

Part I: Representations and Learning in Computer Vision

Aleš Leonardis

ViCoS
Visual Cognitive Systems Laboratory
University of Ljubljana



VisionTrain Thematic School
Les Houches, France, 30 March, 2007



Outline

Part I:

- Motivation: Computer vision for cognitive assistants
- Evolution of object representations (models)
 - Generic (category-based) versus exemplar-based
 - Object-centered versus viewer-centered
 - Shape-based, appearance-based
 - Global features, local features



Computer vision for cognitive assistants



MORPHA-Video (www.morpha.de)



Tasks

- Objects
 - Object recognition
 - Object categorization
 - Object segmentation (notion of an object)
 - Pose estimation (understanding the layout)
 - Object manipulation (affordances)
- Actions
 - Action recognition/categorization/segmentation
 - Interaction
- Places
 - Recognition/categorization of places
 - Understanding the spatial relations
 - Affordances (interaction with the environment)



Representation-learning-recognition

- Representation-learning-recognition (three inseparable parts of visual perception)
- Visual recognition seems to be an easy task for humans.
 - How does human brain learn and store visual information?
 - How is the recognition performed?
- Psychology, psychophysics, neuroscience, computer (cognitive) vision;
 - Workshop on generic object recognition and categorization, CVPR 2004
 - Workshop on object categorization, ICCV 2007



Complexity of recognition




- complex objects/scenes
- intra-category variability
- varying pose (3D rotation, scale)
- cluttered background/foreground
- occlusions (noise)
- varying illumination



Intra-category variability


- Women, Fire, and Dangerous Things by G. Lakoff
- Prototypical versus exemplar models



ETH-80 database

Pose and intra-variability

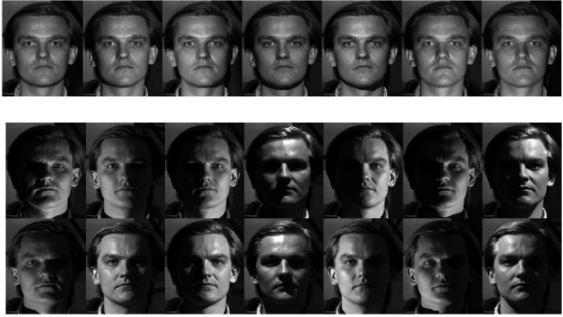
Pose/Shape:



A Physician Riding a Donkey, by Niko Pirosmanashvili


You Who Can't Do Anything, by Francisco Goya

Illumination

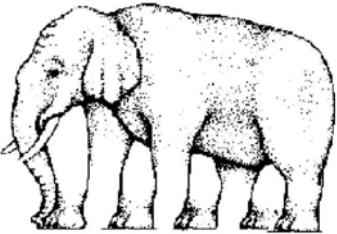


Yale Face Database


Illumination - Outdoor environment



Global visual consistency



Global visual consistency



Magritte: "Carte Blanche"

Occlusion is one of the most powerful cues to depth order.

Tampering with occlusions yields violent disorder of pictorial space, only local regions are coherent, global order is lost.

www.phys.uu.nl/~wwwpm/HumPerc/koenderink.html

Global recognition preceding local



Components of a recognition system

- Object representations
- Feature extraction
 - How reliable/stable are the features?
 - How difficult is it to extract them?
- Object database organization
- Model matching/indexing

Visual cues - Intrinsic properties

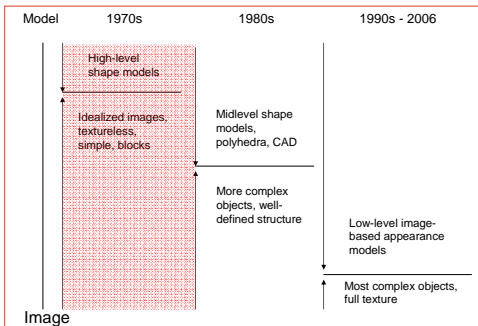
- Visual cues
 - Contours
 - Color
 - Texture
 - Shading
 - Depth (Stereo)
- Intrinsic properties
 - Shape
 - Reflectance properties
 - Illumination

Representations

- Prototypical models (abstract descriptions)
- Exemplar models (e.g., 2D or 3D templates, exact geometry)
- Object centered approaches (a single 3D model)
 - Compact, efficient, but hard to extract from the data
 - Comparing 2D to 3D (viewpoint invariant features)
- Viewer centered approaches (reduces to 2D)
 - Easy to extract from the data, but complexity!
- Simple versus complex features (Gestalt)
 - Power of complex indexing features versus
 - difficult recovery from images

Evolution of object models

Adapted from Y. Keselman and S. Dickinson, *Generic model abstraction from examples*, PAMI 2005



Object-centered, volumetric models

- Generalized cylinders
 - Sizes, shapes, positions, orientations
- Considerable variations within a class
- Examples:
 - All coffee mugs
- Hierarchically defined models (weak constrains to exemplars)
- Different levels of abstraction
- Major drawback: recovery of these high level models from images
- Brooks's ACRONYM system [1983]

ACRONYM (Brooks & Binford, 1981)

The diagram illustrates the ACronym system. On the left is a grayscale image of an airplane. On the right is a skeletal structure of the same airplane, with parts labeled: 'FORELAGE (1)', 'WING (2)', 'TAIL (3)', 'LANDING GEAR (4)', and 'ENGINE (5)'. The ACronym system is a method for representing objects as a set of labeled parts.

Generalized Cylinders

Structural Description in Terms of Volumetric Primitives

The diagram shows a photograph of a person on the left and its structural description on the right. The structural description uses volumetric primitives (cylinders) to represent the person's body and limbs. A red arrow points from the photograph to the structural description. The structural description is labeled with numbers 1 through 10, representing different parts of the person's body.

(Nevatia & Binford, 1973)

From image to 3-D description

The flowchart illustrates the process of converting an image into a 3-D description. It starts with an image of a teapot. The process involves several steps: Edges and corners, Curves and junctions, Patches, part hypotheses, and 3D parts and connectivity graph. The process is divided into three main stages: Grouping (generic properties), Grouping (properties of GCs), and Grouping (object level).

G. Medioni, *Generic shape learning and recognition*, Workshop on Generic Object Recognition and Categorization (CVPR 2004)

Geons, superquadrics

- A restricted set of generalized cylinders
 - Geons (Biederman; human vision)
 - Superquadrics (Pentland; computer vision)
- Recovery from image data has met with very little success
- Grouping and abstraction is needed
- Top-down and bottom-up

The diagram shows two columns of geometric shapes. The left column is labeled 'All Geons' and the right column is labeled 'All Superquadrics'. The shapes are used for object recognition.

Segmentation and modeling of range images

The diagram shows the segmentation and modeling of range images. It includes a 3-D model of a teapot and its corresponding range image. The process involves segmenting the range image into parts and modeling them using superquadrics.

- A. Leonardis, A. Jaklic, and F. Solina, "Superquadrics for segmentation and modeling range data", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19, pages 1289-1295, 1997.

Interpretation trees

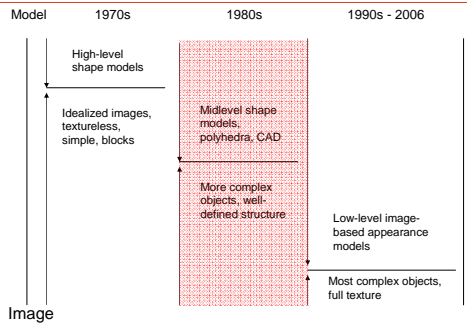
- Given
 - The list of feature descriptors from a given object model
 - The list of feature descriptors detected in the image
 - A list of (geometric) constraints that model features must satisfy
- Find a mapping between model features and image features such that the constraints satisfied by the model features are satisfied by the corresponding image features.

The diagram shows a 3-D Model of a teapot and its corresponding range image. The 3-D Model is represented by a set of oriented cylinders. The range image is represented by a set of pixels. A mapping is shown between the 3-D Model and the range image.

3-D Model
 • Represents: 3D Structure
 • Primitives: Oriented cylinders
 generalized

Evolution of object models

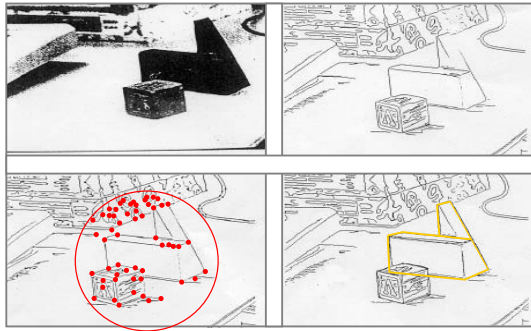
Adapted from Y. Keselman and S. Dickinson, *Generic model abstraction from examples*, PAMI 2005



Object-centered, feature-based

- Correspondence between 2D features in images and 3D features in models
- Properties
 - Viewpoint invariance
 - Locality
 - Ease of recovery
- Lines (not only due to shape but also reflectance and illumination)
- Corners (triplets of corners)

Huttenlocher & Ullman (1987)



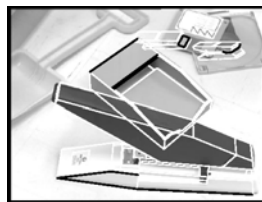
Object-centered, using perceptual groups

- More discriminative features (to reduce a search space)
- Gestalt principles
 - Parallelism
 - Collinearity
 - Proximity
 - Symmetry
- Example: David Lowe's approach
 - Still polyhedral objects
 - Still relying on one-to-one correspondence (exemplar-based approach)
 - Faster indexing, more complex detection

Object-centered, using perceptual groups

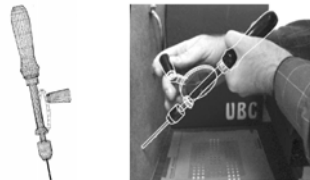
- 3D object recognition with multiple 2D views
- Extract feature groupings
- Indexing 3D object from 2D images
- Efficient search to validate matches

Black lines indicate feature groupings, white lines indicate possible matches (Beis, Lowe 1999)



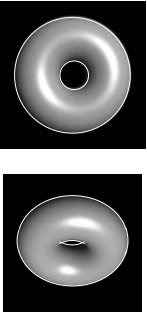
3-D Model-Based Approach

- Calibration/pose estimation problem (Lowe 1991)



- Issues:
 - Model construction, indexing
 - Class generalization
 - Occlusion, articulation

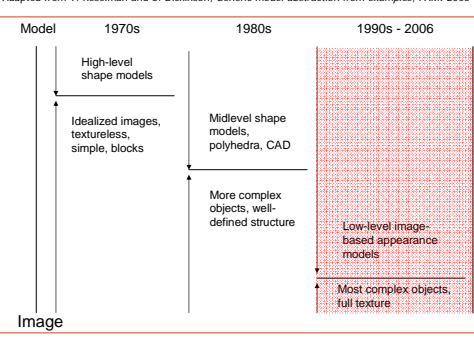
Formal geometry is nearly intractable

$$\begin{aligned}
 &13u^4r_1^4 - 8r_1^4v^4 + 24u^6v^2 - 6u^2r_1^6 + 40v^2r_2^4u^2 \\
 &- 40v^2r_2^2u^4 - 88v^4u^2r_2^2 - 40r_1^2v^4u^2 \\
 &- 40r_1^2v^4r_2^2 - 20r_1^4v^2r_2^2 - 24v^2r_2^6 + 52v^4r_2^4 \\
 &- 48v^6r_2^2 + 48v^6u^2 + 52u^4v^4 - 20r_1^2r_2^4 \\
 &+ 44r_1^2v^2r_2^4 - 8r_2^4u^4 + r_1^8 - 44u^4v^2r_1^2 + 13r_2^4r_1^4 - 12r_2^6r_1^2 \\
 &- 6r_2^2r_1^6 + 22r_2^2u^2r_1^4 - 20r_2^2u^4r_1^2 + 20v^2u^2r_1^4 + 4r_2^8 + 4u^8 \\
 &+ 16v^8 - 12u^6r_1^2 = 0
 \end{aligned}$$


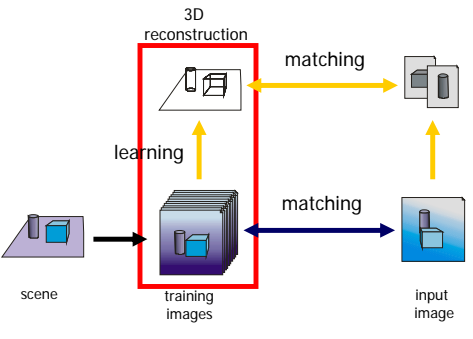
(Ponce & Kriegman, 1990)

Evolution of object models

Adapted from Y. Kesselman and S. Dickinson, *Generic model abstraction from examples*, PAMI 2005



Object-centered versus viewer-centered




Viewer-centered, global appearance-based

- Encompass combined effects of:
 - shape,
 - reflectance properties,
 - pose in the scene,
 - illumination conditions.
- Acquired through an automatic learning phase

Data acquisition

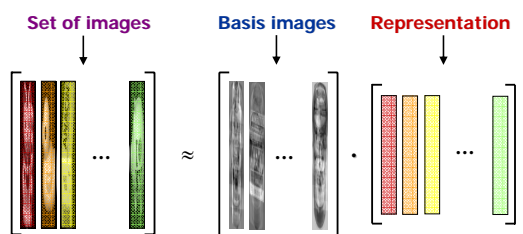
Objects are represented by a **large number of views**:



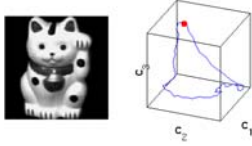
COIL Database

Subspace methods

- Images as points in high dimensional spaces
- A set of images occupies a small subspace
- Characterization of the subspace

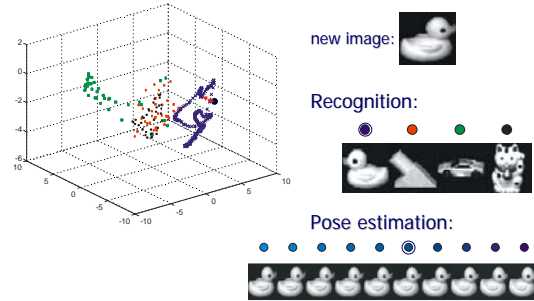


Object recognition and pose estimation

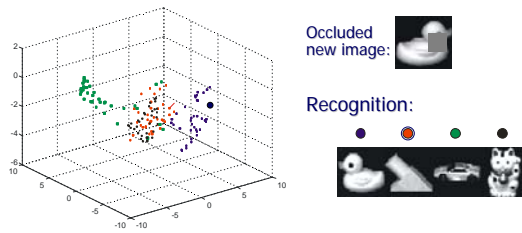


An object is represented as a manifold in the principal subspace.

Object recognition and pose estimation



Shortcomings of standard PCA



- PCA coefficients are calculated using the standard projection of the image onto the principal vectors
- all pixels are used
- inherently **non-robust!**

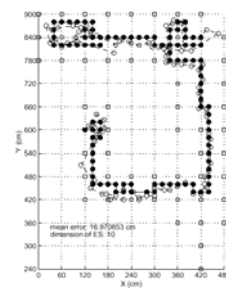
Limitations and extensions

- Suitable for object exemplars but not for object categories or prototypes
- Extensions
 - Scale invariance
 - Coping with occlusions
 - Illumination invariance
 - Incremental and robust learning

Mobile Robot



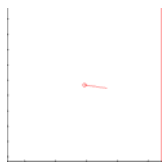
Localisation



On-line learning



- Application on a mobile robot
 - On-line learning
 - Odometry, GPS



Path (GPS) in XY plane

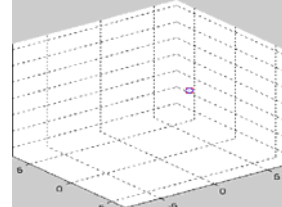


First 6 basis vectors

Eigenspace

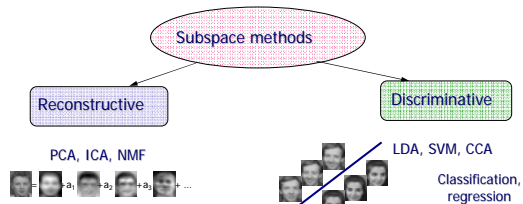


- Built incrementally



Subspace (first three dimensions)

Subspace methods



- Reconstructive
 - Enable (partial) reconstruction of input images (hallucinations).
 - More general, not specific task-dependent.
 - Enable two way processing (feedback loop)
- Discriminative
 - Store only information necessary for a specific task.
 - More specialized, specific task-dependent.
 - Do not enable (partial) reconstruction.

Internal representations



- "The reason for trying to recover the low-dimensional manifold in which the data live, instead of constructing a decision surface for a given classification problem involving these data, has to do with transfer of learning or expertise across tasks. The hyperplane constructed may be easy to learn, may afford good generalization to new examples of the same problem, however, it is useless for generalization of expertise to different sets of labels for the same data."* (S. Edelman)
- "A characterization of the (class-conditional) probability density of the data is much more informative and potentially useful than a characterization of the decision surface for a given task."* (G. Hinton)

Viewer-centered, local appearance-based



- Local photometric features
- Distinctive features
- Robust to occlusion and clutter
- Scale and affine invariance

Viewer-centered, local appearance-based

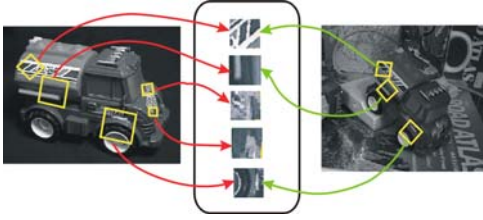


- Region detectors:
 - Difference of Gaussian (DOG)
 - Laplacian
 - Harris-Affine & Hessian Affine*: K. Mikolajczyk and C. Schmid, Scale and affine invariant interest point detectors. In IJCV 1(60):63-86, 2004.
 - MSER*: J.Matas, O. Chum, M. Urban, and T. Pajdla, Robust wide baseline stereo from maximally stable extremal regions. In BMVC p. 384-393, 2002.
 - IBR & EBR*: T. Tuytelaars and L. Van Gool, Matching widely separated views based on affine invariant regions. In IJCV 1(59):61-85, 2004.
 - Salient regions*: T. Kadir, A. Zisserman, and M. Brady, An affine invariant salient region detector. In ECCV p. 404-416, 2004.
- Region descriptor
 - Differential invariants
 - Steerable filters
 - Moments
 - SIFT*: D. Lowe, Distinctive image features from scale invariant keypoints. In IJCV 2(60):91-110, 2004.

Viewer-centered, local appearance-based



- Local invariant features: SIFT (Lowe, IJCV 2004)
 - Scale, rotation invariant key-points
 - Select and match key-points



Viewer-centered, local appearance-based



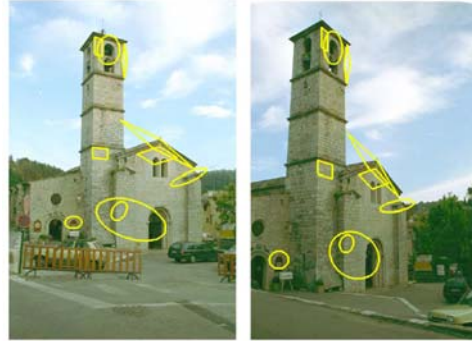
- Local invariant features: SIFT (Lowe, IJCV 2004)
 - Scale, rotation invariant key-points
 - Select and match key-points



Example



Example

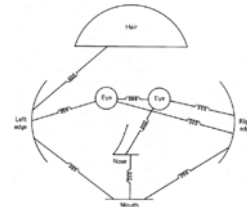


Viewer-centered, local appearance-based



- Trainable visual models for object class recognition (categorisation)
- Objectives
 - Recognition (but not perfect segmentation)
 - (Semi) unsupervised learning
- Main issues:
 - Parts
 - appearance, shape
 - Structure
 - model (e.g. implicit or explicit)
 - Model learning
 - from training data
 - Model fitting (recognition)
 - complexity

The "templates and springs" model



(Fischler & Elschlager, 1973)
Ballard & Brown (1980, Fig. 11.5). Courtesy
Bob Fisher and Ballard & Brown on-line.

Probabilistic relaxation algorithms (Rosenfeld et al., 1976)

Various approaches



- Models that learn parts, then add structure
 - Weber, Welling & Perona, Leibe & Sziele, Agarwal & Roth, Borenstein & Ullman
- Models for which the structure is primary
 - Felzenszwalb & Huttenlocher, Ramanan & Forsyth
- Models that learn parts and structure simultaneously
 - Fergus, Perona & Zisserman

Learn part models, then add structure



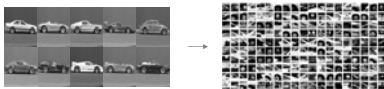
- Recognize class instances under image translation
 - Implicit structure model
 - No inter-part articulation
 - Only single visual aspect
- Extend to image scale change and rotation by exhaustive search over scale and orientation

Learn part models, then add structure

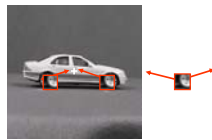
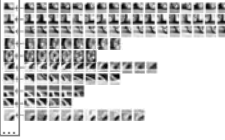


Leibe & Sziele, 2004

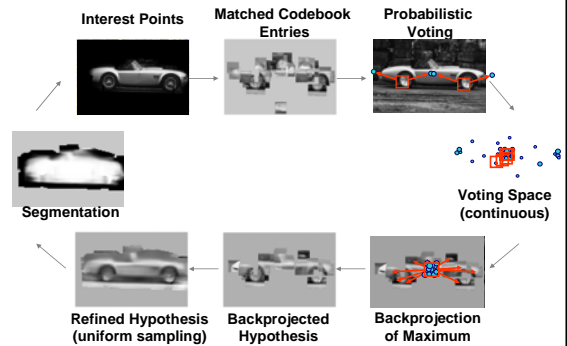
- Collect patches from whole training set



- Appearance codebook



Categorisation & segmentation Leibe & Sziele



Models for which structure model is primary



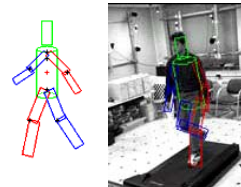
- New ideas
 - Explicit structure model
 - Articulated structure
- Detect and localize multi-part objects at arbitrary locations in a scene
 - Generic object models such as person
 - Allow for articulated objects
 - Combine 2D geometry and appearance
 - Provide efficient and practical algorithms
- Felzenszwalb and Huttenlocher

Matching pictorial structures



Felzenszwalb and Huttenlocher, 2000

- Simultaneous use of appearance and spatial information
- Minimize an energy (or cost) function that reflects both
 - Appearance: how well each part matches at given location
 - Configuration: degree to which model is deformed in placing the parts at chosen locations

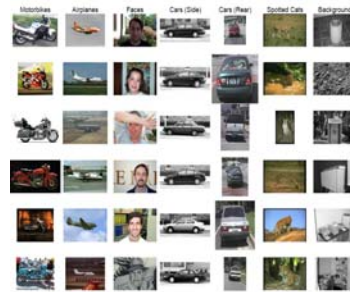


Parts & structure modeled simultaneously

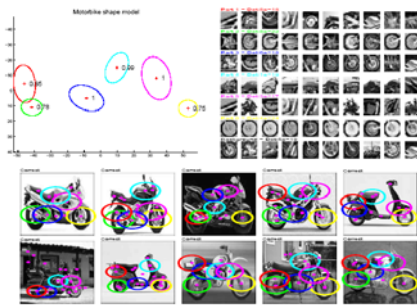


- New ideas
 - Explicit structure model – Joint Gaussian over all part positions
 - Part detector determines position *and* scale
 - Heterogeneous parts
 - Simultaneous learning of parts and structure
- Constellation model of Fergus, Perona & Zisserman 2003

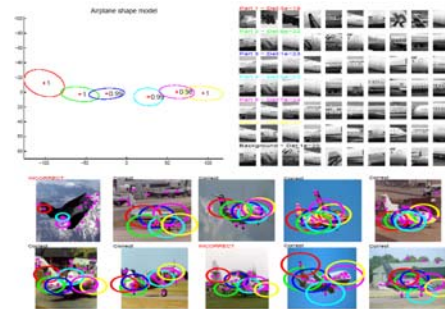
Object classes



Motorbikes



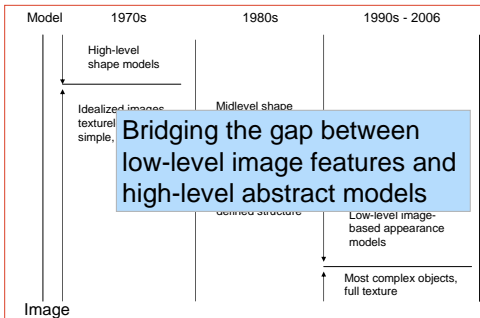
Airplanes



Evolution of object models



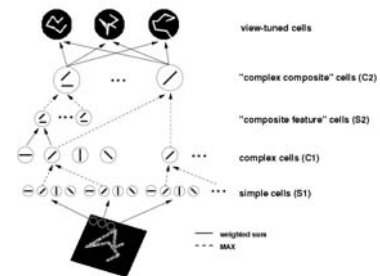
Adapted from Y. Keselman and S. Dickinson, *Generic model abstraction from examples*, PAMI 2005



Hierarchical architecture



• HMAX



Riesenhuber and Poggio, Nature 1999

Classification by feature hierarchies

Boris Epstein, Shimon Ullman
Weizmann Institute of Science

Goal
Improve recognition by using hierarchical features.

Building a hierarchy of sub-fragments
1. Define training set for the fragment.
Fragment to be detected:
Detected in images:
"Almost detected" (detected with lowered threshold):
2. Extract the most informative sub-fragments.
3. Apply the algorithm recursively to the sub-fragments.
Stopping rule: When no information gain is achieved by splitting a fragment, keep it holistic.
4. Optimize receptive field (RF) sizes of the fragments.

Classification by hierarchy
Classification by neural network:
 $r = \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^K \max(0, \frac{p_{ij} - p_{i,j+1}}{1 + p_{ij}})$

Examples & Results
Average 80% accuracy, 90% 100% x 10 100%

Conclusions
Hierarchical features are more informative than original features
- Features and sizes of sub-features need to be chosen adaptively
- Size of Receptive Field need to be chosen adaptively
- Optimal depth of hierarchy: 3-5 levels
- High-level fragments are class-specific
- Low-level fragments are shared between classes

Extracting informative fragments
Decomposition into sub-fragments increases tolerance to deformations and local illumination changes.
RF too small → low number of hits
RF too big → high number of false alarms
Choose RF size to maximize the Mutual Information between the fragment and

Overview of the architecture

Architecture

Summary: Evolution of object models

- From prototypical models (class-based or generic models) to exemplar-based models (template-, appearance-based) to
- prototypical constellations (trainable visual models for object class recognition) to
- hierarchical representations.

Human visual system

Retina: Rods: 120 million (light sensitive – not color)
Cones: 6 million (color sensitive, high acuity)

Brain: V1-V2 complex: Map for edges
V3: Map for form and local movement
V4: Map for colour
V5: Map for global motion
Number of neurons: 10^{10} - 10^{11}
Neuron fan-out: 10^2 - 10^4

Distributed local representations

M. Tanifuji, Nature 2001

References

- S. Dickinson, Object representation and recognition, In: E. Lepore and Y. Pylshyn (eds.) Rutgers University Lectures on Cognitive Science, 1999, pp. 172-207
- Y. Keselman and S. Dickinson, *Generic model abstraction from examples*, PAMI 2005
- B. Leibe, A. Leonardis, and B. Schiele, *Combined Object Categorization and Segmentation with an Implicit Shape Model*, ECCV04 Workshop on Statistical Learning in Computer Vision
- D. G. Lowe, SIFT: Distinctive image features from scale-invariant keypoints, IJCV 2(60):91-110, 2004.
- G. Medioni, *Generic shape learning and recognition*, Workshop on Generic Object Recognition and Categorization (CVPR 2004)
- J. Ponce, *Toward true 3D object recognition*, Workshop on Generic Object Recognition and Categorization (CVPR 2004)
- A. Zisserman, *Trainable visual models for object class recognition*, Pascal Pattern Recognition and Machine Learning in Computer Vision Workshop
- J. Tsotsos, "Analyzing vision at the complexity level", Behavioral and Brain Sciences, 13(3), 1990, pp. 423-445.
- Tsunoda K, Yamane Y, Nishizaki M, Tanifuji M. *Complex objects are represented in macaque inferotemporal cortex by the combination of feature columns*. Nature Neuroscience, 2001, 4(8):832-8