İTÜ

3D Vision
BLG 634E
Spring 2022

Lecture: Introduction to 3D Vision

Professor: Gozde UNAL

Some material courtesy of :
- Greg Slabaugh @ Queen's Mary University, London
- Stanford University CS231A Lectures
- Angjoo Kanazawa CS294
- S Seitz "3D Computer Vision: Past, Present, and Future", 2012.

1

Computer Vision

⇨ **Computer vision** is a scientific discipline that studies how computers can efficiently perceive, process, and evaluate visual data such as images and video in order to understand the surroundings, objects, scenes, actions, etc…

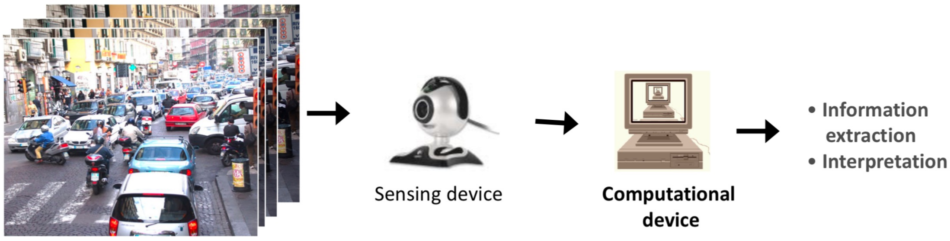
Artificial vision:

- potential of relieving humans of tasks that are
  - dangerous,
  - monotonous, boring or unnecessary time consuming such as driving/self driving cars; surveillance
  - infeasible to quickly process and sort through, extract info from big visual data
- Replace lost vision skills or augment capabilities
- Automate tasks such as navigation, obstacle avoidance, recognition, etc.

Alan Turing's remarks in 1950's, 'Computing Machinery and Intelligence', Mind: The problem was assumed to be easier in the beginning and **only processing power and limited storage was regarded as the major hurdle.**

2

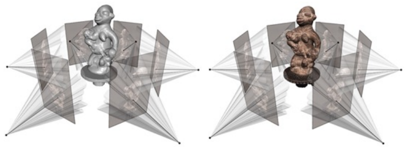
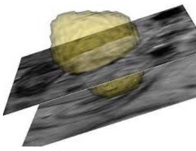
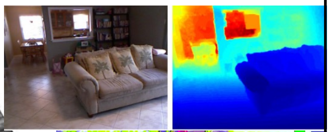
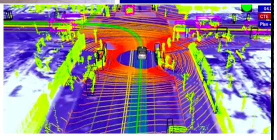
Computer Vision



1. **Information extraction:** features, 3D structure, motion flows, etc…
2. **Interpretation:** recognize objects, scenes, actions, events

What is 3D Computer Vision?

3D computer vision can be described as *extraction of 3D information from digital images, and 3D models*

Computer Vision vs Computer Graphics

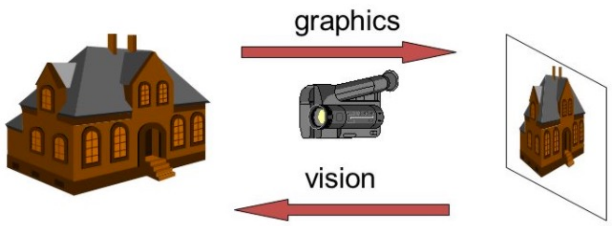


Image formation: how objects give rise to images: 3D to 2D (loss of information)

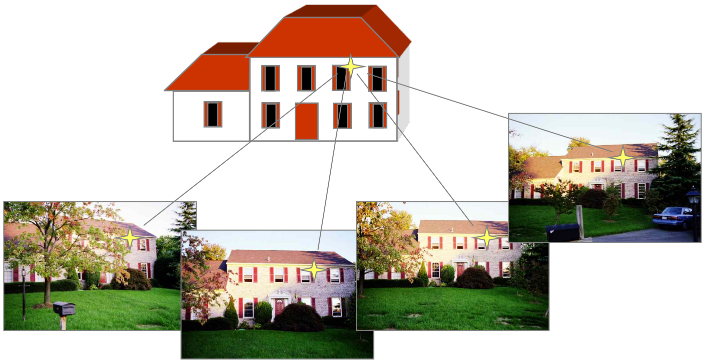
Computer Vision: Inverse problem of image formation: 2D to 3D
uses images to recover a description of objects in space
Ill-posed since we lose 1 dimension coming from 3D world to 2D images

5

3D Vision– The Fundamental Problem

Problem: how to recover the 3-D geometry of the scene?
Q: What makes this problem difficult?

A: we typically do NOT know the viewpoints from which the images were taken. Furthermore, some of the camera parameters such as the focal length is also unknown.



Input: Corresponding “features” in multiple perspective images.
Output: Camera pose, calibration, scene structure representation.

Difficult problem, but under some reasonable assumptions, we can recover the output

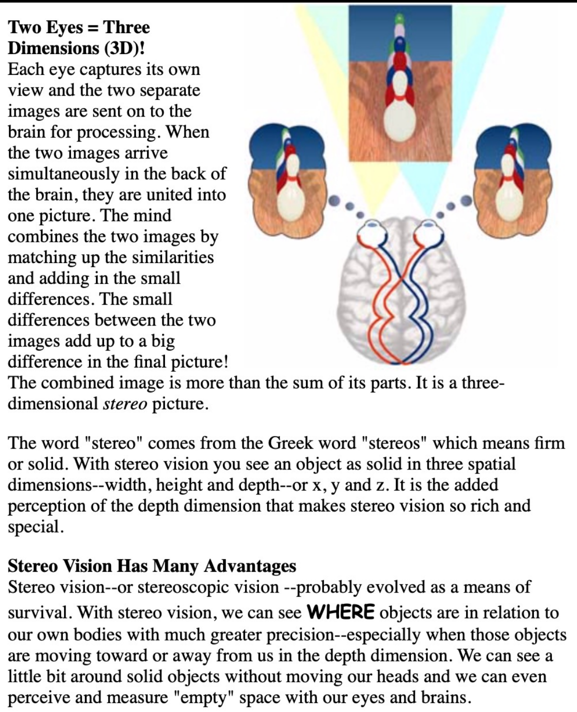
6

Stereovision

- In stereovision, *two or more* views of the scene are available from different vantage points, for example
  - **From two separate cameras**
  - **A single camera that is translated**
- By finding matches (correspondences) between the images, the relative distance of objects from the camera (depth) can be determined.
- The human visual system achieves stereopsis with **binocular vision**.
  - Two images of the world are captured by the eyes (~65 mm apart)
  - Receptive fields fuse images and recognise different positions
  - *Disparity* (the amount of displacement) is used to infer depth

7

Human 3D Vision



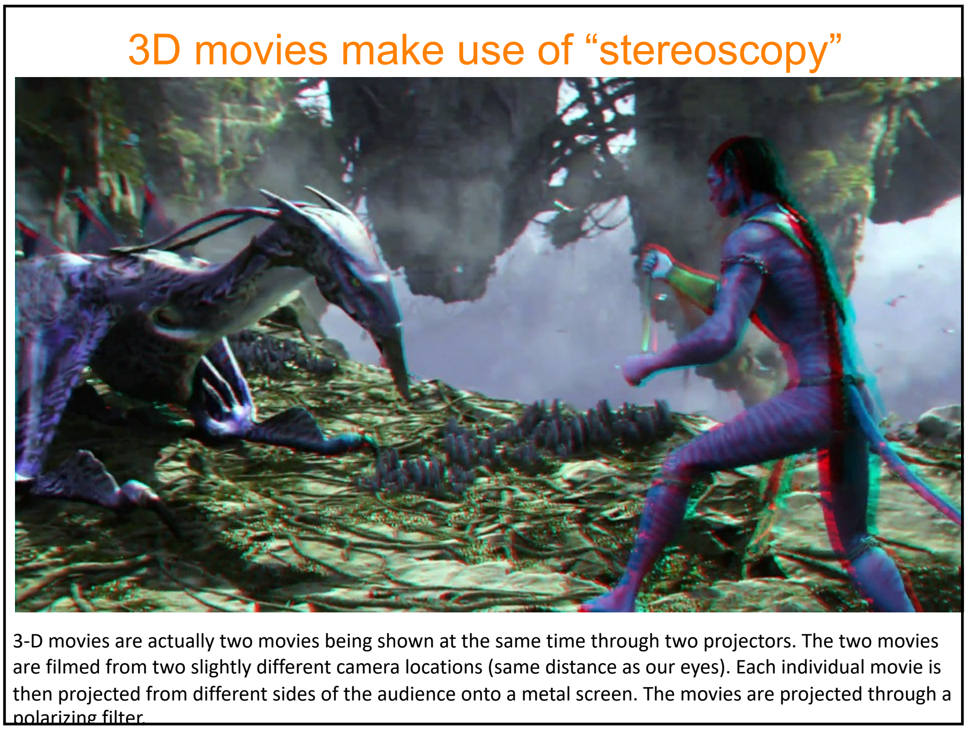
Two Eyes = Three Dimensions (3D)!
Each eye captures its own view and the two separate images are sent on to the brain for processing. When the two images arrive simultaneously in the back of the brain, they are united into one picture. The mind combines the two images by matching up the similarities and adding in the small differences. The small differences between the two images add up to a big difference in the final picture! The combined image is more than the sum of its parts. It is a three-dimensional *stereo* picture.

The word "stereo" comes from the Greek word "stereos" which means firm or solid. With stereo vision you see an object as solid in three spatial dimensions--width, height and depth--or x, y and z. It is the added perception of the depth dimension that makes stereo vision so rich and special.

Stereo Vision Has Many Advantages
Stereo vision--or stereoscopic vision --probably evolved as a means of survival. With stereo vision, we can see **WHERE** objects are in relation to our own bodies with much greater precision--especially when those objects are moving toward or away from us in the depth dimension. We can see a little bit around solid objects without moving our heads and we can even perceive and measure "empty" space with our eyes and brains.

https://www.vision3d.com/stereo.html

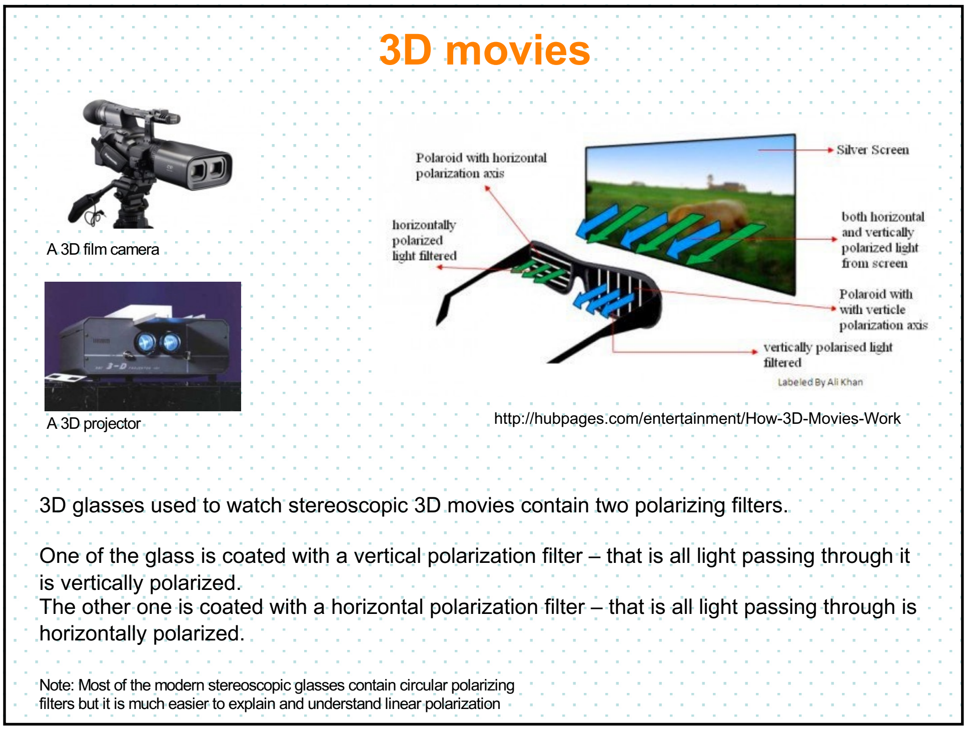
8



3D movies make use of “stereoscopy”

3-D movies are actually two movies being shown at the same time through two projectors. The two movies are filmed from two slightly different camera locations (same distance as our eyes). Each individual movie is then projected from different sides of the audience onto a metal screen. The movies are projected through a polarizing filter

10



3D movies

A 3D film camera

A 3D projector

Polaroid with horizontal polarization axis
horizontally polarized light filtered
Silver Screen
both horizontal and vertically polarized light from screen
Polaroid with with verticle polarization axis
vertically polarised light filtered
Labeled By Ali Khan

http://hubpages.com/entertainment/How-3D-Movies-Work

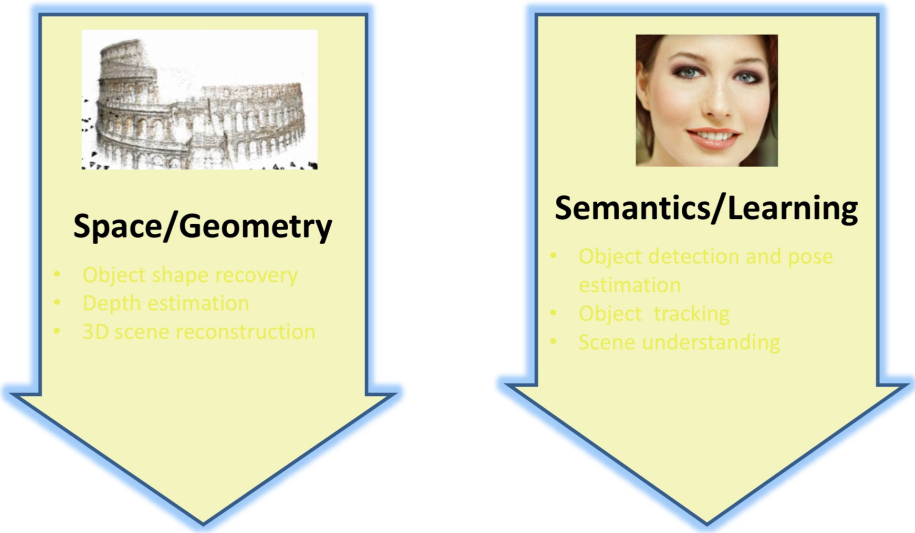
3D glasses used to watch stereoscopic 3D movies contain two polarizing filters.

One of the glass is coated with a vertical polarization filter – that is all light passing through it is vertically polarized.
The other one is coated with a horizontal polarization filter – that is all light passing through is horizontally polarized.

Note: Most of the modern stereoscopic glasses contain circular polarizing filters but it is much easier to explain and understand linear polarization

11

Major Areas in Computer Vision

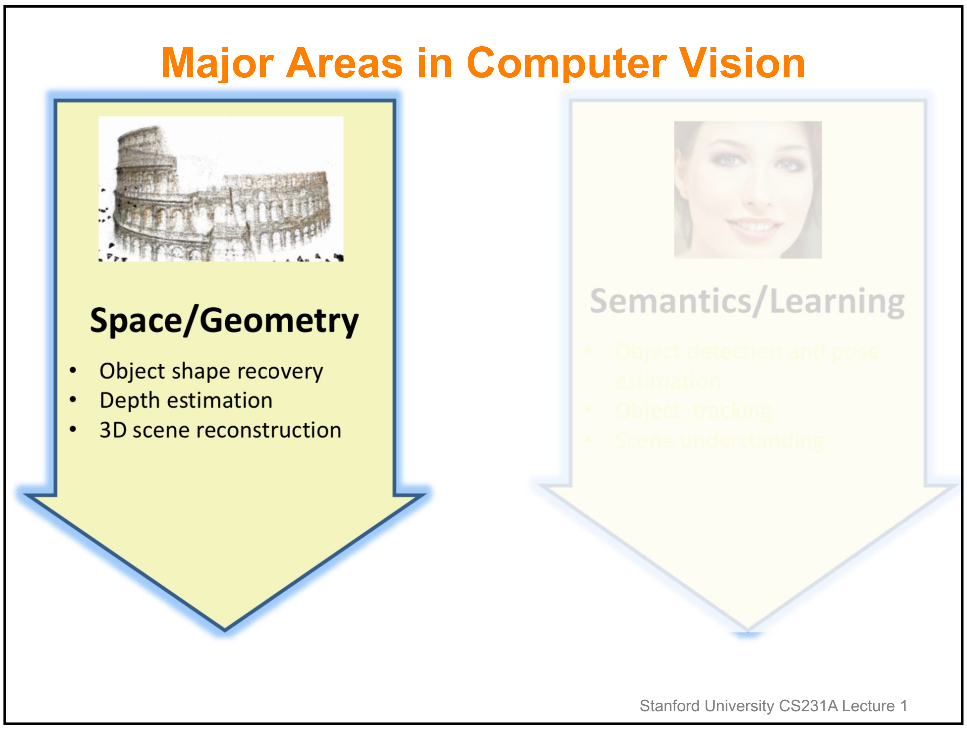


Stanford University CS231A Lecture 1

12

A few mixed examples in Semantics/Learning & Space/Geometry

13



Major Areas in Computer Vision

Space/Geometry

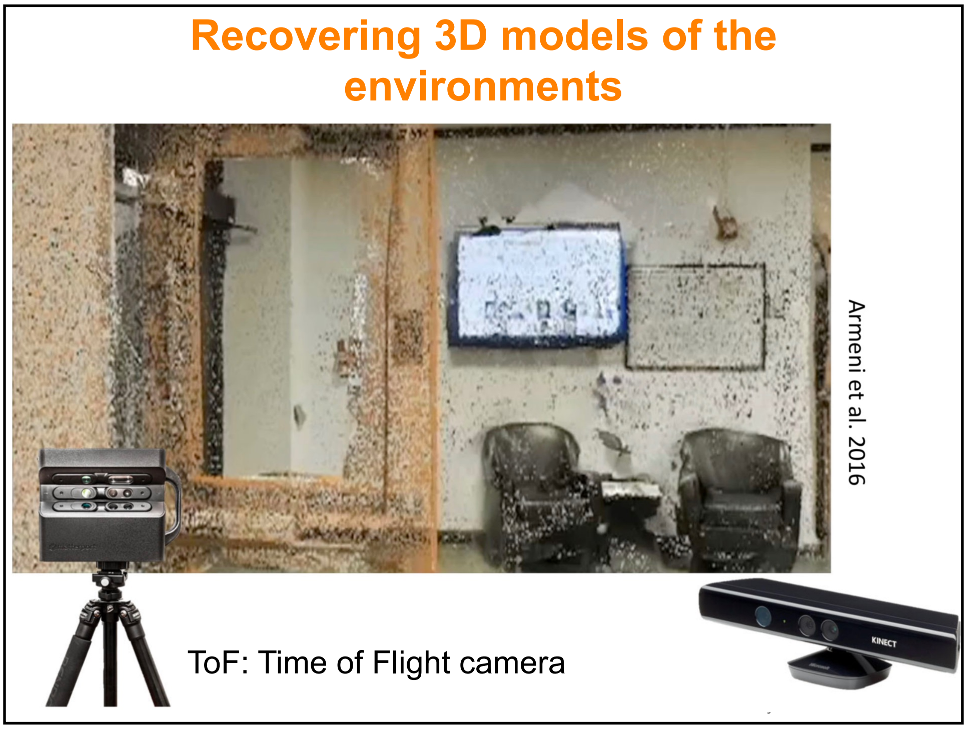
- Object shape recovery
- Depth estimation
- 3D scene reconstruction

Semantics/Learning

- Object detection and pose estimation
- Object tracking
- Scene understanding

Stanford University CS231A Lecture 1

14

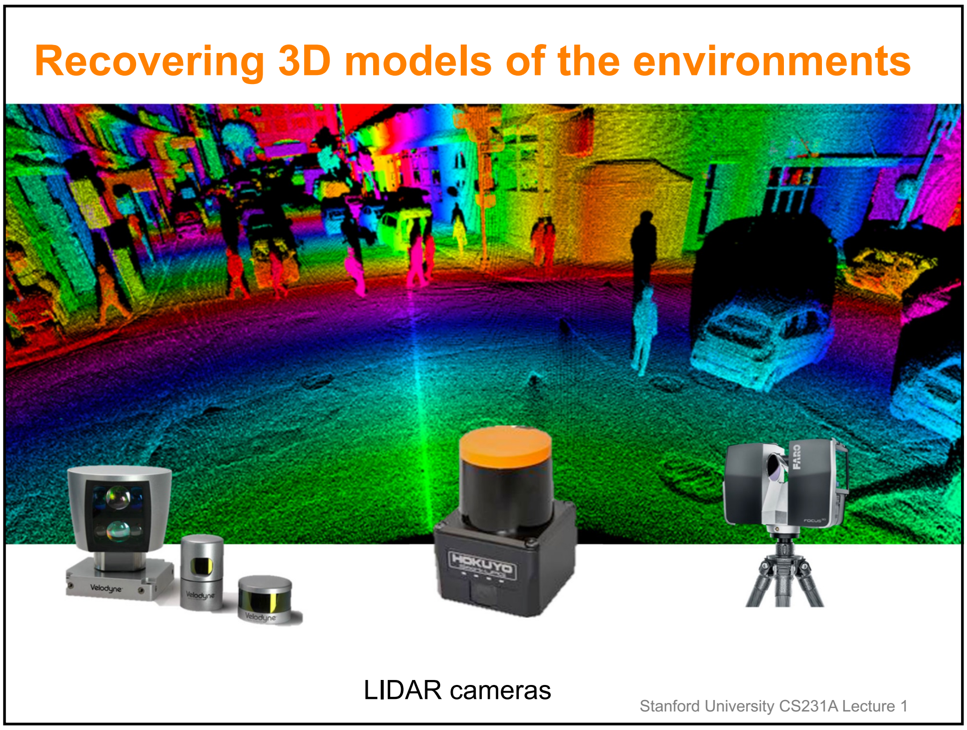


Recovering 3D models of the environments

Armeni et al. 2016

ToF: Time of Flight camera

15

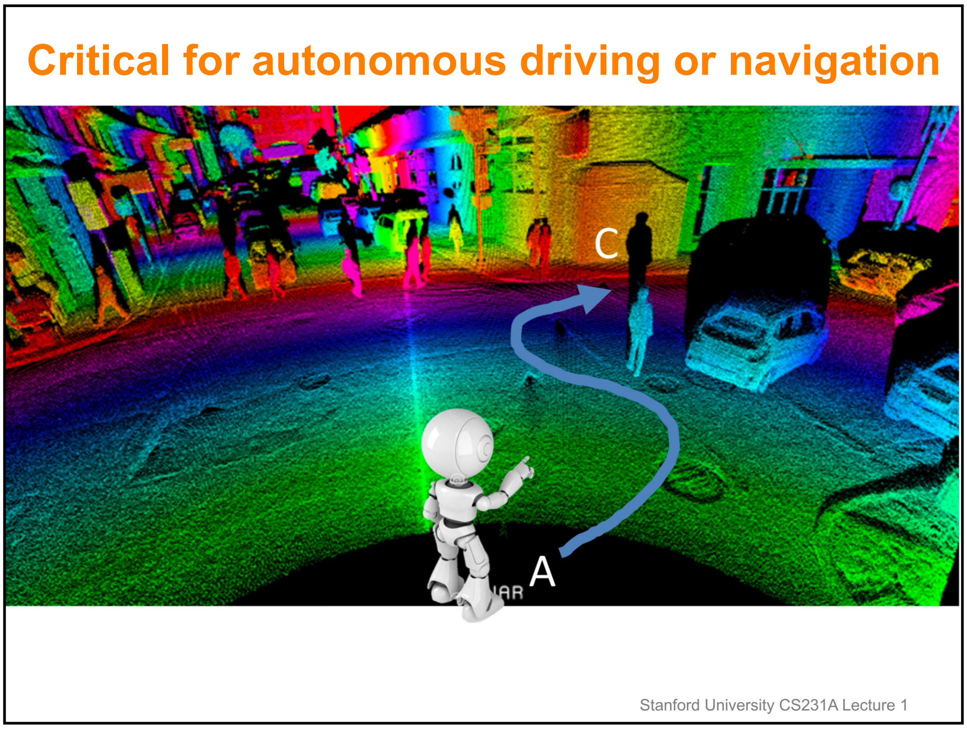


Recovering 3D models of the environments

LIDAR cameras

Stanford University CS231A Lecture 1

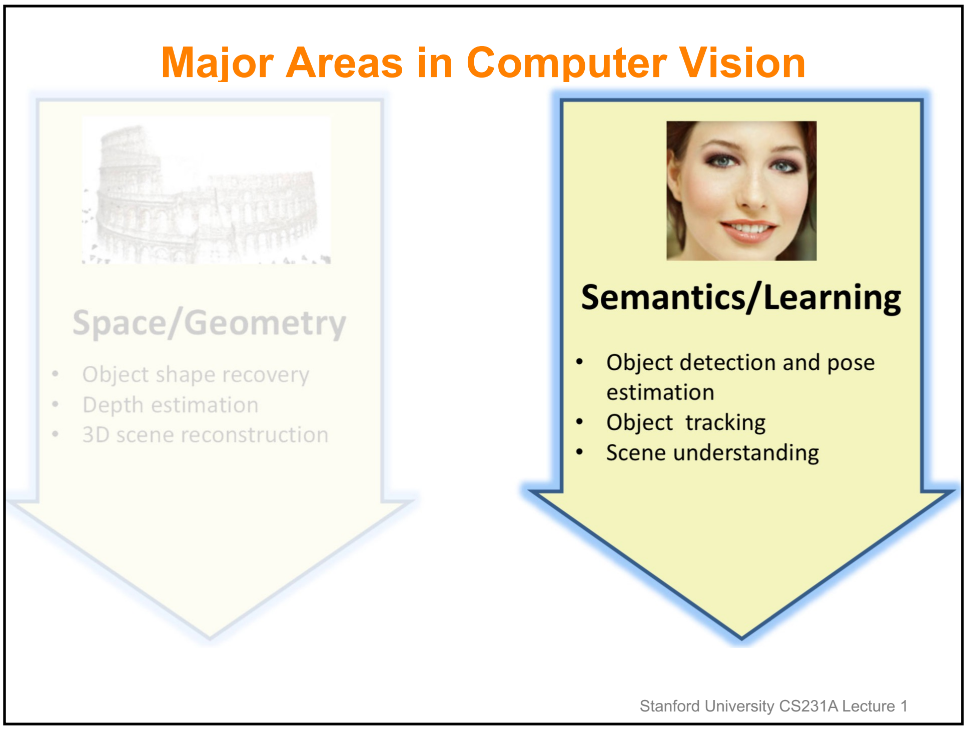
16



Critical for autonomous driving or navigation

Stanford University CS231A Lecture 1

17



Major Areas in Computer Vision

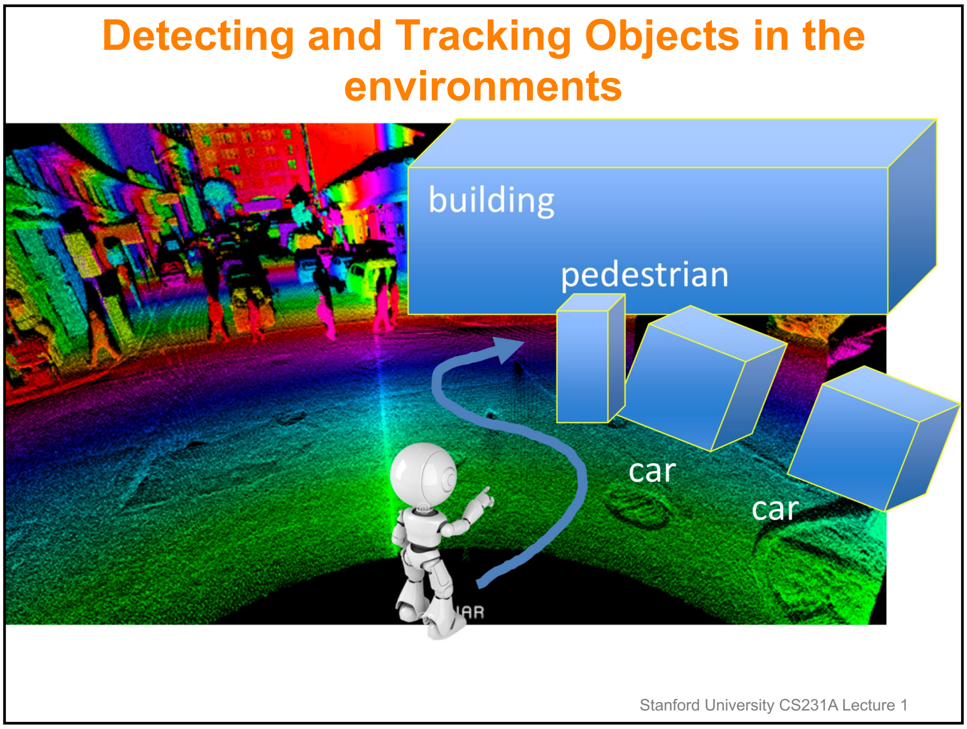
Space/Geometry

- Object shape recovery
- Depth estimation
- 3D scene reconstruction

Semantics/Learning

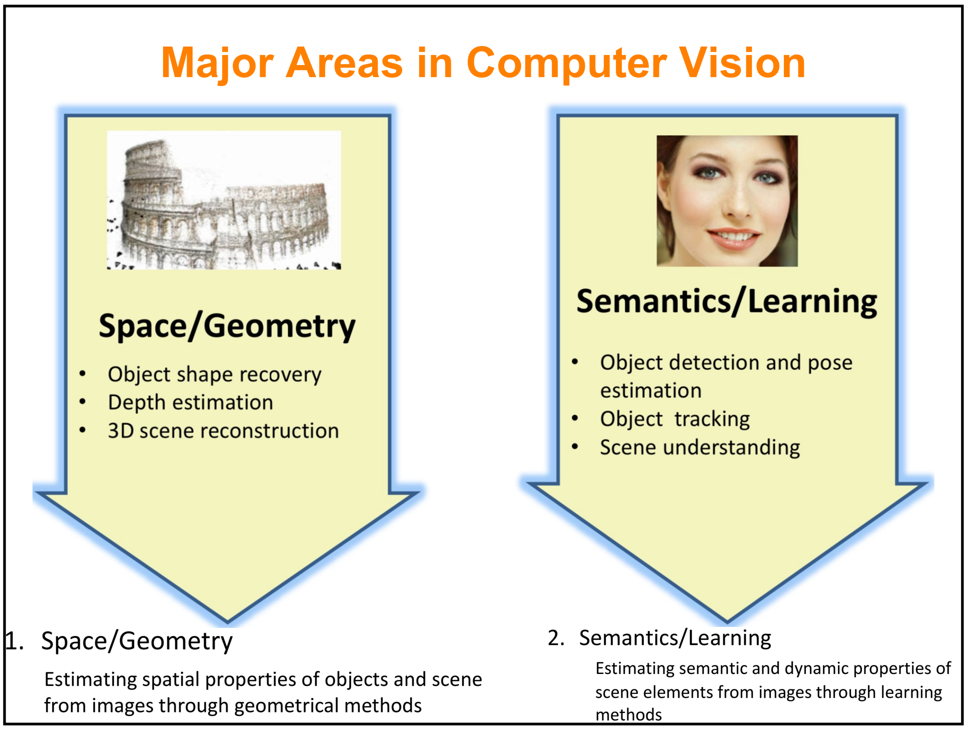
- Object detection and pose estimation
- Object tracking
- Scene understanding

18



Detecting and Tracking Objects in the environments

19



Major Areas in Computer Vision

Space/Geometry

- Object shape recovery
- Depth estimation
- 3D scene reconstruction

Semantics/Learning

- Object detection and pose estimation
- Object tracking
- Scene understanding

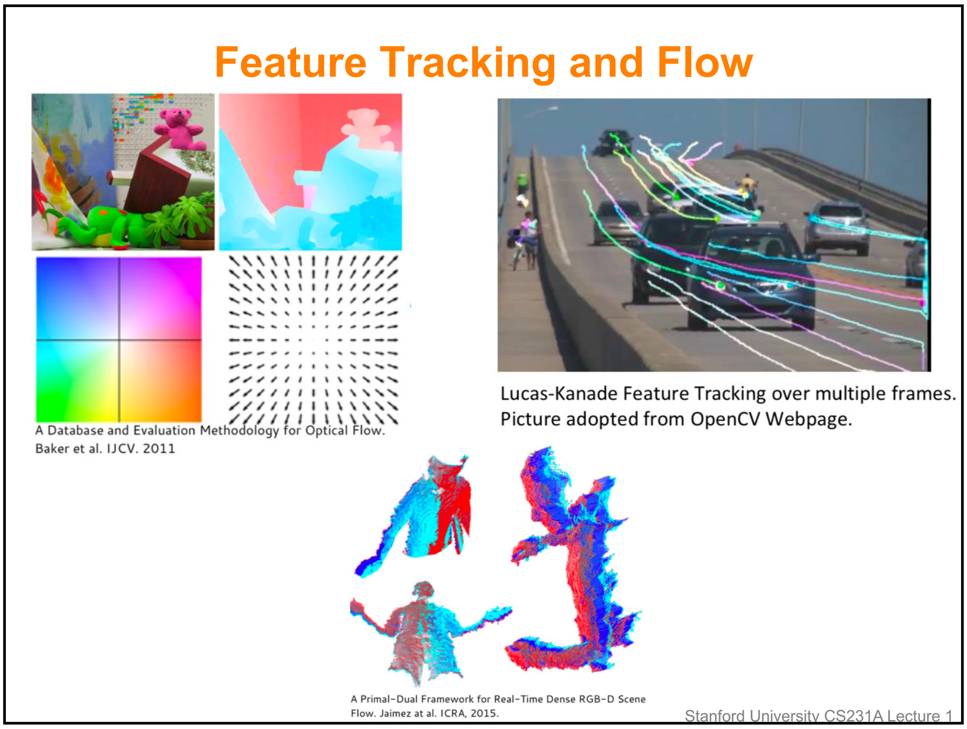
1. Space/Geometry

   Estimating spatial properties of objects and scene from images through geometrical methods

2. Semantics/Learning

   Estimating semantic and dynamic properties of scene elements from images through learning methods

21



Feature Tracking and Flow

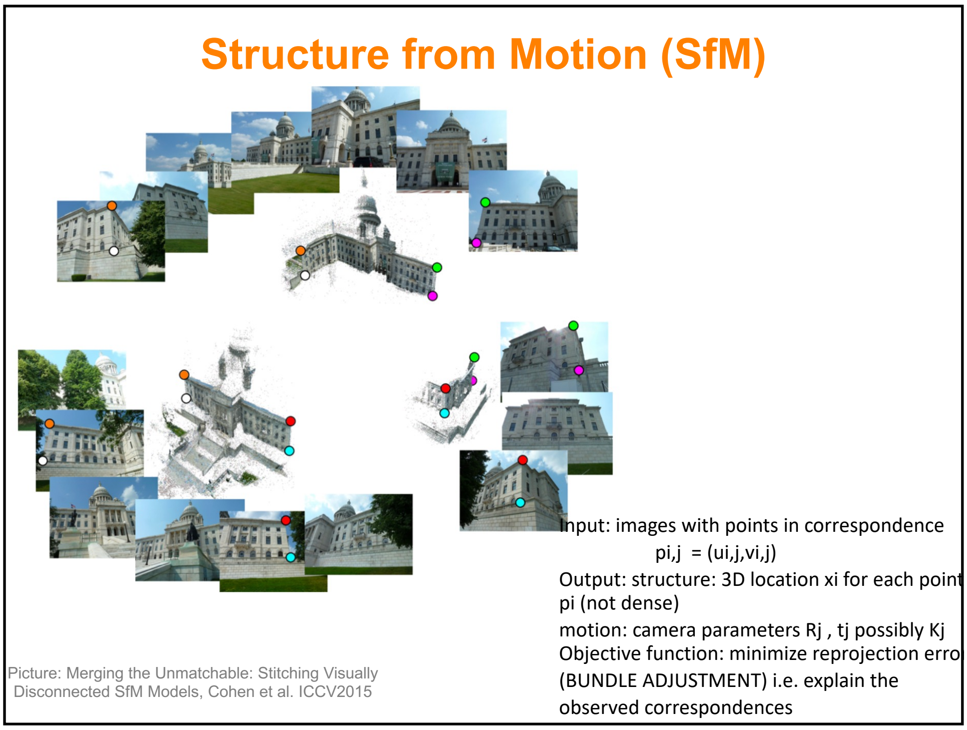
A Database and Evaluation Methodology for Optical Flow. Baker et al. IJCV. 2011

Lucas-Kanade Feature Tracking over multiple frames. Picture adopted from OpenCV Webpage.

A Primal-Dual Framework for Real-Time Dense RGB-D Scene Flow. Jaimez at al. ICRA, 2015.

Stanford University CS231A Lecture 1

22



Structure from Motion (SfM)

Input: images with points in correspondence

pi,j = (ui,j,vi,j)

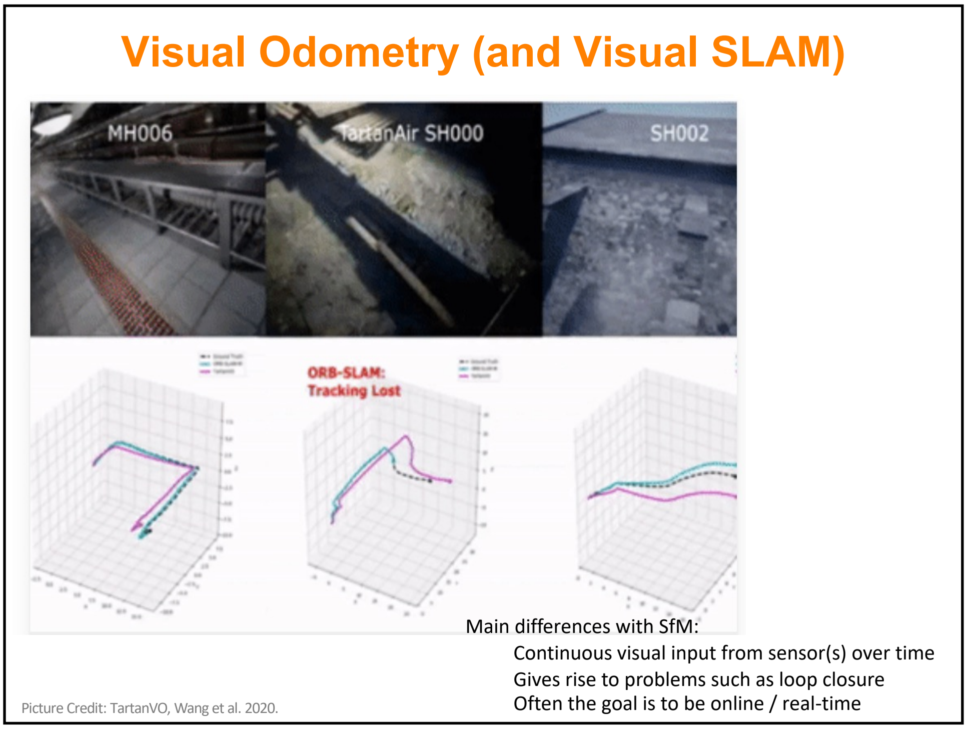
Output: structure: 3D location xi for each point pi (not dense)

motion: camera parameters Rj , tj possibly Kj

Objective function: minimize reprojection erro (BUNDLE ADJUSTMENT) i.e. explain the observed correspondences

Picture: Merging the Unmatchable: Stitching Visually Disconnected SfM Models, Cohen et al. ICCV2015

23



Visual Odometry (and Visual SLAM)

Main differences with SfM:

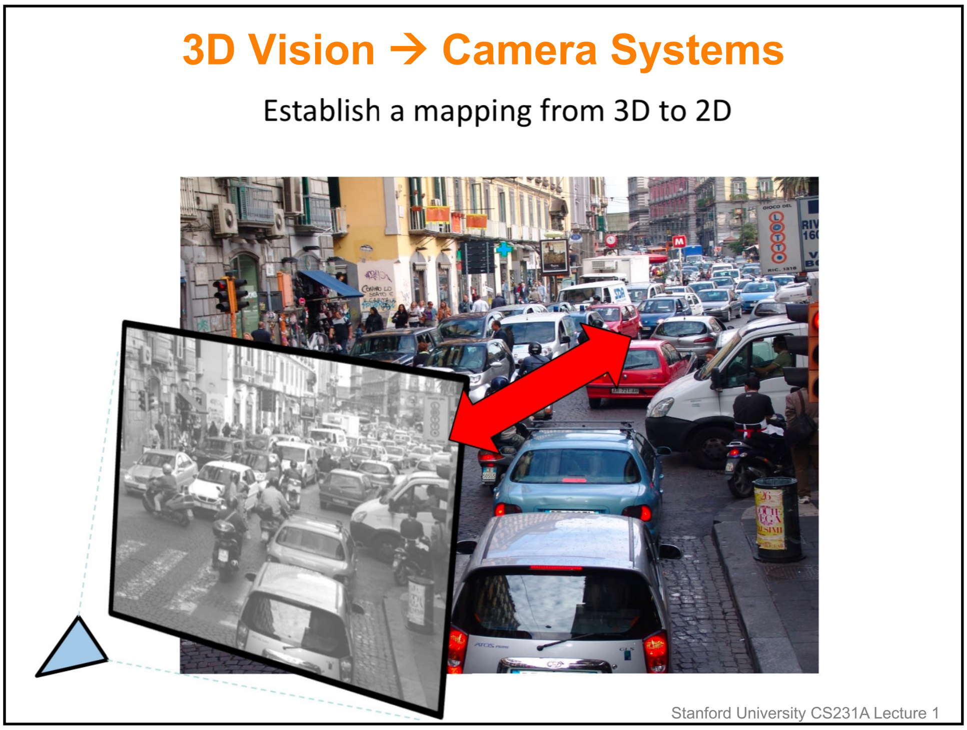
Continuous visual input from sensor(s) over time

Gives rise to problems such as loop closure

Often the goal is to be online / real-time

Picture Credit: TartanVO, Wang et al. 2020.

24

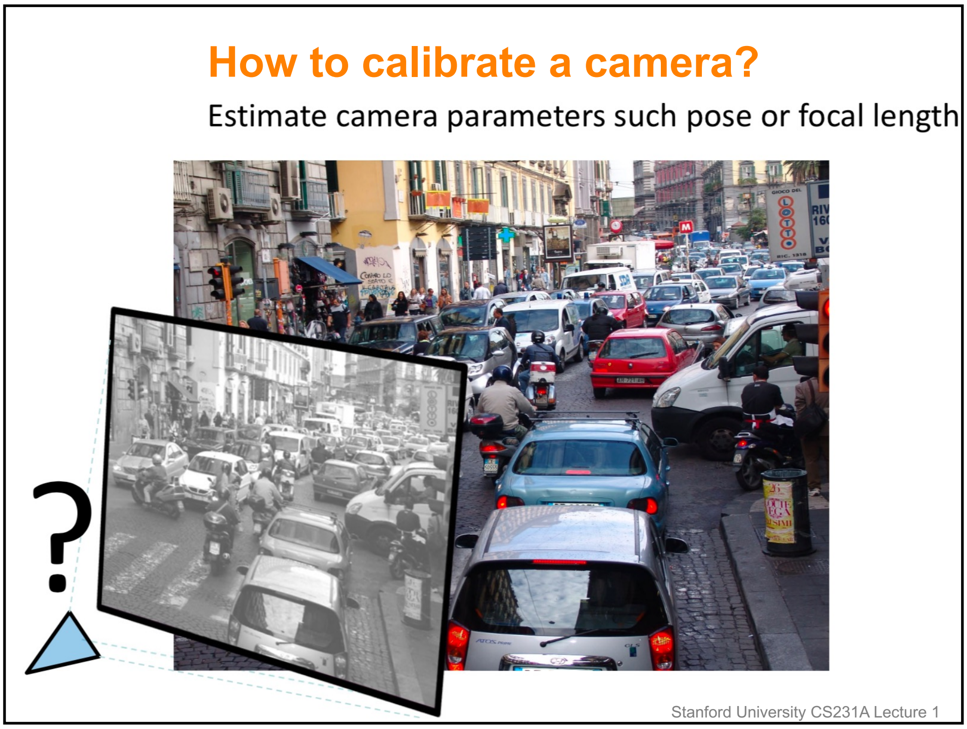


3D Vision → Camera Systems

Establish a mapping from 3D to 2D

Stanford University CS231A Lecture 1

25



How to calibrate a camera?

Estimate camera parameters such pose or focal length

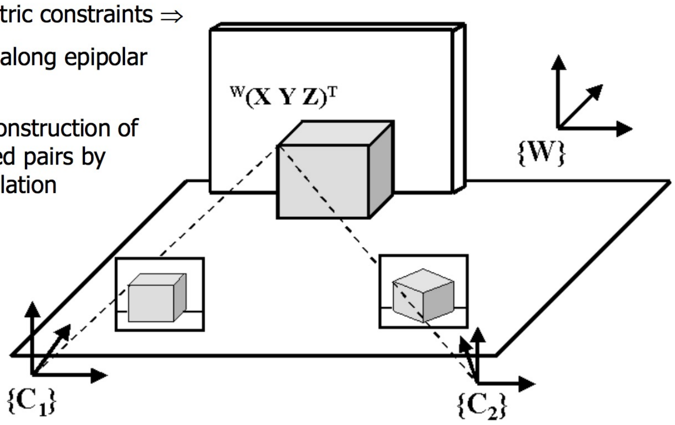
Stanford University CS231A Lecture 1

26

StereoVision: Passive Triangulation

- Correspondence problem
- Geometric constraints ⇒

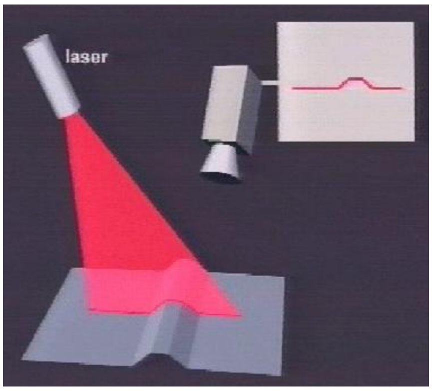
  search along epipolar lines
- 3D reconstruction of matched pairs by triangulation



Picture: G. Gerig Lecture Notes: C6320

27

Active Vision

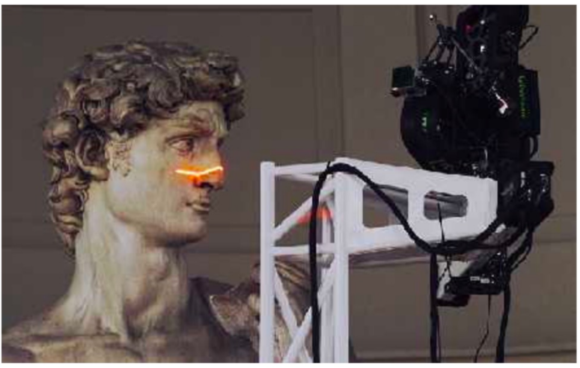


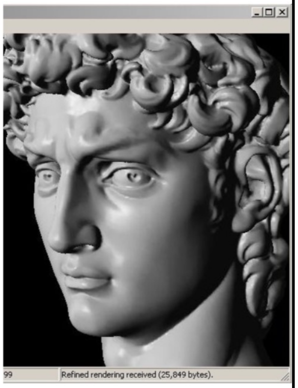
Active manipulation of scene: Project Light Pattern on object. Observe geometry of pattern via camera → 3D geometry

Picture: G. Gerig Lecture Notes: C6320

28

Accurate 3D Object Prototyping





99 Refined rendering received (25,849 bytes).

Scanning Michelangelo's *"The David"*

- The Digital Michelangelo Project
  - http://graphics.stanford.edu/projects/mich/
- 2 BILLION polygons, accuracy to .29mm

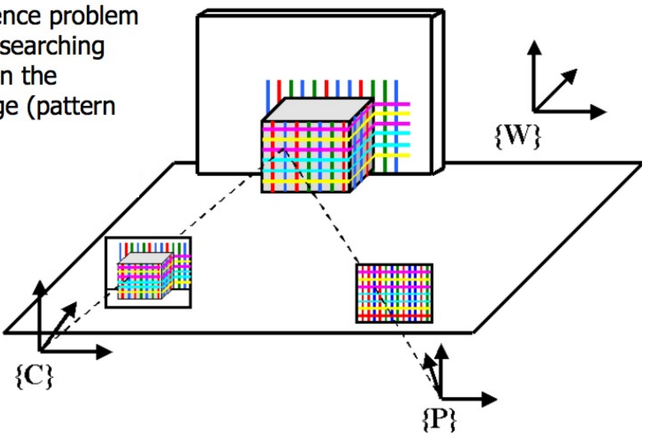
as of 2009

model of the entire 5-meter statue

29

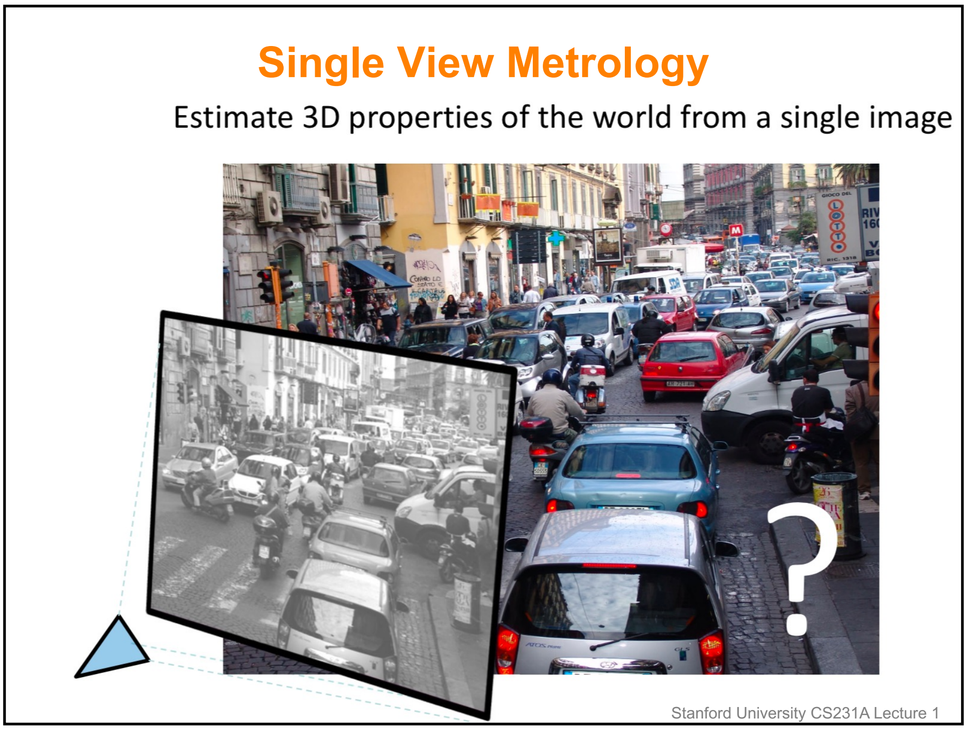
Active Triangulation: Structured Light

- One of the cameras is replaced by a light emitter
- Correspondence problem is solved by searching the pattern in the camera image (pattern decoding)



G. Gerig Lecture Notes: C6320

30

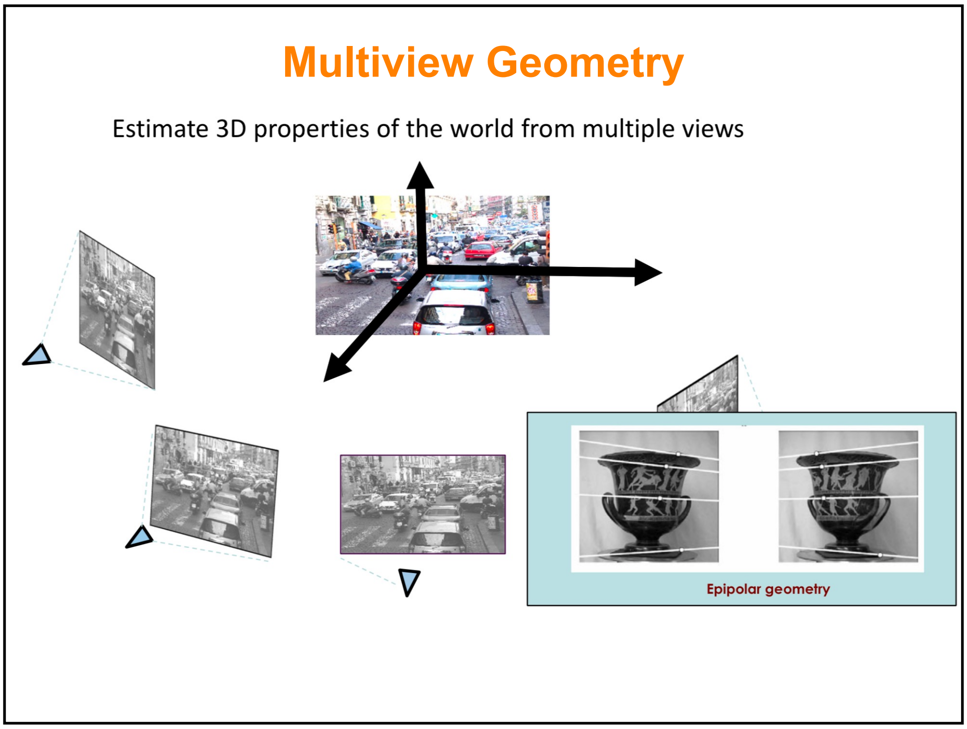


Single View Metrology

Estimate 3D properties of the world from a single image

Stanford University CS231A Lecture 1

31



Multiview Geometry

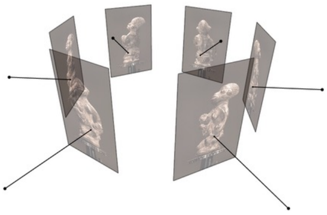
Estimate 3D properties of the world from multiple views

32

Multi-view Stereo

Input: calibrated images from several viewpoints (**known camera**: intrinsics and extrinsics)

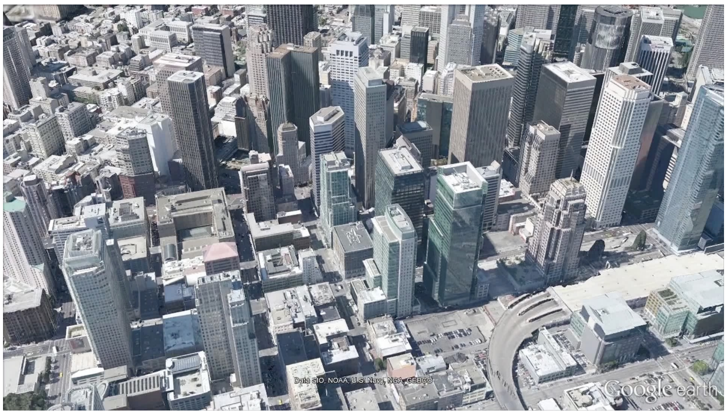
Output: 3D object model



Figures by Carlos Hernandez

Slide credit: Noah Snavely

33



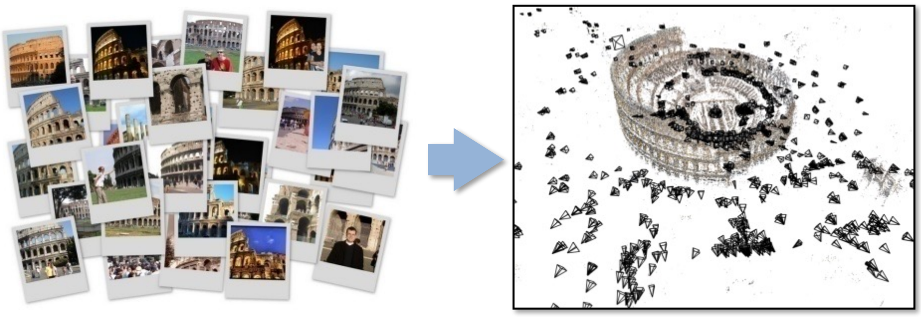
Slide credit: Noah Snavely

For example, Google maps and google earth make heavy use of this to reconstruct cities

34

Photo Tourism

http://grail.cs.washington.edu/projects/rome/



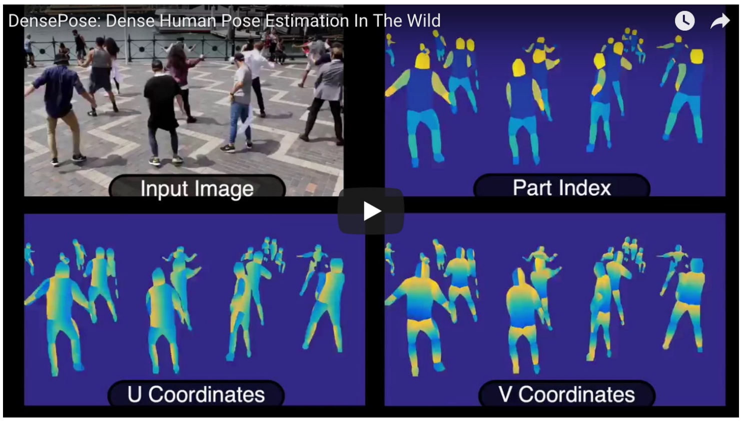
Takes as input large collections of images from either personal photo collections or Internet photo sharing sites (a), and automatically computes each photo's viewpoint and a sparse 3D model of the scene (b).

http://phototour.cs.washington.edu/

35

Human body mapping: 2D to 3D

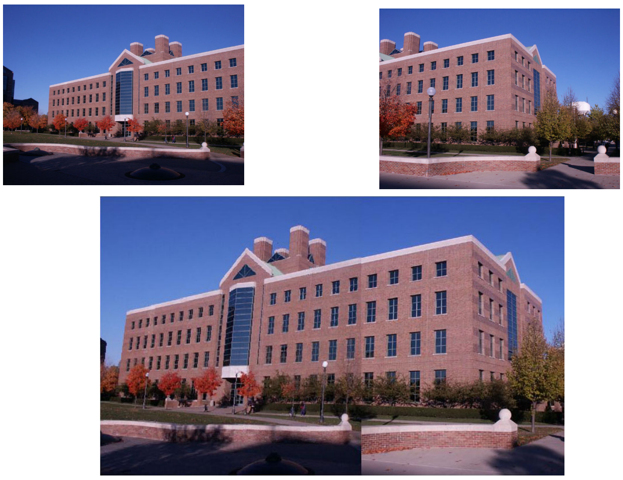
http://densepose.org/ by R. Alp Güler et al.



Dense human pose estimation aims at mapping all human pixels of an RGB image to the 3D surface of the human body.

36

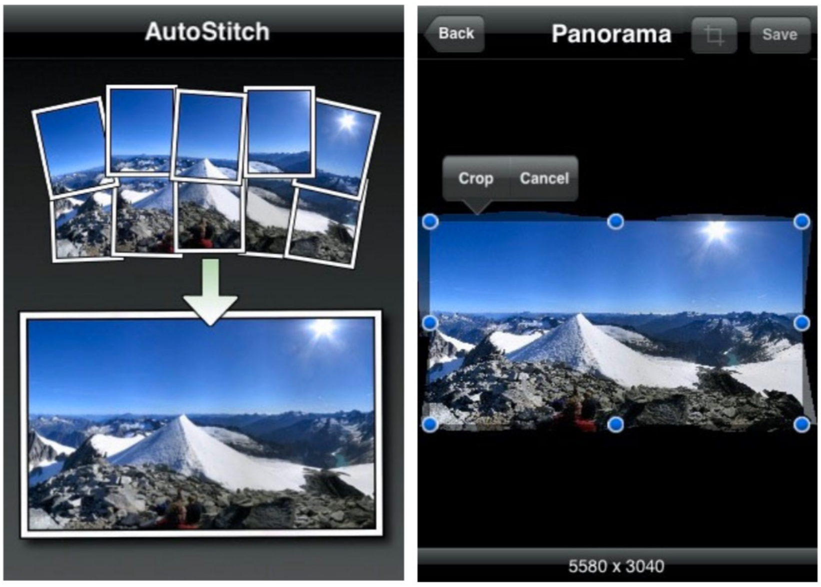
3D Vision– Image Alignment and Mosaicing



Even if our final goal is not for a 3-D model, understanding the 3-D geometry encoded in the images facilitates many image processing tasks

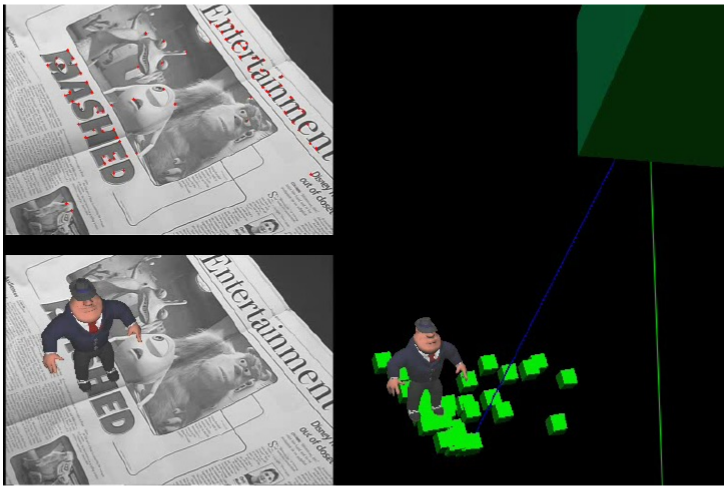
37

Panaromic Photography



38

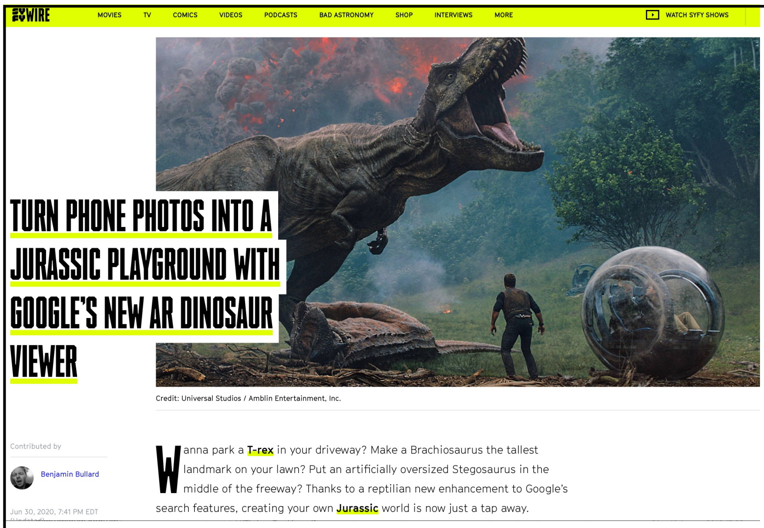
3D Vision: Augmented Reality AR



e.g. Real-Time Virtual Object Insertion into your own home-made video

UCLA Vision Lab

39



TURN PHONE PHOTOS INTO A JURASSIC PLAYGROUND WITH GOOGLE'S NEW AR DINOSAUR VIEWER

Credit: Universal Studios / Amblin Entertainment, Inc.

Contributed by

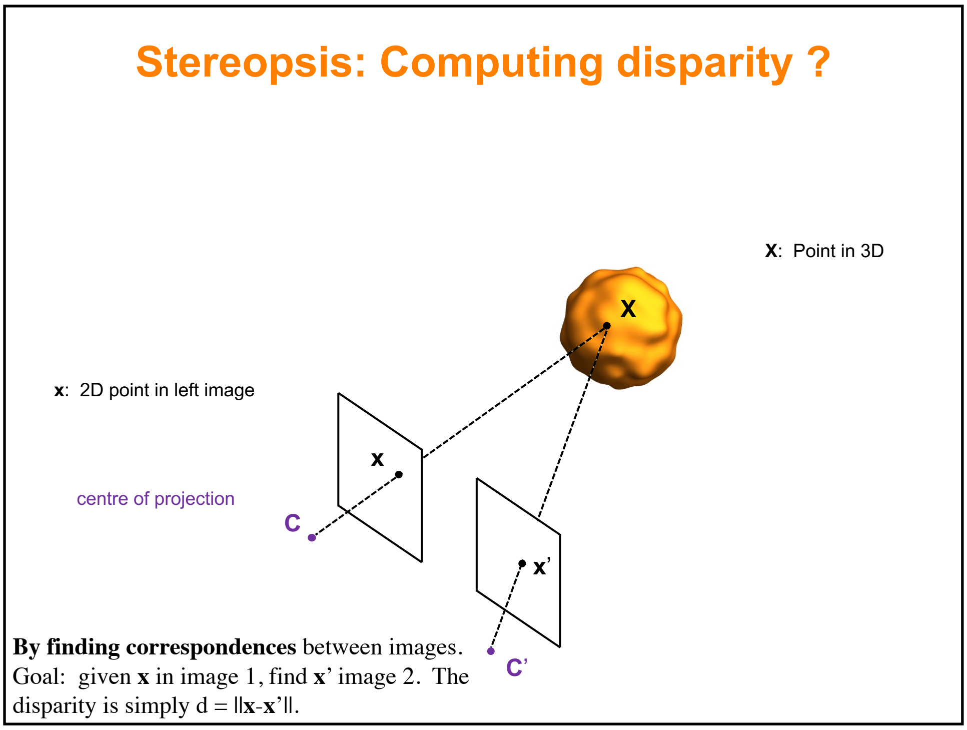
Benjamin Bullard

Jun 30, 2020, 7:41 PM EDT

Wanna park a T-rex in your driveway? Make a Brachiosaurus the tallest landmark on your lawn? Put an artificially oversized Stegosaurus in the middle of the freeway? Thanks to a reptilian new enhancement to Google's search features, creating your own Jurassic world is now just a tap away.

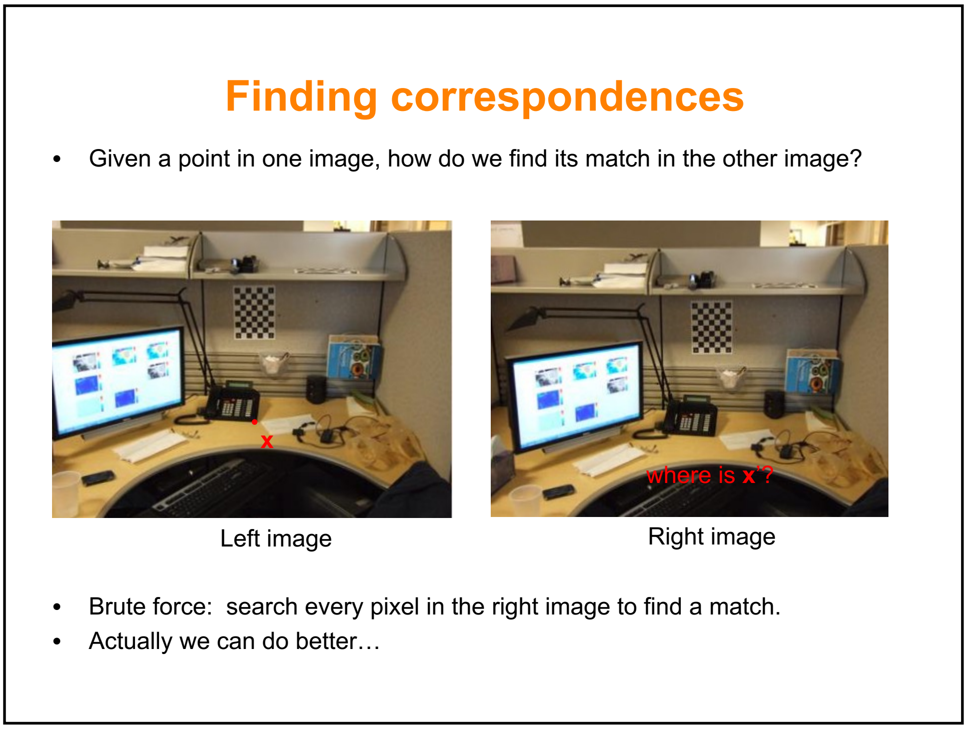
https://www.syfy.com/syfywire/google-search-adds-augmented-reality-dinosaurs

40



Stereopsis: Computing disparity ?

By finding correspondences between images. Goal: given **x** in image 1, find **x**’ image 2. The disparity is simply $d = \|\mathbf{x}-\mathbf{x}'\|$.



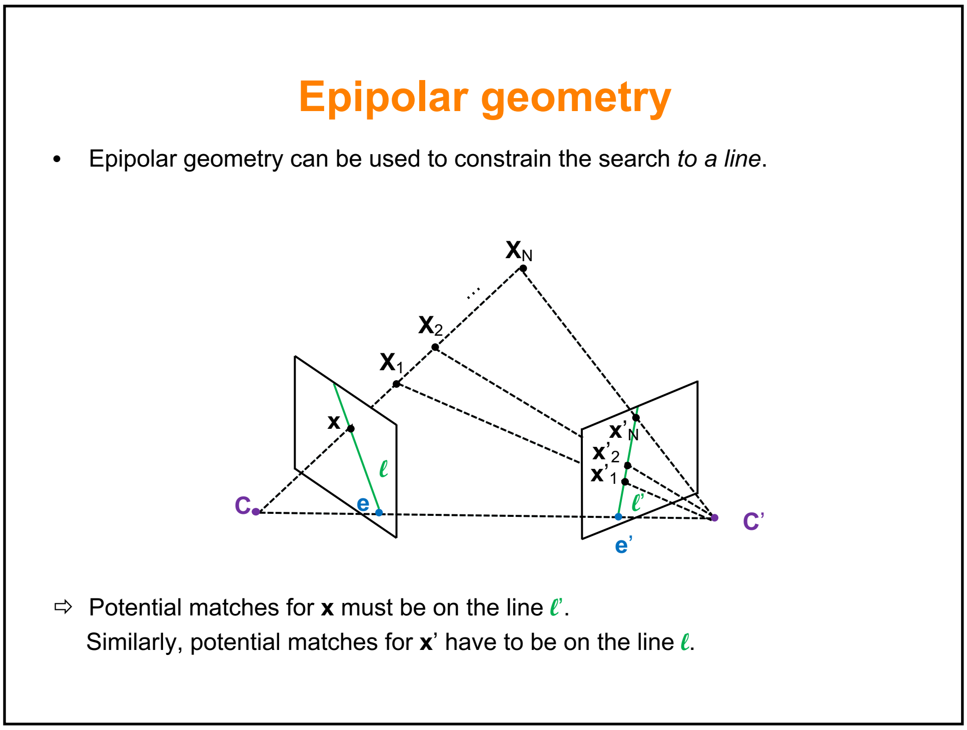
Finding correspondences

- Given a point in one image, how do we find its match in the other image?

Left image

Right image

- Brute force: search every pixel in the right image to find a match.
- Actually we can do better…

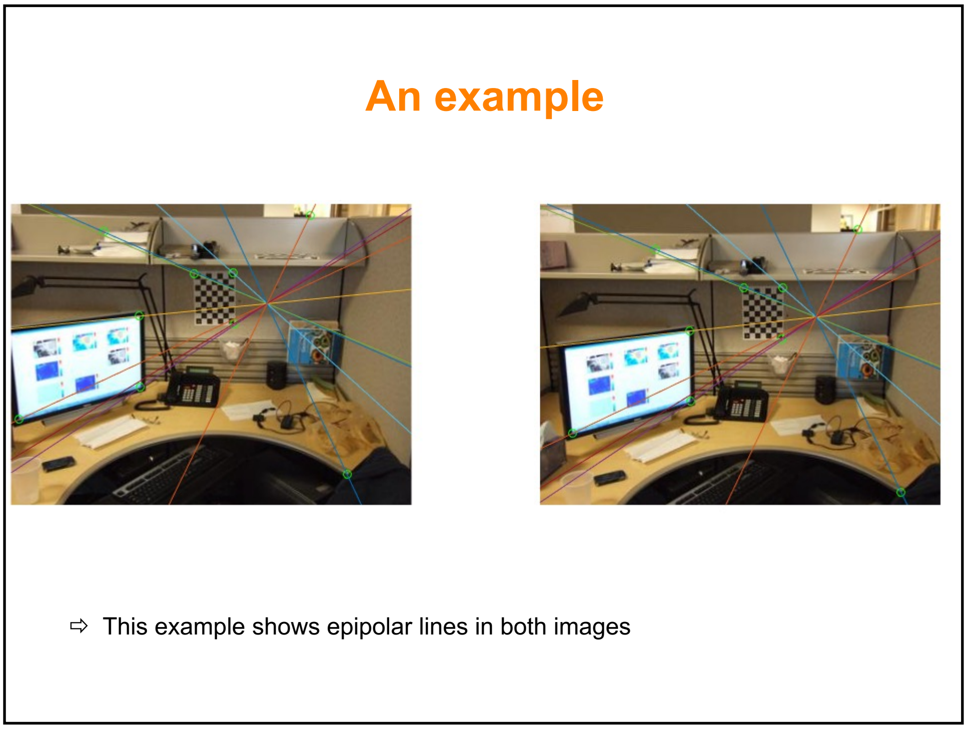


Epipolar geometry

- Epipolar geometry can be used to constrain the search *to a line*.

⇨ Potential matches for **x** must be on the line ℓ'.
Similarly, potential matches for **x**’ have to be on the line ℓ.

43



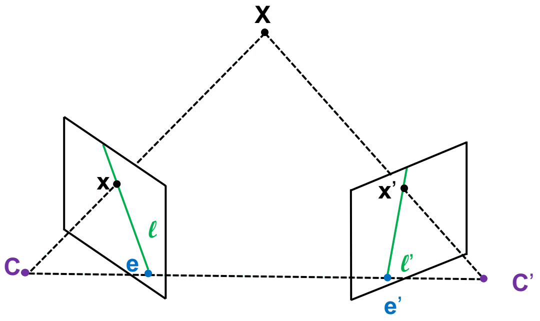
An example

⇨ This example shows epipolar lines in both images

44

Epipolar Geometry

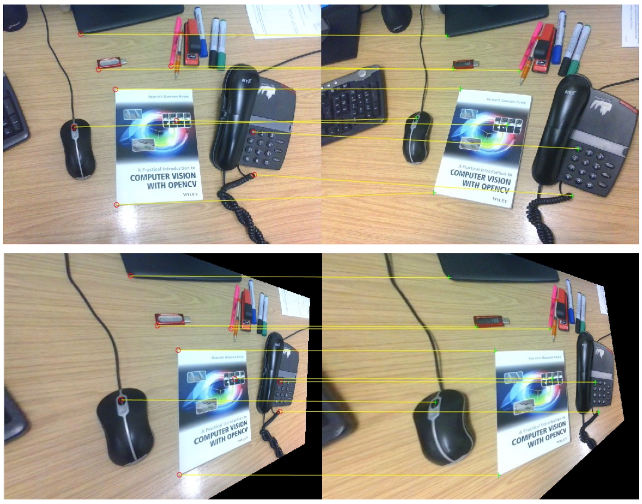
- Epipolar geometry can be used to constrain the search *to a line*.



- **C**, **C'**, and **X** are three 3D points that form the *epipolar plane*.
- The baseline connects the two camera centres. The *epipoles* **e** and **e'** are where the baseline intersects the image plane.
- All epipolar lines in an image will go through the epipole, regardless of **X**.

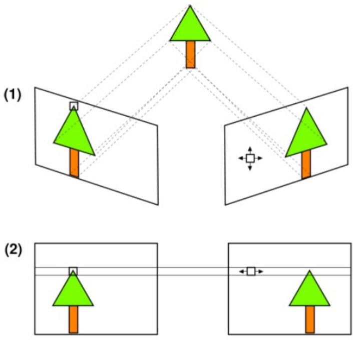
45

Example



46

Image Rectification



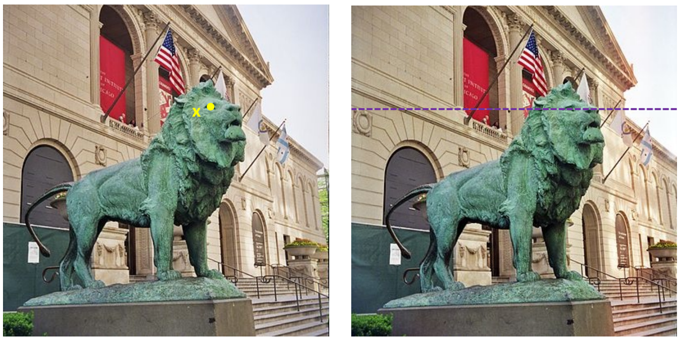
If the two cameras are aligned to be coplanar, the search is simplified to one dimension - a horizontal line parallel to the baseline between the cameras

http://en.wikipedia.org/wiki/Image_rectification

47

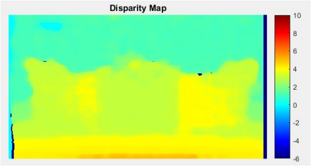
Dense stereo matching

- To compute disparity. The standard algorithm to this is
  1. Find a sparse set of inlier correspondences, estimate **F**
  2. If necessary, rectify the stereo images
  3. For each pixel **x** in image 1, search along the epipolar line for a match **x**’
  4. Save the disparity d = ||**x**-**x**’||



48

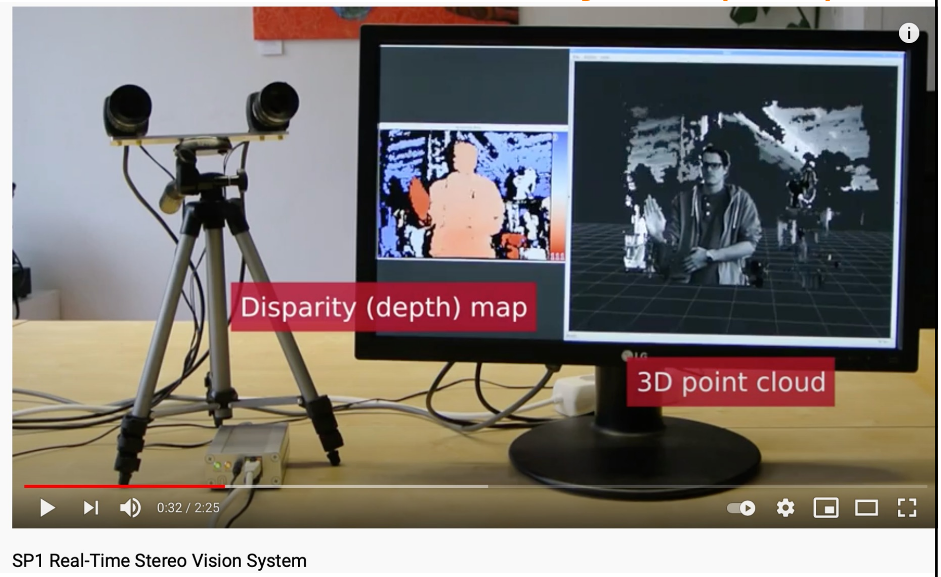
Example



Depth/disparity estimation from stereo images
Middlebury dataset:
http://vision.middlebury.edu/stereo/
with GT disparities
http://vision.middlebury.edu/stereo/data/scenes2014/

Kitti
http://www.cvlibs.net/datasets/kitti/

Real time stereo vision system (2015)

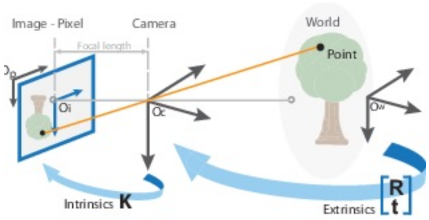


SP1 Real-Time Stereo Vision System

https://www.youtube.com/watch?v=vVVjFqUkG4E

Upgrading to depth

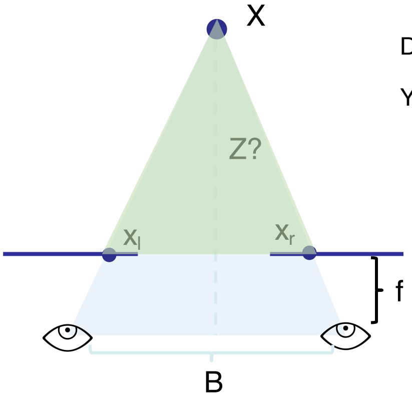
- Depth will be inversely proportional to disparity.
- To upgrade to depth, we must know the baseline *B* and the focal length *f*. But how to determine these?
- Normally these will come from **camera calibration**.



Picture: www.mathworks.com

51

Once you have disparity aka how much did the pixel move, you can get (relative) depth



Do we have enough to know what is Z?

Yes, similar triangles!

$$\frac{B-(x_l+x_r)}{Z-f}=\frac{B}{Z}$$

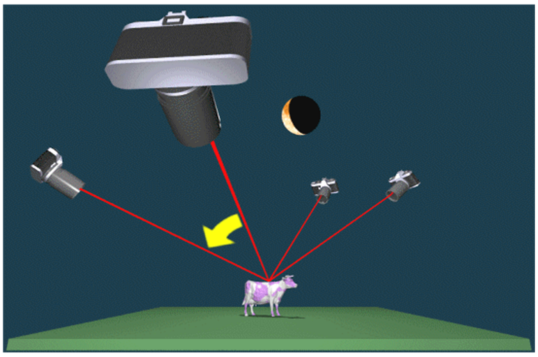
$$Z=\frac{fB}{x_r-x_l}$$

disparity

52

Summary: 3D Vision

- ❑ Want to infer some representation of the world from collection of images
- ❑ Complexity of the physical world >> complexity of the image measurements
- ❑ Projecting from 3D world to 2D images: 1 dimension is lost



- Cannot simply invert and reconstruct the “true” scene from a number of images
- We can reconstruct at best a “MODEL” of the world
- Modeling: a form of engineering art
- Depends on the applications/tasks at hand
  - Visualize the scene from different viewpoints
  - Recognize objects, their shape or motion, actions,
  - **Where am I? How should I navigate in the world? What to do next? …**

Now, new norm in 3D Vision is becoming Learning Methods (just past 1-2 years)

Deep Methods in 3D

We will explore these together in the last part of the course.

Course Requirements

Basic Computer Vision and/or Image Processing knowledge is required.

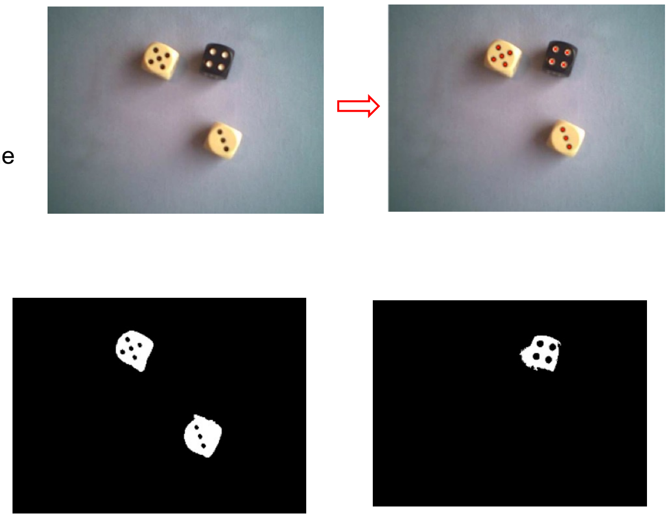
See the Syllabus

TEST YOUR Basic Computer Vision Skills

HW0

Using the image on the left, perform image processing:

Calculate and output the numbers on both the lighter dice and the dark dice.



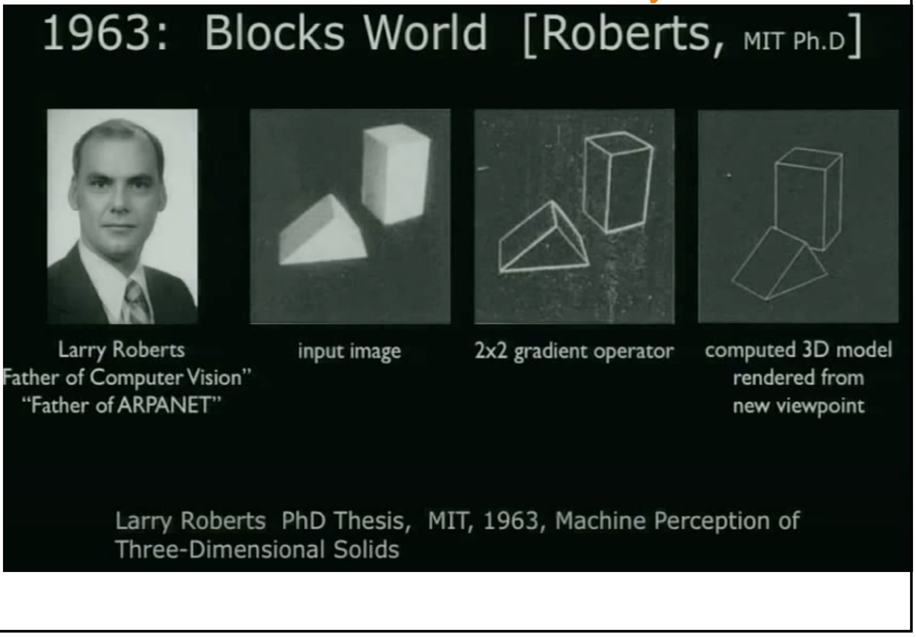
3D Vision A little bit history

some slides
are due:

https://www.youtube.com/watch?v=kyIzMr917Rc

58

3D Vision: A little bit History



1963: Blocks World [Roberts, MIT Ph.D]

Larry Roberts
Father of Computer Vision”
“Father of ARPANET”

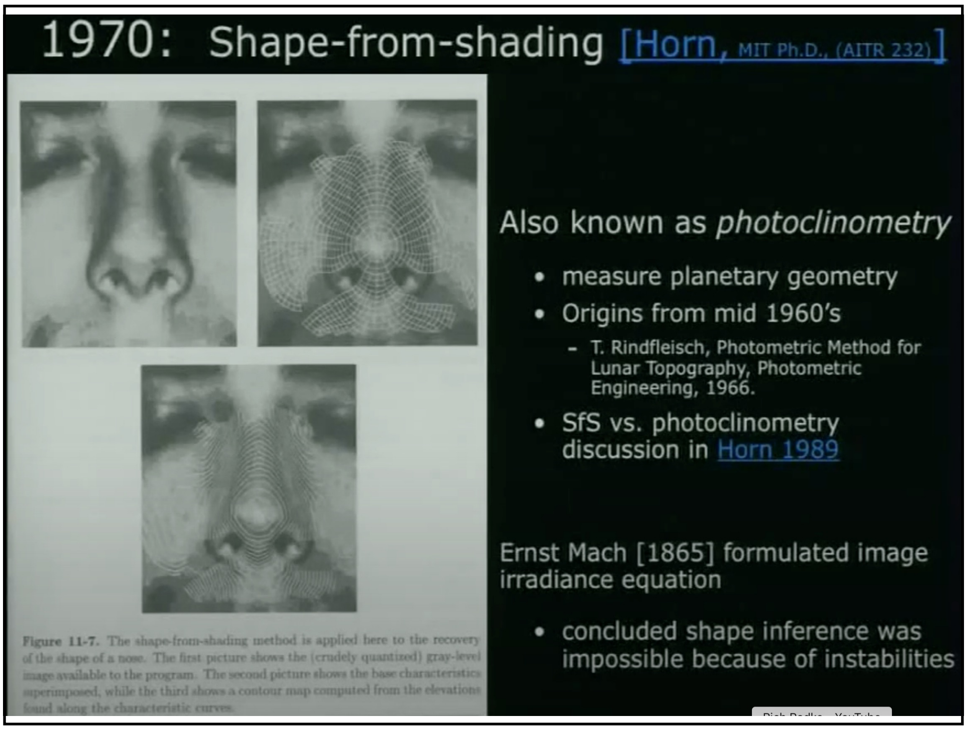
input image

2x2 gradient operator

computed 3D model
rendered from
new viewpoint

Larry Roberts PhD Thesis, MIT, 1963, Machine Perception of Three-Dimensional Solids

59



1970: Shape-from-shading [Horn, MIT Ph.D., (AITR 232)]

Figure 11-7. The shape-from-shading method is applied here to the recovery of the shape of a nose. The first picture shows the (crudely quantized) gray-level image available to the program. The second picture shows the base characteristics superimposed, while the third shows a contour map computed from the elevations found along the characteristic curves.

Also known as *photoclinometry*

- measure planetary geometry
- Origins from mid 1960's
  - T. Rindfleisch, Photometric Method for Lunar Topography, Photometric Engineering, 1966.
- SfS vs. photoclinometry discussion in Horn 1989

Ernst Mach [1865] formulated image irradiance equation

- concluded shape inference was impossible because of instabilities

Rich Radke – YouTube

60

Photometric Stereo

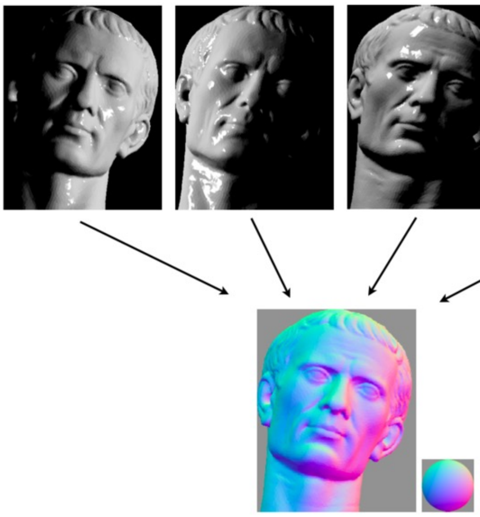
Robert Woodham 1980

Single camera, multiple-lights, what is the surface normal?

D. Forsyth book Section 2.2.4

61

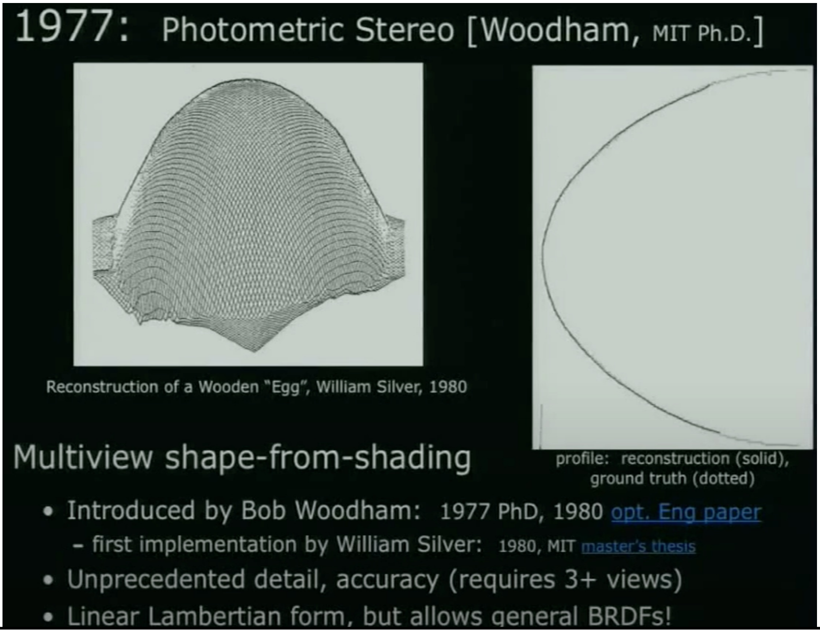
Photometric Stereo



Photometric Stereo results in per-pixel high resolution capture

62

3D Vision: A little bit History



1977: Photometric Stereo [Woodham, MIT Ph.D.]

Reconstruction of a Wooden "Egg", William Silver, 1980

profile: reconstruction (solid), ground truth (dotted)

Multiview shape-from-shading

- Introduced by Bob Woodham: 1977 PhD, 1980 opt. Eng paper
  - first implementation by William Silver: 1980, MIT master's thesis
- Unprecedented detail, accuracy (requires 3+ views)
- Linear Lambertian form, but allows general BRDFs!

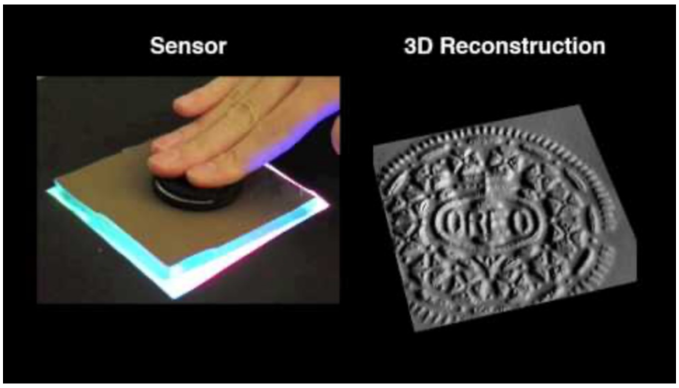
63

The president in 3D:
https://www.youtube.com/watch?v=4GiLAOtjHNo&t=167s

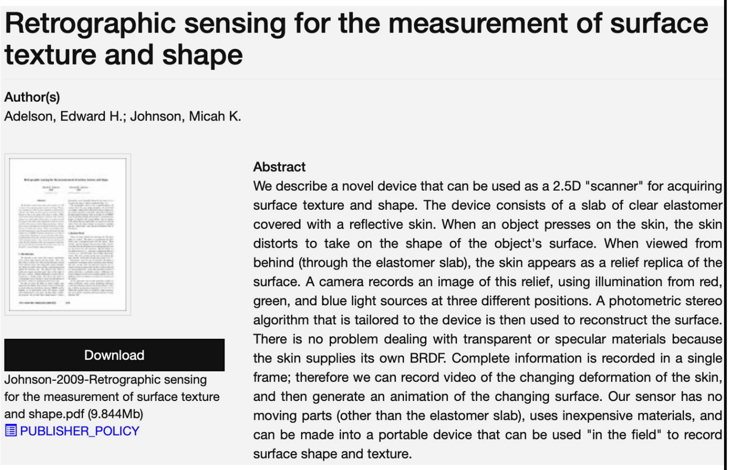
High-resolution 3D Capture

Recent instantiation of Photometric Stereo: GelSight

Sensing Surfaces with Gelsight:
https://www.youtube.com/watch?v=S7gXih4XS7A



Uses of Photometric Stereo



Retrographic sensing for the measurement of surface texture and shape

Author(s)
Adelson, Edward H.; Johnson, Micah K.

Abstract
We describe a novel device that can be used as a 2.5D "scanner" for acquiring surface texture and shape. The device consists of a slab of clear elastomer covered with a reflective skin. When an object presses on the skin, the skin distorts to take on the shape of the object's surface. When viewed from behind (through the elastomer slab), the skin appears as a relief replica of the surface. A camera records an image of this relief, using illumination from red, green, and blue light sources at three different positions. A photometric stereo algorithm that is tailored to the device is then used to reconstruct the surface. There is no problem dealing with transparent or specular materials because the skin supplies its own BRDF. Complete information is recorded in a single frame; therefore we can record video of the changing deformation of the skin, and then generate an animation of the changing surface. Our sensor has no moving parts (other than the elastomer slab), uses inexpensive materials, and can be made into a portable device that can be used "in the field" to record surface shape and texture.

Download

Johnson-2009-Retrographic sensing for the measurement of surface texture and shape.pdf (9.844Mb)

PUBLISHER_POLICY

Sensing
Surfaces with
GelSight

https://www.youtube.com/watch?v=S7gXih4XS7A

Johnson, M.K., and E.H. Adelson. “Retrographic sensing for the measurement of surface texture and shape.” Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. 2009. 1070-1077.

66

3D Vision: A little bit History

1974: shape from silhouettes [Baumgart, Ph.D]

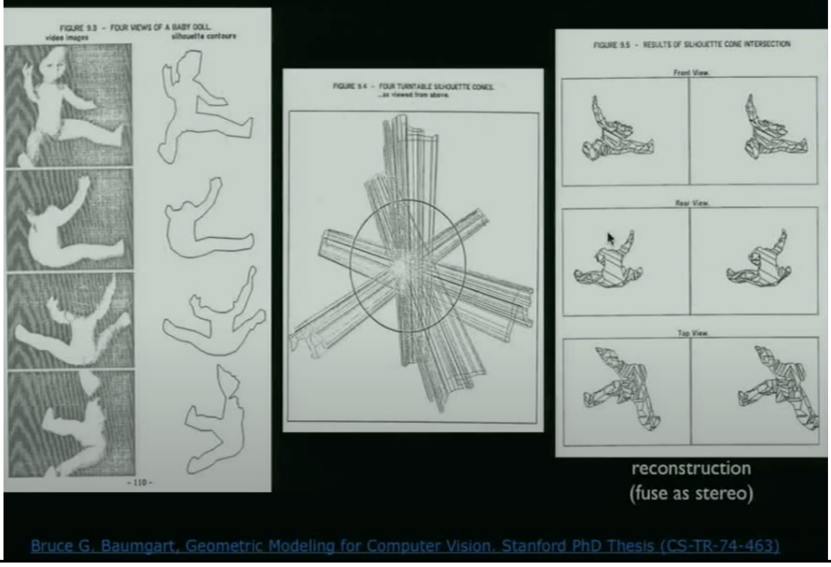
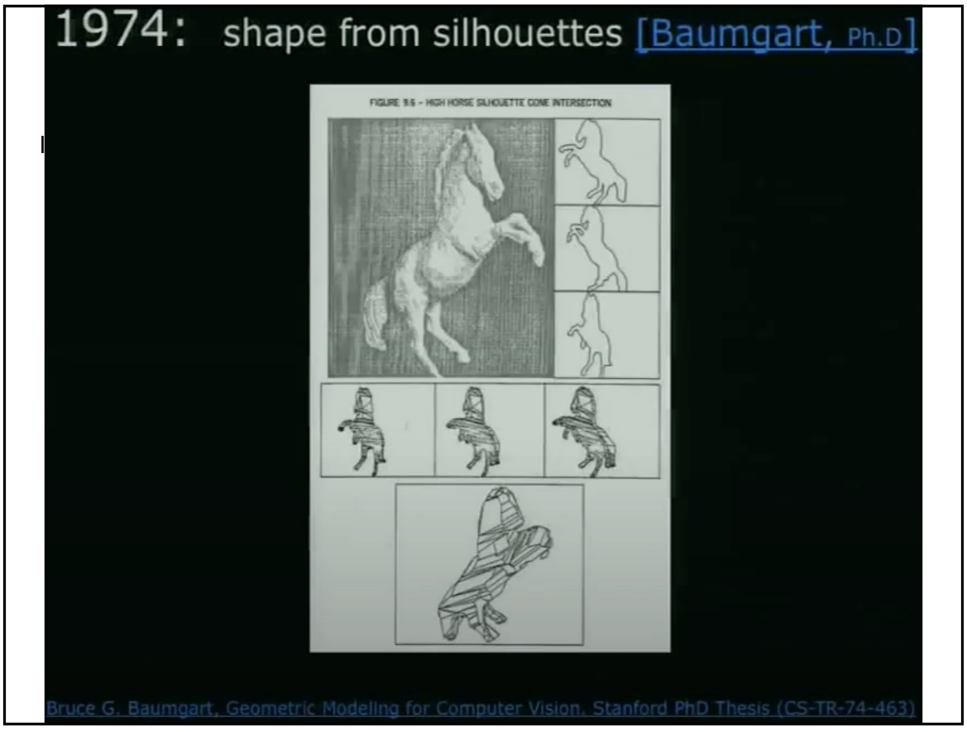


FIGURE 9.3 - FOUR VIEWS OF A BABY DOLL.
video images
silhouette contours
- 110 -
FIGURE 9.4 - FOUR TURNTABLE SILHOUETTE CONES.
...as viewed from above.
FIGURE 9.5 - RESULTS OF SILHOUETTE CONE INTERSECTION
Front View
Rear View
Top View

reconstruction
(fuse as stereo)

Bruce G. Baumgart, Geometric Modeling for Computer Vision, Stanford PhD Thesis (CS-TR-74-463)

67



1974: shape from silhouettes [Baumgart, Ph.D]

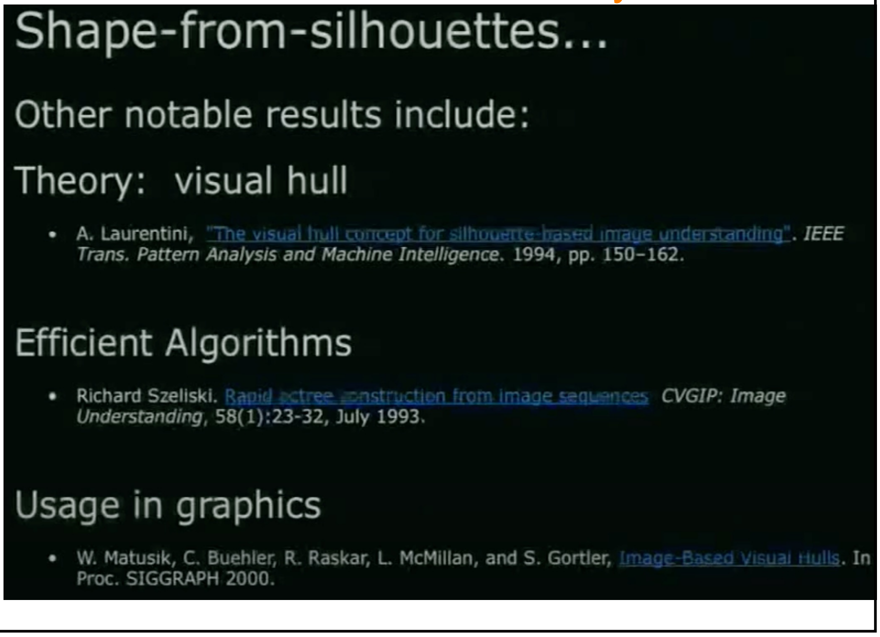
I

Bruce G. Baumgart, Geometric Modeling for Computer Vision, Stanford PhD Thesis (CS-TR-74-463)

68

3D Vision: A little bit History

In



Shape-from-silhouettes...

Other notable results include:

Theory: visual hull

- A. Laurentini, "The visual hull concept for silhouette-based image understanding". *IEEE Trans. Pattern Analysis and Machine Intelligence*. 1994, pp. 150-162.

Efficient Algorithms

- Richard Szeliski. Rapid octree construction from image sequences *CVGIP: Image Understanding*, 58(1):23-32, July 1993.

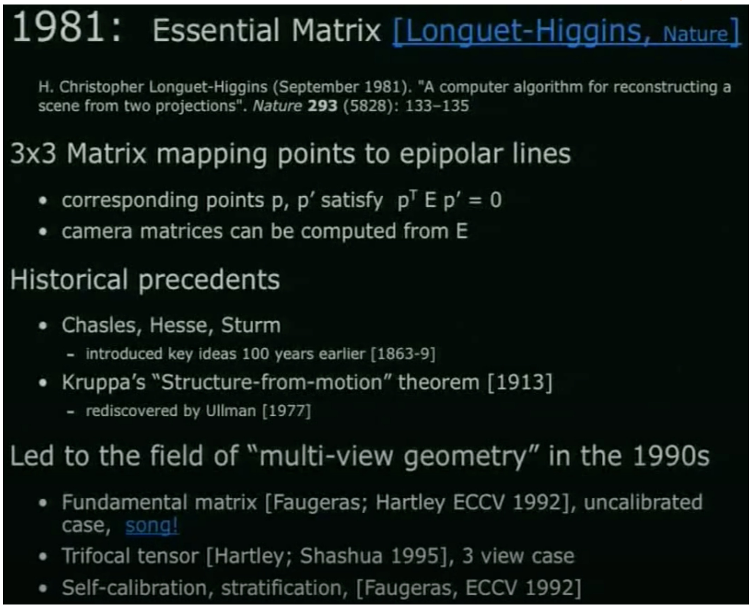
Usage in graphics

- W. Matusik, C. Buehler, R. Raskar, L. McMillan, and S. Gortler, Image-Based Visual Hulls. In Proc. SIGGRAPH 2000.

69

Next major landmark:

3D Vision: A little bit History



1981: Essential Matrix [Longuet-Higgins, Nature]

H. Christopher Longuet-Higgins (September 1981). "A computer algorithm for reconstructing a scene from two projections". *Nature* **293** (5828): 133–135

3x3 Matrix mapping points to epipolar lines

- corresponding points p, p' satisfy $p^T E p' = 0$
- camera matrices can be computed from E

Historical precedents

- Chasles, Hesse, Sturm
  - introduced key ideas 100 years earlier [1863-9]
- Kruppa's "Structure-from-motion" theorem [1913]
  - rediscovered by Ullman [1977]

Led to the field of "multi-view geometry" in the 1990s

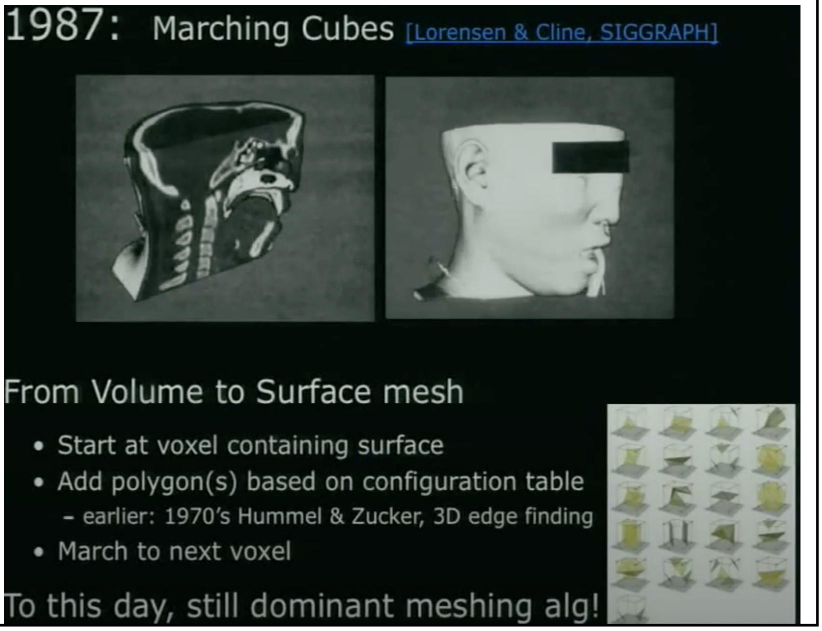
- Fundamental matrix [Faugeras; Hartley ECCV 1992], uncalibrated case, song!
- Trifocal tensor [Hartley; Shashua 1995], 3 view case
- Self-calibration, stratification, [Faugeras, ECCV 1992]

This work inspired an whole area of 3D computer vision, now known as Multiview Geometry

70

3D Vision: A little bit History

Q: Maybe not 3D Vision?



1987: Marching Cubes [Lorensen & Cline, SIGGRAPH]

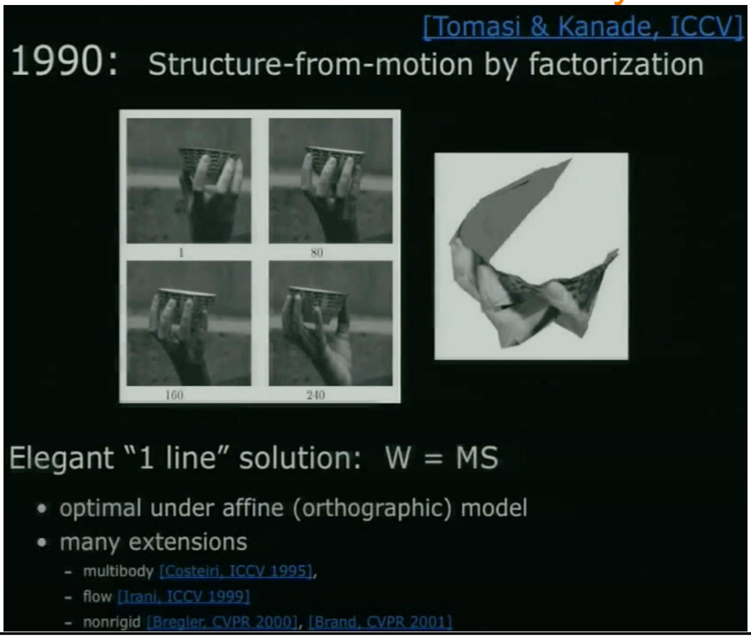
From Volume to Surface mesh

- Start at voxel containing surface
- Add polygon(s) based on configuration table
  - earlier: 1970's Hummel & Zucker, 3D edge finding
- March to next voxel

To this day, still dominant meshing alg!

71

3D Vision: A little bit History



[Tomasi & Kanade, ICCV]

1990: Structure-from-motion by factorization

Elegant "1 line" solution: W = MS

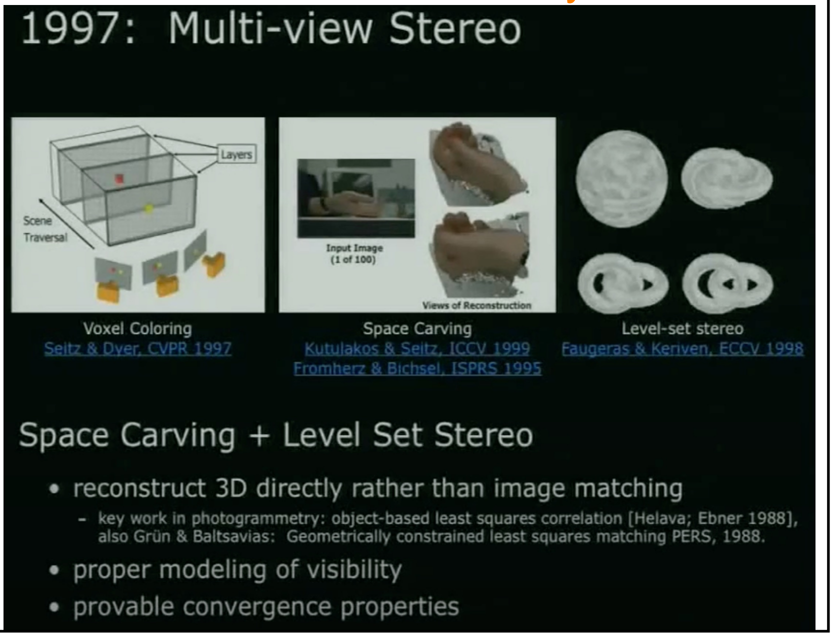
- optimal under affine (orthographic) model
- many extensions
  - multibody [Costeiri, ICCV 1995],
  - flow [Irani, ICCV 1999]
  - nonrigid [Bregler, CVPR 2000], [Brand, CVPR 2001]

W matrix: positions of points in a set of video images. M motion of the camera (affine model), S is the shape of the scene.

Not the modern SfM

72

3D Vision: A little bit History



1997: Multi-view Stereo

Voxel Coloring
Seitz & Dyer, CVPR 1997

Space Carving
Kutulakos & Seitz, ICCV 1999
Fromherz & Bichsel, ISPRS 1995

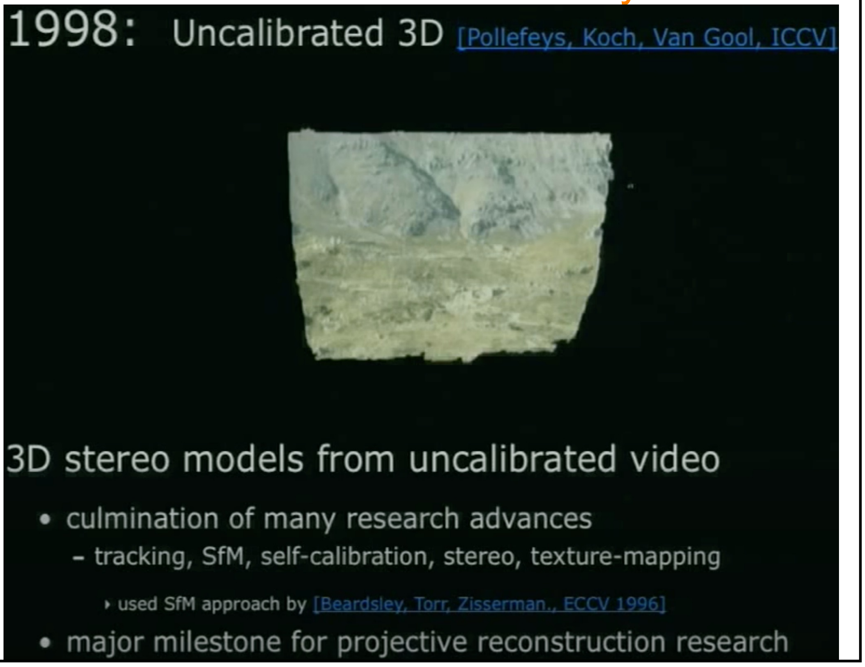
Level-set stereo
Faugeras & Keriven, ECCV 1998

Space Carving + Level Set Stereo

- reconstruct 3D directly rather than image matching
  - key work in photogrammetry: object-based least squares correlation [Helava; Ebner 1988], also Grün & Baltsavias: Geometrically constrained least squares matching PERS, 1988.
- proper modeling of visibility
- provable convergence properties

73

3D Vision: A little bit History



1998: Uncalibrated 3D [Pollefeys, Koch, Van Gool, ICCV]

3D stereo models from uncalibrated video

- culmination of many research advances
  - tracking, SfM, self-calibration, stereo, texture-mapping
    - used SfM approach by [Beardsley, Torr, Zisserman., ECCV 1996]
- major milestone for projective reconstruction research

74

THE END

75