



Video Fields: Fusing Multiple Surveillance Videos Into a Dynamic Virtual Environment

Ruofei Du, Sujal Bista, and Amitabh Varshney

www.Video-Fields.com

www.Augmentarium.com

Augmentarium | Department of Computer Science | UMIACS
University of Maryland, College Park

In Proceedings of the 21st Annual ACM SIGGRAPH Web3D Conference, 2016

Vocal: Sai Yuan; BGM: Ukulele by Bensound CC

Video Fields: Fusing Multiple Surveillance Videos into a Dynamic Virtual Environment

Ruofei Du, Sujal Bista, Amitabh Varshney

The Augmentarium | UMIACS | University of Maryland, College Park

{ruofei, suj, varshney} @ cs.umd.edu

www.VideoFields.com



Introduction

Surveillance Videos - Monitoring



UNIVERSITY OF MARYLAND • DEPARTMENT OF PUBLIC SAFETY



image courtesy: university of maryland, college park

Introduction

Surveillance Videos – Shopping Centers



image courtesy: www.icsc.org

Introduction

Surveillance Videos - Airports



image courtesy: wikipedia

Introduction

Surveillance Videos – Train stations



image courtesy: wikipedia

Introduction

Surveillance Videos - Campuses



image courtesy: university of maryland, college park

Introduction

Surveillance Videos - Conventional



UNIVERSITY OF MARYLAND • DEPARTMENT OF PUBLIC SAFETY



image courtesy: university of maryland, college park

Introduction

Surveillance Videos – Cognitive Burden



image courtesy: theimaginativeconservative.org

Introduction

Surveillance Videos – Fusing & Interpreting

UNIVERSITY OF MARYLAND • DEPARTMENT OF PUBLIC SAFETY



image courtesy: university of maryland, college park

Related Work

Fusing Multiple Static Photographs

Related Work

Fusing Multiple Static Photographs

Photo Tourism: Exploring Photo Collections in 3D

Noah Snavely
University of Washington

Steven M. Seitz
University of Washington

Richard Szeliski
Microsoft Research

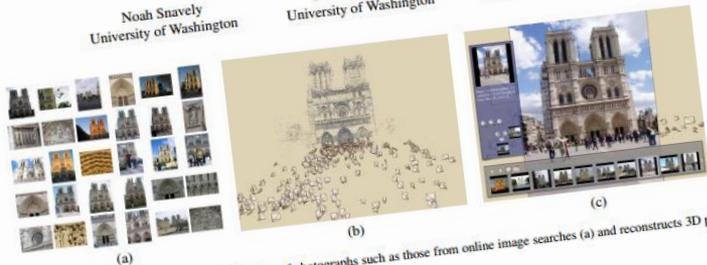


Figure 1: Our system takes unstructured collections of photographs such as those from online image searches (a) and reconstructs 3D points and viewpoints (b) to enable novel ways of browsing the photos (c).

Abstract

We present a system for interactively browsing and exploring large unstructured collections of photographs of a scene using a novel 3D interface. Our system consists of an image-based modeling front end that automatically computes the viewpoint of each photograph as well as a sparse 3D model of the scene and image to model correspondences. Our *photo explorer* uses image-based rendering techniques to smoothly transition between photographs, while also enabling full 3D navigation and exploration of the set of images and world geometry, along with auxiliary information such as overhead maps. Our system also makes it easy to construct photo tours of scenic or historic locations, and to annotate images. We demonstrate automatically transferred to other relevant images as well as images gathered from several large personal photo collections as well as images gathered from Internet photo sharing sites.

CR Categories: H.5.1 [Information Interfaces and Presentation]: Multimedia Information Systems—Artificial, augmented, and virtual realities 1.2.10 [Artificial Intelligence]: Vision and Scene Understanding—Modeling and recovery of physical attributes

Keywords: image-based rendering, image-based modeling, photo browsing, structure from motion

is that these approaches will one day allow virtual tourism of the world's interesting and important sites.

During this same time, digital photography, together with the Internet, have combined to enable sharing of photographs on a truly massive scale. For example, a Google image search on "Notre Dame Cathedral" returns over 15,000 photos, capturing the scene from myriad viewpoints, levels of detail, lighting conditions, seasons, decades, and so forth. Unfortunately, the proliferation of shared photographs has outpaced the technology for browsing such collections, as tools like Google (www.google.com) and Flickr (www.flickr.com) return pages and pages of thumbnails that the user must comb through.

In this paper, we present a system for browsing and organizing large photo collections of popular sites which exploits the common 3D geometry of the underlying scene. Our approach is based on computing, from the images themselves, the photographers' locations and orientations, along with a sparse 3D geometric representation of the scene, using a state-of-the-art image-based modeling system. Our system handles large collections of unorganized photographs taken by different cameras in widely different conditions. We show how the inferred camera and scene information enables the following capabilities:

- **Scene visualization.** Fly around popular world sites in 3D by morphing between photos.
- **Object-based photo browsing.** Show me more images that contain object or part of the scene.

Related Work

Fusing Multiple Static Photographs

Experience the new Photosynth 3D [preview it →](#)

Microsoft **Photosynth** | Tech Preview | Explore | About | My Synths | Search | New Account | Sign In | Create

Kyoto graveyard
Outsid3r 6/9/2016 17565 Views 82 PHOTOS 0 5



Capture your world in 3D
Shoot wraparound panoramas or full synths and share them with friends.

Gear up by checking out some of the best:

- Bridges
- Towers
- Collections
- Museums
- National Parks
- Markets
- Insects
- Forests
- Archaeology
- Aerial Views
- Beaches

Featured

Arch Angle 4
original 6/23/2015
★7 📍9 @GEOTAG
Panorama - 4.65 Megapixels 17757 Views

The Treasury, Petra
RohitJayakaran 7/4/2015
★8 📍11 @GEOTAG
Panorama - 4.82 Megapixels 8513 Views

Gothic Basin - Jun 2015
ericotm 6/14/2015
★4 📍6
Panorama - 152 Megapixels 5087 Views

VØRINGSFOSSEN WATERFALL, Norway - Handrgervidda plateau
RobertOyle 7/6/2015
★7 📍9 @GEOTAG
Panorama - 1.24 Gigapixels 6587 Views

Recommended

Experience the new dreamlike Photosynth 3D



Go ahead. Take it for a spin.

 Create your Synth

 About Photosynth

 Explore Synths

 Latest Synth News

 Discussion Forum

Figure 1: Our system and viewpoints (b)

Abstract

We present a system for creating 3D virtual worlds from unstructured 2D photographs. Our system takes a set of photographs and automatically generates a 3D scene that can be explored from any viewpoint. This is achieved by fusing multiple static photographs into a 3D scene. Our system is designed to be user-friendly and easy to use. It allows users to create their own 3D worlds from their own photographs. This is a significant step towards creating a more immersive and interactive virtual reality experience. Our system is designed to be user-friendly and easy to use. It allows users to create their own 3D worlds from their own photographs. This is a significant step towards creating a more immersive and interactive virtual reality experience.

Related Work

Fusing Multiple Static Photographs

Experience the new Photosynth



Figure 1: Our system and viewpoints (b)

Abstract

We present a system for acquiring and reconstructing temporally dynamic data. The system enables spatiotemporal 3D photography using commodity devices, assisted by their auxiliary sensors and network functionality. It engages users, making them active rather than passive participants in data acquisition.

Featured



Social Snapshot: A System for Temporally Coupled Social Photography

Robert Patro, Cheuk Yiu Ip, Sujal Bista, and Amitabh Varshney • University of Maryland, College Park

Since the invention of photography, taking pictures of people, places, and activities has become integral to our lives. In the past, only purposeful, precious moments were the primary subjects of photography. But technological advances have brought photography to our everyday lives in the form of compact cameras and even cell phone cameras.

Social Snapshot's Contributions

Social Snapshot's contributions fit naturally into two categories: technical and social. The technical contributions are improved algorithms and techniques that enhance our system's novelty and scalability. For example, Social Snapshot produces a textured and colored-mesh reconstruction from a loosely ordered photo collection, rather than the sparse or dense point reconstructions produced by related approaches. In addition, it features locally optimized mesh generation and viewing. Finally, it provides camera network capabilities to support synchronized capture of temporally dynamic data.

Social Snapshot actively acquires and reconstructs temporally dynamic data. The system enables spatiotemporal 3D photography using commodity devices, assisted by their auxiliary sensors and network functionality. It engages users, making them active rather than passive participants in data acquisition.

The next phase in the photography revolution, 3D photography, can bring users together to socialize and collaboratively take pictures in an entirely new way. However, transforming a photographic scene from 2D to 3D requires introducing multiple images of the same underlying geometry from different viewpoints. The reconstruction of 3D geometry from multiple overlapping images is the classic structure-from-motion (SfM) problem in computer vision. Typically, the instruments used to acquire photographs are tediously calibrated to produce precise measurements.

To simplify 3D photography, our Social Snapshot system performs active acquisition and reconstruction of temporally dynamic data. For a look at some of the previous research...

Related Work

Fusing Multiple Static Photographs

Experience the new Photosynth

Microsoft
Photosynth | Tech Preview

Kyoto graveyard
Outsid3r 6/9/2016 17565 Views

Figure 1: Our system and viewpoints (b)

Abstract

We present a system for capturing and visualizing unstructured 3D data. The system's front end that acquires data as well as its back end that performs correspondence techniques to enable full 3D world geometry maps. Our system is automatic and we demonstrate our system as images generated by the system.

CR Categories
Multimedia
Virtual Reality
Understanding

Keyword
browsing

Create your Synth About

Social Snapshot: A System for Temporally Coupled Social Photography

Robert Patro, Cheuk Yiu Ip

Since pictures become the only purpose of many subjects in our day lives in the cell phone

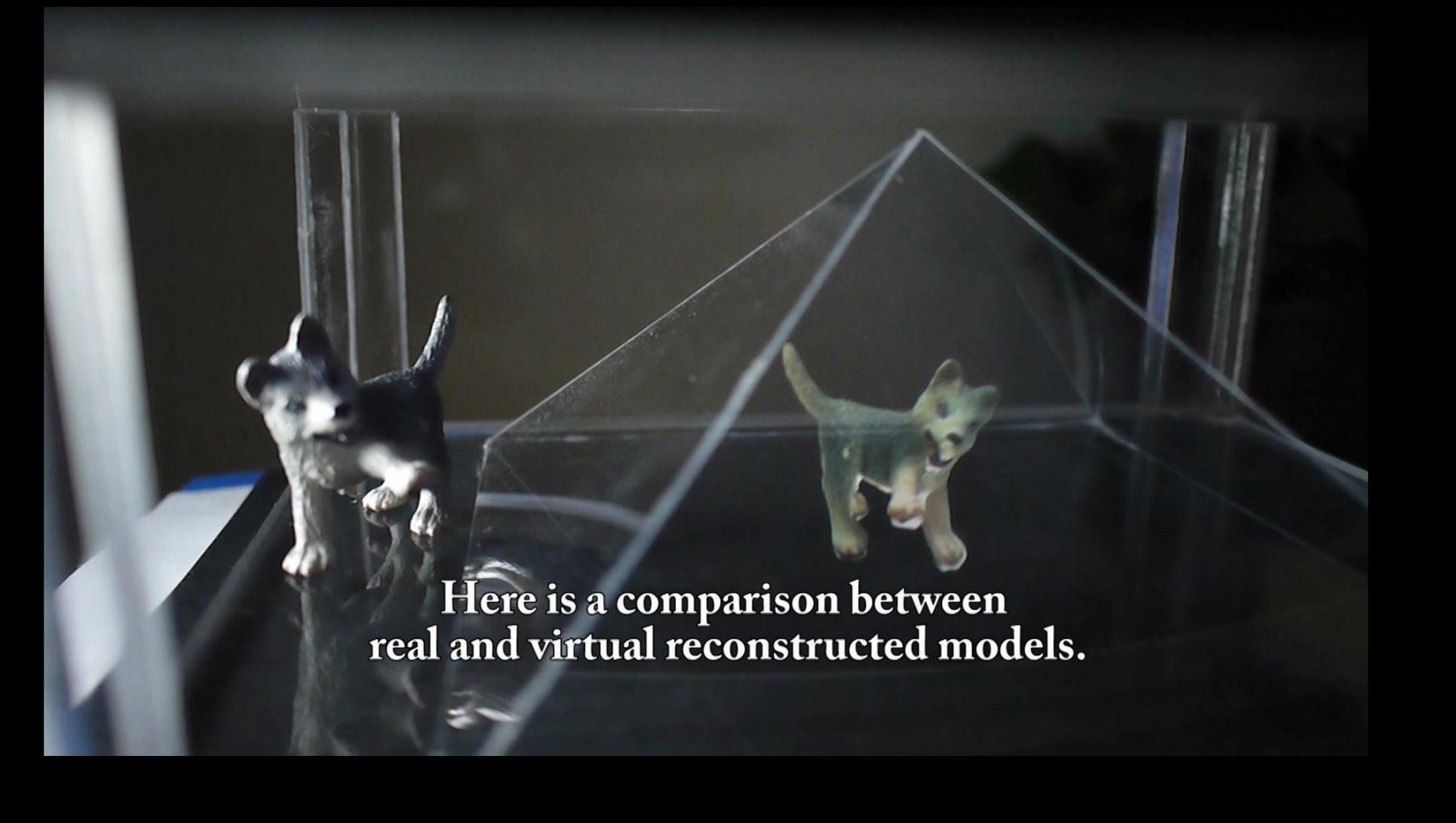
Social Snapshot actively acquires and reconstructs temporally dynamic data. The system enables spatiotemporal 3D photography using commodity devices, assisted by their auxiliary sensors and network functionality. It engages users, making them active rather than passive participants in data acquisition.

etry from multiple overlapping images is the classic structure-from-motion (SfM) problem in computer vision. Typically, the instruments used to acquire photographs are tediously calibrated to produce precise measurements. To simplify 3D photography, our Social Snapshot system performs active acquisition and reconstruction of dynamic scenes

endeavor, letting users capture dynamic scenes by synchronizing their photographs. It leverages social trends such as online media sharing and event organization to spur a novel data acquisition mode.

For a look at some of the previous research



The image shows two small dog figurines on a dark, reflective surface. On the left is a realistic, dark-colored dog figurine. On the right is a translucent, wireframe-like model of a dog, representing a virtual reconstruction. The background is dark with some vertical light streaks, possibly from a window or lighting setup. The text is centered at the bottom of the image.

**Here is a comparison between
real and virtual reconstructed models.**

Related Work

Fusing Multiple Dynamic Videos

Related Work

Fusing Multiple Dynamic Videos

RGB

Related Work

Fusing Multiple Dynamic Videos

RGB

RGBD

Related Work

Fusing Multiple Dynamic Videos

Dense 3D Motion Capture from Synchronized Video Streams

Yasutaka Furukawa¹
Department of Computer Science
and Beckman Institute
University of Illinois at Urbana-Champaign, USA¹

Jean Ponce^{2,1}
Willow Team
LIENS (CNRS/ENS/INRIA UMR 8548)
Ecole Normale Supérieure, Paris, France²

Abstract: This paper proposes a novel approach to non-rigid, markerless motion capture from synchronized video streams acquired by calibrated cameras. The instantaneous scene geometry of the observed scene is represented by a polyhedral mesh with fixed topology. The initial mesh is constructed in the first frame using the publicly available PMVS software for multi-view stereo [7]. Its deformation is captured by tracking its vertices over time, using two optimization processes at each frame: a local one using a rigid motion model in the neighborhood of each vertex, and a global one using a regularized nonrigid model for the whole mesh. Qualitative and quantitative experiments using seven real datasets show that our algorithm effectively handles complex nonrigid motions and severe occlusions.

1. Introduction

The most popular approach to motion capture today is to attach distinctive markers to the body and/or face of an actor, and track these markers in images acquired by multiple calibrated video cameras. The marker tracks are then matched, and triangulation is used to reconstruct the corresponding position and velocity information. The accuracy of any motion capture system is limited by the temporal and spatial resolution of the cameras. In the case of marker-based technology, it is also limited by the number of markers available: Although relatively few (say, 50) markers are available, they can be used to recover skeletal body configurations, and to recover skeletal body configurations, and to recover skeletal body configurations.

estimates of nonrigid motion. Markerless technology using special make-up is indeed emerging in the entertainment industry [15], and several approaches to local *scene flow* estimation have also been proposed to handle less constrained settings [4, 13, 16, 19, 23]. Typically, these methods do not fully exploit global spatio-temporal consistency constraints. They have been mostly limited to relatively simple and slow motions without much occlusion, and may be susceptible to error accumulation. We propose a different approach to motion capture as a 3D tracking problem and show that it effectively overcomes these limitations.

1.1. Related Work

Three-dimensional *active appearance models* (AAMs) are often used for facial motion capture [11, 14]. In this approach, parametric models encoding both facial shape and appearance are fitted to one or several image sequences. AAMs require an a priori parametric face model and are, by design, aimed at tracking relatively coarse facial motions rather than recovering fine surface detail and subtle expressions. *Active sensing* approaches to motion capture use a projected pattern to independently estimate the scene structure in each frame, then use optical flow and/or surface matches between adjacent frames to recover the three-dimensional motion field, or *scene flow* [10, 25]. Although qualitative results are impressive, these methods typically do not exploit the redundancy of the spatio-temporal information, and may be susceptible to error accumulation over time. Several *passive* approaches to scene flow computation, and may be susceptible to error accumulation over time. Several *passive* approaches to scene flow computation, and may be susceptible to error accumulation over time. Several *passive* approaches to scene flow computation, and may be susceptible to error accumulation over time.

Related Work

Fusing Multiple Dynamic Videos

To appear in the ACM SIGGRAPH conference proceedings

Performance Capture from Sparse Multi-view Video

Edilson de Aguiar* Carsten Stoll* Christian Theobalt† Naveed Ahmed* Hans-Peter Seidel* Sebastian Thrun†

*MPI Informatik, Saarbruecken, Germany
†Stanford University, Stanford, USA



Figure 1: A sequence of poses captured from eight video recordings of a capoeira turn kick. Our algorithm delivers spatio-temporally coherent geometry of the moving performer that captures both the time-varying surface detail as well as details in his motion very faithfully.

Abstract

This paper proposes a new marker-less approach to capturing human performances from multi-view video. Our algorithm can jointly reconstruct spatio-temporally coherent geometry, motion and textural surface appearance of actors that perform complex and rapid moves. Furthermore, since our algorithm is purely mesh-based and makes as few as possible prior assumptions about the type of subject being tracked, it can even capture performances of people wearing wide apparel, such as a dancer wearing a skirt. To serve this purpose our method efficiently and effectively combines the power of surface- and volume-based shape deformation techniques with a new mesh-based analysis-through-synthesis framework. This framework extracts motion constraints from video and makes the laser-scan of the tracked subject mimic the recorded performance. Also small-scale time-varying shape detail is re-performed. Our method delivers captured performance data at high level of detail, is highly versatile, and is applicable to many complex types of scenes that could not be handled by alternative marker-based or marker-free recording techniques.

CR Categories: 1.3.7 [Computer Graphics]: Three-Dimensional Graphics and Realism; 1.4.8 [Image Processing and Computer Vision]: Scene Analysis

1 Introduction

The recently released photo-realistic CGI movie Beowulf [Paramount 2007] provides an impressive foretaste of the way how many movies will be produced as well as displayed in the future. In contrast to previous animation movies, the goal was not the creation of a cartoon style appearance but a photo-realistic display of the virtual sets and actors. Today it still takes a tremendous effort to create authentic virtual doubles of real-world actors. It remains one of the biggest challenges to capture human performances, i.e. motion and possibly dynamic geometry of actors in the real world in order to map them onto virtual doubles. To measure body and facial motion, the studios resort to marker-based optical motion capture technology. Although this delivers data of high accuracy, it is still a stopgap. Marker-based motion capture requires a significant setup time, expects subjects to wear unnatural skin-tight clothing with optical beacons, and often makes necessary many hours of manual data cleanup. It therefore does not allow for what both actors and directors would actually prefer: To capture human performances *densely in space and time* - i.e. to be able to jointly capture accurate dynamic shape, motion and textural appearance of actors in arbitrary everyday apparel.

In this paper, we therefore propose a new marker-less dense performance capture technique. From only eight multi-view video recordings of a performer moving in his normal and even loose or wavy clothing, our algorithm is able to reconstruct his motion and his statio-temporally coherent time-varying geometry (i.e. geometry that captures even subtle deformation

University

Abstract: The rigid, marker streams acquire geometry of the triangular mesh constructed in the software for tracking motion processes one using a qualitative datasets show complex nonrigid

1. Intro

The mos to attach da actor, and multiple calibr matched, a sponding p of any mo spatial res based tech ers availa

marker-less scene reconstruc-

Related Work

Fusing Multiple Dynamic Videos

Edilson de Aguiar*



Figure 1: A sequence of frames showing a person in a dynamic pose, illustrating the concept of coherent geometry.

Abstract

This paper proposes a framework for joint reconstruction and texture synthesis of dynamic scenes. It is based on a novel marker-based method that makes the last step of the pipeline more robust to occlusions and complex nonrigid motions. This framework makes the last step of the pipeline more robust to occlusions and complex nonrigid motions. This framework makes the last step of the pipeline more robust to occlusions and complex nonrigid motions.

CR Categories: I.3.1 [Computer Graphics]: Scene

Probabilistic Deformable Surface Tracking From Multiple Videos

Cedric Cagniard¹, Edmond Boyer², and Slobodan Ilic¹

¹ Technische Universität München

² Grenoble Universités - INRIA Rhône-Alpes

{cagniard, slobodan.ilic}@in.tum.de, edmond.boyer@inrialpes.fr

Abstract. In this paper, we address the problem of tracking the temporal evolution of arbitrary shapes observed in multi-camera setups. This is motivated by the ever growing number of applications that require consistent shape information along temporal sequences. The approach we propose considers a temporal sequence of independently reconstructed surfaces and iteratively deforms a reference mesh to fit these observations. To effectively cope with outlying and missing geometry, we introduce a novel probabilistic mesh deformation framework. Using generic local rigidity priors and accounting for the uncertainty in the data acquisition process, this framework effectively handles missing data, relatively large reconstruction artefacts and multiple objects. Extensive experiments demonstrate the effectiveness and robustness of the method on various 4D datasets.

1 Introduction

Inferring shapes and their temporal evolutions from image data is a central problem in computer vision. Applications range from the visual restitution of live events to their analysis, recognition and even synthesis. The recovery of shapes using multiple images has received considerable attention over the last decade and several approaches can build precise static 3D models from geometric and photometric information, sometimes in real time. However, when applied to temporal sequences of moving objects, they provide temporally inconsistent shape models by treating each frame independently hence ignoring the dynamic nature of the observed event.

Most methods interested in tracking deformable surfaces in multi-camera systems deform a reference template mesh to fit observed geometric cues as well as possible at each time frame. These cues appear in the literature as photo-

Related Work

Fusing Multiple Dynamic Videos

Edilson de Aguiar*



Figure 1: A sequence of coherent geometry

Abstract

This paper proposes a method for jointly reconstructing the geometry and texture of a deforming object from multiple dynamic videos. The method is based on a template-less 4D reconstruction method that incrementally fuses highly-incomplete 3D observations of a deforming object, and generates a complete, temporally-coherent shape representation of the object. To this end, we design an online algorithm that alternatively registers new observations to the current model estimate and updates the model. We demonstrate the effectiveness of our approach at reconstructing non-rigidly moving objects from highly-incomplete measurements on both sequences of partial 3D point clouds and Kinect videos.

CR Categories: [Graphics and Visualization]: Scene

Deformable 3D Fusion: From Partial Dynamic 3D Observations to Complete 4D Models

Weipeng Xu^{1,2} Mathieu Salzmann^{2,3} Yongtian Wang¹ Yue Liu¹

¹Beijing Institute of Technology, China

²NICTA, Canberra, Australia

³CVLab, EPFL, Switzerland

{xuwp,wyt,liuyue}@bit.edu.cn, mathieu.salzmann@epfl.ch

Abstract

Capturing the 3D motion of dynamic, non-rigid objects has attracted significant attention in computer vision. Existing methods typically require either mostly complete 3D volumetric observations, or a shape template. In this paper, we introduce a template-less 4D reconstruction method that incrementally fuses highly-incomplete 3D observations of a deforming object, and generates a complete, temporally-coherent shape representation of the object. To this end, we design an online algorithm that alternatively registers new observations to the current model estimate and updates the model. We demonstrate the effectiveness of our approach at reconstructing non-rigidly moving objects from highly-incomplete measurements on both sequences of partial 3D point clouds and Kinect videos.

1. Introduction

In this paper, we introduce an approach to estimating a temporally-coherent 3D model of a non-rigid object given a dynamic sequence of highly-incomplete 3D observations of the object undergoing large deformations. Capturing the 3D motion of dynamic objects, or 4D reconstruction, has been a longstanding goal of computer vision. Ultimately, the resulting methods should yield a temporally-coherent shape representation of the observed deformable object.

Multiview reconstruction methods have been well-studied to address 4D reconstruction. While current methods achieve impressive results [12, 9, 6, 27, 32, 36], they typically require well-engineered and complete obser-

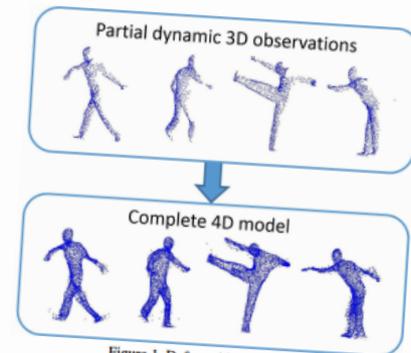


Figure 1. Deformable 3D fusion.

of the object, depicting, at best, half of its 3D surface.

In the case of rigid motion, several fusion techniques have been proposed to combine multiple partial 3D observations [15, 28, 41]. However, when it comes to capturing a dynamically deforming object, the literature remains very sparse. More specifically, most existing methods [17, 7, 18, 35, 39, 42] rely on a pre-processing stage, where the object undergoes (quasi-)rigid motion, to acquire a complete 3D template of the object, which will then be deformed to match new non-rigid data.

By contrast, in this paper, we introduce a template-less 4D reconstruction method that

Related Work

Fusing Multiple Dynamic Videos

Edilson de Aguiar*



Figure 1: A sequence of frames showing coherent geometry.

Abstract

This paper proposes a novel method for joint pose estimation and texture reconstruction of articulated objects using a single depth camera. Qualitative and quantitative evaluations on publicly available datasets show that our method achieves state-of-the-art performance. It is based on a novel shape adaptation algorithm based on a probabilistic model automatically captures the shape of the subjects during the dynamic pose estimation process. Experiments show that our shape estimation method achieves comparable accuracy with state-of-the-art methods, yet requires neither parametric model nor extra calibration procedure.

CR Categories: Graphics and Visualization; Scene Analysis

Real-time Simultaneous Pose and Shape Estimation for Articulated Objects Using a Single Depth Camera

Mao Ye
mao.ye@uky.edu

Ruigang Yang
ryang@cs.uky.edu

University of Kentucky
Lexington, Kentucky, USA, 40506

Abstract

In this paper we present a novel real-time algorithm for simultaneous pose and shape estimation for articulated objects, such as human beings and animals. The key of our pose estimation component is to embed the articulated deformation model with exponential-maps-based parametrization into a Gaussian Mixture Model. Benefiting from the probabilistic measurement model, our algorithm requires no explicit point correspondences as opposed to most existing methods. Consequently, our approach is less sensitive to local minimum and well handles fast and complex motions. Extensive evaluations on publicly available datasets demonstrate that our method outperforms most state-of-art pose estimation algorithms with large margin, especially in the case of challenging motions. Moreover, our novel shape adaptation algorithm based on the same probabilistic model automatically captures the shape of the subjects during the dynamic pose estimation process. Experiments show that our shape estimation method achieves comparable accuracy with state-of-the-art methods, yet requires neither parametric model nor extra calibration procedure.

1. Introduction

The topic of pose estimation for articulated objects, in particular human pose estimation [17, 22], has been actively studied by the computer vision community for decades. In recent years, due to the increasing popularity of depth sensors, studies have been conducted to capture the pose of articulated objects using one or more such depth sensors (detailed in Sec. 2). Despite of the substantial progress that have been achieved, there are still various limitations. Discriminative approaches [23, 25, 21] in general are capable of handling large body shape variations. However, they have been shown that most of these methods require either parametric model nor extra calibration procedure.



Figure 1. Our novel algorithm effectively estimates the pose of articulated objects using one single depth camera, such as human and dogs, even with challenging cases.

When a template model is used, as in generative or hybrid approaches, the consistency of body shape (limb lengths and girths) between the model and the subject is critical for accurate pose estimation. Most existing approaches either require given shapes [11, 29], small variations from the template [12], or specific initialization [27, 13]. Apparently, these requirements limit the applicability of these methods in home environments. To overcome the limitations mentioned above, we propose a novel (generative) articulated pose estimation algorithm that **does not require explicit point correspondences and captures the subject's shape automatically during the pose estimation process.** Our algorithm relates the observed data with our template using Gaussian Mixture Model (GMM), without explicitly building point correspondences. The pose is then estimated through probability density estimation under articulated deformation model parameterized with exponential maps [2]. Consequently, the algorithm is **less sensitive to local minimum and well accommodates fast and complex motions.** In addition, we develop a novel shape estimation algorithm based on a probabilistic framework. It automatically captures the shape of the subjects during the dynamic pose estimation process. Experiments show that our shape estimation method achieves comparable accuracy with state-of-the-art methods, yet requires neither parametric model nor extra calibration procedure.

Related Work

Fusing Multiple Dynamic Videos

Edilson de Aguiar*



Figure 1: A scene with coherent geometry

Abstract

This paper proposes a method for jointly reconstructing and texturing surfaces of dynamic scenes with a high level of detail. The method is based on a novel multi-view stereo (MVS) algorithm that makes use of a novel point cloud registration method to handle non-rigid motion. The method is able to reconstruct and texture surfaces of dynamic scenes with a high level of detail. The method is based on a novel multi-view stereo (MVS) algorithm that makes use of a novel point cloud registration method to handle non-rigid motion.

CR Categories: I.3.3 [Computer Graphics]: Picture/Image Generation; I.5.1 [Information Interfaces and Presentation]: Multimedia Information Systems

In this paper, we present a system for accurate real-time mapping of complex and moving indoor scenes in variable lighting conditions, using only a single low-cost depth camera and commodity graphics hardware. We fuse all of the depth data streamed from a Kinect sensor into a single global implicit surface model of the observed scene in real-time. The current sensor pose is simultaneously obtained by tracking the live depth frame relative to the global model using a coarse-to-fine iterative closest point (ICP) algorithm, which uses all of the observed depth data available. We demonstrate the advantages of tracking against the growing full surface model compared with frame-to-frame tracking, obtaining tracking and mapping results in constant time within room sized scenes with limited drift and high accuracy. We also show both qualitative and quantitative results relating to various aspects of our tracking and mapping system. Modelling of natural scenes, in real-time with only commodity sensor and GPU hardware, promises an exciting step forward in augmented reality (AR), in particular, it allows dense surfaces to be reconstructed in real-time, with a level of detail and robustness beyond any solution yet presented using passive computer vision.

1. Introduction

The particular challenge of real-time dense surface reconstruction is that the scene geometry is often highly complex and dynamic. This makes the task of reconstructing the scene geometry in real-time a significant challenge. In this paper, we present a system for accurate real-time mapping of complex and moving indoor scenes in variable lighting conditions, using only a single low-cost depth camera and commodity graphics hardware.

Keywords: Real-Time, Dense Reconstruction, Tracking, GPU, SLAM, Depth Cameras, Volumetric Representation, AR

Index Terms: I.3.3 [Computer Graphics]: Picture/Image Generation - Digitizing and Scanning; I.4.8 [Image Processing and Computer Vision]: Scene Analysis - Tracking, Surface Fitting; H.5.1 [Information Interfaces and Presentation]: Multimedia Information Systems - Artificial, augmented, and virtual realities

*This work was performed at Microsoft Research

KinectFusion: Real-Time Dense Surface Mapping and Tracking*

Richard A. Newcombe
Imperial College London

Shahram Izadi
Microsoft Research

Andrew J. Davison
Imperial College London

Pushmeet Kohli
Microsoft Research

Otmar Hilliges
Microsoft Research

Jamie Shotton
Microsoft Research

David Molyneux
Microsoft Research
Lancaster University
Steve Hodges
Microsoft Research

David Kim
Microsoft Research
Newcastle University
Andrew Fitzgibbon
Microsoft Research

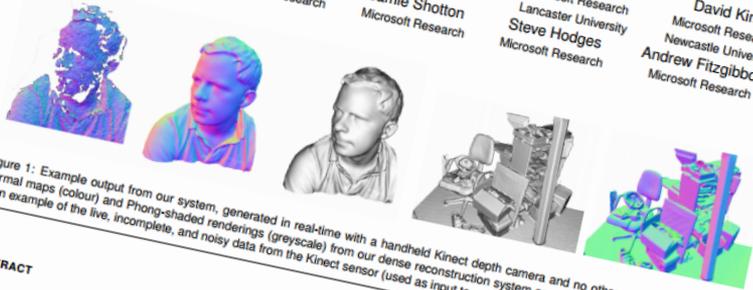


Figure 1: Example output from our system, generated in real-time with a handheld Kinect depth camera and no other sensing infrastructure. Normal maps (colour) and Phong-shaded renderings (grayscale) from our dense reconstruction system are shown. On the left for comparison is an example of the live, incomplete, and noisy data from the Kinect sensor (used as input to our system).

ABSTRACT

We present a system for accurate real-time mapping of complex and moving indoor scenes in variable lighting conditions, using only a single low-cost depth camera and commodity graphics hardware. We fuse all of the depth data streamed from a Kinect sensor into a single global implicit surface model of the observed scene in real-time. The current sensor pose is simultaneously obtained by tracking the live depth frame relative to the global model using a coarse-to-fine iterative closest point (ICP) algorithm, which uses all of the observed depth data available. We demonstrate the advantages of tracking against the growing full surface model compared with frame-to-frame tracking, obtaining tracking and mapping results in constant time within room sized scenes with limited drift and high accuracy. We also show both qualitative and quantitative results relating to various aspects of our tracking and mapping system. Modelling of natural scenes, in real-time with only commodity sensor and GPU hardware, promises an exciting step forward in augmented reality (AR), in particular, it allows dense surfaces to be reconstructed in real-time, with a level of detail and robustness beyond any solution yet presented using passive computer vision.

1 INTRODUCTION

Real-time infrastructure-free tracking of a handheld camera whilst simultaneously mapping the physical scene in high-detail promises new possibilities for augmented and mixed reality applications. In computer vision, research on structure from motion (SfM) and multi-view stereo (MVS) has produced many compelling results, in particular accurate camera tracking and sparse reconstructions (e.g. [10]), and increasingly reconstruction of dense surfaces (e.g. [24]). However, much of this work was not motivated by real-time applications. Research on simultaneous scene reconstruction focused more on real-time localisation and mapping (SLAM) based construction based on the input of a single commodity sensor—a monocular RGB camera. Such 'monocular SLAM' systems such as MonoSLAM [8] and the more accurate Parallel Tracking and Mapping (PTAM) system [17] allow researchers to investigate flexible infrastructure- and marker-free AR applications. But while these systems perform real-time mapping, they were optimised for efficient camera tracking, with the sparse point cloud models they produce enabling only rudimentary scene reconstruction. In the past year, systems have begun to emerge that combine PTAM's handheld camera tracking capability with dense surface occlusion prediction and surface interaction [19, 26]. Most recently in this line of research, iterative image alignment based dense surface reconstructions has also been used to replace point feature dense surface tracking [20]. While this work is very promising, dense surface reconstruction in real-time with a single commodity sensor and monocular stereo is still a significant challenge.

Related Work

Fusing Multiple Dynamic Videos

Edilson de Aguiar*



Figure 1: A sequence of coherent geometry

Abstract

This paper proposes a method for jointly reconstructing and texturing surface geometry of a person in a sequence of dynamic scenes. The method is based on a multi-view geometry approach and makes use of a novel multi-view geometry based method for recovering the geometry of a person in a sequence of dynamic scenes. The method is based on a multi-view geometry approach and makes use of a novel multi-view geometry based method for recovering the geometry of a person in a sequence of dynamic scenes.

CR Categories: [Computer Graphics and Animation]: Scene

University

Abstract: This paper proposes a method for jointly reconstructing and texturing surface geometry of a person in a sequence of dynamic scenes. The method is based on a multi-view geometry approach and makes use of a novel multi-view geometry based method for recovering the geometry of a person in a sequence of dynamic scenes.

1. Introduction

The motivation for this work is to provide a method for recovering the geometry of a person in a sequence of dynamic scenes. The method is based on a multi-view geometry approach and makes use of a novel multi-view geometry based method for recovering the geometry of a person in a sequence of dynamic scenes.

Capturing the geometry of a person in a sequence of dynamic scenes is a challenging task. This paper proposes a method for jointly reconstructing and texturing surface geometry of a person in a sequence of dynamic scenes. The method is based on a multi-view geometry approach and makes use of a novel multi-view geometry based method for recovering the geometry of a person in a sequence of dynamic scenes.

1. Introduction

The motivation for this work is to provide a method for recovering the geometry of a person in a sequence of dynamic scenes. The method is based on a multi-view geometry approach and makes use of a novel multi-view geometry based method for recovering the geometry of a person in a sequence of dynamic scenes.

In this paper, we present a method for recovering the geometry of a person in a sequence of dynamic scenes. The method is based on a multi-view geometry approach and makes use of a novel multi-view geometry based method for recovering the geometry of a person in a sequence of dynamic scenes.

1. Introduction

The motivation for this work is to provide a method for recovering the geometry of a person in a sequence of dynamic scenes. The method is based on a multi-view geometry approach and makes use of a novel multi-view geometry based method for recovering the geometry of a person in a sequence of dynamic scenes.

ABSTRACT

We present a system for recovering the geometry of a person in a sequence of dynamic scenes. The method is based on a multi-view geometry approach and makes use of a novel multi-view geometry based method for recovering the geometry of a person in a sequence of dynamic scenes.

Keywords: SLAM, Depth

Index Terms: Digital Computer Vision (Informatics)

*This work was performed at Microsoft

KinectFusion

Richard A. Newcombe
Imperial College London

Andrew J. Davison
Imperial College London



Figure 1: Example of our Normal maps (colour) is an example of the live

We present the first dense SLAM system capable of reconstructing non-rigidly deforming scenes in real-time, by fusing together RGBD scans captured from commodity sensors. Our DynamicFusion approach reconstructs scene geometry whilst simultaneously estimating a dense volumetric 6D motion field that warps the estimated geometry into a live frame. Like KinectFusion, our system produces increasingly denoised, detailed, and complete reconstructions as more measurements are fused, and displays the updated model in real time. Because we do not require a template or other prior scene model, the approach is applicable to a wide range of moving objects and scenes.

3D scanning traditionally involves separate capture and off-line processing phases, requiring very careful planning of the capture to make sure that every surface is covered. In practice, it's very difficult to avoid holes, requiring several iterations of capture, reconstruction, identifying holes, and recapturing missing regions to ensure a complete model. Real-time 3D reconstruction systems like KinectFusion [18, 10] represent a major advance, by providing users the ability to instantly see the reconstruction and identify regions that remain to be scanned. KinectFusion spurred a flurry of follow up research aimed at improving the tracking [9, 32] and expanding its spatial mapping capabilities to larger environments [22, 19, 34, 31, 9].

However, as with all traditional SLAM and dense reconstruction systems, the most basic assumption behind KinectFusion is that the observed scene is largely static. The core question we tackle in this paper is: How can we generalise KinectFusion to reconstruct dynamic scenes?

DynamicFusion: Reconstruction and Tracking of Non-rigid Scenes in Real-Time

Richard A. Newcombe
newcombe@cs.washington.edu

Dieter Fox
fox@cs.washington.edu
University of Washington, Seattle

Steven M. Seitz
seitz@cs.washington.edu

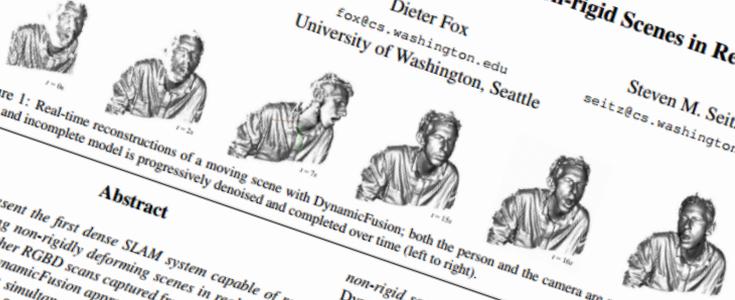


Figure 1: Real-time reconstructions of a moving scene with DynamicFusion; both the person and the camera are moving. The initially noisy and incomplete model is progressively denoised and completed over time (left to right).

Abstract

We present the first dense SLAM system capable of reconstructing non-rigidly deforming scenes in real-time, by fusing together RGBD scans captured from commodity sensors. Our DynamicFusion approach reconstructs scene geometry whilst simultaneously estimating a dense volumetric 6D motion field that warps the estimated geometry into a live frame. Like KinectFusion, our system produces increasingly denoised, detailed, and complete reconstructions as more measurements are fused, and displays the updated model in real time. Because we do not require a template or other prior scene model, the approach is applicable to a wide range of moving objects and scenes.

DynamicFusion, an approach based on solving for a volumetric flow field that transforms the state of the scene at each time instant into a fixed, canonical frame. In the case of a moving person, for example, this transformation undoes the pose of the first frame, warping each body configuration dates is effectively rigid, and standard KinectFusion updates can be used to obtain a high quality, denoised reconstruction. This progressively denoised reconstruction can then be transformed back into the live frame using the inverse map; each point in the canonical frame is transformed to its location in the live frame (see Figure 1).

Defining a canonical "rigid" space for a dynamically moving scene is not straightforward. A key contribution of our work is an approach for non-rigid transformation and fusion that retains the optimality properties of volumetric scan fusion [5], developed originally for rigid scenes. The main insight is that undoing the scene motion to enable fusion of all observations into a single fixed frame can be achieved efficiently by computing the inverse map alone. Under this transformation, each canonical point projects along a line of sight in the live camera frame. Since the optimality arguments of [5] (developed for rigid scenes) depend only on lines of sight, we can generalize their validity results to the non-rigid case.

Our second key contribution is a method for efficiently warping the non-rigid scene into the canonical frame, even a relatively large scene.

Related Work

Fusing Multiple Dynamic Videos

Edilson de Aguiar*



Figure 1: A sequence of coherent geometry

Abstract

This paper proposes a method for joint reconstruction and texture synthesis of dynamic scenes. We focus on capturing the geometry of a scene in real-time. Our method is based on a novel implicit representation of the scene geometry. We use a neural network to learn a model of the scene geometry from a sequence of images. The model is then used to reconstruct the scene geometry in real-time. Our method is able to handle complex scenes with dynamic objects and is robust to noise and occlusion.

CR Categories: Computer graphics, Animation, Virtual reality, Performance.

Capturing dynamic scenes has attracted increasing attention. We introduce a novel method for capturing dynamic scenes. Our method is based on a novel implicit representation of the scene geometry. We use a neural network to learn a model of the scene geometry from a sequence of images. The model is then used to reconstruct the scene geometry in real-time. Our method is able to handle complex scenes with dynamic objects and is robust to noise and occlusion.

1. Introduction

In this paper, we propose a method for capturing dynamic scenes. Our method is based on a novel implicit representation of the scene geometry. We use a neural network to learn a model of the scene geometry from a sequence of images. The model is then used to reconstruct the scene geometry in real-time. Our method is able to handle complex scenes with dynamic objects and is robust to noise and occlusion.

In this paper, we propose a method for capturing dynamic scenes. Our method is based on a novel implicit representation of the scene geometry. We use a neural network to learn a model of the scene geometry from a sequence of images. The model is then used to reconstruct the scene geometry in real-time. Our method is able to handle complex scenes with dynamic objects and is robust to noise and occlusion.

1. Introduction

The particular challenge of capturing dynamic scenes is that the scene is constantly changing. This makes it difficult to capture the scene geometry in real-time. Our method is based on a novel implicit representation of the scene geometry. We use a neural network to learn a model of the scene geometry from a sequence of images. The model is then used to reconstruct the scene geometry in real-time. Our method is able to handle complex scenes with dynamic objects and is robust to noise and occlusion.

Keywords: SLAM, Deep Learning, Computer Graphics, Animation, Virtual Reality, Performance.

*This work



Figure 1: Example of our normal maps (colour) is an example of the

ABSTRACT

We present a system for capturing dynamic scenes in real-time. Our method is based on a novel implicit representation of the scene geometry. We use a neural network to learn a model of the scene geometry from a sequence of images. The model is then used to reconstruct the scene geometry in real-time. Our method is able to handle complex scenes with dynamic objects and is robust to noise and occlusion.

KinectFusion

Richard A. Newcombe
Imperial College London

Andrew J. Davison
Imperial College London

DynamicFusion: Reconstruction and Tracking of Non-rigid Scenes in Real-Time

Richard A. Newcombe
newcombe@cs.washington.edu

Dieter Fox
fox@cs.washington.edu
University of Washington, Seattle

Steven M. Seitz
seitz@cs.washington.edu



Figure 1: noisy a

Fusion4D: Real-time Performance Capture of Challenging Scenes

Mingsong Dou
Adarsh Kowdle*

Sameh Khamis
Sergio Orts Escolano*
Pushmeet Kohli

Yury Degtyarev
Christoph Rhemann*
Vladimir Tankovich
Microsoft Research

Philip Davidson*
David Kim
Shahram Izadi!
Sean Ryan Fanello*
Jonathan Taylor



Figure 1: We present a new method for real-time high quality 4D (i.e. spatio-temporally coherent) performance capture, allowing for incremental nonrigid reconstruction from noisy input from multiple RGBD cameras. Our system demonstrates unprecedented reconstructions of challenging nonrigid sequences, at real-time rates, including robust handling of large frame-to-frame motions and topology changes.

Abstract

We contribute a new pipeline for live multi-view performance capture generating temporally coherent high-quality reconstructions in real-time. Our algorithm supports both incremental reconstruction and improving the surface estimation over time, as well as parameterizing the nonrigid scene motion. Our approach is highly robust to both large frame-to-frame motion and topology changes, allowing us to reconstruct extremely challenging scenes. We demonstrate advantages over related real-time techniques that either deform an online generated template or continually fuse depth data nonrigidly into a single reference model. Finally, we show geometric reconstruction results on par with offline methods which require orders of magnitude more processing time and many more RGBD cameras.

Keywords: nonrigid, real-time, 4D reconstruction, multi-view

1 Introduction

Whilst real-time 3D reconstruction with the ubiquity of modern sensors and processing power, capturing dynamic scenes in real-time remains a significant challenge. This is due to the inherent complexity of dynamic scenes, which are constantly changing and often contain non-rigid objects. Our method is based on a novel implicit representation of the scene geometry. We use a neural network to learn a model of the scene geometry from a sequence of images. The model is then used to reconstruct the scene geometry in real-time. Our method is able to handle complex scenes with dynamic objects and is robust to noise and occlusion.

Despite these challenges, there is clear value in reconstructing non-rigid motion and surface deformations in real-time. In particular, performance capture, where multiple cameras are used to reconstruct human motion and shape, and even object interactions, is currently constrained to offline processing. When if this processing could happen live in real-time, performance is happening could happen live in real-time. This can be achieved by remotely capturing the scene in real-time. Our method is based on a novel implicit representation of the scene geometry. We use a neural network to learn a model of the scene geometry from a sequence of images. The model is then used to reconstruct the scene geometry in real-time. Our method is able to handle complex scenes with dynamic objects and is robust to noise and occlusion.

Related Work

Fusing Multiple Dynamic Videos

Edilson de Aguiar*



Figure 1: A sequence of coherent geometry

Abstract

This paper proposes a system for capturing high quality 3D surface geometry of complex type marker-based motion capture sequences. We demonstrate that our system can capture high quality 3D surface geometry of complex type marker-based motion capture sequences. We demonstrate that our system can capture high quality 3D surface geometry of complex type marker-based motion capture sequences.

CR Categories: [Human Factors] [Motion Capture] [Computer Graphics] [Motion Capture]

Capturing high quality 3D surface geometry of complex type marker-based motion capture sequences. We demonstrate that our system can capture high quality 3D surface geometry of complex type marker-based motion capture sequences.

1. Introduction

The particular challenge of capturing high quality 3D surface geometry of complex type marker-based motion capture sequences. We demonstrate that our system can capture high quality 3D surface geometry of complex type marker-based motion capture sequences.

In this paper, we present a system for capturing high quality 3D surface geometry of complex type marker-based motion capture sequences. We demonstrate that our system can capture high quality 3D surface geometry of complex type marker-based motion capture sequences.

Keywords: SLAM, Depth

Index Terms: Motion Capture, 3D Reconstruction, SLAM, Depth

*This work

KinectFusion

Richard A. Newcombe
Imperial College London

Andrew J. Davison
Imperial College London



Figure 1: Example of our normal maps (colour) is an example of the

ABSTRACT

We present a system for capturing high quality 3D surface geometry of complex type marker-based motion capture sequences. We demonstrate that our system can capture high quality 3D surface geometry of complex type marker-based motion capture sequences.

1 Introduction

Whilst real-time 3D reconstruction, multi-view stereo, and depth-based SLAM are becoming increasingly important in a wide range of applications, the current state-of-the-art is limited by the high processing time and memory requirements of these methods.

DynamicFusion: Reconstruction and Tracking of Non-rigid Scenes in Real-Time

Richard A. Newcombe
newcombe@cs.washington.edu

Dieter Fox
fox@cs.washington.edu
University of Washington, Seattle

Steven M. Seitz
seitz@cs.washington.edu



Figure 1: noisy a

Fusion4D: Real-time Performance Capture of Challenging Scenes

Mingsong Dou
Adarsh Kowdle*

Sameh Khamis
Sergio Orts Escolano*
Pushmeet Kohli

Yury Degtyarev
Christoph Rhemann*
Vladimir Tankovich
Microsoft Research

Philip Davidson*
David Kim
Shahram Izadi!
Sean Ryan Fanello*
Jonathan Taylor



Figure 1: We present a new method for real-time high quality 4D (i.e. spatio-temporally coherent) performance capture of challenging nonrigid sequences, at real-time rates, including robust handling of large frame-to-frame motions and

Abstract

We contribute a new pipeline for live multi-view performance capture generating temporally coherent high-quality reconstructions in real-time. Our algorithm supports both incremental reconstruction, improving the surface estimation over time, as well as parameterizing the nonrigid scene motion and topology changes, allowing us to reconstruct extremely challenging scenes. We demonstrate advantages over related real-time techniques that either deform an online generated template or continually fuse depth data nonrigidly into a single reference model. Finally, we show geometric reconstruction results on par with offline methods which require orders of magnitude more processing time and many more RGBD cameras.

Keywords: nonrigid, real-time, 4D reconstruction, multi-view

1 Introduction

Whilst real-time 3D reconstruction, multi-view stereo, and depth-based SLAM are becoming increasingly important in a wide range of applications, the current state-of-the-art is limited by the high processing time and memory requirements of these methods.

Despite these challenges, there has been significant progress in real-time performance capture of human motion and surface deformation. This progress has been constrained to offline processing, where multiple cameras are used to capture performance in real-time before processing. When if this processing is done in real-time, performance is hampered by the need to reconstruct scene geometry in real-time. This is a challenging task, as it requires the ability to watch a scene in real-time and reconstruct it in real-time. This is a challenging task, as it requires the ability to watch a scene in real-time and reconstruct it in real-time.

Related Work

Fusing Multiple Dynamic Videos



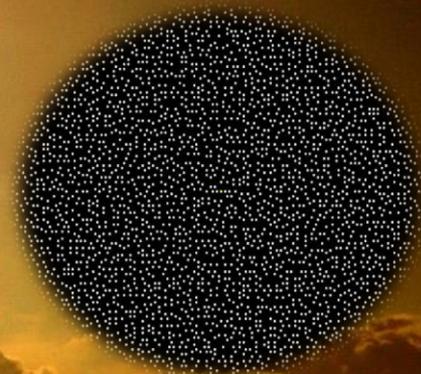
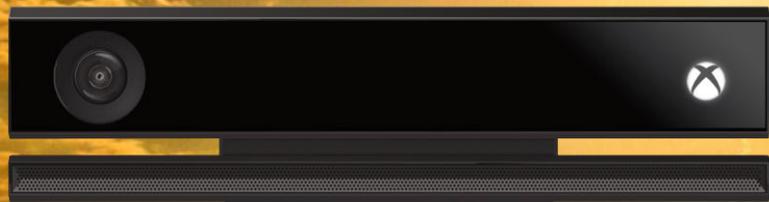
Related Work

Fusing Multiple Dynamic Videos



Related Work

Fusing Multiple Dynamic Videos



Our Approach?

Video Fields

Video Fields



*Fusing multiple RGB surveillance
videos without feature matching
algorithms nor calibration patterns*

Introduction

Video Field



Introduction

Surveillance Videos



They monitors a variety of activities in shopping centers, airports, train stations, and university campuses.

Introduction

Video Field



Video-Fields

Overview



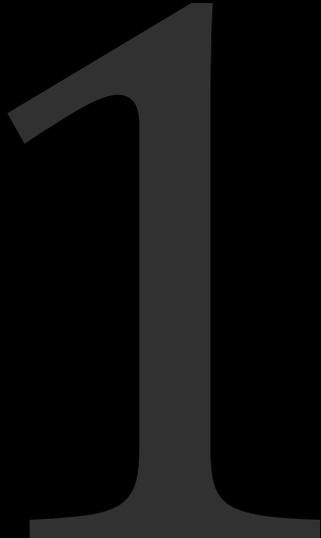
ground	Static
bd.visible	<input checked="" type="checkbox"/>
bd.transparent	<input checked="" type="checkbox"/>
pm.rotate	<input type="checkbox"/>
pm.attract	<input type="checkbox"/>
test0	<input type="checkbox"/>
test1	<input type="checkbox"/>
test2	<input checked="" type="checkbox"/>
filter	Gaussian
threshold0	<input type="range" value="0.08"/> 0.08
threshold1	<input type="range" value="0.12"/> 0.12
threshold2	<input type="range" value="0"/> 0
video	<input type="range" value="17"/> 17
Close Controls	

In this paper we introduce, Video Fields, a novel web-based interactive system to create, calibrate, and render ...

Conception, architecting & implementation

Video Fields

A mixed reality system that fuses multiple surveillance videos into an immersive virtual environment,



Integrating automatic segmentation of
moving entities

2

Video Fields Rendering

Real-time fragment shader processing

Two algorithms to fuse multiple videos

3

Early & deferred pruning

These methods use voxels and meshes respectively to render moving entities in the video fields

Achieving cross-platform compatibility by

4

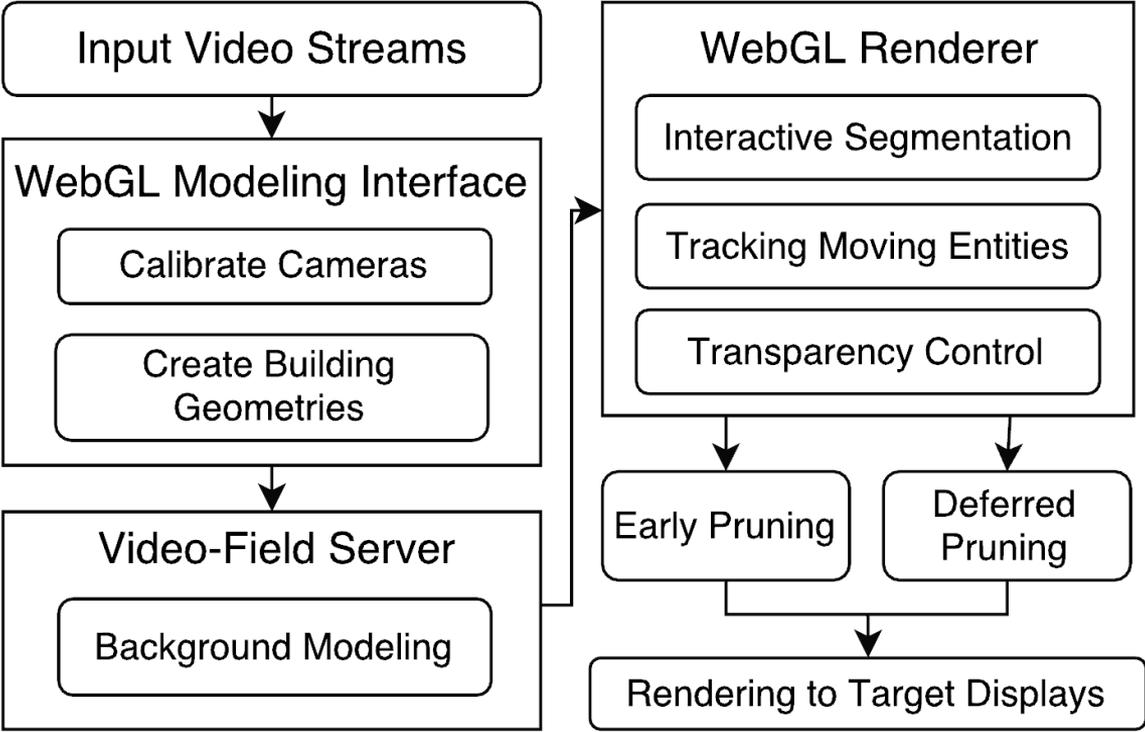
WebGL + Three.js

smartphones, tablets, desktop, high-resolution
large-area wide field of view tiled display walls, as
well as head-mounted displays.

System Overview

Architecture

Video Fields Flowchart

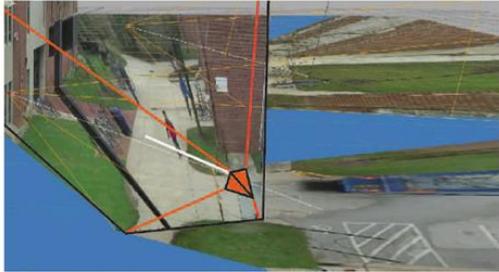


Architecture

Video Fields Flowchart



surveillance video streams



calibration of camera world matrices



static 3D models and satellite image

Video Fields
Mapping



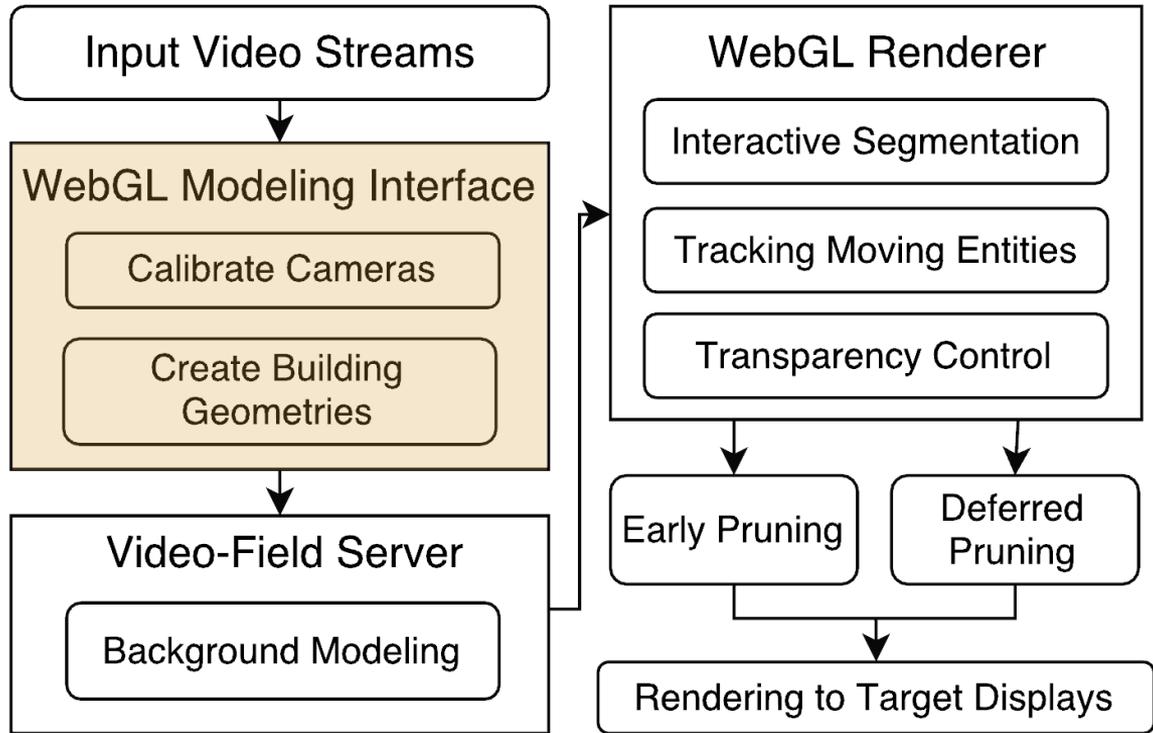
dynamic virtual environment



automatic segmentation and view-dependent rendering

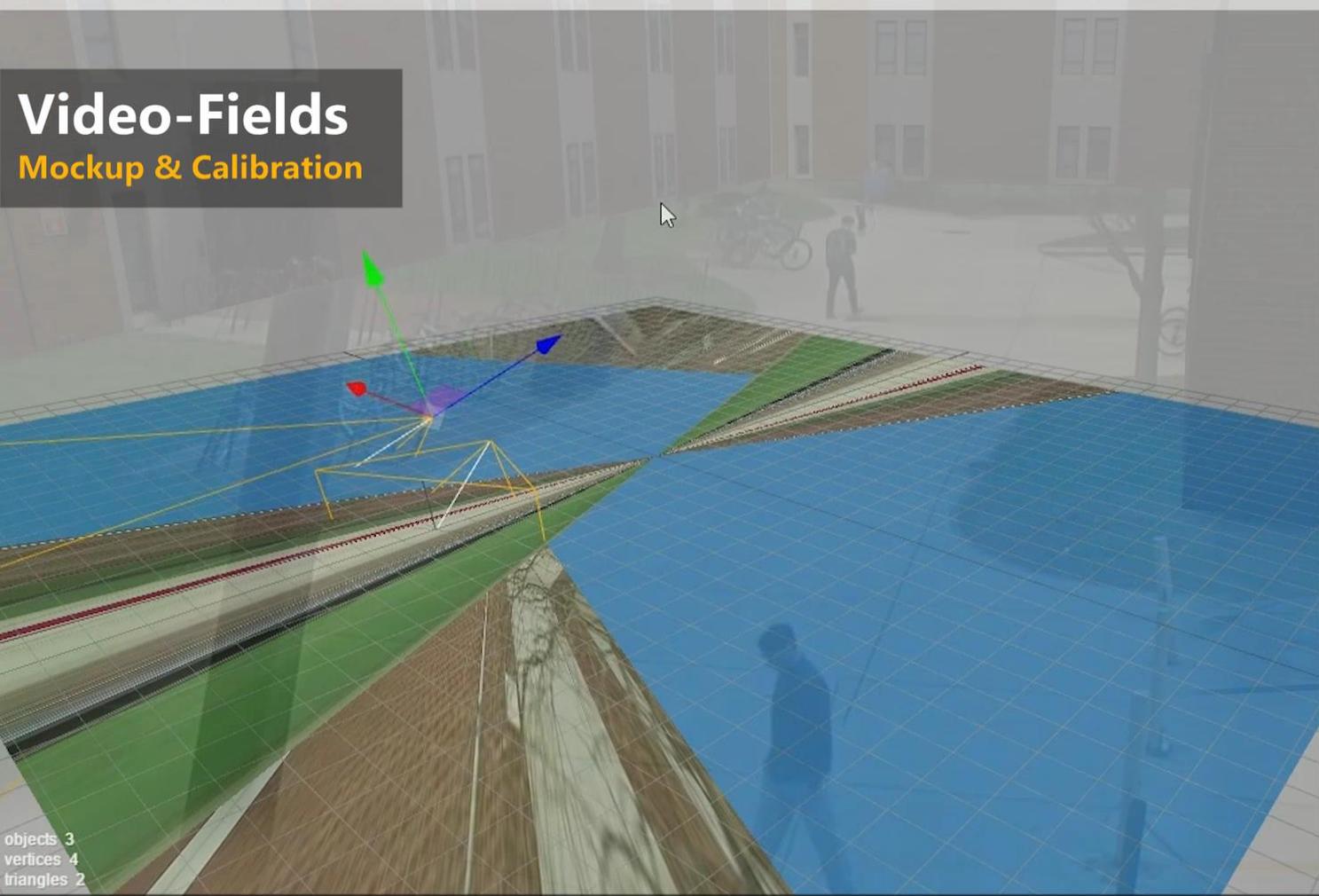
Architecture

Video Fields Flowchart



Video-Fields

Mockup & Calibration



PROJECT

- ground
- hd visible
- hd transparent
- pen.rotate
- pen.translate
- pen.rotate
- pen.translate
- test0
- test1
- test2
- filter
- threshold0
- threshold1
- threshold2
- video

Close Controls

Fog

PERSPECTIVECAMER

UUID 567D3649-3AD6-4B61

Name

Position

Rotation

Scale

Fov

Near

Far

Shadow cast receive

Visible

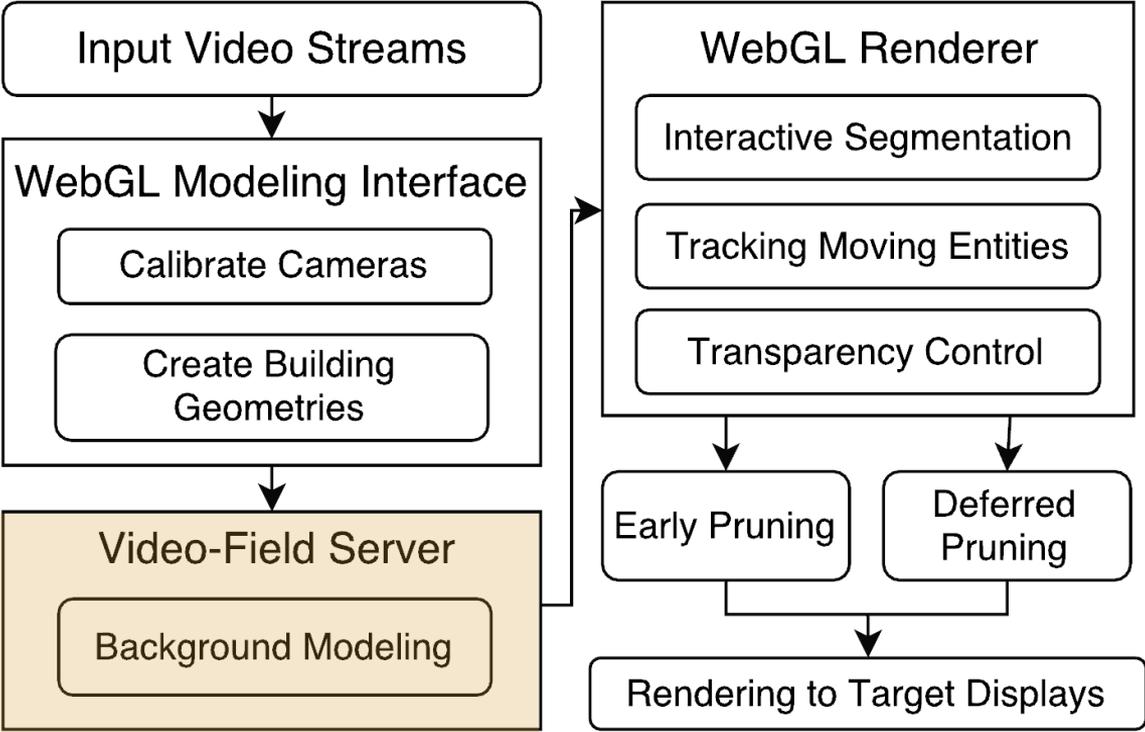
User data

objects 3
vertices 4
triangles 2

dynamic video-based virtual reality scenes in head-mounted displays, as well as high-resolution wide-field-of-view tiled display walls.

Architecture

Video Fields Flowchart



Background Modeling

Motivation

- Provide a background texture for each camera
- Identify moving entities in the rendering stage
- Reduce the network bandwidth requirements

Background Modeling

Gaussian Mixture Models (GMM)

$$\mathbf{T}(u, v)_i = \{\mathbf{T}(u, v, j), 1 \leq j \leq i\}$$

$$\mathcal{P}(\mathbf{T}(u, v)_i) = \sum_{j=1}^N \mathcal{N}(\mathbf{T}(u, v)_i | \mu_{ij}, \Sigma_{ij}) \cdot \omega_{ij} \quad (1)$$

$$\mathcal{N}(\mathbf{T}(u, v)_i | \mu, \Sigma) = \frac{1}{(2\pi)^{\frac{n}{2}}} \cdot \frac{1}{\Sigma^{\frac{1}{2}}} \cdot e^{-\frac{1}{2}(\mathbf{T}(u, v)_i - \mu_{ij})^T \sigma_{ij}^{-1} (\mathbf{T}(u, v)_i - \mu_{ij})} \quad (2)$$

$$\omega_{ij} \leftarrow (1 - \alpha)\omega_{i(j-1)} + \alpha\mathcal{M}_{ij} \quad (3)$$

Background Modeling

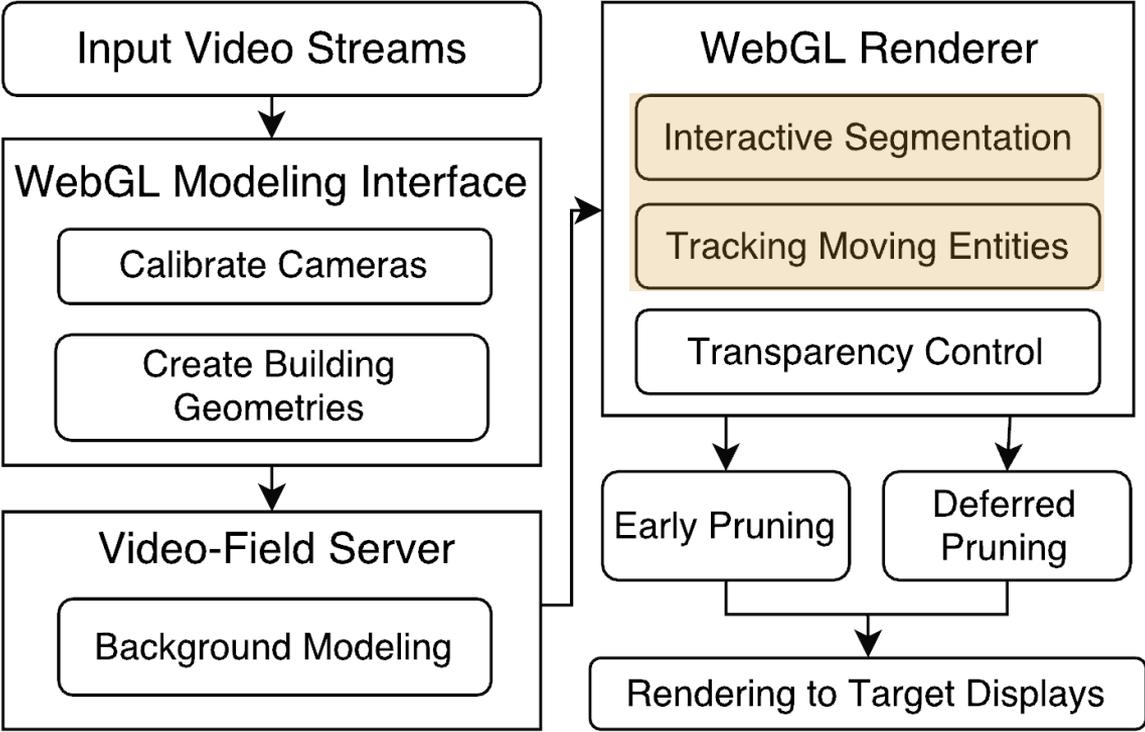
Advantages [Stauffer and Grimson]

More adaptive with:

- different lighting conditions,
- repetitive motions of scene elements,
- moving entities in slow motion

Architecture

Video Fields Flowchart



Segmentation

Moving Entities

$$\mathbf{T}' \leftarrow \mathcal{G}(\sigma) \otimes \mathbf{T}, \mathbf{B}' \leftarrow \mathcal{G}(\sigma) \otimes \mathbf{B}$$

$$\mathbf{F} \leftarrow \delta(|\mathbf{I}' - \mathbf{B}'|)$$

Background Modeling

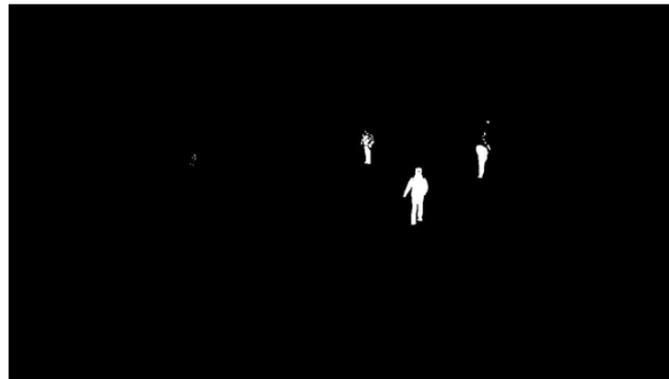
Gaussian Mixture Models (GMM)



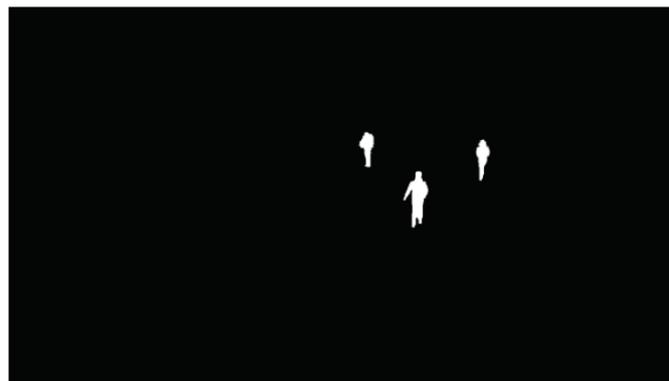
(a) source video texture



(b) background model by GMM



(c) segmentation without Gaussian convolution



(d) segmentation with Gaussian convolution

Video-Fields

Real-time Segmentation

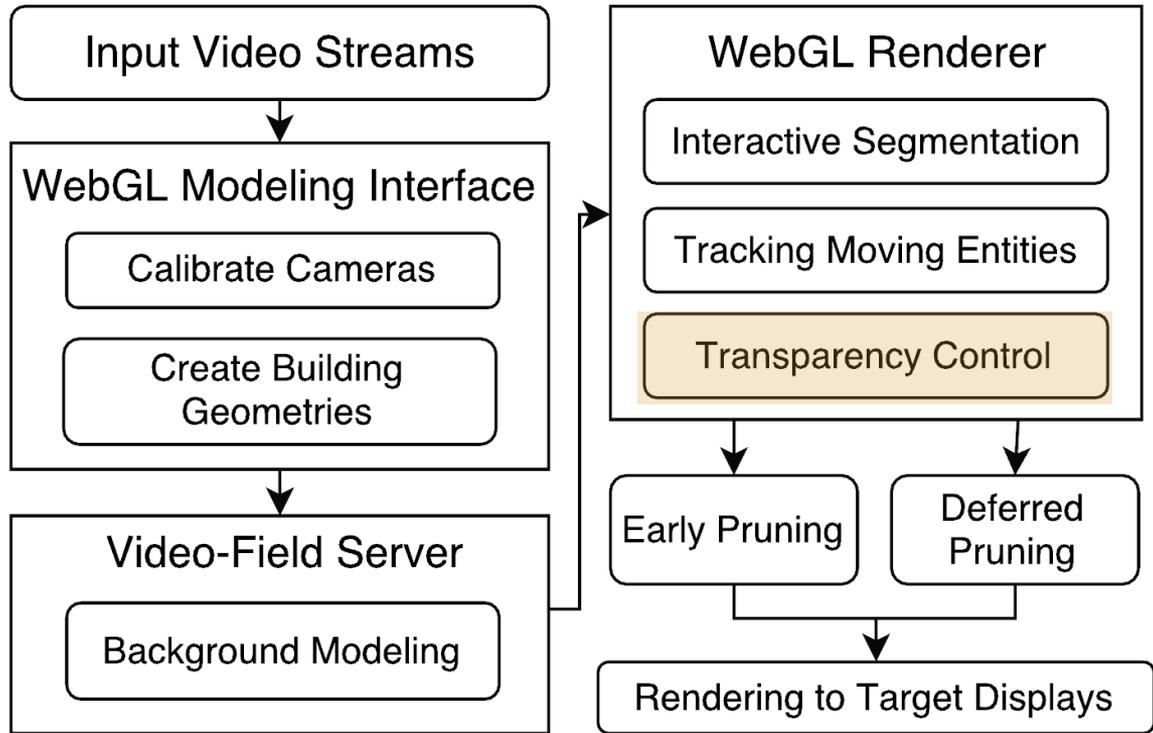


ground	Static	
bd.visible	<input checked="" type="checkbox"/>	
bd.transparent	<input checked="" type="checkbox"/>	
pm.rotate	<input type="checkbox"/>	
pm.attract	<input type="checkbox"/>	
test0	<input type="checkbox"/>	
test1	<input type="checkbox"/>	
test2	<input checked="" type="checkbox"/>	
filter	Gaussian	
threshold0	<input type="range" value="0.08"/>	0.08
threshold1	<input type="range" value="0.12"/>	0.12
threshold2	<input type="range" value="0"/>	0
video	<input type="range" value="5"/>	5
Close Controls		

Our system integrates background modeling and automatic segmentation of moving entities with rendering of video fields.

Architecture

Video Fields Flowchart



Visibility Test

Plus Opacity Modulation



(a) Rendering before visibility testing and opacity modulation



(b) Rendering after visibility testing and opacity modulation

Video-Fields

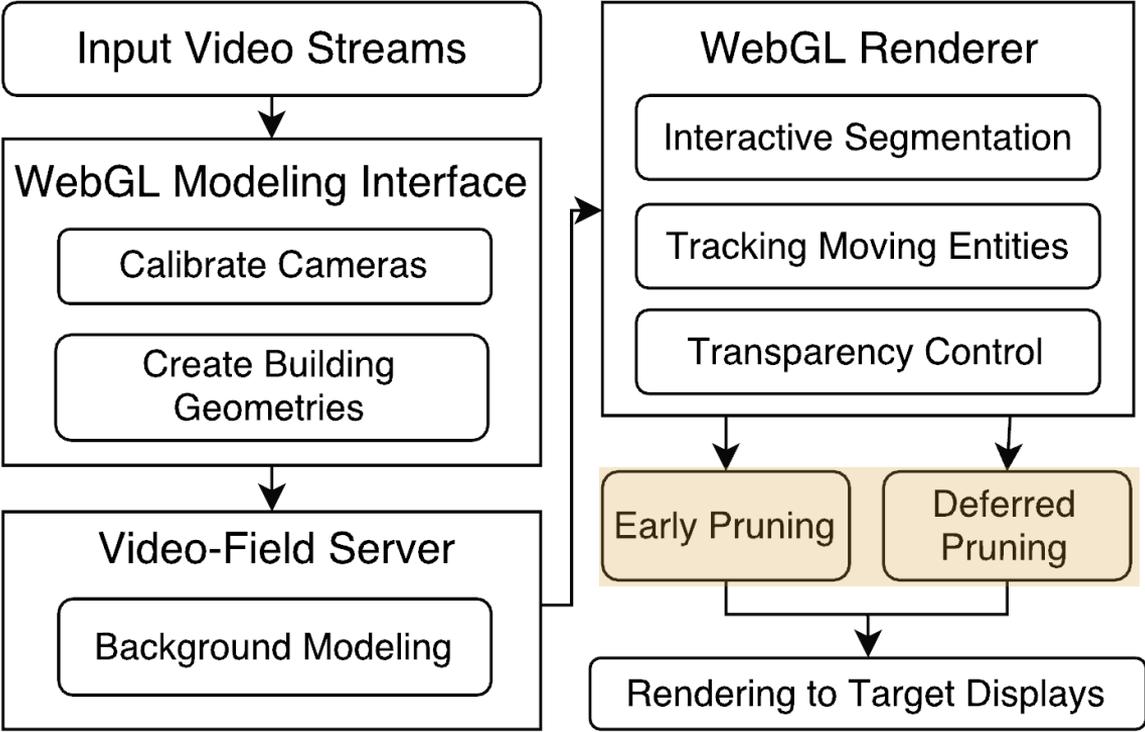
See-through Buildings



It allows users to adjust camera parameters, navigate through time, walk around the scene, and see through the buildings.

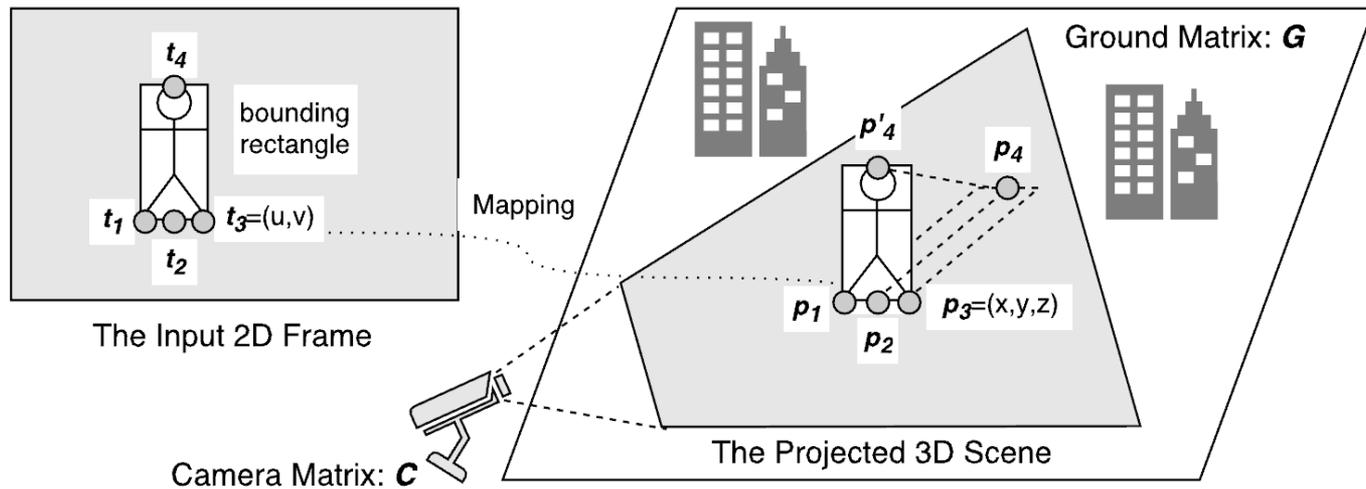
Architecture

Video Fields Flowchart



Video Fields Mapping

Overview



Video Fields Mapping

Challenges

1. Vertex in the 3D models -> Pixel in the texture space
 2. Pixel in the texture space -> Vertex on the ground
- The second is useful for projecting a 2D segmentation of a moving entity to the 3D world

Video Fields Mapping

Projection Mapping

$$\hat{\mathbf{p}}_{xyzw} \leftarrow \mathbf{C} \cdot \mathbf{G} \cdot (\mathbf{p}_{xyz}, 1.0) \quad (6)$$

$$\mathbf{t}_{uv} \leftarrow \left(\frac{\hat{\mathbf{p}}_x + \hat{\mathbf{p}}_w}{2\hat{\mathbf{p}}_w}, \frac{\hat{\mathbf{p}}_y + \hat{\mathbf{p}}_w}{2\hat{\mathbf{p}}_w} \right) \quad (7)$$

Video Fields Mapping

Perspective correction



(a) Video Fields mapping before perspective correction



(b) Video Fields mapping after perspective correction

Video Fields Mapping

Depth Map / Hashing Function

$$\mathcal{H} : \mathbf{t}_{uv} \mapsto \mathbf{p}_{xyz}$$

Early Pruning for Rendering Moving Entities

Voxels

ALGORITHM 1: Early Pruning for Rendering Moving Entities

Input: foreground F and the set of bounding rectangles R of moving entities

Output: a 3D point cloud P visualizing the moving entities

- 1 Initialize a set of points for the video visualization. (Run once);
 - 2 For each pixel t inside the bounding box, calculate the intersection point t_{\perp} between its perpendicular line and $t_1 t_3$;
 - 3 **for** *each pixel t from the video* **do**
 - 4 **if** $t \notin F$ **then**
 - 5 | discard t and **continue**;
 - 6 set the color of the pixel: $c \leftarrow \text{texture2D}(F, t)$;
 - 7 look up the corresponding projected points in the 3D scene:
 $p \leftarrow \mathcal{H}(t), p_{\perp} = \mathcal{H}(t_{\perp})$;
 - 8 update the z coordinate of the 3D point:
 $p_z \leftarrow |p - p_{\perp}| \cdot \tan(\theta_p)$;
 - 9 use the x, y coordinates of t_{\perp} to place the point vertically:
 $p_{xy} \leftarrow t_{uv}$;
 - 10 render the point p ;
-

Deferred Pruning for Rendering Moving Entities

Billboards

ALGORITHM 2: Deferred Pruning for Rendering Moving Entities

Input: foreground F and the set of bounding rectangles R of moving entities

Output: a set of billboards rendering the moving entities

- 1 Initialize a set of billboards to display moving objects. (Run once);
 - 2 **for** each detected bounding box r in R **do**
 - 3 calculate the bottom-left, bottom-middle, bottom-right and top-middle points t_1, t_2, t_3, t_4 in r , as illustrated in Fig. 5;
 - 4 look up the corresponding projected points in the 3D scene:
 $p_i \leftarrow \mathcal{H}(t_i), i \in \{1, 2, 3, 4\}$;
 - 5 calculate the width of the billboard in the 3D space:
 $w \leftarrow |p_3 - p_1|, h \leftarrow |p_4 - p_2| \cdot \tan(\theta_{p_4})$;
 - 6 Reposition a billboard to the position $\frac{p_1 + p_3}{2}$ with width and height w and h ;
 - 7 In the fragment shader of the billboard, sample the color from I as described in Equation. 6 and 7, but replace G with the current billboard's model matrix; discard pixels which does not belong to the foreground F ;
-

Visual Comparison

Early Pruning vs. Deferred Pruning



(a) early pruning
for rendering
moving entities



(b) deferred pruning
for rendering
moving entities

View-dependent Rendering



View-dependent Rendering



View-dependent Rendering



View-dependent Rendering



Experimental Results

Early Pruning vs. Deferred Pruning

Render Algorithm	Resolution	WebVR	Framerate
Early Pruning	2560×1440	No	60.0 fps
	$2 \times 960 \times 1080$	Yes	55.2 fps
	6000×3000	No	48.6 fps
Deferred Pruning	2560×1440	No	60.0 fps
	$2 \times 960 \times 1080$	Yes	41.5 fps
	6000×3000	No	32.4 fps



Video Fields: Fusing Multiple Surveillance Videos Into a Dynamic Virtual Environment

Ruofei Du, Sujal Bista, and Amitabh Varshney
www.Video-Fields.com
www.Augmentarium.com

Augmentarium | Department of Computer Science | UMIACS
University of Maryland, College Park
In Proceedings of the 21st Annual ACM SIGGRAPH Web3D Conference, 2016

Vocal: Sai Yuan; BGM: Ukulele by Bensound CC

Experimental Results

Early Pruning vs. Deferred Pruning

Render Algorithm	Resolution	WebVR	Framerate
Early Pruning	2560 × 1440	No	60.0 fps
	2 × 960 × 1080	Yes	55.2 fps
	6000 × 3000	No	48.6 fps
Deferred Pruning	2560 × 1440	No	60.0 fps
	2 × 960 × 1080	Yes	41.5 fps
	6000 × 3000	No	32.4 fps

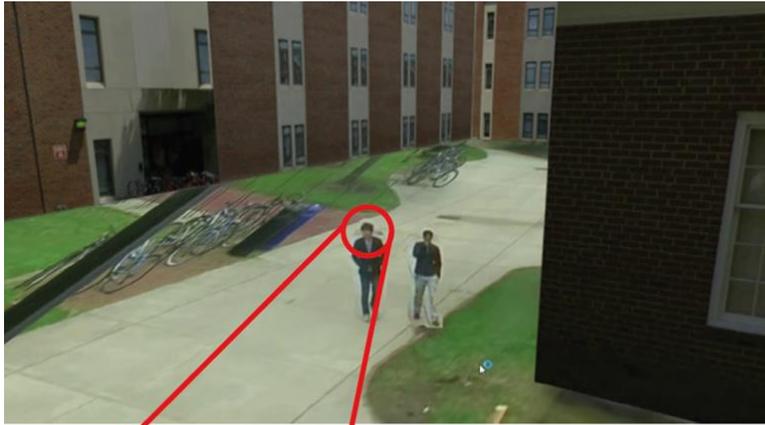
Experimental Results

Early Pruning vs. Deferred Pruning

Render Algorithm	Resolution	WebVR	Framerate
Early Pruning	2560 × 1440	No	60.0 fps
	2 × 960 × 1080	Yes	55.2 fps
	6000 × 3000	No	48.6 fps
Deferred Pruning	2560 × 1440	No	60.0 fps
	2 × 960 × 1080	Yes	41.5 fps
	6000 × 3000	No	32.4 fps

Visual Comparison

Early Pruning vs. Deferred Pruning



(a) early pruning
for rendering
moving entities



(b) deferred pruning
for rendering
moving entities

Video-Fields

Overview



ground	Static
bd.visible	<input checked="" type="checkbox"/>
bd.transparent	<input checked="" type="checkbox"/>
pm.rotate	<input type="checkbox"/>
pm.attract	<input type="checkbox"/>
test0	<input type="checkbox"/>
test1	<input type="checkbox"/>
test2	<input checked="" type="checkbox"/>
filter	Gaussian
threshold0	<input type="range" value="0.08"/> 0.08
threshold1	<input type="range" value="0.12"/> 0.12
threshold2	<input type="range" value="0"/> 0
video	<input type="range" value="17"/> 17
Close Controls	

In this paper we introduce, Video Fields, a novel web-based interactive system to create, calibrate, and render ...

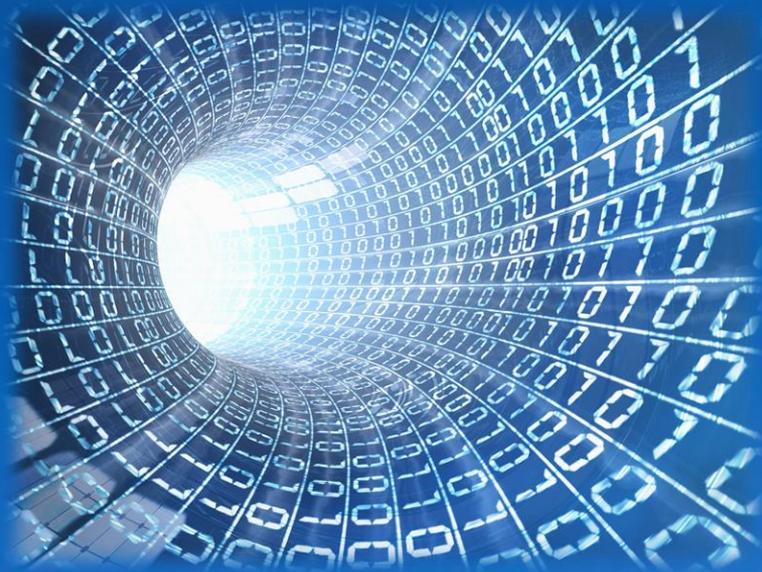
Future Work

Scale Up - Hundreds of cameras



Future Work

Bandwidth Problem



Future Work

Holoportation with RGB cameras



Acknowledgement

Augmentarium Lab | GVIL | UMIACS



Acknowledgement

NSF | Nvidia | MPower | UMIACS



UMIACS
University of Maryland
Institute for Advanced
Computer Studies



UNIVERSITY OF
MARYLAND



UNIVERSITY OF MARYLAND
MPOWERING THE STATE

Video Fields

www.Video-Fields.com

Thank you! Questions or comments?

Ruofei Du and Amitabh Varshney

Augmentarium Lab | GVIL | UMIACS

Web3D 2016