Big Picture Ideas

Willie

April 16, 2012

Abstract

This document contains big picture ideas and directions for research. Currently there are three related but distinct directions for projects that I am looking into: (1) A nonparametric Bayesian treatment of learning object types within a Dependent Dirichlet Process (DDP) tracking scheme, (2) a sophisticated appearance model for tracking a specific object type with a DDP tracker (based on the work of Sigal), and (3) ways to represent and infer the structure of abstract objects within a DDP tracking scheme.

1 NPBayes Treatment of Object Types

Here, I am trying to extend the current "simple" Multivariate-Normal-Multinomial object appearance model (MNM model) of the GPUDDPM such that each object has an assignment, or "type". Generatively speaking, we start with a mixture model (where each mixture component is an object type), we draw object appearance parameters given a component of this mixture, and we draw image observations (foreground pixels) given these appearance parameters. I.e., in order to incorporate object types, we make the prior distribution over the object appearance parameters a mixture model, where we might seek to infer the parameters and/or number of mixture components. To start off, I am fixing the number of type-clusters and inferring component parameters only (and I'll do inference on a video with, for example, two types of objects to gauge performance). Next, I want to incorporate a NPB prior to estimate the number of types.

We are left with an important modeling consideration: as each object-cluster is actually a sequence of object appearance parameters in the GPUDDPM model, should the parameters of a single object be allowed to be drawn from different "types" at different points in time? In the framework of the GPUDDPM, I believe this is necessary, since each $\theta_{k,t}$ is drawn from the base distribution $\mathbb{G}_0$; thus, if the base distribution is a mixture, the cluster parameters are able to switch between modes of the mixture at different points in time. This may prove to be ok, as the transition kernel should keep parameters similar between adjacent time steps, and therefore keep an object within a mode of a mixture. An alternate solution could be to have a separate GPUDDPM for each "type" (i.e. a mixture of GPUDDPMs)—however, I really don't think inference would be feasible in a model like this.

NOTE: I think I might've come up with a way to solve this problem in the framework of the GPUDDPM. I have graphical models in notebook to show to Frank. Basically, consider a DPM with a mixture model base distribution (i.e. each cluster parameter now has an assignment variable); then consider a GPUDDPM, where we extend each cluster parameter to a sequence of parameters, but still keep a single assignment variable for each sequence.

The idea in this line of research is to figure out a way to add object type clusters to the GPUDDPM, using (nearly) the same representation as before. Once this is developed, I'll be able to assign types to objects in the current model (i.e. this will yield object clusters with similar oval / color parameters) and it may be possible to apply this technique to the other developments below.

2 Sophisticated Appearance Model for Specific Tracker

Sigal has developed a sophisticated model for estimating human pose (i.e. estimating the position and orientation, in $\mathbb{R}^3$, of a collection of nodes, each representing a part of the body). His "Loose-limbed Human Body Model" is structured as an undirected graphical model, with kinematic and penetration constraints between the nodes, and specific likelihood distributions for each node defined on foreground pixel and edge image features.

Sigal's model uses a lot of information to make really good estimates about the pose of a single, upright human in a video. The model is able to incorporate multiple camera views, externally trained body-part detectors (e.g. head, hands, feet), separate likelihoods for image features common to each body part, and constraints that are justified under assumptions about the orientation, body structure, and behavior of the typical human walking in a video. Even then, the model sometimes makes mistakes.

I believe that estimating the 3D pose of a deformable figure to a high precision is probably very hard or impossible without many of the pieces of information and dependencies that Sigal is making use of. An initial plan was to infer parameters of a model similar to Sigal's (instead of fully specifying certain parts of the model as Sigal does) for different animals by setting up a general model with more latent parameters (e.g. something like a "general vertebrate model"); the strategy would've been, given foreground-pixel data from different animals, to infer distributions over and/or contraints between the parameters of the model. I still think this project might give results, but I am not optimistic about our ability to accurately estimate pose parameters that match up with the true pose (mostly because these Sigal-esque models are very high dimensional, and we use only a single camera, no edge image data, noisier foreground-pixel data, often much smaller and lower resolution figures, there is high variability between the poses of different vertebrates, and we don't expect to incorporate the many object-specific assumptions as Sigal is able to do with humans). Relating to this topic, I've been working with ideas on inferring more-general representations of arbitrary objects (instead of the 3D pose of vertebrates), which I write about in the next section.

However, while I am more interested in developing methods to automatically learn object structure from a video, I still think the GPUDDPM could contribute to Sigal's work. He developed a very good appearance distribution for humans, and we have a model for localizing/segmenting and tracking multiple objects simultaneously, which makes use of (i.e. needs) an appearance distribution. I don't think Sigal has developed a way to track multiple humans at once (in fact, all the tracking that he does on single humans is very basic—through, it has also shown to be very effective). To apply Sigal's appearance distribution to the GPUDDPM tracking scheme we need a prior over the model parameters (for a single human) and a transition kernel between these model parameters that has the prior as its stationary distribution).

I define the chief aspects of Sigal's Loose-limbed Human Body Model here:

The model is a 10-node undirected graphical model, where each node represents a body part. The i_{th} node has 6 fixed parameters $\Phi_i = [l_i, w_i^p, w_i^d, o_i^p, o_i^d, \epsilon_i^d]$ (length, width at both ends, offset at both ends, and eccentricity) and six parameters to be inferred, $\mathbf{X}_i \in \mathbb{R}^6$.

Sigal defines $\mathbf{X}_i = [\mathbf{x}_i, \mathbf{q}_i]$, where $\mathbf{x}_i = [\mathbf{x}_{x,i}, \mathbf{x}_{y,i}, \mathbf{x}_{z,i}]$, and $\mathbf{q} = \mathrm{SO}(3) = [q_{x,i}, q_{y,i}, q_{z,i}, q_{w,i}] \in \mathbb{R}^4$ (the rotation group). All together, $\mathbf{X}_i$ is in 7 continuous dimensions, but lies on a 6 dimensional manifold.

Each (undirected) edge between parts i and j has a potential function $\psi_{ij}(\mathbf{X}_i, \mathbf{X}_j)$, which is composed of kinematic contraints $\psi_{ij}^K(\mathbf{X}_i, \mathbf{X}_j)$, and interpenetration constraints $\psi_{ij}^P(\mathbf{X}_i, \mathbf{X}_j)$.

The foreground pixel likelihood is defined for each part configuration $\mathbf{X}_i$ to be a function of pixels in an image partitioned into three subsets: those representing the given body part configuration, those correlated with the body part (e.g. pixels representing the feet if the body part is a leg), and pixels that have nothing to do with the given body part.

Overall, this is a 60 parameter model. I need to define reasonable priors over these parameters and a transition kernel whose stationary distribution is equal to this prior. Then we could try and see if we get any accuracy in pose estimation with our data.

3 Representing and Inferring Object Structure

Depending on the success we have inferring the three-dimensional pose of objects, we could try to infer other structural representations that might provide a less-realistic 3D depiction, but be more applicable to wider (and more arbitrary) classes of objects. For example, suppose we chose to model (the foreground pixels of) a single object in a video with a DDP mixture of bivariate Gaussians. This model would incorporate dependencies between adjacent frames for each body part, and would automatically infer a wide scale of structures and complexities—it would, in fact, be a nonparametric representation of object structure.

Didn't finish writing. We can talk more about this in person.