Fast Online Upper Body Pose Estimation from Video

Ming-Ching Chang[1,2]
changm@ge.com

Honggang Qi[1,3]
hgqi@ucas.ac.cn

Xin Wang[1]
xwang26@albany.edu

Hong Cheng[4]
hcheng@uestc.edu.cn

Siwei Lyu[1]
slyu@albany.edu

[1] Computer Science Department
University at Albany, State University of New York
Albany, USA

[2] Computer Vision Lab
GE Global Research Center
Niskayuna, USA

[3] University of Chinese Academy of Sciences
Beijing, China

[4] Center for Robotics
University of Electronic Science and Technology of China
Chengdu, China

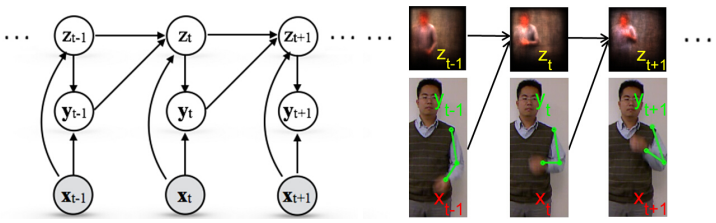

Figure 1: *(**Left**) The CDBN model structure and (**Right**) an example of the CDBN-MODEC model applied to online video pose estimation. Variable* $\mathbf{x}_t$ *corresponds to observations at time* t, i.e., *image features in individual video frames;* $\mathbf{y}_t$ *corresponds to poses,* i.e., *joint locations, and* z_t *is the latent pose modes. See text for detailed explanations.*

Estimation of human body poses, represented as the ensemble of joint locations, is an important problem in computer vision. In this work, we describe a fast *online* method for video pose estimation. We aim to extend existing single frame methods for *online* use, where latent pose modes (or "poselets") can be directly leveraged to improve motion consistency, in a paradigm similar to detect-and-track for object tracking.

Our method is based on a general *conditional dynamic Bayesian network* (CDBN) model, which is a combination of two widely used probabilistic graphical models, namely the dynamic Bayesian network (DBN) [3] and conditional random field (CRF) [2]. The DBN aspect of our model captures the temporal correlations between variables in a sequence, and the CRF aspect incorporates the complex relations between the observations and latent variables.

Specifically, we consider the following dynamic model that involves three time series variables: (1) an input sequence of observation variables $\mathbf{x}_{0:t}$, (2) an output temporal sequence of latent state variables $\mathbf{y}_{0:t}$, and (3) the sequence of latent mode selection variables $z_{0:t}$ with $z_t \in \{1,\cdots,M\}$ indicating one of the M input/output relation modes is active at time t. The CDBN model represents the dependencies of these variables with a dynamic probabilistic model, which corresponds to a factorization of probability distribution $p(z_{0:t},\mathbf{y}_{0:t}|\mathbf{x}_{0:t})$ according to the graphical structure in Fig.1(a), as:

$$p(z_{0:t},\mathbf{y}_{0:t}|\mathbf{x}_{0:t}) = p(z_0|\mathbf{x}_0)\prod_{\tau=0}^{t} p(\mathbf{y}_\tau|z_\tau,\mathbf{x}_\tau) \times \prod_{\tau=0}^{t-1} p(z_{\tau+1}|z_\tau,\mathbf{y}_\tau,\mathbf{x}_{\tau+1}). \quad (1)$$

The joint model in Eq.(1) can be used for dynamic Bayesian inference, implemented with particle filtering.

A key characteristic of the CDBN model is that it is an "open architecture", as it can incorporate different underlying CRF models (including future ones) into the DBN structure. When applying to *online* video pose estimation, this becomes an advantage, as it allows the resulting algorithm to incorporate effective single frame pose estimation method into the dynamic framework that models intra-frame correlations. In this work, we adopt part of the efficient CRF pipeline from the MODEC single pose estimation of Sapp and Taskar [4] as the CRF model in our CDBN framework. We term our method CDBN-MODEC for video pose estimation.

Fig.1 illustrates the structure of CDBN as a graphical model. To better evaluate *online* pose estimation from practical video streams, we also create a new high frame rate labeled dataset that calls for run time efficiency. When evaluated on this dataset and the VideoPose2 benchmark dataset, CDBN-MODEC achieves considerable improvements in both performance and efficiency over several state-of-the-art *online* video pose estimation methods. Qualitative results of using CDBN-MODEC estimating poses in several frames of videos from the HFR and VideoPose2 datasets are presented in Fig.2. In Fig.3 we compare the running time versus accuracy of MODEC [4], MODEC+S [5], CDBN-MODEC, and Cherian *et.al.* [1].

Figure 2: *Estimated upper body poses from frames of two videos in the HFR (top) and VideoPose2 (bottom) using CDBN-MODEC.*

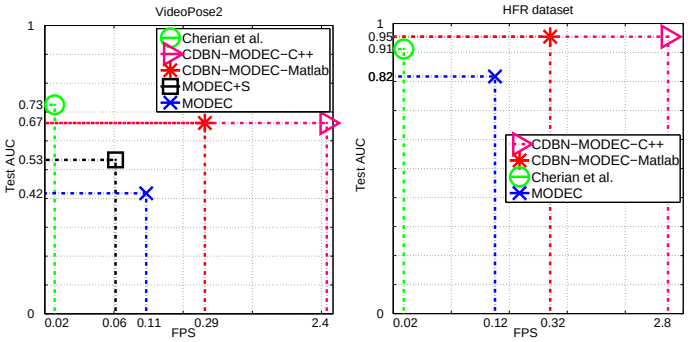

Figure 3: *Comparison of running time versus accuracy. See text for details.*

[1] A. Cherian, J. Mairal, K. Alahari, and C. Schmid. Mixing Body-Part Sequences for Human Pose Estimation. In *Computer Vision and Pattern Recognition (CVPR)*, pages 2361–2368, 2014.

[2] John Lafferty. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *International Conference on Machine Learning (ICML)*, pages 282–289, 2001.

[3] Kevin Patrick Murphy. Dynamic Bayesian Networks: Representation, Inference and Learning. In *PhD Thesis*, 2002.

[4] Benjamin Sapp and Ben Taskar. Multimodal Decomposable Models for Human Pose Estimation. In *Computer Vision and Pattern Recognition (CVPR)*, pages 3674–3681, 2013.

[5] D. Weiss, B. Sapp, and B. Taskar. Dynamic Structured Model Selection. In *International Conference on Computer Vision (ICCV)*, pages 2656–2663, 2013.