Machine Vision

Lecture 8
Visual Recognition

Based on lecture of Ehud Rivlin,
Michael Rudzsky

Outline

- Introduction
- Pattern recognition
- Alignment
- Invariants
- Recognition by parts
- Function and context
- Recognition and cognition
- Applications

Motivations for Computer Vision

- "Computational Vision":

  Modeling biological visual processes

  -[Vision as a challenge]

- "Machine Vision":

  Performing specific tasks with the aid of visual data -[Vision as an engineering tool]

- "Image Understanding":

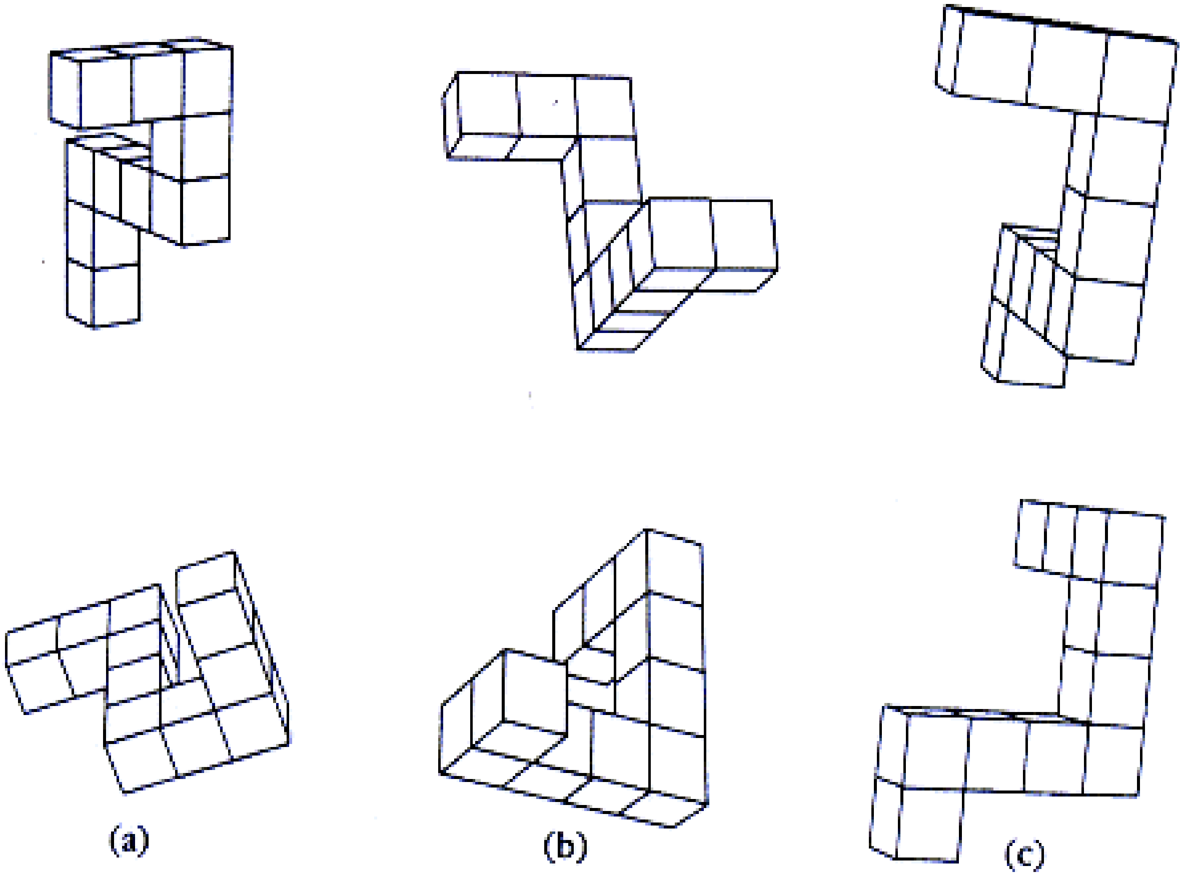
  Recovering the structure of a scene (geometry photometry) from images

  -[Vision as a science]

Vision is

- To see – to know what is where by looking Aristotle
- Vision – an information processing task
- Representation and processing
- To understand computers is different from understanding computations

Mental Rotations Shepard and Mezler 1979

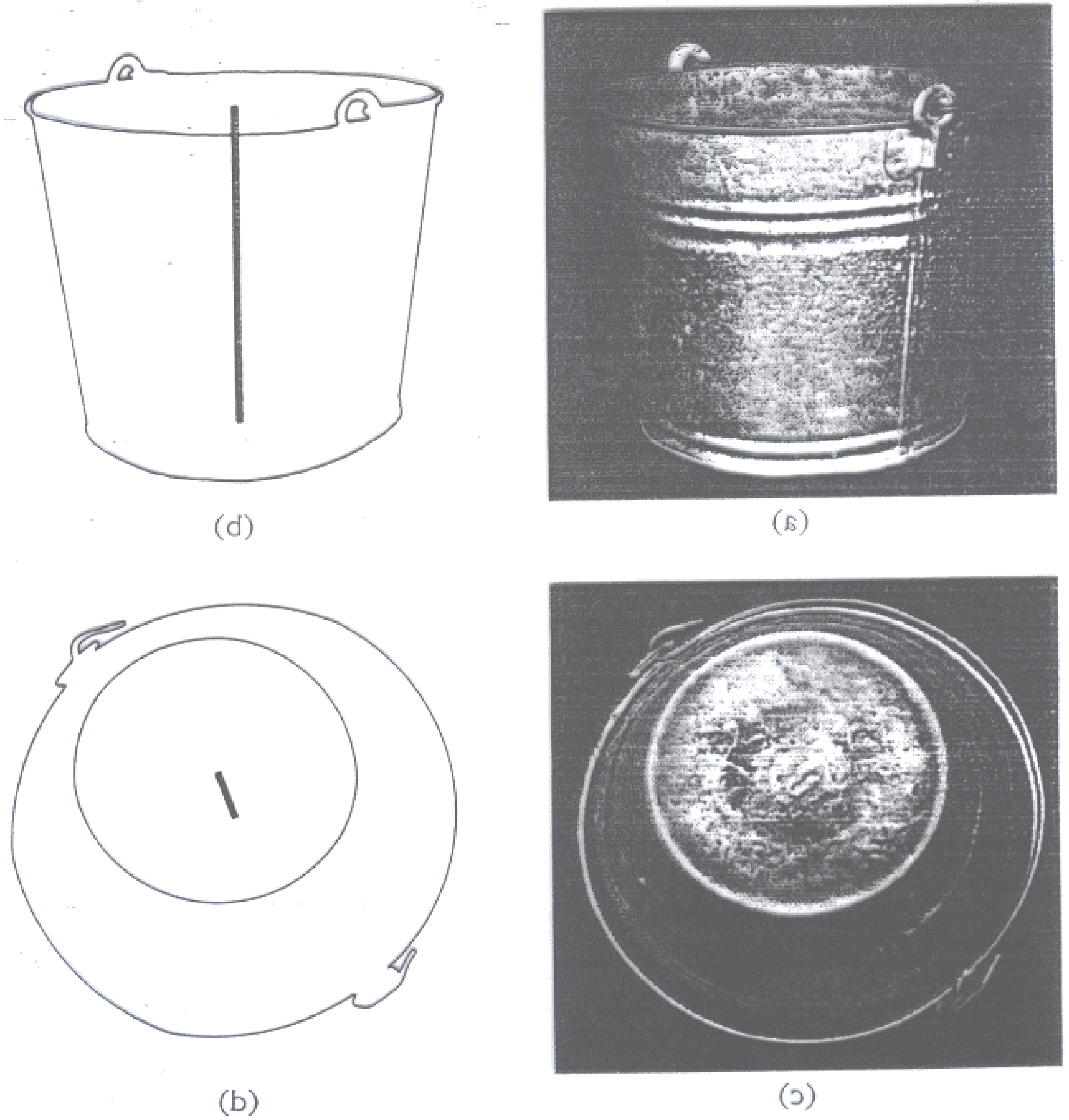


Understanding an IP Device

- Computational theory:
  - What is the goal of the computation, why is it appropriate?
- Representation and algorithm:
  - How can this computational theory be implemented? In particular, what is the representation (input/output) and the algorithm for the transformation? (e.g. Psychophysics)
- Hardware implementation:
  - How can the representation and algorithm be realized physically? (e.g. Neuroanatomy)

What is (human) Vision for?

- What kind of information is vision delivering?
- What are the representational issues involved?
- Warrington (1973): patient who had suffered left or right parietal lesions
  - Right side: provided a conventional view -> name, semantics. O.w. – fail to recognize
  - Left side: unable to name, or state purpose and semantics. Correctly perceived geometry (un/conventional)
- ***Main job of vision – to derive representation of shape***


(b)
(a)
(d)
(c)

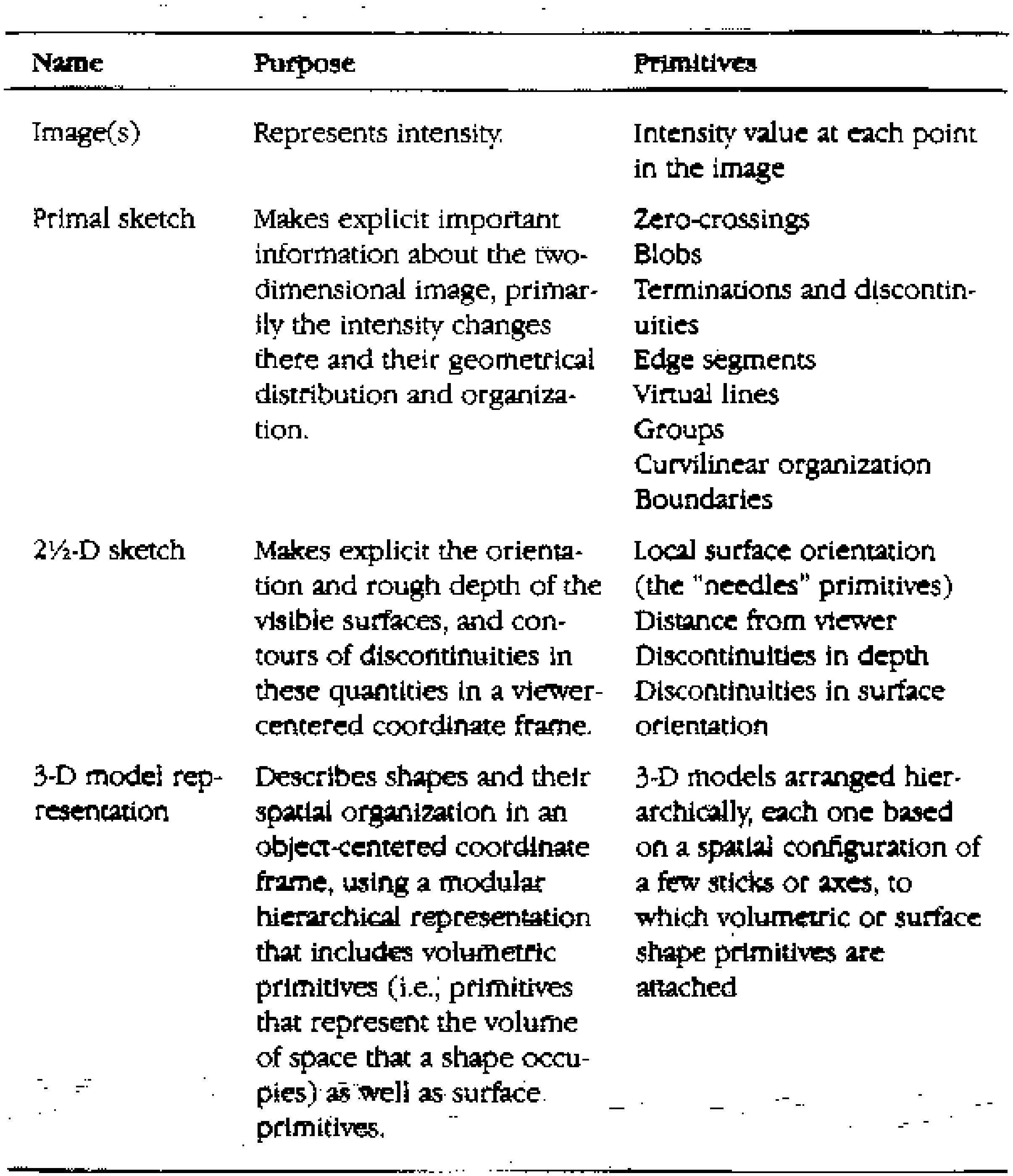
Vision as Recovery

If we can recover we can:

- Navigate and avoid obstacles
- Recognize classes of objects

Use the following methodology:

- Computational Theory
- Algorithms and Data structures
- Implementation



Name	Purpose	Primitives
Image(s)	Represents intensity.	Intensity value at each point in the image
Primal sketch	Makes explicit important information about the two-dimensional image, primarily the intensity changes there and their geometrical distribution and organization.	Zero-crossings Blobs Terminations and discontinuities Edge segments Virtual lines Groups Curvilinear organization Boundaries
2½-D sketch	Makes explicit the orientation and rough depth of the visible surfaces, and contours of discontinuities in these quantities in a viewer-centered coordinate frame.	Local surface orientation (the "needles" primitives) Distance from viewer Discontinuities in depth Discontinuities in surface orientation
3-D model representation	Describes shapes and their spatial organization in an object-centered coordinate frame, using a modular hierarchical representation that includes volumetric primitives (i.e., primitives that represent the volume of space that a shape occupies) as well as surface primitives.	3-D models arranged hierarchically, each one based on a spatial configuration of a few sticks or axes, to which volumetric or surface shape primitives are attached

Vision processes

Are divided into three levels:

- Low level processes (e.g., edge detection, stereo, motion analysis, shape from shading):

  recovering surfaces from the image, extracting features.

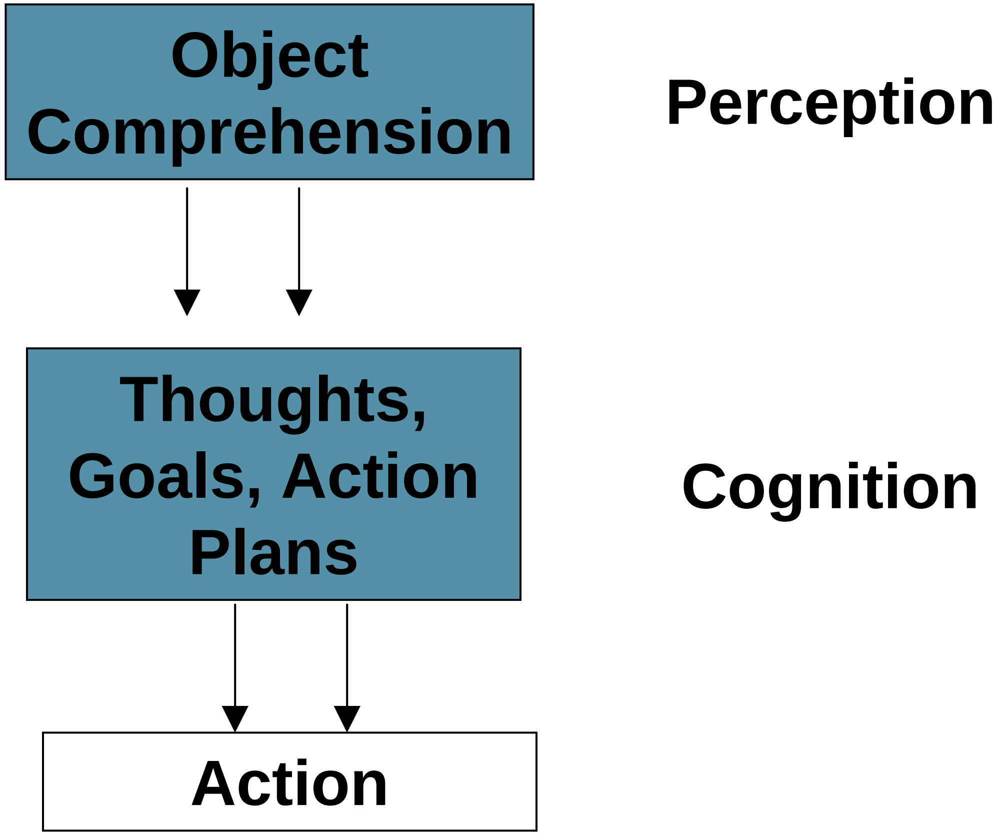
- Mid level processes (e.g., attention, segmentation, grouping, object-background separation) :

  focusing on the objects.

- High level processes (e.g., recognition, localization, navigation) :

  building a map of the environment

Perception and cognition



Object Comprehension

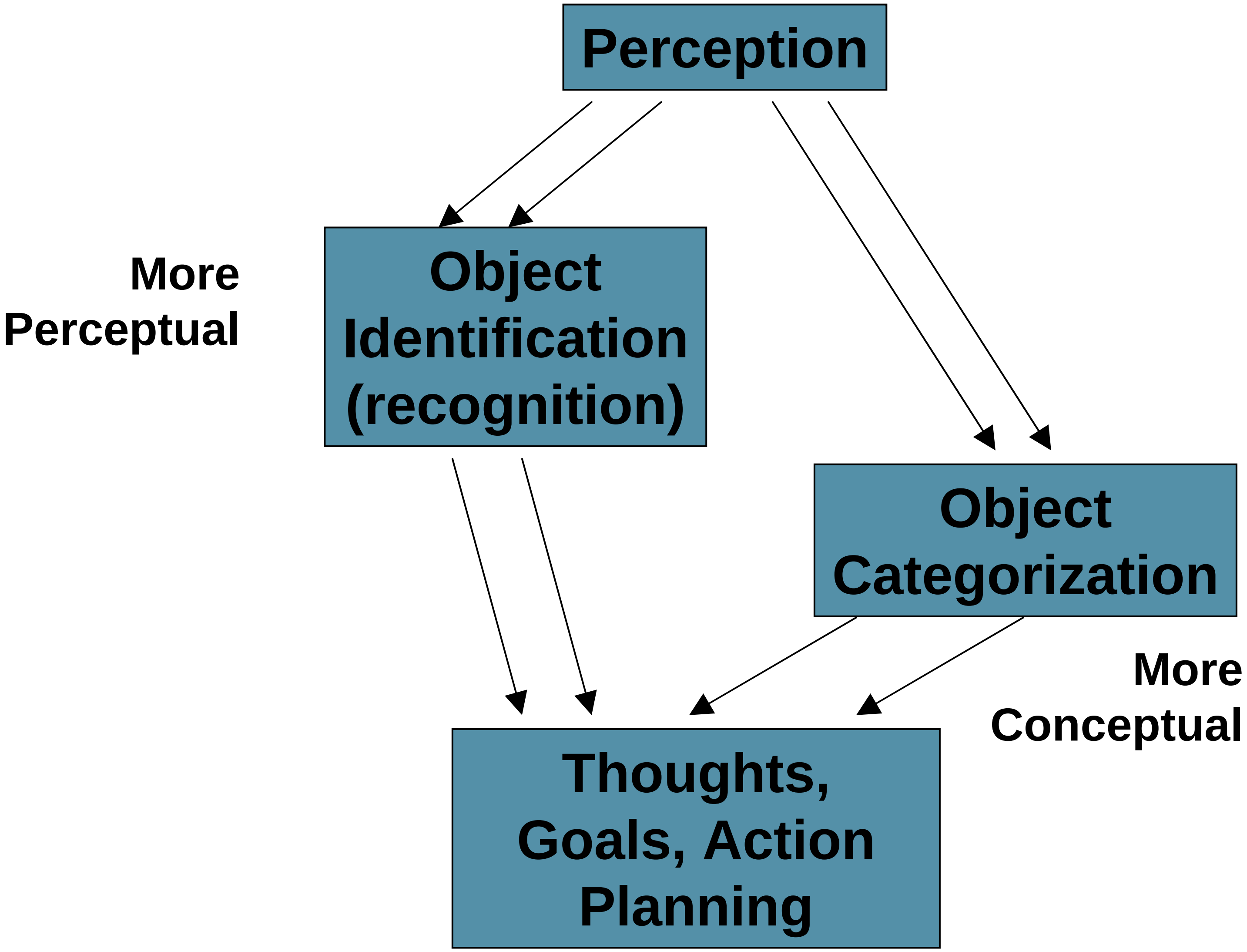
Perception

Thoughts, Goals, Action Plans

Cognition

Action

Object identification and categorization



Perception

More Perceptual

Object Identification (recognition)

Object Categorization

More Conceptual

Thoughts, Goals, Action Planning

Marr Paradigm

To understand a visual process consider the following levels:

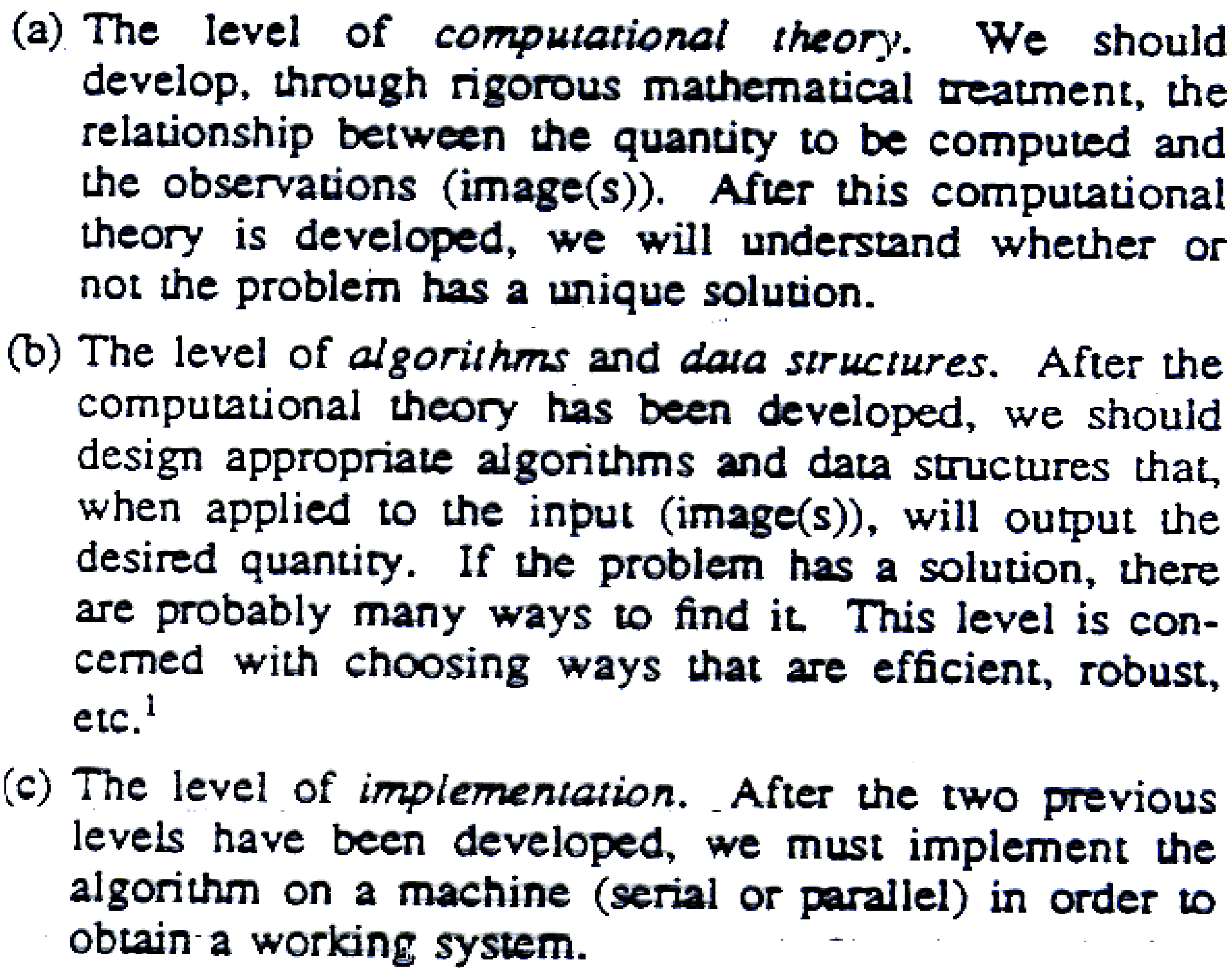
(a) The level of computational theory

(b) The level of algorithms and data structures

(c) The level of implementation

Marr Paradigm

To understand a visual process consider the follwoing levels:



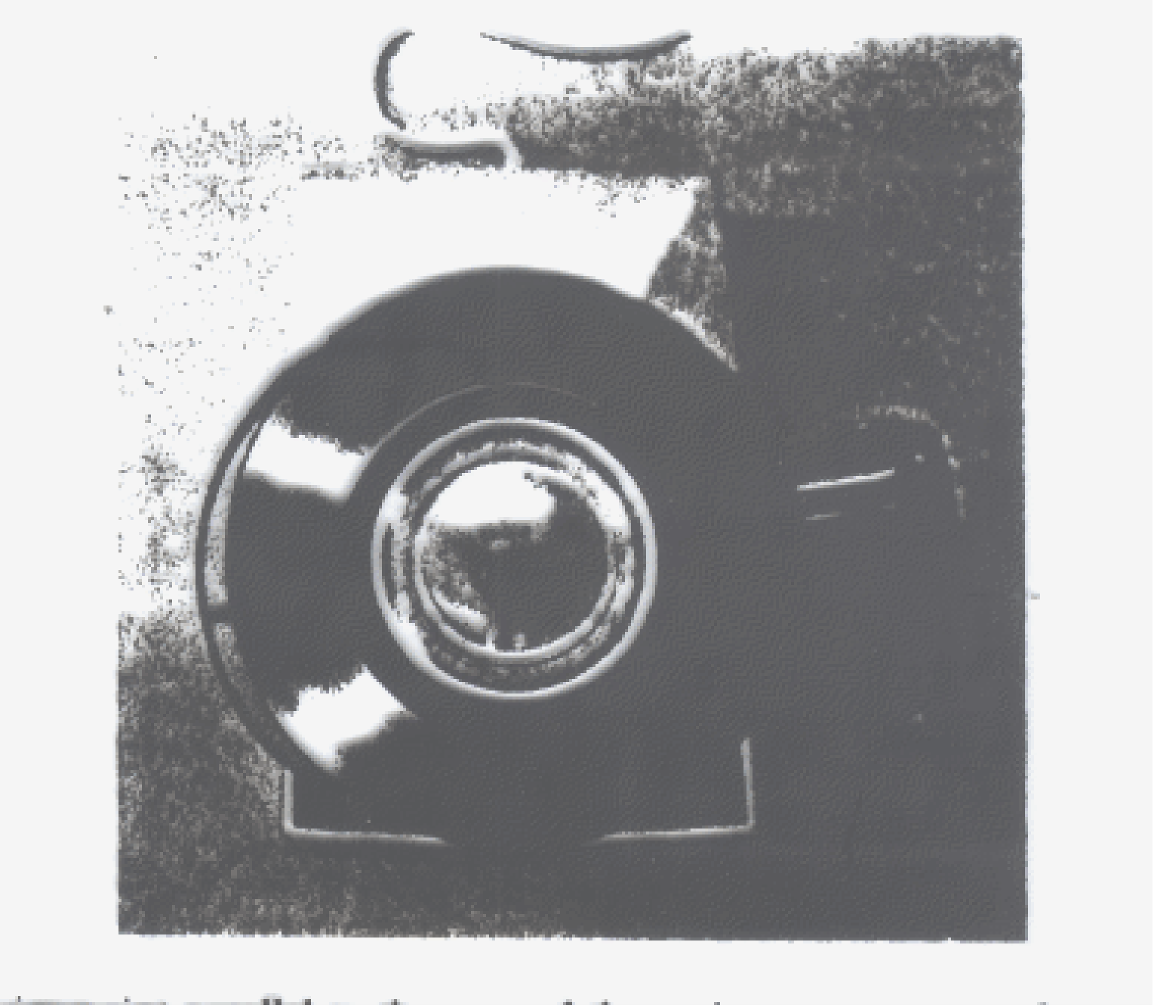
(a) The level of *computational theory.* We should develop, through rigorous mathematical treatment, the relationship between the quantity to be computed and the observations (image(s)). After this computational theory is developed, we will understand whether or not the problem has a unique solution.

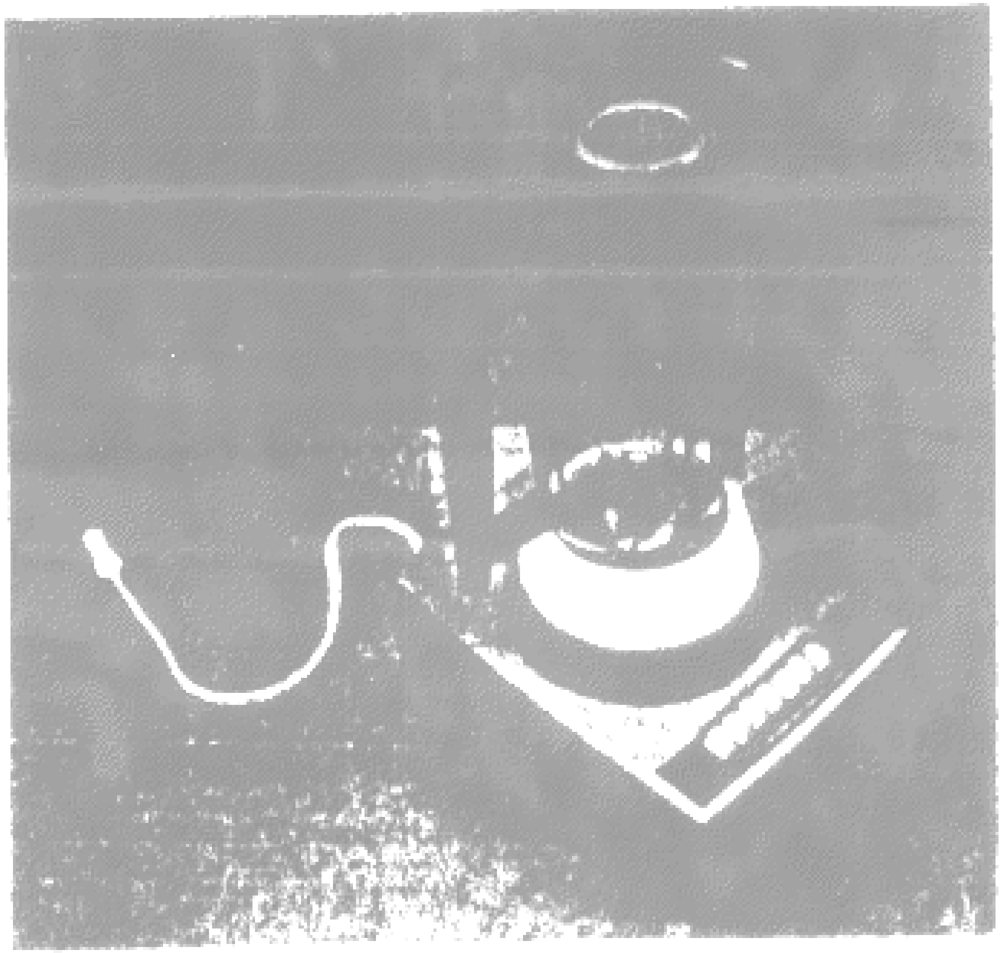
(b) The level of *algorithms* and *data structures*. After the computational theory has been developed, we should design appropriate algorithms and data structures that, when applied to the input (image(s)), will output the desired quantity. If the problem has a solution, there are probably many ways to find it. This level is concerned with choosing ways that are efficient, robust, etc.[1]

(c) The level of *implementation*. After the two previous levels have been developed, we must implement the algorithm on a machine (serial or parallel) in order to obtain a working system.

Conjectures

- For purposes of rapid recognition humans represent a 3-D object by a set of characteristic views or “aspects”, by a set of commonly occurring 2-D projections
- To a first approximation, humans describe the image of an object as consisting of a set of “primitive” parts. There are two types of such parts: pieces of regions and pieces of boundaries
- The properties used by humans to describe parts for purposes of rapid recognition are local property values, or simple combinations of such values…….





Vision is Hard

- Some pictures are difficult to interpret
- A large size of the brain is devoted to vision
- 45 years of research in computer vision and we are still nowhere near a solution

Why Vision is Hard?

- Ill-definedness
- Ill-posedness
- Intractability

Ill - definedness

- The standard scene models used in general recovery tasks are not fully defined:
  - “Piecewise” simple means nothing unless we impose a lower bound on the piece sizes.
  - “Noise” is not easy to model (or to distinguish from the “signal”); it’s not Gaussian!

- Many easily recognizable object classes do not have simple definitions (A’s, chairs, bushes, dogs, …)

Ill - posedness

Recovery problems are usually underconstrained -
e.g., ambiguity of illumination/photometry/ geometry

Questionable approaches:

Add constraints - e.g.,

- smoothness (regularization)
- description length

Critique: the actual scene may not satisfy the constraints!

Regularization

The equation $L(\bar{\omega}) = 0$ is not enough to recover

Assume $S(\bar{\omega}) = 0$

- For surface shape – surface is smooth
- In color processing – reflectance is described with small number of basis functions (retinal pigments)
- For non-rigid motion – small deviation from rigid motion
- In segmentation – as simple as possible with respect to some complexity measure

- Minimize: $\iint\limits_{\Omega} \left(L^2(\bar{\omega}) + \lambda S^2(\bar{\omega}) \right) dxdy$

Issues

- The coefficient determines the relative importance of smoothing
- Tends to smooth over discontinuities
- How to pick the coefficient systematically?
- Small values increase sensitivity to noise

Solutions:

- Segment into homogeneous regions and regularize within each region

- Divide the image into boundary and nonboundary points

- etc…. (e.g. Active Vision)

Intractability

- Recovery and recognition tasks are of combinatorial complexity.

- Parallelism can speed up the early stages of the vision process (e.g., image operations), but little is known about how to speed up the potentially combinatorial stages.

- We are usually forced to solve vision problems in real time with inadequate computational resources; hence we are always cutting corners (modularity, suboptimality).

What can be done

- **Define your domain**
  Work in a domain that can be adequate (e.g., specialized)

- **Pick your problem**
  Attempt only partial recovery ("qualitative") for recognition
  Purposive (functional) vision

- **Improve your inputs**
  - Sensory redundancy (Multisensor fusion, Active vision
  - Processing redundancy (consensus)

- **Take your time**
  Use adequate computational resources

Active vision

- An active vision system is one that is able to interact with its environment by altering its viewpoint rather than passively observing it, and by operating on sequences of images rather than on a single frame.
- Moreover, since its foveas can scan over the scene, the range of the visual scene is not restricted to that of the static view.
- The ability to physically track a target reduces motion blur, increasing target resolution for higher level tasks such as classification.
- Active Vision is close, in principle, to the biological systems that inspired it and so it seems intuitively acceptable that as a visual sensor (especially augmented with color) it is perfectly suited to human/robot interaction and autonomous robot navigation in human environments.

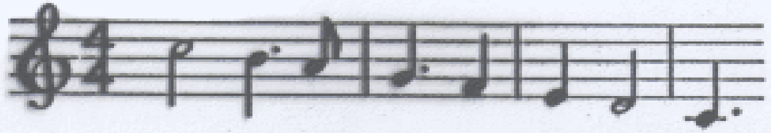
3

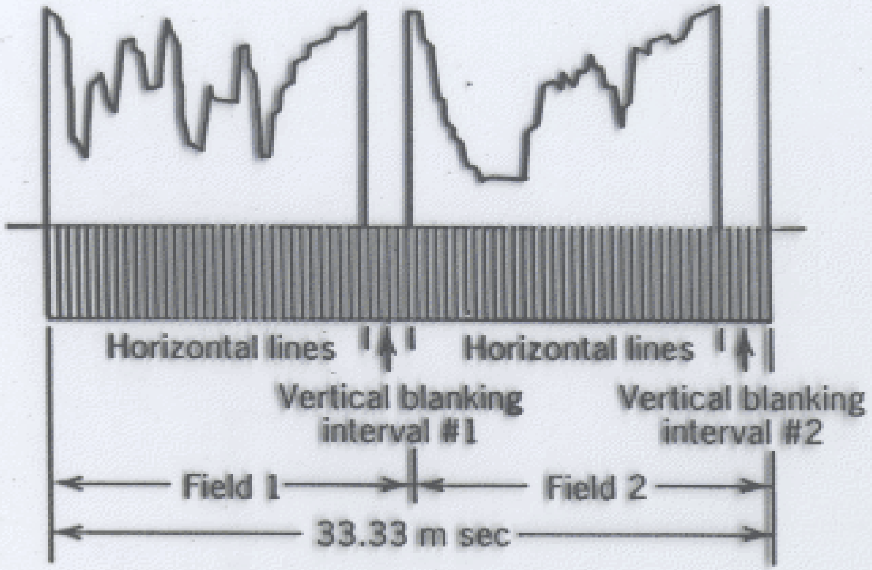
Recognition

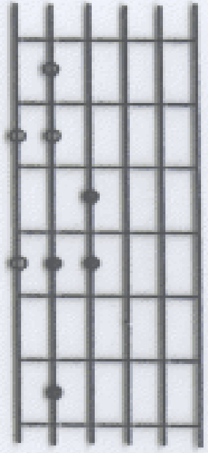
Associating consistent labels to objects ignoring variations due to illumination conditions, viewing position, posture (non-rigidities), occlusion, background, etc.

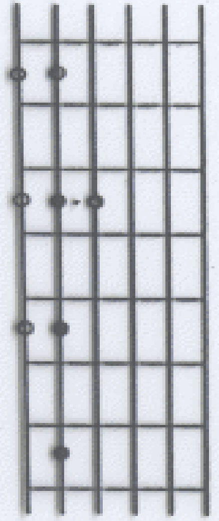
Main approaches to recognition:

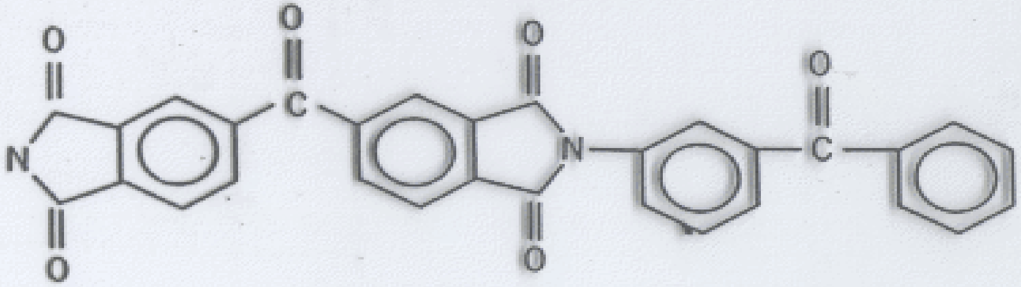
- Pattern recognition
- Alignment
- Invariants
- Part decomposition
- Functional description











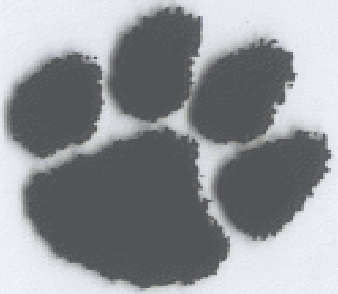
ISBN 0-471-50536-6

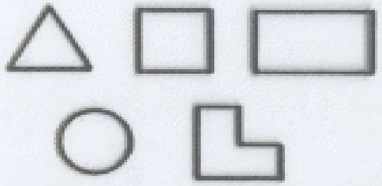
This is a pattern.

This is too.

654–3731

656–5921





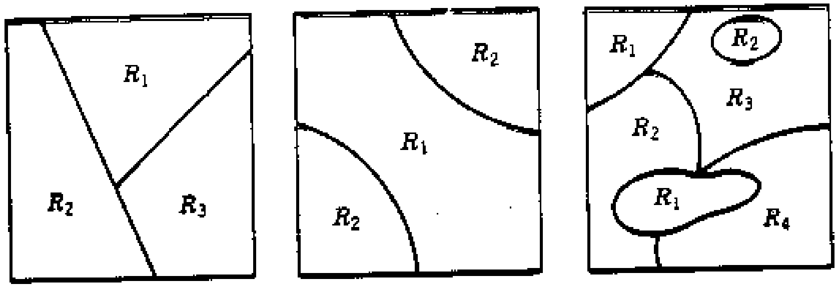
XXXOOXXOOOXXXOO

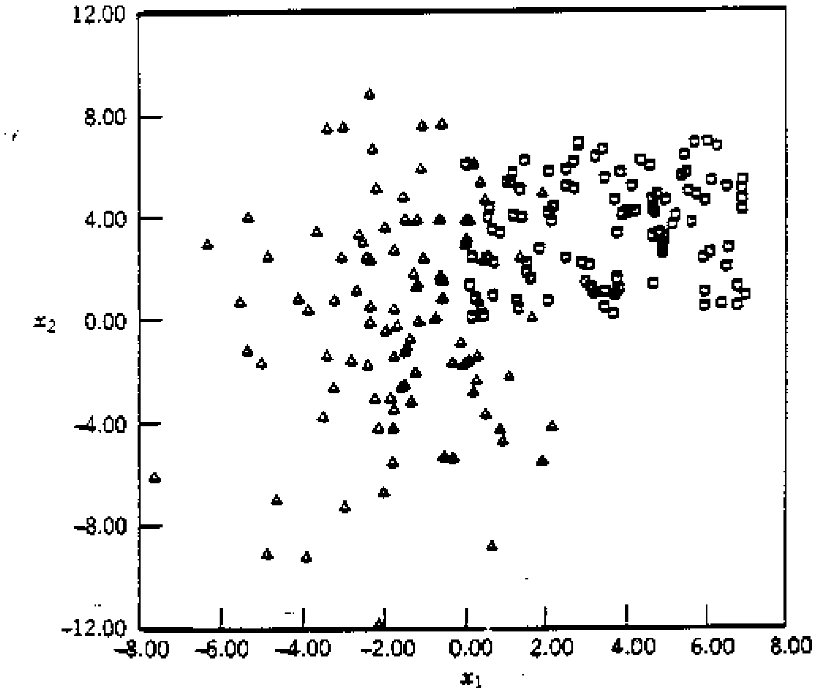
Pattern Recognition

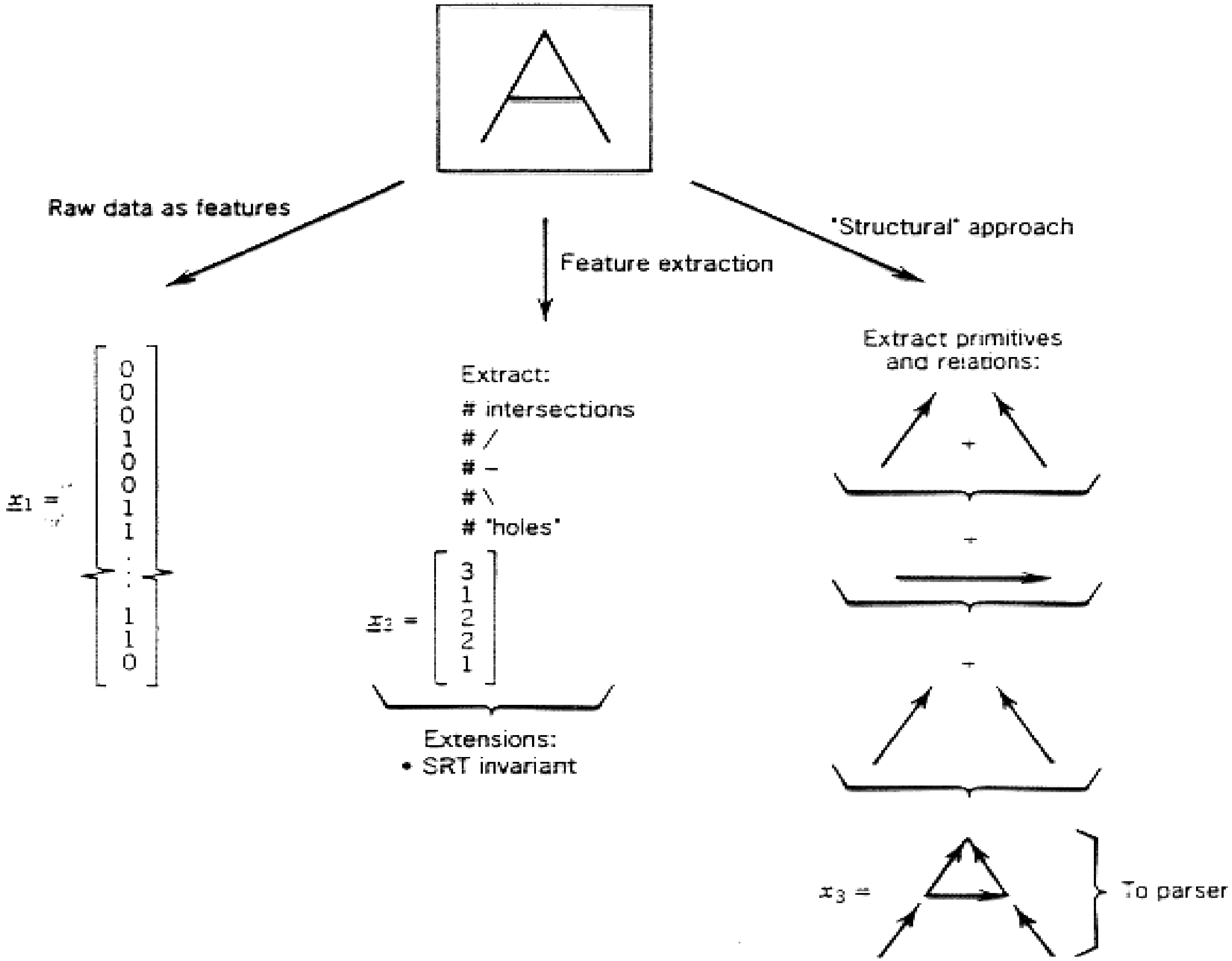
- Recognizing objects by their distinguished features

Method:

- A set of characteristic features is extracted from the image.
- Characteristic features of the same object are clustered together using statistical modeling.
- Recognition is achieved by finding the cluster to which the image features belong


R_1
R_2
R_3
R_2
R_1
R_2
R_1
R_2
R_3
R_2
R_1
R_4


12.00
8.00
4.00
0.00
-4.00
-8.00
-12.00
x_2
-8.00
-6.00
-4.00
-2.00
0.00
2.00
4.00
6.00
8.00
x_1



Raw data as features

Feature extraction

"Structural" approach

$$\underline{x}_1 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 1 \\ 1 \\ \vdots \\ 1 \\ 1 \\ 0 \end{bmatrix}$$

Extract:

# intersections

# /

# –

# \

# "holes"

$$\underbrace{\underline{x}_2 = \begin{bmatrix} 3 \\ 1 \\ 2 \\ 2 \\ 1 \end{bmatrix}}$$

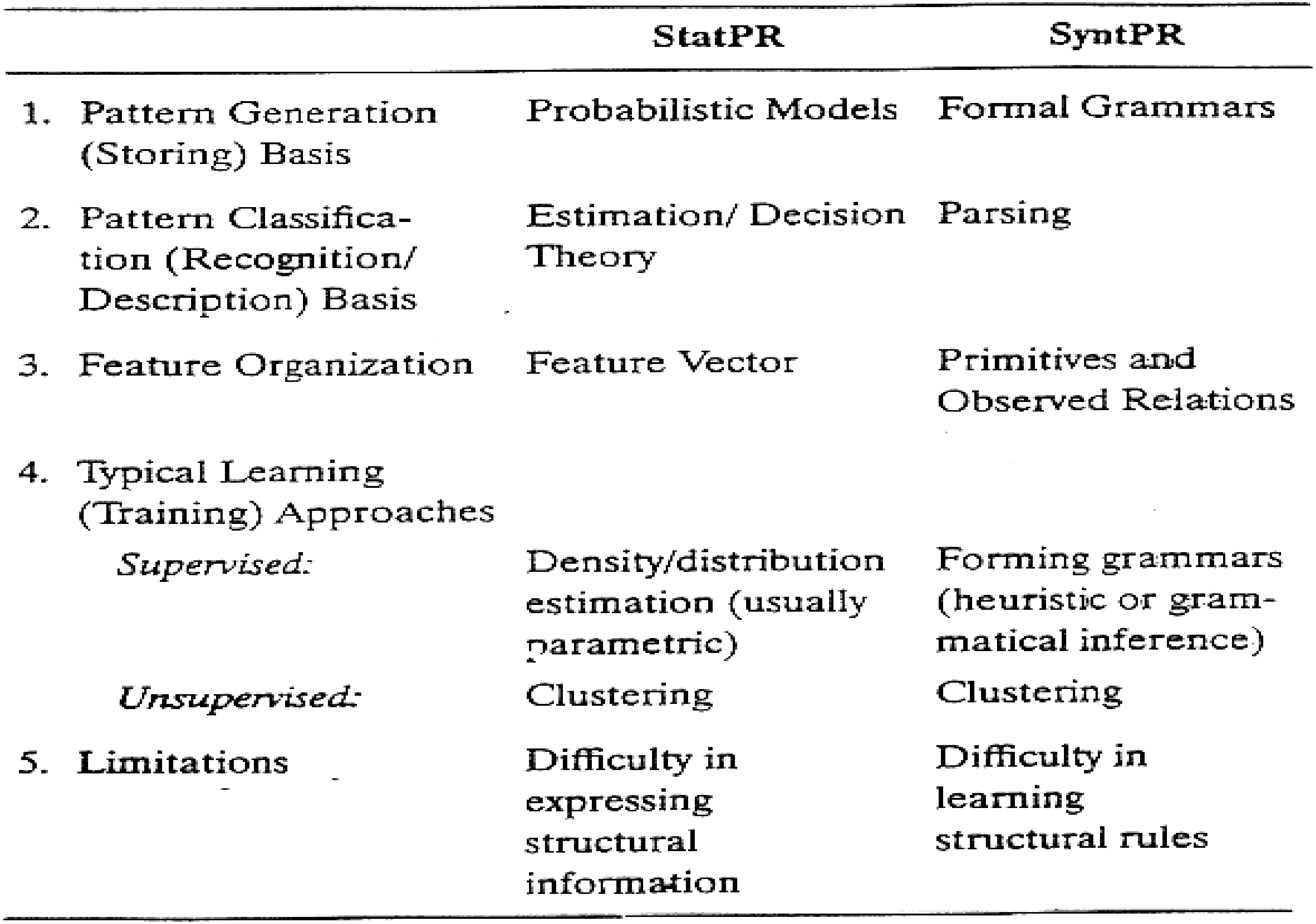
Extensions:

- SRT invariant

Extract primitives and relations:

$x_3 =$

To parser



	StatPR	SyntPR
1. Pattern Generation (Storing) Basis	Probabilistic Models	Formal Grammars
2. Pattern Classification (Recognition/ Description) Basis	Estimation/ Decision Theory	Parsing
3. Feature Organization	Feature Vector	Primitives and Observed Relations
4. Typical Learning (Training) Approaches		
Supervised:	Density/distribution estimation (usually parametric)	Forming grammars (heuristic or grammatical inference)
Unsupervised:	Clustering	Clustering
5. Limitations	Difficulty in expressing structural information	Difficulty in learning structural rules

Pattern recognition (cont.)

Domain:

- Suitable mainly for 2D objects.

Problems:

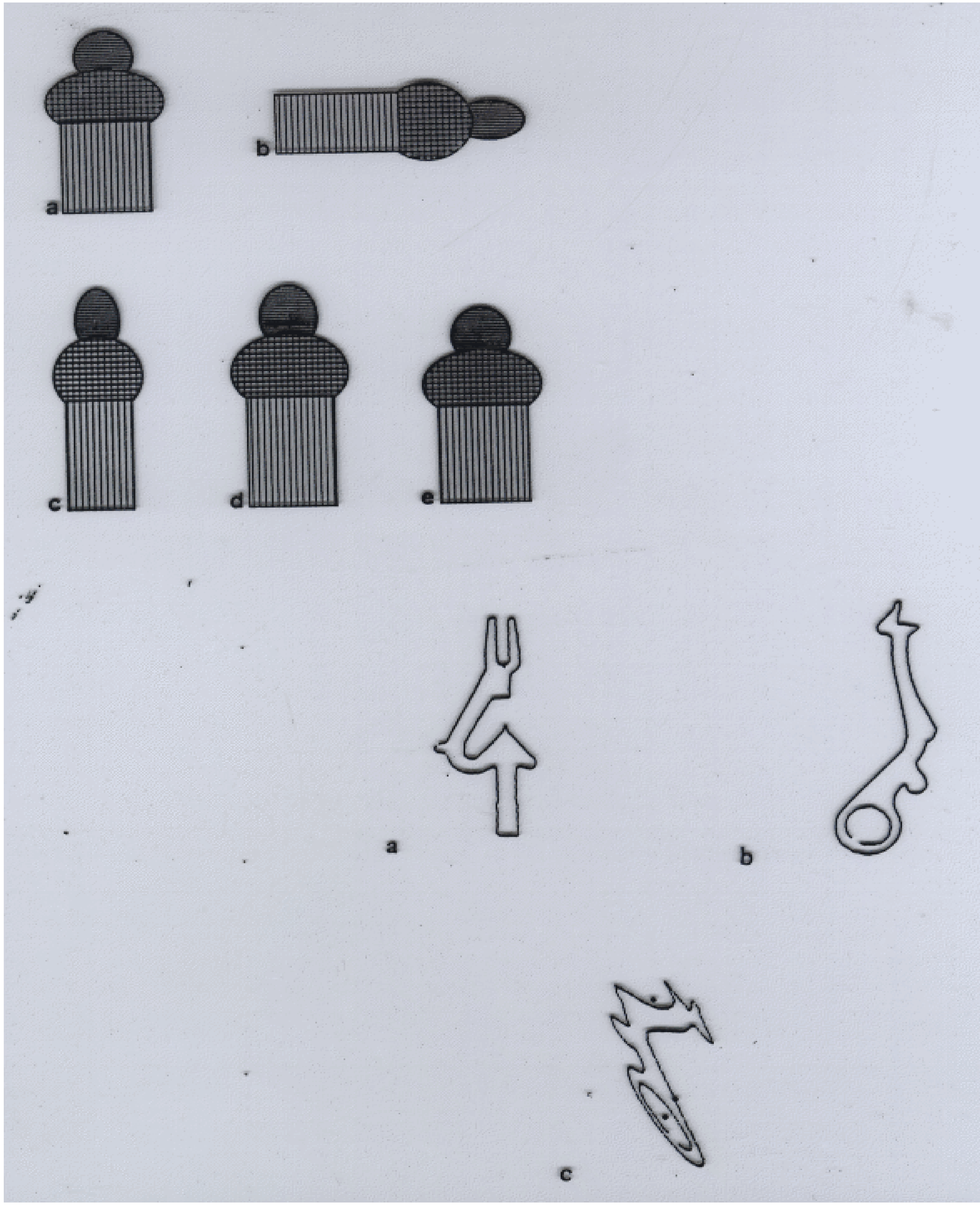
- Limited domain.
- Can statistical approaches effectively account for variations between images?

Alignment

- Recognizing objects by compensating for variations

Method:

- The stored library of objects contains their shape and allowed transformations.
- Given an image and an object model, a transformation is sought that brings the object to appear identical to the image.


a
b
c
d
e
a
b
c

Alignment (cont.)

Domain:

- Suitable mainly for recognition of specific object.

Problems:

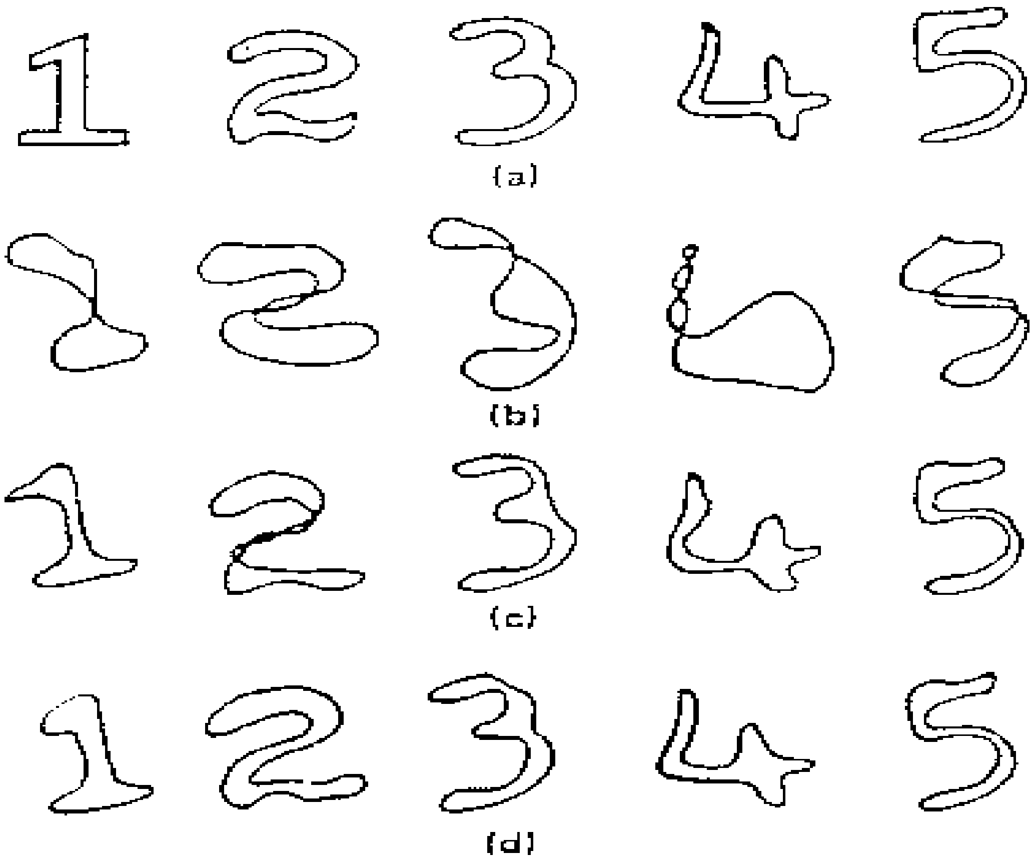
- Complexity: recovering the transformation is time-consuming.
- Indexing: library is searched serially.
- Non rigidities are difficult to model.

Invariance

- Removing the effect of transformation from the image.

Method:

- Compute a set of measurement that are independent of image variations



Fourier approximations. (a) Original, (b) Five harmonics, (c) Ten harmonics, (d) Fifteen harmonics (From E. L. Brill, 1969; reproduced by permission).

Invariance (cont.)

Domain:

- Suitable mainly for planar objects.

Problems:

- Limited domain: problematic when 3D objects are considered.
- Non rigidities are difficult to model.

Part decomposition

Recognizing objects by their structural descriptions.

Method:

- Given an image, identify basic parts in the image.
- Generate a graph in which the nodes represent the parts and the edges
- Represent the spatial relations between the parts.
- Compute the obtained graph with stored graph representations.

Part decomposition (cont.)

Domain:

- Suitable mainly for categorization.

Problems:

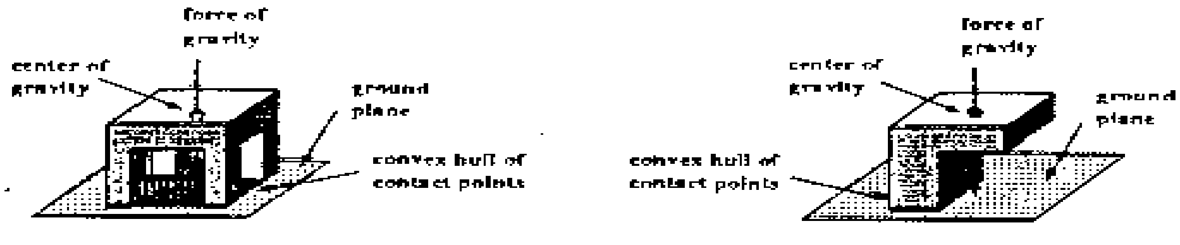
- Extracting parts from the image is often difficult and unreliable.
- Many objects cannot be distinguished by their part structure only.
- Metric information is essential in these cases.

Functional description

Recognizing objects by their function.

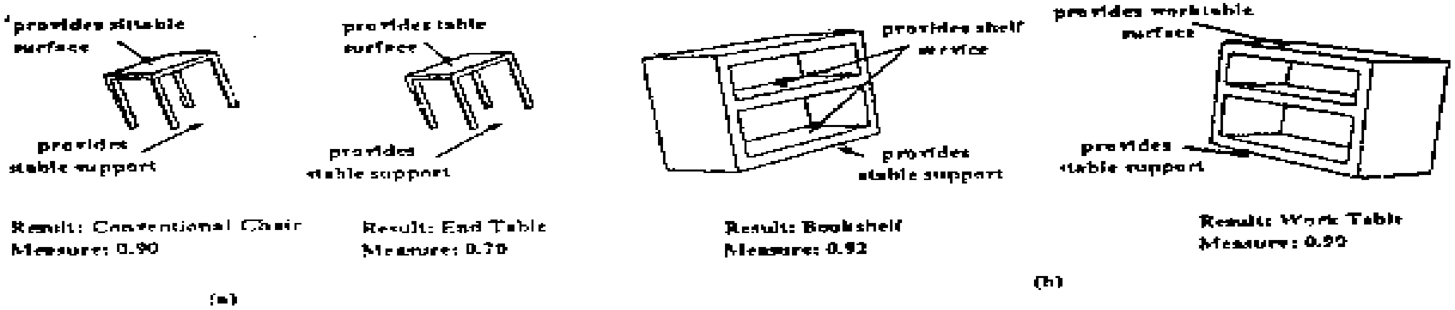
Method:

- Identify parts of the object that have certain functional use.
- Compare the set of functions extracted from the image with similar sets of stored objects (or with a needed one).



(a) stable situation

(b) unstable situation



Result: Conventional Chair
Measure: 0.90

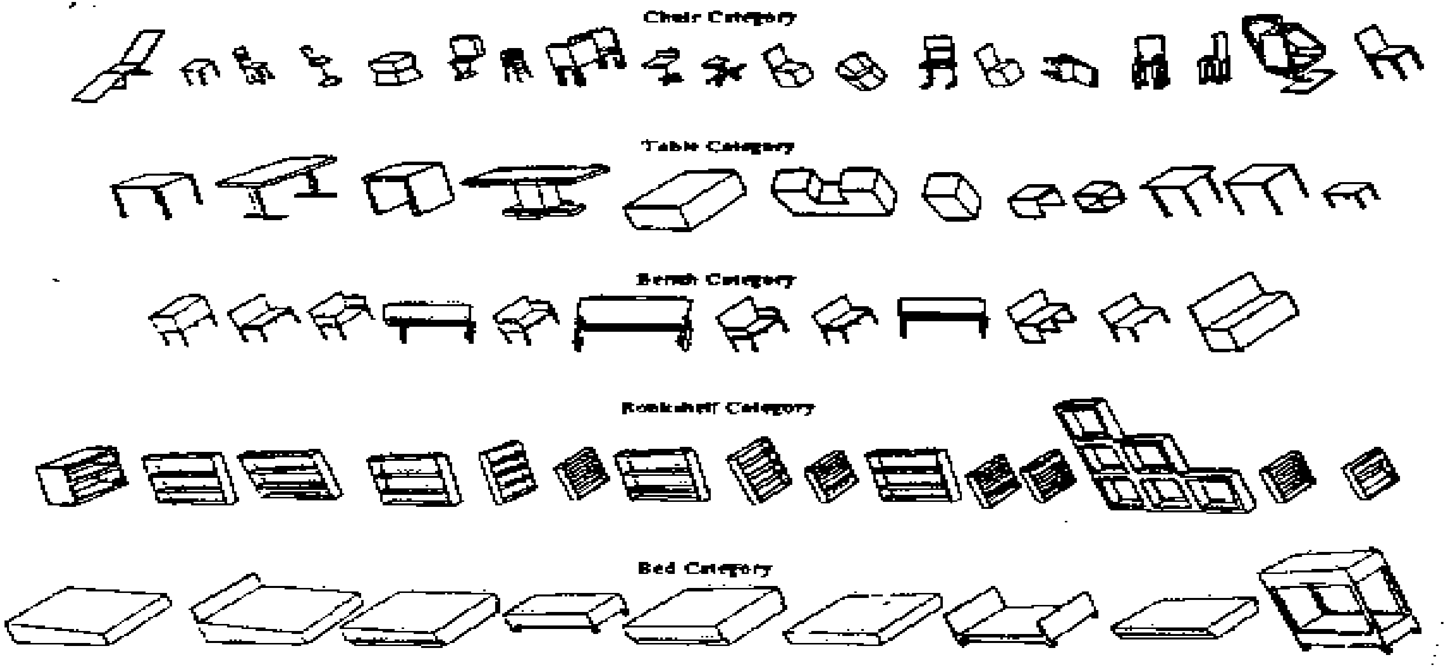
Result: End Table
Measure: 0.70

Result: Bookshelf
Measure: 0.92

Result: Work Table
Measure: 0.99

(a)

(b)



Functional description (cont.)

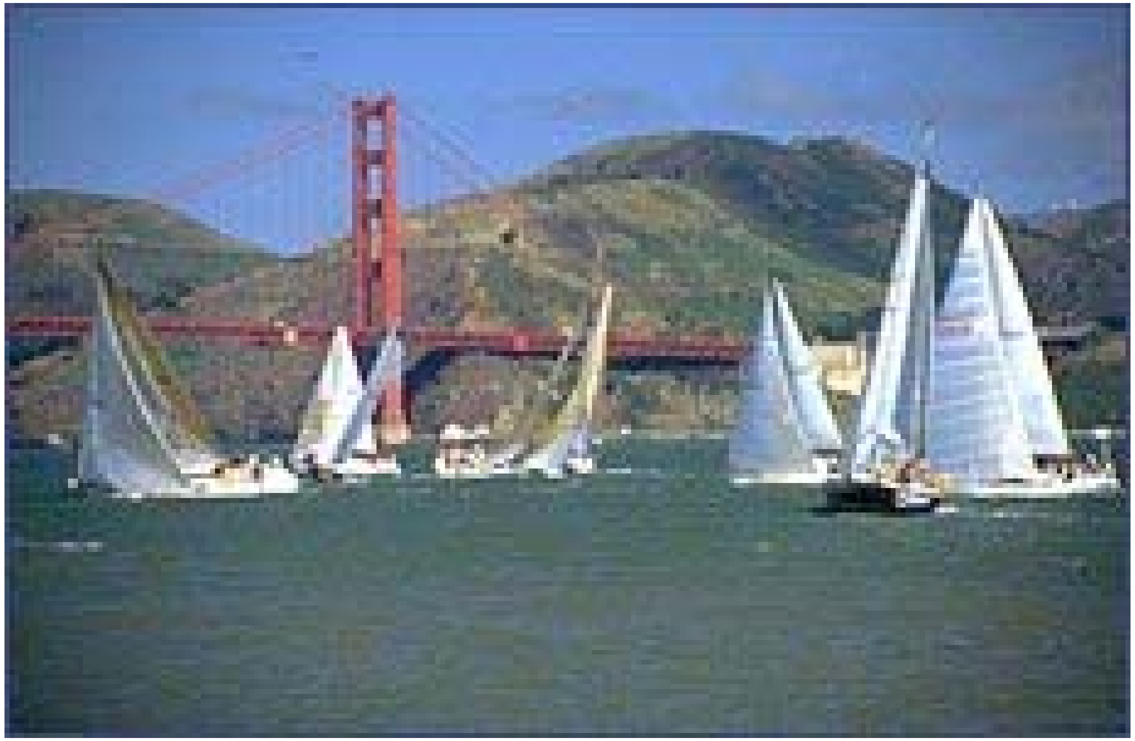
Domain:

- Easy to apply to man-made objects (design).

Problems:

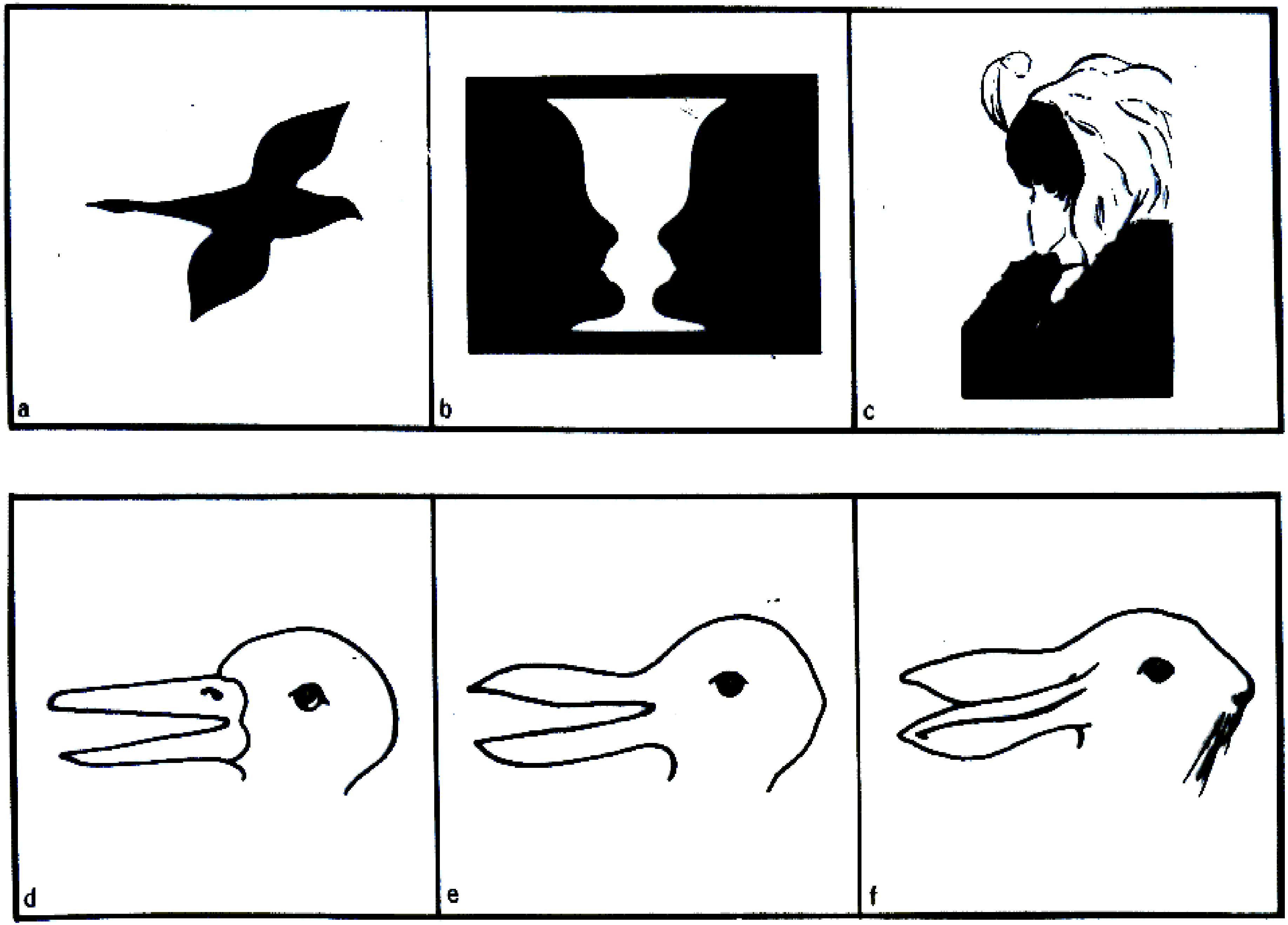
- Function is often difficult to extract
- Mapping Shape to Function

Recognize !



Recognition is harder

- What constitutes an object?
- Categorization vs. identification
- Recognition of entire scenes
- Context.


a
b
c
d
e
f

Context

The Question

- How does our visual system enable us to gain reliable information about the environment?
- What kind of information can or should a visual system derive from the image?
- Are descriptions of 3-D spatial structure and location the only descriptions that should be produced?

What is vision for?

- Bottom-up vs. Top-down
- Complete and accurate recovery of the scene is not necessary Full recovery is not always

Purposive Recognition

- Perception for action: formulating questions that are directly related to visual tasks
- Recognition is defined with respect to a given task
- Complete and accurate recovery of the scene is not necessary

Visual Tasks \\ Assumptions	General	Specific
General	Object recognition from 2D views	Structure from motion with rigid motion, or for planar environment
Specific	Deriving time-to-collision repositioning from 2D views	Recognizing specific objects in a restricted industrial environment

Applications

- Inspection
- Guiding tools for the blind.
- Autonomous machines
- Security systems
- Optical character recognition
- Image data bases
- Image stabilizers, video editing