Scene Grammars, Factor Graphs, and Belief Propagation

Pedro Felzenszwalb

Brown University

Joint work with Jeroen Chua

Probabilistic Scene Grammars

General purpose framework for image understanding and machine perception.

- What are the objects in the scene, and how are they related?
- Scene have regularities that provide context for recognition.
- Objects have parts that are (recursively) objects.
- Relationships are captured by compositional rules.

Vision as Bayesian Inference

The goal is to recover information about the world from an image.

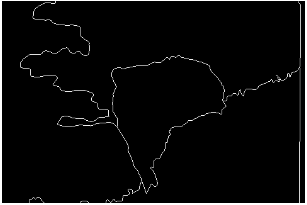
- Hidden structure X (the world/scene).
- Observations Y (the image).
- Consider the posterior distribution and Bayes Rule

$$p(X|Y) = \frac{p(Y|X)p(X)}{p(Y)}$$

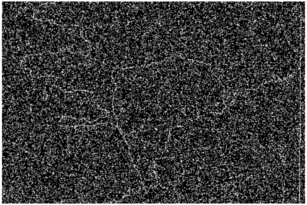
- The approach involves an imaging model $p(Y|X)$
- And a prior distribution $p(X)$

Image Restoration

Clean image x



Measured image $y = x + n$.



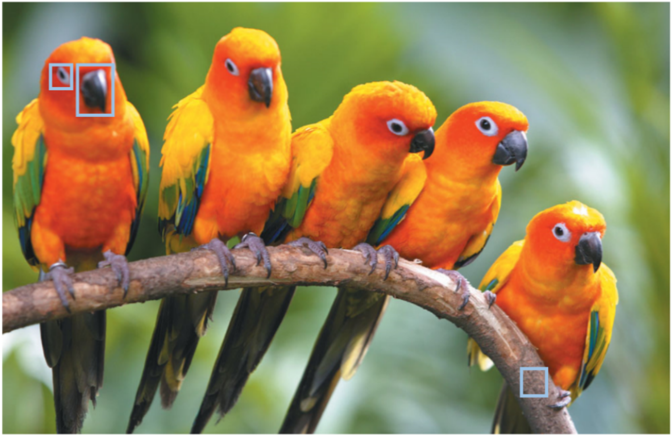
Ambiguous problem.

Impossible to restore a pixel by itself.

Requires modeling relationships between pixels.

Object Recognition

Object Recognition



Context is key for recognition.

Captured by relationships between objects.

Modeling scenes

$$p(X)$$

Scenes are complex high-dimensional structures.

The number of possible scenes is very large (infinite), yet scenes have regularities.

- Faces have eyes.
- Boundaries are piecewise smooth.
- etc.

A set of regular scenes forms a "Language".

Regular scenes can be defined using stochastic grammars.

The Framework

- Representation: Probabilistic scene grammar.
- Transformation: Grammar model to factor graph.
- Inference: Loopy belief propagation.
- Learning: Maximum likelihood (EM).

Scene Grammar

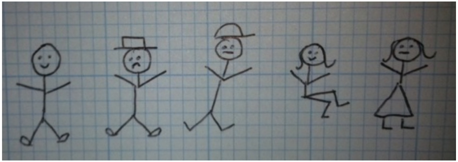
Scenes are structures generated by a stochastic grammar.

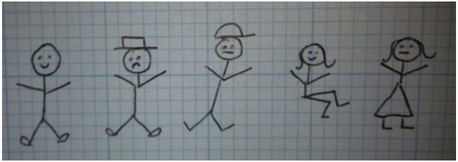
Scenes are composed of objects of several types.

Objects are composed of parts that are (recursively) objects.

Parts tend to be in certain relative locations.

The parts that make up an object can vary.





$\text{PERSON} \to \{\text{FACE}, \text{ARMS}, \text{LOWER}\}$

$\text{FACE} \to \{\text{EYES}, \text{NOSE}, \text{MOUTH}\}$

$\text{FACE} \to \{\text{HAT}, \text{EYES}, \text{NOSE}, \text{MOUTH}\}$

$\text{EYES} \to \{\text{EYE}, \text{EYE}\}$

$\text{EYES} \to \{\text{SUNGLASSES}\}$

$\text{HAT} \to \{\text{BASEBALL}\}$

$\text{HAT} \to \{\text{SOMBRERO}\}$

$\text{LOWER} \to \{\text{SHOE}, \text{SHOE}, \text{LEGS}\}$

$\text{LEGS} \to \{\text{PANTS}\}$

$\text{LEGS} \to \{\text{SKIRT}\}$

Scene Grammar

- Finite set of symbols (object types) Σ.
- Finite pose space Ω_A for each symbol.
- Finite set of productions $\mathcal{R}$.

$$A_0 \rightarrow \{A_1, \ldots, A_K\} \quad A_i \in \Sigma$$

- Rule selection probabilities $p(r)$.
- Conditional pose distributions associated with each rule.

$$p_i(\omega_i | \omega_0)$$

- Self-rooting probabilities ϵ_A.

Scene

Set of building blocks, or bricks,

$$\mathcal{B} = \{(A, \omega) \,|\, A \in \Sigma, \omega \in \Omega_A\}.$$

A scene is defined by

- A subset of bricks $O \in \mathcal{B}$.
- For each brick in $(A, \omega) \in O$ a rule $A \to \{A_1, \ldots, A_K\}$ and poses $\omega_1, \ldots, \omega_K$ such that $(A_i, \omega_i) \in O$.

Generating a scene

Brick (A, ω) is *on* if the scene has an object of type A in pose ω.

Stochastic process:

- Initially all bricks are off.
- Independently turn each brick (A, ω) on with probability ϵ_A.
- The first time a brick is turned on, expand it.

Expanding (A, ω):

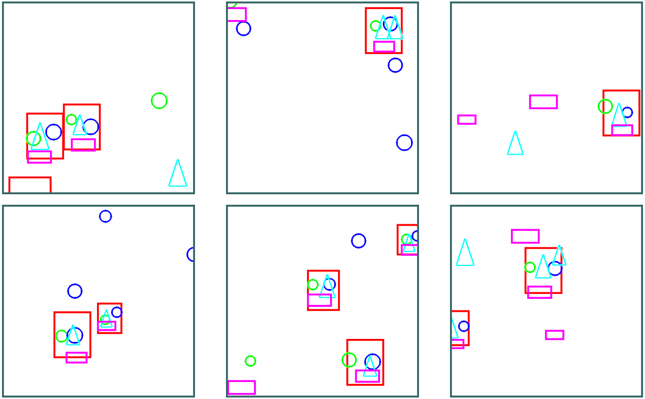
- Select a rule $A \to \{A_1, \ldots, A_K\}$.
- Select K poses $(\omega_1, \ldots, \omega_K)$ conditional on ω.
- Turn on bricks $(A_1, \omega_1), \ldots, (A_K, \omega_K)$.

A grammar for scenes with faces

- Symbols $\Sigma = \{\mathrm{FACE}, \mathrm{EYE}, \mathrm{NOSE}, \mathrm{MOUTH}\}$.
- Poses space $\Omega = \{(x, y, \mathrm{size})\}$.
- Rules:

  (1) $\mathrm{FACE} \to \{\mathrm{EYE}, \mathrm{EYE}, \mathrm{NOSE}, \mathrm{MOUTH}\}$
  (2) $\mathrm{EYE} \to \{\}$
  (3) $\mathrm{NOSE} \to \{\}$
  (4) $\mathrm{MOUTH} \to \{\}$
- Conditional pose distributions for (1) specify typical locations of face parts within a face.
- Each symbol has a small self rooting probability.

Random scenes with face model



A grammar for images with curves

- Symbols $\Sigma = \{\mathrm{C}, \mathrm{P}\}$.
- Pose of C specifies position and orientation.
- Pose of P specifies position.
- Rules:

(1) $\mathrm{C}(x, y, \theta) \to \{\mathrm{P}(x, y)\}$

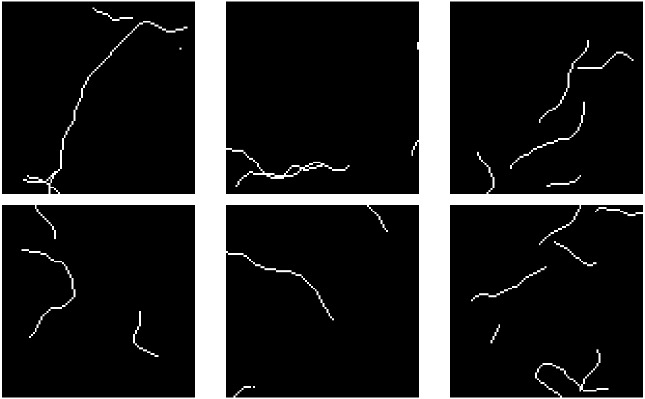
(2) $\mathrm{C}(x, y, \theta) \to \{\mathrm{P}(x, y), \mathrm{C}(x + \Delta x_\theta, y + \Delta y_\theta, \theta)\}$

(3) $\mathrm{C}(x, y, \theta) \to \{\mathrm{C}(x, y, \theta + 1)\}$

(4) $\mathrm{C}(x, y, \theta) \to \{\mathrm{C}(x, y, \theta - 1)\}$

(5) $\mathrm{P} \to \{\}$

Random images

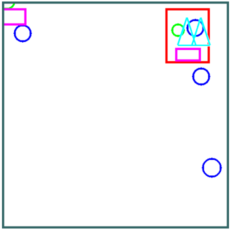


Computation

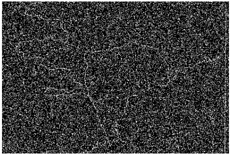
Grammar defines a distribution over scenes.

A key problem is computing conditional probabilities.

What is the probability that there is a nose near location $(20, 32)$ given that there is an eye at location $(15, 29)$?



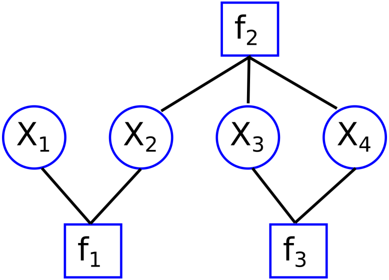
What is the probability that each pixel in the clean image is on, given the noisy observations?



Factor Graphs

A factor graph represents a factored distribution.

$$p(X_1, X_2, X_3, X_4) = f_1(X_1, X_2) f_2(X_2, X_3, X_4) f_3(X_3, X_4)$$

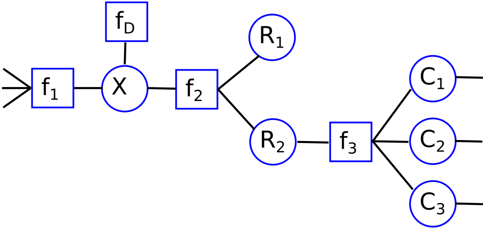


Variable nodes (circles)

Factor nodes (squares)

Factor Graph Representation for Scenes

"Gadget" represents a brick



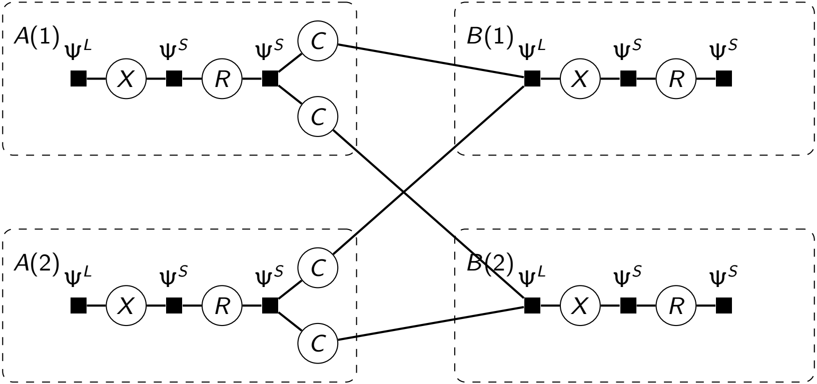
Binary random variables

- X brick on/off
- R_i rule selection
- C_i child selection

Factors

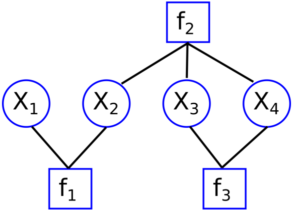
- f_1 Leaky-or
- f_2 Selection
- f_3 Selection
- f_D Data model

$\Sigma = \{A, B\}$.
$\Omega = \{1, 2\}$.
$A(x) \rightarrow B(y)$
$B(x) \rightarrow \{\}$.



Loopy belief propagation

Inference by message passing.



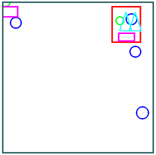
$$\mu_{f\to v}(x_v) = \sum_{x_{N(f)\setminus v}} \Psi(x_{N(f)}) \prod_{u\in N(f)} \mu_{u\to f}(x_u)$$

In general message computation is *exponential* in degree of factors.

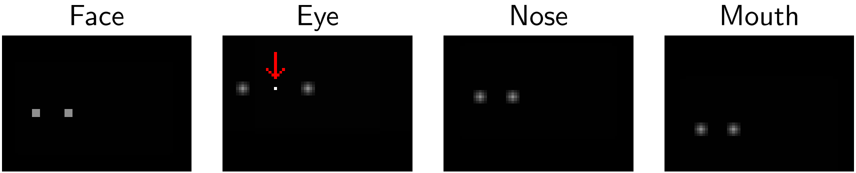
For our factors, message computation is *linear* in degree.

Conditional inference with LBP

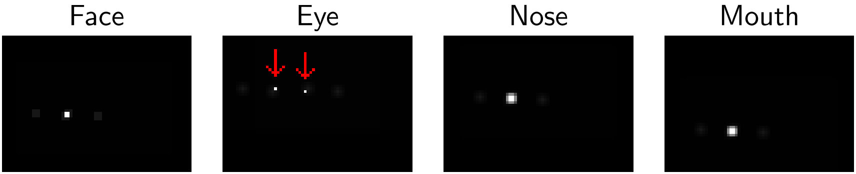
$$\Sigma = \{\mathrm{FACE}, \mathrm{EYE}, \mathrm{NOSE}, \mathrm{MOUTH}\}$$
$$\mathrm{FACE} \rightarrow \{\mathrm{EYE}, \mathrm{EYE}, \mathrm{NOSE}, \mathrm{MOUTH}\}$$



Marginal probabilities conditional on one eye.

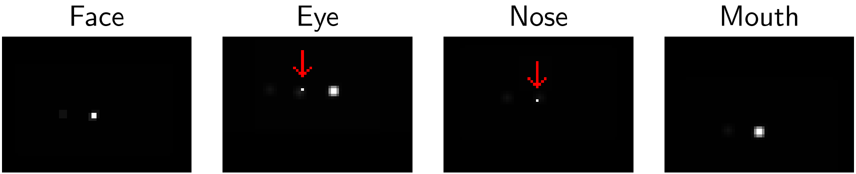


Marginal probabilities conditional on two eyes.



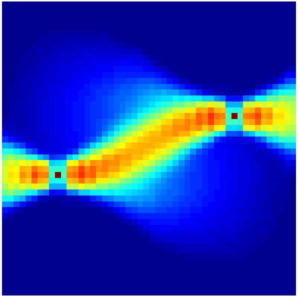
Conditional inference with LBP

- Evidence for an object provides context for other objects.
- LBP combines “bottom-up” and “top-down” influence.
- LBP captures chains of contextual evidence.
- LBP naturally combines multiple contextual cues.



Conditional inference with LBP

Contour completion with curve grammar.



Face detection

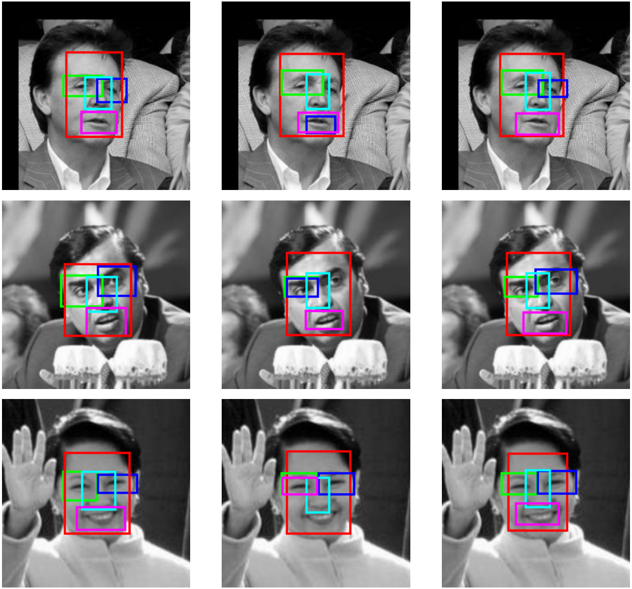
$$p(X|Y) \propto p(Y|X)p(X)$$

$p(Y|X)$ defined by templates for each symbol.

Defines local evidence for each brick in the factor graph.

Belief Propagation combines “weak” local evidence from all bricks.

Face detection results



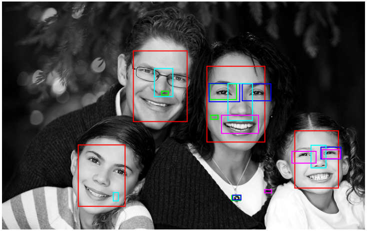
Ground Truth

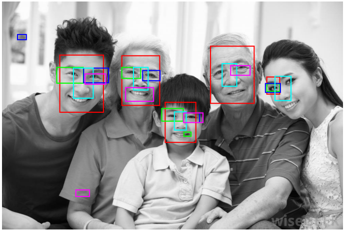
HOG Filters

Face Grammar

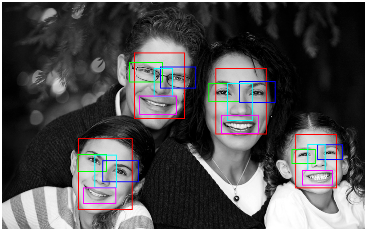
Scenes with several faces

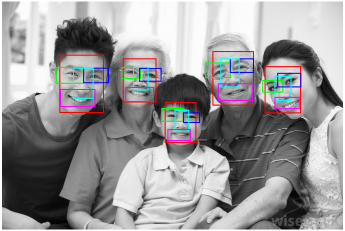
HOG filters





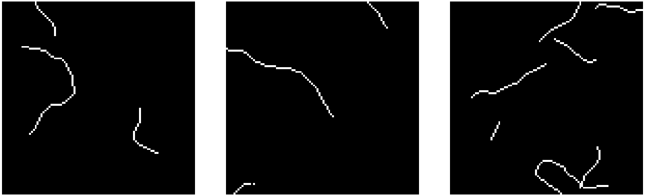
Grammar



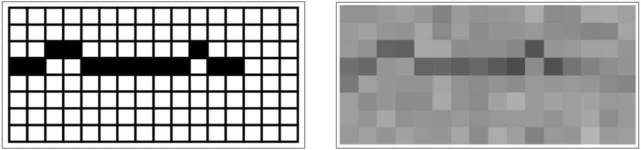


Curve detection

$p(X)$ defined by a grammar for curves.



$p(Y|X)$ defined by noisy observations at each pixel

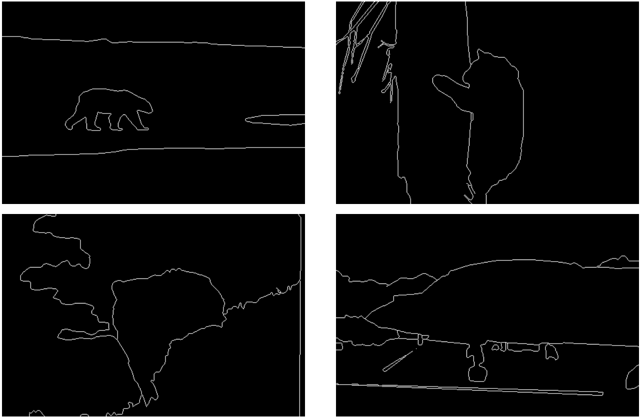


X

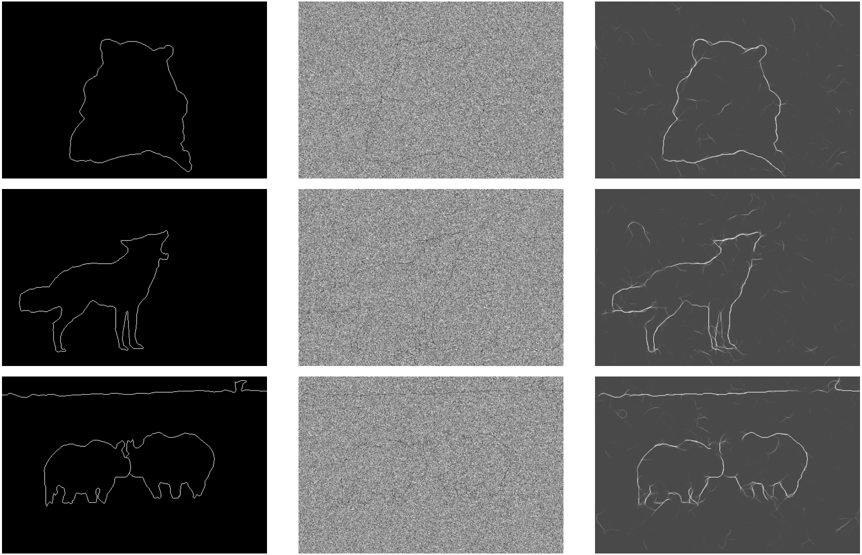
Y

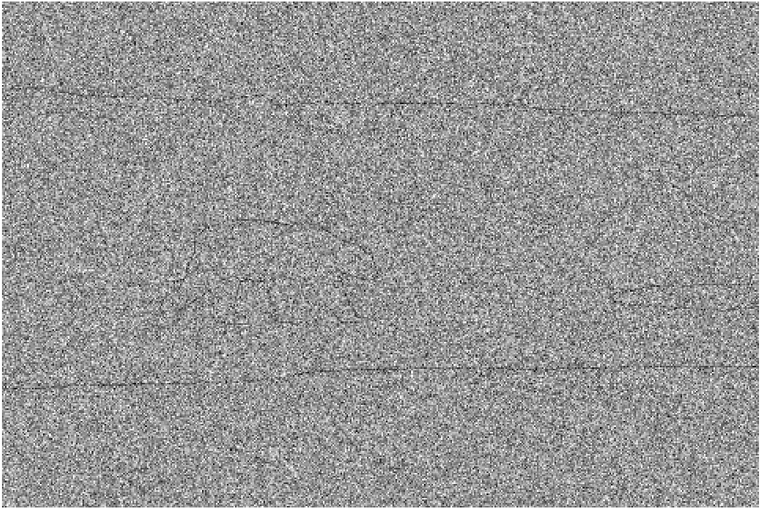
Curve detection dataset

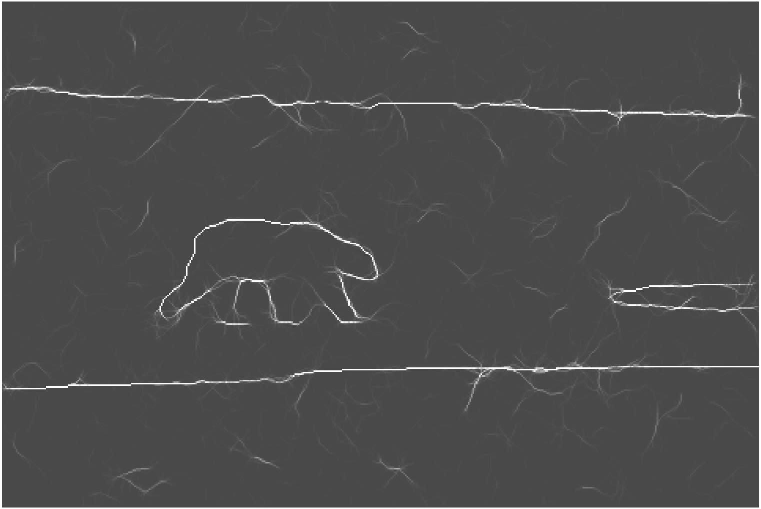
Ground-truth: human-drawn object boundaries from BSDS.

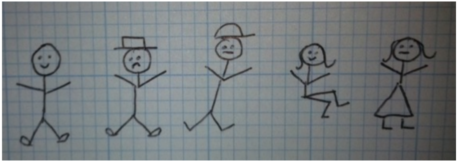


Curve detection results









PERSON → {FACE, ARMS, LOWER}

FACE → {EYES, NOSE, MOUTH}

FACE → {HAT, EYES, NOSE, MOUTH}

EYES → {EYE, EYE}

EYES → {SUNGLASSES}

HAT → {BASEBALL}

HAT → {SOMBRERO}

LOWER → {SHOE, SHOE, LEGS}

LEGS → {PANTS}

LEGS → {SKIRT}