization and Transmission (3DIMPVT)*, pages 260–267.

[75] Izadi, S., Kim, D., Hilliges, O., Molyneaux, D., Newcombe, R., Kohli, P., Shotton, J., Hodges, S., Freeman, D., Davison, A., and Fitzgibbon, A. (2011). Kinectfusion: Real-time 3d reconstruction and interaction using a moving depth camera. In *Annual ACM Symposium on User Interface Software and Technology*, pages 559–568.

[76] Ji, D., Dunn, E., and Frahm, J. M. (2014). 3d reconstruction of dynamic textures in crowd sourced data. In *European Conference on Computer Vision (ECCV)*, volume 8689, pages 143–158.

[77] Jiang, H., Liu, H., Tan, P., Zhang, G., and Bao, H. (2012). 3d reconstruction of dynamic scenes with multiple handheld cameras. In *European Conference on Computer Vision (ECCV)*, pages 601–615.

[78] Joo, H., Liu, H., Tan, L., Gui, L., Nabbe, B., Matthews, I., Kanade, T., Nobuhara, S., and Sheikh, Y. (2015). Panoptic studio: A massively multiview system for social motion capture. In *The IEEE International Conference on Computer Vision (ICCV)*.

[79] Joo, H., Soo Park, H., and Sheikh, Y. (2014). Map visibility estimation for large-scale dynamic 3d reconstruction. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[80] Kanade, T., Rander, P., and Narayanan, P. J. (1997). Virtualized reality: Constructing virtual worlds from real scenes. *IEEE MultiMedia*, 4(5):34–47.

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

[82] Kim, H., Guillemaut, J., Takai, T., Sarim, M., and Hilton, A. (2012). Outdoor Dynamic 3-D Scene Reconstruction. *IEEE transactions on Circuits and Systems for Video Technology (T-CSVT)*, 22(11):1611–1622.

[83] Kim, J. and Grauman, K. (2011). Boundary preserving dense local regions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1153–1560.

[84] Klaudiny, M. and Hilton, A. (2012). High-fidelity facial performance capture with non-sequential temporal alignment. In *3rd Symposium on Facial Analysis and Animation (FAA)*, pages 3:1–3:1.

[85] Kohli, P., Rihan, J., Bray, M., and Torr, P. H. (2008). Simultaneous segmentation and pose estimation of humans using dynamic graph cuts. *International Journal of Computer Vision (IJCV)*, 79(3):285–298.

[86] Kolev, K., Klodt, M., Brox, T., and Cremers, D. (2009). Continuous global optimization in multiview 3d reconstruction. *International Journal of Computer Vision (IJCV)*, 84(4):80–96.

[87] Kolmogorov, V., Criminisi, A., Blake, A., Cross, G., and Rother, C. (2006). Probabilistic fusion of stereo with color and contrast for bilayer segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 28(9):2006.

[88] Koniusz, P. and Mikolajczyk (2009). Segmentation based interest points and evaluation of unsupervised image segmentation methods. In *The British Machine Vision Conference (BMVC)*.

[89] Kowdle, A., Sinha, S., and Szeliski, R. (2012). Multiple view object cosegmentation using appearance and stereo cues. In *European Conference on Computer Vision (ECCV)*, pages 789–803.

[90] Krähenbühl, P. and Koltun, V. (2011). Efficient inference in fully connected crfs with gaussian edge potentials. In *Advances in Neural Information Processing Systems 24*, pages 109–117.

[91] Kruskal, J. B. (1956). On the Shortest Spanning Subtree of a Graph and the Traveling Salesman Problem. In *Proceedings of the American Mathematical Society, 7*.

[92] Kundu, A., Li, Y., Dellaert, F., Li, F., and Rehg, J. M. (2014). Joint semantic segmentation and 3d reconstruction from monocular video. In *European Conference on Computer Vision (ECCV)*, volume 8694, pages 703–718.

[93] Kutulakos, K. N. and Seitz, S. M. (2000). A theory of shape by space carving. *International Journal of Computer Vision (IJCV)*, 38(3):199–218.

[94] Kyle, W. and Snavely, N. (2014). Robust global translations with 1dsfm. In *European Conference on Computer Vision (ECCV)*, pages 61–75.

[95] Lai, P. L. and Yilmaz, A. (2009). Shape recovery using rotated slicing planes. In *Image and Signal Processing*, pages 1–5.

[96] Larsen, E., Mordohai, P., Pollefeys, M., and Fuchs, H. (2007). Temporally consistent reconstruction from multiple video streams using enhanced belief propagation. In *The IEEE International Conference on Computer Vision (ICCV)*, pages 1–8.

[97] Laurentini, A. (1994). The visual hull concept for silhouette-based image understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 16(2):150–162.

[98] Lee, W., Woo, W., and Boyer, E. (2011). Silhouette segmentation in multiple views. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 33(9):1429–1441.

[99] Lei, C., Chen, X. D., and Yang, Y. H. (2009). A new multiview spacetime-consistent depth recovery framework for free viewpoint video rendering. In *The IEEE International Conference on Computer Vision (ICCV)*, pages 1570–1577.

[100] Lepetit, V. and Fua, P. (2006). Keypoint recognition using randomized trees. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 28(9):1465–1479.

[101] Leutenegger, S., Chli, M., and Siegwart, R. Y. (2011). Brisk: Binary robust invariant scalable keypoints. In *The IEEE International Conference on Computer Vision (ICCV)*, pages 2548–2555.

[102] Levenberg, K. (1944). A method for the solution of certain non-linear problems in least squares. *Quarterly Journal of Applied Mathmatics*, 2(2):164–168.

[103] Locher, A., Perdoch, M., and Van Gool, L. (2016). Progressive prioritized multi-view stereo. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[104] Long, J., Shelhamer, E., and Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[105] Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision (IJCV)*, 60(2):91–110.

[106] Magnor, M., Grau, O., Sorkine-Hornung, O., and Theobalt, C. (2015). *Digital Representations of the Real World: How to Capture, Model, and Render Visual Reality*. CRC Press, 1 edition.

[107] Maire, M., Arbelaez, P., Fowlkes, C., and Malik, J. (2008). Using contours to detect and localize junctions in natural images. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8.

[108] Malleson, C., Klaudiny, M., Guillemaut, J.-Y., and Hilton, A. (2014). Structured representation of non-rigid surfaces from single view 3d point tracks. In *International Conference on 3D Vision (3DV)*.

[109] Matas, J., Chum, O., Urban, M., and Pajdla, T. (2002). Robust wide baseline stereo from maximally stable extremal regions. In *The British Machine Vision Conference (BMVC)*, pages 36.1–36.10.

[110] Matsuyama, T., Xiaojun, W., Takai, T., and Wada, T. (2004). Real-time dynamic 3-d object shape reconstruction and high-fidelity texture mapping for 3-d video. *IEEE transactions on Circuits and Systems for Video Technology (T-CSVT)*, 14(3):357–369.

[111] Matthies, L. (1992). Stereo vision for planetary rovers: Stochastic modeling to near real-time implementation. *International Journal of Computer Vision (IJCV)*, 8(1):71–91.

[112] Matusik, W., Buehler, C., and McMillan, L. (2001). Polyhedral visual hulls for real-time rendering. In *Eurographics Workshop on Rendering*, pages 115–125.

[113] Menze, M. and Geiger, A. (2015). Object scene flow for autonomous vehicles. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[114] Merrell, P., Akbarzadeh, A., Wang, L., Frahm, J.-M., Yang, R., and Nistér, D. (2007). Real-time visibility-based fusion of depth maps. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[115] Meyer, F. (2001). An overview of morphological segmentation. *International Journal of Pattern Recognition and Artificial Intelligence*, 15(2):1089–1118.

[116] Mikolajczyk, K. and Schmid, C. (2004). Scale and affine invariant interest point detectors. *International Journal of Computer Vision (IJCV)*, 60:63–86.

[117] Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., Kadir, T., and Gool, L. V. (2005). A comparison of affine region detectors. *International Journal of Computer Vision (IJCV)*, 65(1-2):43–72.

[118] Miller, G. and Hilton, A. (2007). Safe hulls. In *European Conference on Visual Media Production (CVMP)*, pages 1–8.

[119] Mitchelson, J. and Hilton, A. (2003). Wand-based multiple camera studio calibration. Technical report.

[120] Mokhtarian, F. and Suomela, R. (1998). Robust Image Corner Detection Through Curvature Scale Space. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 20(12):1376–1381.

[121] Moravec, H. (1980). Obstacle avoidance and navigation in the real world by a seeing robot rover. Technical report.

[122] Mustafa, A., Kim, H., Guillemaut, J., and Hilton, A. (2015a). General dynamic scene reconstruction from wide-baseline views. In *The IEEE International Conference on Computer Vision (ICCV)*.

[123] Mustafa, A., Kim, H., Guillemaut, J.-Y., and Hilton, A. (2016a). Temporally coherent 4d reconstruction of complex dynamic scenes. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Oral*.

[124] Mustafa, A., Kim, H., and Hilton, A. (2016b). 4d match trees for non-rigid surface alignment. In *European Conference on Computer Vision (ECCV)*.

[125] Mustafa, A., Kim, H., Imre, E., and Hilton, A. (2014). Initial disparity estimation using sparse matching for wide-baseline dense stereo. In *European Conference on Visual Media Production (CVMP)*.

[126] Mustafa, A., Kim, H., Imre, E., and Hilton, A. (2015b). Segmentation based features for wide-baseline multi-view reconstruction. In *International Conference on 3D Vision (3DV)*.

[127] Narayana, M., Hanson, A., and Learned-Miller, E. (2013). Coherent motion segmentation in moving camera videos using optical flow orientations. In *The IEEE International Conference on Computer Vision (ICCV)*, pages 1577–1584.

[128] Nebehay, G. and Pflugfelder, R. (2015). Clustering of static-adaptive correspondences for deformable dbject tracking. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[129] Newcombe, R., Fox, D., and Seitz, S. (2015). Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[130] Newcombe, R., Lovegrove, S., and Davison, A. (2011). DTAM: Dense Tracking and Mapping in Real-Time. In *The IEEE International Conference on Computer Vision (ICCV)*.

[131] Niessner, M., Zollhofer, M., Izadi, S., and Stamminger, M. (2013). Real-time 3d reconstruction at scale using voxel hashing. *ACM Transactions in Graphics*, 32(6):169:1–169:11.

[132] Nistér, D. (2004). An efficient solution to the five-point relative pose problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 26(6):756–777.

[133] Nurutdinova, I. and Fitzgibbon, A. (2015). Towards pointless structure from motion: 3d reconstruction and camera parameters from general 3d curves. In *The IEEE International Conference on Computer Vision (ICCV)*, pages 2363–2371.

[134] Oswald, M., Stühmer, J., and Cremers, D. (2014). Generalized connectivity constraints for spatio-temporal 3d reconstruction. In *European Conference on Computer Vision (ECCV)*, pages 32–46.

[135] Ozden, K., Schindler, K., and Van Gool, L. (2007). Simultaneous segmentation and 3d reconstruction of monocular image sequences. In *The IEEE International Conference on Computer Vision (ICCV)*, pages 1–8.

[136] Papazoglou, A. and Ferrari, V. (2013). Fast object segmentation in unconstrained video. In *The IEEE International Conference on Computer Vision (ICCV)*, pages 1777–1784.

[137] Perona, P. and Malik, J. (1990). Scale-space and edge detection using anisotropic diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 12(7).

[138] Poelman, C. and Kanade, T. (1994). *A paraperspective factorization method for shape and motion recovery*, pages 97–108. Springer Berlin Heidelberg.

[139] Pollefeys, M., Koch, R., Vergauwen, M., and Gool, L. V. (2000). Automated reconstruction of 3d scenes from sequences of images. *ISPRS Journal Of Photogrammetry And Remote Sensing*, 55(4):251–267.

[140] Pollefeys, M., Nistér, D., Frahm, J.-M., Akbarzadeh, A., Mordohai, P., Clipp, B., Engels, C., Gallup, D., Kim, S.-J., Merrell, P., Salmi, C., Sinha, S., Talton, B., Wang, L., Yang, Q., Stewénius, H., Yang, R., Welch, G., and Towles, H. (2007). Detailed Real-Time Urban 3D Reconstruction from Video. volume 78, pages 143–167.

[141] Prada, F., Kazhdan, M., Chuang, M., Collet, A., and Hoppe, H. (2016). Motion graphs for unstructured textured meshes. *ACM Transaction in Graphics*, 35(4):108:1–108:14.

[142] Prim, R. C. (1957). Shortest connection networks and some generalizations. *The Bell Systems Technical Journal*, 36(11):1389–1401.

[143] Pritchett, P. and Zisserman, A. (1998). Wide baseline stereo matching. In *The IEEE International Conference on Computer Vision (ICCV)*, pages 754–760.

[144] Rhodin, H., Roberitini, N., Casas, D., Richardt, C., Seidel, H.-P., and Theobalt, C. (2016). General automatic human shape and motion capture using volumetric contour cues. In *European Conference on Computer Vision (ECCV)*.

[145] Richardson, A. and Olson, E. (2013). Learning convolutional filters for interest point detection. In *The International Conference in Robotics and Automation (ICRA)*.

[146] Roerdink, J. B. and Meijster, A. (2000). The watershed transform: Definitions, algorithms and parallelization strategies. *Fundam. Inf.*, 41(3):187–228.

[147] Rosten, E. and Drummond, T. (2005). Fusing points and lines for high performance tracking. In *The IEEE International Conference on Computer Vision (ICCV)*, pages 1508–1511.

[148] Rosten, E. and Drummond, T. (2006). *Machine Learning for High-Speed Corner Detection*, pages 430–443.

[149] Rosten, E., Porter, R., and Drummond, T. (2010). Faster and better: A machine learning approach to corner detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 32(1):105–119.

[150] Rublee, E., Rabaud, V., Konolige, K., and Bradski, G. (2011). Orb: An efficient alternative to sift or surf. In *The IEEE International Conference on Computer Vision (ICCV)*, pages 2564–2571.

[151] Russell, C., Yu, R., and Agapito, L. (2014). Video pop-up: Monocular 3d reconstruction of dynamic scenes. In *European Conference on Computer Vision (ECCV)*, pages 583–598.

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

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

[154] Scharstein, D. and Szeliski, R. (2002). A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms. *International Journal of Computer Vision (IJCV)*, 47(1-3):7–42.

[155] Schonberger, J. L. and Frahm, J.-M. (2016). Structure-from-motion revisited. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[156] Seitz, S., Curless, B., Diebel, J., Scharstein, D., and Szeliski, R. (2006). A comparison and evaluation of multi-view stereo reconstruction algorithms. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 519–528.

[157] Shi, J. and Tomasi, C. (1994). Good features to track. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 593–600.

[158] Shin, Y. M., Cho, M., and Lee, K. M. (2013). Multi-object reconstruction from dynamic scenes: An object-centered approach. *Computer Vision and Image Understanding (CVIU)*, 117(11):1575 – 1588.

[159] Sinha, S., Mordohai, P., and Pollefeys, M. (2007). Multi-view stereo via graph cuts on the dual of an adaptive tetrahedral mesh. In *The IEEE International Conference on Computer Vision (ICCV)*, pages 1–8.

[160] Slabaugh, G. G., Culbertson, W. B., M., T., Stevens, M. R., and Schafer, R. W. (2003). Methods for volumetric reconstruction of visual scenes. *International Journal of Computer Vision (IJCV)*, 57(3):179–199.

[161] Smith, S. M. and Brady, J. M. (1997). SUSAN - A New Approach to Low Level Image Processing. *International Journal of Computer Vision (IJCV)*, 23(1):45–78.

[162] Snavely, N., Seitz, S. M., and Szeliski, R. (2008). Modeling the world from internet photo collections. *International Journal of Computer Vision (IJCV)*, 80(2):189–210.

[163] Sobel, I. (2014). History and definition of the sobel operator.

[164] Starck, J. and Hilton, A. (2003). Model-based multiple view reconstruction of people. In *The IEEE International Conference on Computer Vision (ICCV)*, pages 915–922.

[165] Starck, J. and Hilton, A. (2007). Surface Capture for Performance-Based Animation. *IEEE Computer Graphics and Applications*, 27(3):21–31.

[166] Starck, J., Maki, A., Nobuhara, S., Hilton, A., and Matsuyama, T. (2009). The Multiple-Camera 3-D Production Studio. *IEEE transactions on Circuits and Systems for Video Technology (T-CSVT)*, 19(6):856–869.

[167] Strecha, C., Von Hansen, W., Van Gool, L., Fua, P., and Thoennessen, U. (2008). On benchmarking camera calibration and multi-view stereo for high resolution imagery. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8.

[168] Sundaram, N., Brox, T., and Keutzer, K. (2010). Dense point trajectories by gpu-accelerated large displacement optical flow. In *European Conference on Computer Vision (ECCV)*, pages 438–451.

[169] Szeliski, R. and Golland, P. (1998). Stereo matching with transparency and matting. In *The IEEE International Conference on Computer Vision (ICCV)*, pages 517–524.

[170] Taneja, A., Ballan, L., and Pollefeys, M. (2011). Modeling dynamic scenes recorded with freely moving cameras. In *Asian Conference on Computer Vision (ACCV)*, pages 613–626.

[171] Tappen, M. F. and Freeman, W. T. (2003). Comparison of graph cuts with belief propagation for stereo, using identical mrf parameters. In *The IEEE International Conference on Computer Vision (ICCV)*, pages 900–908.

[172] Tevs, A., Berner, A., Wand, M., Ihrke, I., Bokeloh, M., Kerber, J., and Seidel, H.-P. (2012). Animation cartography; intrinsic reconstruction of shape and motion. *ACM Transactions on Graphics*, 31(2):12:1–12:15.

[173] Tomasi, C. (1992). Shape and motion from image streams under orthography: a factorization method. *International Journal of Computer Vision (IJCV)*, 9(2):137–154.

[174] Tomasi, C. and Manduchi, R. (1998). Bilateral filtering for gray and color images. In *The IEEE International Conference on Computer Vision (ICCV)*, pages 839–846.

[175] Toshev, E., Shi, J., and Daniilidis, K. (2007). Image matching via saliency region correspondences. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[176] Tran, S. and Davis, L. (2006). 3D Surface Reconstruction Using Graph Cuts with Surface Constraints. In *European Conference on Computer Vision (ECCV)*, pages 219–231.

[177] Triggs, B., McLauchlan, P. F., Hartley, R. I., and Fitzgibbon, A. W. (2000). Bundle adjustment - a modern synthesis. In *Proceedings of the International Workshop on Vision Algorithms: Theory and Practice*, The IEEE International Conference on Computer Vision (ICCV), pages 298–372.

[178] Tung, T., Nobuhara, S., and Matsuyama, T. (2009). Complete multi-view reconstruction of dynamic scenes from probabilistic fusion of narrow and wide baseline stereo. In *The IEEE International Conference on Computer Vision (ICCV)*, pages 1709–1716.

[179] Tuytelaars, T. and Mikolajczyk, K. (2008). Local invariant feature detectors: A survey. *Foundation and Trends in Computer Graphics and Visualization*, 3(3):177–280.

[180] Vedula, S., Rander, P., Collins, R., and Kanade, T. (2005). Three-dimensional scene flow. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 27(3):475–480.

[181] Veksler, O. (2008). Star shape prior for graph-cut image segmentation. In *European Conference on Computer Vision (ECCV)*, pages 454–467.

[182] Verdie, Y., Moo Yi, K., Verdie, Y., Fua, P., and Lepetit, V. (2015). Tilde: A temporally invariant learned detector. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5279–5288.

[183] Verdie, Y., Yi, K. M., Fua, P., and Lepetit, V. (2014). TILDE: A temporally invariant learned detector. *CoRR*, abs/1411.4568.

[184] Vicente, S., Kolmogorov, V., and Rother, C. (2008). Graph cut based image segmentation with connectivity priors. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8.

[185] Vlasic, D., Baran, I., Matusik, W., and Popović, J. (2008). Articulated mesh animation from multi-view silhouettes. *ACM Transactions on Graphics*, 27(4):97:1–97:9.

[186] Vogiatzis, G., Hernandez, C., Torr, P. H. S., and Cipolla, R. (2007). Multiview stereo via volumetric graph-cuts and occlusion robust photo-consistency. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 29(12).

[187] Wedel, A., Brox, T., Vaudrey, T., Rabe, C., Franke, U., and Cremers, D. (2011). Stereoscopic scene flow computation for 3d motion understanding. *International Journal of Computer Vision (IJCV)*, 95(1):29–51.

[188] Wei, L., Huang, Q., Ceylan, D., Vouga, E., and Li, H. (2016). Dense human body correspondences using convolutional networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[189] Weinzaepfel, P., Revaud, J., Harchaoui, Z., and Schmid, C. (2013). Deepflow: Large displacement optical flow with deep matching. In *The IEEE International Conference on Computer Vision (ICCV)*, pages 1385–1392.

[190] Wu, C. (2013). Towards linear-time incremental structure from motion. In *International Conference on 3D Vision (3DV)*, pages 127–134.

[191] Yang, R., Pollefeys, M., and Welch, G. (2003). Dealing with textureless regions and specular highlights - a progressive space carving scheme using a novel photo-consistency measure. In *The IEEE International Conference on Computer Vision (ICCV)*, pages 576–584.

[192] Yu, R., Russell, C., Campbell, N. D. F., and Agapito, L. (2015). Direct, dense, and deformable: Template-based non-rigid 3d reconstruction from RGB video. In *The IEEE International Conference on Computer Vision(ICCV)*, pages 918–926.

[193] Zach, C., Cohen, A., and Pollefeys, M. (2013). Joint 3d scene reconstruction and class segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[194] Zach, C., Pock, T., and Bischof, H. (2007). A duality based approach for realtime tv-l1 optical flow. In *In Annual Symposium German Association Pattern Recognition*, pages 214–223.

[195] Zanfir, A. and Sminchisescu, C. (2015). Large displacement 3d scene flow with occlusion reasoning. In *The IEEE International Conference on Computer Vision (ICCV)*.

[196] Zeng, G. and Quan, L. (2004). Silhouette extraction from multiple images of an unknown background. In *Asian Conference on Computer Vision (ACCV)*.

[197] Zhang, C., Li, Z., R.Cai, Chao, H., and Rui, Y. (2016). Joint multiview segmentation and localization of rgb-d images using depth-induced silhouette consistency. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[198] Zhang, D., Javed, O., and Shah, M. (2013). Video object segmentation through spatially accurate and temporally dense extraction of primary object regions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[199] Zhang, G., Jia, J., Hua, W., and Bao, H. (2011). Robust bilayer segmentation and motion/depth estimation with a handheld camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 33(3).

[200] Zhang, X., Wang, H., Smith, A. W. B., Xu, L., Lovell, B. C., and Yang, D. (2010). Corner detection based on gradient correlation matrices of planar curves. *Pattern Recognition*, 43(4):1207–1223.

[201] Zhang, Z. (1996). On the epipolar geometry between two images with lens distortion. In *International Conference on Pattern Recognition (ICPR)*, pages 407–411.

[202] Zhang, Z. (2000). A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 22(11):1330–1334.

[203] Zheng, E.and Ji, D., Dunn, E., and Frahm, J.-M. (2015). Sparse dynamic 3d reconstruction from unsynchronized videos. In *The IEEE International Conference on Computer Vision (ICCV)*.