Literature Review: Graphical Models for Articulated Forms, Humans, and Other Objects

Willie

March 23, 2012

Abstract

This is a review of graphical models for representing humans and other articulated bodies. Models of both deformable and non-deformable objects allow us, given image and video data, to infer the pose and/or orientation of objects. We aim to either learn the structure and pose of objects, or assume a structure and learn the pose of objects, and incorporate this appearance model into the GPUDDPM detection and tracking scheme.

1 Data

We take our data to be the same as in the previous paper: frame difference (i.e. foreground, i.e. silhouette) pixel positions, and additional local appearance features

2 Sigal Thesis

This thesis (Sigal [2008]) involves articulated pose estimation and tracking using graphical models. Nodes of the graph correspond to parts or limbs of the body. Edges correspond to "kinematic, inter-penetration, and occlusion constraints imposed by the structure of the body and the imaging process". The model allows one to infer 3D pose from multiple synchronized views, or 2D pose from a single monocular image. Sigal also develops a hierarchical model where 2D pose is inferred from a monocular source, and then 3D pose is inferred given this 2D pose; finally, tracking is carried out in 3D. Inference is carried out using Particle Message Passing (PaMPas).

2.1 Chapter 4: graphical models for rigid objects

Sigal breaks down his framework into five components [p103]

1. a graphical model

2. an inference algorithm that provides the ability to infer a state of each node in the graph

3. a local evidence distribution (or image likelihood)

4. a proposal process for some or all nodes in a graphical model

5. a set of spatial and/or temporal constraints corresponding to the edges in a graph

Each object is represented by a undirected graphical model. The model for each object contains a set of four loosely connected regions; for example, a human is represented by head, left arm, right arm, and leg regions. Spatial constraints govern the relative positions of the regions for a given object type, and are modeled with a mixture of Gaussians. Sigal defines a likelihood, which models the probability of image data given one of the regions. Features for the likelihood for each region for a given object are found in a supervised manner, by using a boosted classifier.

2.2 Chapter 5: graphical models for loose-limbed objects

Each node represents the 3D position and orientation of a body part (in a 6D continuous domain). This approach also incorporates bottom-up information from object detectors, which in this case are used for detection of body parts (i.e. face, heads, and limbs). The probabilistic relationships between joints (nodes) in the loose-limbed graphical model are specified using mixture models learned from a database of motion capture sequences. The model also incorporates an image likelihood for each limb. The domain of the likelihood is foreground silhouette and edge features. Sigal presents two loose-limbed models for humans–a 10 node model, and a finer, 15 node model (which also has more dependencies).

Kinematic constraints refer to the geometry of and position between limbs. Kinematic constraints between limb positions [p123] and between limb orientations [p124] are modeled by mixtures of Gaussians. For limb orientations, this mixtures of Gaussians representation can be converted back into the rotation group, SO(3), space in which the orientation representation resides. These contraints are learned from ground truth motion capture sequences.

Penetration contraints refer to the inability for solid limbs to penetrate each other. Sigal only constrains limbs that are likely to come into contact in his model. Penetration constraints are learned computationally by making a model human out of basic shapes and and computing how much given parts have a tendency to intersect.

Sigal defines image likelihoods which give the probability of image data given the pose of a limb. A main component of the domain over which this likelihood is defined is binary foreground / silhouette. For each limb, he assumes three groups of pixels are present in the binary foreground image: those corresponding with a limb, those correlated with a limb, and those completely uncorrelated with

the limb; these groups are used to define his limb likelihood on foreground pixel positions. He also defines an edge likelihood (edge data gathered from Canny edge transformation of the image) that is a function of distances between edges. Note that this likelihood is non-clothes-specific.

Sigal also incorporates part detectors (such as head and limb detection), referred to as "shouters", which give noises guesses as to the positions of parts of the body.

Belief propagation is used for inference; in particular, Sigal uses a nonparametric algorithm called Particle Message Passing (PaMPas), which he also says is a generalization of a particle filter (that allows for inference over arbitrary graphs and not necessarily chains).

2.3 Chapter 6: hierarchical model for 3D pose from monocular source

Inferred 2D pose from monocular sources gives proposals in inference of 3D pose.

A mixture of experts model is used to learn a mapping from inferred 2D poses to 3D poses (including joint angles and foreshortening / perspective information). In particular, Sigal uses a mixture of regularized linear regression models that are trained from a set of 2D-3D pose pairs obtained from motion capture data.

The main difference between the 2D model used here and the loose-limbed model described in Section 2.2 is that Sigal develops "Occlusion-Sensitive Local Likelihoods", which incorporates whether foreground / edge pixels could be explained by other body parts in the image.

3 Sudderth Paper

In this paper (Sudderth et al. [2004]), Sudderth presents a three-dimensional geometric hand model that incorporates geometric contraints, has an image likelihood defined on color and edge data, and infers a model of the hand using a nonparametric inference algorithm (similar to that used by Sigal). Various information about hand geometry and results of edge detection on videos of hands are gained from training data. This model incorporates kinematic contraints between parts of the hand, hand dynamics, and occlusion constraints / variables.

4 Ying Wu Paper

In this paper (Wu et al. [2003]), articulated bodies are modeled and tracked via a dynamic markov random field. This paper begins by defining how general constraints between articulated parts are encoded. In the experimental section, Wu defines a 2D "cardboard" model where each subpart of the human body is represented as a planar object. For his motion model, there are dependencies

between adjacent time steps for each limb. The image observation associated with each limb is taken to be the detected edges of the shape contour of the limb. A sequential mean field monte carlo algorithm is used to perform inference in this model.

5 Plan of Attack

The papers reviewed here provide models that encode information about certain articulated bodies (humans, hands) and give strategies for inferring correct positions of these bodies from image data (either foreground pixels, edges, or colors). Dependencies in these papers were primarily defined using undirected graphical models.

For incorporation into the GPUDDPM, we need to define some articulated body likelihood structure with parameters $\theta_{k,t}$ for frame difference pixel positions $\mathbf{x}_{i,t}$, and the distributions

$$\begin{aligned} \mathbf{x}_{i,t} &\sim \mathrm{F}(\theta_{k,t}) \\ \theta_{k,t} &\sim \mathbb{G}_0(\theta_{k,t}) \\ P(\theta_{k,t}|\theta_{k,t-1}) & \end{aligned} \tag{1}$$

Additionally, it would be cool if we could

1. create a generative pose model to sample the pose of articulated bodies. I'm not sure if these undirected graphical model setups allow us to do this.

2. have a model which learns articulated structure from frame differencing data. Might be possible since we have an unsupervised system which can segment individual objects. It would probably be hard to capture specific dependencies / occlusion relations between limbs.

3. have a model that learns types of objects based on the different structures it infers.

References

Leonid Sigal. *Continuous-state graphical models for object localization, pose estimation and tracking*. PhD thesis, Providence, RI, USA, 2008. AAI3318361.

Erik B. Sudderth, Michael I. M, William T. Freeman, and Alan S. Willsky. Distributed occlusion reasoning for tracking with nonparametric belief propagation. In *In NIPS*, pages 1369–1376. MIT Press, 2004.

Ying Wu, Gang Hua, and Ting Yu. Tracking articulated body by dynamic markov network. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 1094 –1101 vol.2, oct. 2003. doi: 10.1109/ICCV.2003.1238471.