Perceptual Grouping in Computer Vision

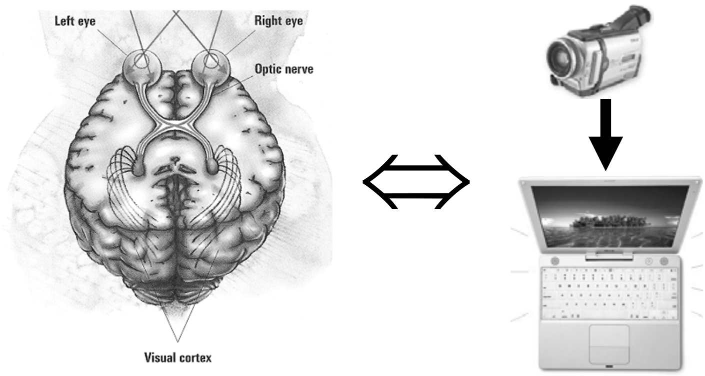
Gérard Medioni

University of Southern California

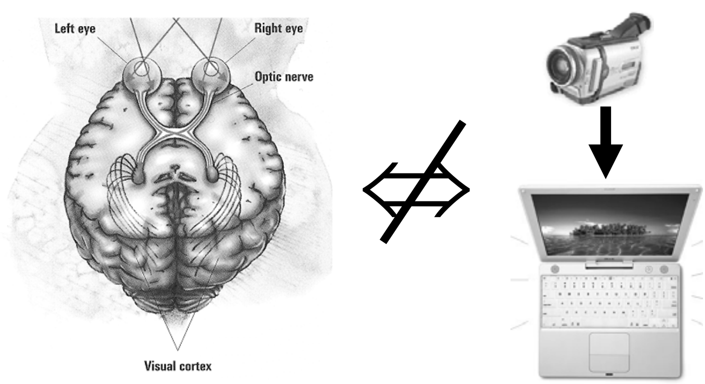
What is Computer Vision?

- Computer Vision
  - Attempt to emulate Human Visual System
  - Perceive visual stimuli with cameras instead of eyes
  - Apply laws and constraints to analyze and interpret what is being seen

Human vs. Computer Vision

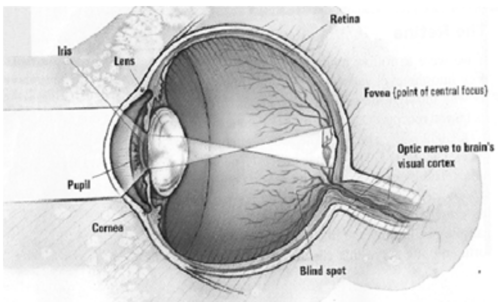


Human vs. Computer Vision



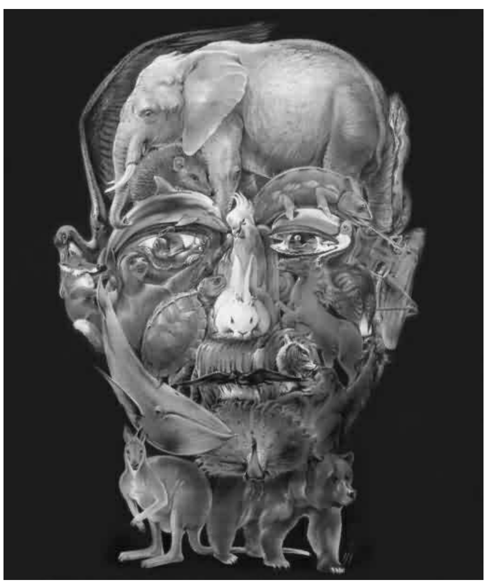
Wishful thinking, but incorrect

Human Visual System



- Eyes
  - Perceive stimuli
  - Perform low-level processing
  - Dispatch processed information to the brain
- Brain
  - Recalls relevant memories
  - Incorporates prior knowledge in the analysis

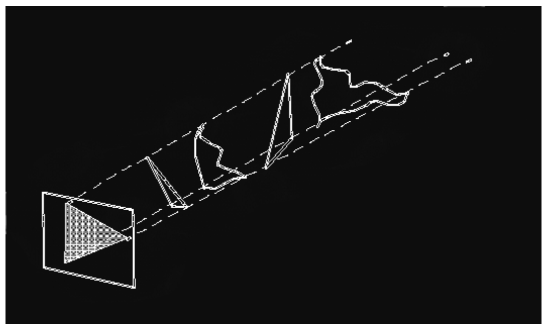
Vision is not that easy!



Do you see a bunch of animals or a human face?

What is hard about vision?

- Mapping from 2-D to 3-D
  - Numerous shapes can generate the same projections on the image plane
- Inverse, ill-posed problem
- Mathematically provably impossible



The human visual system is not perfect

Which of the two stripes is larger?

The human visual system is not perfect

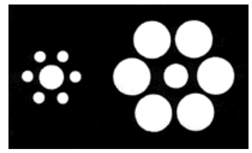
Which of the two stripes is larger?

Need for Constraints

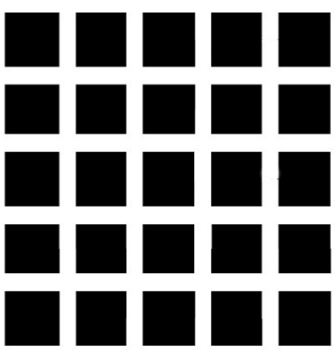
- Since the problem has infinite solutions, hard constraints need to be imposed
- Restrict the class of objects and scenes to the “most usual” or “simplest”
- For instance, the “matter is cohesive” constraint holds for the majority of natural scenes and is a basis for many algorithms
- Not all constraints apply to all scenes, leading to misinterpretations

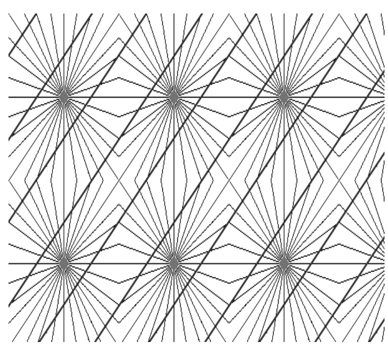
Illusions

- Visual illusions violate some of the constraints imposed by the Human Visual System
- They result from conflicts among the various constraints

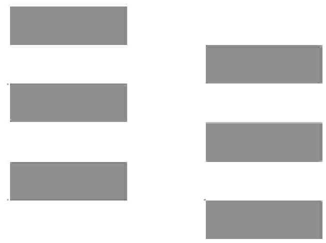


Visual Illusions

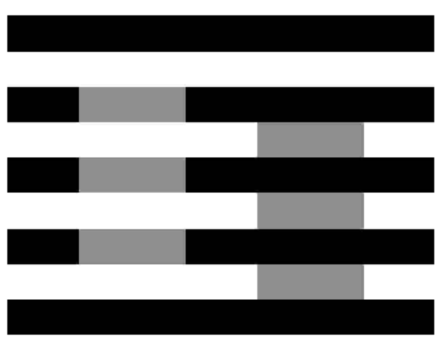




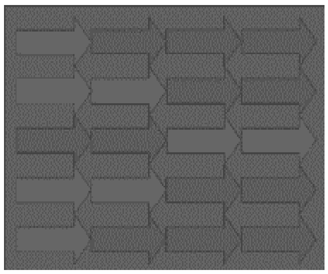
Simultaneous Contrast

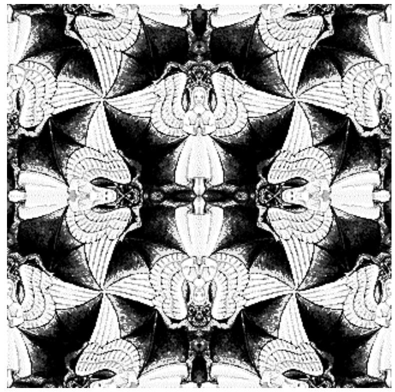


Simultaneous Contrast



Figure/Ground Segmentation





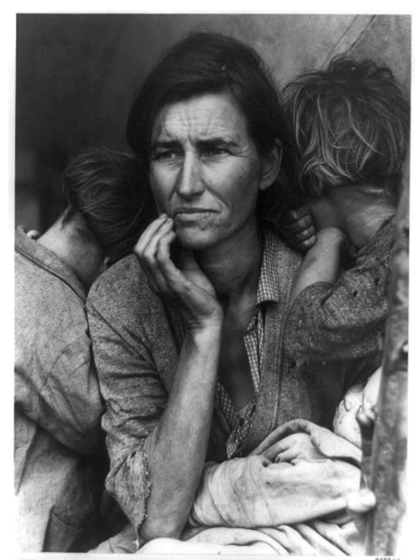
Impossible Drawings

Low level Vision

- Determination of local image properties
  - Smoothing
  - Thresholding
  - Edge detection
  - Color
  - Texture
- Pre-attentive retinal processes
- Occurs at the image level (2-D)

High level Vision

- Inference of scene description
- Semantic analysis and interpretation
- Inference of unseen details based on experience
- Usually 3-D processes



Mid level Vision

Low level Vision

Points
Curves
Edges
Colors
Textures

?

↓

High level Vision

Scene description
Interpretation

Mid level Vision

- Low level vision processes can be performed by image processing methods
- High level analysis is feasible given complete and accurate data
- The missing link is mid level vision that can convert low level data into objects and scene descriptions

Mid level Vision

Low level Vision

↓

Mid level Vision

↓

High level Vision

Boundaries
Regions
Surfaces
Objects

Perceptual Grouping

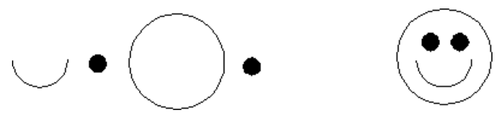
- Occurs after local primitives have been detected
- Pre-attentive process
- Does not involve cognition, it occurs without conscious effort

Perceptual grouping illusions

Perceptual grouping illusions

This is not a pre- attentive illusion

Gestalt Principles

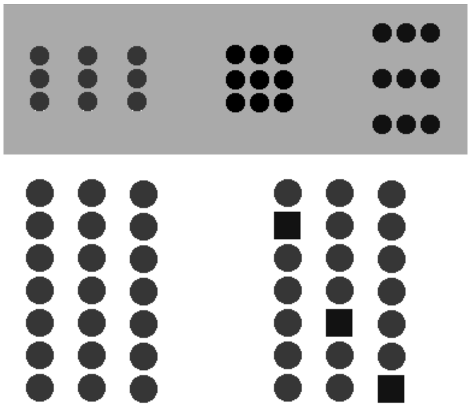


- "The whole is different from the sum of the parts"
- Established by psychological experiments
- Often conflicting

Gestalt Laws

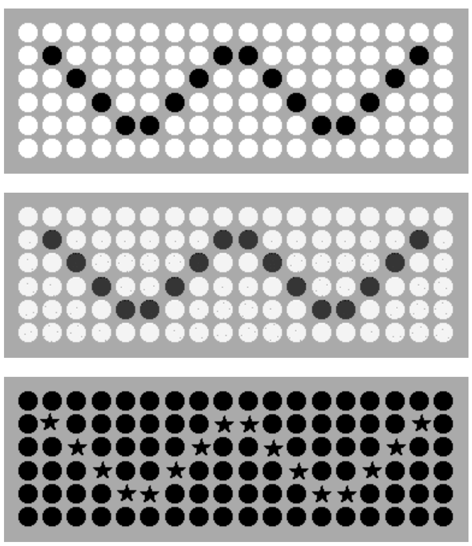
- Proximity
- Similarity
- Closure
- Good Continuation
- Simplicity or Good Form

Proximity



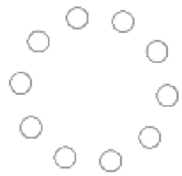
Objects that are close to one another tend to be grouped together

Similarity

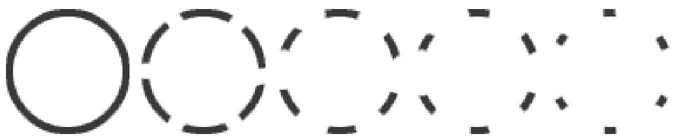


Objects that are more similar to one another tend to be grouped together

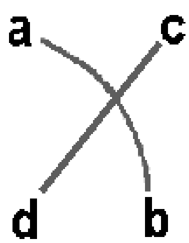
Closure



Stimuli tend to be grouped into complete figures



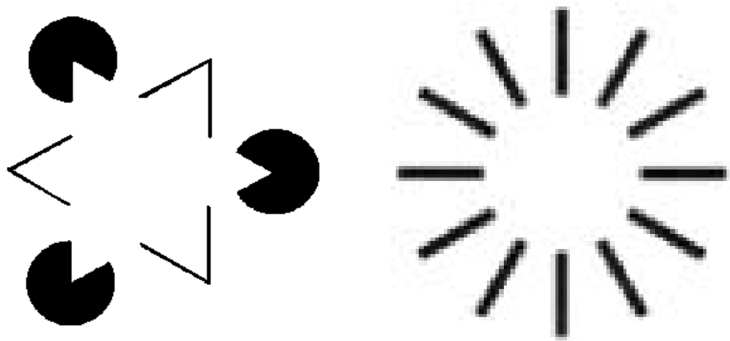
Good Continuation



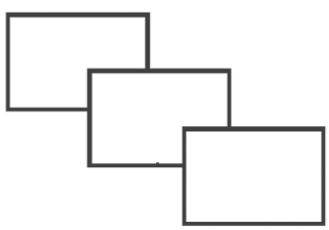
Objects organized in a straight line or a smooth curve tend to be perceived as a unit

The lines from a to b and c to d are the most salient perceptual groups in this image

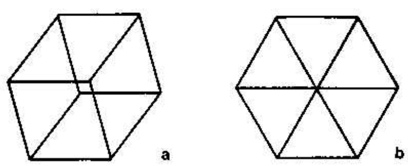
Illusory Contours



Simplicity or “good form”



The simplest interpretation organization of stimuli will be perceived



(a) is perceived as 3-D and (b) as 2-D

Approaches to Grouping in Computer Vision

- Regularization
- Relaxation labeling
- Clustering
- Robust methods
- Level Sets

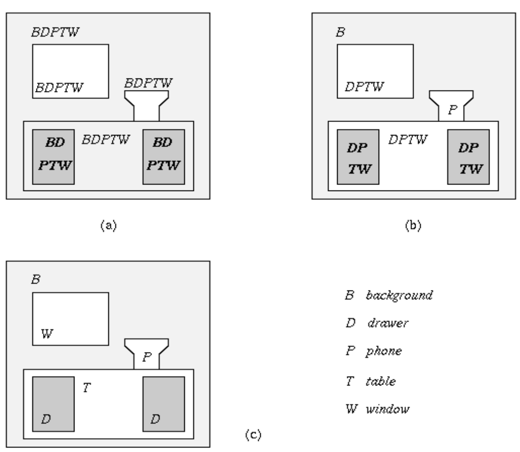
Regularization Approaches

- Define a “quality of fit” or error metric
- Express it as a function of some parameters
- Maximize objective or minimize error using numerical optimization techniques

Relaxation Labeling

- Begin by assigning all possible labels to every scene object
- Remove labels that are inconsistent based on unary constraints
- Remove labels that are inconsistent based on N-ary constraints
- Converge to consistent labeling of all objects

Relaxation Labeling Example



(a) Input with all possible labels
(b) Labels after unary constraints
(c) Labels after binary constraints

Clustering and Robust Methods

- Use statistical methods to explore the tendency of a point pattern to form compact groups
- Examine data sets for the presence of pre-specified configurations
- Detection is possible even in excessive corruption by noise
- For example, the Hough transform can be used to detect straight lines in point clouds

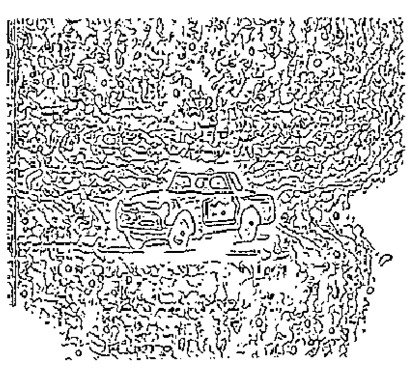
Level Set Methods

- Implicit representation of curves or surfaces
- Points on curves are on the zero level of function
- Common function used is the distance function
- Inherently multi-resolution representation
- Can handle topological changes

Structural Saliency

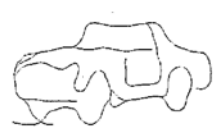
- Saliency literally means the quality of jumping out, being prominent
- Structural Saliency is a property of the structure as a whole
  - Parts of the structure are not salient in isolation
- Sha'ashua and Ullman defined a saliency measure based on curvature and curvature variation

Structural Saliency
(Sha'ashua and Ullman)



(a) Input image

(b) Saliency map



(c) The 5 most salient curves

Drawbacks of these methods

- Inherently exponential
  - With respect to the number of:
    - Images
    - Features
    - Pixels
- Iterative
- Not general due to enforcement of strict constraints

Overview of Tensor Voting

- Introduction
- Salient feature inference
  - tensor voting
  - 2D and 3D systems
- Applications
  - shape from shading
  - shape from stereo
  - optical flow
- Perspectives

Computational Approach

Grouping/Matching

$\Downarrow$

Optimization

$\Downarrow$

Constraints/Evaluation Criteria

Choice of Criteria

Matter is cohesive $\Rightarrow$

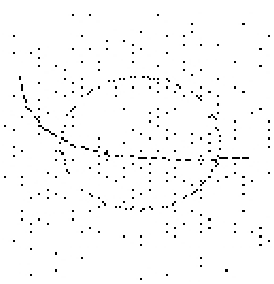
Smoothness

Objects are opaque $\Rightarrow$

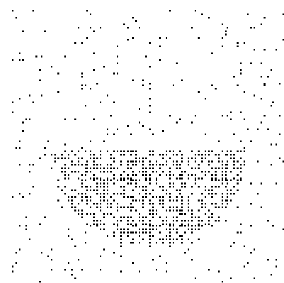
Uniqueness along line of sight

Implementation at the image level ?

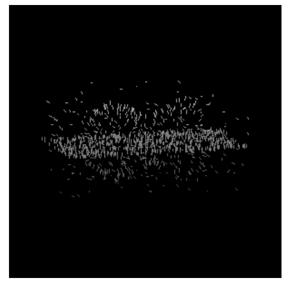
Examples



2-D curves



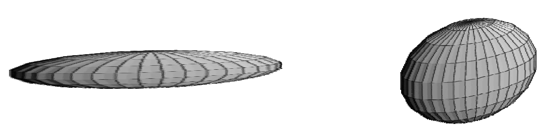
2-D regions



3-D surfaces

Tensor Voting

- Representation: 2nd order symmetric **Tensor**
  - shape: orientation certainty



  - **size:** feature saliency

Tensor Voting

- Constraint Representation: Voting fields
  - tensor fields
  - encode smoothness criteria

- Communication: **Voting**
  - non-iterative
  - no initialization

Our approach in a nutshell

- Each input site propagates its information in a neighborhood

- Each site collects the information cast there

- Salient features correspond to local extrema

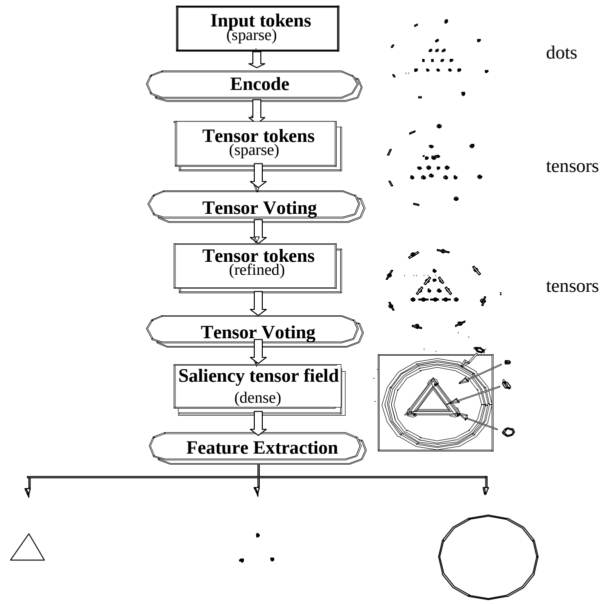
Properties of Tensor Voting

- Linear
- Non-Iterative
- Extract all features *simultaneously*
- 1 parameter (scale)
- Objective thresholds
- Efficient
  - O(1) for parallel computation

Overview

- Introduction
- Salient feature inference
  - tensor voting
  - 2D and 3D systems
- Applications
  - shape from shading
  - shape from stereo
  - optical flow
- Perspectives
  - curvature
  - *N*D

Overview

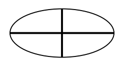


2-D Tensor Voting

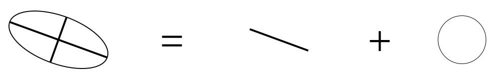
- Representation: 2-D **Tensor**
- Constraints: 2-D **Voting fields**
- Data communication: **Voting**

2-D Tensor

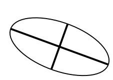
- ∘ and — are extreme cases of a tensor



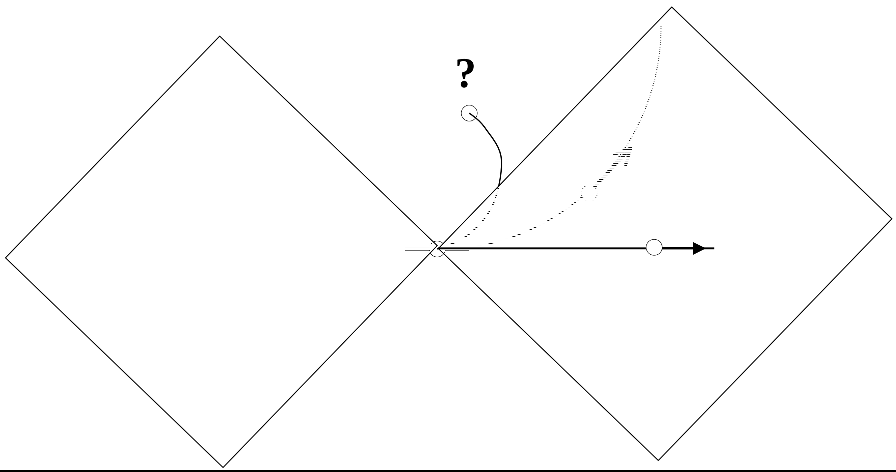
- Conversely, any tensor can be expressed as



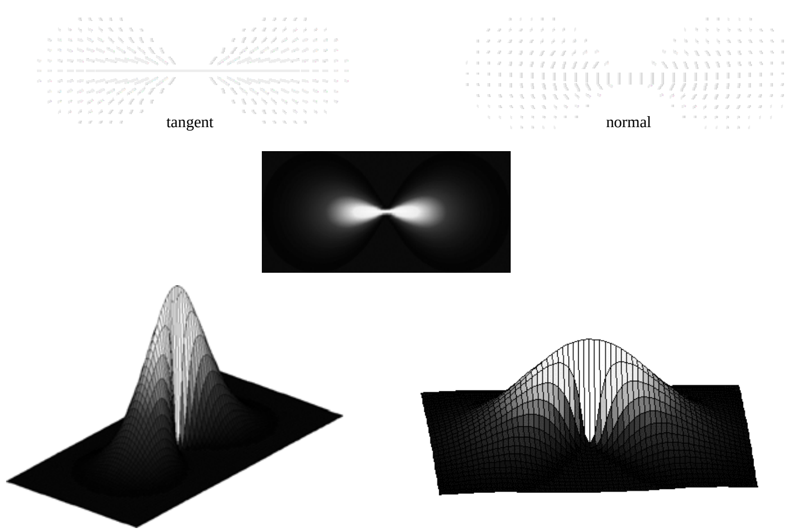
Decomposition & Interpretation

- in 2-D  is 3 numbers λ_{max}, λ_{min}, θ
- λ_{min} represents **orientation uncertainty**
- (λ_{max}- λ_{min}) represents **orientation saliency**

Design of Voting Field



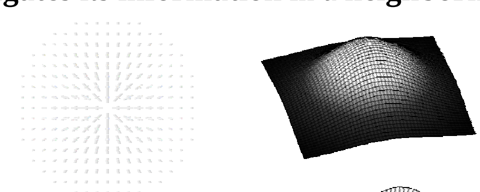
Fundamental Stick Voting Field

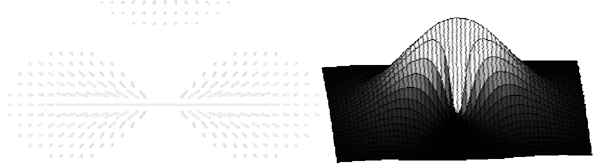


2-D Ball Field

2D Voting Fields

Each input site **propagates** its **information** in a **neighborhood**

votes with 

votes with 

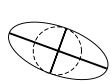
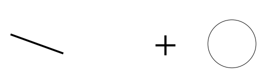
votes with 

Illustration of Voting?

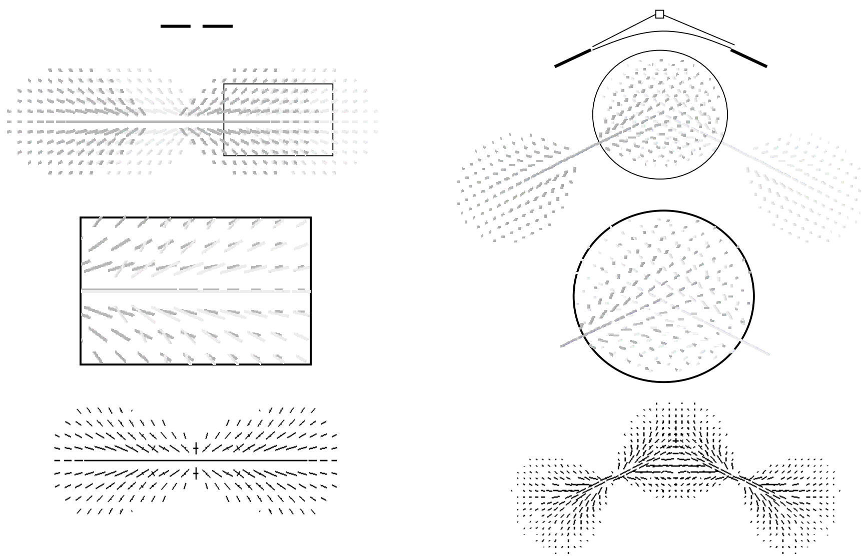
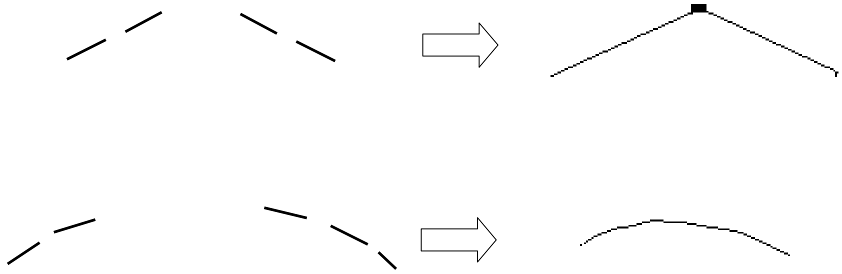


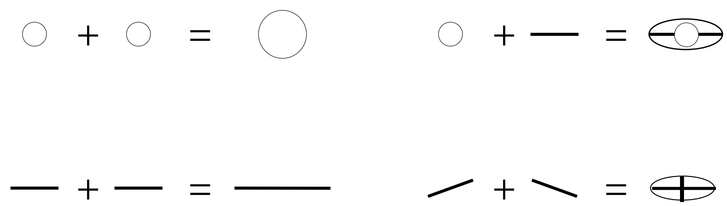
Illustration of Voting



Vote Collection

Each site **collects** the information cast there

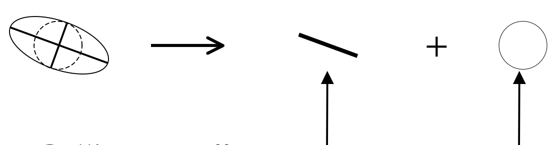
By tensor **addition** :



Vote Interpretation

Salient features correspond to local extrema

At each site

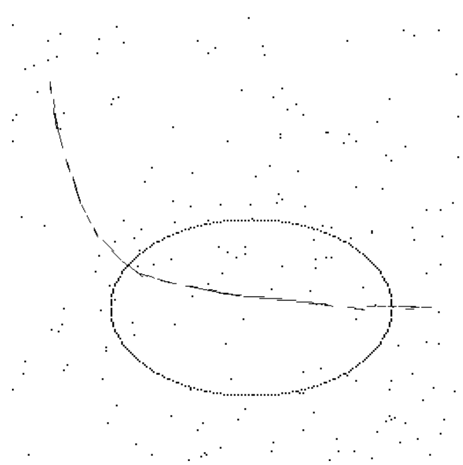


$\Rightarrow$ 2 “images” CMap JMap

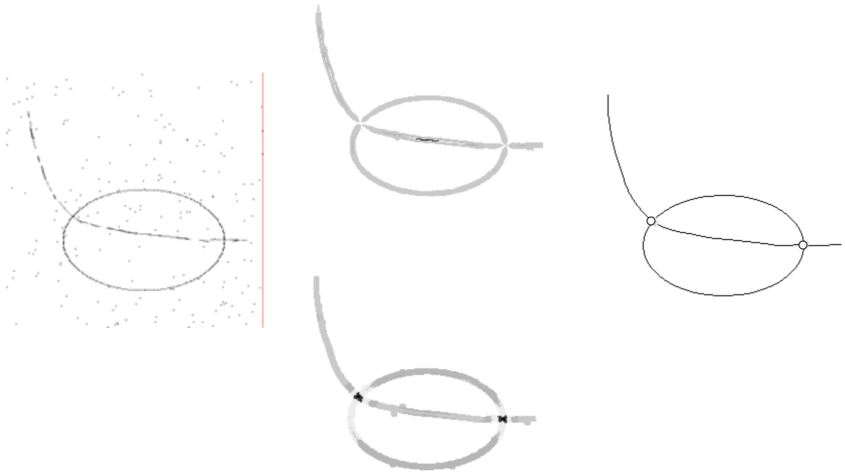
Feature Extraction

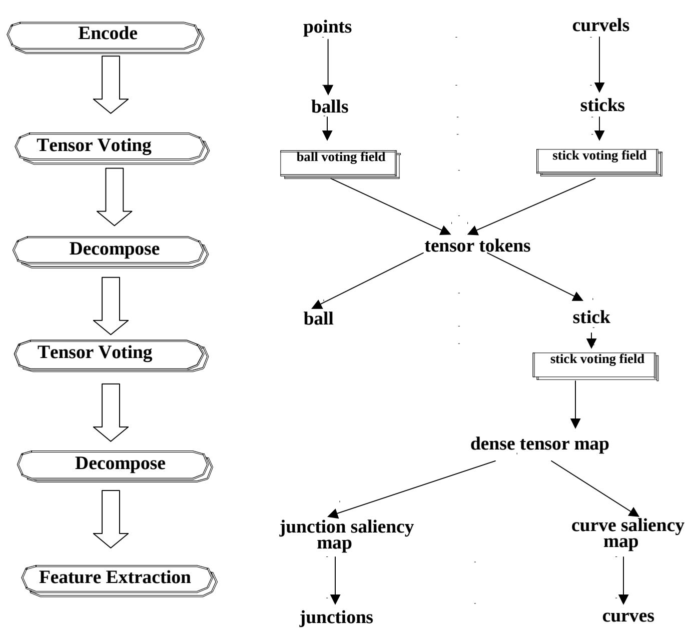
- **Curves** are local maxima of Cmap
- **Junctions** are local maxima of Jmap
- performed by a local marching process

2-D Feature Inference



2-D Feature Inference

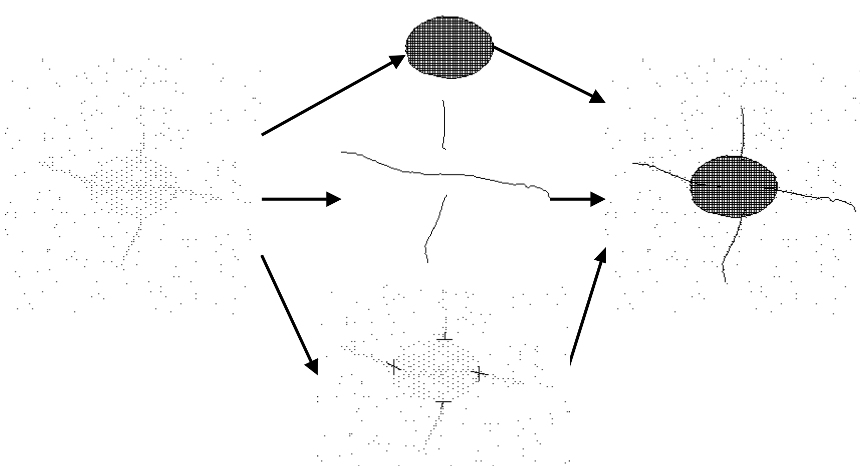




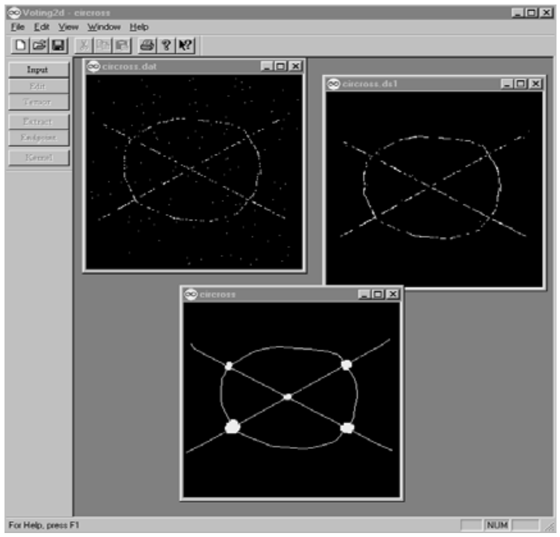
Results

input points　　region boundariness　　curve and junctions

Results



2-D system demo



3-D Tensor Voting

- Representation: 3-D Tensor
- Constraints: 3-D Voting fields
- Data communication: Voting

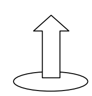
3-D Tensor

Input may consist of

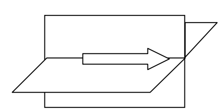
point   curvel   surfel

3-D Tensor = Ellipsoid

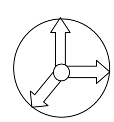
SMOOTH SURFACE



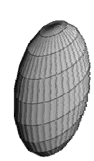
+ CURVE JUNCTION



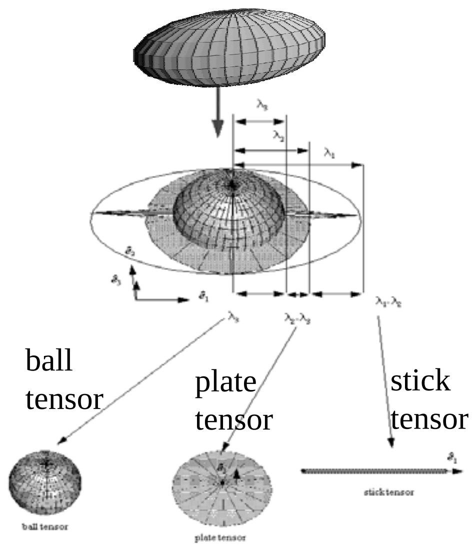
+ POINT JUNCTION



= ELLIPSOID (TENSOR)



Decomposition



3 eigenvalues
$(\lambda_{max}\,\lambda_{mid}\,\lambda_{min})$

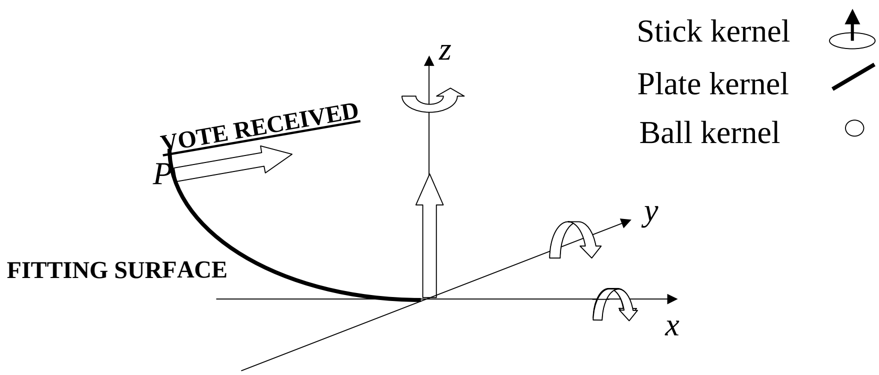
3 eigenvectors
$(\mathbf{V}_{max}\,\mathbf{V}_{mid}\,\mathbf{V}_{min})$

Interpretation

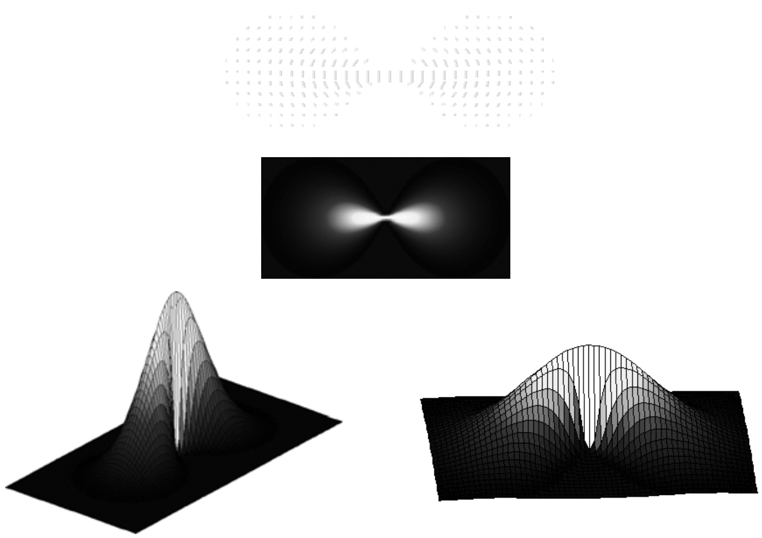
	Saliency	Direction
Surface	$\lambda_{max} - \lambda_{mid}$	$\mathbf{V}_{max}$ **= normal**
Curve	$\lambda_{mid} - \lambda_{min}$	$\mathbf{V}_{min}$ **= tangent**
Junction	λ_{min}	**arbitrary**

3-D Voting Fields

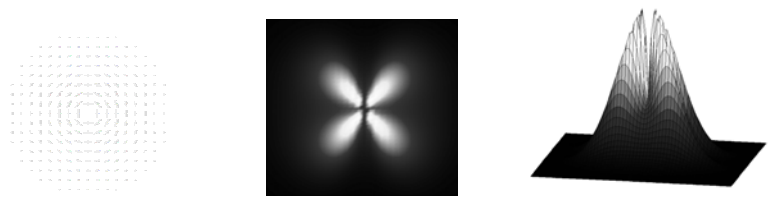
Derived from the Fundamental 2-D Stick Field



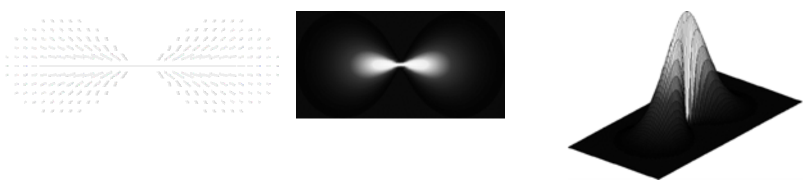
3-D Stick Voting Field



3-D Plate Voting Field



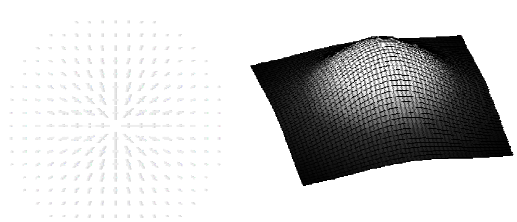
$\lambda_{max} - \lambda_{mid}$

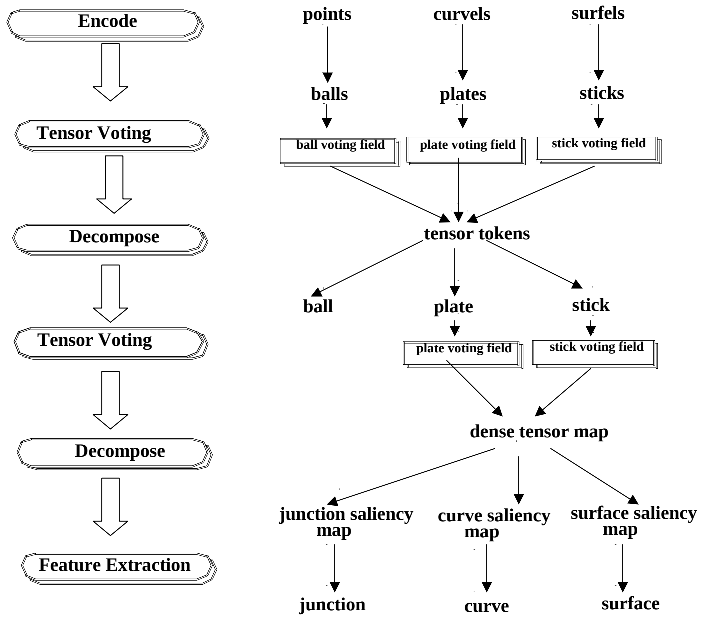


$\lambda_{mid} - \lambda_{min}$

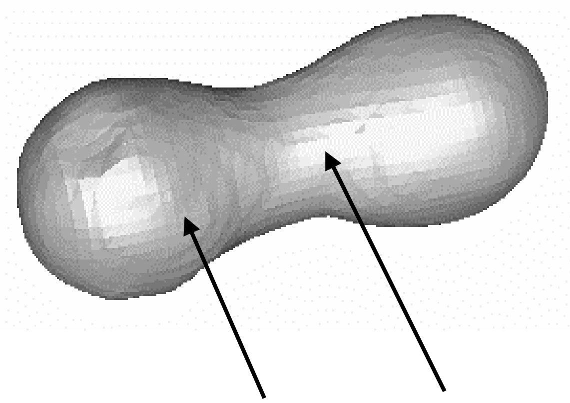
3-D Ball Voting Field

- Isotropic tensor field

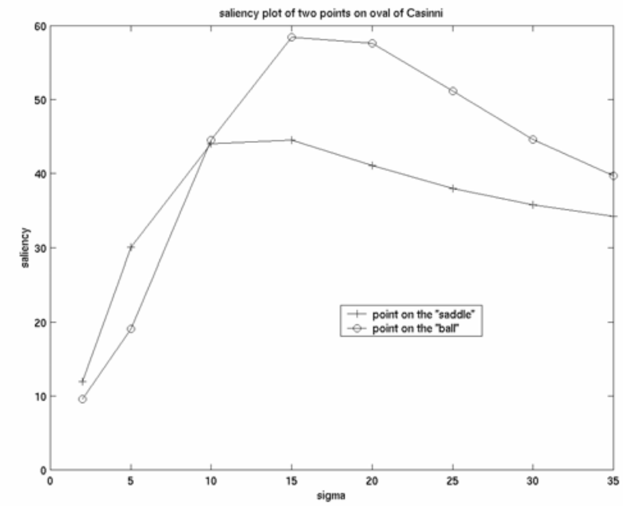




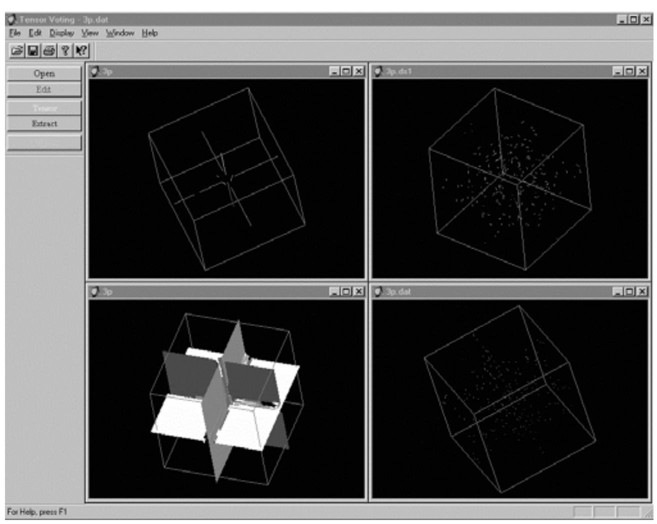
Issue of Scale



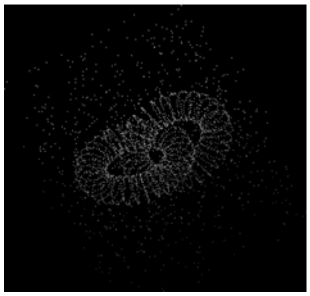
Scale

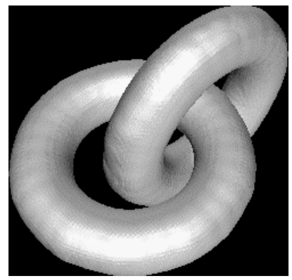


3-D System

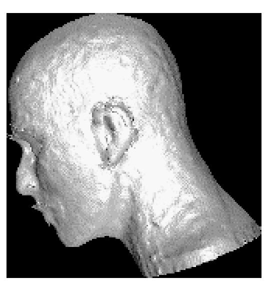


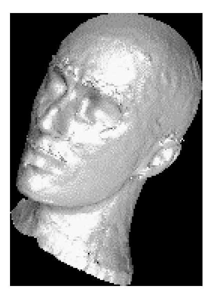
3-D Feature Inference





Results

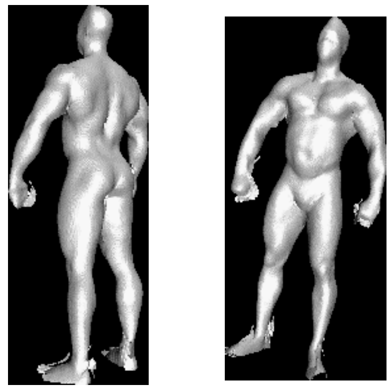




noisy input two views of output surface

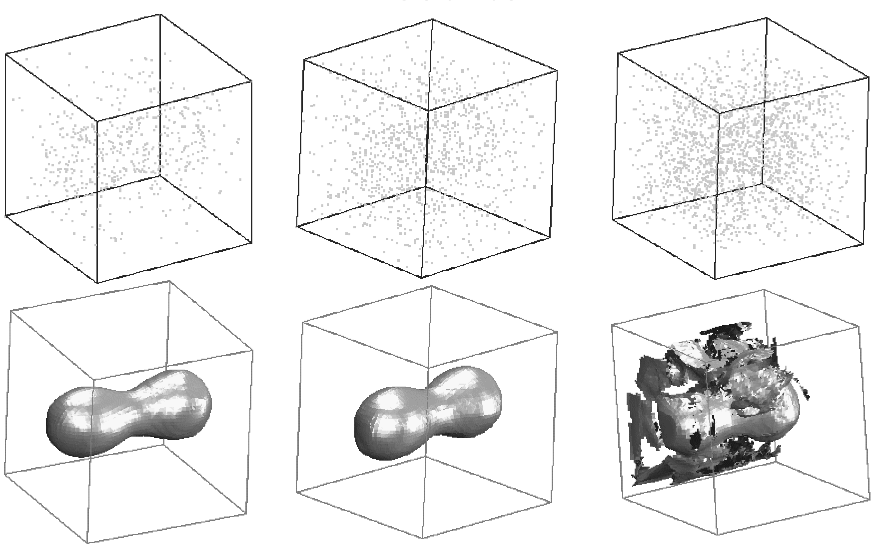
Results

noisy input



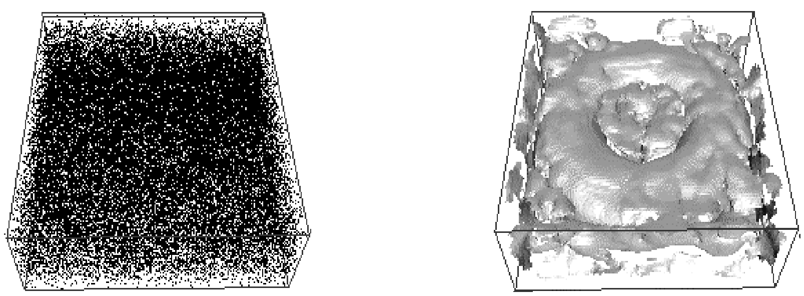
two views of output surface

Results



Results

1200% noise



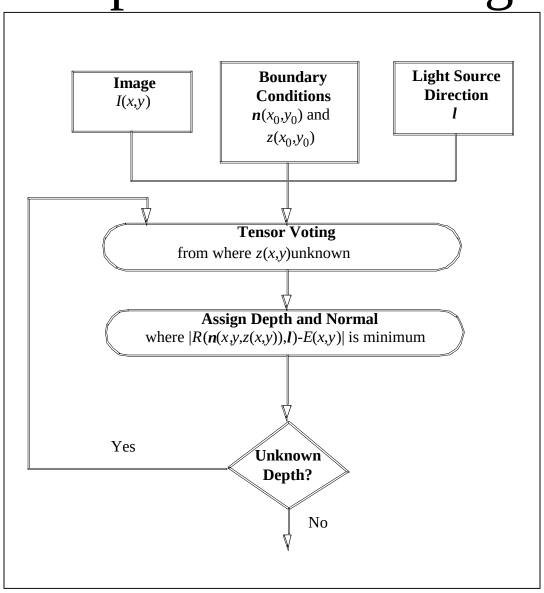
Overview

- Introduction
- Salient feature inference
  - tensor voting
  - 2D and 3D systems
- Applications
  - shape from shading
  - shape from stereo
  - optical flow
- Perspectives

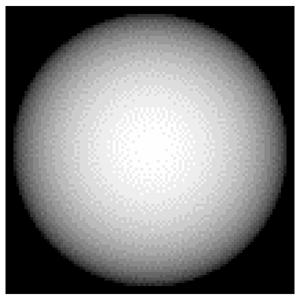
Applications to Low-level vision

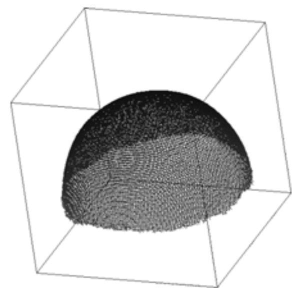
- Shape from Shading
- Shape from stereo
- optical flow

Shape From Shading

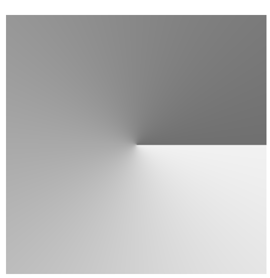


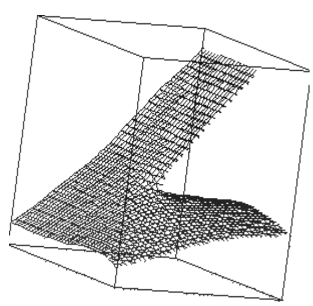
Results





Results

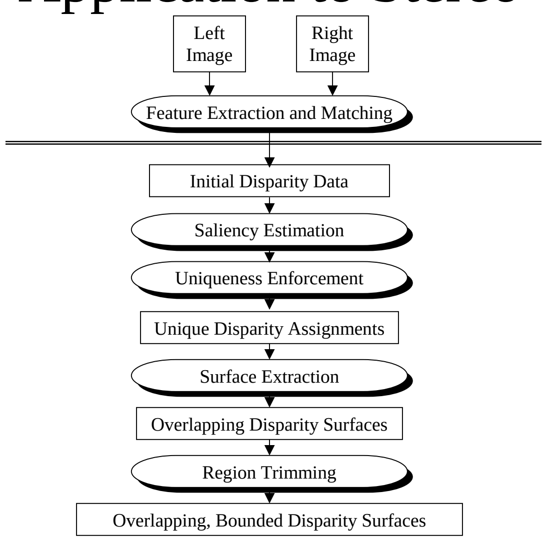




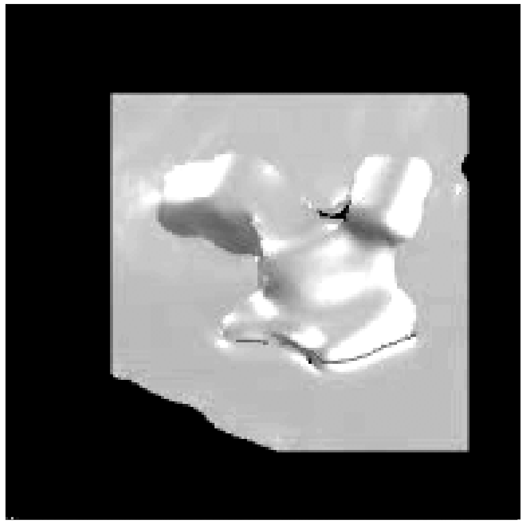
Applications to Low-level vision

- Shape from Shading
- Shape from stereo
- optical flow

Application to Stereo



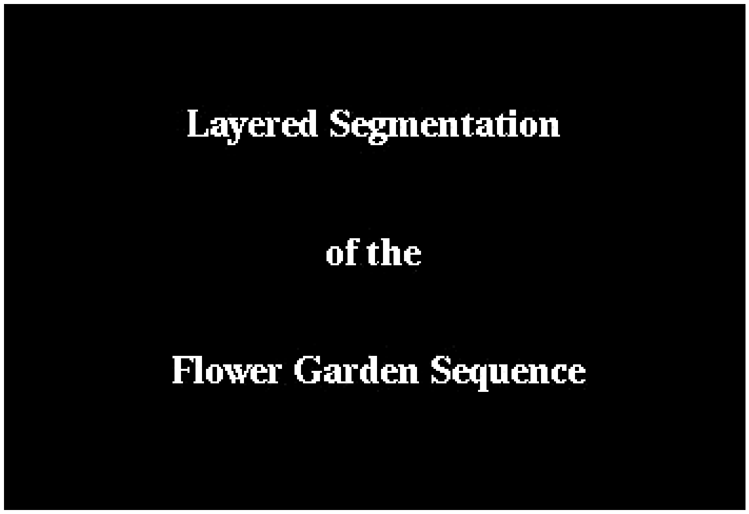
Results



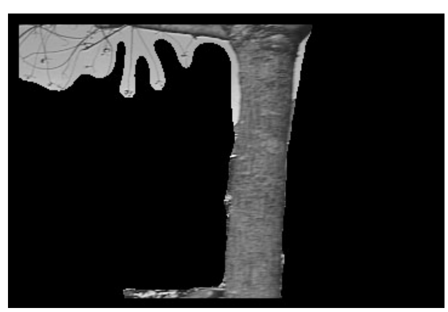
Applications to Low-level vision

- Shape from Shading
- Shape from stereo
- Optical flow

Flower Garden Sequence



Flower Garden Sequence Layers



Conclusions

- Simultaneous determination of motion boundaries and accurate optical flow
- No need for iterative global optimization
- Layered description resulting from segmentation, not an *a priori* model or mathematical fit

Conclusion

- Unified framework
- Applicable to many problems
- Non-iterative optimization
- Promising results
- Issues ...