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, sujal, varshney} @ cs.umd.edu
www.VideoFields.com

Introduction

Surveillance Videos - Monitoring

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

image courtesy: university of maryland, college park

Introduction

Surveillance Videos – Cognitive Burden

image courtesy: theimaginativeconservative.org

Introduction

Surveillance Videos – Fusing & Interpreting

image courtesy: university of maryland, college park

Related Work

Fusing Multiple Static Photographs

Related Work

Fusing Multiple Static Photographs

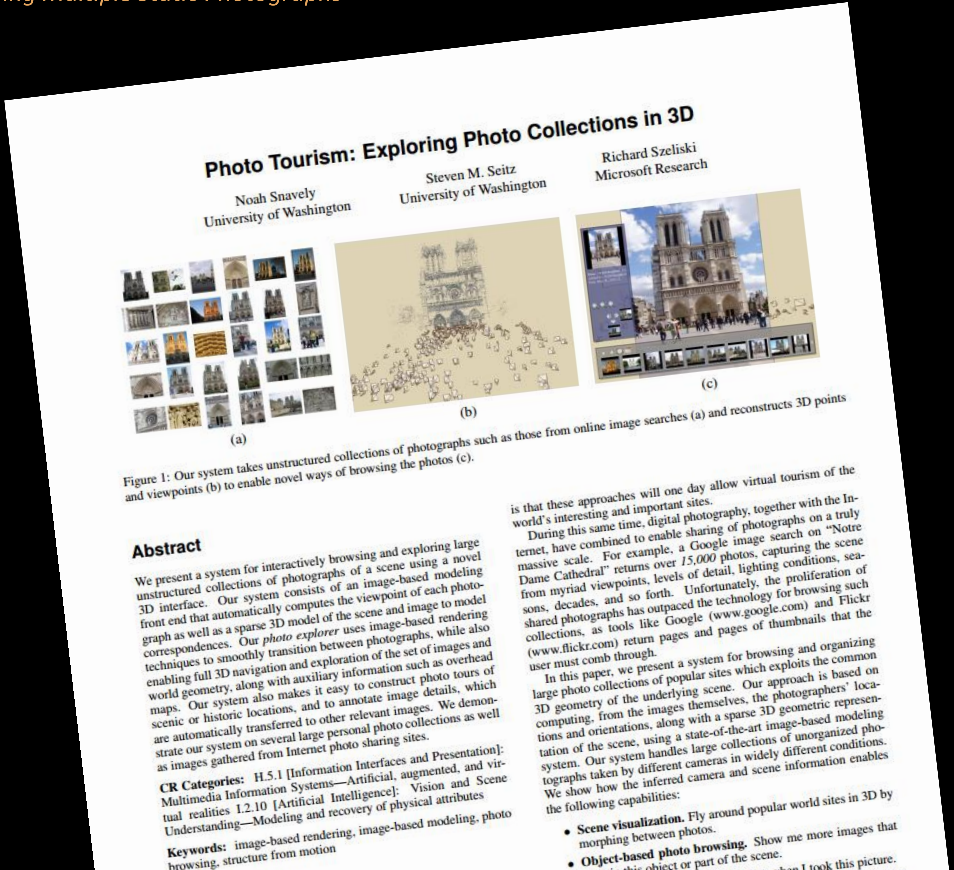


Photo Tourism: Exploring Photo Collections in 3D

Noah Snavely
University of Washington

Steven M. Seitz
University of Washington

Richard Szeliski
Microsoft Research

(a) (b) (c)

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 image details, which are automatically transferred to other relevant images. We demonstrate our system on 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 I.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.

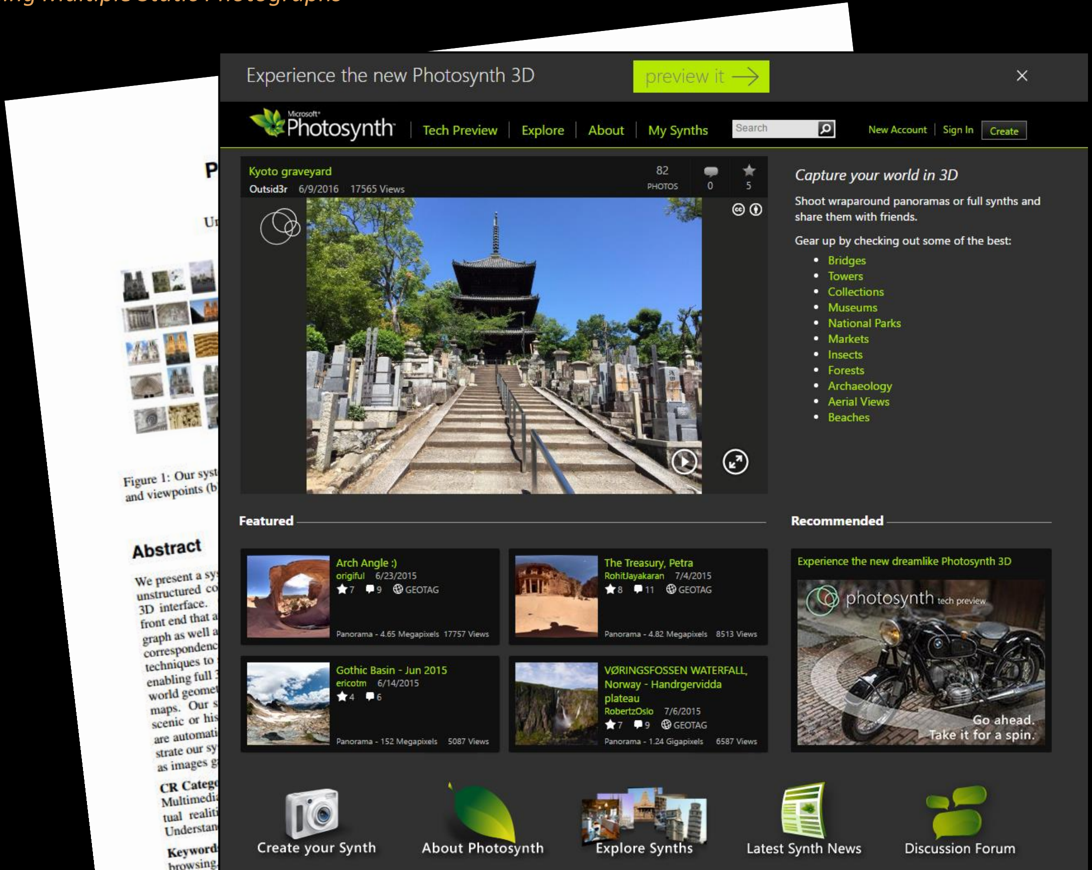
During this same time, digital photography, together with the Internet, have combined to enable sharing of photographs on a truly massive scale. For example, a Google image search on "Notre Dame Cathedral" returns over *15,000* photos, capturing the scene from myriad viewpoints, levels of detail, lighting 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 this object or part of the scene.

Related Work

Fusing Multiple Static Photographs



Related Work

Fusing Multiple Static Photographs

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 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 ...

Social Snapshot's Contributions

Social Snapshot's contributions fit naturally into two categories: technical and social.

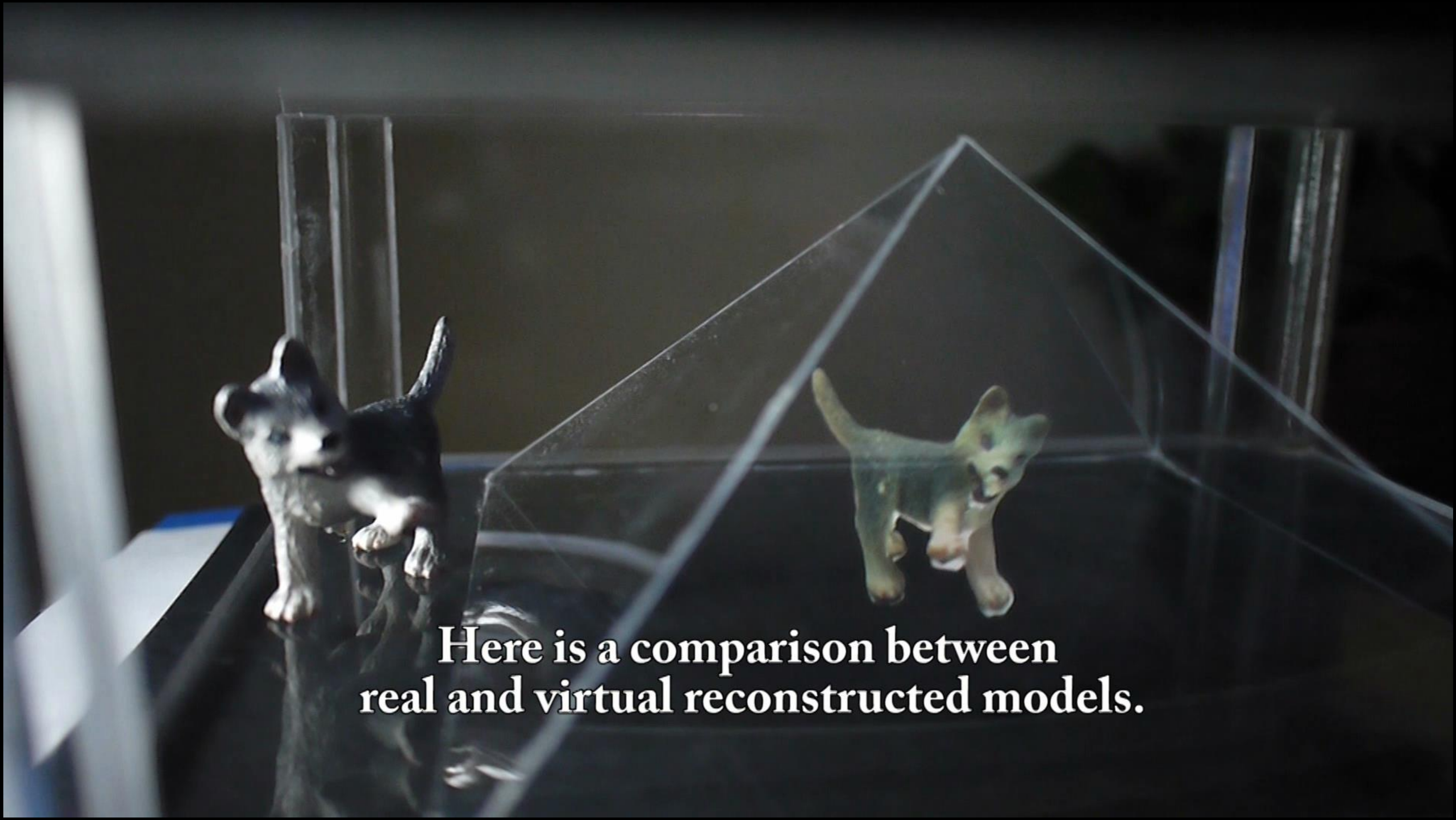
The technical contributions are improved algorithms and techniques that enhance our system's novelty and scalability. For example, Social Snapshot produces a textured and colored-mesh reconstruction from a loosely ordered photo collection, rather than the sparse or dense point reconstructions produced 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.

The social contributions lead to a new way of thinking about the interplay between data acquisition and social interactions. They also let us define social photography as an active, rather than a passive, endeavor. For example, Social Snapshot encourages collaborative photography as a social endeavor, letting users capture dynamic action 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 ...

Related Work

Fusing Multiple Static Photographs


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 Furukawa[1]
Department of Computer Science
and Beckman Institute
University of Illinois at Urbana-Champaign, USA[1]

Jean Ponce[2,1]
Willow Team
LIENS (CNRS/ENS/INRIA UMR 8548)
Ecole Normale Supérieure, Paris, France[2]

Abstract: *This paper proposes a novel approach to nonrigid, markerless motion capture from synchronized video streams acquired by calibrated cameras. The instantaneous 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 ... to recover skeletal body configurations, ... recover the com-

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 have also been proposed [4, 13, 16, 19, 23]. Some ... optical flow in each image inde- ... recovered

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 recovered by applying model-guided multi-view stereo to refine the model surface. 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: I.3.7 [Computer Graphics]: Three-Dimensional Graphics and Realism; I.4.8 [Image Processing and Computer Vision]: Scene Analysis

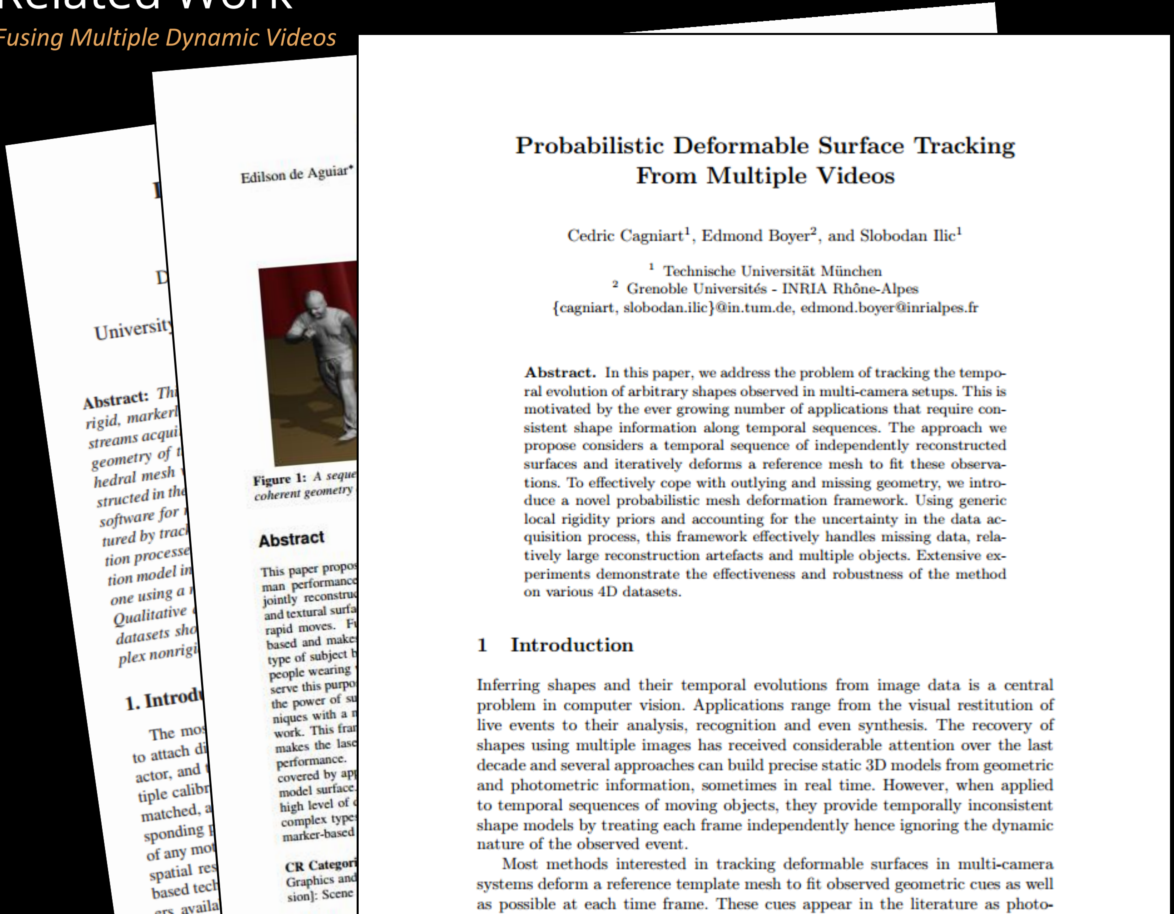
1 Introduction

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 spatio-temporally coherent time-varying geometry (i.e. geometry

Related Work

Fusing Multiple Dynamic Videos



Probabilistic Deformable Surface Tracking From Multiple Videos

Cedric Cagniart[1], Edmond Boyer[2], and Slobodan Ilic[1]

[1] Technische Universität München
[2] Grenoble Universités - INRIA Rhône-Alpes
{cagniart, 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.

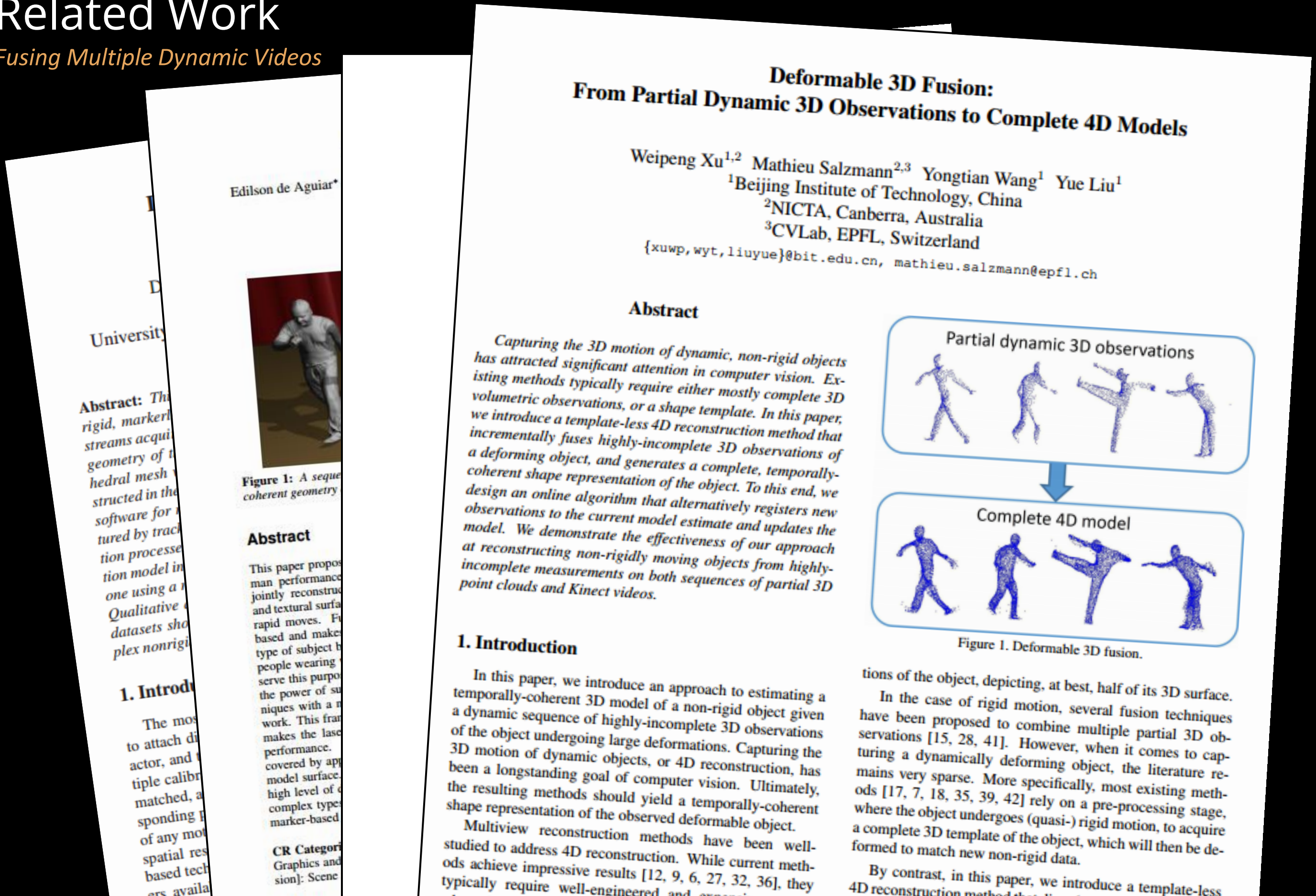
1 Introduction

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



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

Weipeng Xu[1,2] Mathieu Salzmann[2,3] Yongtian Wang[1] Yue Liu[1]
[1]Beijing Institute of Technology, China
[2]NICTA, Canberra, Australia
[3]CVLab, 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.

Figure 1. Deformable 3D fusion.

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 ...

tions 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 ...

Related Work

Fusing Multiple Dynamic Videos

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 arts, 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 ...

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.

mum, especially in the case of fast and complex motions. 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 ...

Related Work

Fusing Multiple Dynamic Videos

KinectFusion: Real-Time Dense Surface Mapping and Tracking*

Richard A. Newcombe, Imperial College London
Shahram Izadi, Microsoft Research
Otmar Hilliges, Microsoft Research
David Molyneaux, Microsoft Research, Lancaster University
David Kim, Microsoft Research, Newcastle University
Andrew J. Davison, Imperial College London
Pushmeet Kohli, Microsoft Research
Jamie Shotton, Microsoft Research
Steve Hodges, Microsoft Research
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 (greyscale) 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 arbitrary indoor scenes in variable lighting conditions, using only a moving 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.

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 Mi

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 localisation and mapping (SLAM) has focused more on real-time markerless tracking and *live* scene reconstruction 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 MVS-style reconstruction modules, enabling more sophisticated occlusion prediction and surface interaction [19, 26]. Most recently in this line of research, iterative image alignment against dense reconstructions has also been used to replace point features for camera tracking [20]. While this work is very

Related Work

Fusing Multiple Dynamic Videos

Related Work

Fusing Multiple Dynamic Videos

Related Work

Fusing Multiple Dynamic Videos

SIGGRAPH 2016

Wednesday, 3:30-4:00 PM

Related Work

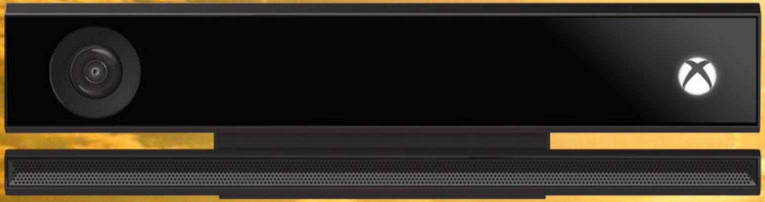
Fusing Multiple Dynamic Videos

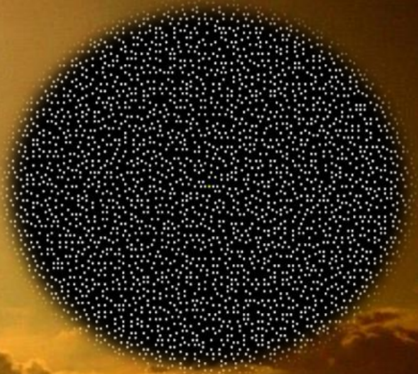
Related Work

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

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

Conception, architecting & implementation

1

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

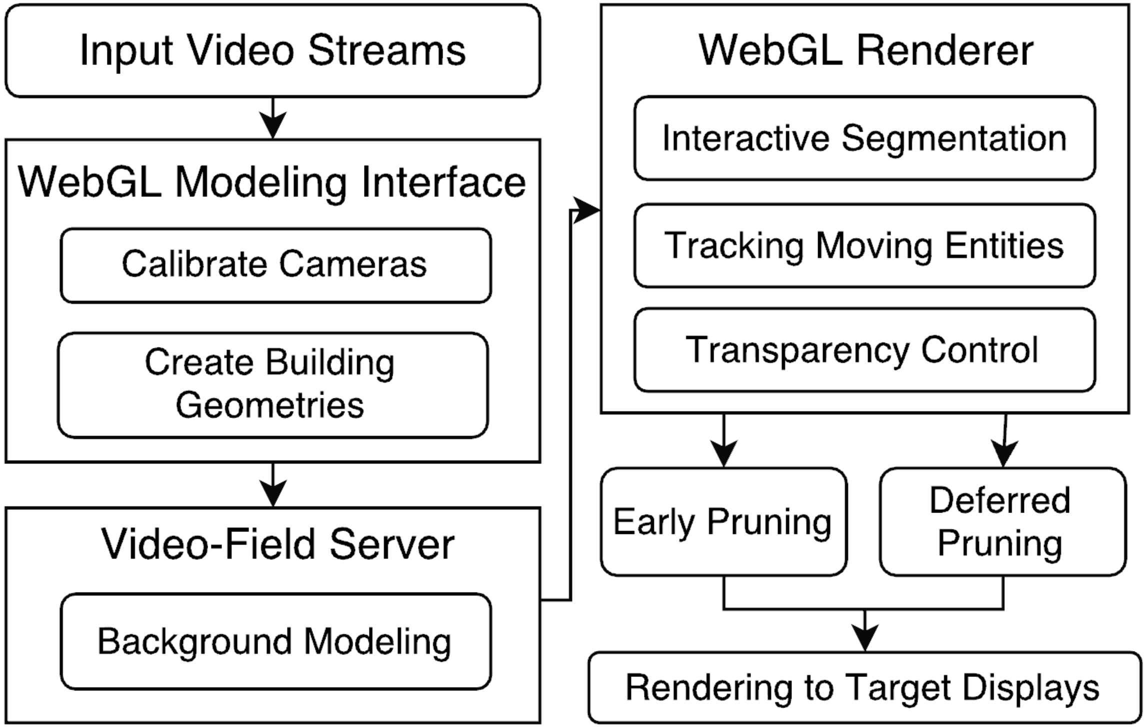
WebGL + Three.js

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

System Overview

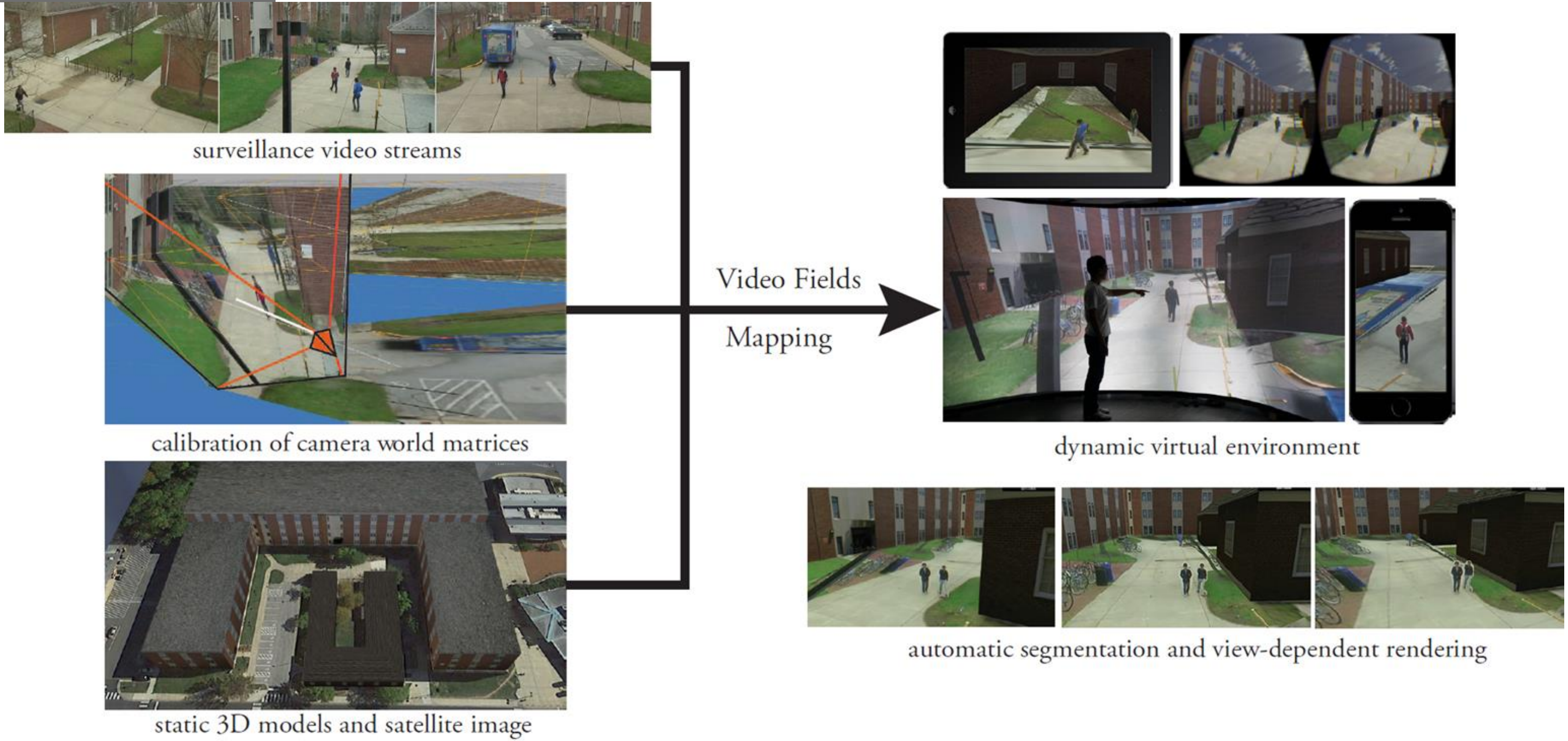
Architecture

Video Fields Flowchart



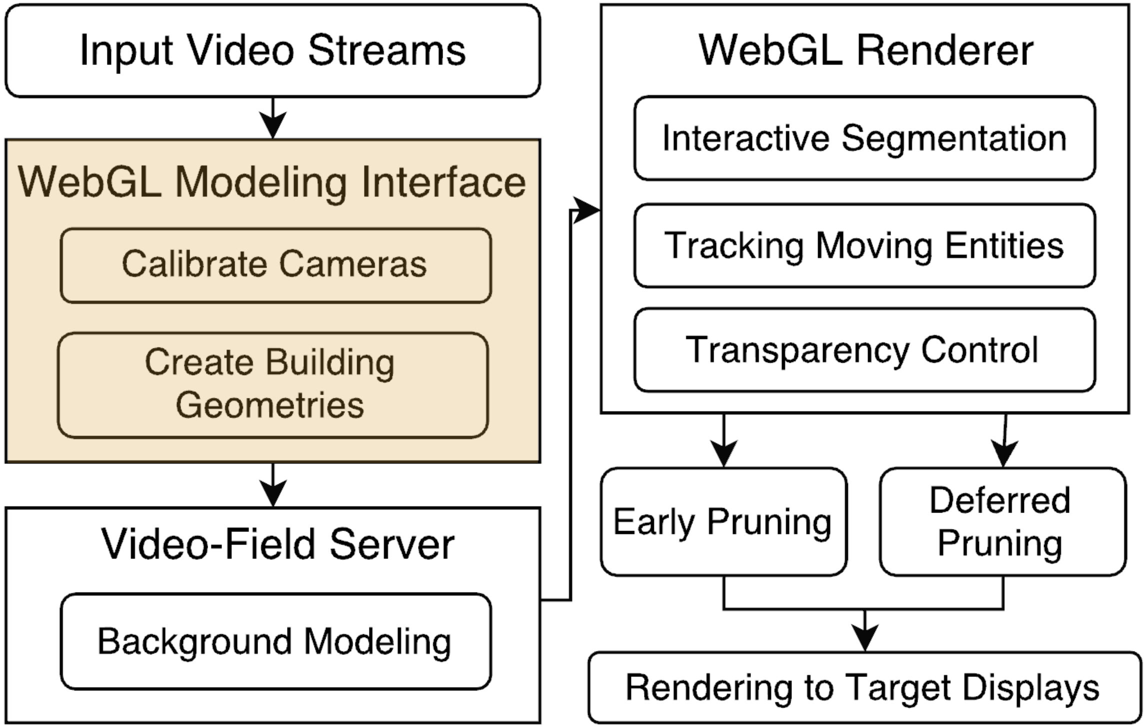
Architecture

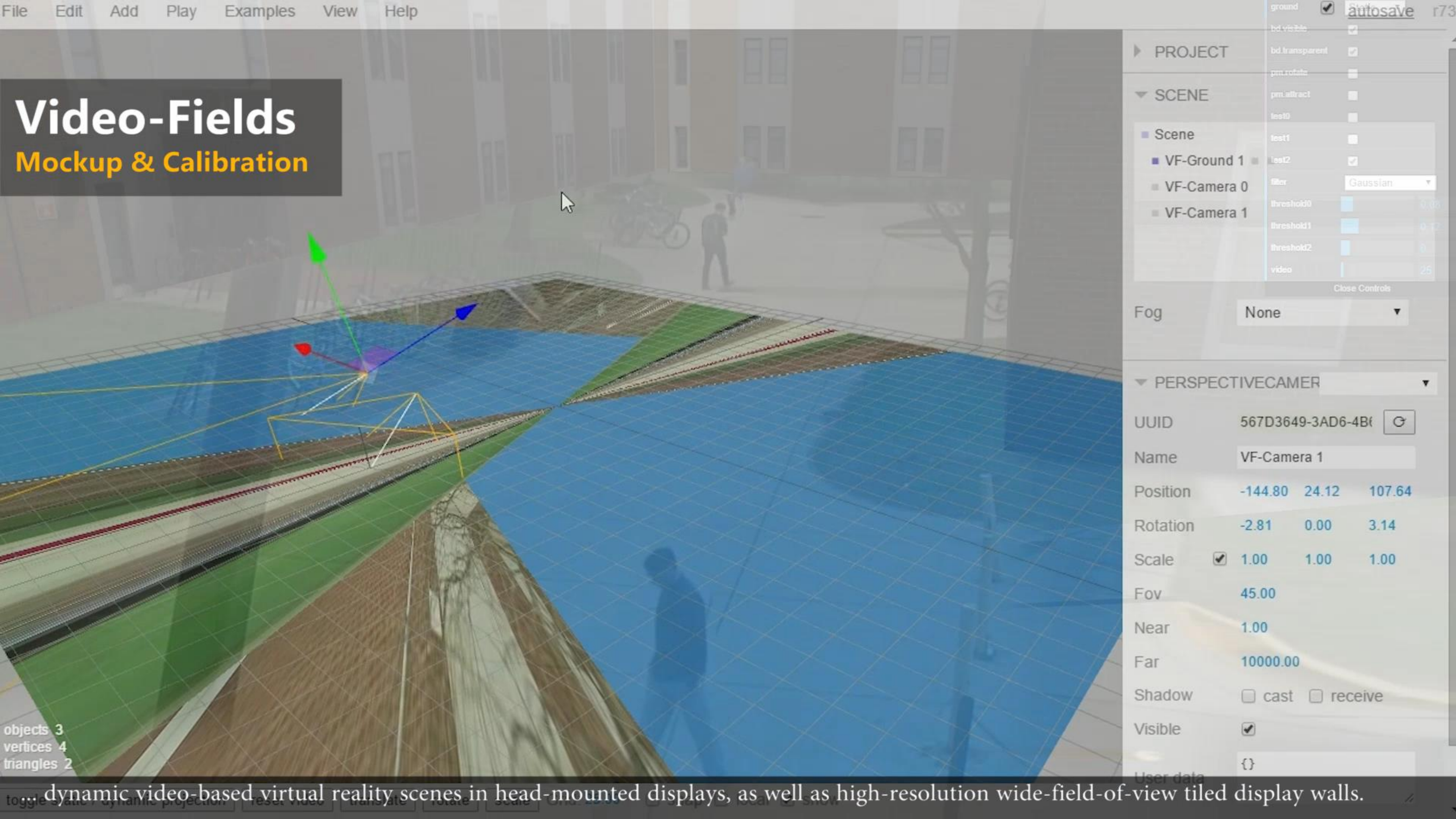
Video Fields Flowchart



Architecture

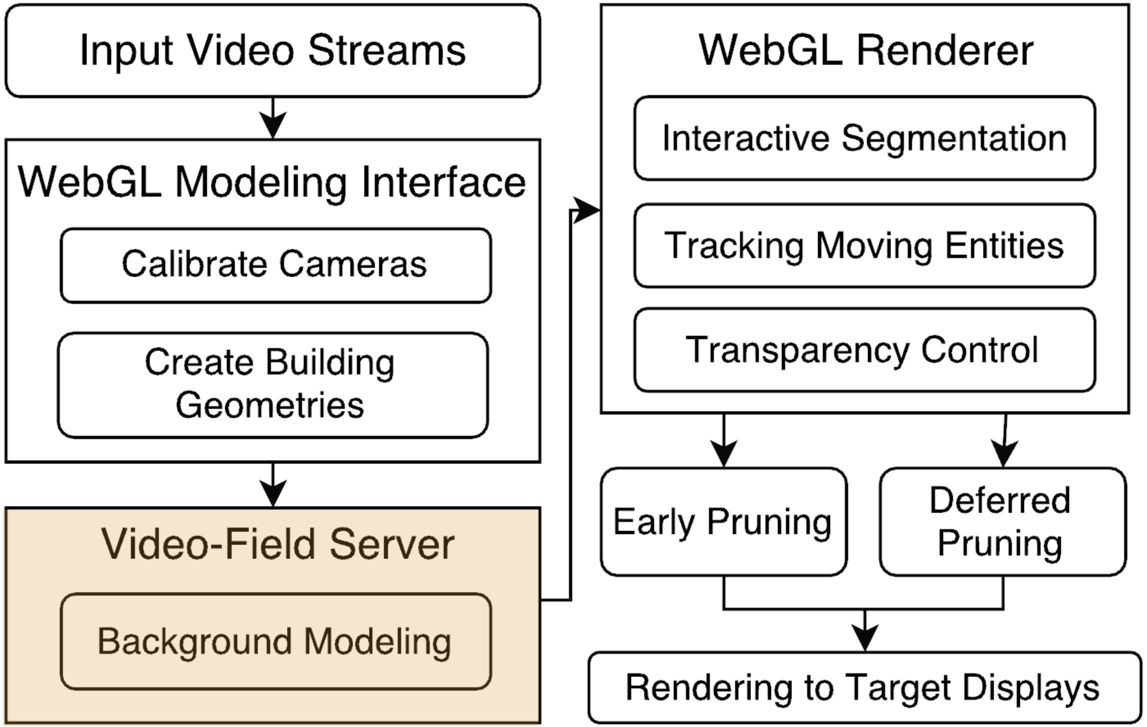
Video Fields Flowchart




File Edit Add Play Examples View Help
autosave
Video-Fields
Mockup & Calibration
PROJECT
SCENE
Scene
VF-Ground 1
VF-Camera 0
VF-Camera 1
Close Controls
Fog None
PERSPECTIVECAMER
UUID 567D3649-3AD6-4B(
Name VF-Camera 1
Position -144.80 24.12 107.64
Rotation -2.81 0.00 3.14
Scale 1.00 1.00 1.00
Fov 45.00
Near 1.00
Far 10000.00
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)

$$\boldsymbol{T}(u,v)_i = \{\boldsymbol{T}(u,v,j), 1 \leq j \leq i\}$$
$$\mathscr{P}(\boldsymbol{T}(u,v)_i) = \sum_{j=1}^{N} \mathscr{N}(\boldsymbol{T}(u,v)_i | \mu_{ij}, \Sigma_{ij}) \cdot \omega_{ij} \tag{1}$$

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

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

Background Modeling

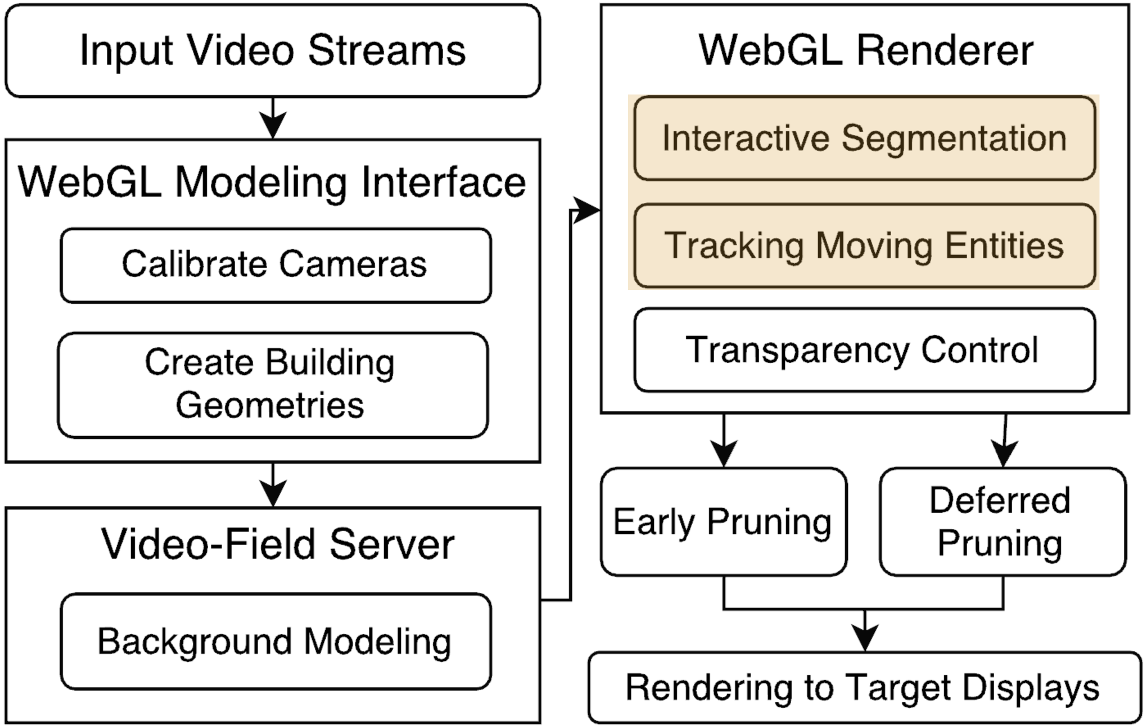
Advantages [Stauffer and Grimson]

More adaptive with:

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

Architecture

Video Fields Flowchart



Segmentation

Moving Entities

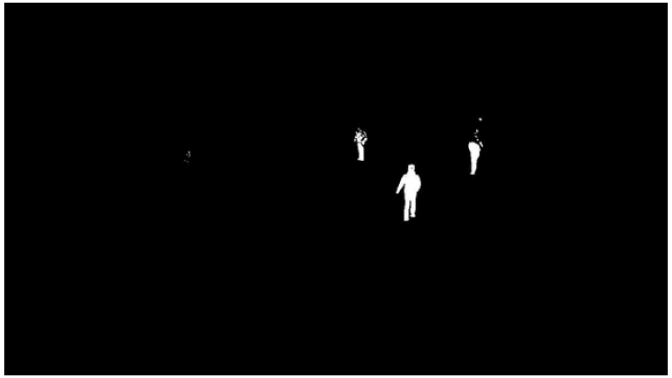
$$\boldsymbol{T}' \leftarrow \mathscr{G}(\sigma) \otimes \boldsymbol{T}, \boldsymbol{B}' \leftarrow \mathscr{G}(\sigma) \otimes \boldsymbol{B}$$

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

Background Modeling

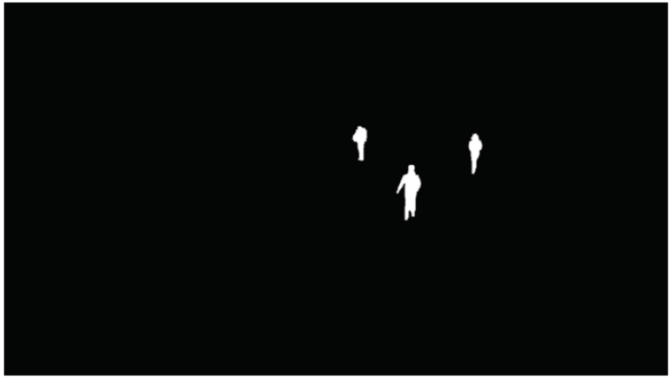
Gaussian Mixture Models (GMM)

(a) source video texture



(c) segmentation without Gaussian convolution

(b) background model by GMM

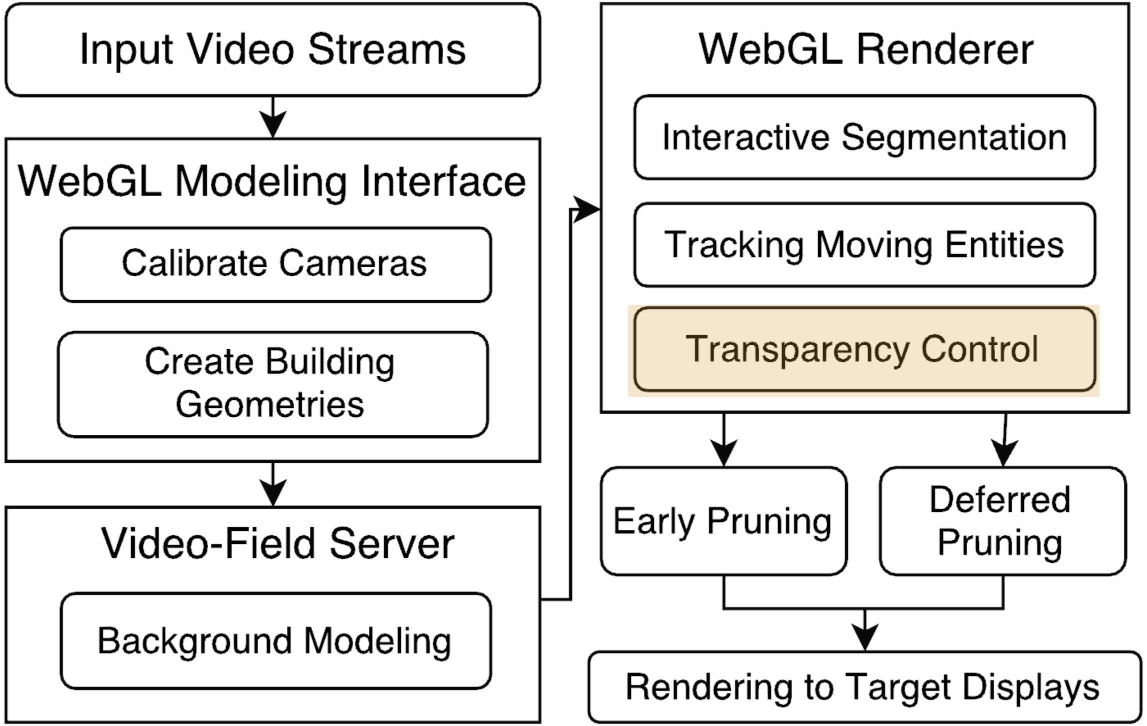


(d) segmentation with Gaussian convolution

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

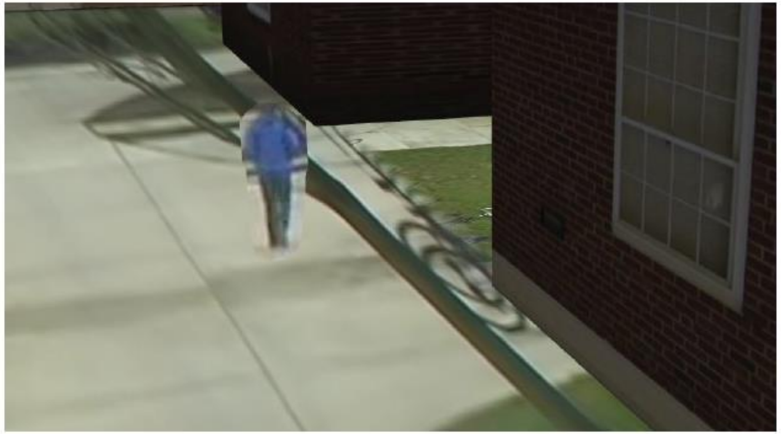
Architecture

Video Fields Flowchart

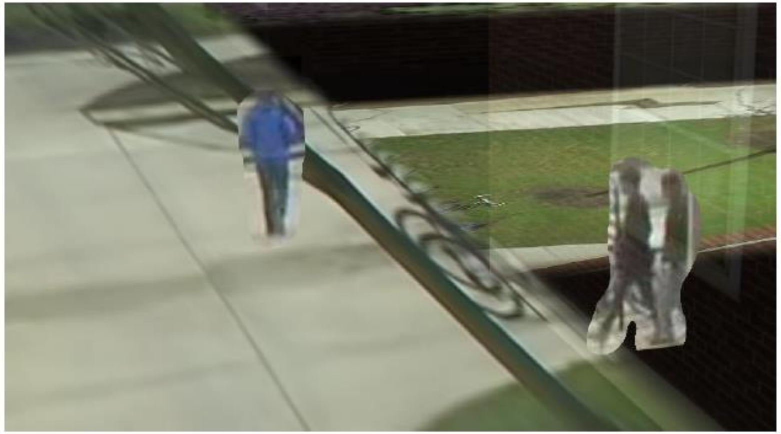


Visibility Test

Plus Opacity Modulation



(a) Rendering before visibility testing and opacity modulation



(b) Rendering after visibility testing and opacity modulation

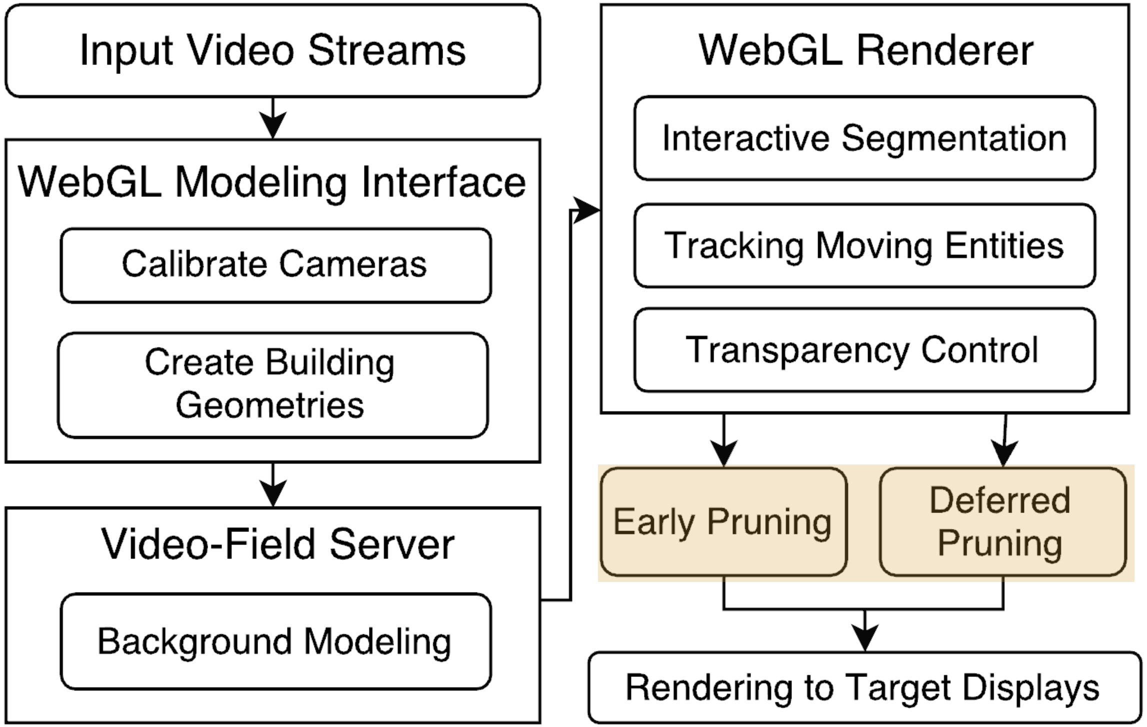
Video-Fields

See-through Buildings

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

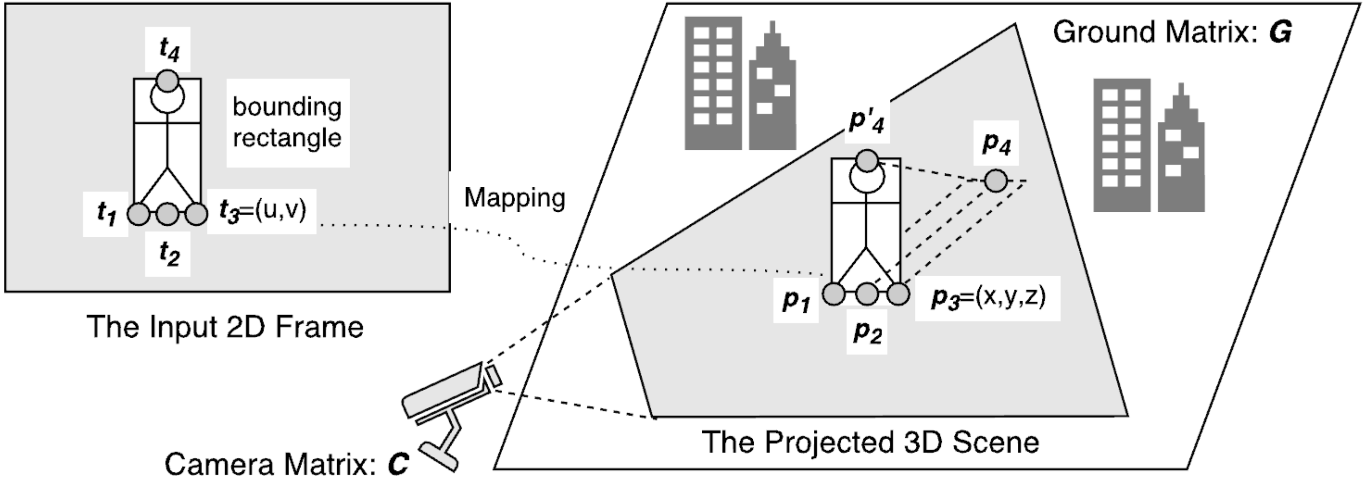
Architecture

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

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

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

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

$$\mathscr{H} : \boldsymbol{t}_{uv} \longmapsto \boldsymbol{p}_{xyz}$$

Early Pruning for Rendering Moving Entities

Voxels

ALGORITHM 1: Early Pruning for Rendering Moving Entities

Input: foreground $\boldsymbol{F}$ and the set of bounding rectangles R of moving entities

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

Initialize a set of points for the video visualization. (Run once);

For each pixel $\boldsymbol{t}$ inside the bounding box, calculate the intersection point $\boldsymbol{t}_\perp$ between its perpendicular line and $\boldsymbol{t}_1\boldsymbol{t}_3$;

for *each pixel* **t** *from the video* **do**

  if $\boldsymbol{t} \notin \boldsymbol{F}$ **then**

    discard $\boldsymbol{t}$ and **continue**;

  set the color of the pixel: $c \leftarrow \mathbf{texture2D}(\boldsymbol{F}, \boldsymbol{t})$;

  look up the corresponding projected points in the 3D scene: $\boldsymbol{p} \leftarrow \mathscr{H}(\boldsymbol{t}), \boldsymbol{p}_\perp = \mathscr{H}(\boldsymbol{t}_\perp)$;

  update the z coordinate of the 3D point: $\boldsymbol{p}_z \leftarrow |\boldsymbol{p} - \boldsymbol{p}_\perp| \cdot \tan(\theta_{\boldsymbol{p}})$;

  use the x, y coordinates of $\boldsymbol{t}_\perp$ to place the point vertically: $\boldsymbol{p}_{xy} \leftarrow \boldsymbol{t}_{uv}$;

  render the point $\boldsymbol{p}$;

Deferred Pruning for Rendering Moving Entities

Billboards

ALGORITHM 2: Deferred Pruning for Rendering Moving Entities

Input: foreground $\boldsymbol{F}$ and the set of bounding rectangles $\boldsymbol{R}$ 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* $\boldsymbol{r}$ *in* $\boldsymbol{R}$ **do**

3 calculate the bottom-left, bottom-middle, bottom-right and top-middle points $\boldsymbol{t}_1, \boldsymbol{t}_2, \boldsymbol{t}_3, \boldsymbol{t}_4$ in $\boldsymbol{r}$, as illustrated in Fig. 5;

4 look up the corresponding projected points in the 3D scene: $\boldsymbol{p}_i \leftarrow \mathscr{H}(\boldsymbol{t}_i), i \in \{1, 2, 3, 4\}$.;

5 calculate the width of the billboard in the 3D space: $w \leftarrow |\boldsymbol{p}_3 - \boldsymbol{p}_1|, h \leftarrow |\boldsymbol{p}_4 - \boldsymbol{p}_2| \cdot \tan(\theta_{\boldsymbol{p}_4})$.;

6 Reposition a billboard to the position $\frac{\boldsymbol{p}_1 + \boldsymbol{p}_3}{2}$ 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

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
	$2 \times 960 \times 1080$	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

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

Visual Comparison

Early Pruning vs. Deferred Pruning

Video-Fields

Overview

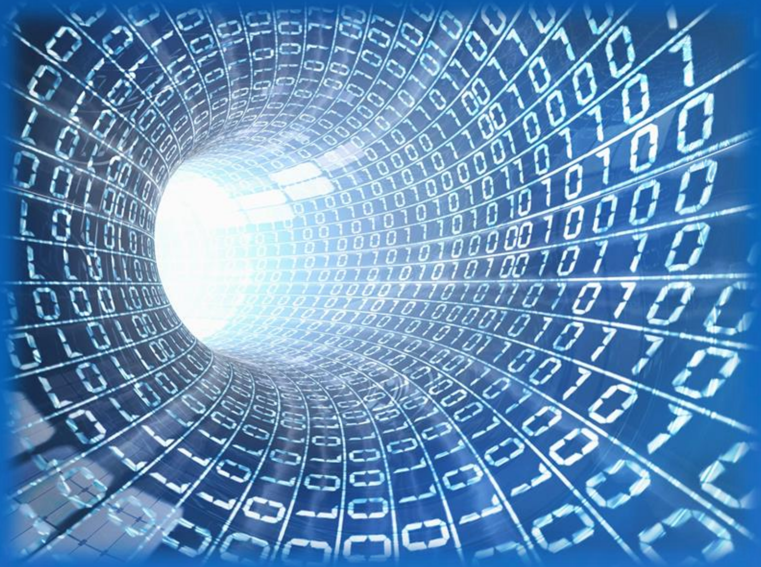
In this paper we introduce, Video Fields, a novel web-based interactive system to create, calibrate, and render ...

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

NSF

NVIDIA

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