Bidirectional Model Learning for Visual Inference

Cristian Sminchisescu[†], Atul Kanaujia[‡] and Dimitris Metaxas[‡]
[†]TTI-C, *crismin@nagoya.uchicago.edu, ttic.uchicago.edu/~crismin*
[‡]Rutgers University, *{kanaujia,dnm}@cs.rutgers.edu,*
www.cs.rutgers.edu/~{kanaujia,dnm}

We present our ongoing work on visual processing that combines top-down and bottom-up processing in order to learn consistent visual models useful for recognition and synthesis. We will show applications to three-dimensional dynamic human pose reconstruction in video, but the methods apply more generally to other types of objects or scene factors.

The pose inference problem has been initially addressed using the machinery of top-down, *generative modeling*, with feedback provided within an image analysis-by-synthesis loop. Despite being a natural way to model the appearance of complex articulated structures, the success of generative models has been partly shadowed because it is computational demanding to infer the distribution on their hidden states (here human joint angles) and because their parameters are unknown and variable across many real scenes. This has motivated the emergence of bottom-up, feed-forward *recognition methods* that predict state distributions directly from image features [6, 7]. Despite their speed and simplicity, recognition methods tend to assume that the object of interest is segmented and could suffer from the lack of feedback. This makes them prone to hallucination.

We will describe one possible way to combine the strengths of both approaches using a bidirectional model with both recognition and generative sub-components. (A remarkable precursor with a different structure, cost function and properties can be found in [3]) The generative model predicts human like images represented as spatial distributions of local histograms of edge orientations [5]. In turn, the recognition model, based on conditional Bayesian mixtures of sparse experts, is trained to predict complex hidden state distributions from the same edge-based image descriptors. *Learning* the parameters of the bidirectional model alternates self-training stages in order to maximize the probability of the observed evidence (images of humans). During one step, the recognition model is trained to invert the generative model using samples drawn from it. In the next step, the generative model is trained to have a state distribution close to the one predicted by the recognition model. At local equilibrium, which is guaranteed within a variational approximation framework, the two models have consistent, registered parameterizations. We train on artificially generated human silhouettes obtained from a computer graphics human model, animated using human motion capture data, and rendered on backgrounds drawn from a database of natural indoor and outdoor images.

During *on-line inference*, the pose estimates are driven mostly by the fast recognition model but implicitly include generative feedback for consistency. The resulting 3d temporal pose predictor operates similarly to existing 2d object detectors [8]. It searches the image at different locations and uses the recognition model to hypothesize 3d configurations. Feedback from the generative model helps to downgrade incorrect competing 3d pose hypotheses and to decide on the detection status (human or not) at the analyzed image sub-window. This strategy provides a uniform treatment of object detection, pose initialization and tracking.

Topic: visual processing and pattern recognition
Preference: oral/poster

Figure 1: Examples of images included in our training database that consists of synthetically generated poses (from motion capture) rendered on natural indoor and outdoor image backgrounds. The rightmost plot shows masks used to generate partial occlusion within the detection window (regions shown in gray) views. We use only simple combinations restricted to intermediate horizontal and vertical positions, but more complex arrangements are possible.

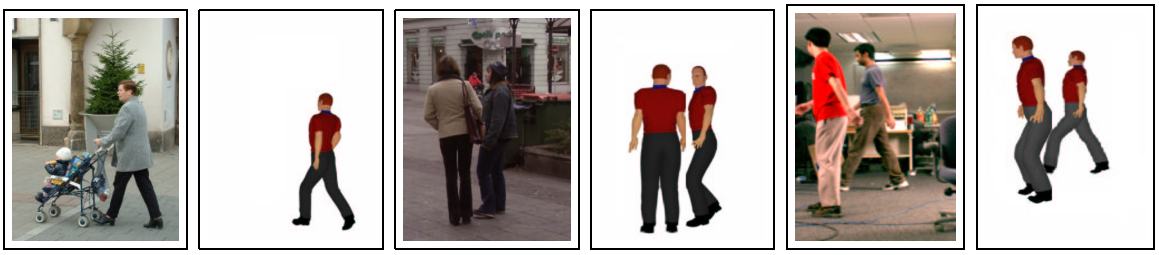

Figure 2: Several reconstructions in outdoor and indoor scenes. Notice the difficulty of the poses that involve self-occlusion, clutter and sometimes low limb contrast. The position of the arms is not always estimated precisely (especially for cases atypical of the training distribution), but often the overall reconstruction still captures some of the important qualitative aspects in the posture of the human subjects.

References

[1] C. Bishop and M. Svensen. Bayesian mixtures of experts. In *Uncertainty in Artificial Intelligence*, 2003.

[2] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2005.

[3] G. Hinton, P. Dayan, B. Frey, and R. Neal. The wake-sleep algorithm for unsupervised neural networks. *Science*, 1995.

[4] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proc. of IEEE*, 1998.

[5] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2), 2004.

[6] R. Rosales and S. Sclaroff. Learning Body Pose Via Specialized Maps. In *Advances in Neural Information Processing Systems*, 2002.

[7] C. Sminchisescu, A. Kanaujia, Z. Li, and D. Metaxas. Discriminative Density Propagation for 3D Human Motion Estimation. In *IEEE International Conference on Computer Vision and Pattern Recognition*, volume 1, pages 390–397, 2005.

[8] P. Viola, M. Jones, and D. Snow. Detecting Pedestrians using Patterns of Motion and Appearance. In *IEEE International Conference on Computer Vision*, 2003.