Complexity and Compositionality

Alan Yuille (UCLA)

Dept. Statistics,

Computer Science, Psychiatry, Psychology

Neuro-inspired? Weakly.

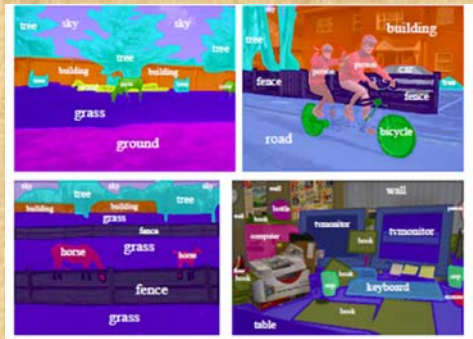
- (I) Develop vision systems that have the same visual abilities as humans (or better). Pass visual Turing tests. Develop practical computer vision systems.

- (II) Get some inspiration from known properties of the brain. Try to explain some aspects of the brain.

- Can we derive the structure of the human visual system from first principles?

Human Visual Abilities

- The Human Visual System has many more abilities than current computer vision systems. These include:
- (I) Robustness and Flexibility.
- (III) Unsupervised Learning from small numbers of examples: 1-shot learning.
- (IV) Most Importantly the ability to perform an enormous number of **visual tasks**.



Complexity – Fundamental Problem of Vision

The fundamental problem of vision is how to deal with the astronomical complexity of images, scenes, and visual tasks.

For example, considering the enormous input space of images and output space of objects, how can a human observer obtain a coarse interpretation of an image within less than 150 msec?

How can the observer, given more time, be able to parse the image into its components (objects, object parts, and scene structures) and reason about their relationships and actions?

Research Program

- (I) Theoretical Framework. Visual Architectures which are capable of performing all these tasks.
- Key Problem – dealing with complexity.
- (II) Applying this framework to natural images. Developing working computer vision systems. Compete on challenging datasets.
- Key Problem – scaling from "toy world" to "real world".
- (III) Relate the framework to behavioral, electrophysiological, imaging, and other studies.
- Key Problem – how much do we really know about how the brain does visual processing? TS Lee. D.K. Kersten.
- (IV) Hardware: Vision Cortex in Silicon. NSF Expedition in Computing.

  PI: Vijaykrishnan Narayanan.

This Talk.

- Concentrate on (I) the Theoretical Framework.

  Basic Principle: compositionality. Build models recursively from re-usable components.

- (II) Illustrative examples showing that these ideas can work on complex images and tasks.
- (III) Speculations about neuroscience. Beyond McCullough and Pitts neurons? Dynamic binding.
- (IV) Hardware Implementations? Prof. V. Narayanan.
- Related Work – compositional models (S. Geman), probabilistic grammars (S.C. Zhu and D.B. Mumford). Analysis by synthesis (T.S. Lee and D.B. Mumford).

Probabilistic Grammar Markov Models (PGMM)

- PGMMs (L. Zhu et al. 2006, 2009).

- Visual Tasks:
- Unsupervised learning of object class models.
- Learning in presence of clutter.
- Learning models of different object classes without supervision.
- Learning from small numbers of examples (towards one-shot learning).

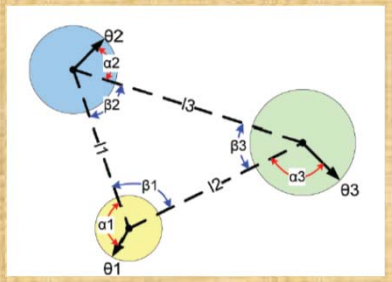
PGMMs: Interest Points and Attributed Features.

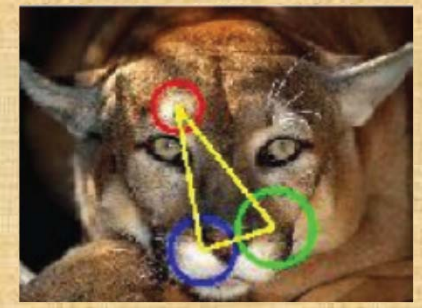
- Basic Ingredients.
- Interest Points (IPs).
- IPs described by position X , orientation Theta, attributed features (AF).
- Learn dictionary of IPs by clustering on the AFs.
- AF-1, AF-2,….,AF-N

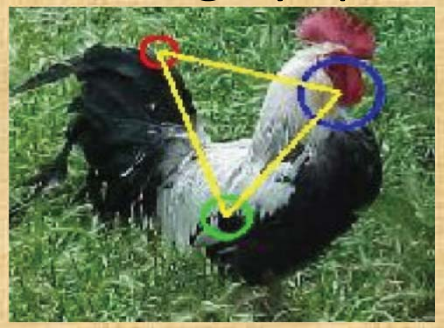
- Technical Details: choices of IP detectors and AF descriptors.

PGMMs: Spatial Relations and Triplets

- Objects are compositions of IP's.
- Basic Building block – triplets of IPs grouped based on spatial relations. {(AF-1, AF-2, AF-3, spatial relations),….}.
- Dictionary of IP-Triplets – obtained by clustering by spatial relations.



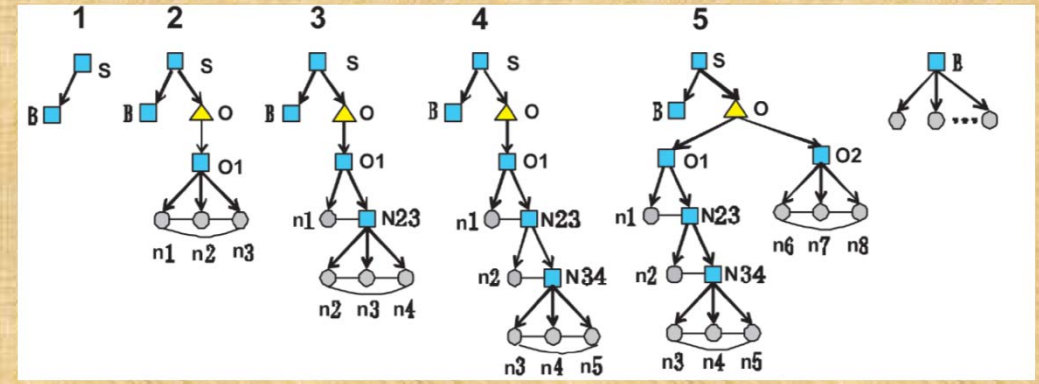




- Note: much more mathematics in L. Zhu et al. (2009). Here I give the executive-level summary.

Unsupervised Learning by Structure Induction

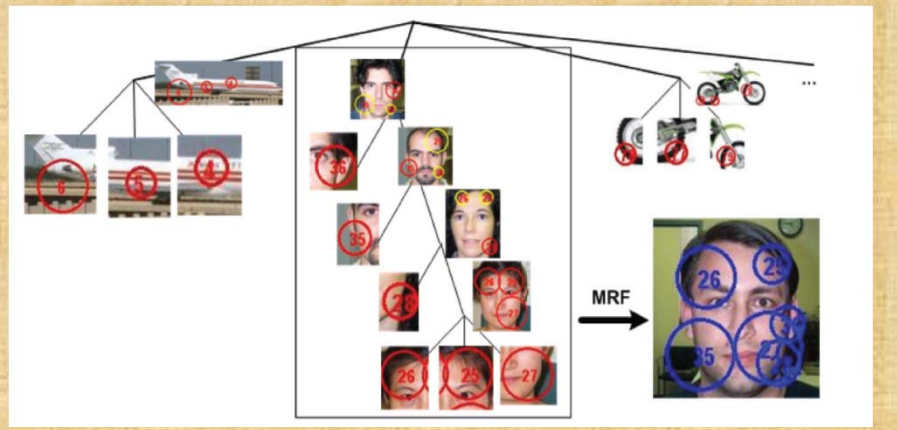
- Start with default background model (independent IP's).
- Search for frequently occurring IP-Triplets.
- Grow object models by: (i) adding OR nodes, (ii) add triplets.

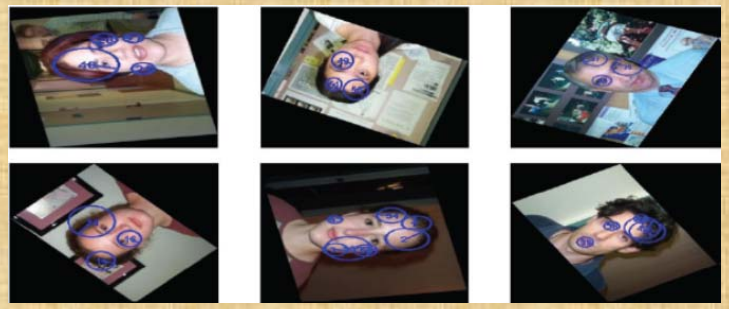


- Intuition: MDL. Better encoding of images. Constructing representations dynamically.

PGMMs: Results for Learning

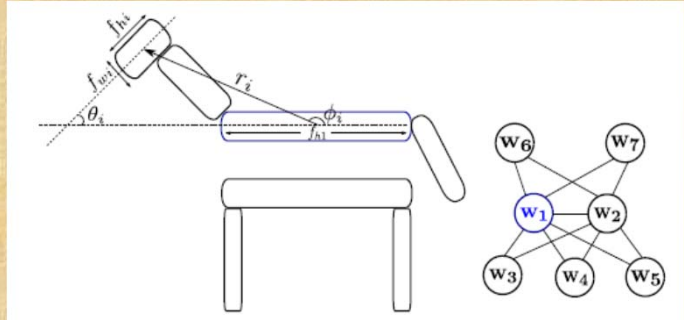
- Unsupervised learning of object classes in noisy background (cocktail party problem).
- Unsupervised learning of mixtures of object classes – without knowing how many classes exist.
- Learning models allowing for spatial transformations.
- Learning from small numbers of examples (only partially explored).

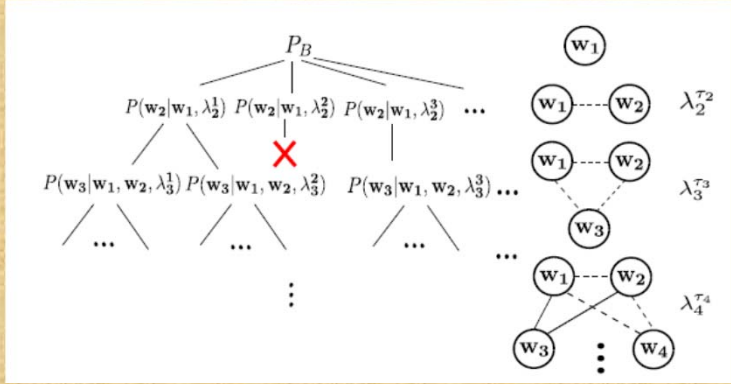


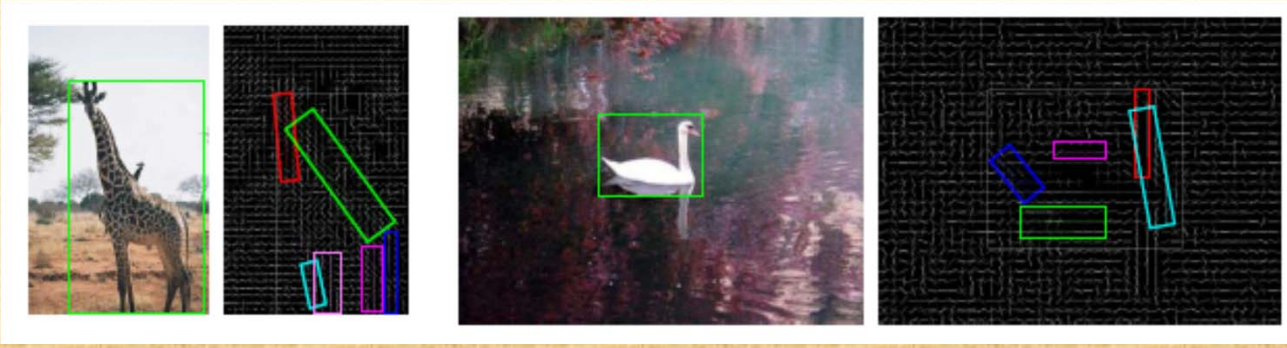


Mottaghi and Yuille 2011.

- Other examples – different features (HOG Bundles) – databases.







PGMMs and Neurons?

- PGMMs include spatial relations and clustering. This differs from the classical perceptron – artificial neural network model.

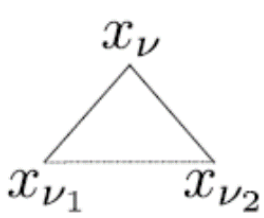
- Relate to classical neural ideas: (i) Binding problem/mechanisms (Von der Malsburg, Abeles), (ii) shifter circuits (Olshausen and Anderson)

- Closest neural network model – L. Valiant "Circuits of the Mind". Valiant shows that clustering, memorization, and binding can be performed by small variants of perceptron models.
- Can PGMMs be implemented by neurons? "Alan, neurons are so complex that they can probably do all the things you want them to". Bartlett Mel (USC).

Hierarchies and Compositional Models

- .Addressing the problem of complexity.

- Recursively build hierarchies of dictionaries – triplets of triplets of triplets…
- Starting point – edgelets. Dictionaries building by recursive clustering.
- Basic building block – part-subpart composition.



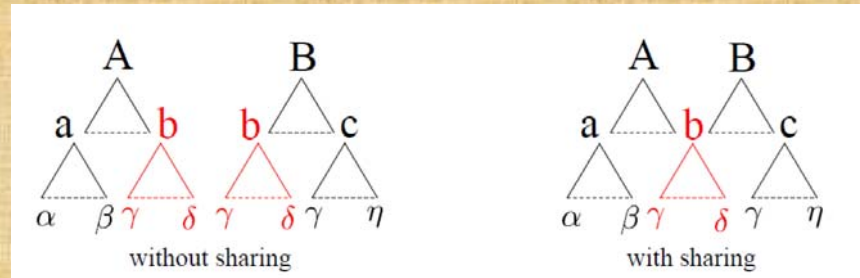
$\tau_{\nu_1} = H$ $\quad T = (H, V, \lambda_T)$ $\quad \lambda_T = (\mu_T, \sigma_T)$

$\tau_{\nu_2} = V$ $\quad L = (H, V, \lambda_L)$ $\quad \lambda_L = (\mu_L, \sigma_L)$

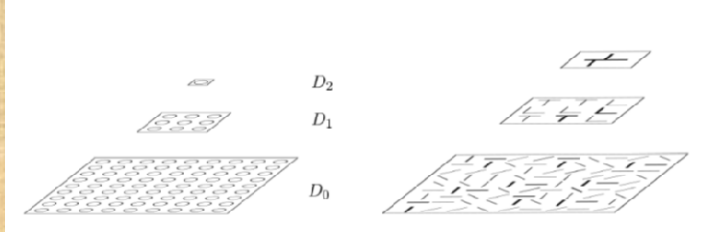
Key Concept: Compositionality

- Compositionality. *The probability models for reasoning tasks are built by combining elementary models by composition.*
- Allow the system to answer many different questions based on the same underlying knowledge structure. No need to learn every task separately (e.g., detect, locate, recognize, parse objects by the same model).

Advantages of Hierarchies



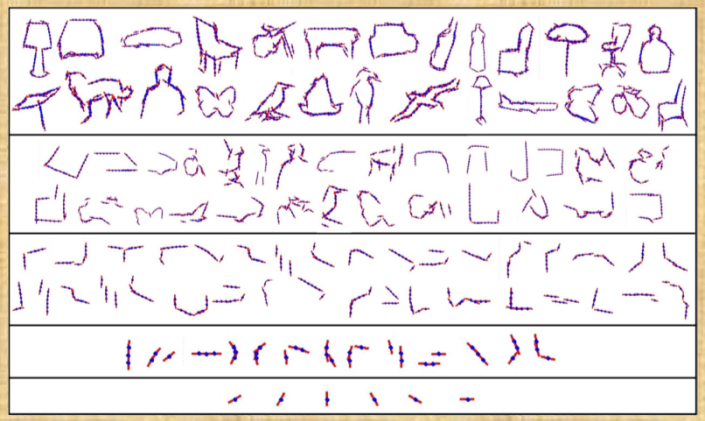
- Key Concepts.
- Part-sharing – efficiency of representation, inference, and learning.
- Executive-level summary – part descriptions are less detailed than subpart description. Hierarchical representation. (Boss knows there is an ‘elephant in the room”, low-level executive knows where its foot is).
- Corporation Metaphor.

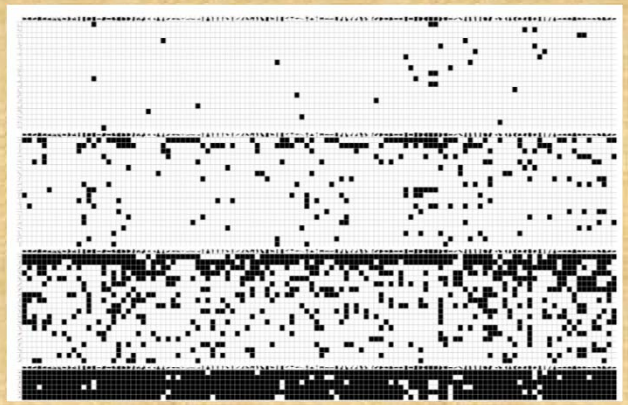


- Objects are represented in a hierarchical distributed manner.

Hierarchies: Unsupervised Learning

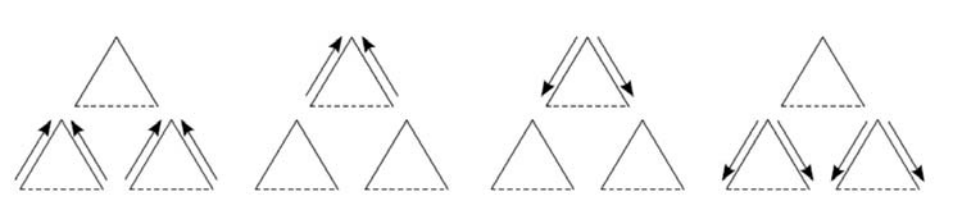
- Unsupervised learning by recursive clustering (L. Zhu et al. 2008, 2010).





- Breadth first search over space of generative models of images.
- Constructing hierarchical dictionaries. Dictionaries can be re-used to learn novel objects.

Hierarchical Architecture



- .
- Inference Algorithm.
- Bottom-up Pass – propagates hypotheses for low-level parts to form hypotheses for high-level parts.
- Low-level parts – local context – ambiguous. Many hypotheses needed.
- Higher-level parts – more structured, less ambiguous.
- Top-down resolves ambiguities in bottom-up stage. Removes low-level hypotheses which are not validated by high-level hypotheses.
- Formally – this can be formulated as dynamic programming (e.g., like the inside-outside algorithm for stochastic context free grammars in Natural Language Processing).

Hierarchies: Complexity Analysis.

- Serial Formulation (e.g., as in L. Zhu et al. 2008, 2010).
- Parallel Formulation (Yuille and Mottaghi 2013, 2015):
- Make copies of nodes in the graph. “Convolutional Compositional Models”.
- Analysis: inference complexity (serial formulation) and memory complexity (parallel formulation).
- Exponential gains – depending on how many parts can be shared, and gains from “executive summary”.

Compositional Complexity

- Analysis in Yuille and Mottaghi (2013, 2015).

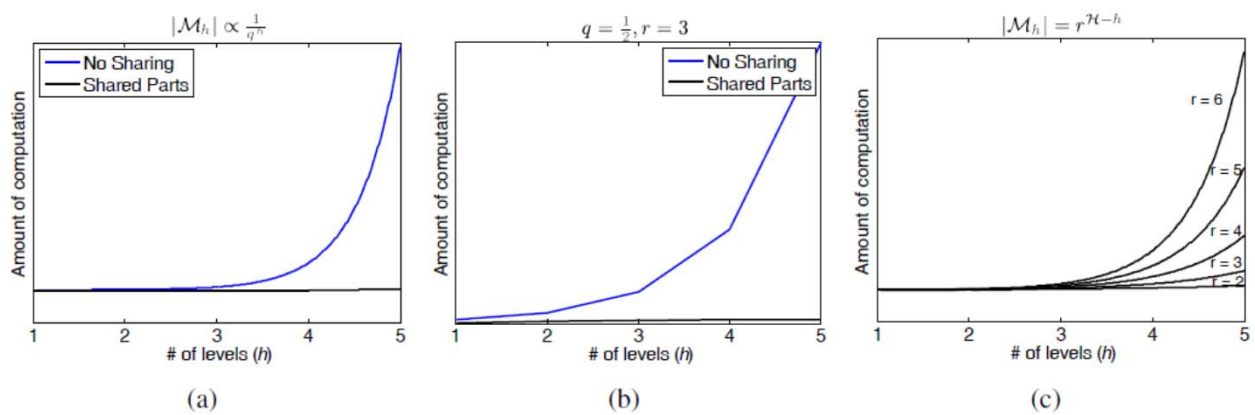
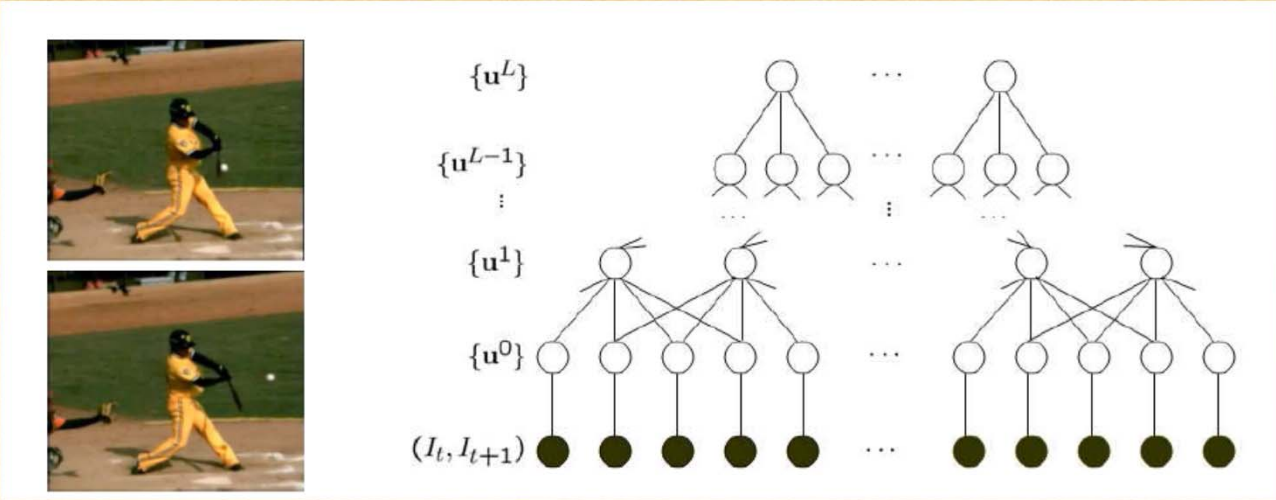
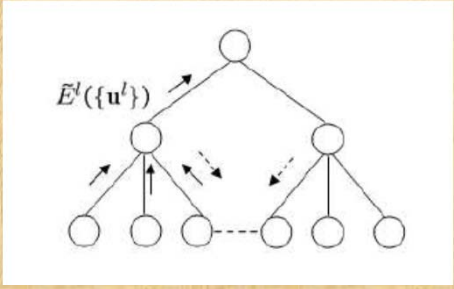


Figure 12: The curves are plotted as a function of h. Left panel: The first plot is the case where $M_h = a/(q^h)$. So we have a constant cost for the computations, when we have shared parts. Center panel: This plot is based on the experiment of Zhu et al. (2010). Right panel: The third plot is the case where M_h decreases exponentially. The amount of computation is the same for the shared and non-shared cases.

Hierarchical Algorithms for Motion Inference

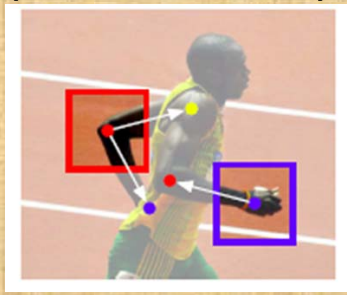
- Motion Estimation (S. Wu el al. NIPS 2010).
- Horizontal Inference – or Feedforward and Feedback?
- Psychophysics.

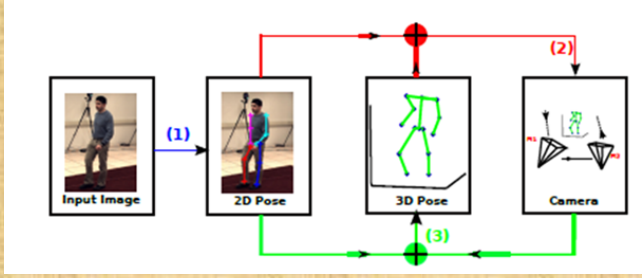




Applications – Objects, Parts, Segmentation.

- Human Pose Detection. Human 3D structure estimation. Semantic Parsing (Caveat – uses range of techniques including DCNNs).
- X. Chen and A. Yuille (NIPS 2014), C. Wang et al. (CVPR 2014), X. Lian at el. (BMVC 2014), X. Chen et al. (CVPR 2014). J. Mao et al. (NIPS 2013)





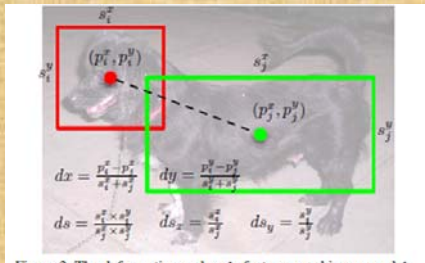


Figure 3. The deformation and scale features used in our model.

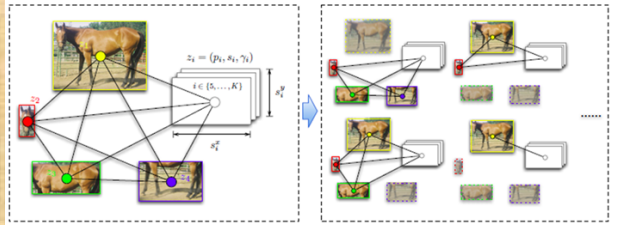
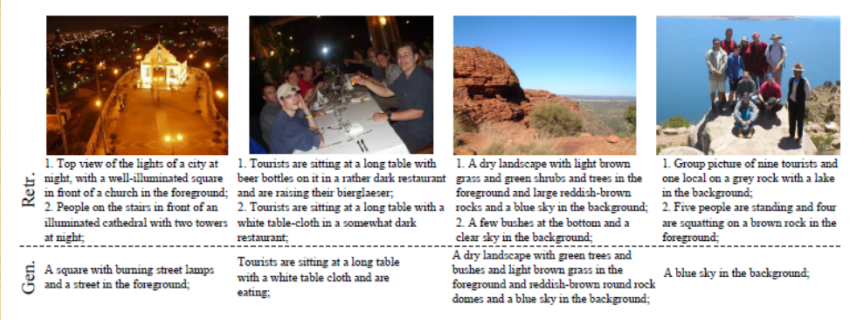


Figure 2. (a) Our graphical model where the nodes represent the holistic object and its body parts. Their state variables are position, scale, and switch variable (γ). The holistic object is shown in yellow and some example body parts are shown in red, green, and blue. The rest of the body parts are shown with white rectangles. (b) The switch variables decouple nodes from the graph, depending on which parts are detected, and enable the model to deal with different detectability patterns. Boxes with dashed border are those decoupled from the graph.



Summary

- (I) Theoretical Issues – Key Problem of Vision. Complexity.
- Compositionality, Part-Sharing, Spatial Relations, Executive Summary.
- Visual Architecture – Representation, Inference, Learning.
- Ability to perform large range of visual tasks.
- (II) Natural Images and Dataset Performance. Task – scale up the algorithms to the complexity of natural images.
- (III) Relations to Neuroscience. Consistent with neural models –dynamic binding (von der Malsburg, Abeles, Hummel), shifter circuits (Olshausen and Field), circuits of the mind (Valiant), complex models of neurons (Mel).
- (IV) Relations to Hardware. In progress. Vision Cortex in Silicon.

References:

- L. Zhu, Y. Chen, and A.L. Yuille. Unsupervised Learning of a Probabilistic Grammar for Object Detection and Parsing. NIPS. December 2006.
- L. Zhu, C. Lin, H. Huang, Y. Chen, A.L. Yuille. Unsupervised Structure Learning: Hierarchical Recursive Composition, Suspicious Coincidence and Competitive Exclusion. ECCV. 2008.
- L. Zhu, Y. Chen, A.L. Yuille. Unsupervised Learning of Probabilistic Grammar-Markov Models for Object Categories. PAMI. 2009
- L. Zhu, Y. Chen, A. Torrable, W. Freeman, A.L. Yuille. Part and Appearance Sharing: Recursive Compositional Models for Multi-View Multi-Object Detection. CVPR. June. 2010.
- S. Wu, X. He, H. Lu, A.L. Yuille. A Unified model of short-range and long-range motion perception. NIPS. 2010.
- A.L. Yuille and R. Mottaghi. Complexity of Representation and Inference in Compositional Models with Part SharingICML. 2013.
- A.L. Yuille and R. Mottaghi. Compositionality and Complexity. JMLR. 2015.

Local Ambiguity: Psychophysics in the Wild.

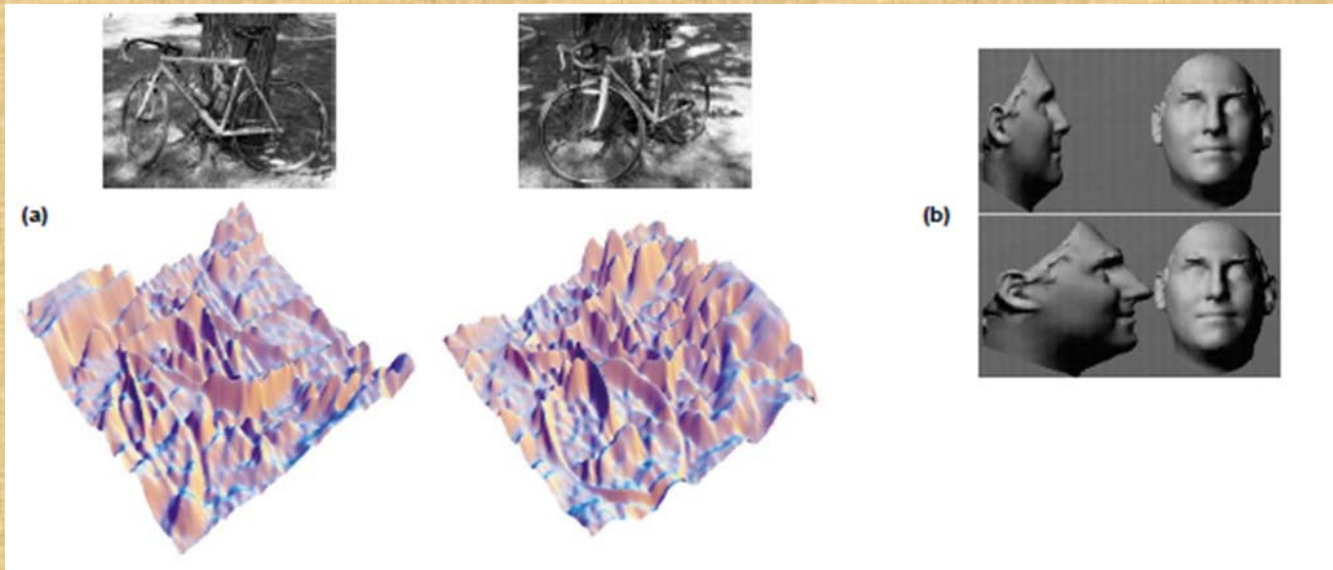
Roozbeh Mottaghi et al. 2013, 2014.

Image regions are locally very ambiguous.

Airplane
Car
Boat
Sign
Building

Why is Vision Hard?

- Look at the raw input displayed as a set of numbers which plot the intensity as a function of position (bottom left).
- The images are very complex. They are of the same bike and tree. But the raw intensities are very different.



Vision is unconscious inference (Helmholtz)

We interpret images by reasoning about the factors that caused them.

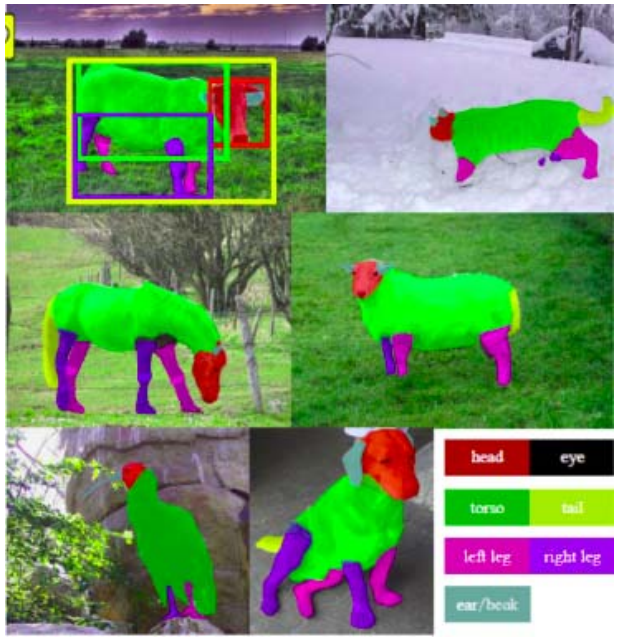
Sometimes we make mistakes due to accidental alignments:
Flying carpets and Levitation.

Theoretical Results

- Studies of Compositional Models (Yuille and Mottaghi 2013).
- Show that you can represent and rapidly access an exponential (roughly) number of objects and access them rapidly. Key Ideas: (i) part sharing, (ii) executive summary.
- Models which are generative – enabling top down analysis by synthesis, prediction, attention – but can also enable rapid feedforward inference.

Object and Parts: detection and segmentation

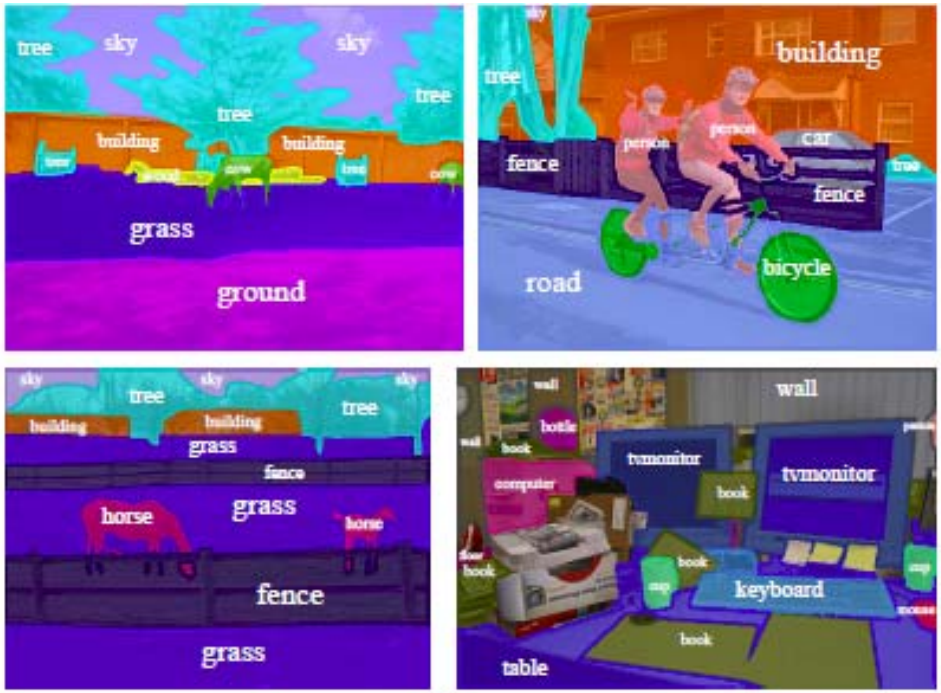
- Pascal Images 20,000.
- Detailed labeling of objects and object parts.
- Silhouettes of 20 objects.
- Parts/subparts of 18 objects.
- *Status: Released at CVPR 2014.*



Datasets: Image Labeling.

- Pascal Images 20,000.
- Pixels labeled as 1 of 35,60,600
- Categories
- Objects + background regions.

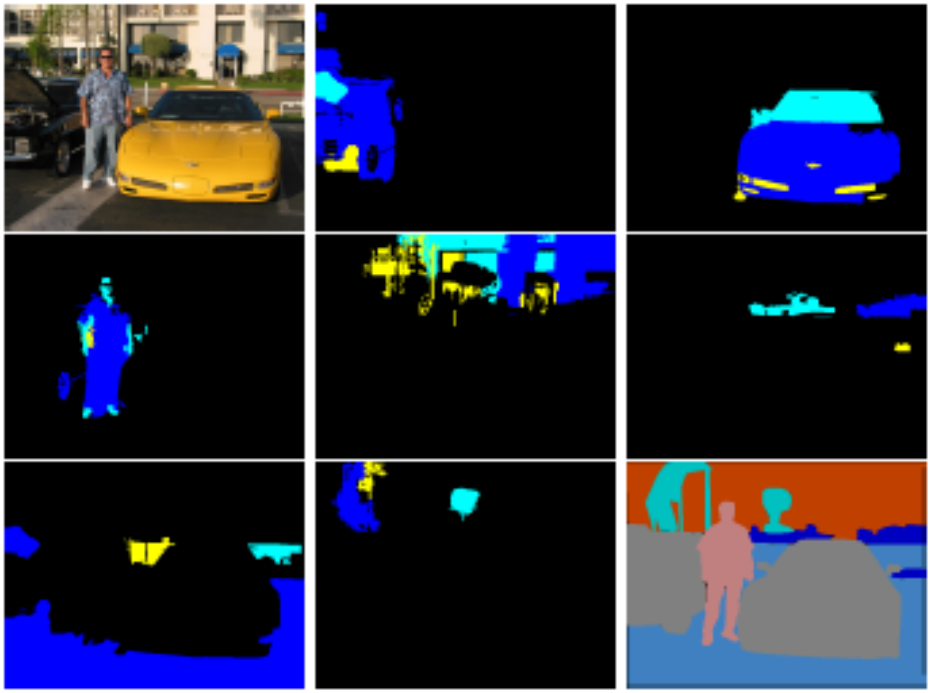
- *Status: released*
- *at CVPR 2014.*



.

Candidate Regions for Objects and Object Parts

- Examples from Bonev and Yuille (ECCV 2014).



The spatial relations.

- Spatial relations between parts.

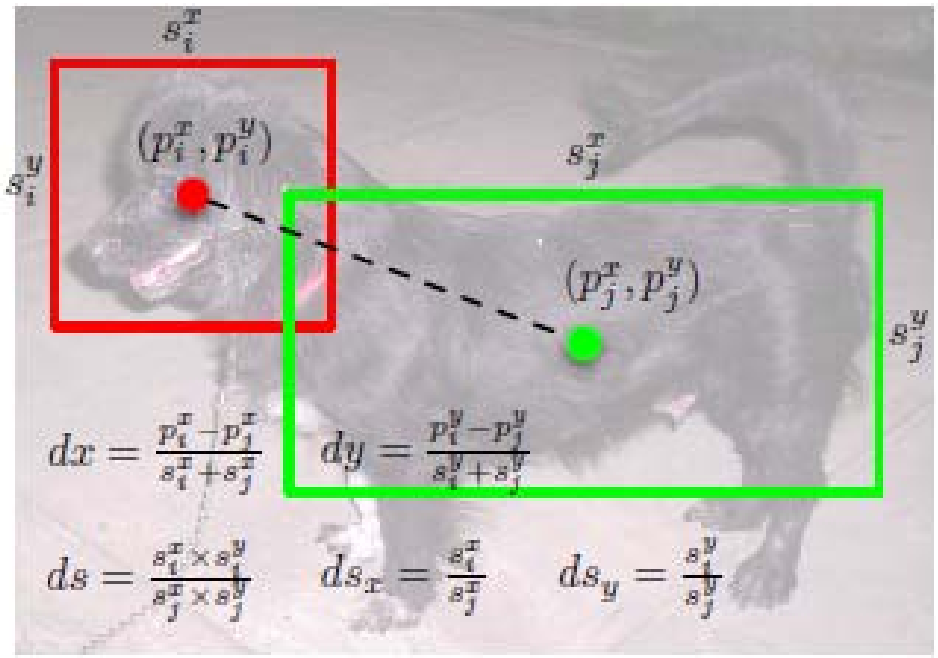
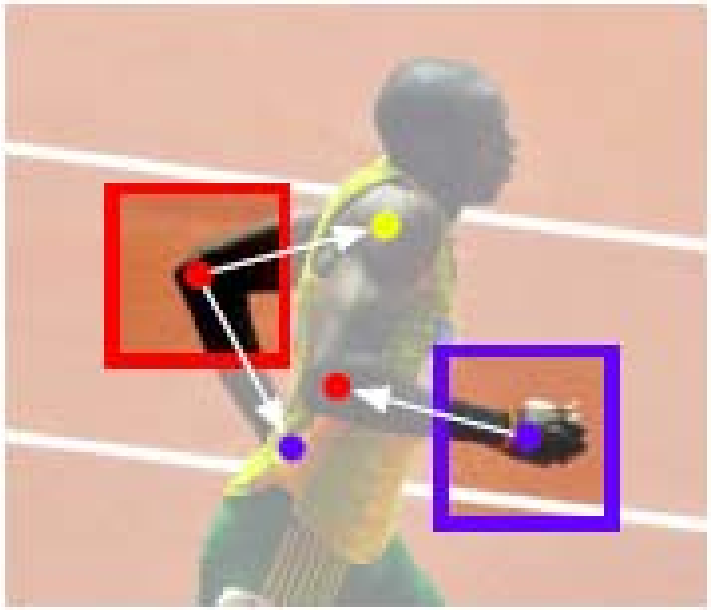


Figure 3. The deformation and scale features used in our model.

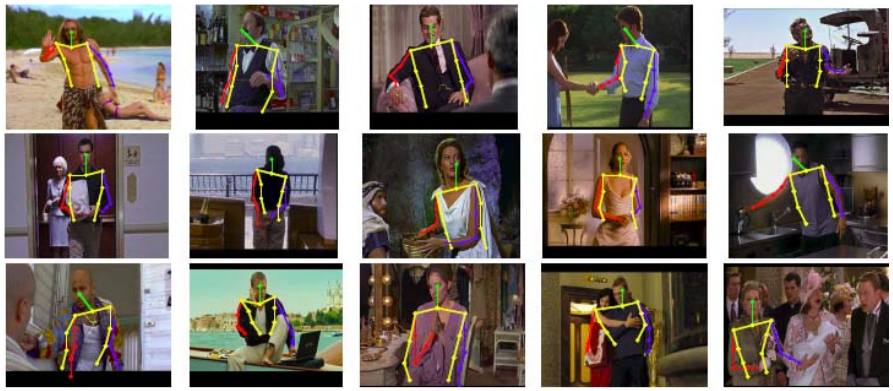
Joints and Joint Poses

- The local image information at joints can reliably predict the presence/absence oi a joint AND the relative positions of neighboring joints.



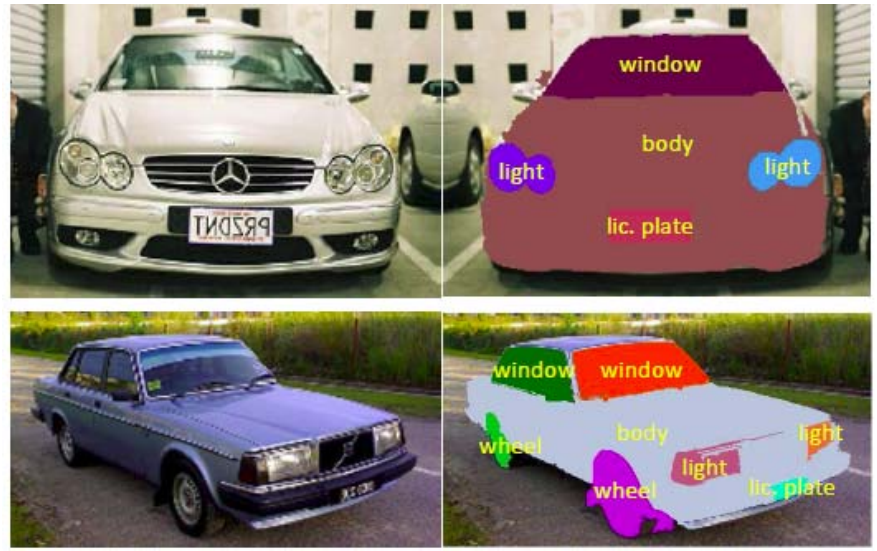
Graphical Model: Compositions

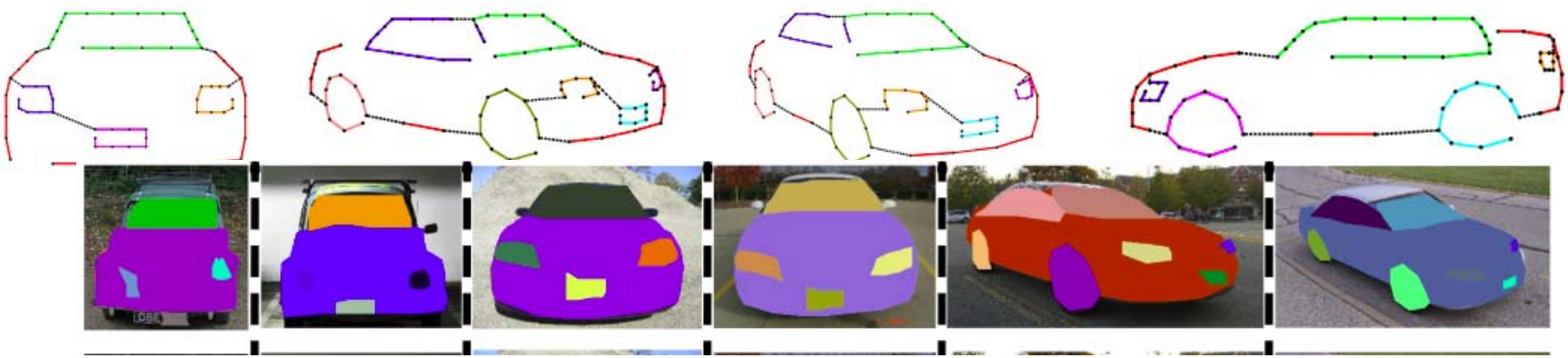
- Search for compositions of joints and joint poses which satisfy the prior spatial relations between joints (as before).



Project 3. Extras.

- Mixture Models
- Results – body, wheels, windows,
- License plates.





Project 3: Extras

- Horses,…

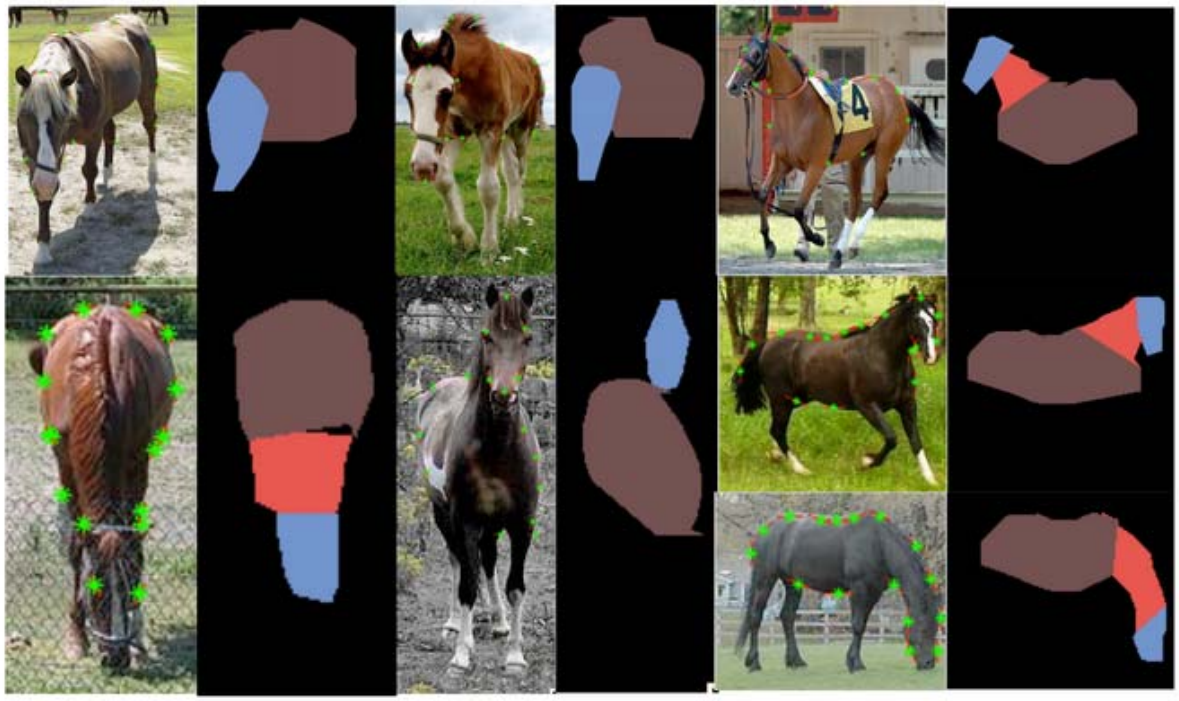
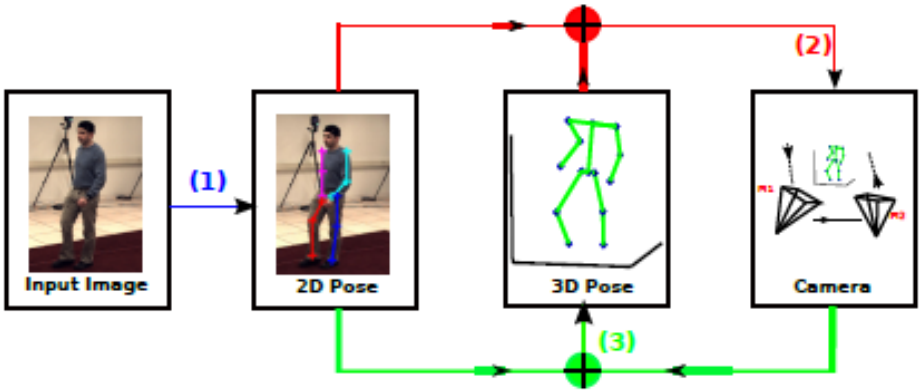


Figure 3: Seven part segmentation examples. For each example, the left is the original image superimposed by the detected leaf node location (green star represents location and red line represents orientation). The right is part segmentation result. Best viewed in color.

Project 4: 3D object models -- humans.

- Chunyu Wang et al. (CVPR 2014).



(1) Estimate the 2D joint locations and initialize a 3D pose.

(2) Estimate camera parameters and 3D pose