Grammar Induction in Vision

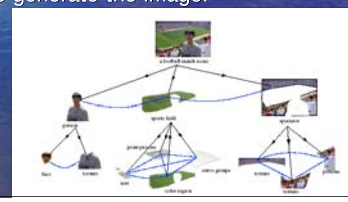
Alan Yuille (UCLA).

Research program for unsupervised learning of probability grammars for objects.

L. Zhu et al. NIPS 2006.
Kokkinos & Yuille, ICCV 2007.

Parse an Image by decomposing it into its constituent visual patterns.

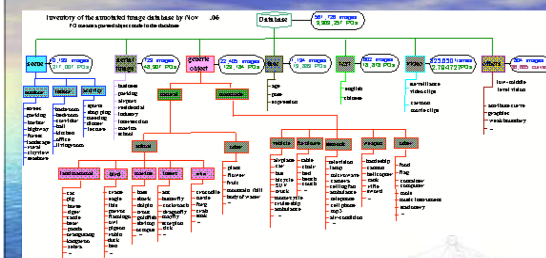
- (I) Discriminative models give proposals:
- (II) The proposals are validated/rejected by top-down generative models which compete and cooperate to generate the image.



Where do the Grammars come from?

- Want to learn grammars for Computer Vision (CV) applications.
- How do babies learn?

Lotus Hill Database: Hand Parsed.



Learning Object Models in Vision

- Can we automatically learn object models for vision?
- Input: examples of the objects in cluttered background (e.g. Caltech 101, Fergus et al).
- Strategy: *too hard to directly learn generative models for the full object appearance – try learning generative models for image features, and gradually increase the complexity of the image features.*

PGMM 1 (NIPS 06)

- Goal: learn a probabilistic grammatical markov model (PGMM) for attributed feature points.
- Dataset: Caltech 101. Each image contains a known object with unknown (random) background.

Artificially vary the pose = position, scale, orientation of object.

- Tasks: Detection and Classification.

PGMM 1

- How to represent the images?
- Too difficult (initially) to represent the image intensities – for computational and modeling reasons.

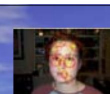
PGMM 1: Dots-World.

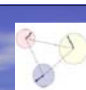
- Represent the images in terms of attributed points. $x_i = (z_i, \theta_i, A_i)$
- Where: z_i is the location of the feature
  θ_i is the orientation
  A_i is an appearance vector
- These attributed points can be extracted by the Brady-Kadir operation. They are represented with SIFT features (Lowe).

PGMM 1

- Images with attributed points.
- Problems:
  (i) some points are background.
  (ii) the object may have variable number of points.
  (iii) the object may have different appearances (e.g. due to changing viewpoint).
  (iii) the pose of the object is unknown.

PGMM 1: Triplets.

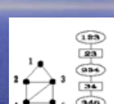




- To deal with the pose (position, scale, and orientation).
- Represent the object in terms of *oriented triplets* of points.
- Define invariant shape vector $\vec{l}(z_i, \theta_i, z_j, \theta_j, z_k, \theta_k)$ which is invariant to pose.
- The probability distribution will be defined on the invariant shape vector (invariant to pose).
- Gaussian distribution (with missing points).

PGMM 1: Combining Triplets.

- Build the model by combining triplets.

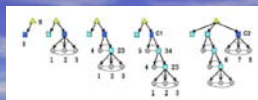


- The distribution on each triplet is a Gaussian defined on the shape invariant vector.
- This representation enables efficient inference (dynamic programming by junction trees).

PGMM 1: Background and Aspects.

- Need to model the background points. The number of these is unknown. Dirichlet process.
- Need to model different appearance of the object (e.g. different viewpoints).
- This gives a model with OR nodes.

PGMM 1:



- Possible Models:
- Triangles represent OR nodes. Squares represent AND nodes. Circles represent attributed points.
- Let *y* denote the configuration of the tree: positions, orientations, and attributes of points.
- *y* depends on topological, structure, and appearance parameters *omega* and *Omega*.
- (E.g. parameters of the Gaussian models, probabilities of OR nodes,...).

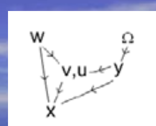
PGMM 1.

- Relating the model to the image.
- The leaf nodes of the tree correspond to background points or object points.
- Spatial assignment vectors tau.
- The full model is:

$$P(x, y, v, \omega, \Omega) = P(x|y, v, \omega_A)P(v|y, \omega_z)P(y|\Omega)P(\omega)P(\Omega).$$

- Where x denotes the positions and attributs of the image points.

Grammar for PGMM 1.



- To generate the image:
- (i) First generate the structure of the graph.
- (ii) Second, generate the image properties.

$$P(x, y, v, \omega, \Omega) = P(x|y, v, \omega_A)P(v|y, \omega_z)P(y|\Omega)P(\omega)P(\Omega)$$

- Graph Structure y, Observed Data x,
- Model parameters, Omega and omega,
- Internal variables u & v

PGMM 1

- Three Tasks:
- (1) Inference. Detect the object in a single image.
- (2) Learning the model parameters.
- (3) Learning the structure of the model (e.g. how many triples, how many OR nodes). Structure pursuit.

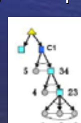
PGMM 1 Inference.

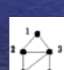
- The model and its parameters are known.

Inference requires estimating the parse tree (y, v) from input x.

This requires solving $(y^*, v^*) = \arg\max_{y,v} P(y, v|x, \omega, \Omega)$.

- Intuitively: match the points extracted from the image to the object and background points generated by the model.
- Dynamic Programming.





PGMM 1: Parameter Learning

- Know the structure of the model.

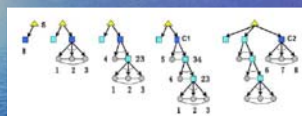
Parameter learning we specify a set W of parameters (ω, Ω) which we estimate by MAP.

Hence we estimate $(\omega^*, \Omega^*) = \arg\max_{\omega,\Omega \in W} \sum_{y,v} P(\omega, \Omega, y, v|x)$.

- EM algorithm. The variables y, nu are hidden.
- DP used to sum out over nu.

PGMM 1: Structure Pursuit.

- Learn the structure of the model.

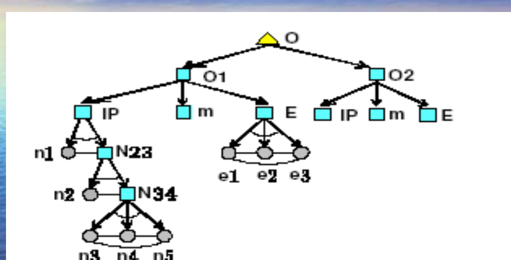


- Strategy: structure pursuit. Initialize with simplest model (everything is background).
- Grow the model by proposing new triplets. Accept, or reject, by model selection.

PGMM 1: Structure Pursuit.

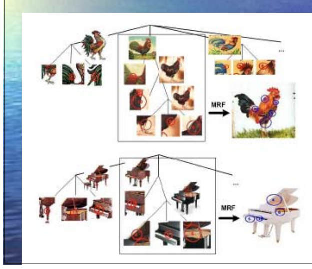
- Propose new triplets.
- Intuition: suspicious coincidences.
- First – *determine a feature vocabulary attributed points which frequently occur in the images. These are plausible candidates to be points on the object (background is variable).*
- Second – *determine a triplet vocabulary of tripes of attributed features (from the feature vocabulary set).*
- Use this triplet vocabulary to generate proposals.

PGMM 1 Grammar



Example Models

- Grand Piano, Rooster, Faces, Motorbikes, Airplanes.



Class of Models

- Harder Task:
- Image can contain airplanes, faces, or motorbikes.

Invariance to Rotation and Scale

- Invariance to image rotation and scale (range).

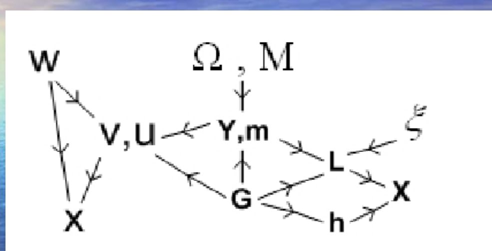
PGMM 1: Success & Limitations.

- Performance on Detection/Recognition is good (~ state of the art).
- Inference speed is very fast (seconds).
- But detection/recognition performance is not optimal, because we ignore many cues. (Can still recognize object if interest points removed).
- *But PGMM1 can only do detection and recognition because of its limited representation.*

From PGMM 1 to PGMM 2.

- Use PGMM1 to teach a new model PGMM 2.
- PGMM 1 gives rough estimation of position, orientation, scale of object (and identity).
- PGMM 2 includes a *mask for the shape* of the object, and edge features.
- *Enables segmentation and some crude parsing.*

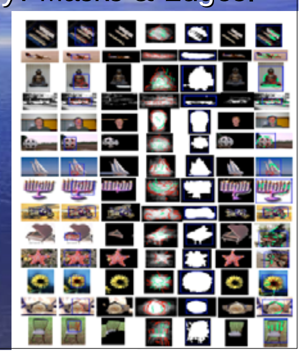
Bayes Net for PGMM 1 & PGMM 2



Pose variable is used to couple PGGM 1 & 2.

Richer Vocabulary: Masks & Edges.

- Combination of features.
- Interest-points.
- Masks.
- Edgelets.
- *Enables segmentation and limited parsing.*



Results: Recognizing, Segmenting, and Parsing.

- Measures of Success

Dataset	Size	Class. 1	Class. 2	Det. 1	Det. 2	Seg. 1	Seg. 2	Pre./Rec. 1	Pre./Rec. 2	Parsing
Accordion	55	97.5	97.5	28.8	79.5	37.8	40.5	80.6 / 43.0	88.2 / 44.0	79.5
Airplane	800	91.3	91.8	42.2	60.7	51.4	62.5	61.4 / 75.9	75.2 / 75.4	87.6
Buddha	85	88.8	91.3	40.6	70.6	67.4	68.5	76.0 / 85.4	80.9 / 83.4	87.1
Car	123	89.2	90.3	45.6	70.5	23.1	32.5	28.0 / 61.6	50.0 / 54.3	67.3
Face	435	95.8	96.8	41.7	71.7	66.2	69.0	72.6 / 87.0	73.5 / 89.6	92.0
Football	64	68.3	78.3	35.1	65.1	49.0	58.9	96.8 / 50.5	93.0 / 62.6	97.7
Ketch	114	85.0	87.0	37.9	63.1	50.9	53.6	67.9 / 69.7	69.8 / 71.0	90.8
Menorah	87	71.3	73.8	30.9	63.6	24.8	30.2	73.2 / 35.4	74.2 / 38.3	97.1
Motorbike	798	86.1	94.6	62.0	63.6	58.4	72.6	80.9 / 71.8	82.8 / 86.3	97.4
Grand Piano	90	87.0	93.0	28.7	76.8	57.2	73.4	86.2 / 61.5	87.8 / 81.3	95.7
Starfish	86	78.5	78.5	43.2	56.6	57.3	61.5	71.5 / 77.5	77.1 / 78.5	84.1
Sunflower	85	86.3	88.8	42.8	72.8	71.7	73.8	87.9 / 79.4	87.9 / 81.8	95.8
Watch	239	86.5	90.5	47.8	79.7	59.8	66.5	94.0 / 63.4	95.4 / 69.2	98.8
Windsor Chair	56	97.5	97.5	31.3	79.0	48.9	57.3	84.7 / 55.8	94.4 / 56.0	97.6

Alternative Models.

- Deformable Models with Parts. (Iasonas Kokkinos)
- Hierarchical Models. (Long Zhu -- Leo).

Deformable Objects with Parts.

- The previous approach will not work on objects like cows, horses, or yaks.
- Richer representation based on edges and ridge features (Lindeberg's primal sketch).
- Generative model includes multiple parts (e.g. legs) which may, or may not, be present.
- Kokkinos & Yuille (ICCV 07).

Objects with parts.

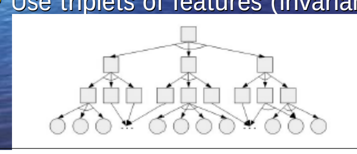
- Our previous models assume that the shape was fixed (mask). Now we allow deformations and movable parts.
- Proposals to add new parts.

Hierarchical Models.

- Previous models represented the object at a single scale only.
- *The structure of the representation was motivated partly by whether we could compute it or not (i.e. use DP).*
- Need to enhance the representation by introducing hierarchy.

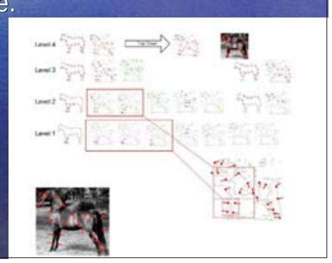
Hierarchical Model.

- Hierarchical Model. AND nodes only (extending to AND/OR graphs).
- Encode spatial relations at all scales.
- Input: edgelets in the image.
- Use triplets of features (invariance).



Representation

- Variables assigned to node: position, orientation, scale.



Inference Algorithm: Bottom-Up

- Bottom-Up and Top-Down.
- Bottom-Up: make proposals for sub-configurations of the object by combining proposals for elementary proposals.
- Prune out proposals to prevent combinatorial explosion: surround suppression, local goodness of fit.
- Surround suppression – loss of resolution, recovered by top-down. (Tai Sing Lee).

Inference Algorithm: Top-Down

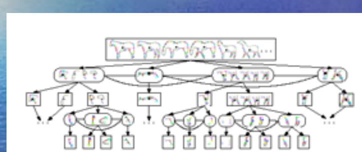
- Bottom-Up makes proposals for possible configurations of the object.
- Top-down process refines and validates (or rejects) these proposals.
- Top-down explores proposals that were rejected by the bottom-up process due to surround suppression.

Examples of the Hierarchy

- Good performance results.
- Detection & Segmentation.
- Evaluated on 100's images.

Extend to AND/OR Graph.

- Horses as AND/OR graphs.



Parsed Results for AND/OR graph

- The OR nodes enable the model to account for different configurations of the horse.

Summary:

- Research Program for unsupervised learning of probabilistic grammars for objects.
- Difficulties: complexities of images. Variable pose, variable appearance, cluttered background.
- Strategy is incremental. Points, Masks, Deformable parts, hierarchies.
- Structure Learning: Proposals generated by suspicious coincidences.