Computational 3D Photography
Extracting Shape, Motion and Appearance from Images

Marc Pollefeys
ETH Zurich

International Symposium on Visual Computing
29 November 2010
Video → 3D model

Recorded at archaeological site of Sagalassos in Turkey
accuracy ~1/500 from DV video (i.e. 140kb jpegs 576x720)
Talk outline

• Introduction
• Object modeling
• Scene modeling
• People/event modeling
• Summary and conclusion
Application areas and motivation

Visualization and metrology

Virtual worlds

Biometry

Industrial metrology

Forensics

Cultural heritage and archaeology

Robot navigation

Convergence of computer vision, graphics and photogrammetry
Application areas and motivation

- Medical (training, tele-medicine)
- VR & Games (dynamic content capture)
- Surveillance (3D sensor fusion)
- Tele-immersion/3DTV
- Motion analysis
- Intangible Heritage
- and more ...

convergence of computer vision, graphics and photogrammetry
Basic camera model: perspective projection

- Camera center
- Image plane
- Image point
- Line of sight
- World point
2D → 3D reconstruction

Triangulation
- calibration
- correspondences
3D from Video

(Pollefeys et al. ICCV’98)

... (Pollefeys et al. IJCV’04)
Talk outline

• Introduction
• Object modeling
• Scene modeling
• People/event modeling
• Summary and conclusion
2D → 3D reconstruction: silhouette constraints

Additional constraint for closed objects

Silhouettes
- object inside cone (visual hull)
- object tangent to cone (rim)
Multi-view 3D object reconstruction

• Combine dense matching with silhouette constraints (Compute graph min-cut to obtain watertight surface)
  – Exact silhouettes (Sinha & Pollefeys ICCV’05)
  – Photo-consistency adaptive tetrahedral mesh (Sinha et al. ICCV’07)
Talk outline

• Introduction
• Object modeling
• Scene modeling
• People/event modeling
• Summary and conclusion
Modeling the world

- Need for 3D models of real world

  e.g. interactive 3D modeling of architectural (Sinha et al. Siggraph Asia 08)

  collaboration with Microsoft Research
Fast automated video-based modeling of cities

2x4 cameras
1024x768@30Hz

capture ≈1TB/hour raw video data

GPS/INS system
Fast video-based modeling of cities

Fast video processing pipeline
- up to 26Hz on single CPU/GPU
- Most image processing on GPU (x10-100 faster)

- Exploits urban structure

- Generates textured 3D mesh
  (Pollefeys et al. IJCV, 2008)
2D Feature Tracker

fast GPU-based feature tracking
(Sinha et al. MVA’07, Zach et al.08)

+ tracking of exposure changes
(Kim et al. ICCV07)

tracks 1000 features at 200Hz
3D Tracker / Geo-location

- Fusion of 2D video tracks and INS/GPS

or use 2D video tracks only (need to deal with drift, see later)

Interesting option to use vertical orientation (Fraundorfer et al. ECCV2010) or vehicle motion (Scaramuzza et al. ICCV2009) to facilitate motion estimation
Dense multi-view matching

- Plane-sweep multi-view depth estimation on GPU
  (Yang & Pollefeys, CVPR’03)

Blend:
\[
\frac{(I_0 + I_1 + I_2 + I_3 + I_4)}{5}
\]
(correct depth=in focus)

Sum of Absolute Differences:
\[
|I_1 - I_0| + |I_2 - I_0| + |I_3 - I_0| + |I_4 - I_0|
\]
(correct depth=small value =dark)
Dense 3D surface reconstruction

- Multi-Directional plane-sweeping stereo
  - Sweep along façade & ground-plane directions
    - Choose best-cost solution over depth and orientation

- Fuse depth-maps to obtain consensus depth map by minimizing visibility conflicts
  - Free-space violation
  - Occlusion conflict

3D model from 11 video frames (hand-held)

(Merrell et al., ICCV07)

(Gallup et al., CVPR07)
3D-from-video evaluation: Firestone building

**RMS error:** 13.4cm
**mean error:** 6.8cm
**median error:** 3.0cm

error histogram
3D-from-video evaluation: Middlebury Multi-View Stereo Evaluation Benchmark

Ring datasets: 47 images

Results competitive but much, much faster (30 minutes → 30 seconds)
1.3 million video frames
(Chapel Hill, NC)
• 1.3 million frames (2 cams per side)
• 26 Hz reconstruction frame rate

Computation time:
1PC (3Ghz CPU+ Nvidia 8800 GTX):
  14hrs @ 26fps
  2 weeks @ 1fps
  2.5 years @ 1fpm
• 1.3 million frames (2 cams per side)
• 26 Hz reconstruction frame rate

Computation time:
IPC (3GHz CPU+ Nvidia 8800 GTX):
14hrs @ 26fps
Real-time stereo limitations

Street-Side Video

Notice problems at windows and homogeneous areas

Real-Time Stereo
Including planar prior for urban scenes

(Gallup et al. CVPR10)
Including planar prior for urban scenes

(Gallup et al. CVPR10)
n-layer heightmap fusion

(Gallup et al. DAGM10)
Video-only large-scale reconstruction?

Challenge:
Error accumulation yields drift of relative scale, orientation and position

Solution:
cancel drift by closing loops (e.g. at intersections)
Need to visually recognize locations
Solving 3D puzzles with VIPs

**SIFT features**
- Extracted from 2D images
- Variation due to viewpoint

**VIP features** *(Wu et al., CVPR08)*
- Extracted from 3D model
- Viewpoint invariant
3D Models with VIPs
Geo-location from images

(Baatz et al., ECCV2010)

Images + 3D Database

Building ortho-textures

Rectification of query image

descriptor database

rectified features

promising candidates

Geometric verification

Collaboration with NOKIA

Computational 3D Photography

scale

x translation

y translation

38
Minimal relative pose with known vertical

(Fraundorfer et al., ECCV2010)

Vertical direction can often be estimated
- inertial sensor
- vanishing point

\[ E = \begin{bmatrix} t_z \sin(y) & -t_z \cos(y) & t_y \\ t_z \cos(y) & t_z \sin(y) & -t_x \\ -t_y \cos(y) - t_x \sin(y) & t_y \sin(y) & 0 \end{bmatrix} \]

5 linear unknowns $\rightarrow$ linear 5 point algorithm
3 unknowns $\rightarrow$ quartic 3 point algorithm
Challenge: repetition ambiguity

→ result in incorrect correspondences!
Disambiguating visual relations using loop constraints

(Zach et al CVPR'10)
Towards Parsing Urban Scenes

• Detecting symmetries and repetitions (Wu et al ECCV‘10)

• Applications:
  – Extracting architectural grammars
  – Matching repeating structures
  – Shape from symmetry and repetition
Real-Time Stereo Visual SLAM

(Clipp et al., IROS2010)

- Stereo KLT for local motion estimation
- SIFT for feature re-detection and loop closure
- Local and global bundle adjustment
More applications of SLAM

OmniTour
(Saurer et al., 3DPVT2010)

MAVs
PixHawk student team
1st place autonomy EMAV09
(http://pixhawk.ethz.ch/)

Funded with Google award

Marc Pollefeys
Rome on a cloudless day

(Frahm et al. ECCV 2010)

Dense Reconstruction (1h58)

Some numbers

- 1PC
- 2.88M images
- 100k clusters
- 22k SfM with 307k images
- 63k 3D models
- Largest model 5700 images
- Total time 23h53

- GIST & clustering (1h35)

- SIFT & Geometric verification (11h36)

- SfM & Bundle (8h35)
Talk outline

• Introduction
• Object modeling
• Scene modeling
• People/event modeling
• Summary and conclusion
Monocular Articulated Motion and Shape Recovery
(Yan & Pollefeys, CVPR05/ECCV06/CVPR06 & PAMI08)

- Feature tracks of articulated bodies span multiple intersecting 4D linear subspaces (under affine imaging conditions)
- Motion segmentation using local subspace affinity
  - Best in recent comparison (Tron & Vidal, CVPR07)
- Kinematic chain recovery
- Articulated 3D motion and shape recovery

Extension to **multi-camera configurations** where **13D subspaces** are obtained independently of the number of cameras. Points need not be observed in more than one view. Formulation as third order tensor. (Angst & Pollefeys, ICCV09/ECCV10)
Camera network calibration from silhouettes

(Sinha et al., CVPR04; Sinha and Pollefeys ICPRo4/IJCV10)

calibrate—and synchronize—camera network without requiring specific calibration data

Our approach is robust and efficient

http://cs.unc.edu/~ssinha/Research/silcalib/
Dynamic shape from silhouettes

- Unreliable silhouettes: do not make decision about their location
- Do sensor fusion: use all image information simultaneously

(Franco and Boyer, ICCV05)
Bayesian formulation

- Idea: find the content of the scene from images, as a probability grid.
- Modeling the forward problem - explaining image observations given the grid state - is easy. It can be accounted for in a sensor model.
- Bayesian inference enables the formulation of our initial inverse problem from the sensor model.
- Simplification for tractability: independent analysis and processing of voxels.
Visualization

(Franco and Boyer, ICCV05)
Dynamic 3D capture in the real world

(Guan et al., CVPR07; IJCV10)

• Enable capture in real environment with occlusions
  – Robust inference of shape from partially occluded silhouettes
  – Inference of occluder shape from free-space & discrepancies
Occluder shape from incomplete silhouettes: experiments
(Guan et al., CVPR07; IJCV10)
3D tracking of multiple persons

(Guan et al., CVPR08; IJCV10)

• Separate foreground model for each person (GMM trained using EM)

• Multiple grids: (person 1, ..., person n, unmodeled, background)

• Perform 3D analysis using all views, assume people consist of single connected component

• ‘Unmodeled’ foreground class catches shadows, new persons, ...

bench
Occupancy Flow

(Guan et al., CVPR10)

Concept

• Work in 4D space (3D + time)

• Recover the dense flow of the object motion and refine object shape simultaneously

• Possibility for motion segmentation
Modeling dynamic scenes with hand-held cameras

(Taneja et al., ACCV10)
Modeling Dynamic Scenes Recorded With Freely Moving Cameras
Visual Exploration of Casually Captured Events

Unstructured Video-Based Rendering (Ballan et al. SIGGRAPH10)

Input: Casually captured videos (+some photos)

Output: Explore the event using our VBR

Starting Grant 4D Video

http://cvg.ethz.ch/research/unstructured-vbr/
Conclusion

• Possibility to compute shape, motion and appearance from video, as well as camera system calibration

• Challenges:
  – Large-scale scenes
  – Dynamic objects, people in particular, in cluttered scenes

• Opportunities:
  – Advances in camera, processing, network and storage technologies
  – Lots of interesting applications in many different areas
Thank you for your attention!

Questions?