1. Introduction

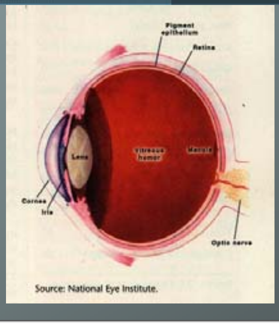
Computer Vision

Zoltan Kato

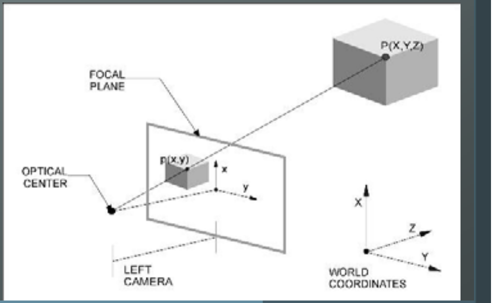
http://www.inf.u-szeged.hu/~kato/

What is Vision?

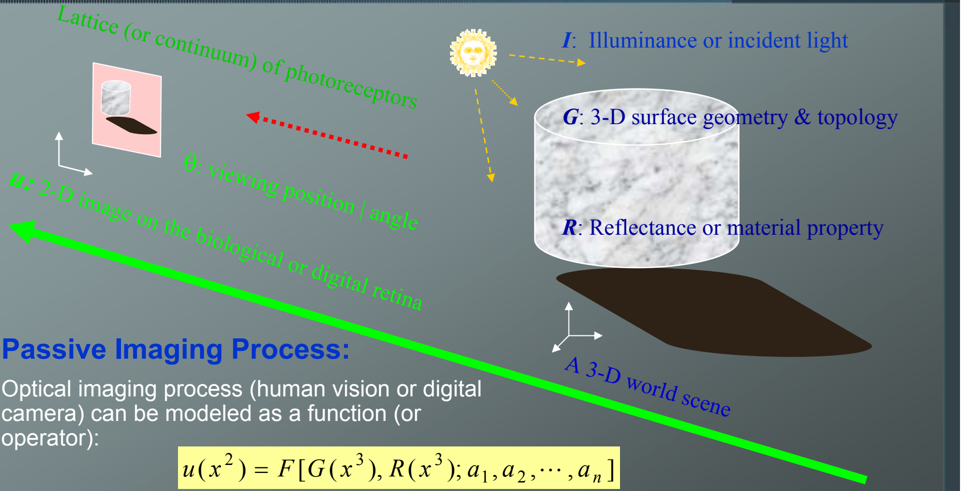
- The perception of the 3-D world from its 2-D partial projections onto the left and right retinas
  - fundamentally an **illposed inverse** problem.
    - But after millions of years' of evolution, human vision has become astonishingly accurate and satisfactory.
  - How could it have become such a remarkable inverse problem solver, and
  - what are the hidden (or subconscious) regularization techniques it employs?



Source: National Eye Institute.



Biological or Digital (Passive) Imaging Process


Lattice (or continuum) of photoreceptors
I: Illuminance or incident light
G: 3-D surface geometry & topology
θ: viewing position | angle
R: Reflectance or material property
u: 2-D image on the biological or digital retina
A 3-D world scene

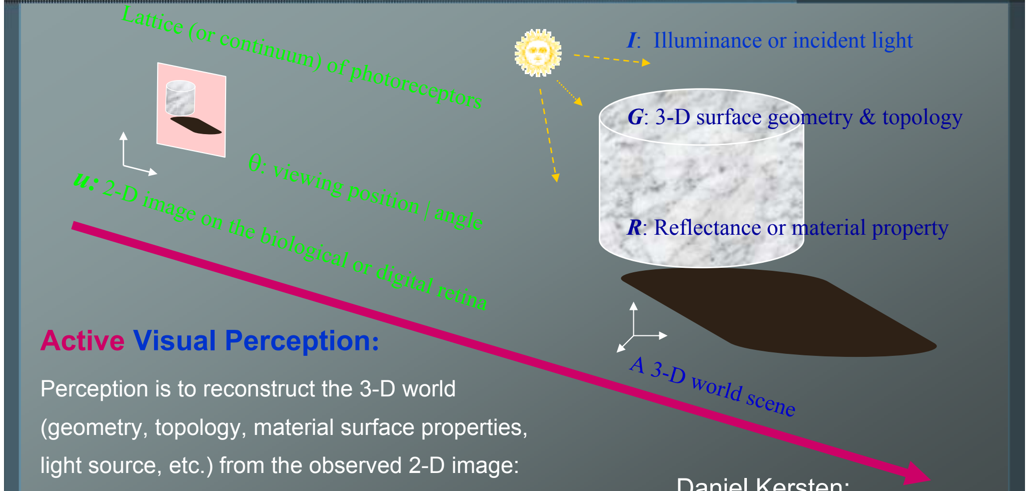
Passive Imaging Process:

Optical imaging process (human vision or digital camera) can be modeled as a function (or operator):

$$u(x^2) = F[G(x^3), R(x^3); a_1, a_2, \cdots, a_n]$$

Here, 2 and 3 indicate the dimension of the spatial variables. ***G*** and ***R*** denote the configuration (scene | geometry) and the reflectance. All the ***a***'s are parameters such as ***I*** and ***θ***.

Active Visual Perception


Lattice (or continuum) of photoreceptors
I: Illuminance or incident light
G: 3-D surface geometry & topology
θ: viewing position | angle
R: Reflectance or material property
u: 2-D image on the biological or digital retina
A 3-D world scene

Active Visual Perception:

Perception is to reconstruct the 3-D world (geometry, topology, material surface properties, light source, etc.) from the observed 2-D image:

$$[G(x^3), R(x^3); a_1, a_2, \cdots, a_n] = V[u(x^2)], \quad \text{or}$$
$$[G(x^3), R(x^3)] = V[u(x^2); a_1, a_2, \cdots, a_n].$$

Daniel Kersten:
Visual Perception is an *Inverse* Computer Graphics Problem

Vision is an *Illposed* Inverse Problem

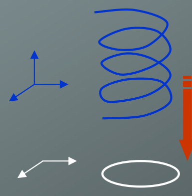
A. Geometry is not invertible: depth or range is lost !

Mathematical Model (1): *Projective Imaging*

P: $(x_1, x_2, x_3) \rightarrow (x_1, x_2)$ is not invertible!

For any given 2-D curve γ_2, there are infinitely many 3-D curves γ_3, so that $P(\gamma_3) = \gamma_2$.

Example.

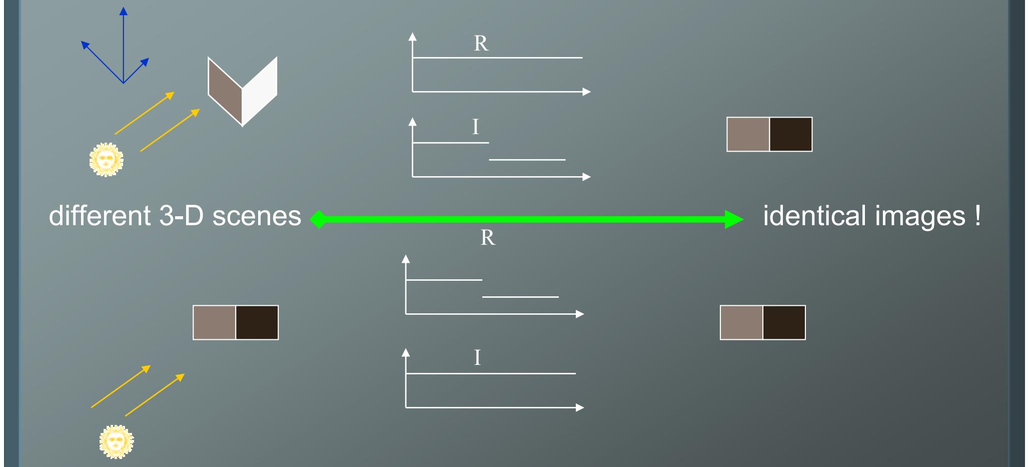


$\gamma_3 = (\cos t, \sin t, t)$

γ_2 is just like the projection of a circle

Vision is an *Illposed* Inverse Problem

B. The **Reflectance-Illuminance** entanglement.



different 3-D scenes ↔ identical images !

Human Vision

The human vision system is a well developed system of software and hardware, which can solve this highly ill-posed inverse problem efficiently and robustly.

Fundamental Questions:

- What kind of regularization techniques the human vision system employs to conquer the illposedness?
- What kind of features or variables to regularize ?

Perception as Bayesian Inference

Bayesian Perception (D. Kersten, A. Yuille, D. Mumford,…): X = (geometry G, reflectance R, …).

$$\max_{X} \text{Prob}(X \mid U_0) = \frac{\text{Prob}(U_0 \mid X)\,\underline{\text{Prob}(X)}}{\text{Prob}(U_0)}$$

It is our *a priori* knowledge of the world (i.e., Prob(X)) that regularizes our visual perception!

A Priori Knowledge (Common Sense) of the World Regularizes Vision

- G: Knowledge of curve and shape geometry;
- I: Knowledge of light sources & illuminance (sun, lamps, indoor or outdoor, …);
- R: Knowledge of materials (wood, bricks, …) and surface reflectance (metal shines and water sparkles, …)
- Θ: Knowledge of the viewers (often standing perpendicularly to the ground, viewing more horizontally, several feet away for indoor scenes, …)
- ……

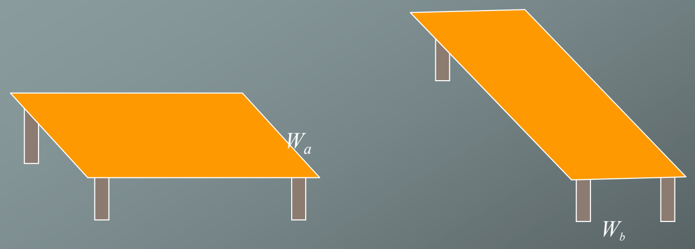
What to Learn & How

- How does the vision system subconsciously choose what knowledge to learn and store, out of massive visual data in daily life?
- For such knowledge, to which degree of regularity or compression that the vision system “decides” to process, in order to achieve maximum efficiency and robustness?
- How to mathematically model (or quantify) such activities?

Perception and Illusion

- Size, length, angles
- Order & depth
- Shade perception
- Grouping
- Interpolation & Continuation
- Light & Surface
- Texture gradient
- Linear perspective

Illusions: Size, Length, Angles

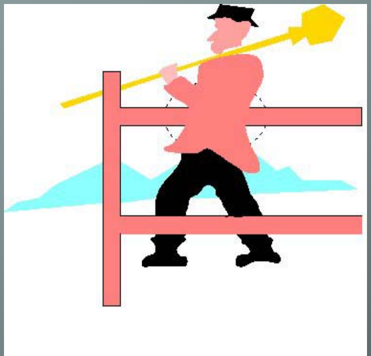


Human visual perception: the length W_a is longer than W_b.

Fact: the two shaded surfaces are *identical* up to a rotation !

Question: How to model humans’ perception of geometry?

Illusion: Order & Depth



Kanizsa's entangled man

G. Kanizsa [1978]

Nitzberg-Mumford-Shiota [1993]

Chan-Shen [*SIAP*, 2001]

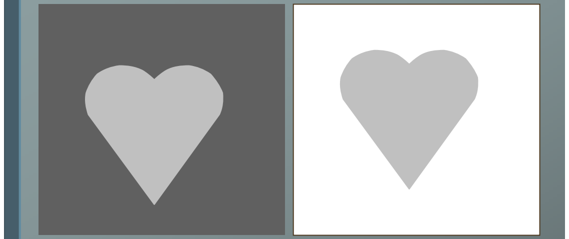
Human visual perception: the poor man is *entangled* in the fence.

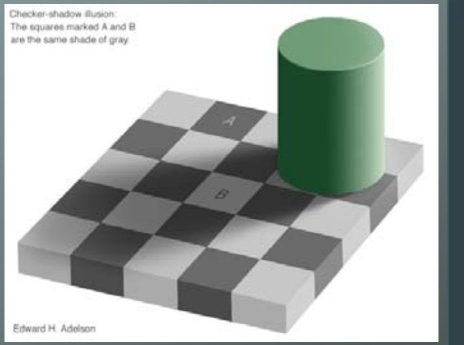
Fact: common sense tells us that he is *behind* the fence.

Question: How to model human's perception of depth? The lost 3rd D.

Illusion: Shade Perception

Ed. Adelson [MIT Cognitive & Comp. Neuron Sciences, 2000]





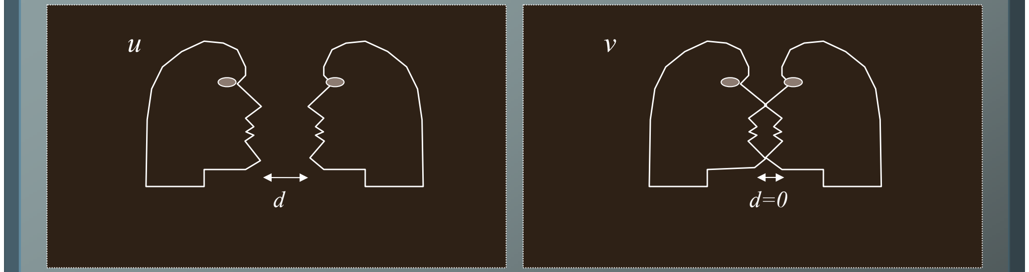
Human visual perception: the left heart is brighter than the right.

Fact: they are *identical*, with the same shape, size, and shade!

Question: How to model humans' perception of light and shades?
(i.e., how do our neurons encode and compute photon inputs?)

Visual Organization: Grouping

D. Kersten [Vision Psychology & Computation, UMN]

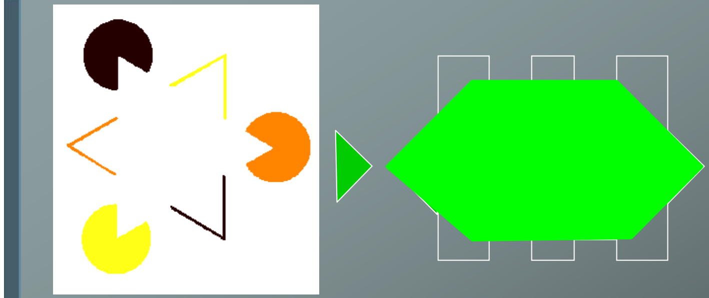


For image u, we "see" two human faces. No doubt.

If we move the face contours closer till touch (image *v*), do we still percept two human faces? Or two simians? Psychologists show that most of us percept the latter. (Gestalt Vision Science).

Question: How to model humans' perception of topology and grouping?

Visual Interpolation & Continuation



G. Kanizsa [1978]

Nitzberg-Mumford-Shiota [1993]

Chan-Shen [*SIAP*, 2001]

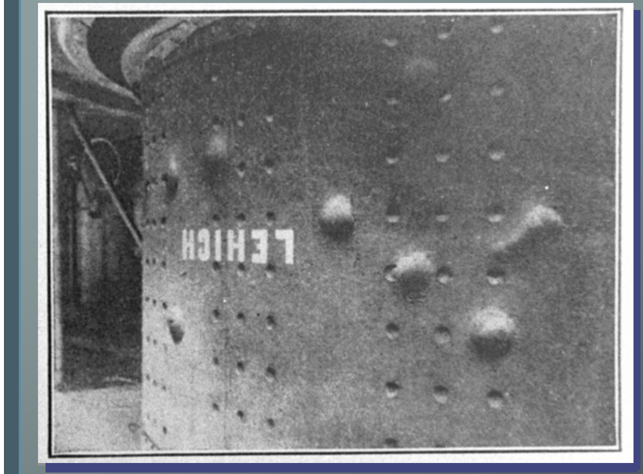
Chan-Kang-Shen [*Visual Comm.*, 2001]

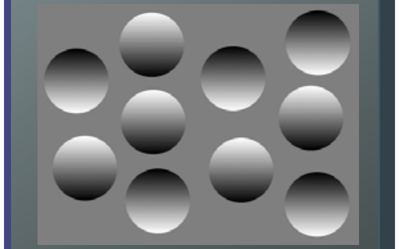
Interpolation is universal in daily life due to object occlusions in 3-D.

Clinical evidences show that it is not born with us. Many patients with vision defects have difficulty in connecting all the broken and separated parts. To them, the world is piece by piece.

Question: How to model the interpolation algorithms of vision neurons?

Depth Cues: Light & Surface

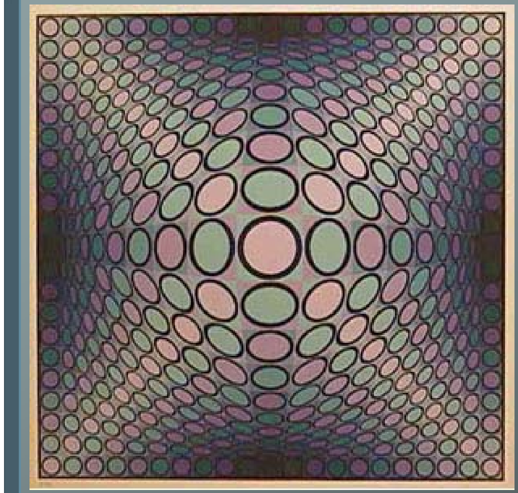


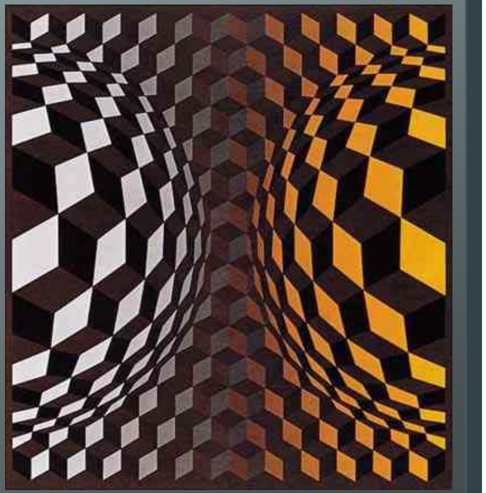


Illumination gradients: gradients and shadow lend a sense of depth

Slide adopted from Jackie Shen: Mathematical Image & Vision Analysis.

Depth Cues: Texture Gradient

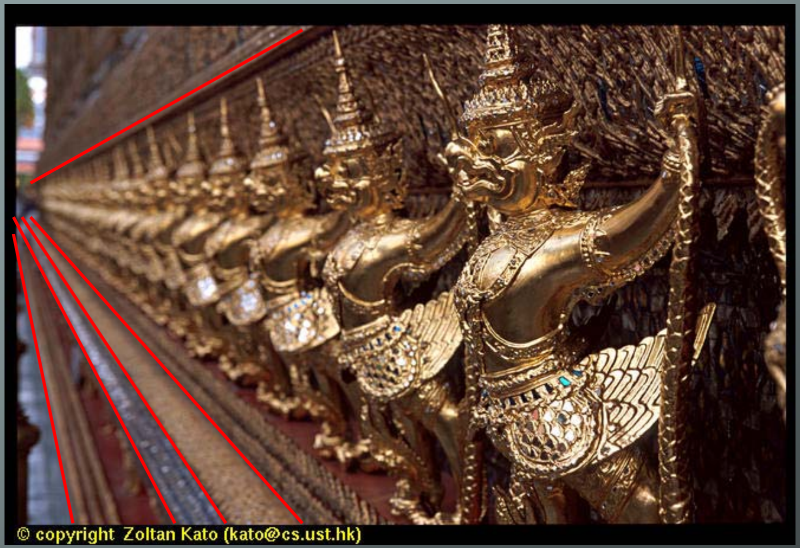




Linear Perspective: objects appear smaller as they recede into the distance

Depth Cues: Linear Perspective

Wat Phra Kaew, Bangkok, Thailand – http://www.cameradigita.com/



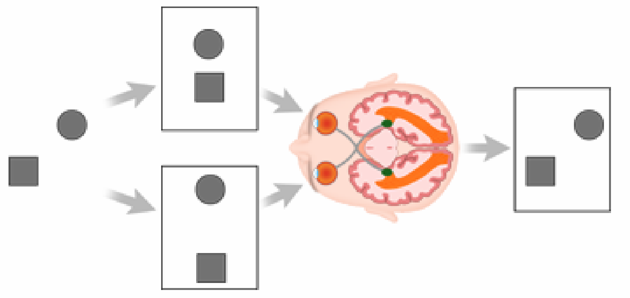
Lines will appear to draw closer together as they go farther into the distance.

Depth Cues: Motion Parallax

- differential perception of motion (speed and direction) as a function of distance from perceiver
  - Objects moving faster are appearing closer

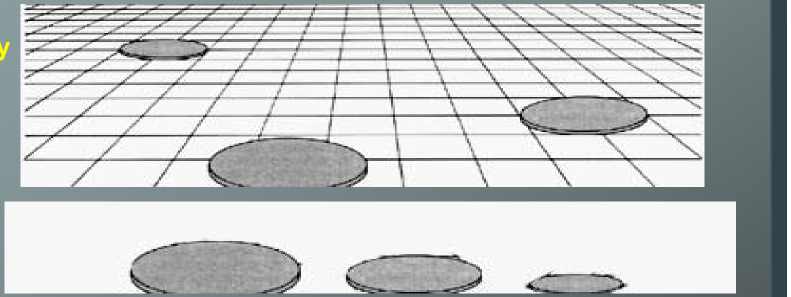
http://epsych.msstate.edu/descriptive/Vision/mparallax/DC1.html

Depth Cues: Binocular Disparity



- Each eye receives a slightly different image of the world from which 3D positions can be inferred.
- **Disparity** - The difference in retinal position between the corresponding points in two images. Disparity is inversely proportional to the depth of the point in space.

Perceptual Constancy

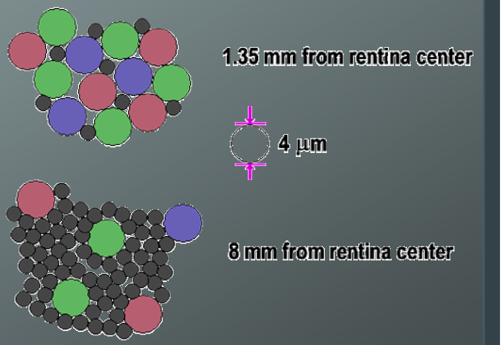


Size constancy

- We tend to experience objects as the same, despite the image they produce on the retina may vary greatly.
  - *Shape constancy* – objects seen from different angles do not appear different or to change shape
  - *Size constancy* – objects do not seem to change size when they move nearer or further away.
  - *Color constancy* – differing illumination does not affect color despite changes in the actual reflected light.

Theory of Visual Perception

- The information we receive by our eyes is relatively impoverished.
  - For example, the retina receives a grainy 2D image of the visual scene…
  - …that includes large gaps (blind spots)…
  - …and an uneven representation of colour (cones) and luminance (rods).
- This information is transformed into a rich visual experience.



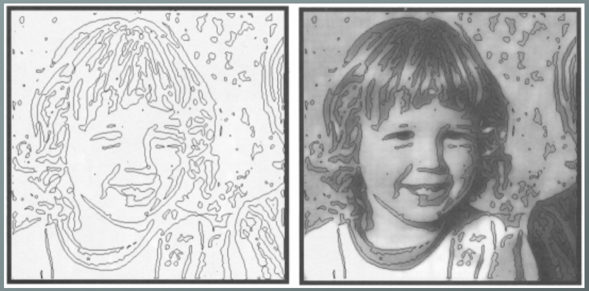
Theory of Visual Perception: Marr

- Theories of visual perception attempt to explain how this happens.
- **David Marr**: *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information, 1982.*
- wanted to understand mechanisms of vision rather than just behaviours associated with it.
  - He took an information processing view of the mind…
  - …and aimed to describe perception in terms of computations on sense data…
  - …to extract high level visual experience.

Marr: Computational Approach

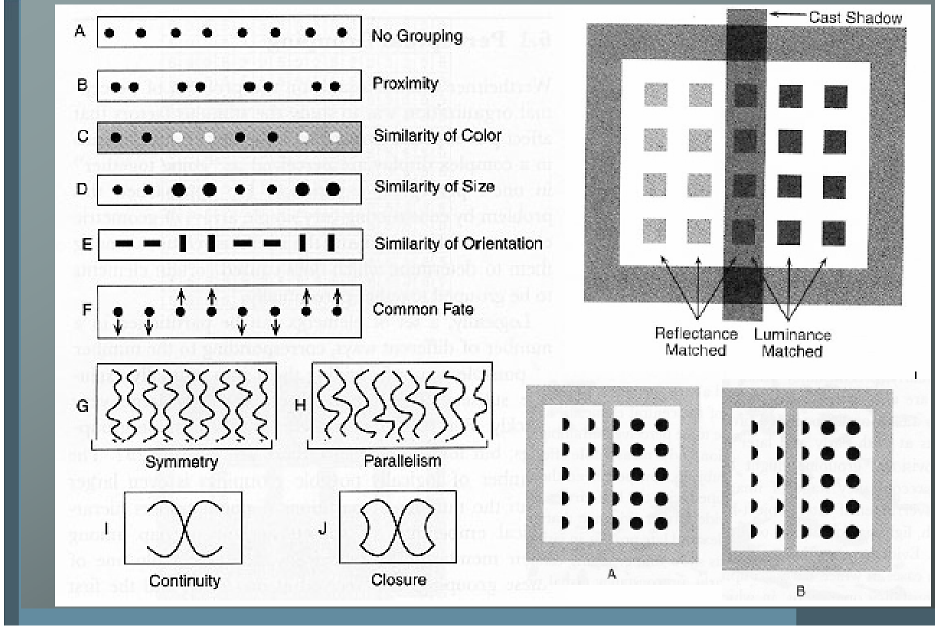
- Marr proposed there were distinct stages of processing in visual perception (Bottom-Up):
  - Raw Primal Sketch
  - Complete Primal Sketch
  - 2½D Sketch
  - Full 3D Representation

Marr: Primal Sketch



- *Early primal sketch* involves the extraction of information regarding edges and intensity changes.
- Then a *Complete Primal Sketch* is created by grouping surfaces and common areas
  - The Gestalt Psychologists of the early 19th Century demonstrated many different ways in which we can group objects.

Gestalt laws of Perceptual Grouping



Marr: 2½D Sketch

- After gaining information about groupings and surfaces the viewer needs some spatial information.
- Marr called this stage the *2½D Sketch* to emphasis that this stage didn’t give a full 3D representation.
- Just an estimate of the spatial locations of objects and materials in relation to the viewer.
  - Depth cues

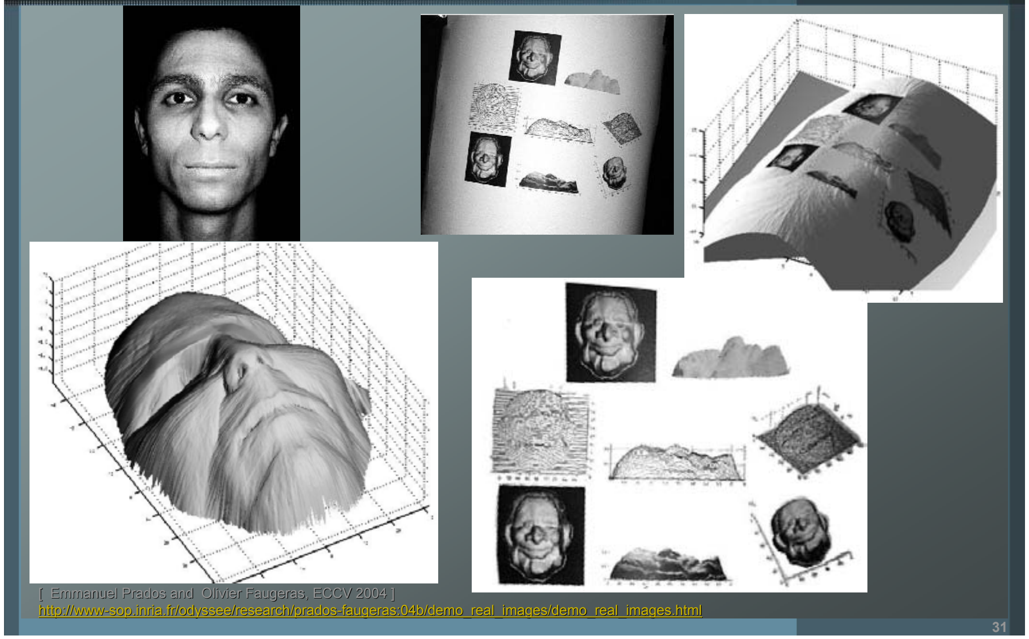
Marr: 3D Model Representation

- A full 3D description of our spatial environment involving the
  - identification of the structure of objects and
  - materials in our visual scene.
- It allows us to work out the 3D environment from a non-egocentric point-of-view.

Can Computers see as Humans?

- Shape from shading
- Shape from texture
- Single view reconstruction
- 2-view reconstruction (stereo vision)
- Multi-view reconstruction
- Tracking moving objects
- Video mosaics
- Video editing/inpainting
- View synthesis

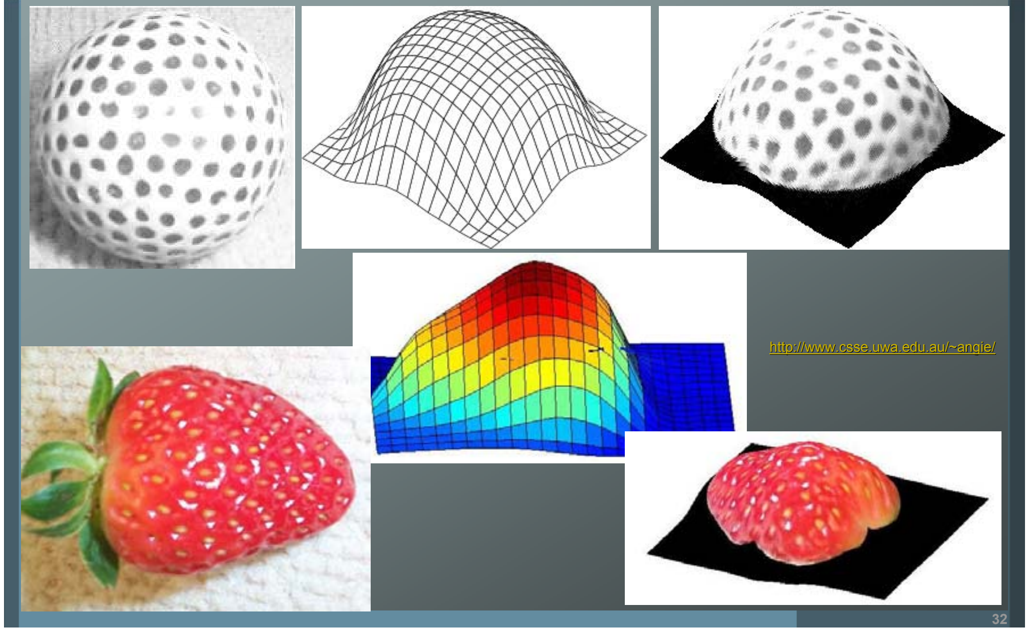
Shape from Shading



[Emmanuel Prados and Olivier Faugeras, ECCV 2004]

http://www-sop.inria.fr/odyssee/research/prados-faugeras:04b/demo_real_images/demo_real_images.html

Shape from Texture



http://www.csse.uwa.edu.au/~angie/

Single View Reconstruction

Reconstructing a 3D scene from recognizable geometric primitives such as lines, planes and spheres by computing their spatial layout given only one view.

Fellows Quad in Merton College, Oxford

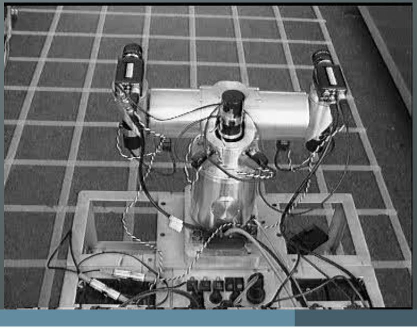
3D reconstruction of a photograph

La Flagellazione di Cristo (1460) by Piero della Francesca (1416-1492)
Galleria Nazionale delle Marche

3D reconstruction of a painting

D.Liebowitz, A. Criminisi, and A. Zisserman, Visual Geometry Group, Oxford University, UK, 1998 - http://www.robots.ox.ac.uk/~vgg/projects/SingleView/

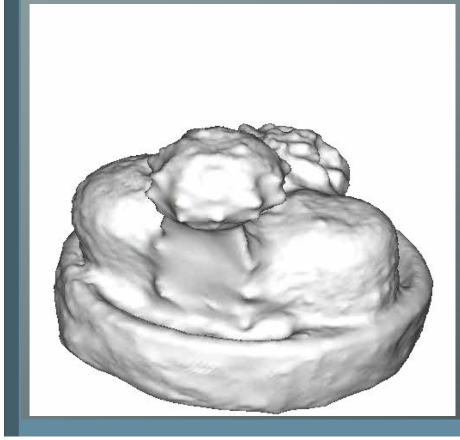
Mobile Robot Navigation

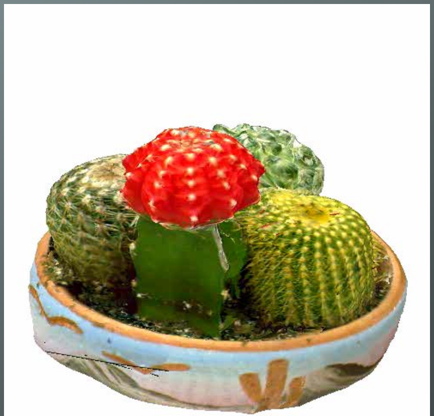


http://www.robots.ox.ac.uk/ActiveVision/Projects/Nav/nav.01/

Multi-view Stereovision

[J.-P. Pons, R. Keriven and O. Faugeras CVPR 2005]
http://www-sop.inria.fr/odyssee/research/pons-keriven-etal:04b/demo/



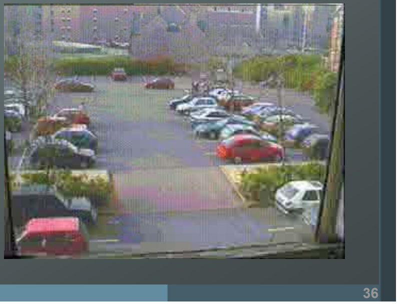


Tracking

Tracking results:
Fitting models to cars and people

An Integrated Traffic and Pedestrian Vision System, Leeds – Reeding, UK, 1997
http://www.scs.leeds.ac.uk/imv/index.html

The Likelihood of a particular trajectory
Dots above the heads indicate likelihood of a particular trajectory.

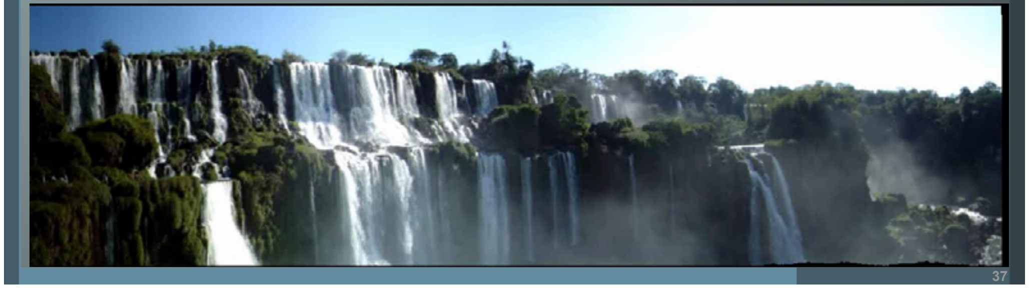


Video Mosaic/Panorama

Dynamosaics: Video Mosaics with Non-Chronological Time [CVPR 2005]
http://www.vision.huji.ac.il/dynmos/

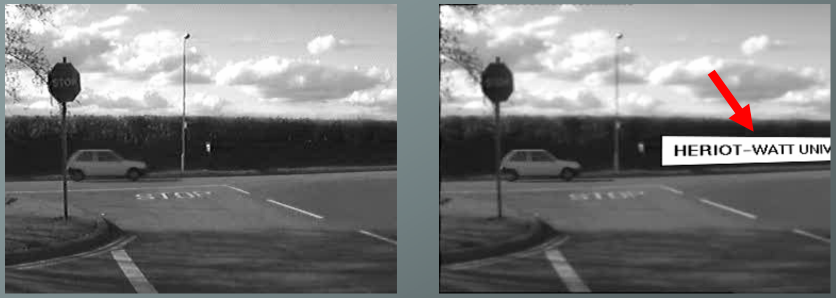
Alex Rav-Acha, Yael Pritch, Dani Lischinski, Shmuel Peleg
Hebrew University of Jerusalem, Israel

With the limited field of view of human vision, our perception of most scenes is built over time while our eyes are scanning the scene. In the case of static scenes this process can be modeled by panoramic mosaicing: stitching together images into a panoramic view. Can a dynamic scene, scanned by a video camera, be represented with a dynamic panoramic video even though different regions were visible at different times?



Video Editing

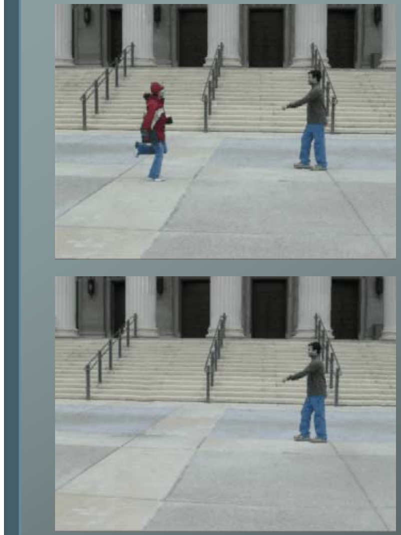
http://profs.sci.univr.it/~fusiello/demo/mosaics/



Insert new objects into a video sequence

Video Inpainting

Guillermo Sapiro, Kedar A. Patwardhan [ICIP 2005]
http://www.tc.umn.edu/~patw0007/icip2005/



Remove unwanted objects

Yunjun Zhang, Jiangjian Xiao, Mubarak Shah –[UCF 2005]
http://www.cs.ucf.edu/%7Evision/projects/ImageVideoCompletion/

View Synthesis: EyeVision of CMU

Takeo kanade [Carnegie Mellon University, USA, 2001]
http://www.ri.cmu.edu/events/sb35/tksuperbowl.html

The action was captured by more than 30 cameras, each with computer-controlled zoom and focus capabilities mounted on a custom-built, robotic pan-tilt head.

These camera heads were controlled in concert so that cameras pointed, zoomed and focused at the same time on the same spot on the field, where a touchdown or fumble occurred.

The detailed geometrical information about a scene is extracted by computer, which enables a person to choose how to view a scene, even from a perspective that was not shot by any camera.