Estimation of 3-d Scene Structure and Motion

Bernd Girod

Image, Video, and Multimedia Systems Group

Information Systems Laboratory

Department of Electrical Engineering

Research Topics: Image, Video, and Multimedia Systems Group

Video Coding Algorithms

- Rate-distortion optimized video compression
- Multiframe prediction
- Error-resilient video coding
- Scalable video coding

Networked Multimedia Systems

- Internet video streaming
- Wireless video
- Voice over IP
- Digital watermarking

3-D Image Analysis and Synthesis

- 3-D motion estimation and structure-from-motion
- Compression of lightfields for image-based rendering
- Facial animation and expression tracking

Research Topics: Image, Video, and Multimedia Systems Group

Video Coding Algorithms

- Rate-distortion optimized video compression
- Multiframe prediction
- Error-resilient video coding
- Scalable video coding

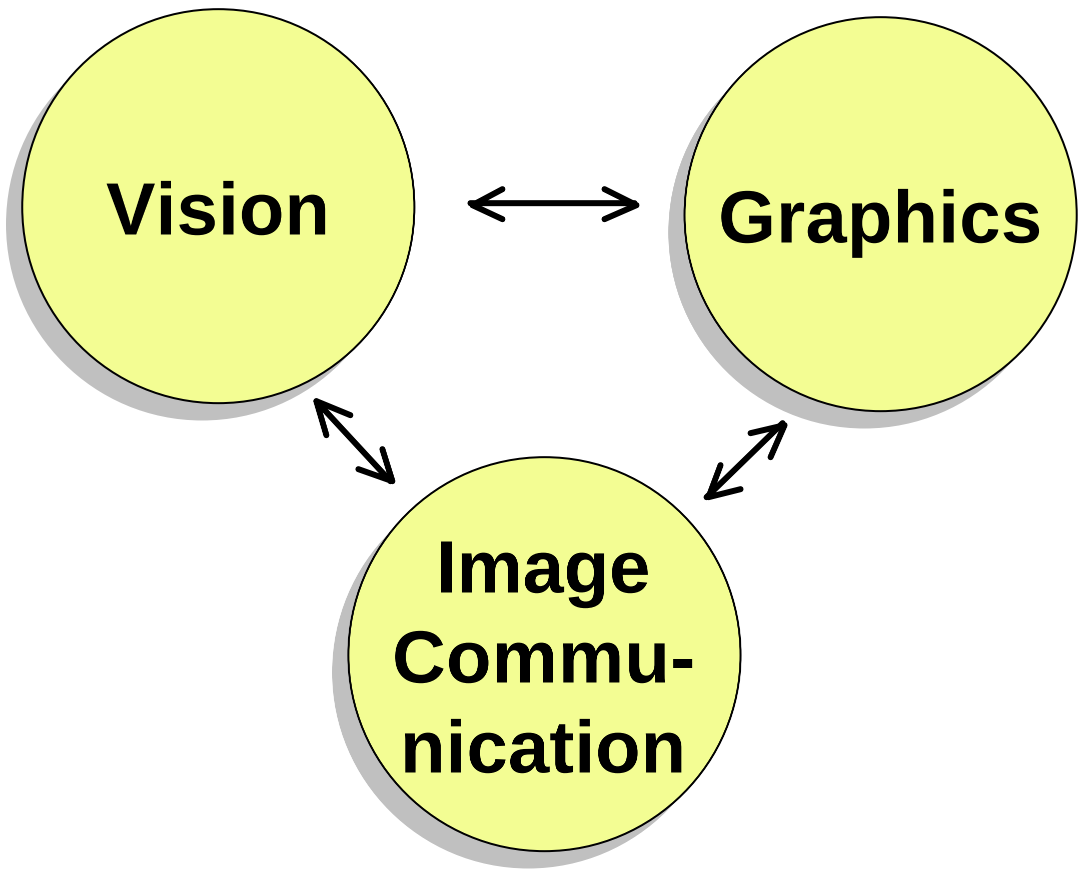
Networked Multimedia Systems

- Internet video streaming
- Wireless video
- Voice over IP
- Digital watermarking

3-D Image Analysis and Synthesis

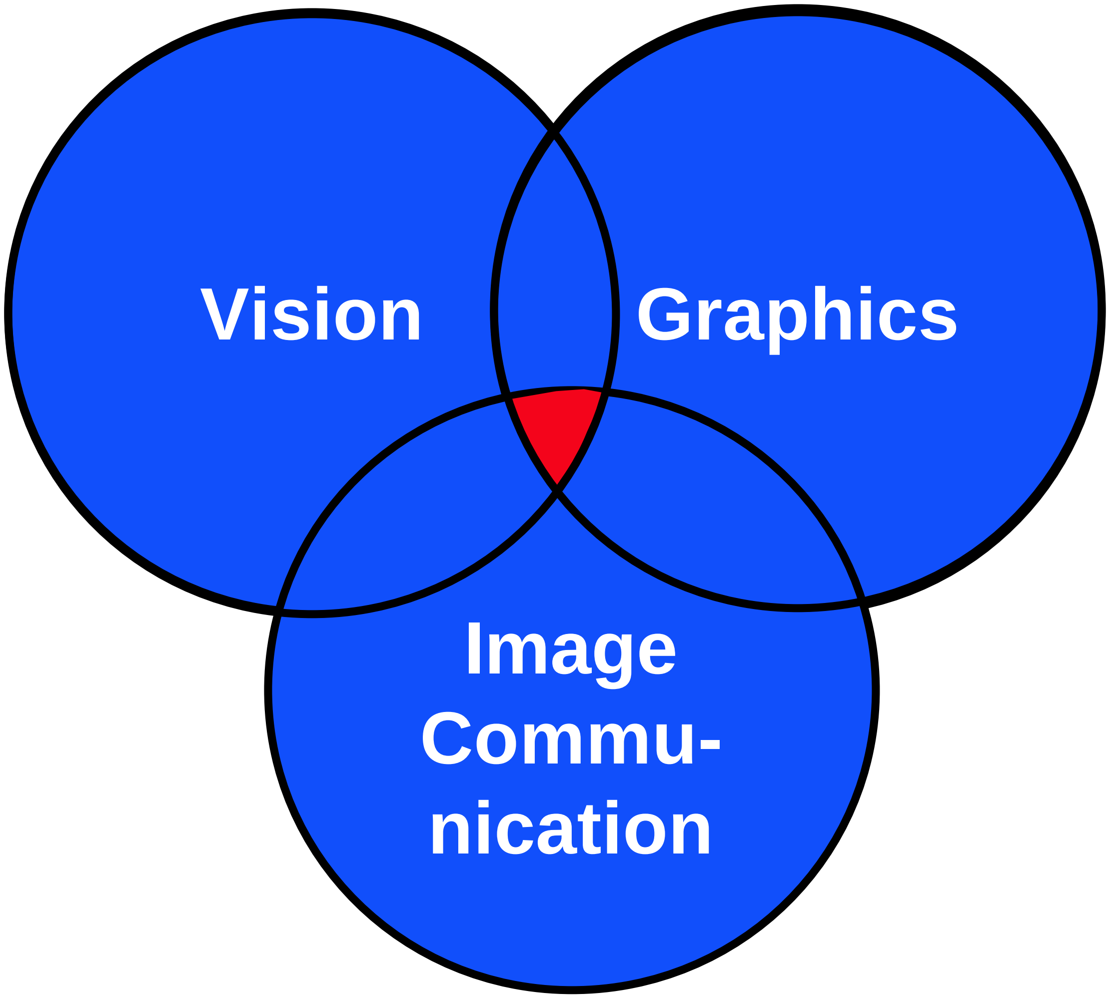
- 3-D motion estimation and structure-from-motion
- Compression of lightfields for image-based rendering
- Facial animation and expression tracking

Vision, Graphics, and Image Communication



1988

Vision, Graphics, and Image Communication

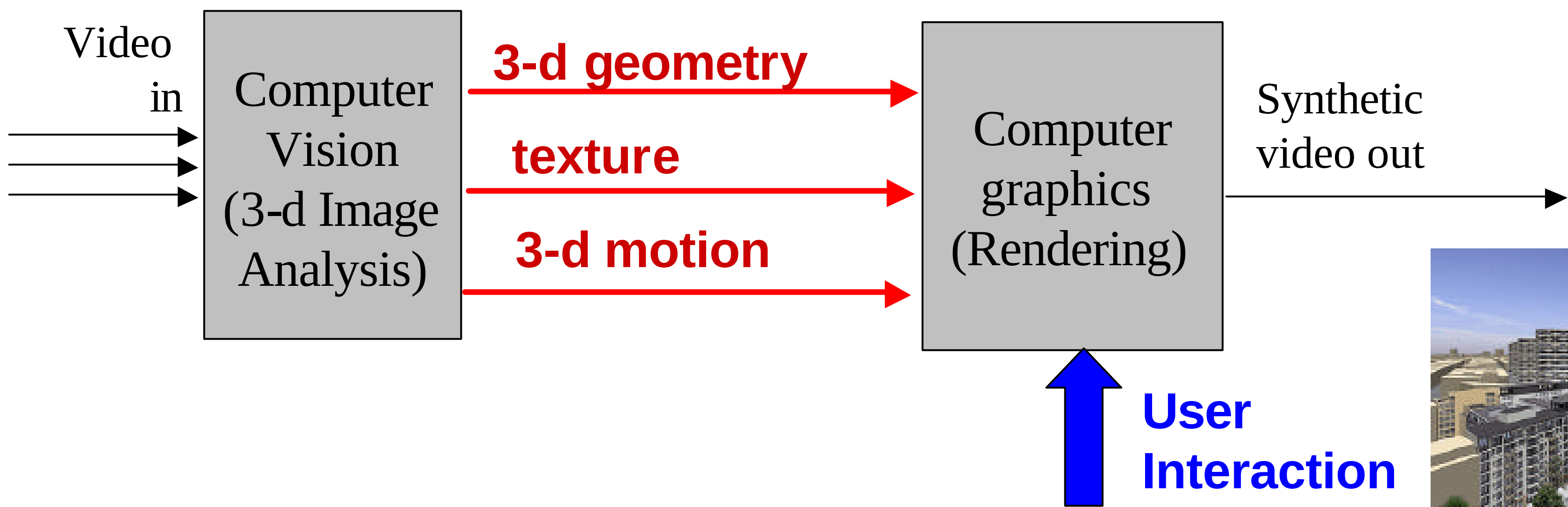


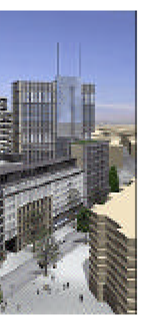
2002

3-D Image Analysis and Synthesis

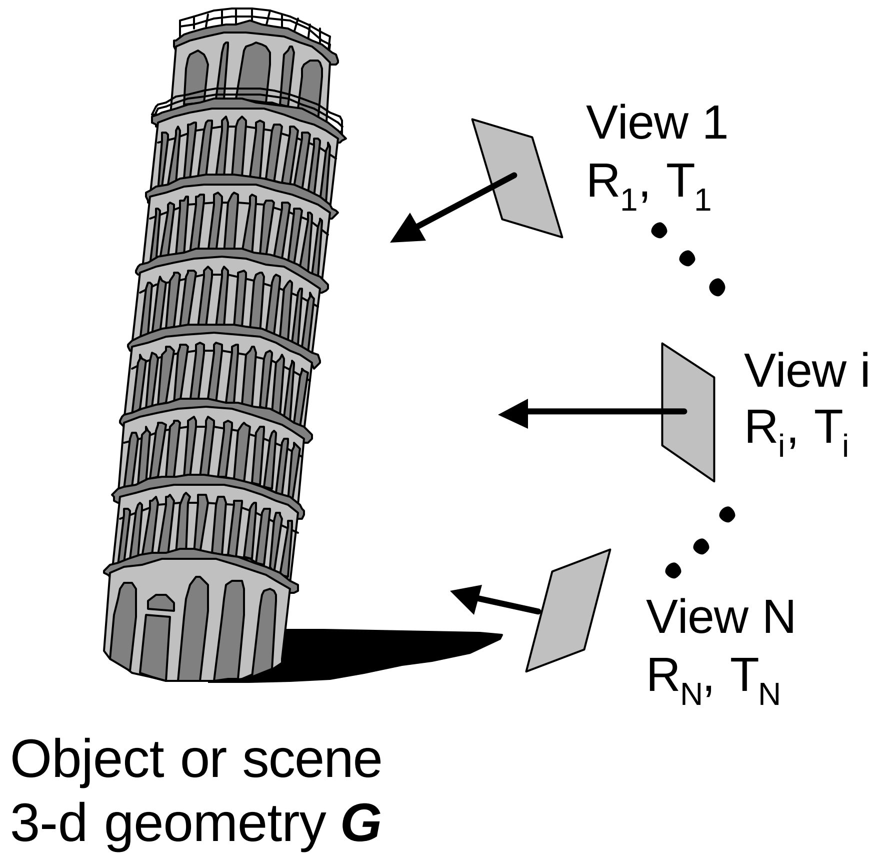
<u>Conjecture</u>

Interactive multimedia systems will make a great leap forward by combining 3-d computer vision and 3-d graphics.





Fundamental Problems of 3-D Image Analysis



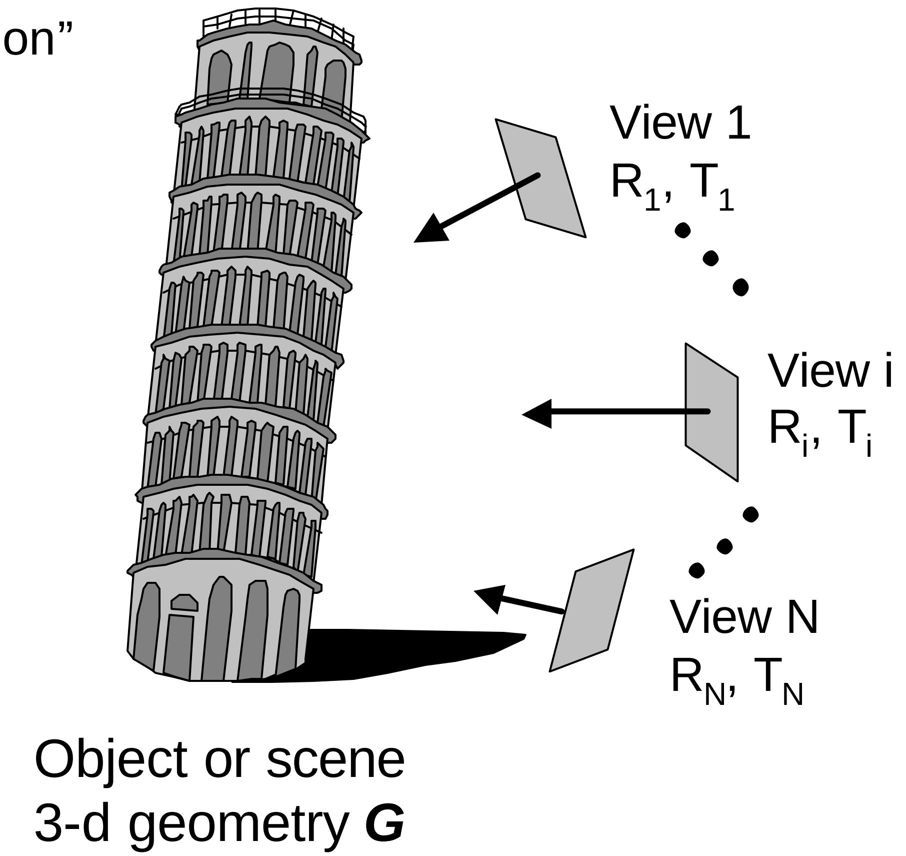
Fundamental Problems of 3-D Image Analysis

Problem 1

"Simultaneous estimation of structure and motion"

"Structure-from-Motion"

G, R_i, T_i unknown



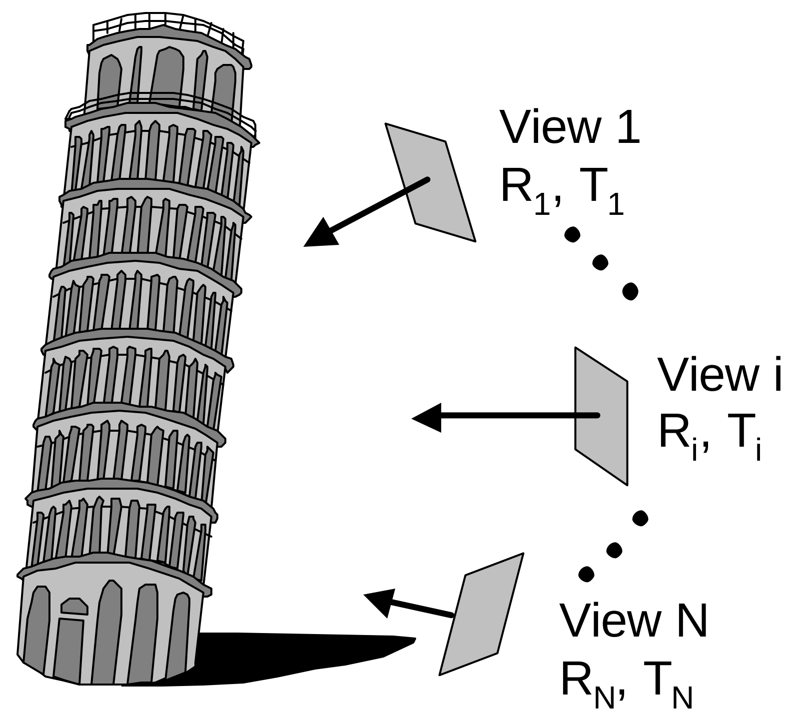
Fundamental Problems of 3-D Image Analysis

Problem 1
"Simultaneous estimation of structure and motion"
"Structure-from-Motion"

G, R_i, T_i unknown

Problem 2
"Model-based 3-d motion estimation"
"Estimation of external camera parameters"

G known, R_i, T_i unknown



Object or scene
3-d geometry ***G***

Fundamental Problems of 3-D Image Analysis

Problem 1
"Simultaneous estimation of structure and motion"
"Structure-from-Motion"

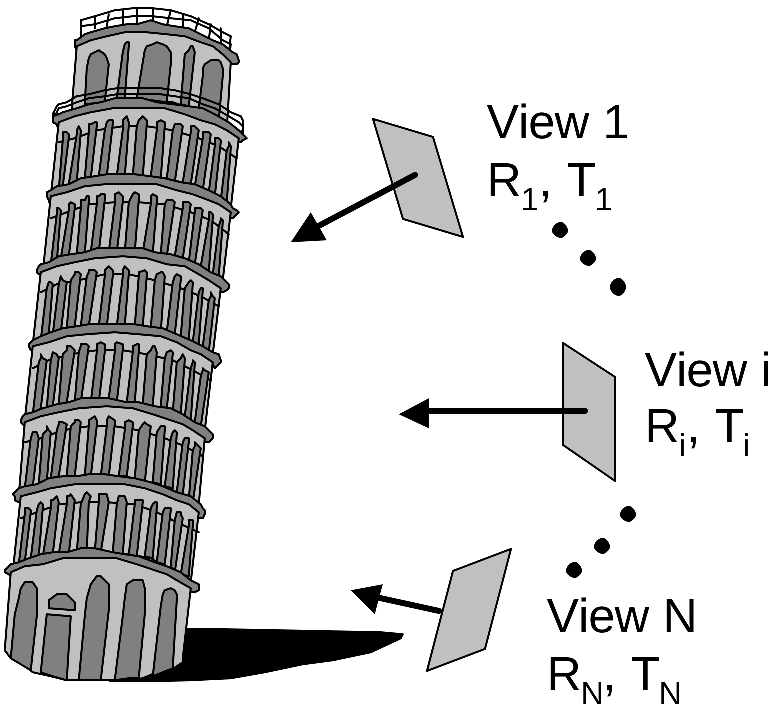
G, R_i, T_i unknown

Problem 2
"Model-based 3-d motion estimation"
"Estimation of external camera parameters"

G known, R_i, T_i unknown

Problem 3
"3-d reconstruction from calibrated views"

G unknown, R_i, T_i known



Object or scene
3-d geometry ***G***

Outline of this talk

- Fundamental problems of 3-d image analysis and synthesis
  - Simultaneous estimation of structure and motion
  - Model-based 3-d motion estimation
  - 3-d reconstruction from calibrated views
- Recent algorithms
- Experimental results
- Application: compression of light-fields

Fundamental Problems of 3-D Image Analysis

Problem 1
"Simultaneous estimation of structure and motion"
"Structure-from-Motion"

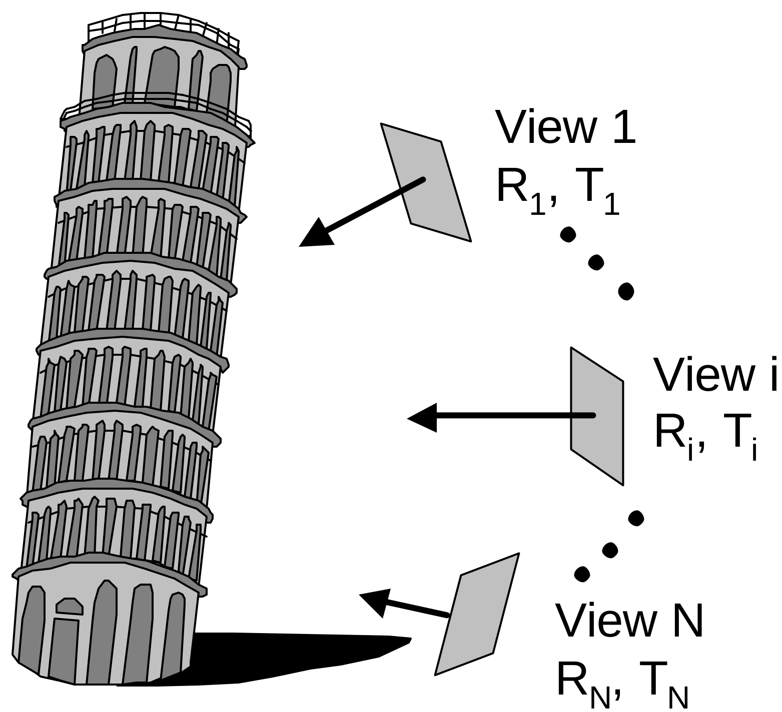
G, R_i, T_i unknown

Problem 2
"Model-based 3-d motion estimation"
"Estimation of external camera parameters"

G known, R_i, T_i unknown

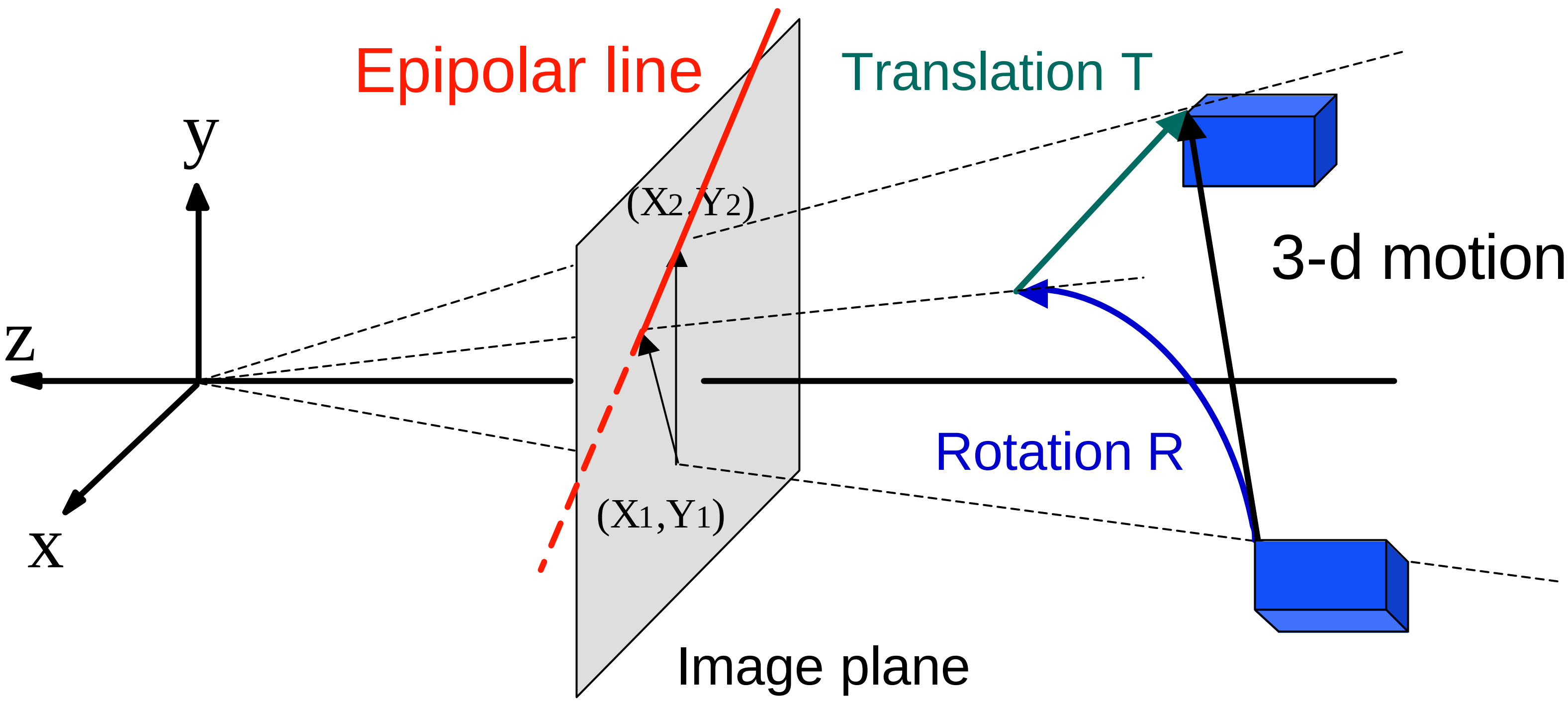
Problem 3
"3-d reconstruction from calibrated views"

G unknown, R_i, T_i known



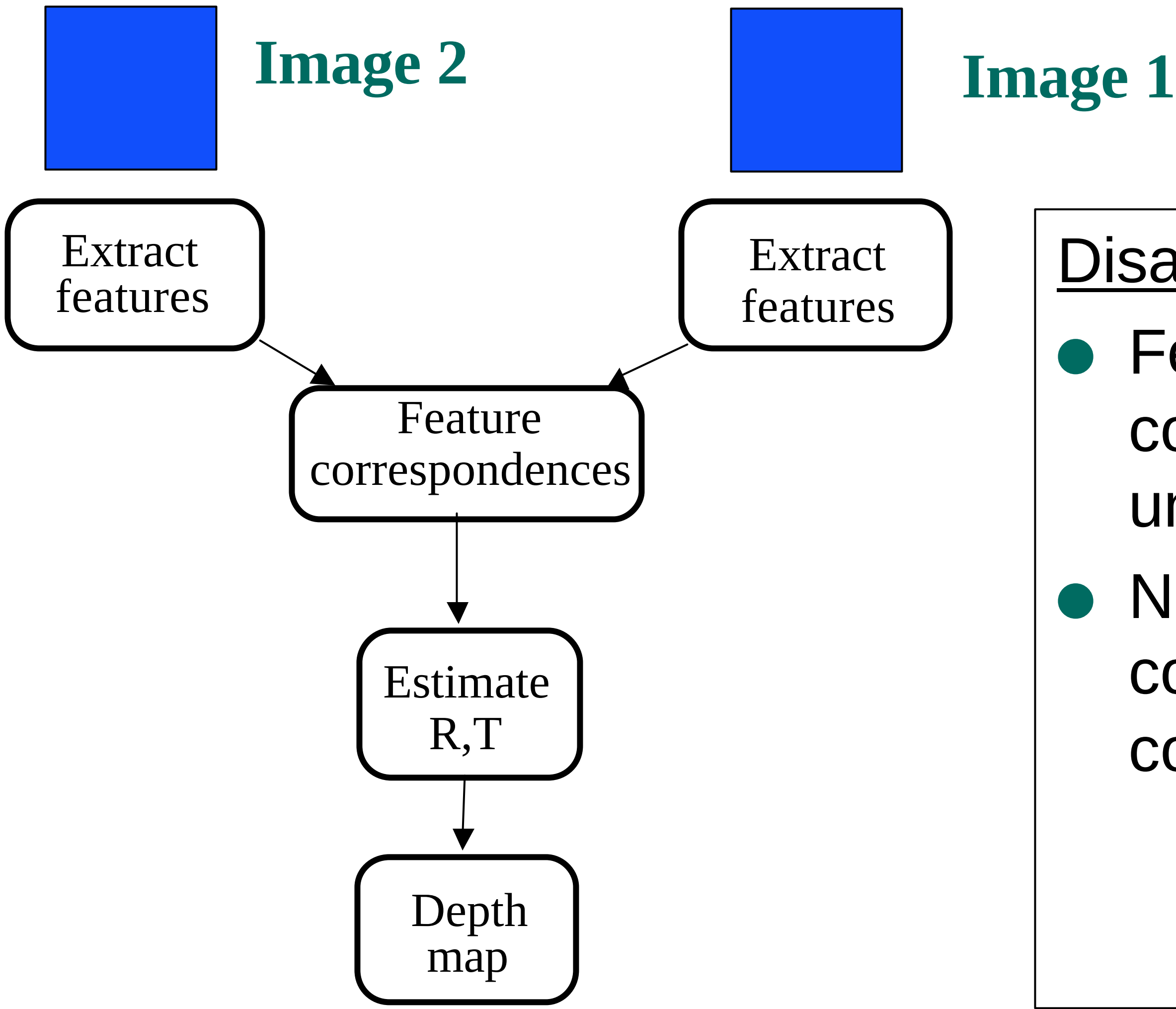
Object or scene
3-d geometry ***G***

Perspective Projection and Epipolar Line



Point correspondences for 3-d rigid body motion must lie on a straight line

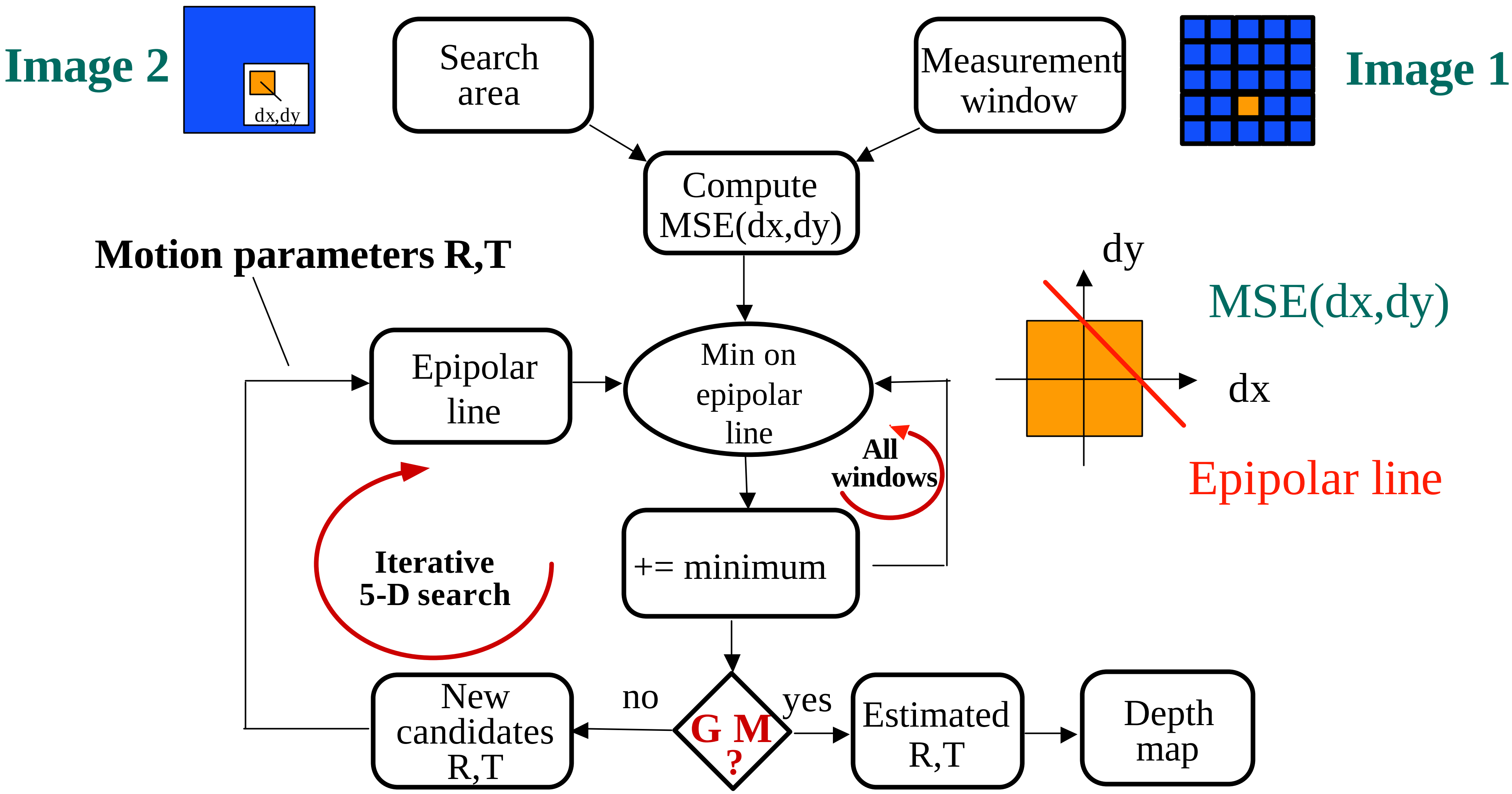
Two-Stage Method



<u>Disadvantages</u>

- Feature extraction / correspondences often unreliable or ambigious
- No rigid-body-motion constraint in feature correspondence stage

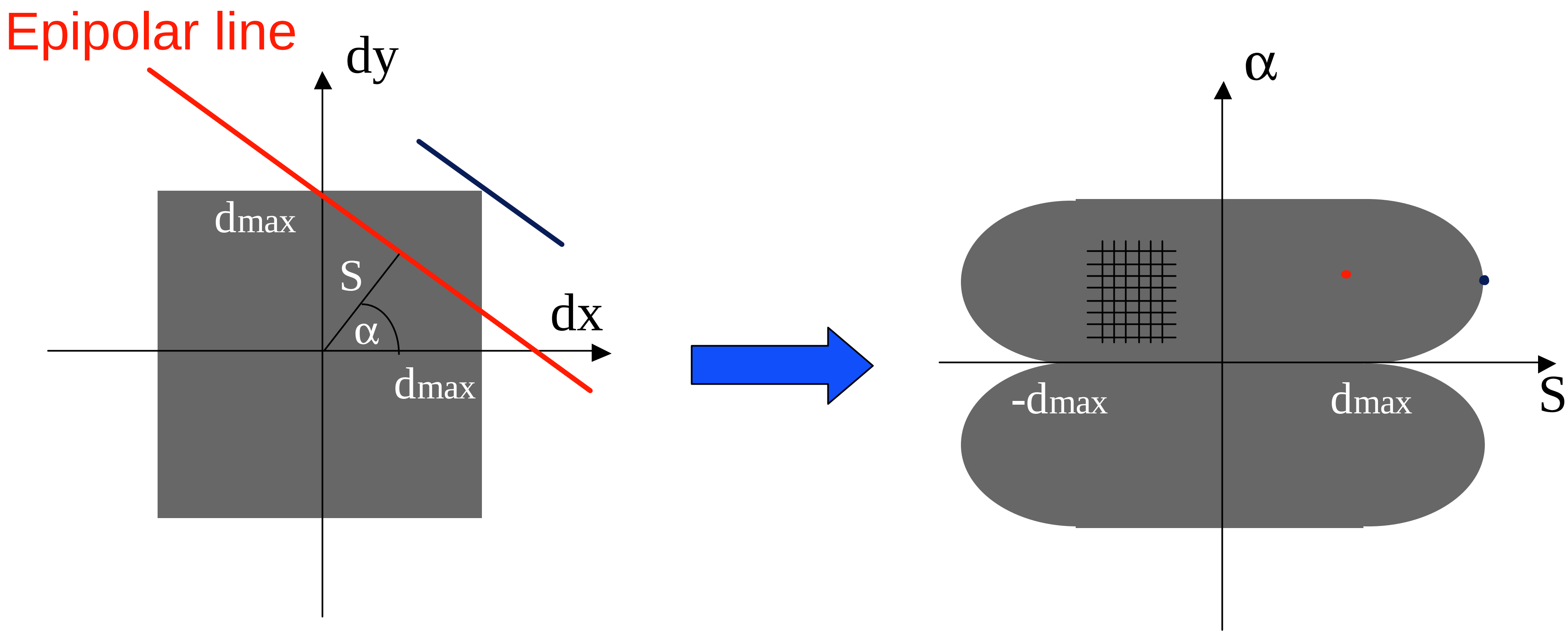
Simultaneous estimation of 3-d structure and motion



[Steinbach, Girod, ICASSP 1996] [Steinbach, Hanjalic, Girod, ICIP 1996]

Pre-computation of minima for all epipolar lines

Displacement space | **Line space**



$$S = dx \cos(\alpha) + dy \sin(\alpha)$$

Example

Image 1

Image 2

Rigid body motion

R, T

Depth map

3-d mosaicing with depth-based segmentation

Image 0

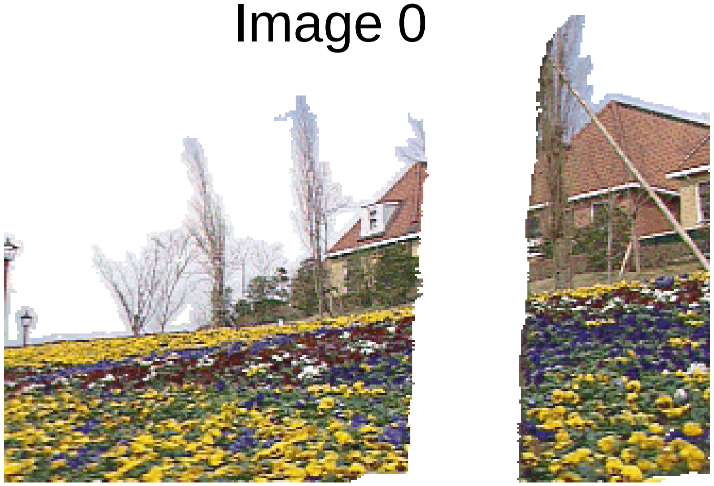


Image 8

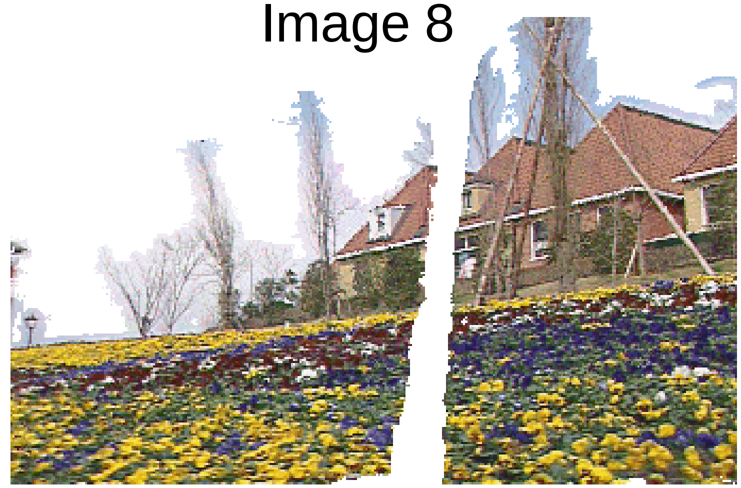
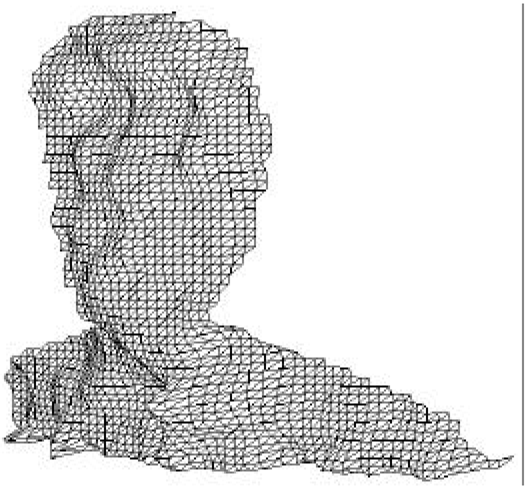


Image 16

Image 24

[Steinbach, Eisert, Girod, Signal Processing, 1998]

3-d motion-based segmentation



[Steinbach, Eisert, Girod, Signal Processing, 1998]

Fundamental Problems of 3-D Image Analysis

Problem 1
"Simultaneous estimation of structure and motion"
"Structure-from-Motion"

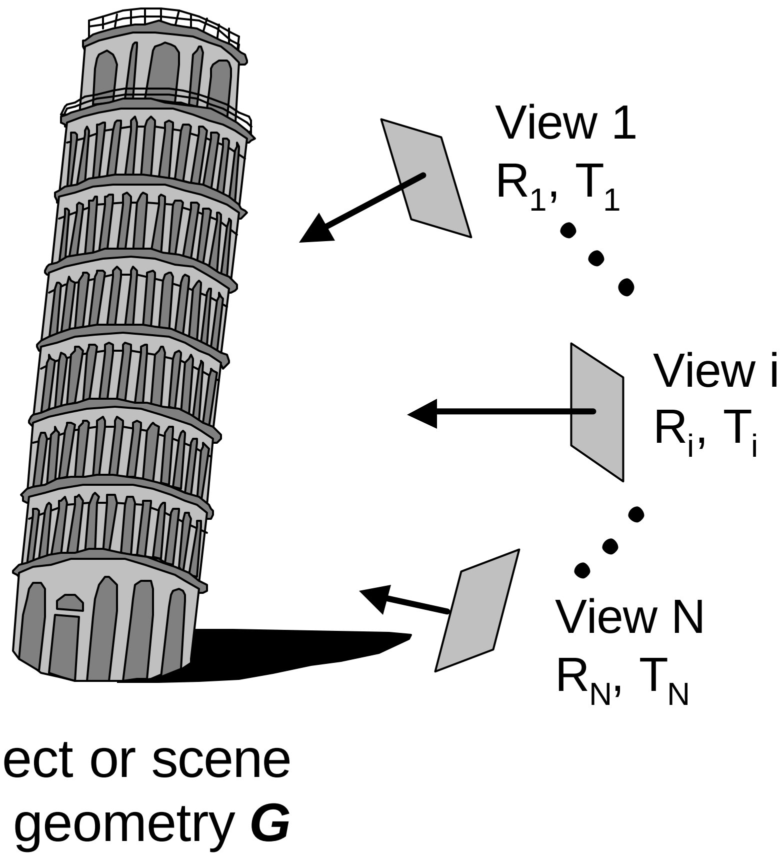
G, R_i, T_i unknown

Problem 2
"Model-based 3-d motion estimation"
"Estimation of external camera parameters"

G known, R_i, T_i unknown

Problem 3
"3-d reconstruction from calibrated views"

G unknown, R_i, T_i known



3-d motion estimation for known geometry

$$\vec{d} = f(R,T,G)$$

Displacement field between $I_1(x,y)$ and $I_2(x,y)$

Linearization for small R,T

$$\vec{d} \approx f_1 \cdot r_x + f_2 \cdot r_y + f_3 \cdot r_z + f_4 \cdot t_x + f_5 \cdot t_y + f_6 \cdot t_z$$

Spatially varying "basis functions"

$$\frac{1}{2}\vec{d}^T \cdot \begin{pmatrix} \frac{\partial I_1}{\partial x} + \frac{\partial I_2}{\partial x} \\ \frac{\partial I_1}{\partial y} + \frac{\partial I_2}{\partial y} \end{pmatrix} \approx I_1 - I_2$$

Assume same brightness of corresponding points *"Optical flow constraint"*

- Solve by linear regression
- Apply iteratively in a resolution pyramid

Extension to flexible bodies

$$\vec{d} = f(R,T,G(\vec{p}))$$

Parametric geometry

Linearization for small *R,T,p*

Spatially varying "basis functions"

$$\vec{d} \approx f_1 \cdot r_x + f_2 \cdot r_y + f_3 \cdot r_z + f_4 \cdot t_x + f_5 \cdot t_y + f_6 \cdot t_z + f_7 \cdot p_1 + f_8 \cdot p_2 + \cdots$$

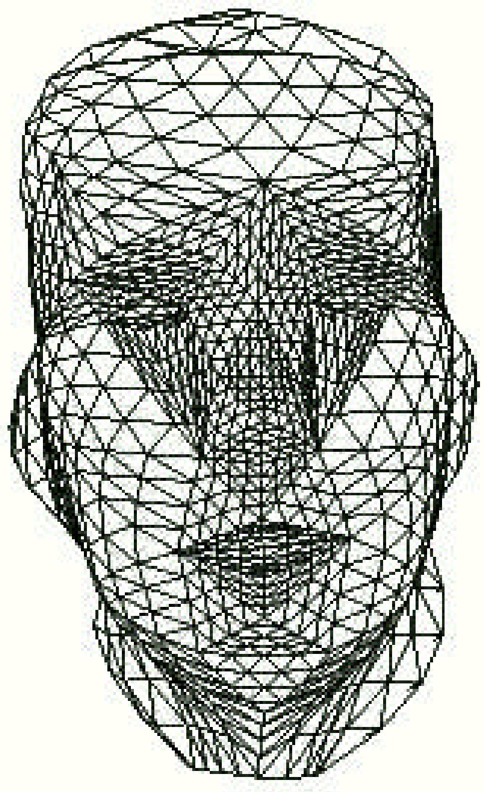
$$\frac{1}{2}\vec{d}^T \cdot \begin{pmatrix} \frac{\partial I_1}{\partial x} + \frac{\partial I_2}{\partial x} \\ \frac{\partial I_1}{\partial y} + \frac{\partial I_2}{\partial y} \end{pmatrix} \approx I_1 - I_2$$

Assume same brightness of corresponding points *"Optical flow constraint"*

- Solve by linear regression
- Apply iteratively in a resolution pyramid

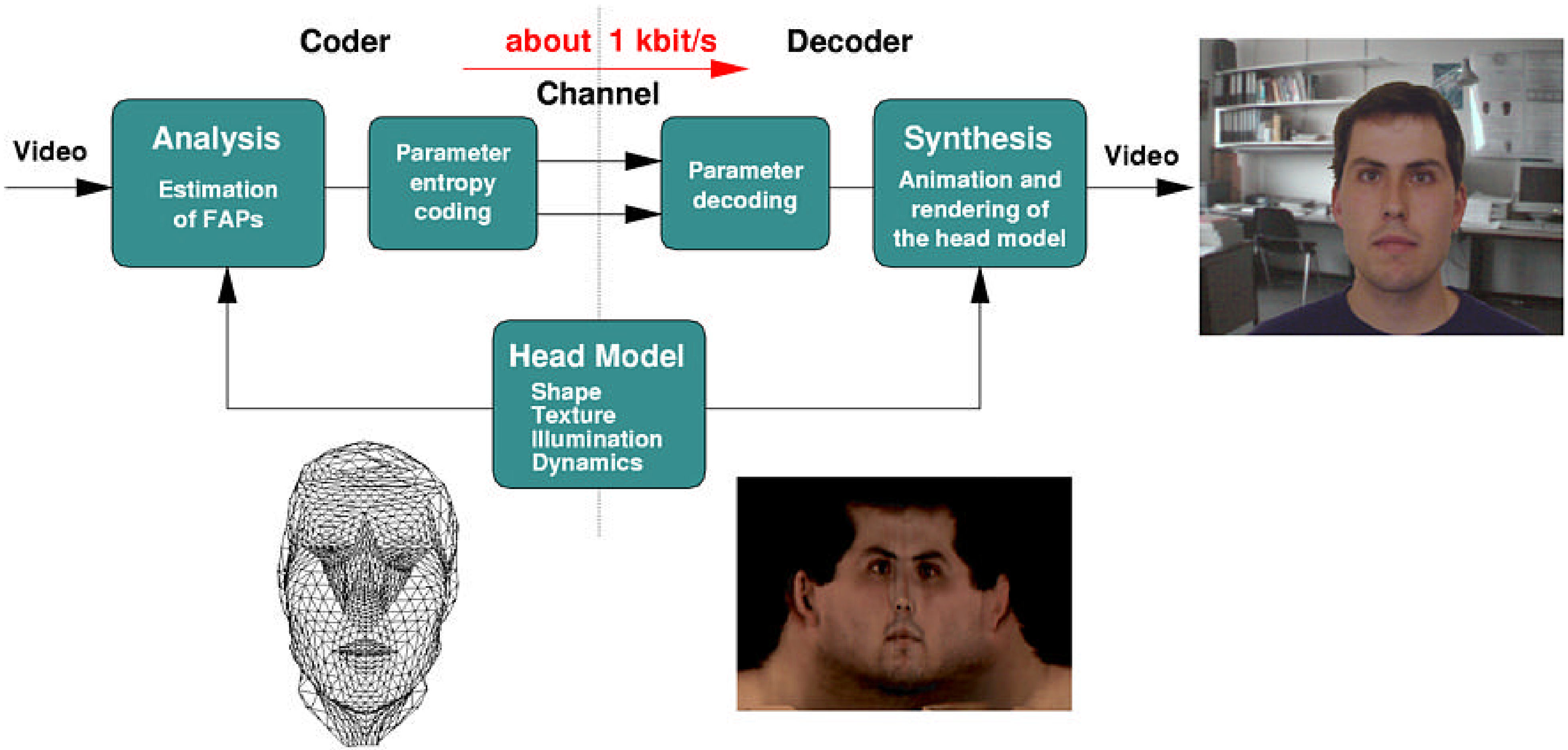
[Eisert, Girod, ICIP 1997] [Eisert, Girod, IEEE CGA, 1998]

Modeling of Facial Expressions



- Head geometry composed of 101 triangular B-spline patches
- Facial expressions by superposition of 66 FAPs (Facial Animation Parameters) according to MPEG-4 standard
- FAPs act on control points of triangular B-spline patches

Model-based videophone



Results: Peter

Original

Synthesized

Sequence: Peter, 230 frames,
CIF resolution, 25 fps

1.2 kbps - 32.8 dB PSNR

Results: Eckehard

Original

Synthesized

Sequence: Eckehard
CIF resolution, 25 fps

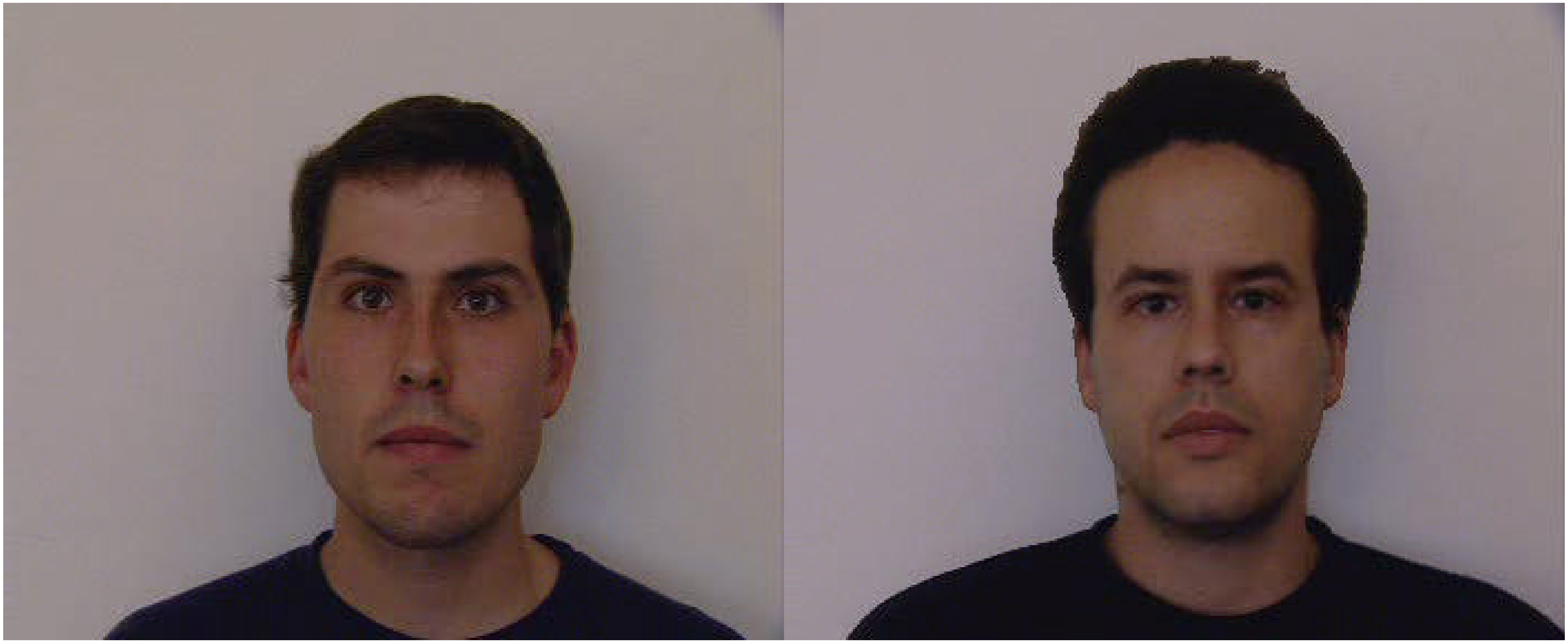
1.1 kbps, 32.6 dB PSNR

Results: Michelle

Original　　　　Synthesized

Results: Peter as Eckehard

Original | Synthesized



Sequence: Peter, 230 frames,
CIF resolution, 25 fps

Results: Eckehard as Peter

Original

Synthesized

Sequence: Eckehard
CIF resolution, 25 fps

Results: Peter as Akiyo

Original

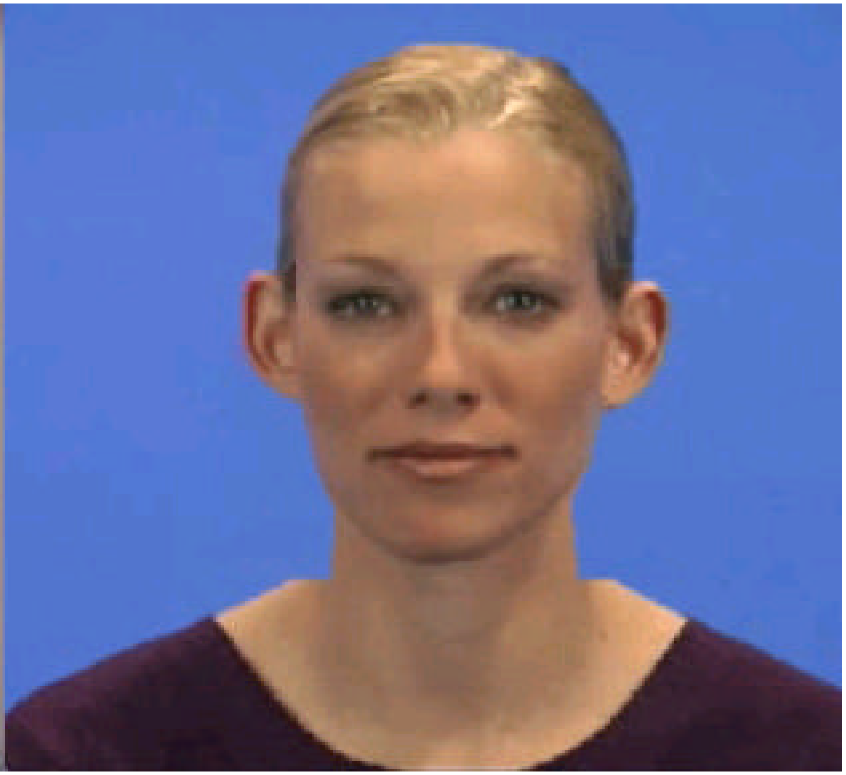
Synthesized

Sequence: Peter, 230 frames,
CIF resolution, 25 fps

Results: Peter as Michelle

Original

Synthesized



Sequence: Peter, 230 frames,
CIF resolution, 25 fps

Fundamental Problems of 3-D Image Analysis

Problem 1
"Simultaneous estimation of structure and motion"
"Structure-from-Motion"

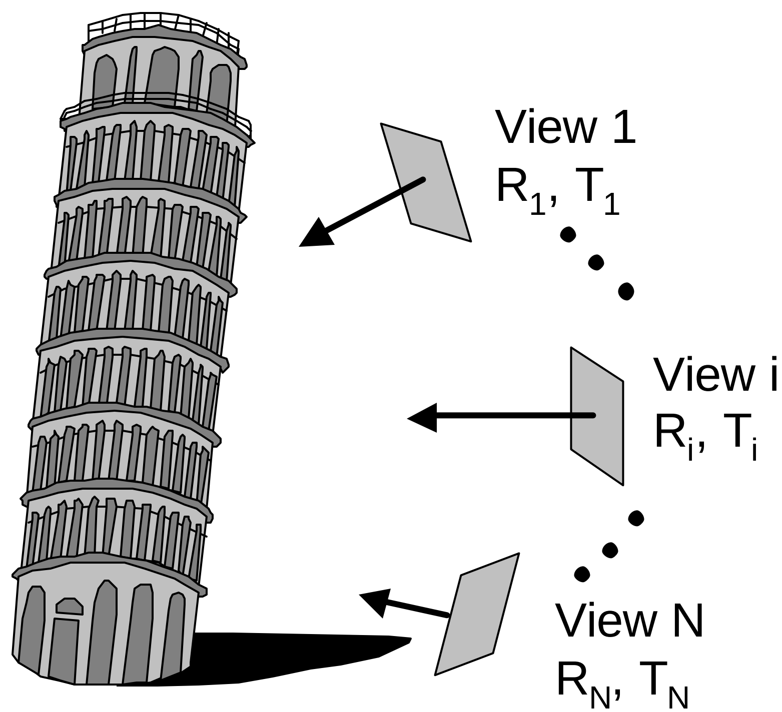
G, R_i, T_i unknown

Problem 2
"Model-based 3-d motion estimation"
"Estimation of external camera parameters"

G known, R_i, T_i unknown

Problem 3
"3-d reconstruction from calibrated views"

G unknown, R_i, T_i known



Object or scene
3-d geometry ***G***

3-D reconstruction from calibrated views: state-of-the-art

- Stereo Methods
  - Depth maps for image pairs ($2^1/_2$-d)
  - Occlusion problem
  - Extension to > 2 views??
  - Good: textured surfaces, parallel to image plane
  - Bad: Depth discontinuities, object silhouette
- Silhouette Methods
  - Backprojection of object silhouettes from many views into 3-space
  - Intersection of backprojected silhouette cones: “Visual hull” approximates object surface
  - Texture not exploited

Geometry Reconstruction from Many Views

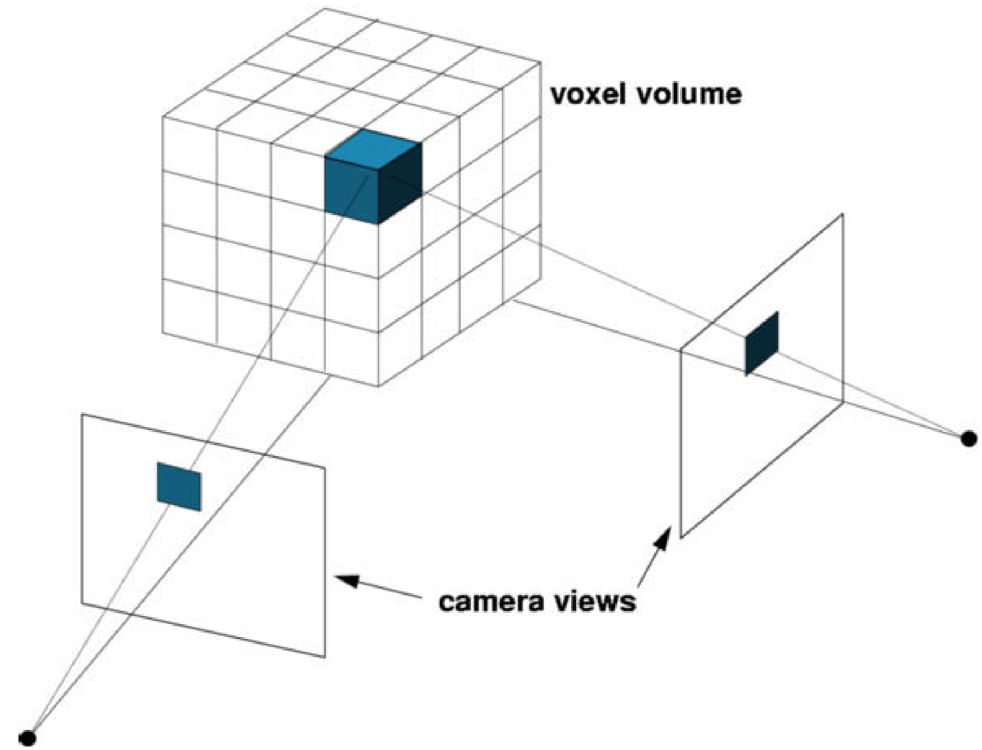
Volumetric Reconstruction

- Subdivide object's bounding box into voxels
- Generation of multiple hypotheses for each voxel
- Hypothesis elimination by projecting visible voxels into all views
- Iterate over all voxels until remaining hypotheses are "photo-consistent"

➡ processes all views simultaneously

➡ exploits texture and silhouette information

➡ yields solid 3-D voxel model



[Eisert, Steinbach, Girod, ICASSP 99]
[Steinbach, Girod, Eisert, Betz, ICIP 2000]

Example

- 11 calibrated views, 352x288 pixels each
- Voxel array: 240 x 240 x 140
- 3.6 $*10^7$ hypotheses generated
- Consistency test: 15 iterations through volume
- Result: 6.8 $*10^4$ visible voxel

Original and Reconstructed Views

Original

Reconstructed for same pose

Interpolated Views

Reconstructed view, not contained in original data set

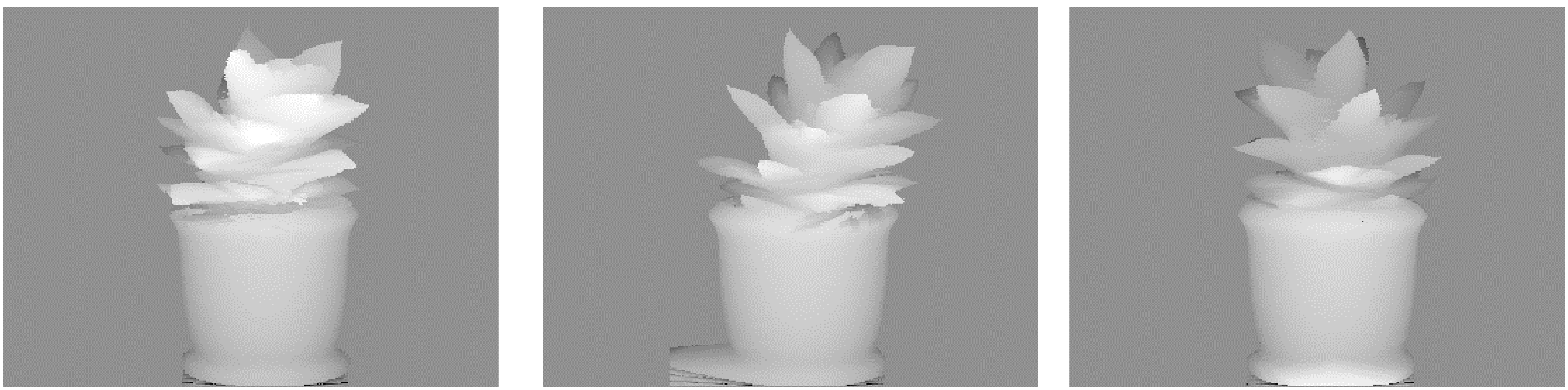
Detail

original

reconstructed

3-D Reconstruction from Many Calibrated Views

Sequence of original camera frames: 15 degree increments



rendered depth maps for the same viewing positions

Problem 1 Revisited: Many Views

Problem 1
"Simultaneous estimation of structure and motion"
"Structure-from-Motion"

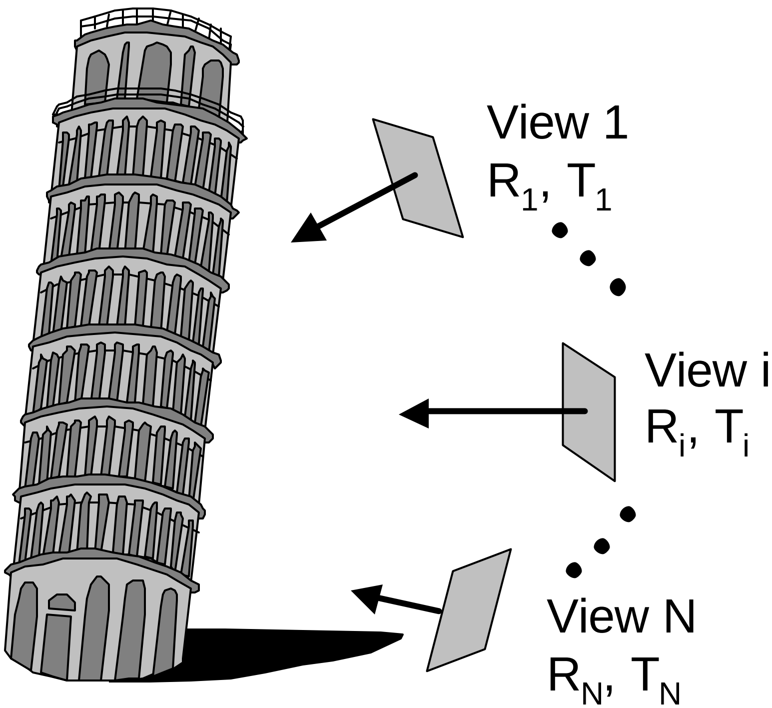
G, R_i, T_i unknown

Problem 2
"Model-based 3-d motion estimation"
"Estimation of external camera parameters"

G known, R_i, T_i unknown

Problem 3
"3-d reconstruction from calibrated views"

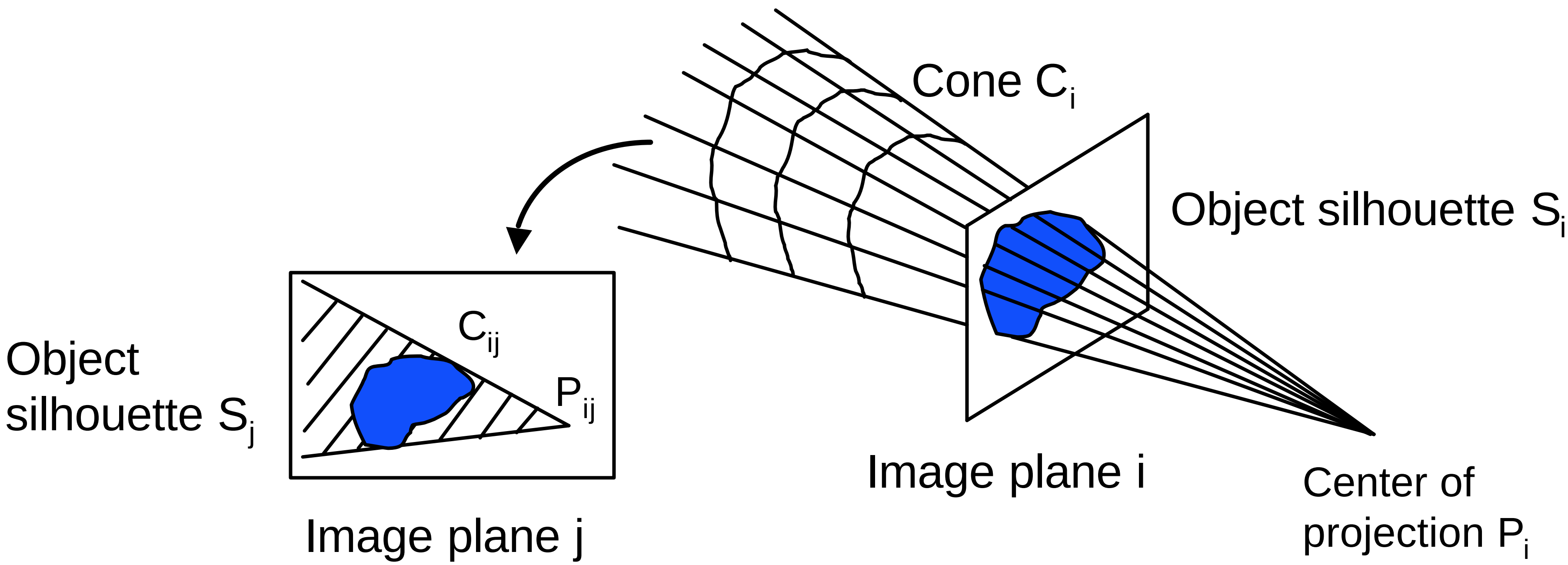
G unknown, R_i, T_i known



Object or scene
3-d geometry ***G***

View Calibration Using Silhouettes

- Exploit mutual consistency in pairs of views



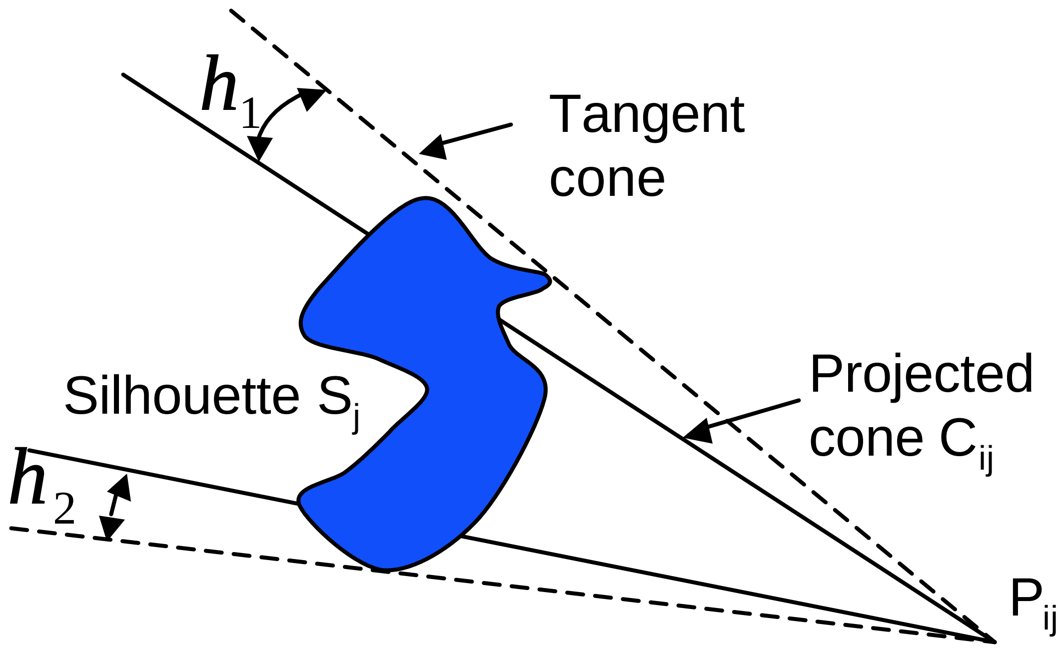
[Ramanathan, Steinbach, Girod, VMV 2000]

Error Measure

- Incorrect calibration parameters lead to difference between tangent and projected 2-D cone

$$e_{ij} = h_1 + h_2$$

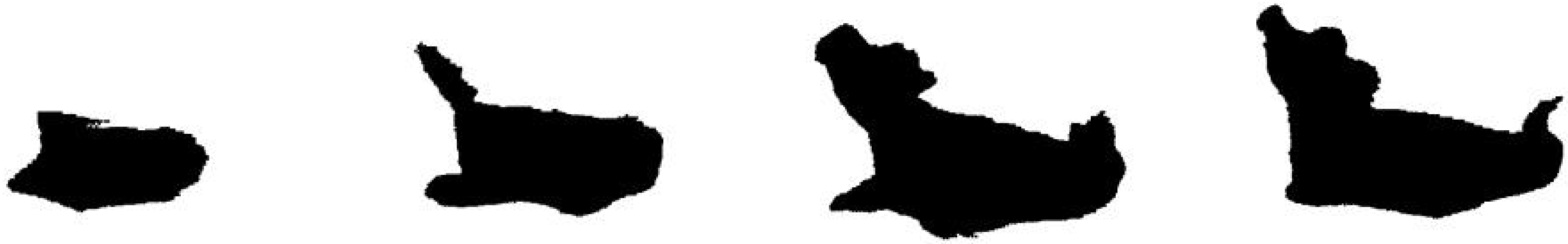
$$E = \sum_{i=1}^{N} \sum_{j=1}^{N} e_{ij} \rightarrow \min.$$



Experimental Results

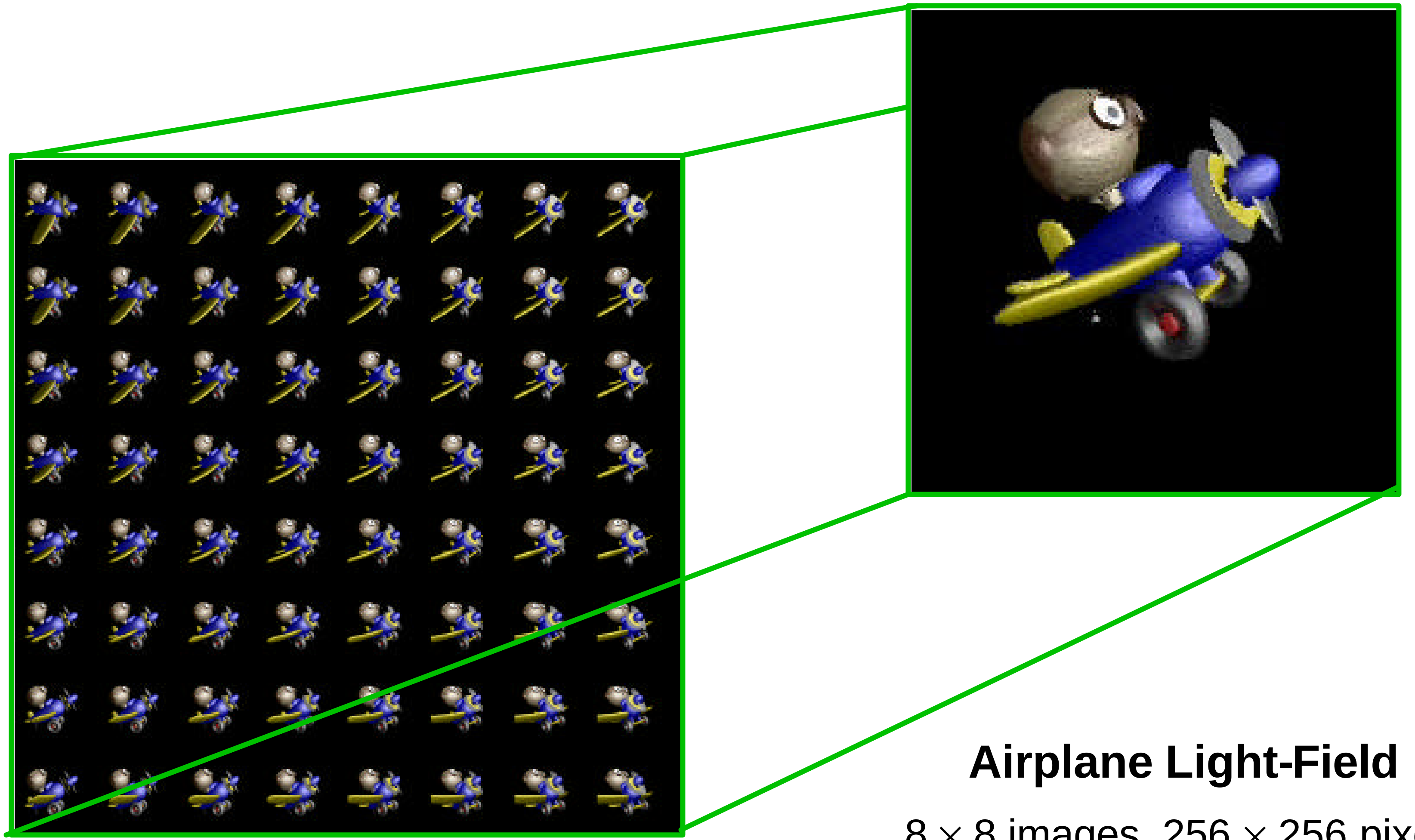
- 32 views from a light-field
- Constrained turntable arrangement
- Translation parameter perturbed
- Projected silhouette of the reconstructed object shown for different stages of the algorithm

Original light-field image



Original uncalibrated parameters — 3 iterations — 7 iterations — Final reconstruction

Image-based Rendering Using Light-Fields



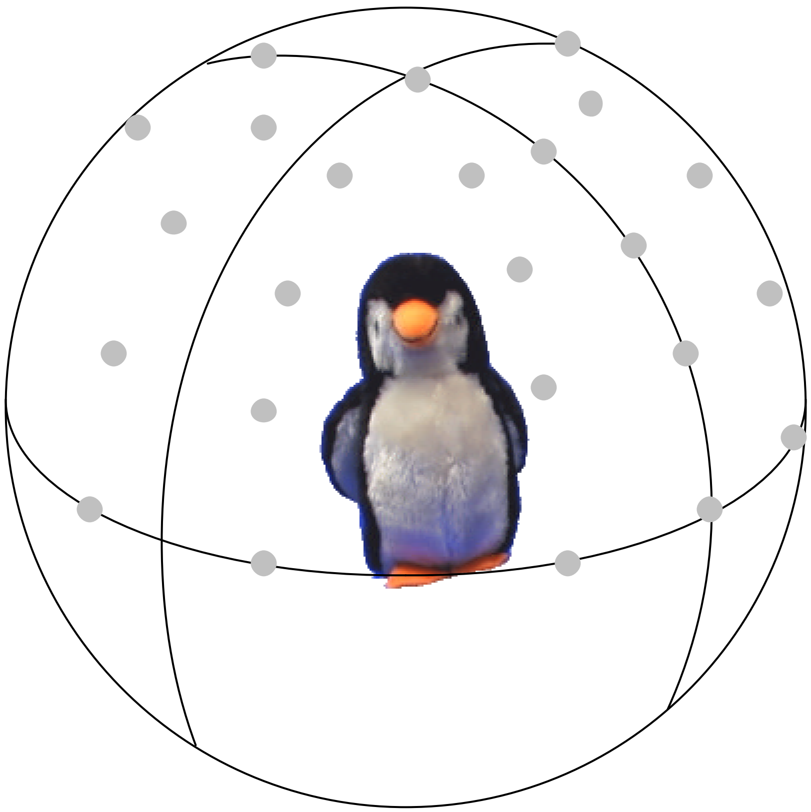
Airplane Light-Field

8×8 images, 256×256 pixels

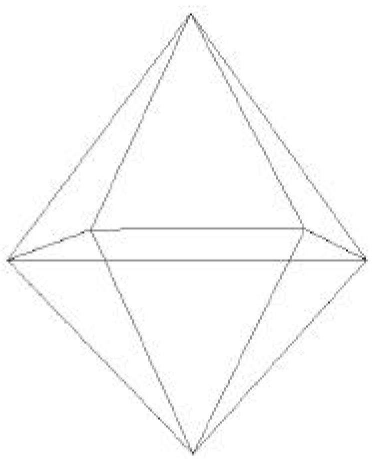
12.6 MByte

Spherical Recording Geometry

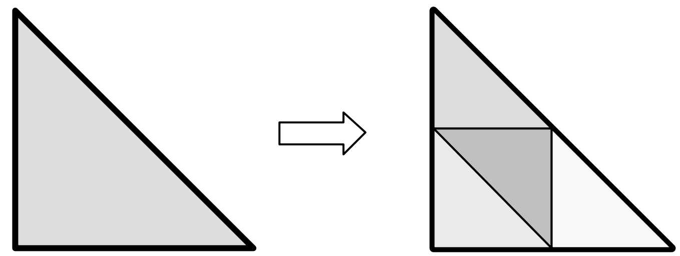
- Calibrated computer-controlled camera mount & turn-table
- 3 test light fields consisting of 32 x 8 calibrated images



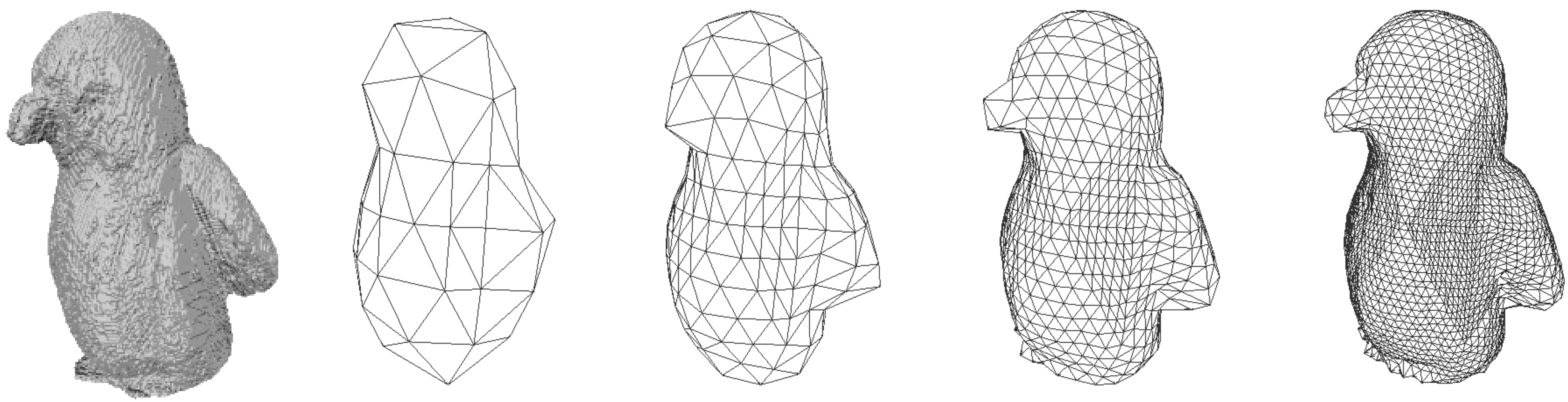
Surface Representation



- Initial octahedral geometry
- Geometry refinement
  - determine vertex normals
  - move vertices to model surface
  - subdivide triangles



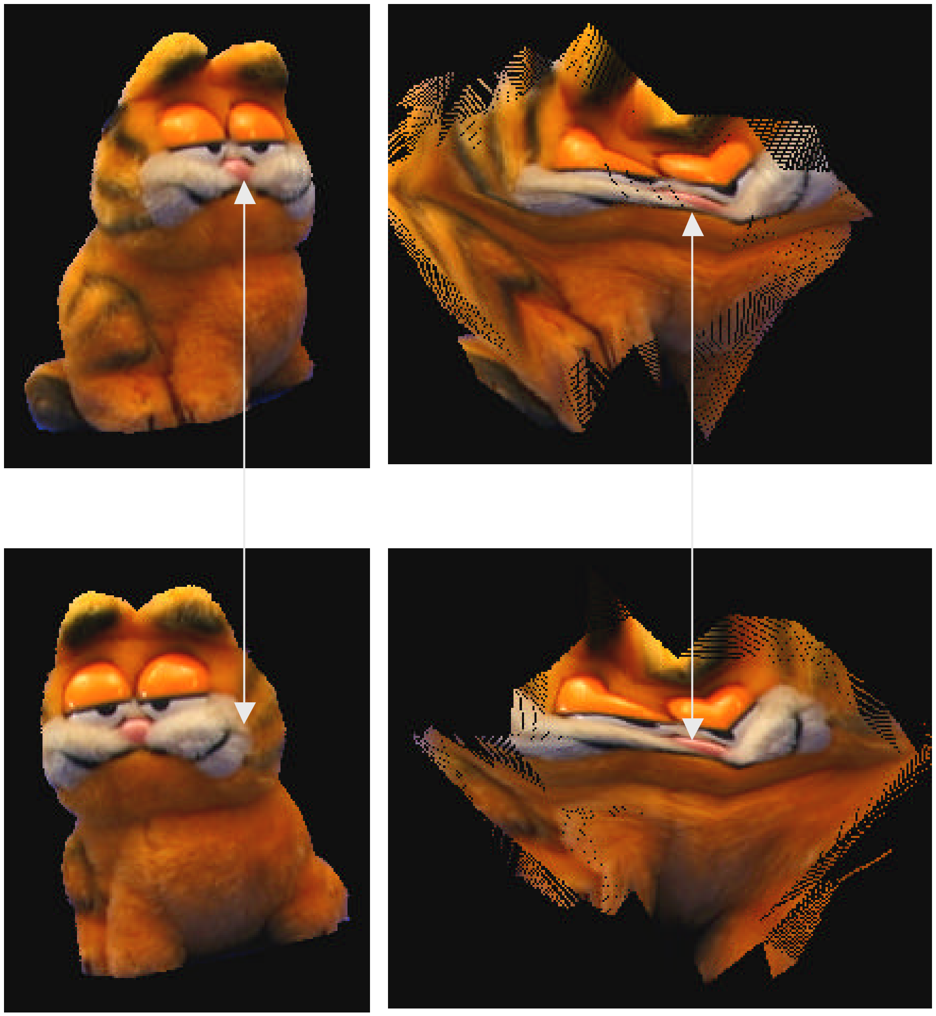
- Encode with Embedded Mesh Coder *[Magnor, Girod, VMV'99]*



voxel model   128 triangles   512 triangles   2048 triangles   8192 triangles

View-dependent texture-map coder

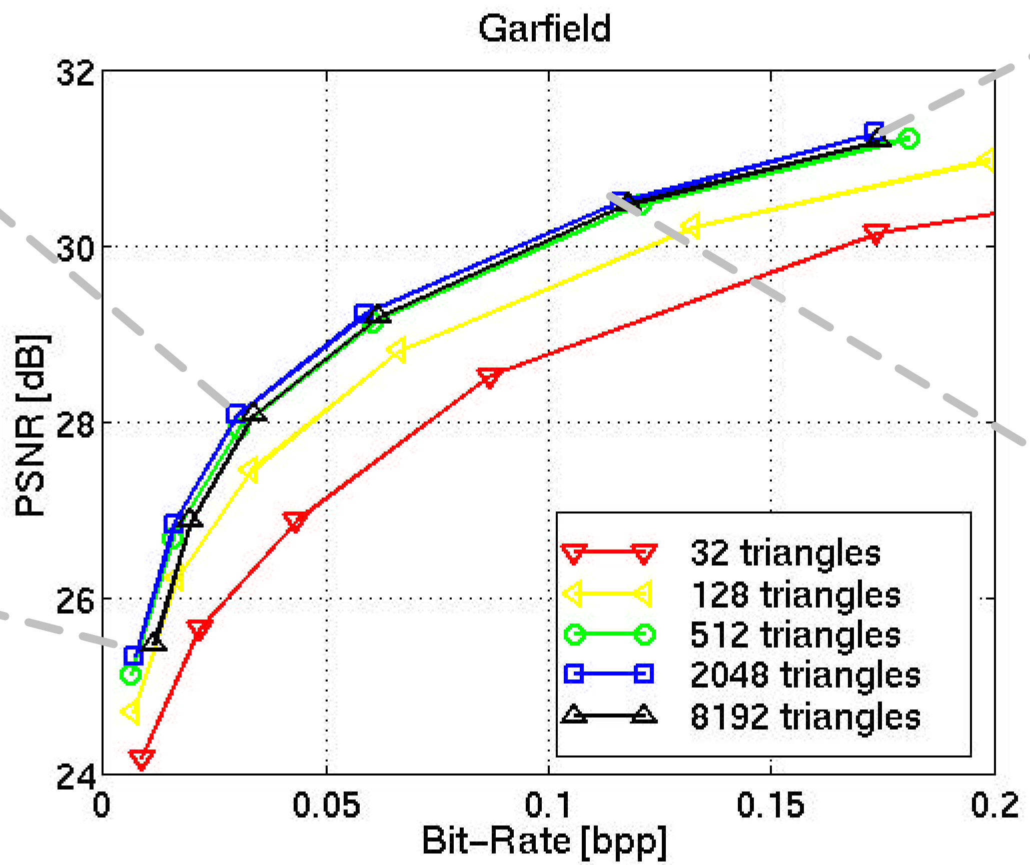
- Warp each image into a texture map
- Arrange texture maps in a 2-d array
- 4-d Haar wavelet decomposition of texture maps
- Quantization and encoding of wavelet coefficients using a 4-d extension of the Set Partioning in Hierarchical Trees (SPIHT) algorithm



[Magnor, Girod, VCIP 2000] [Girod, Magnor, ICIP 2000]

Results: Model-based Coder

Reconstruction quality in *luminance PSNR* (dB)



Conclusions

- Recent algorithms to recover 3-d motion and/or geometry
  - New direct method for structure-from-motion overcomes limitations of two-stage approach
  - Robust model-based motion estimator, extended to non-rigid motion
  - Example: facial expression tracking, videophone at 1kbps
  - Volumetric reconstruction method processing many views simultaneously
- Application example: light-field compression
  - View-dependent texture mapping, 4-d embedded wavelet coder
  - Compression ratios 100…1000:1

Vision, graphics, and image communication are converging!