Using vision to understand the world by the combination of learning and innate mechanisms.

Understanding the world through vision

‘Digital Baby’

Input: image sequences

Innate capacities

→

197	178	172	127	165	173
218	204	202	196	193	194
215	198	186	180	187	184
218	199	203	195	191	156
167	170	134	193	106	110
95	157	160	204	168	151
192	197	203	197	187	175

→

219	188	218	204	202	196
190	235	215	198	186	180
163	223	218	199	203	195
210	224	167	170	134	193
226	179	95	157	160	204
216	193	192	197	203	197
218	221	204	203	186	218

→

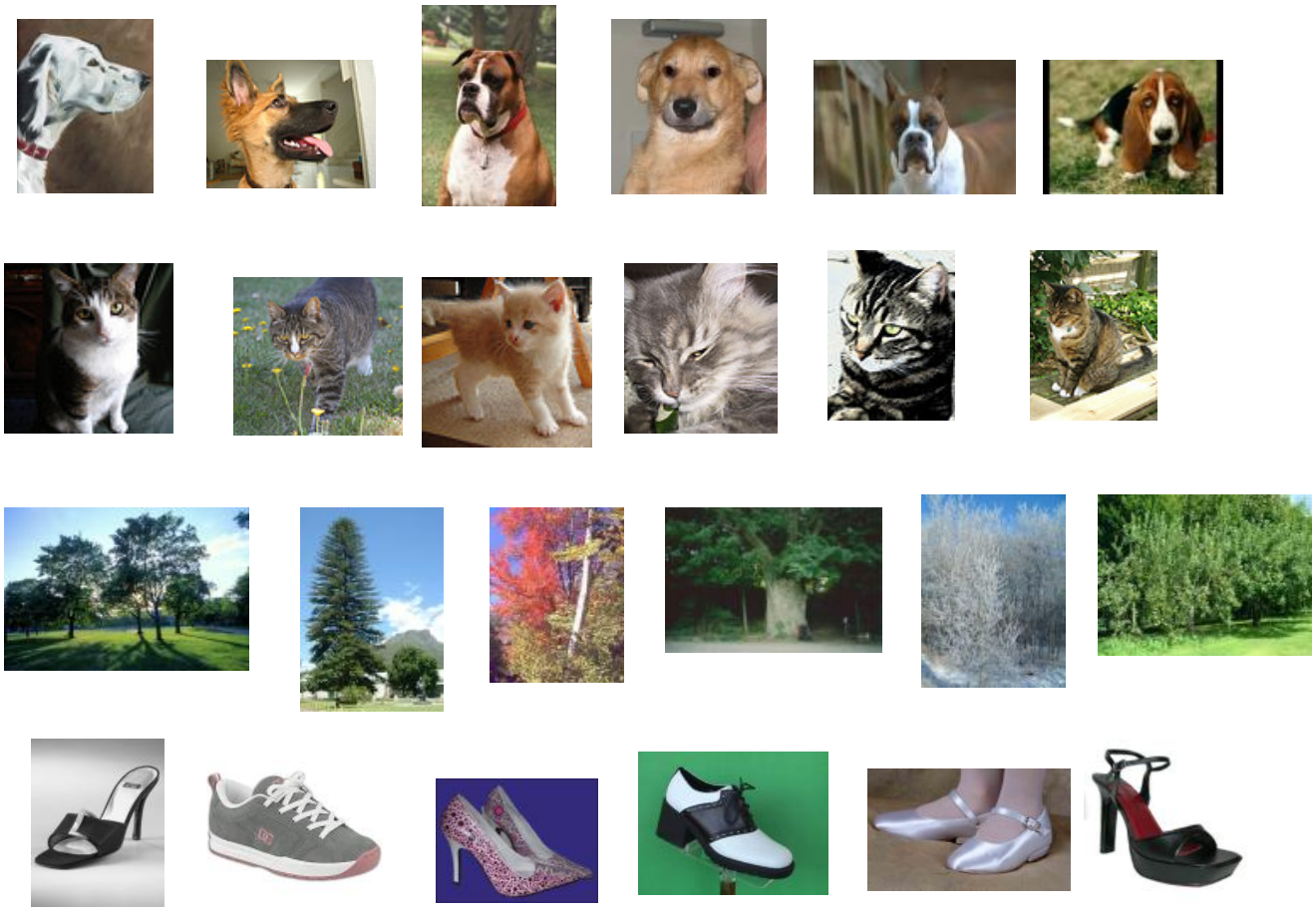
184	113	118	105	117	82
151	95	122	131	87	100
160	156	159	197	178	172
136	219	188	218	204	202
184	190	235	215	198	186
175	163	223	218	199	203
221	210	224	167	170	134

• • •

Objects ←
Actions
Agents
Goals
Tools
Interactions

Learn visual interpretation

Object Categories

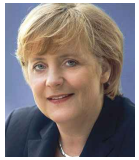


-- We perceive the world in term of objects and classes

-- Large variability within a each class

Individual Recognition

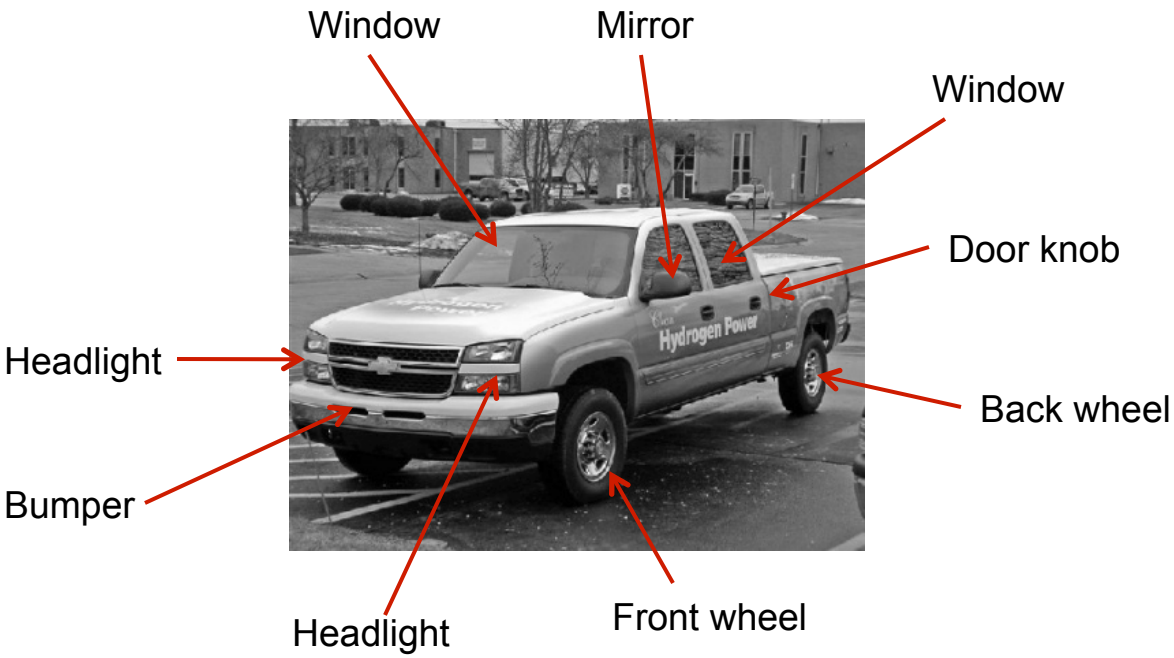




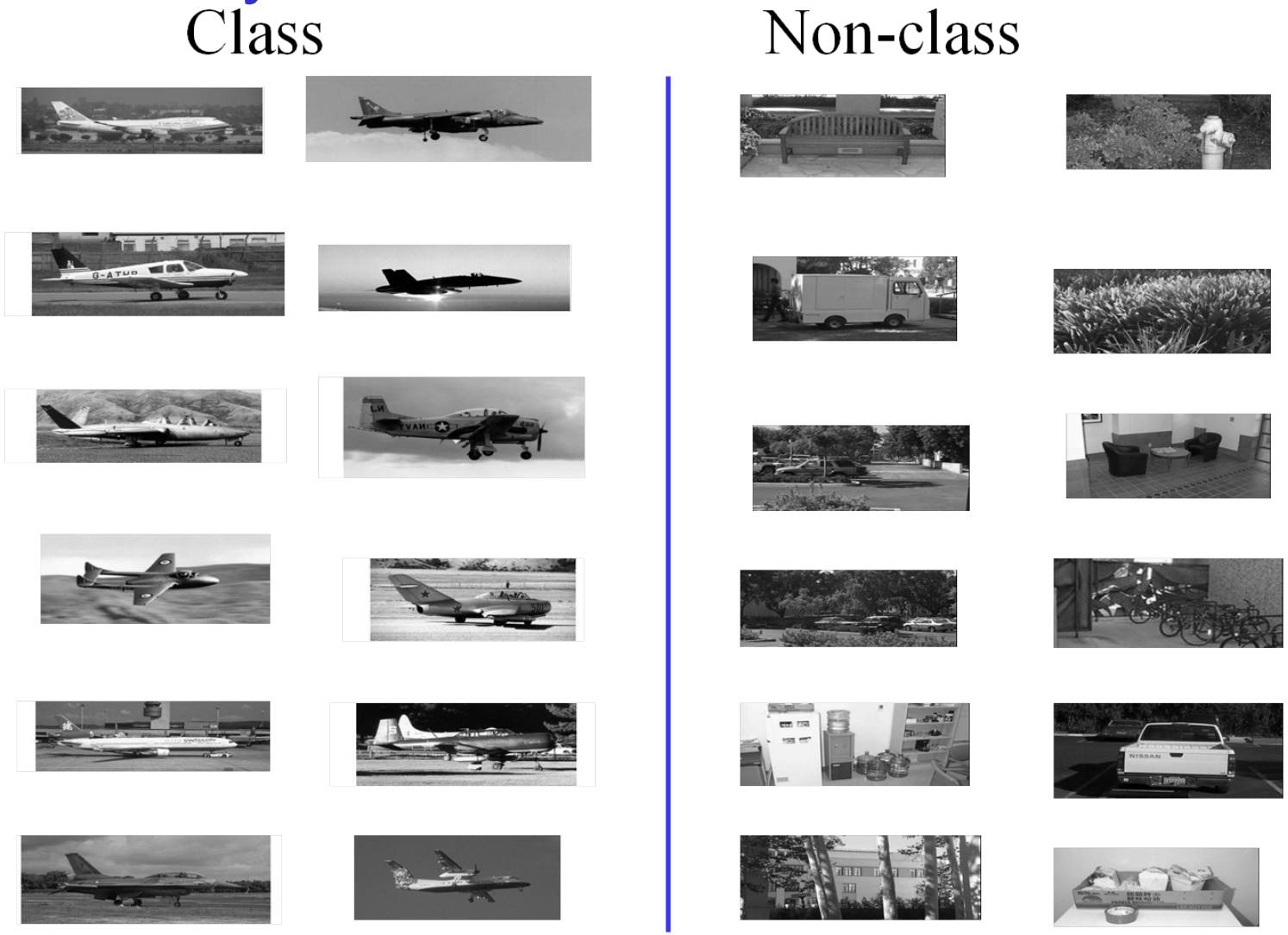





Object parts



I. Categorization: dealing with class variability



We learn categories from examples

Class

81	82	85	70	195	247	91
133	82	85	88	76	89	88
89	110	92	85	85	85	89
160	69	129	87	85	85	88
92	82	125	79	89	89	89
128	129	106	99	93	93	90
185	127	122	135	162	93	90

197	178	172	127	165	173
218	204	202	196	193	194
215	198	186	180	187	184
218	199	203	195	191	156
167	170	134	193	106	110
95	157	160	204	168	151
192	197	203	197	187	175

219	188	218	204	202	196
190	235	215	198	186	180
163	223	218	199	203	195
210	224	167	170	134	193
226	179	95	157	160	204
216	193	192	197	203	197
218	221	204	203	186	218

Non-class

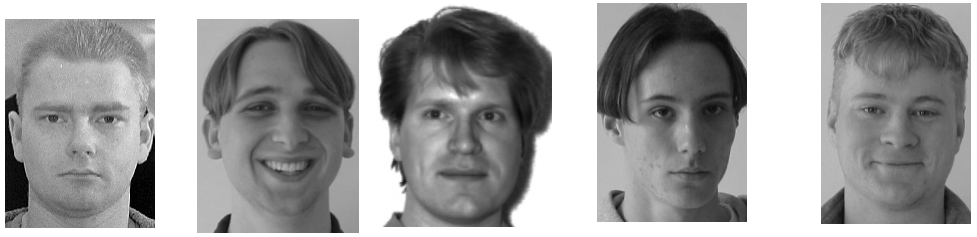
164	171	161	215	88	96	89
174	178	165	255	123	91	89
189	179	189	255	113	88	89
206	151	170	0	86	96	91
81	162	251	252	14	38	97
154	236	252	214	8	33	20
249	244	249	224	16	24	5

184	113	118	105	117	82	
151	95	122	131	87	100	
160	156	159	197	178	172	
136	219	188	218	204	202	
184	190	235	215	198	186	
175	163	223	218	199	203	
221	210	224	167	170	134	

	202	196	193	194	185	127
	186	180	187	184	190	199
	203	195	191	156	176	196
	134	193	106	110	76	174
	160	204	168	151	176	195
	203	197	187	175	169	185
	186	218	183	156	179	201

Natural for the brain, difficult computationally

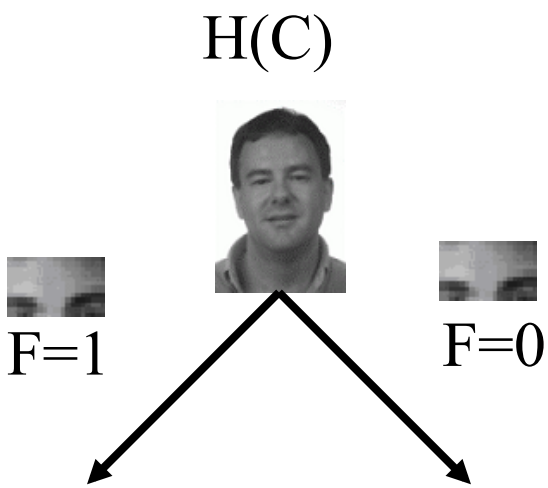
Visual Class: Similar Configurations of Shared Components



What will be optimal building-blocks for the class?

Mutual information

$$H(c) = -\sum P(c)Log(P(c))$$



$$I(C;F) = H(C) - H(C/F)$$

Mutual Information I(C,F)

Class:	1	1	0	1	0	1	0	0
Feature 	1	0	0	1	1	1	0	0

$$I(F,C) = H(C) - H(C|F)$$

$$\sum_{C,F} p(C,F) log \frac{p(C,F)}{p(C)p(F)}$$

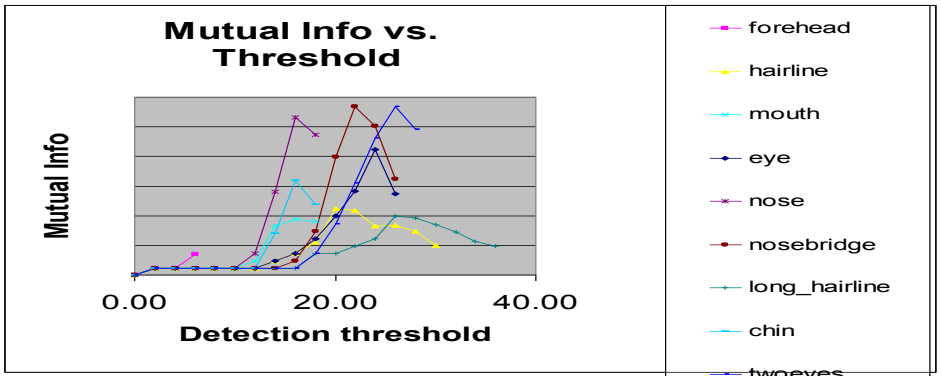
Optimal classification features

- Theoretically: maximizing delivered information minimizes classification error

$$\text{Error} = H - I(C;F)$$

- In practice: informative object components can be identified in training images

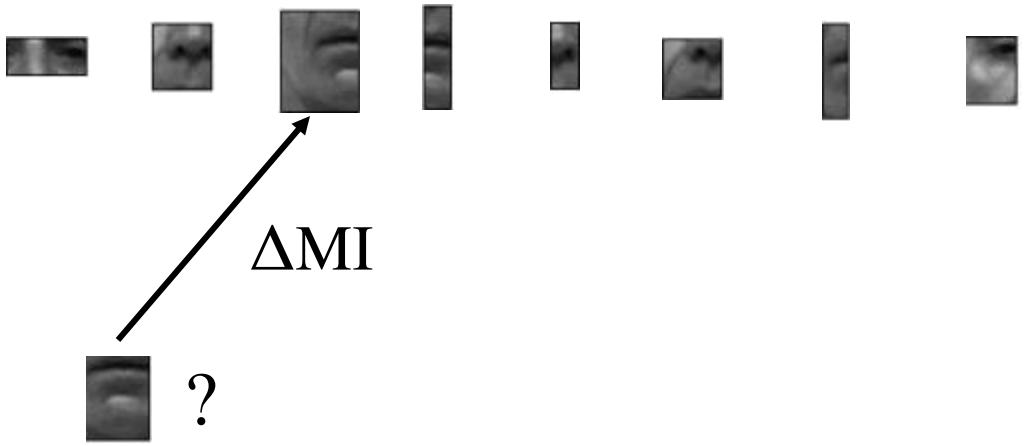
Selecting Fragments



‘Imprinting’ many receptive fields and selecting a subset

Adding a New Fragment

(Avoiding redundancy by max-min selection)



Compare new fragments Fi to all the previous ones.

Select F which maximizes the additional information

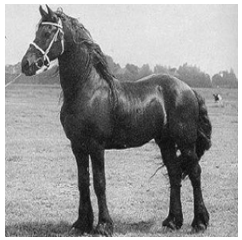
$$\mathrm{Max}_i \, \mathrm{Min}_k \, \Delta MI \, (Fi, Fk)$$

Competition between units with similar responses

Highly Informative Face Fragments

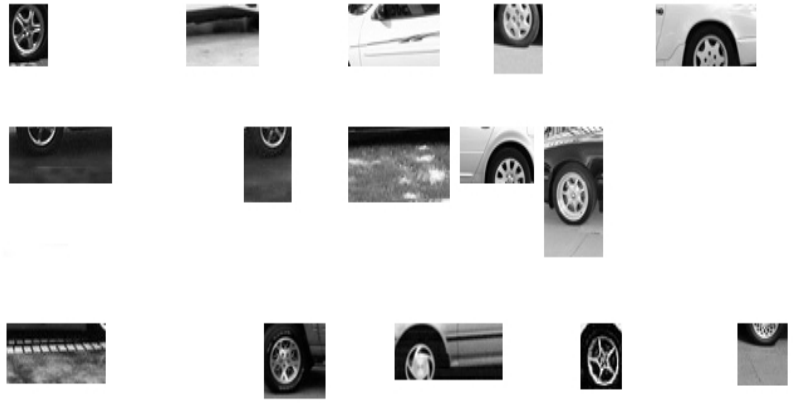
Optimal receptive fields for Faces

Informative class features

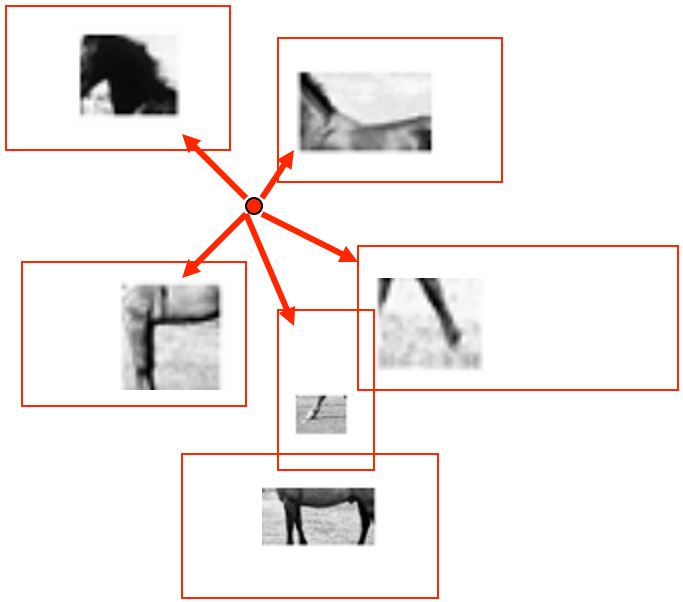


Horse-class features

Car-class features



Fragments with positions



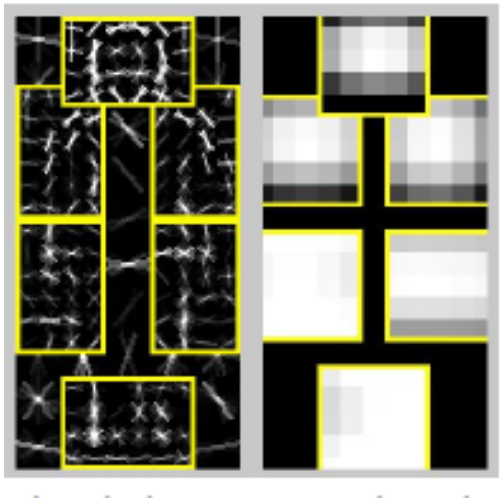
$$\sum w_k F_k > \theta$$

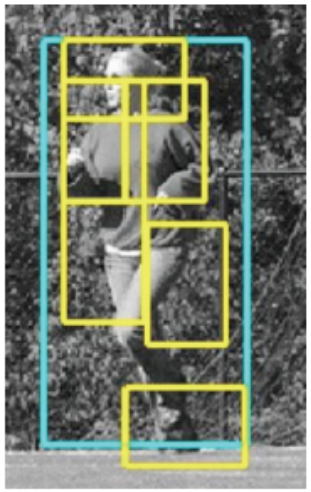
On all detected fragments within their regions

Variability of Horses Detected

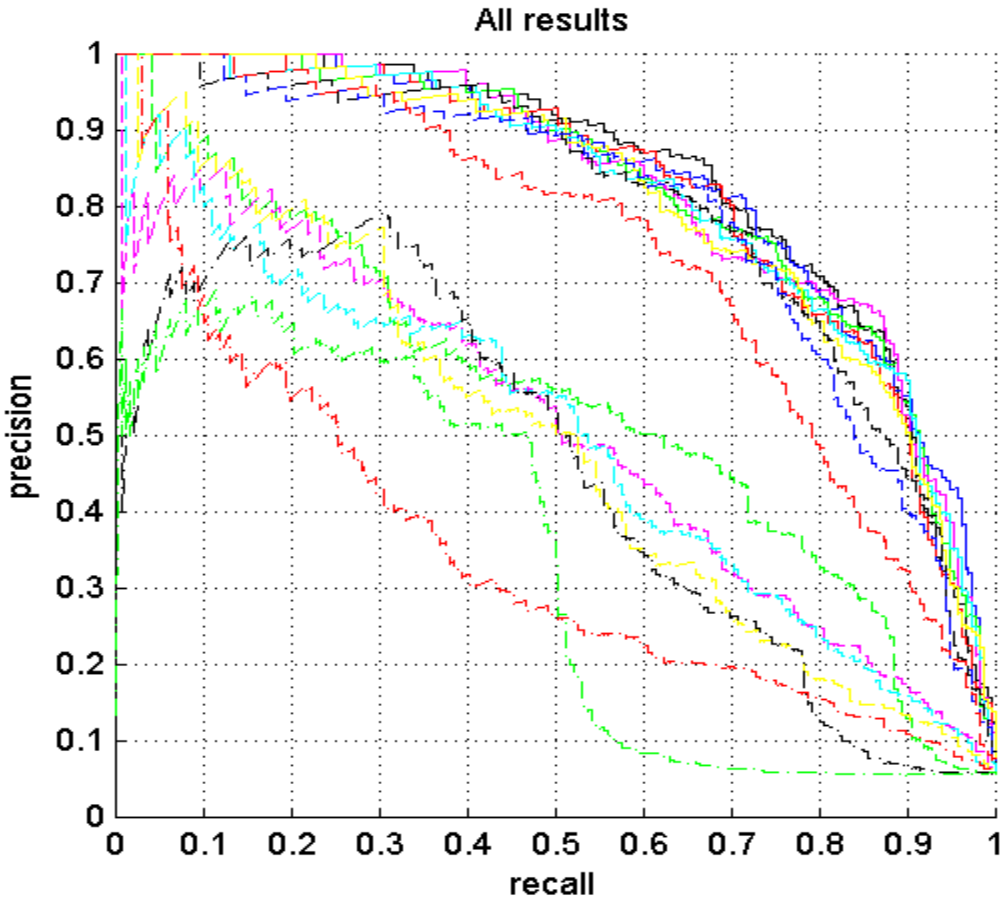
Using informative class-fragments to deal with variability

Felzenszwalb et al (Pascal 2007)





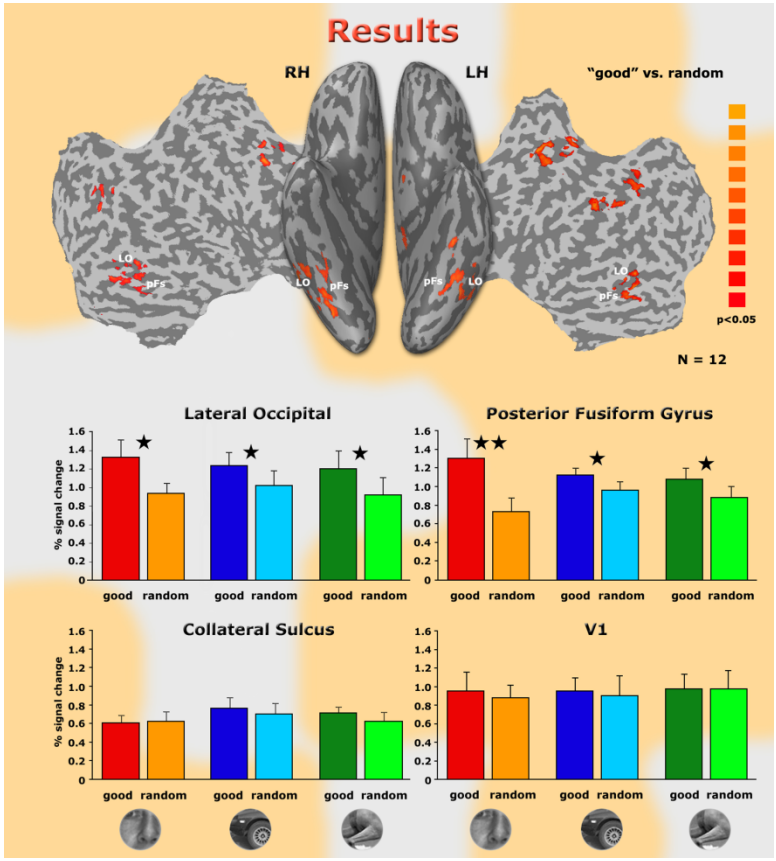
‘Pascal Challenge’ Airplanes



Looking for Class Features in the Brain: fMRI

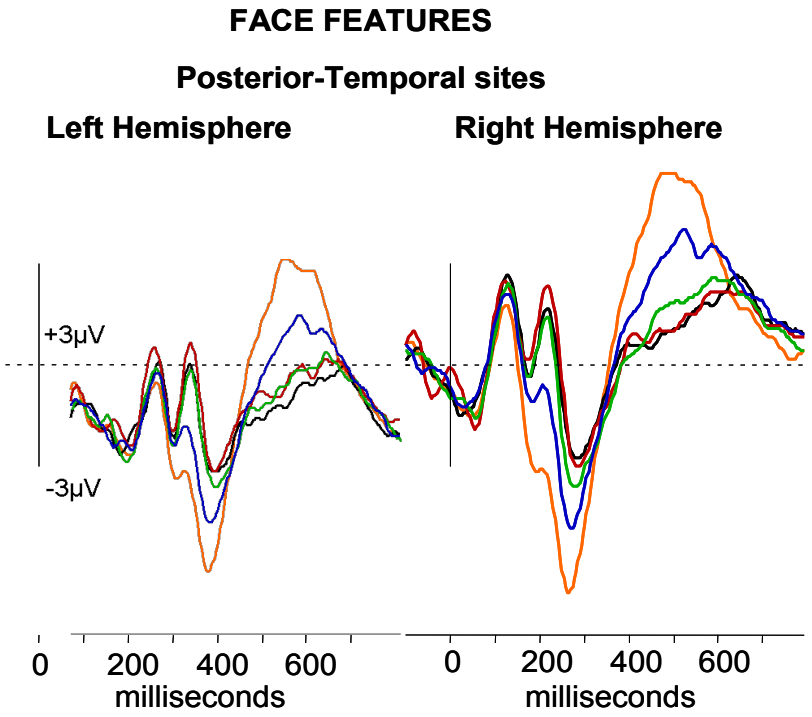
Lerner, Epshtein Ullman Malach JCON 2008

Class-fragments and Activation



Malach et al 2008

ERP

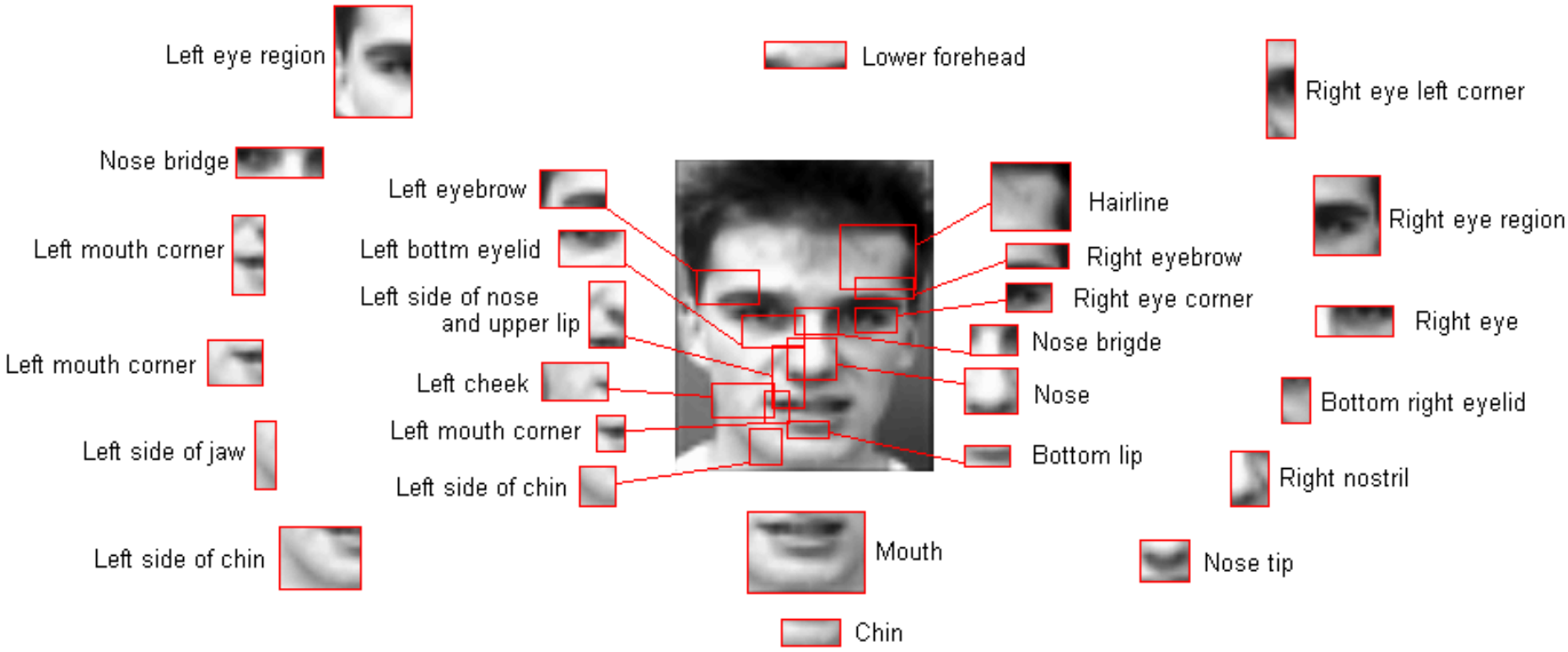


MI 1 —
MI 2 —
MI 3 —
MI 4 —
MI 5 —

Harel, Ullman,Epshtein, Bentin Vis Res 2007

‘Object interpretation’:

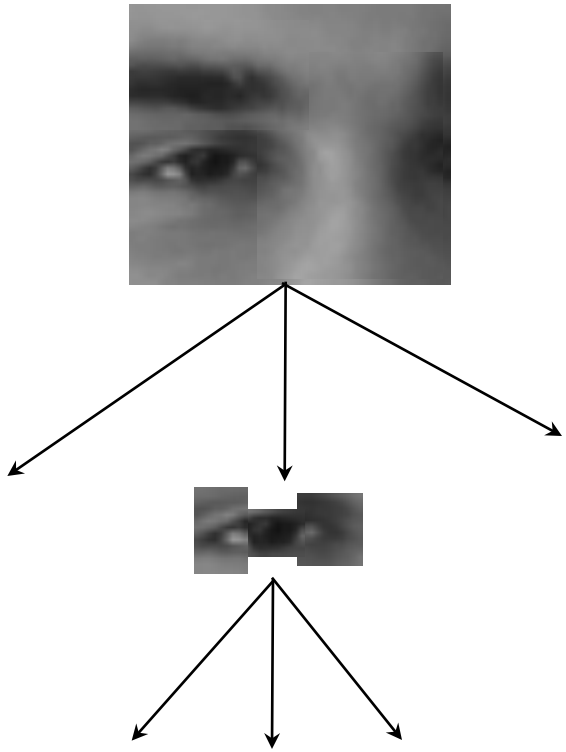
Full recognition of parts and sub-parts at many levels



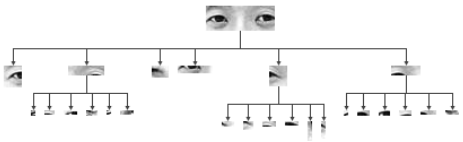
Hierarchies of sub-fragments

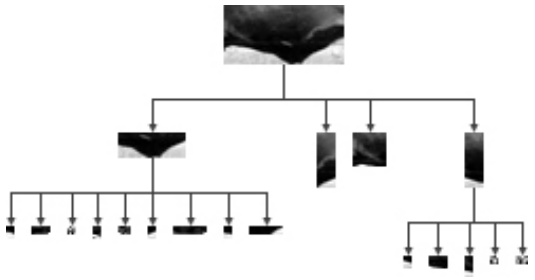
Detect the part by simpler sub-parts

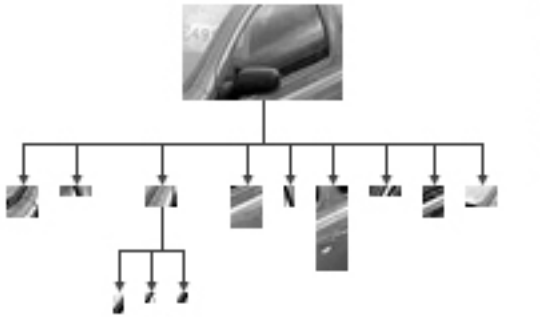
Repeat at multiple levels, to obtain a hierarchy of parts and sub-parts

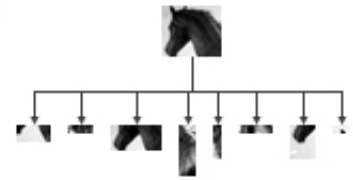


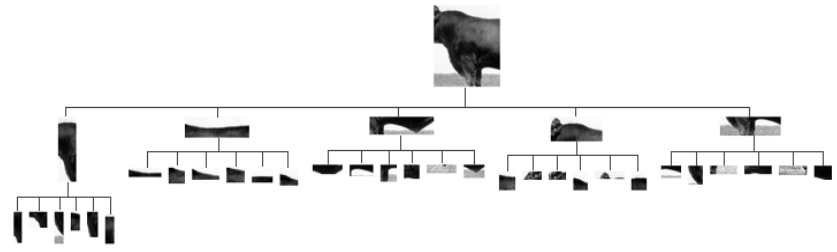
Example Hierarchies



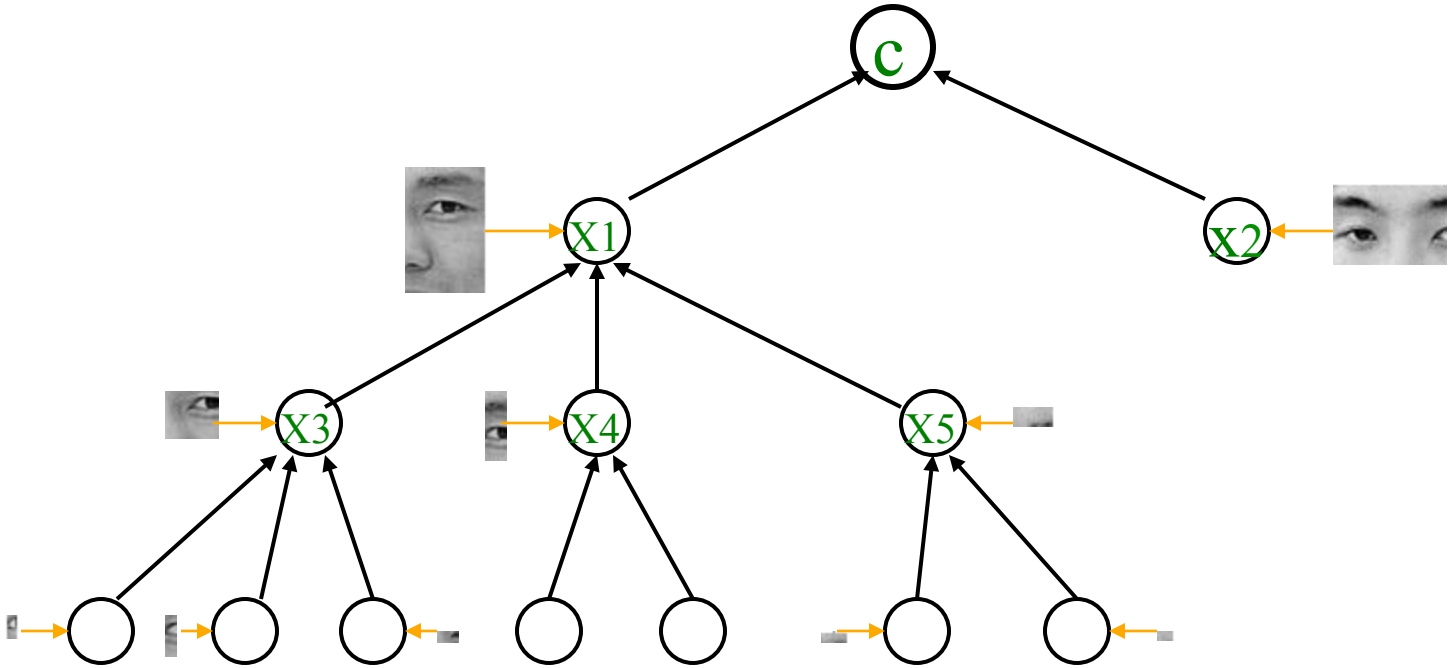








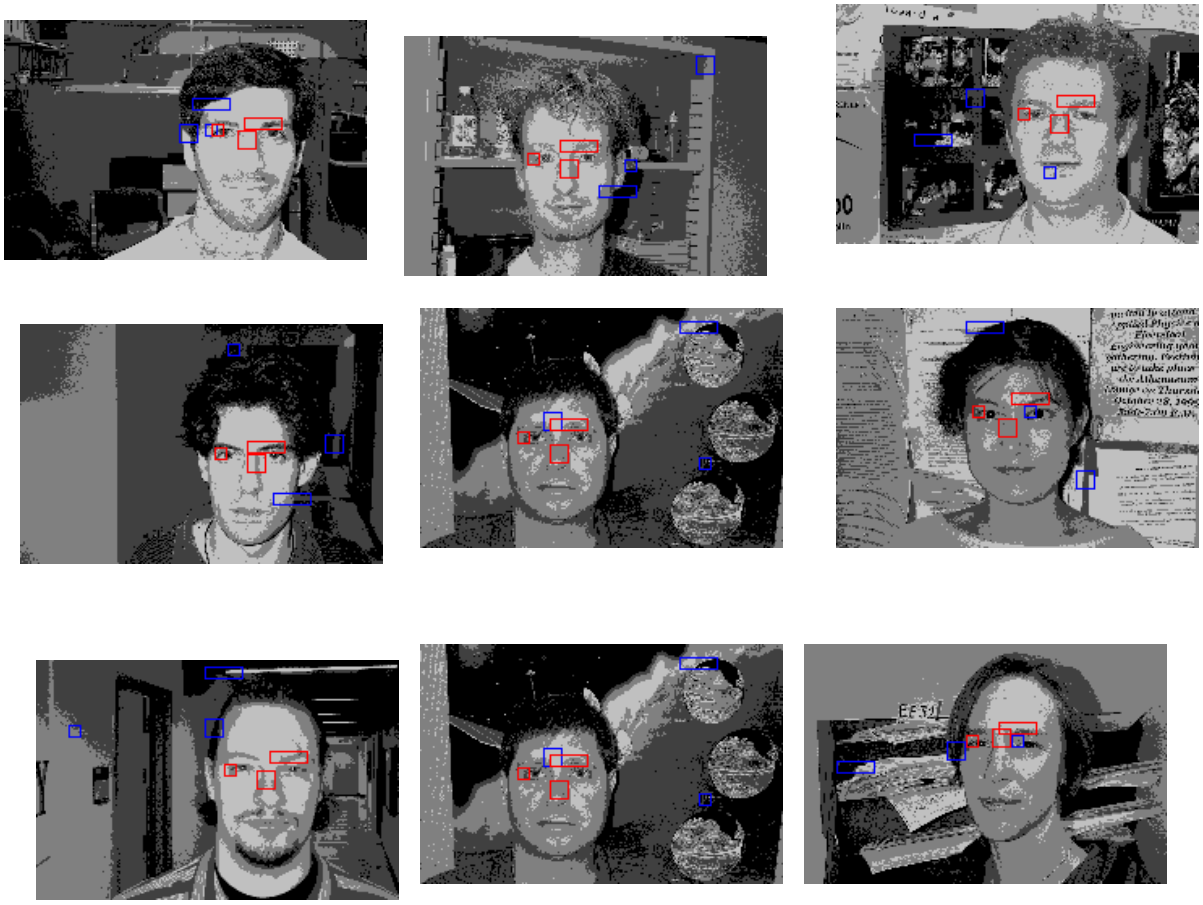
Classification and detecting all the parts
Two pass computation: BU then TD



- The full hierarchy is used for recognition
- A bottom-up then top-down pass

$$p(c, \underline{x}, \underline{F}) = p(c)\Pi p(x_i|x_i^-)\ p(F_i|x_i)$$

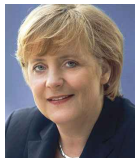
Difficult parts



Eyebrow, nose-bridge, eye-corner
All missed by the feed-forward pass

III. Individual Recognition
Same Object Under Different Views

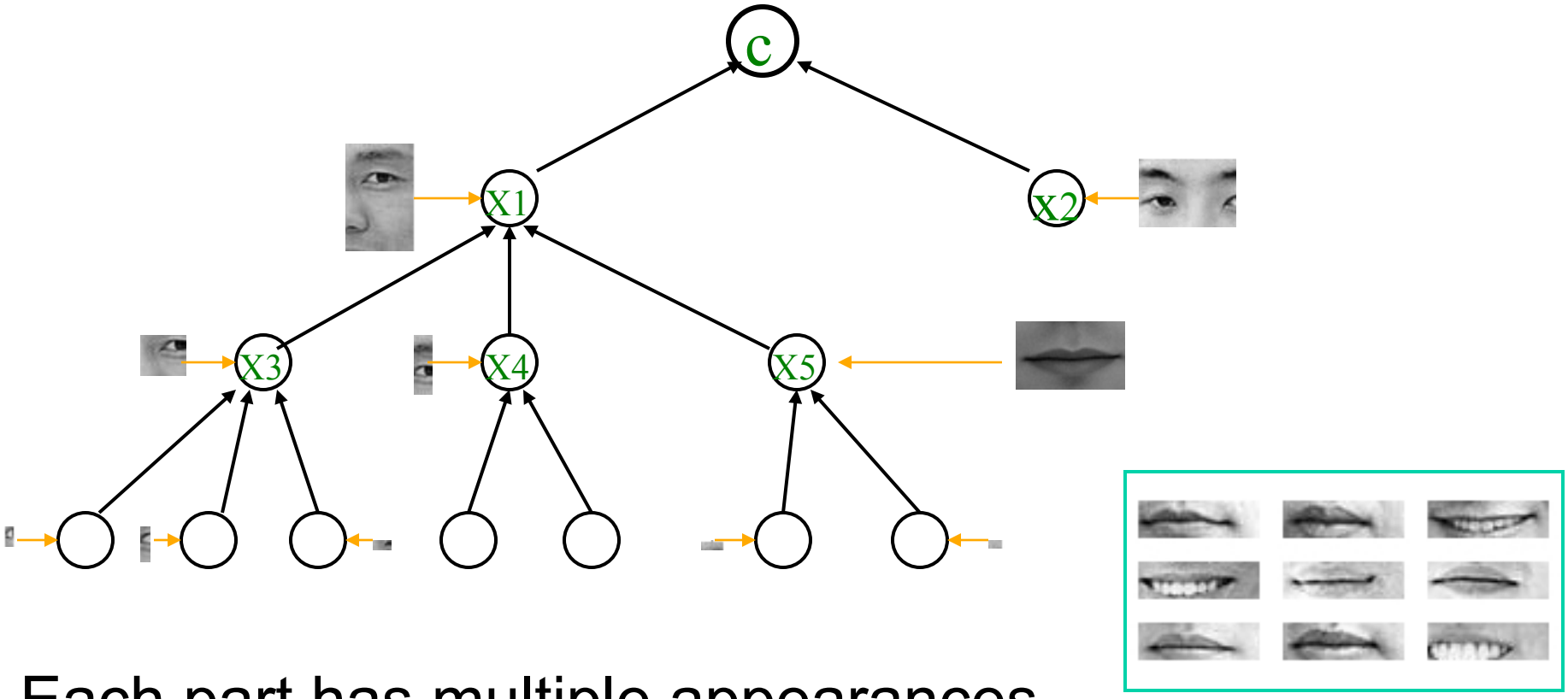









From Appearance to Semantic Features



Each part has multiple appearances

Semantic Features

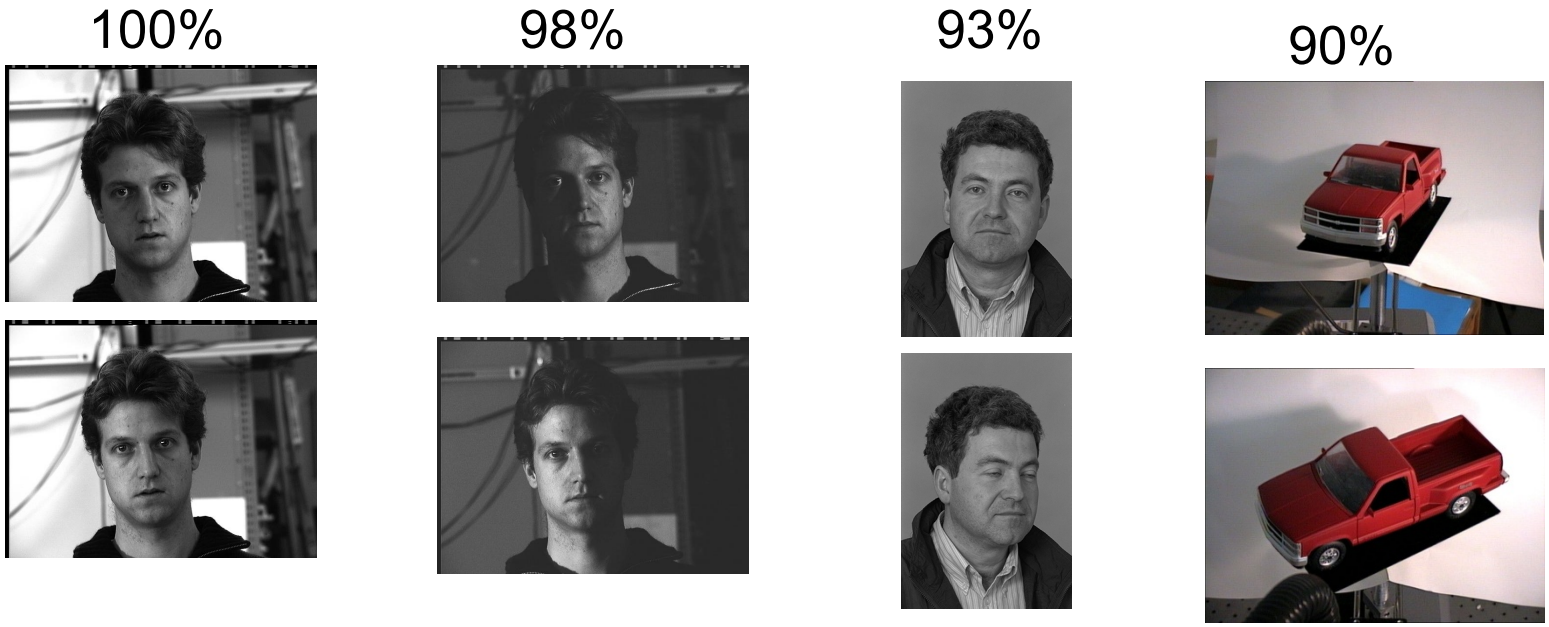
- Group together alternative appearances:

  – *Dynamic:* by motion
  – *Static:* by common context

Semantic Features: Examples

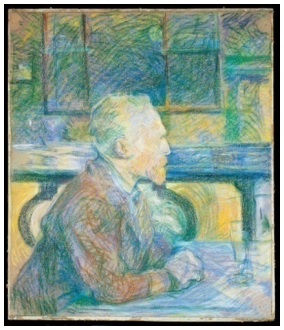
Emergence of abstract representation

Individual Recognition by Semantic Fragments

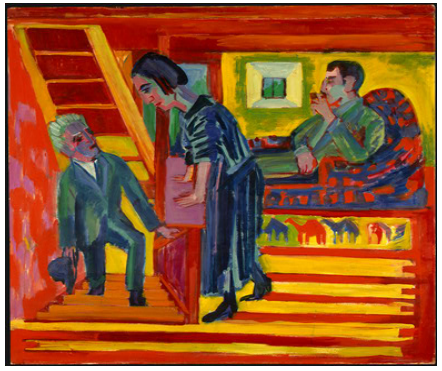


Invariance to individual objects in inherited from the learned invariance of semantic fragments

Full Interpretation, Difficult Parts: Finding Hand by Context

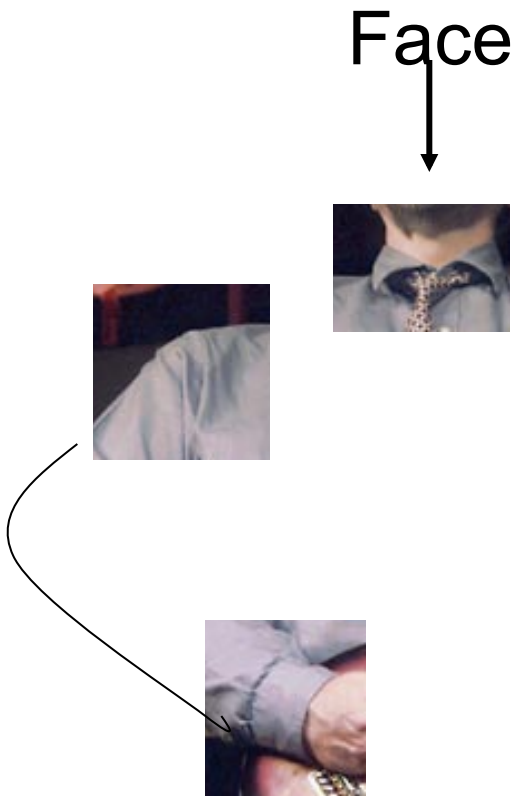


Van Gogh



Kirchner

Hand by Context



Face → Shoulder → Upper-arm → Lower-arm → Hand

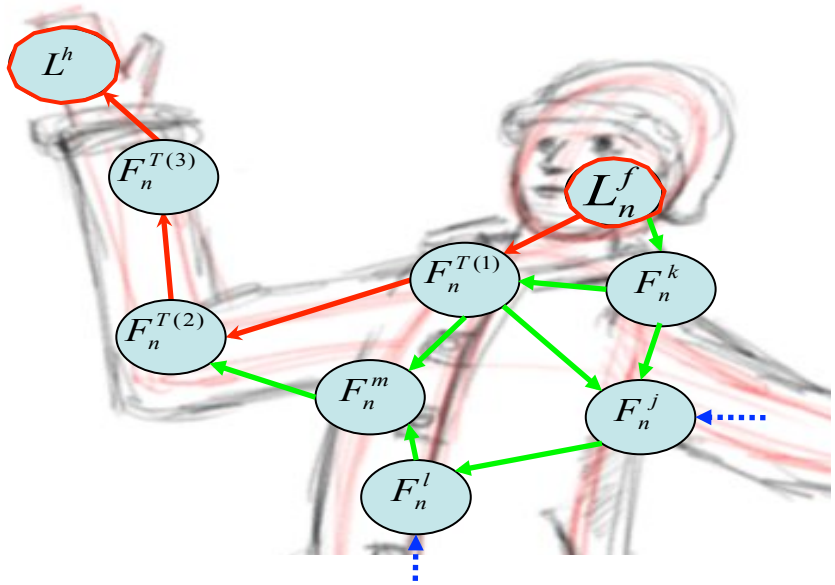
Use appearance to learn the context

The ‘Chains Model’ for learning useful chains from sources to targets

Chains model

$\underline{F} \rightarrow \underline{X}$

$p(\underline{F}, \underline{X})$



p_T(Hand, Face, Features) =

p(Face) $p(F_1|Face)p(F_2|F_1)\ldots p(Hand|F_k)$ $\Pi p(F_j)$

(On-chain) (Off-chain)

$p(Face)\ p(F_1|Face)p(F_2|F_1)\ldots p(Hand|F_k) * \quad \Pi p(F_j)$

After dividing by the product $\Pi p(F_j)$:

p(entire configuration) :

$p_T(Hand, Face, Features) \sim \prod \frac{P_c(F_i, F_{i+1})}{P(F_i) \cdot P(F_{i+1})}$ Links along the chain

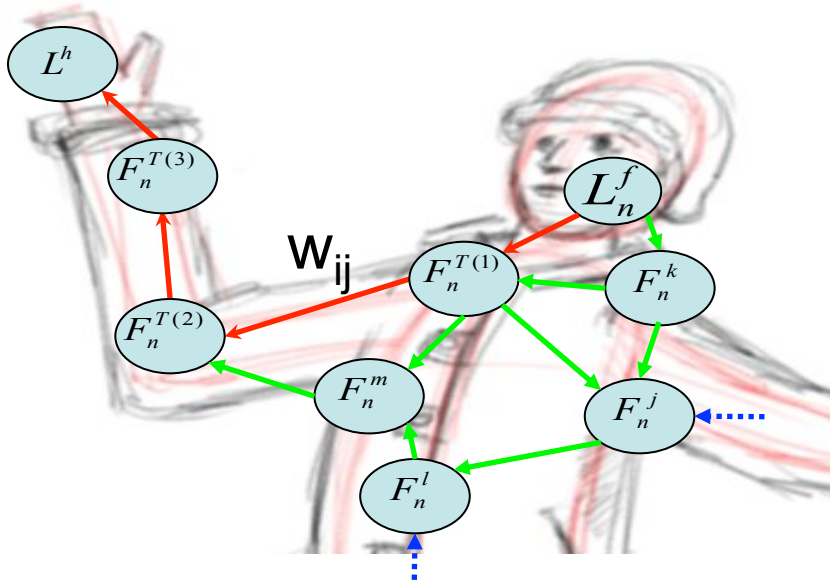
Or: Π 'Suspicious coincidence' (F_i, F_j)

Finally, finding the hand -- combine all chains:

$p(Hand \mid Face, \underline{F}_k) \sim \Sigma_T\ p_T(hand, Face, \underline{F}_k) = \Sigma_T\ \Pi_T\ w_{ij}$

The final computation:

Chains model



$$\frac{P_c(F_i, F_{i+1})}{P(F_i) \cdot P(F_{i+1})}$$

- Features are connected by a lateral weight W_{ij}
- Features that occur frequently together will have high W_{ij}
- The probability of a chain is Πw_{ij} along the chain.
- The computation finds high likelihood chains between the source and the target (face and hand).
- Propagate from known parts to new parts

The chains model inference

$$P_{CH}\left(L_H \middle| L_F, \{F_i\}\right) \propto \sum_T P_{CH}\left(T, L_H, L_F, \{F_i\}\right) = D \cdot C^* \cdot S \approx D \cdot C \cdot S$$

$$D = \left[P_{obj}\left(L_H, F_1\right), \cdots, P_{obj}\left(L_H, F_K\right)\right], C = \left(\sum_{m=0}^{M-1} A^m\right), S = \left[\frac{P_{obj}\left(L_F, F_1\right)}{P_{obj+bg}\left(F_1\right)}, \cdots, \frac{P_{obj}\left(L_F, F_K\right)}{P_{obj+bg}\left(F_K\right)}\right]^T$$

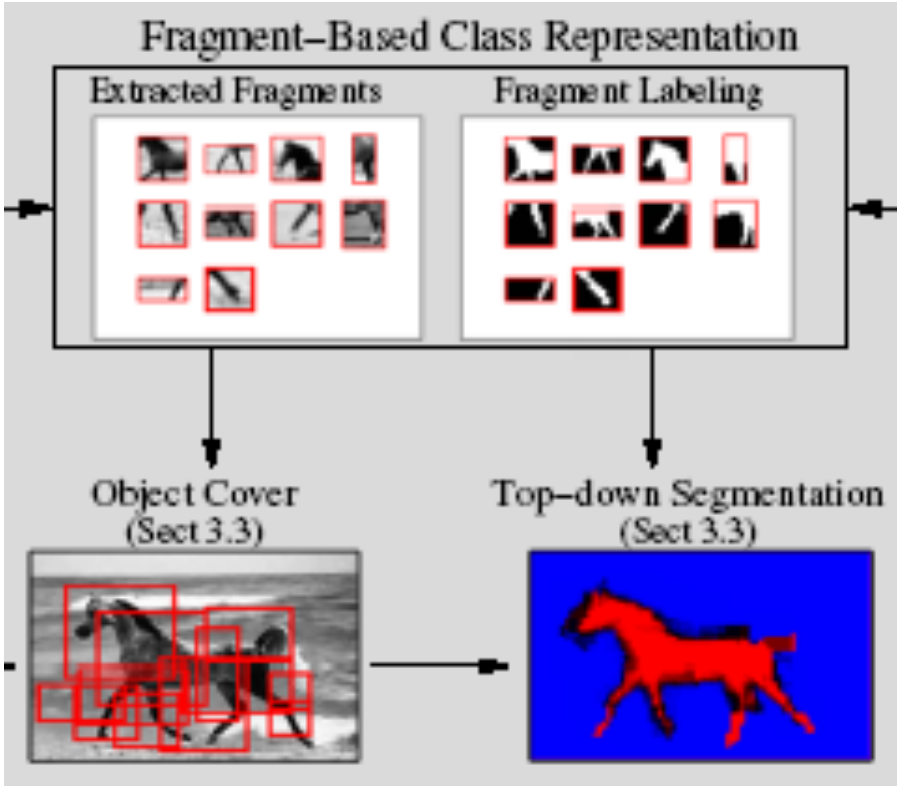
Hand

Internal

Face

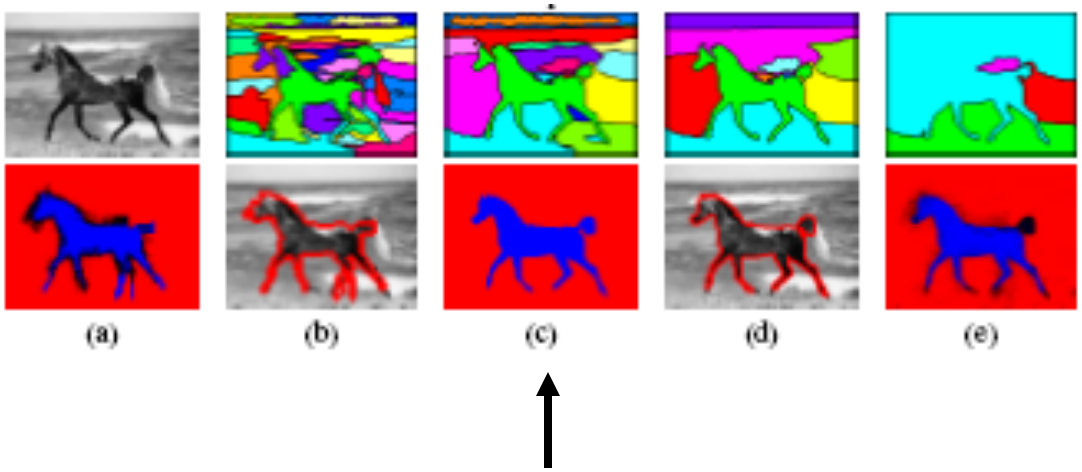
- The inference boils down to matrix multiplication approximating marginalization on chains of length <= m.
- A is the weighted adjacency matrix of the feature graph, w_{ij}.
- C^* is the marginalization over only the simple chains (NP-hard), approximated by C which is the marginalization over all chains.

Class-based Segmentation



Top-down Segmentation

Combining TD with BU



Segmentation: combining TD with BU

Borenstein, Ullman PAMI 2008

Summary

- Informative object fragments
- Class specific
- Continuously extracted from examples
- Hierarchy of informative parts
- Grouped into semantic features (motion, context)
- Categorization, identification, segmentation
- Bi-directional interpretation process BU-TD