Part and Appearance Sharing: Recursive Compositional Models for Multi-View Multi-Object Detection

Presentation based on the paper of the same title by Long Zhu, Yuanhao Chen, Antonio Torralba, William Freeman, and Alan Yuille

University of Toronto Reading Group Presentation
Sagan Bolliger and Zoya Gavrilov
July 19, 2012

Motivation

- Simultaneous detection + parsing (view estimation, part configurations)
- Multi-view, multi-object
- Want to reduce complexity of representation + learning
- High variation of shape + appearance is a challenge for rapid detection (inference) and learning

Contributions

- Part and appearance sharing allows for compact representation and efficient learning (fewer training examples required)
- Hierarchical parts-based object representation

Recursive Compositional Models (RCMs)

- Recursive representation of objects as compositions of parts

$$RCM : (\mathcal{V}, \mathcal{E}, \vec{\psi}, \vec{\phi}, \phi_{\mathcal{R}})$$

- Edges encode parent-child relations of parts
- State variables: $w_\mu = (x_\mu, \theta_\mu, s_\mu), \forall \mu \in \mathcal{V}$

Probability Distribution over RCMs

- Gibbs: $P(W|I) = \frac{1}{Z(\mathbf{I})} exp\{-E(W,\mathbf{I})\}$
- Where:

$$E(W,\mathbf{I}) = \mathbf{\Sigma}_{\mu \in \mathcal{V} \setminus \mathcal{V}_{\mathbf{leaf}}} \psi_\mu(w_\mu, w_{ch(\mu)}) + \Sigma_{\mu \in \mathcal{V}_{leaf}} \phi_\mu(w_\mu, \mathbf{I}) + \phi_{\mathcal{R}}(w_R, \mathbf{I})$$

- Child nodes of μ denoted $w_{ch(\mu)}$
- R is the root node of the RCM

Shape Potentials

$$E(W,\mathbf{I}) = \mathbf{\Sigma}_{\mu\in\mathcal{V}\setminus\mathcal{V}_{\mathbf{leaf}}}\psi_\mu(w_\mu, w_{ch(\mu)}) + \Sigma_{\mu\in\mathcal{V}_{leaf}}\phi_\mu(w_\mu,\mathbf{I}) + \phi_{\mathcal{R}}(w_R,\mathbf{I})$$

- Shape potentials $\vec{\psi}$ encode spatial relationships between states of parent nodes and children
- From [19]: $\psi_\mu(w_\mu, w_{ch(\mu)})$ is 0 provided the average orientations and positions of the child nodes are equal to the orientation and position (x_μ, θ_μ) of the parent node, or a large positive constant otherwise
- Note that the scale (s_μ) of the parent is simply defined by be the sum of the scales of its children (sum of regions in the image) – does not affect $\psi_\mu(w_\mu, w_{ch(\mu)})$

Appearance Potentials

$$E(W,\mathbf{I}) = \mathbf{\Sigma}_{\mu\in\mathcal{V}\backslash\mathcal{V}_{\mathbf{leaf}}}\psi_\mu(w_\mu, w_{ch(\mu)}) + \Sigma_{\mu\in\mathcal{V}_{leaf}}\phi_\mu(w_\mu,\mathbf{I}) + \phi_{\mathcal{R}}(w_R,\mathbf{I})$$

- Appearance potentials relate the leaf nodes and root node to the input image
- Boundary potentials $\vec{\phi}$ at leaf nodes
  - 10 oriented generic boundary segments
- Body potential ϕ_R at root node
  - Average of the texture/material properties of image patches inside the body of the object, specified by an object mask (located, scaled, and rotated by the state of the root node w_R)

$$\phi_{\mathcal{R}}(w_R,\mathbf{I}) = (\frac{\mathbf{1}}{|\mathbf{R}(\mathbf{\Omega},\mathbf{w_R})|}\mathbf{\Sigma}_{\mathbf{x}\in\mathbf{R}(\mathbf{\Omega},\mathbf{w_R})}\phi(\mathbf{x},\mathbf{I})$$

Dictionaries and Object-RCM

- $\mathcal{T}^l$ is a dictionary (a set) of RCMs with l levels
  - $t_a^l = (\mathcal{V}_a^l, \mathcal{E}_a^l, \vec{\psi}_a^l, \vec{\phi}_a^l)$ is the RCM belonging to object a
- constraint: RCM in T^l is a composition of 3 RCMs in T^{l-1}
- $\mathcal{T} = \cup_{l=1}^{L} \mathcal{T}^l$ is a hierarchical dictionary over which inference is performed
- At the top level of the hierarchical dictionary are the object RCMs; they have L levels & comprise shared-RCMs
- each object RCM is specified by a list of its elements in sub-dictionaries (at levels $L-1, L-2, \ldots$) along with the $\psi_R(w_R, w_{ch(R)})$ and $\phi_{\mathcal{R}}(w_R, \mathbf{I})$

Sharing RCMs

- Objects (within a category) share common parts which have the same spatial layout (RCMs with same shape potential)
- Sharing between object categories allows for efficient learning and representation
- Sharing of appearance potentials (one body potential and 10 boundary potentials for one object category)
- Learning a single body appearance model per class
- Thereby, model complexity of RCMs for <u>appearance</u> is linear in size of object classes and invariant to the number of viewpoints
- So, all instances of a given object class (including all viewpoints of each instance) are separately represented by RCMs, but all those RCMs share a single appearance model

Inference

- Find all instances t_a^1 in dictionary T^1 whose energy is below a threshold (and whose states are sufficiently different, selecting those with minimal local energy)
- Form compositions (of size 3) to obtain instances of RCMs from the next level of the dictionary
- Proceed until reach the object-RCMs at top level
- Compute energies for RCMs at each successive level in terms of the RCMs at the previous level, recursively:

$$E(W_\mu, \mathbf{I}) = \psi_\mu(w_\mu, w_{ch(\mu)}) + \Sigma_{v \in ch(\mu)} E(W_v, \mathbf{I})$$

Inference (continued)

- Don’t waste time detecting parts for different objects separately as long as they share parts
- Composition of parts from one level to the next is done over a fixed neighbourhood size at that level
- Performed at different scales of an image pyramid (i.e. at 4 scales) with scaling factor 1.5; resolution of edge features initialized relative to scales of image pyramid

Learning

- Input: dataset (LabelMe) where boundary of shape is known (object and viewpoint not known)
- (1) learn dictionary of RCMs and shape potentials
- (2) learn object/viewpoint masks and appearance potentials

Learning dictionaries of RCMs

- Model: 6 single-node models (dictionary at level $\mathcal{T}^0$), each with a potential favouring boundary edges at a given orientation: $a\pi\backslash 6, a = 0, \ldots, 5$
- Quantize orientation of local segments (3 pixels) into 6 orientation bins
- Detect all instances of level-1 RCMs in dataset
- For each triplet of RCMs in this level, compose them to form hypothesized instances in the next level
- Reject compositions which fail a “spatial test for composition”: 2 (max and min) circles are drawn around the centers of the 3 child instances: if all child instances lie within any of the 3 max circles, but do not all lie in any of the 3 min circles, the composition is valid; otherwise, reject

Learning (continued)

- Cluster compositions in triplet space to obtain a set of prototype triplet clusters
- Estimate the potentials of these clusters to produce a new set of RCMs for the next level
- Prune the dictionary at this level of instances that overlap spatially in the image
- Repeat process for the next level of the dictionary – procedure automatically terminates when it ceases to find new compositions

Learning masks and appearance potentials

- To learn masks: simple averaging over the boundaries of different training examples, for fixed object and viewpoint (supervised learning)
- Appearance potentials learnt using logistic regression techniques using image features as input:
  - Body features: greyscale intensity, color, intensity gradient, Canny edges, response of DOG filters
  - Total of 55 spatial filters
  - Boundary features: edges + corners or edge-based filter responses
- Fixed 10 boundary potentials common to all objects and viewpoints; fixed appearance potential per object class (common to all viewpoints of the class)
- Body and boundary potentials weighted by two scalar parameters (set by cross validation)

Results: some highlights

- Through unsupervised learning, able to learn Y-junctions and T-junctions (perceptual grouping) at levels 2 and 3 of the dictionary
- Multi-view car detection: most errors are either adjacent or symmetric viewpoints
- 26 object classes with a total of 120 objects/viewpoints; unsupervised learning of hierarchical dictionaries produced 119 RCMs – i.e. automatically learned object specificity