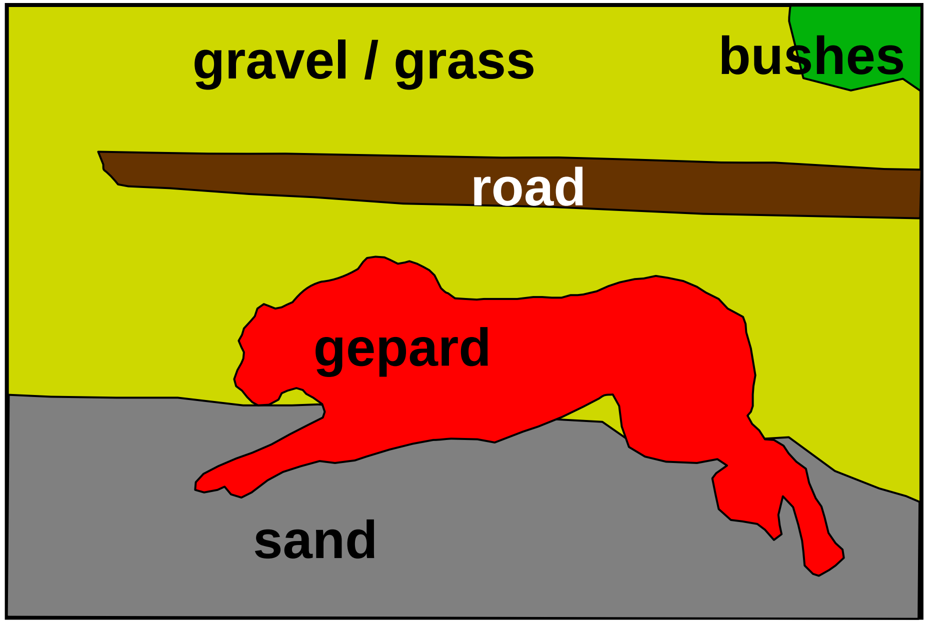
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Compositional Models for Object Recognition / Categorization

Joachim M. Buhmann

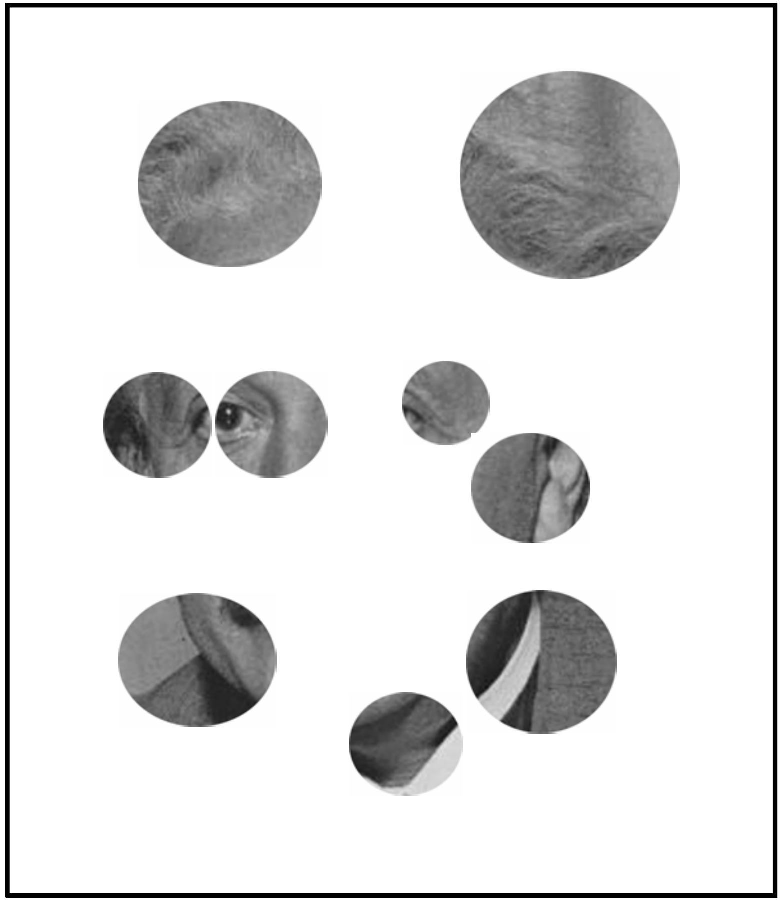
Institute for Computational Science, ETH Zurich

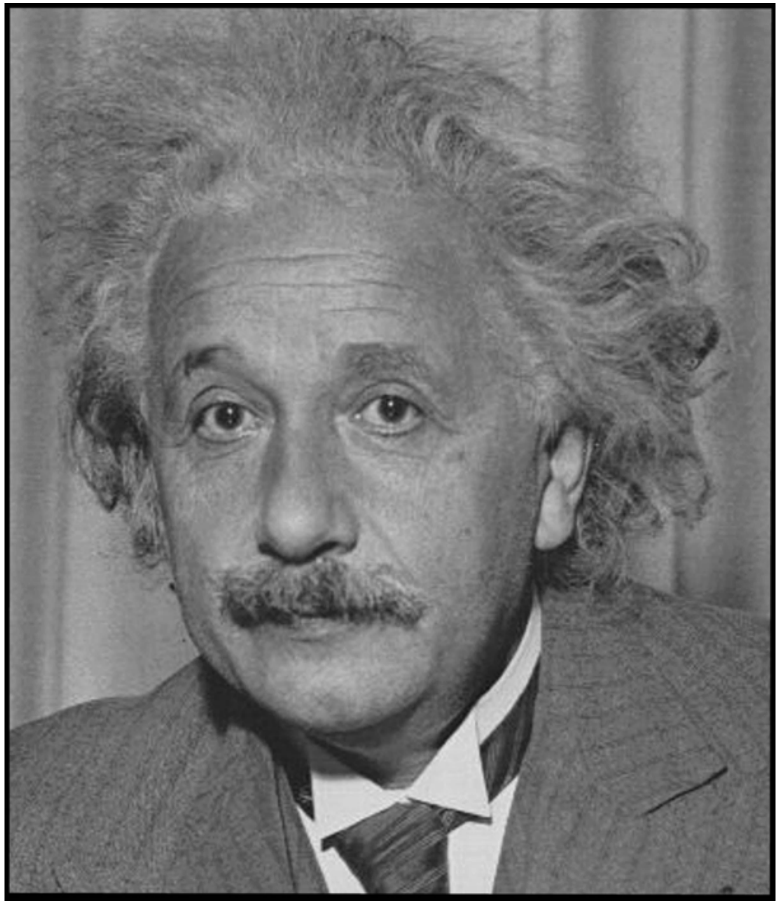
From Images to Objects (M. Minsky 1959, summer project)



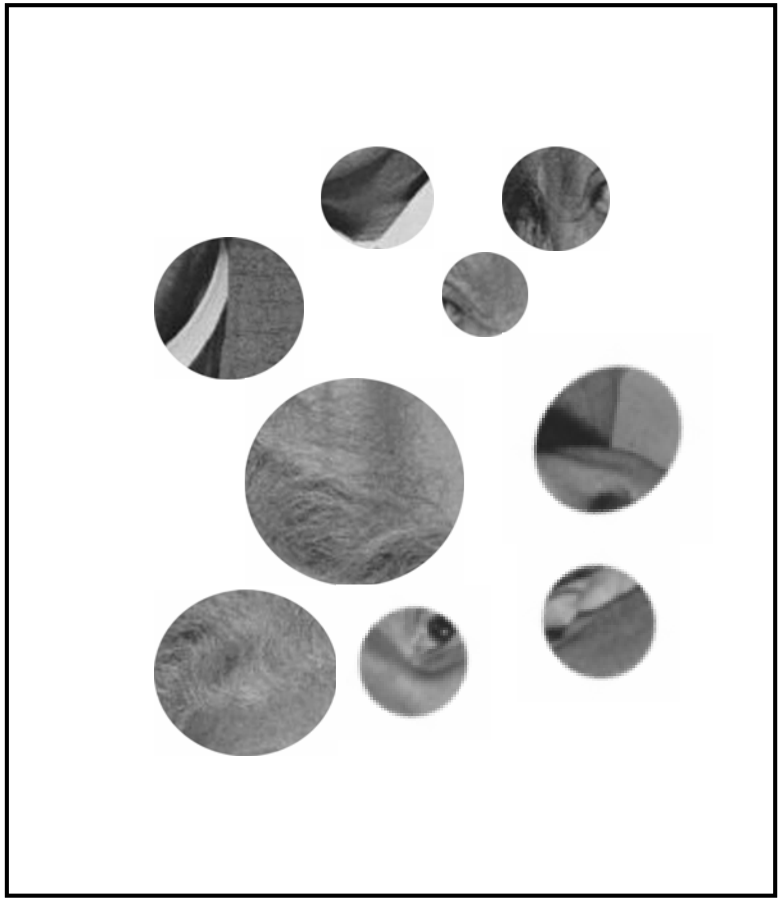
- **segmentation** **Unsupervised**
- **object recognition / categorization** **Supervised**

How Complicated is Object Recognition ?

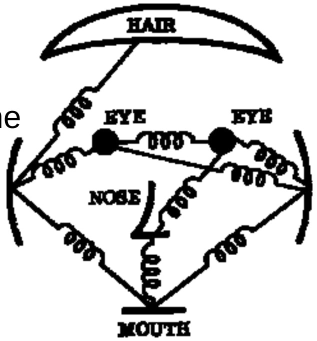




Recognition by Key Features and Spatial Reasoning

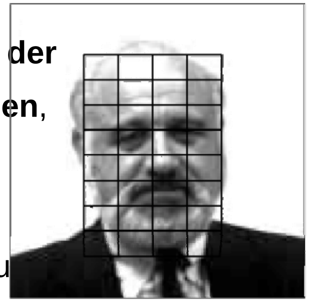


- **Constellation models**



Fischler, M.A., **Elschlager**, R.A.: The representation and matching of pictorial structures. IEEE Tr. Comput. 22 (1973)

Lades, M., **Vorbrüggen**, J.C., **Buhmann,** J.M., **Lange**, J., **von der Malsburg**, C., **Würtz**, R.P., **Konen**, W.: Distortion invariant object recognition in the dynamic link architecture. IEEE Trans. Compu 42 (1993)



Fergus, R., **Perona**, P., **Zisserman**, Object Class Recognition by Unsupervised Scale-Invariant Learning. CVPR (2003)

Position Statement: My Beliefs for Propagation

1. **Vision** requires **complex** (probabilistic) models since the world contains a lot of (stochastic) structure!
2. **Good representations** in vision should work for a **set of tasks** rather than a single task!
3. **Vision problems** are solved by **learning** since the required model complexity is too high for "hand crafting"! => **unsupervised learning**

Requirements on Vision Representations

- **Representations** should have properties like being
  - … **flexible** & **adaptive**, **modular**;
  - … **robust**;
  - … **expressive**;
  - … **explanatory**;
  - … **learnable**.

- **Modelling and algorithmic ingredients** are …
  - **growing**, adaptive, nested **structures** & **self-assembly**;
  - **statistical** models & **inference**;
  - combination of **global** relations with **local** measurements;
  - **generative** models for the "interesting" parts of the image;
  - **complexity** control dependent on sample size.

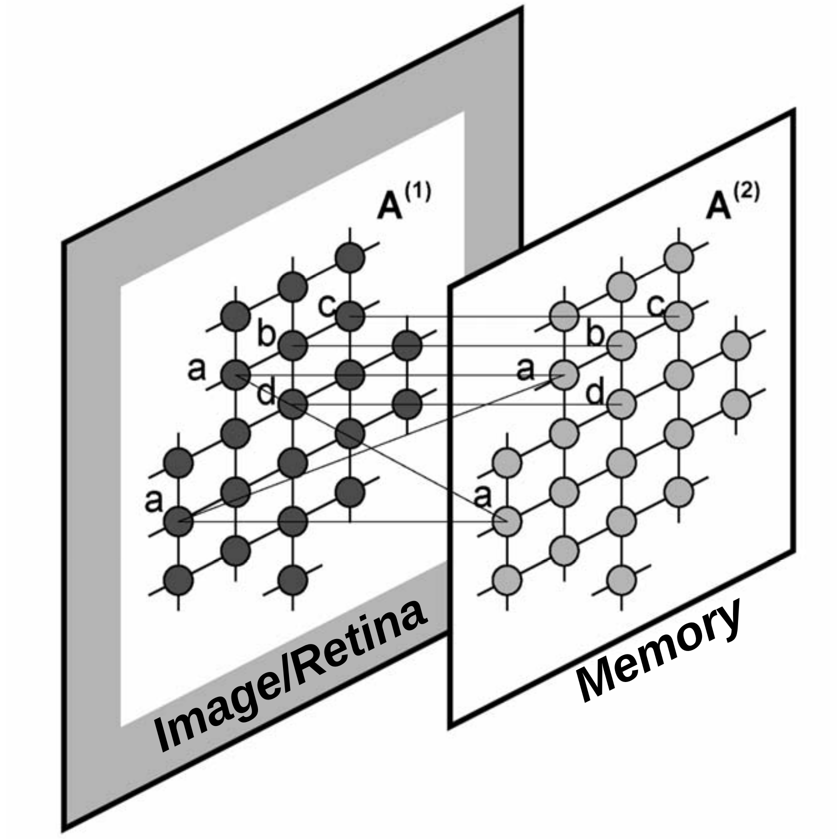
Algorithmic Needs of Vision Systems

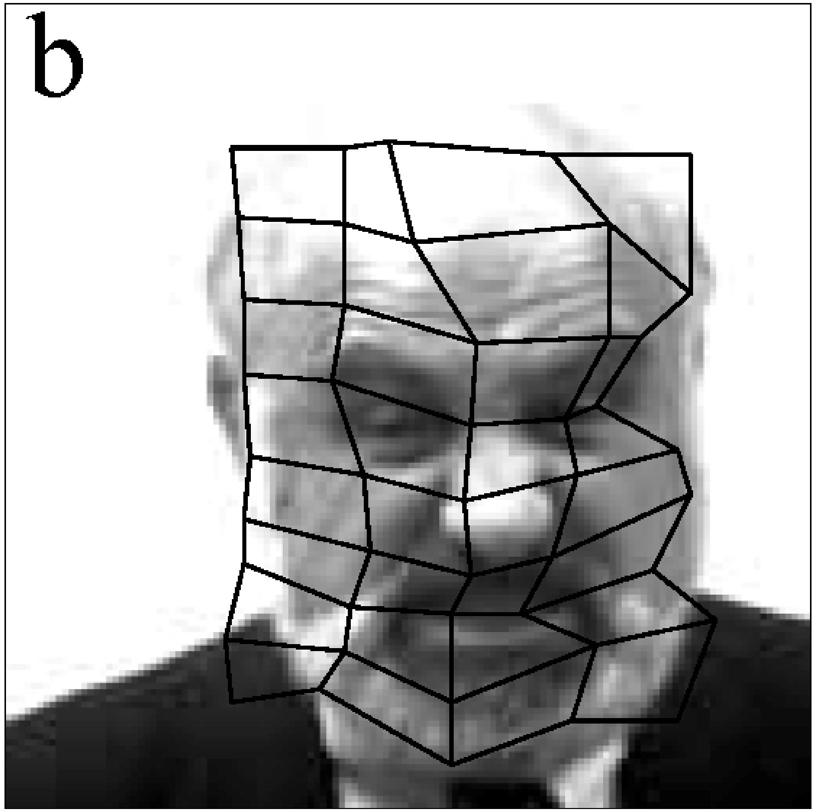
- **Algorithms** should be computationally and statistically **efficient**!
- Nested hypothesis classes $\mathcal{H}_1 \subset \mathcal{H}_2 \subset \dots \subset \mathcal{H}_k \subset \dots$
- Hypothesis class often grows with sample size.
- **Averaging** of statistically equivalent hypotheses.
- **probably approximately correct learning** (PAC)
- **approximative multi-scale** optimization
- extend concepts of learning.
- Bayesian inference, Max. Entropy, nonparametrics

Face Recognition with Dynamic Links

(JB, J. Lange, C. von der Malsburg)

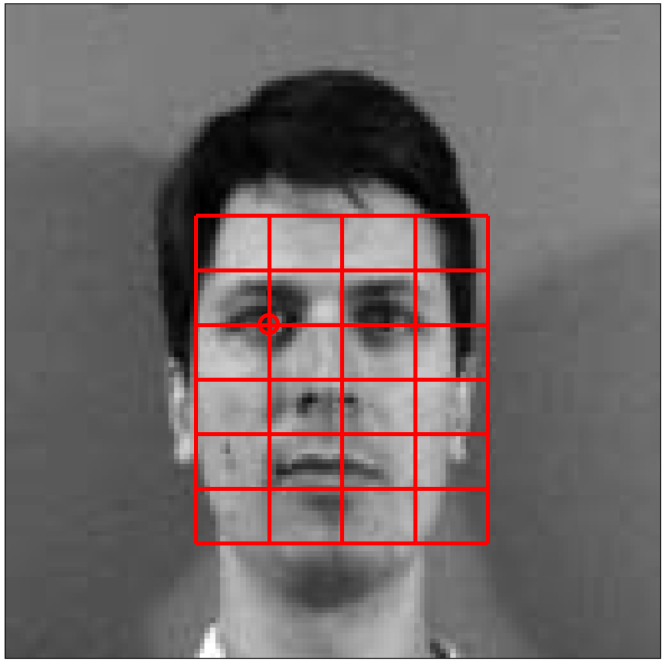
- ***Dynamic Link Architecture*** recognized person (M. Arbib)

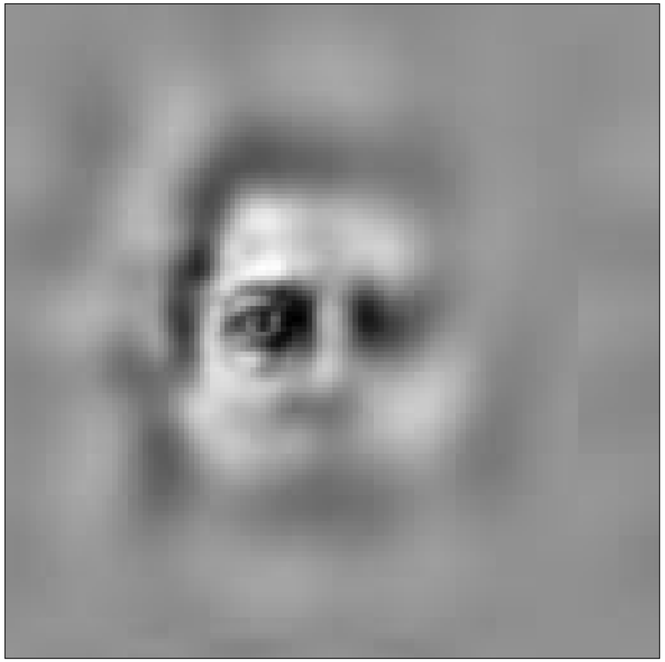




What are flexible/adaptive representations?

- **Object variations** or deformations can be captured, e.g., facial expression, object invariant articulation, perspective distortions …

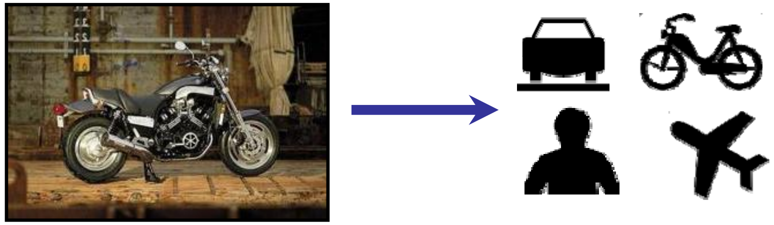




dense
or
sparse
?

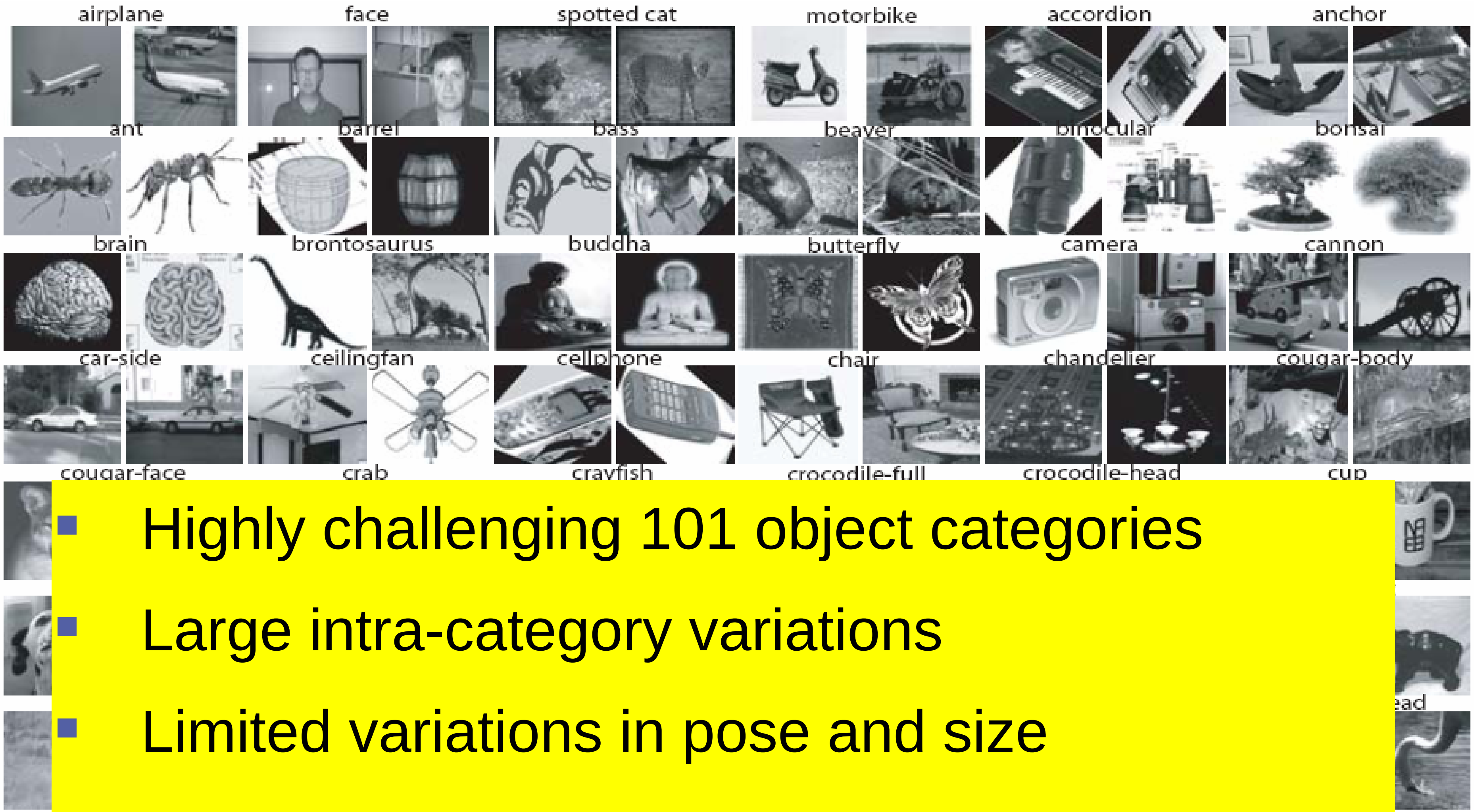
Object Categorization

- **Task**: Learn to classify w/o manual segmentations



- **Challenge**: Large intra-category variations

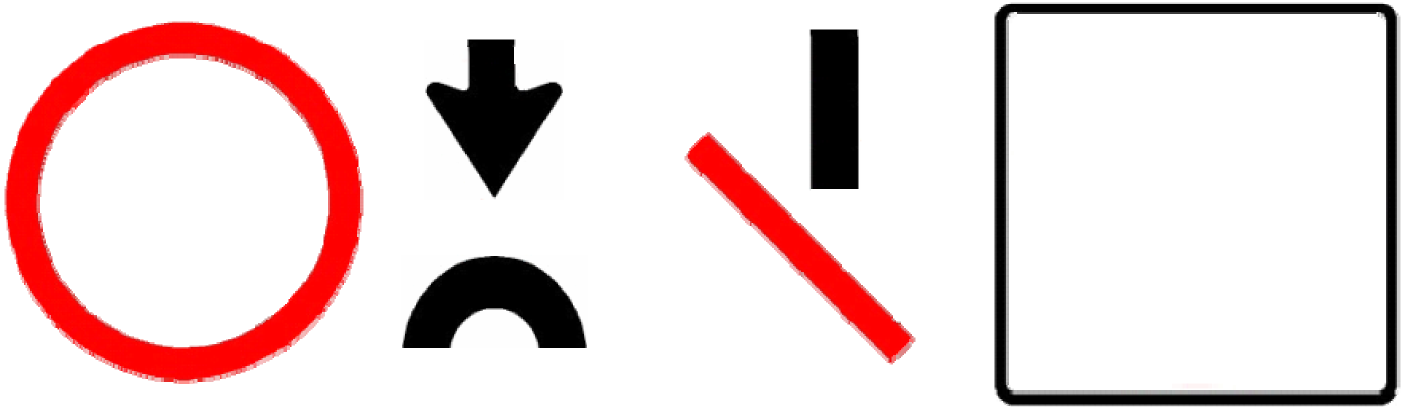
Dealing with many Categories: CalTech 101



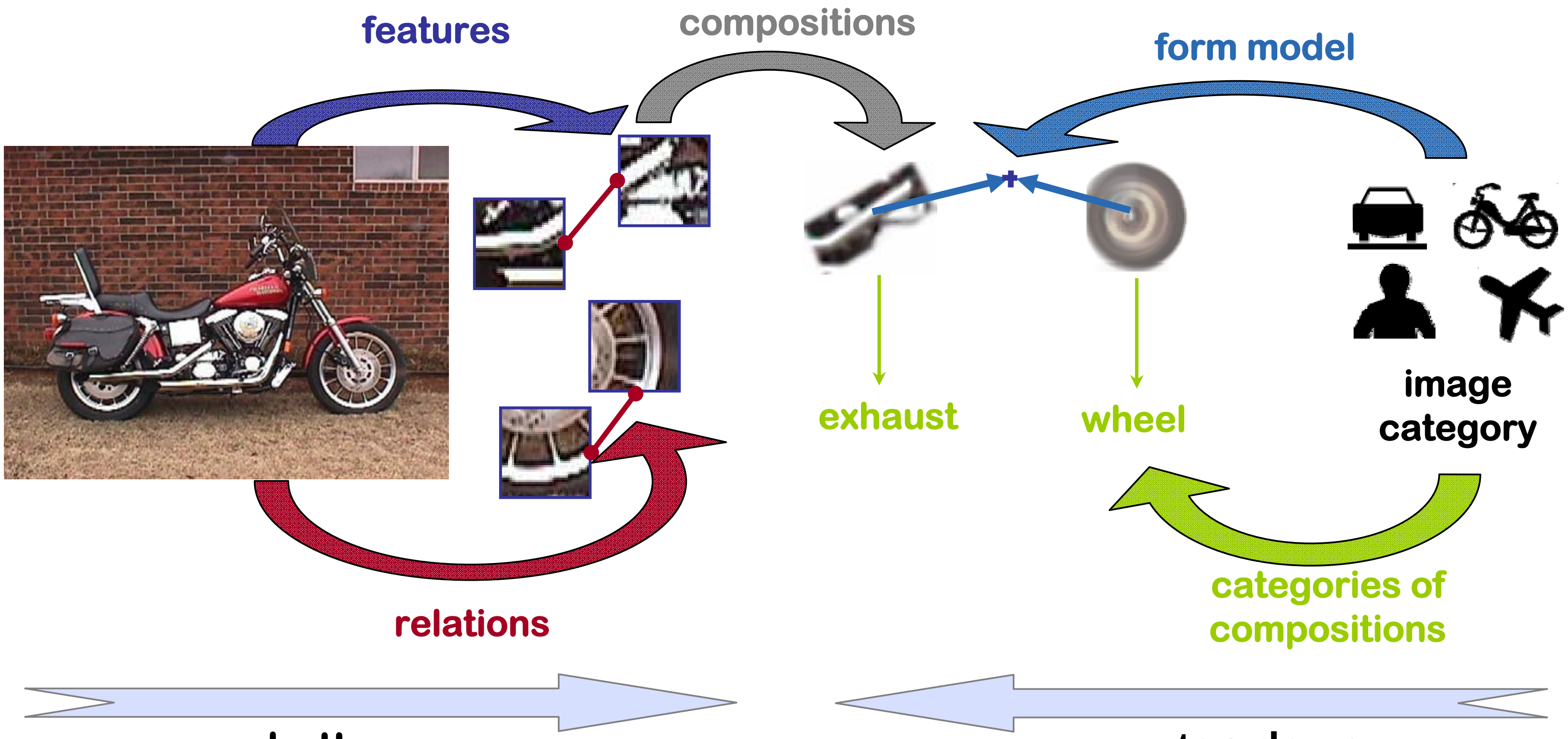
- Highly challenging 101 object categories
- Large intra-category variations
- Limited variations in pose and size

Compositionality (S. Geman)

- Simple, widely reusable parts & relations between them ⇨ Compositions



Information Flow for Image Interpretation



bottom-up:
data driven

top-down:
model driven

Methodology of the Compositional Approach

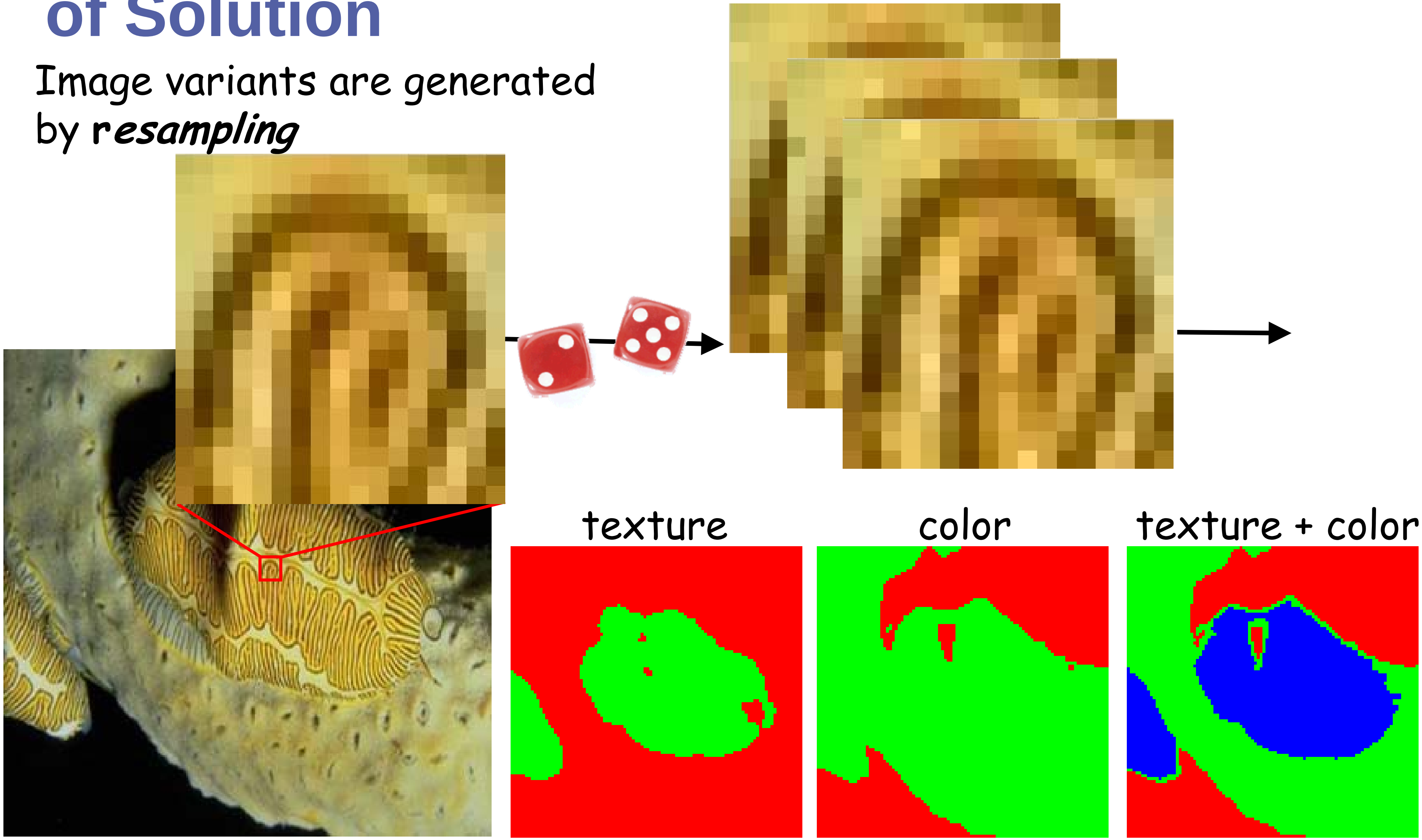
1. **Perceptual grouping** yields initial set of **salient compositions** & limits representation complexity.
2. **Top-down grouping** forms category distinctive **composition hierarchies**.
3. **Unsupervised learning** of top-down grouping probabilities without information on compositions in training images.
4. **Spatial coupling** of compositions using a **probabilistic shape model**.

The Role of Segmentation in Object Recognition

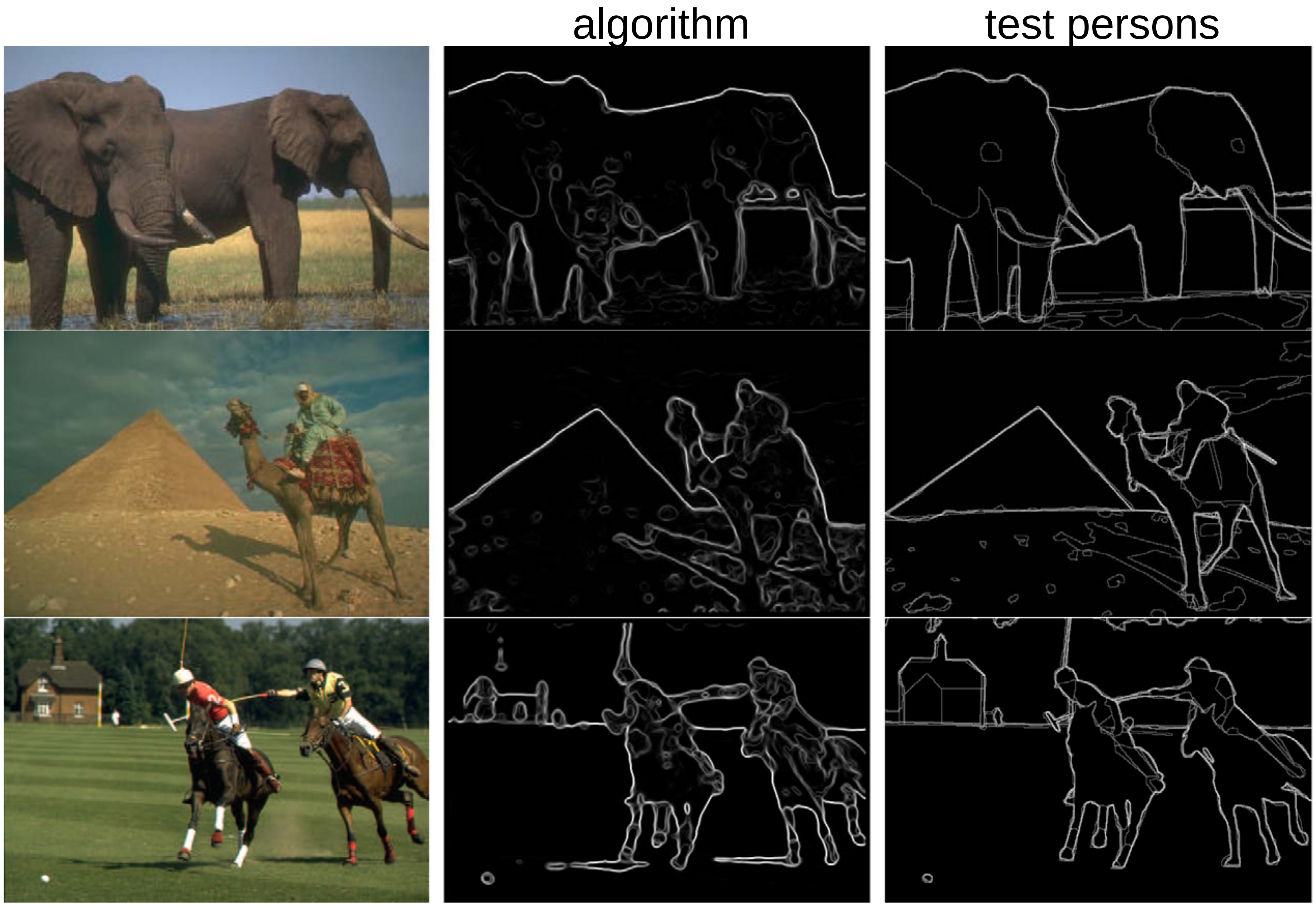
1. Segmentation is a smart preprocessing for **feature extraction**.
2. Segmentation **controls** the **recognition process.**
3. It defines the metric for detecting non-accidentalness and common cause.
4. …?

Image Variations yield Distribution of Solution

Image variants are generated by ***resampling***

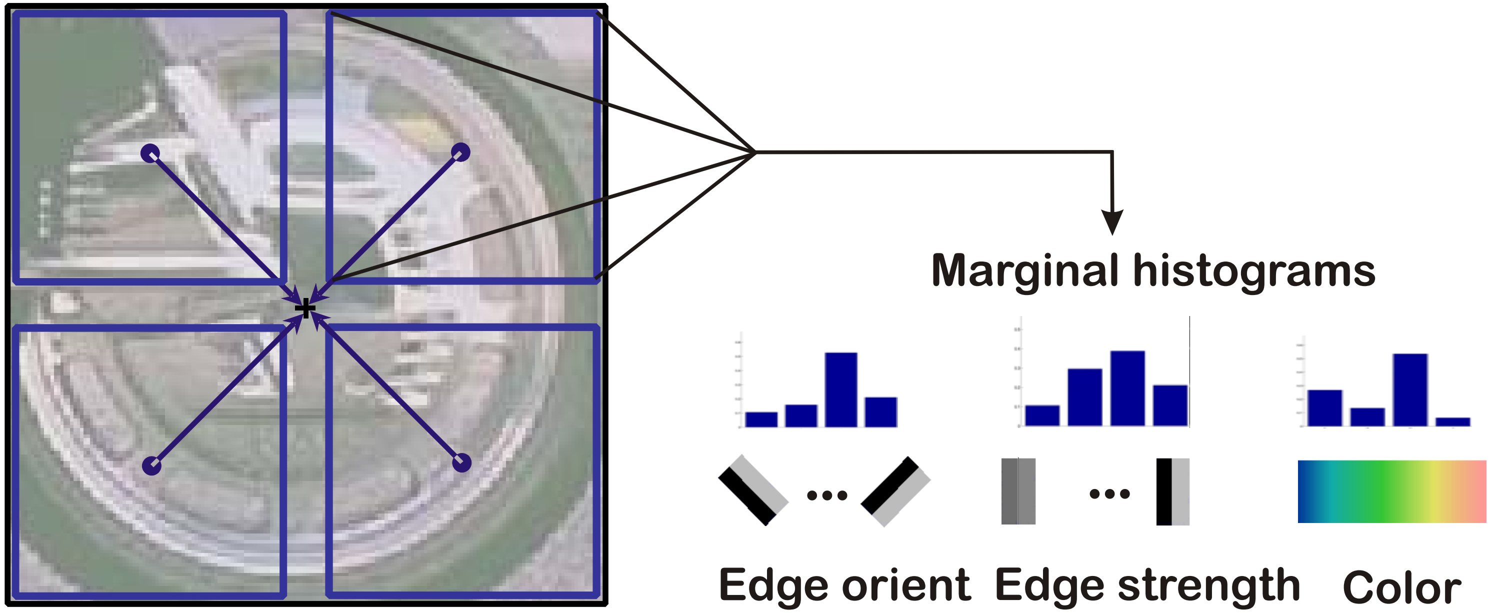


Aggregated Segmentations



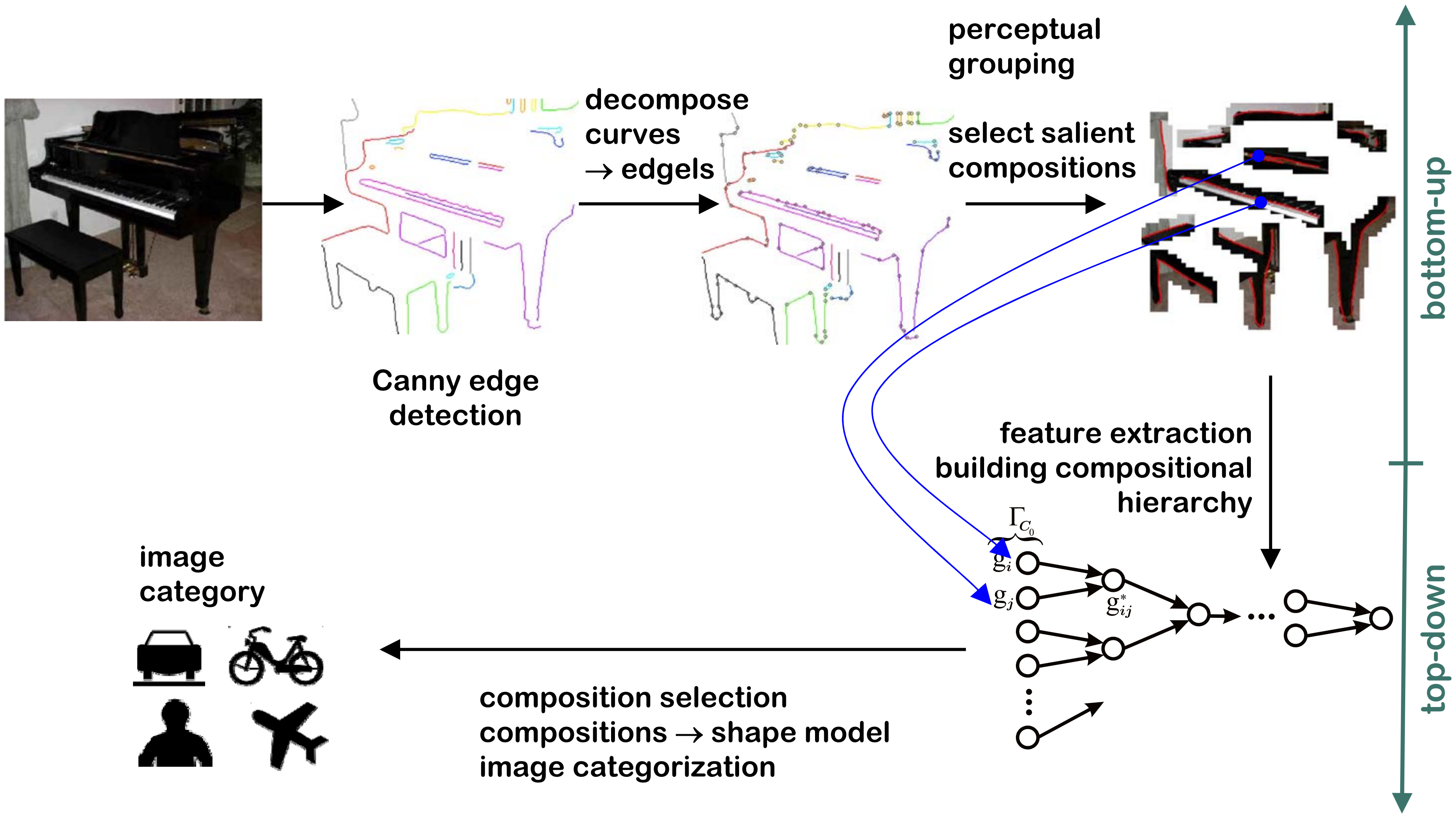
Localized Feature Histograms

- Along grouped curve segments, features are extracted as local part descriptors



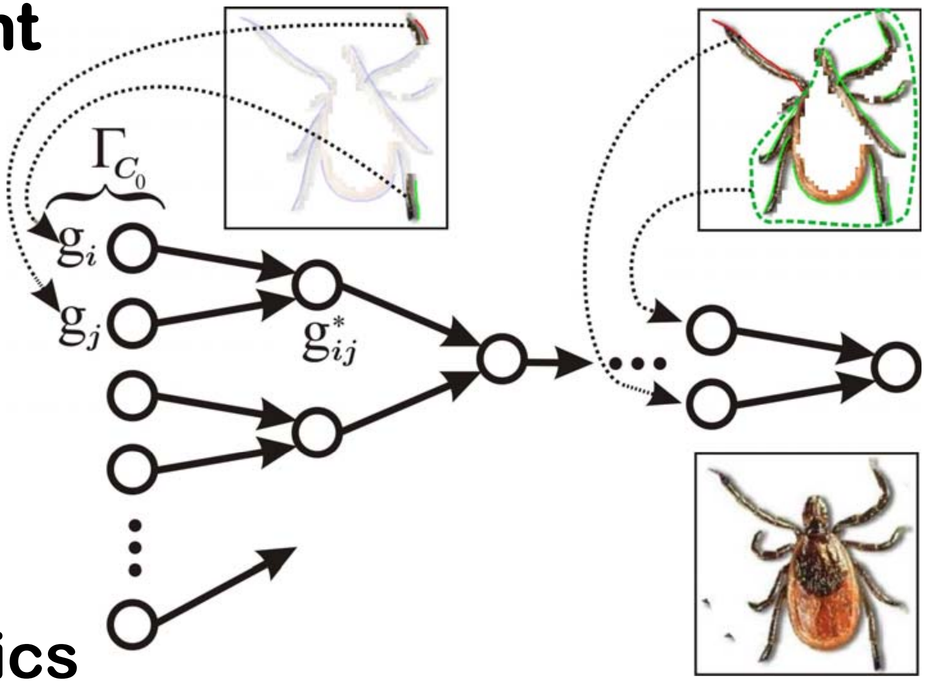
- Local descriptor is Gibbs distrib. over codebook

Recognition Phase



Applying Top-Down Grouping

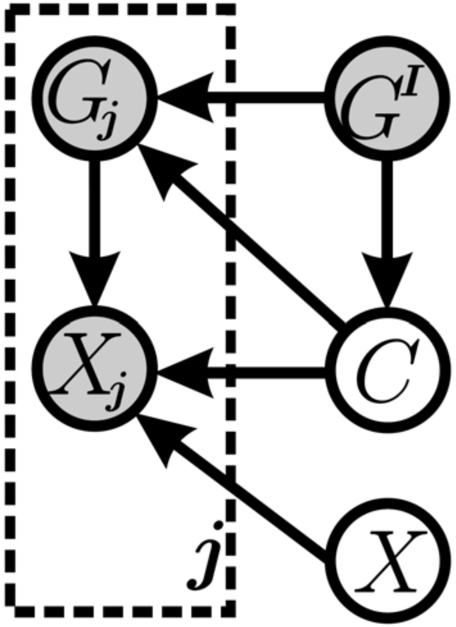
- **Start with set Γ_C of salient compositions from perceptual bottom-up grouping**
- **Recursive grouping of compositions using learned grouping statistics**

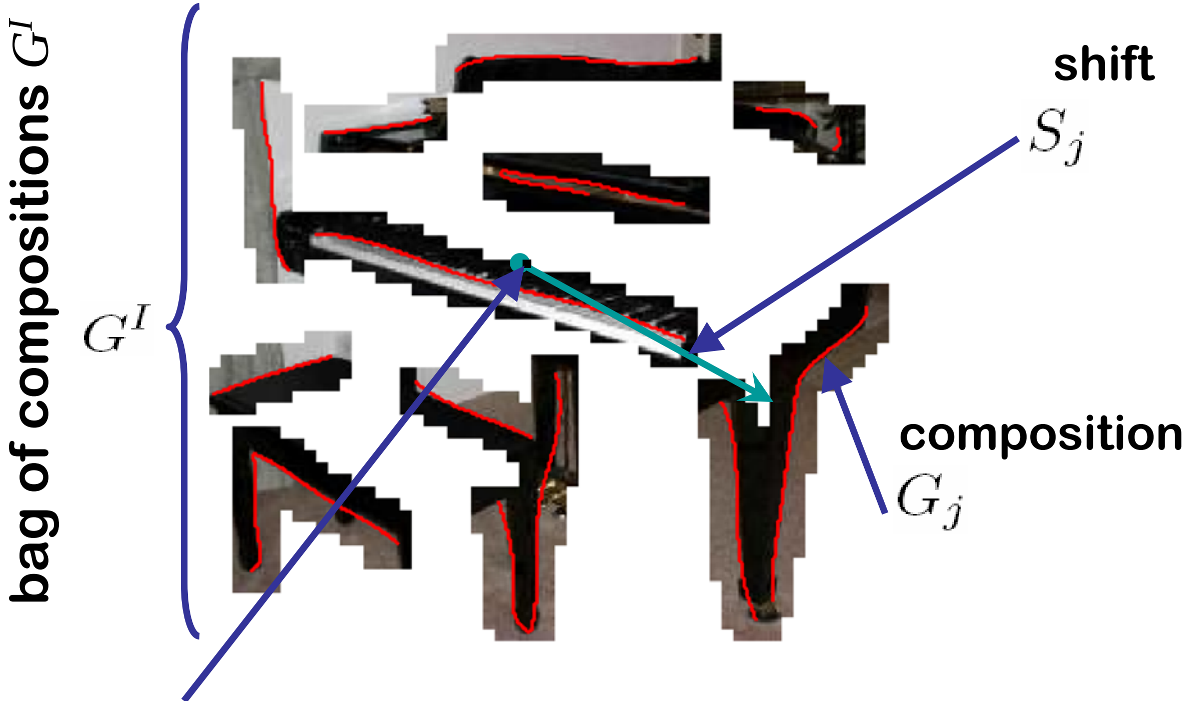


$$\mathbf{g}^*_{ij} = \operatorname*{argmax}_{\mathbf{g}_{ij}:\mathbf{g}_i,\mathbf{g}_j \in \Gamma_C} \max_{c \in \mathcal{L}} P(c|\mathbf{g}_{ij})$$

$$\Gamma_C \leftarrow \Gamma_C \cup \{\mathbf{g}^*_{ij}\} - \{\mathbf{g}_i, \mathbf{g}_j\}$$

Shape Model for Binding Compositions





$$P\left(c \middle| \mathbf{g}^I, \mathbf{x}, \{\mathbf{g}_j, \mathbf{x}_j\}_{j=1:|\Gamma_L|}\right) = \exp\Big[(1 - |\Gamma_L|) \ln P(c|\mathbf{g}^I) + \sum_{\mathbf{g}_j \in \Gamma_L} \ln P\left(c | S_j = \mathbf{x} - \mathbf{x}_j, \mathbf{g}_j, \mathbf{g}^I\right)\Big]$$

object center $\mathbf{x} = \sum_j \mathbf{x}_j \sum_{c \in \mathcal{L}} p(\mathbf{g}_j | c, \mathbf{g}^I)\ P(c|\mathbf{g}^I)$

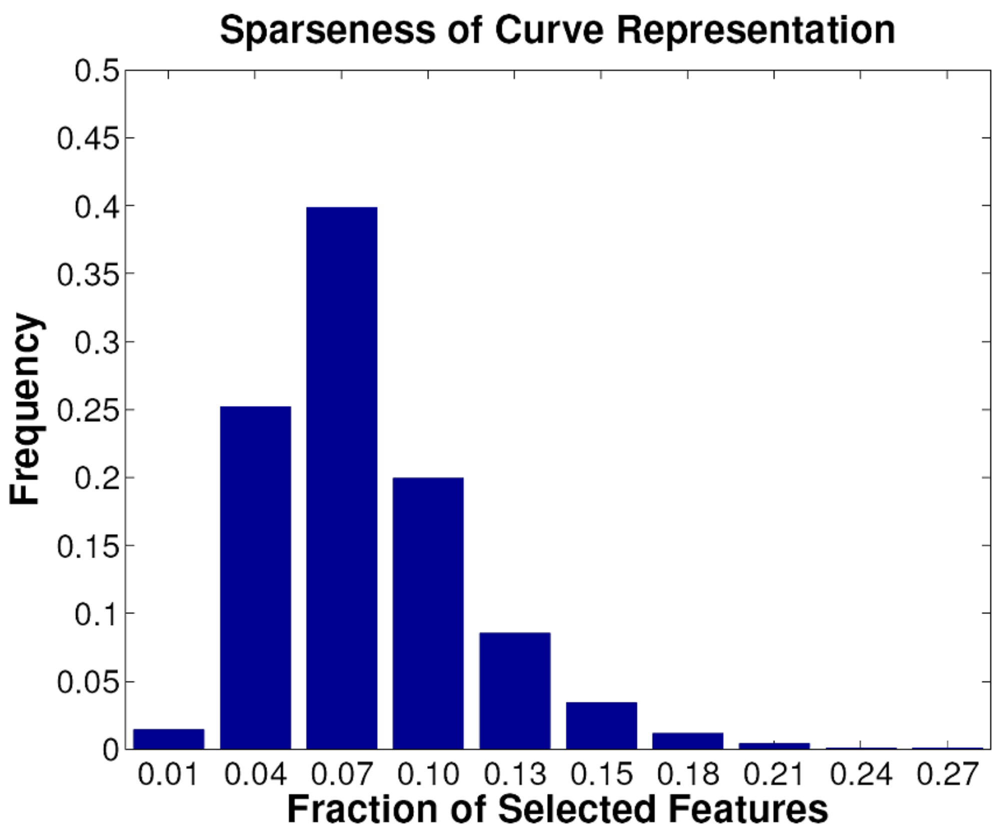
Performance of Compositional Model

- **Retrieval rate**

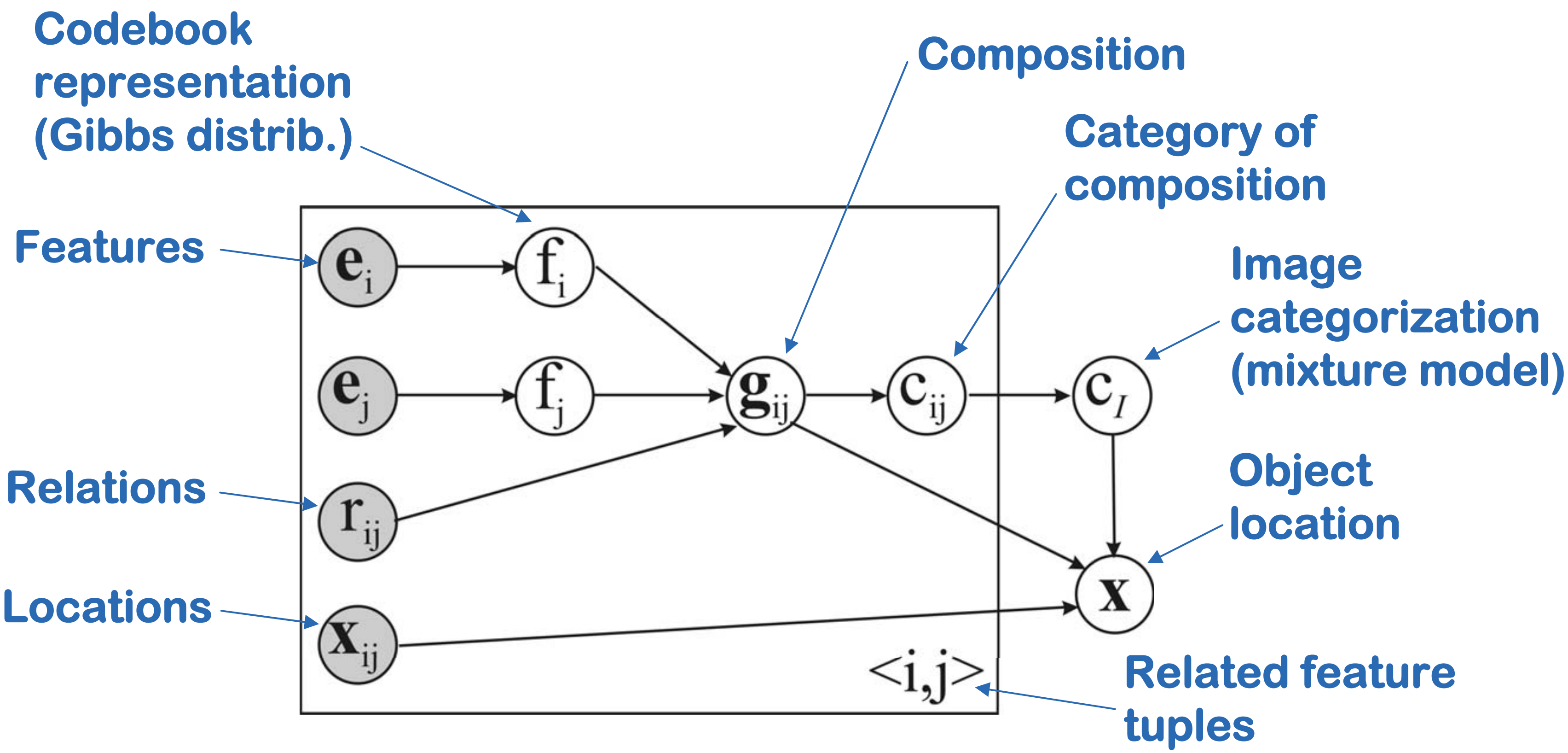
  53.0 ± 0.5 % (single scale)

  58.0 ± 0.8 % (multi-scale)

- **Sparseness** of induced image representation (fraction of selected features)



Bayesian Net of the Architecture



Summary & Perspectives

Learning and **generalization** in vision refers to the general problem of **robust optimization**!

There exist challenges for **unsupervised learning** in vision which are conceptually (much) harder than supervised learning in classification.

Fundamental problem: How is **statistical complexity** related to **computational complexity**?

We have to learn complex models with few data!