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?
- Objects have parts that are (recursively) objects.
- Relationships provide context for recognition.
- Relationships are captured by compositional rules.
Example: Image Restoration

Clean image $x$

Measured image $y = x + n$.

Ambiguous problem.

Impossible to restore a pixel by itself.

*Context* is of key importance.

Requires modeling relationships between pixels.
Example: Object Recognition
Example: Object Recognition

*Context* is of key importance. Requires modeling relationships between objects.
Vision as Bayesian Inference

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

- Hidden variables $X$ (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 $p(X)$
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 of objects are piecewise smooth.

The set of regular scenes forms a “Language”.

Scenes can be generated 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.
Scene Grammar

- Finite set of symbols (object types) $\Sigma$.
- Finite pose space $\Omega$ 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 $\epsilon_A$. 
Generating a scene

Set of building blocks, or bricks, \((A, \omega) \in \Sigma \times \Omega\).

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

Stochastic process:

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

Expanding \((A, \omega)\):

- Select a rule \(A \rightarrow \{A_1, \ldots, A_K\}\).
- Select \(K\) poses \((\omega_1, \ldots, \omega_K)\) conditional on \(\omega\).
- Turn on bricks \((A_1, \omega_1), \ldots, (A_K, \omega_K)\).
A grammar for scenes with faces

- Symbols $\Sigma = \{\text{FACE, EYE, NOSE, MOUTH}\}$.
- Poses space $\Omega = \{(x, y, \text{size})\}$.
- Rules:
  1. $\text{FACE} \rightarrow \{\text{EYE, EYE, NOSE, MOUTH}\}$
  2. $\text{EYE} \rightarrow \{\}$
  3. $\text{NOSE} \rightarrow \{\}$
  4. $\text{MOUTH} \rightarrow \{\}$
- 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 $\Sigma = \{C, P\}$.
- Pose of $C$ specifies position and orientation.
- Pose of $P$ specifies position.
- Rules:

  1. $C(x, y, \theta) \rightarrow \{P(x, y)\}$
  2. $C(x, y, \theta) \rightarrow \{P(x, y), C(x + \Delta x_\theta, y + \Delta y_\theta, \theta)\}$
  3. $C(x, y, \theta) \rightarrow \{C(x, y, \theta + 1)\}$
  4. $C(x, y, \theta) \rightarrow \{C(x, y, \theta - 1)\}$
  5. $P \rightarrow \{\}$
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.

\[
\mu_{f \rightarrow v}(x_v) = \sum_{x_{N(f) \setminus v}} \Psi(x_{N(f)}) \prod_{u \in N(f)} \mu_{u \rightarrow f}(x_u)
\]

In general message computation is \textit{exponential} in degree of factors. For our factors, message computation is \textit{linear} in degree.
Conditional inference with LBP

\[ \Sigma = \{ \text{FACE, EYE, NOSE, MOUTH} \} \]
\[ \text{FACE} \to \{ \text{EYE, EYE, NOSE, MOUTH} \} \]

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.

![Face](image1.png)  ![Eye](image2.png)  ![Nose](image3.png)  ![Mouth](image4.png)
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
Curve detection dataset

Ground-truth: human-drawn object boundaries from BSDS.
Curve detection results
Summary

- Perceptual inference requires models of the world.

- Stochastic grammars can model complex scenes.
  - Part-based representation.
  - Scenes and objects have variable structure.

- Belief propagation provides a general computational engine.

- The same framework applies to many different tasks.
Future work

• Data model $p(Y|X)$.
• Grammar learning.
• Scaling inference to larger grammars.
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}