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

Abstract

Estimation of human body poses from video is an important problem in computer vision with many applications. Most existing methods for video pose estimation are *offline* in nature, where all frames in the video are used in the process to estimate the body pose in each frame. In this work, we describe a fast *online* video upper body pose estimation method (CDBN-MODEC) that is based on a conditional dynamic Bayesian network model, which predicts upper body pose in a frame without using information from future frames. Our method combines fast single image based pose estimation methods with the temporal correlation of poses between frames. We collect a new high frame rate upper body pose dataset that better reflects practical scenarios calling for fast *online* video pose estimation. When evaluated on this dataset and the VideoPose2 benchmark dataset, CDBN-MODEC achieves improvements in both performance and running efficiency over several state-of-art *online* video pose estimation methods.

1 Introduction

Estimation of human body poses, represented as the ensemble of joint locations, is an important problem in computer vision. As the basis for understanding human actions and behaviors from visual imagery, it has many applications, including gesture recognition, human computer interaction, gaming, sign language recognition, and the study of affective states and social behaviors. With the ubiquity of inexpensive video cameras on mobile devices and laptop computers, it has become increasingly easy to capture live feed videos, from which

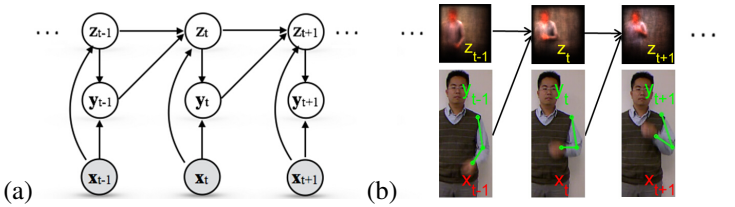

Figure 1: *(**Left**) The CDBN model structure and (**Right**) an example of the CDBN-MODEC model applied to online video pose estimation. Variable* $\mathbf{x}_t$ *corresponds to observations at time t,* i.e.*, image features in individual video frames;* $\mathbf{y}_t$ *corresponds to poses,* i.e.*,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

human poses can be estimated as a continuous time series for further processing. Many practical applications require *online* pose estimation, where the results are obtained from incoming frames without using information from future frames.

With the maturity of efficient single image based human body pose estimation methods [5, 7, 11, 16, 17, 19, 21], one simple solution is to apply such methods to individual frames of a video as if they are independent images. This approach works to certain extent, however, temporal correlations of postures in consecutive frames and the assumption of smooth action in video frames provide strong cues for robust estimations of poses through tracking and prediction. So treating individual frames without considering such temporal correlations usually leads to inefficient algorithm and inaccurate estimations, due to ambiguities and occlusions in a single frame. In contrast, estimating poses from multiple frames in a video provides a better means to handle occlusions and improve robustness of the estimation.

Pose estimation from videos (and particularly the one focusing on upper body postures) has advanced significantly in recent years *e.g.* [3, 8, 9, 15, 18, 20, 22]. However, the majority of these methods are *offline* in nature, *i.e.*, body poses in a frame are inferred from both its past and future frames. Furthermore, the performance of these methods in estimating poses usually comes at the price of complicated inference procedures, which significantly reduce running efficiency. As such, they are not applicable in tasks requiring fast *online* video processing.

In this work, we describe a fast *online* method for upper body pose estimation from videos. [1] 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) [14] and conditional random field (CRF) [12]. 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. Fast *online* estimation are achieved with an efficient particle filtering implementation of the inference.

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

[1] Note that our approach is generic to handle full body, if the underlying per frame detection is full body.

[16] 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.

2 Related Works

Pose Estimation from a Single Image. There is an extensive literature on pose estimation from single images (*e.g.* [2, 5, 7, 11, 16, 17, 19, 21]). Yang and Ramanan [21] introduced a Flexible Mixture-of-Parts model for human pose recognition. This method allows parts to be selected from several types and jointly learns the Deformable Parts Model (DPM) [6] parameters in a tree-based structured model. Sapp and Taskar's *multimodal decomposable models* (MODEC) [16] introduced multimodality at the coarse-body and fine-part (shoulder, elbow and wrist) granularities. They divide the upper body pose into two half-side bodies, and for each side estimate pose modes, which can guide the tracking of arm and joint locations (similar to the DPM scheme) in a single-path tree structure. Such decomposable model achieves improvements on both accuracy and speed. Given a torso bounding box as an input, shoulder positions are estimated more accurately than the elbows and wrists.

***Offline* Video Pose Estimation**. Significant efforts have been invested on the estimation of human pose from videos [3, 8, 9, 15, 18, 20, 22], rather than single images. Motion flow cues and features are key in this category of methods [3, 9, 22]. Many works exploit optical flow for pose estimation. Ferrari *et.al*. [9] first use segmentation to aid the detection of body pose in each frame, and then calculate the motion of lower limbs across frames. The "flow puppets" of Zuffi *et.al*. [22] use a 2D upper-body shape model to track articulated motions, where motion cues are integrated jointly with pose inference. Cherian *et.al*.'s *offline* method [3] consists of two parts. The first part extends the work of [21] by adding motion flow links at the pose inference step between consecutive frames. Across-frame links between elbows and wrists introduce loops in the inference, thus loopy belief propagation (BP) is introduced as a solver. Secondly, candidate poses are decomposed into the modeling of individual limbs, where these limbs are recomposed after temporal smoothing. This method achieves high accuracy among *offline* pose estimation methods (see Fig.2 and 4). However its computation remains burdensome due to the extensive use of dense optical flow and loopy belief propagation inference.

***Online* Video Pose Estimation**. There are relatively fewer works addressing the problem of continuous pose estimation from *online* video streams. Lim *et.al*. [13] proposed an *online* algorithm to jointly segment a person from the background and estimate the upper body pose from a video. Weiss *et.al*. [20] presented a Dynamic Structured Model Selection method based on their MODEC model that uses meta features in structured learning to automatically determine models to choose for inference. Jain *et.al*. [10] used a convolutional neural network to incorporate both color and motion features for video pose estimation. In general, existing methods in this category have difficulties processing real-time video streams due to the extensive use of features such as dense optical flow. To our best knowledge, the development of real-time *online* video pose estimation method is still an open problem, and the

demand of such development continuous to grow.

3 Method

We first describe the CDBN model (§3.1) and its inference (§3.2), where the framework is general and not limited to pose estimation. We then combine CDBN with the MODEC method for *online* video pose estimation (§3.3).

3.1 The CDBN Model

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:

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

The joint model in Eq.(1) can be used for dynamic Bayesian inference. But to further simply the model, we make the following assumption

$$p(z_{t+1}|z_t, \mathbf{y}_t, \mathbf{x}_{t+1}) \propto p(z_{t+1}|\mathbf{x}_{t+1}) \cdot p(z_{t+1}|z_t) \cdot p(z_{t+1}|\mathbf{y}_t). \tag{2}$$

Note that this condition is *different* from the typical assumption in DBN that $\mathbf{x}_{t+1}$ and $(\mathbf{y}_t, z_t)$ is conditionally independent given z_{t+1}[2], but it makes the subsequent computations much easier.

CDBN can be regarded as a conditional random field (CRF), but the output and latent mode variables $\mathbf{y}_t$ and z_t are conditioned on the input observation $\mathbf{x}_t$ from a dynamic Bayesian network (DBN). As such, it is specified with the following conditional distributions given the observations, as in Eqs.(1) and (2), which corresponds to the four arcs in Fig.1(a):

- **state estimation** $p(\mathbf{y}_t|z_t, \mathbf{x}_t)$: conditional probability distribution of current state given current observation and mode;
- **observation-mode estimation** $p(z_t|\mathbf{x}_t)$: conditional probability distribution of current mode given current observable;
- **mode-mode transition estimation** $p(z_{t+1}|z_t)$: conditional probability distribution of next mode given current mode;
- **state-mode estimation** $p(z_{t+1}|\mathbf{y}_t)$: conditional probability distribution of next mode given current state.

The four conditional distributions can be grouped into two categories. Conditional distributions $p(\mathbf{y}_t|z_t, \mathbf{x}_t)$ and $p(z_t|\mathbf{x}_t)$ concern inference using variables of the same time index, as such they form the *inference* module. On the other hand, conditional distributions $p(z_{t+1}|z_t)$ and $p(z_{t+1}|\mathbf{y}_t)$ describe correlations of variables in consecutive time steps, and they form the *dynamic* module in the CDBN model.

[2]Such will be equivalent to $p(z_t, \mathbf{y}_t, \mathbf{x}_{t+1}|z_{t+1}) = p(\mathbf{x}_{t+1}|z_{t+1})p(z_t|z_{t+1})p(\mathbf{y}_t|z_{t+1})$, which cannot be deduced from Eq.(2).

3.2 Inference of the CDBN Model

Inference in the CDBN model corresponds to the computation of posterior distribution of the output given the input $p(\mathbf{y}_{t+1}|\mathbf{x}_{0:t+1})$, which is obtained as

$$p(\mathbf{y}_{t+1}|\mathbf{x}_{0:t+1}) = \sum_{z_{t+1}=1}^{M} p(z_{t+1}, \mathbf{y}_{t+1}|\mathbf{x}_{0:t+1}). \tag{3}$$

We can further expand $p(z_{t+1}, \mathbf{y}_{t+1}|\mathbf{x}_{0:t+1})$ using the joint distribution given in Eq.(1), as

$$p(z_{t+1}, \mathbf{y}_{t+1}|\mathbf{x}_{0:t+1}) = p(\mathbf{y}_{t+1}|z_{t+1}, \mathbf{x}_{t+1}) p(z_{t+1}|\mathbf{x}_{0:t+1}), \tag{4}$$

where the Markovian properties assumed in the joint model are used. The first term in Eq.(4) corresponds to the estimation of the output variable given the current input $\mathbf{x}_{t+1}$ and latent variable z_{t+1}. The second term is mode estimation from the input $\mathbf{x}_{0:t+1}$, which can be further expanded using Eq.(2), as:

$$\begin{aligned} p(z_{t+1}|\mathbf{x}_{0:t}) &= \sum_{z_t} \int_{\mathbf{y}_t} p(z_{t+1}, z_t, \mathbf{y}_t|\mathbf{x}_{0:t}) d\mathbf{y}_t = \sum_{z_t} \int_{\mathbf{y}_t} p(z_{t+1}|z_t, \mathbf{y}_t, \mathbf{x}_{t+1}) p(z_t, \mathbf{y}_t|\mathbf{x}_{0:t}) d\mathbf{y}_t. \\ &= \underbrace{p(z_{t+1}|\mathbf{x}_{t+1})}_{\text{observation-mode est.}} \cdot \sum_{z_t} \underbrace{p(z_{t+1}|z_t)}_{\text{mode-mode est.}} \cdot \int_{\mathbf{y}_t} \underbrace{p(z_{t+1}|\mathbf{y}_t)}_{\text{pose-mode est.}} \cdot \underbrace{p(z_t, \mathbf{y}_t|\mathbf{x}_{0:t})}_{\text{previous posterior}} d\mathbf{y}_t. \end{aligned} \tag{5}$$

Eqs.(4)-(5) provide the recursive Chapman-Kolmogorov update of the posterior distribution $p(z_{t+1}, \mathbf{y}_{t+1}|\mathbf{x}_{0:t+1})$, from which we can build the dynamic inference algorithm for CDBN. However, a straightforward implementation of the dynamic update of CDBN is challenging due to the need to integrate over the space of output variable $\mathbf{y}_t$ in Eq.(5). This step usually does not afford a closed form efficient numerical procedure. Instead, we solve this by adopting a particle filter approach, where the posterior distribution of $p(z_t, \mathbf{y}_t|\mathbf{x}_{0:t})$ is approximated with weighted samples of $\mathbf{y}_t$ (*i.e.*, the particles).

In principle, we should use multiple particles from $p(z_t, \mathbf{y}_t|\mathbf{x}_{0:t})$. But in this work, we use a simpler particle generation scheme to achieve maximum running efficiency. Specifically, we use only one particle $\phi(z_t)$ to represent the continuous output variable $\mathbf{y}_t$ for each value of the latent mode z_t with unnormalized weight $\psi(z_t)$:

$$\phi(z_t) = \text{argmax}_{\mathbf{y}_t} p(z_t, \mathbf{y}_t|\mathbf{x}_{0:t}) = \text{argmax}_{\mathbf{y}_t} p(\mathbf{y}_t|z_t, \mathbf{x}_t), \tag{6}$$

$$\psi(z_t) = \max_{\mathbf{y}_t} p(z_t, \mathbf{y}_t|\mathbf{x}_{0:t}) = p(z_t, \phi(z_t)|\mathbf{x}_{0:t}). \tag{7}$$

That is to say, for each possible value of the mode variable z_t, we represent the posterior distribution $p(\mathbf{y}_t|z_t, \mathbf{x}_{0:t})$ with a particle-weight pair as $(\phi(z_t), \psi(z_t))$ corresponding to the mode of $p(\mathbf{y}_t|z_t, \mathbf{x}_{0:t})$. Using the particle filter approach, Eq.(5) is approximately computed with

$$p(z_{t+1}|\mathbf{x}_{0:t+1}) \approx p(z_{t+1}|\mathbf{x}_{t+1}) \sum_{z_t} p(z_{t+1}|z_t) p(z_{t+1}|\phi(z_t)) \frac{\psi(z_t)}{\sum_{z'} \psi(z')}, \tag{8}$$

which is then combined with Eq.(4) to recursively find $\phi(z_{t+1})$ and $\psi(z_{t+1})$. From the posterior distribution, we use $\text{argmax}_{\mathbf{y}_{t+1}} p(\mathbf{y}_{t+1}|\mathbf{x}_{0:t+1})$ to obtain the optimal estimation of the output variable. The dynamic inference algorithm for CDBN is summarized as:

- Compute $p(z_{t+1}|\mathbf{x}_{0:t+1})$ from Eq.(8) using particles with weights at time step t.
- For each pose mode z_{t+1}, generate a new particle with a weight using Eqs.(6)-(7).
- Compute $p(\mathbf{y}_{t+1}|z_{t+1}, \mathbf{x}_{t+1})$ for the output state $\mathbf{y}_{t+1}$.
- Move on to the next frame $t \leftarrow t+1$.

3.3 CDBN-MODEC for *Online Video Pose Estimation*

We apply CDBN to *online* video pose estimation, where the observation variable $\mathbf{x}_t$ and the latent state variable $\mathbf{y}_t$ correspond to image features and the locations of joints (*i.e.*, shoulder, elbow and wrist) of a person in each frame of the video, respectively. The latent variable z_t corresponds to the "pose mode" or "poselet" [1] that clusters similar poses into groups, Fig.1(b). To implement a CDBN based *online* video pose estimation method, we need to specify the *inference* module, *i.e.*, conditional distributions $p(\mathbf{y}_t|z_t,\mathbf{x}_t)$ and $p(z_t|\mathbf{x}_t)$, and the *dynamic* module, *i.e.*, conditional distributions $p(z_{t+1}|z_t)$ and $p(z_{t+1}|\mathbf{y}_t)$. Because the inference module only relies on variables from a single frame, in principle, we can use any single image pose estimation method that also clusters poses into explicit "modes". In this work, we choose the MODEC model [16] as the basis for the inference module, where $p(z_t|\mathbf{x}_t)$ is computed from matching HOG pyramid features, and $p(\mathbf{y}_t|z_t,\mathbf{x}_t)$ comes from pose estimation using the CRF of the MODEC method. The two conditional distributions $p(z_{t+1}|z_t)$ and $p(z_{t+1}|\mathbf{y}_t)$ in the dynamic module are determined from a machine learning approach using labeled training data.

Revisit MODEC Pose Estimation. To estimate upper body pose from an input image, MODEC first runs torso detection [1] to determine the locations of the upper body of a person in the video. In practice, we found that it is more effective to locate torso by first using a standard face detector and then scale the bounding box of the detected face to determine the bounding box of the person's torso. After that, MODEC computes two sets of HOG feature pyramids in order to employs two cascaded classification steps. In the first step, pose mode (or cluster) probabilities are estimated based on the coarse HOG pyramid (*i.e.* the side model) [4] of the image. For running efficiency, MODEC only models poses of the left half of human upper body, and right body poses are obtained by mirroring the image and treated as left body poses. From training images, MODEC clusters left body poses into 32 pose modes, and chooses the best pose mode as the one with the highest likelihood to represent the coarse HOG side model features. With the mode determined, the second step of MODEC is to estimate the actual poses of the left half body (*i.e.*, locations of limbs and joints) using the other fine-grained HOG feature pyramid (*i.e.* the parts model). The inference using a single-path tree-based CRF is efficient and allows for fast and parallel implementations. [3]

Both steps of the original MODEC method need modifications in order to be incorporated into our CDBN framework. First, instead of only retaining the most likely mode in the first step, we keep all modes with likelihood score above a threshold θ_1, which is a fixed parameter in the original MODEC but can be dynamically adjusted in our method. The likelihood scores of these modes, after normalization to sum to one, model the conditional distribution for the observation-mode estimation, *i.e.*, $p(z_t|\mathbf{x}_t)$. For the second step, instead of just returning the most likely pose, we use the CRF score after normalization to construct the conditional distribution for pose estimation, *i.e.* $p(\mathbf{y}_t|z_t,\mathbf{x}_t)$ in CDBN. These modifications better ensure motion consistency of the estimations in our framework.

Dynamic Mode Prediction. The two conditional distributions $p(z_{t+1}|z_t)$ and $p(z_{t+1}|\mathbf{y}_t)$ of the *dynamic* module are obtained from a learning approach using labeled videos. While ground truth poses are obtained by manual labeling in each frame, we define ground truth modes as the most likely mode interpreting the labeled poses. The ground truth modes can be derived automatically using MODEC mode selection. Specifically, for each frame with

[3] All parameters in the MODEC method are taken as their default value from the publicly available MATLAB code http://vision.grasp.upenn.edu/cgi-bin/index.php?n=VideoLearning.MODEC.

labeled pose $\hat{\mathbf{y}}_t$, we first apply MODEC to obtain the optimal poses corresponding to each of the 32 pose modes, which we denote as $\tilde{\mathbf{y}}^{(i)}$ for $i = 1, \cdots, 32$. Then the ground truth mode for frame t is given as $\hat{z}_t = \text{argmin}_i \|\tilde{\mathbf{y}}_t - \hat{\mathbf{y}}^{(i)}\|_2$, where $\|.\|_2$ denotes the ℓ_2 norm. From the estimated pose modes, we obtain $p(z_{t+1}|z_t)$ as the frequencies of mode transitions across all labeled frames as: $p(z_{t+1} = i|z_t = j) = \frac{\#(\hat{z}_{t+1}=i, \hat{z}_t=j)}{\#(z_t=j)}$, $i, j = 1, \cdots, 32$. In practice, $p(z_{t+1}|z_t)$ is stored as an $M \times M$ matrix.

We learn $p(z_{t+1} = i|\mathbf{y}_t)$ with a simple voting scheme. Specifically, for each mode j at time t, we consider only modes i such that $p(z_{t+1} = i|z_t = j) > 0$. We again use MODEC to generate pose $\tilde{\mathbf{y}}$ for each such mode i. The value of $p(z_{t+1} = i|\mathbf{y}_t)$ is the normalized product of three zero mean Gaussian distributions: (1) the ℓ_2 distance between $\tilde{\mathbf{y}}$ and $\mathbf{y}_t$, (2) the difference between the *scales* of $\tilde{\mathbf{y}}$ and $\mathbf{y}_t$ as come from the resulting layers from the MODEC fine-grained HOG pyramid, and (3) the difference between the joint angles of $\tilde{\mathbf{y}}$ and $\mathbf{y}_t$. Parameters of the CDBN-MODEC include the standard deviations of the above three Gaussian distributions, $\theta_2 = 50$ pixels, $\theta_3 = 2$, and $\theta_4 = 40^\circ$, which are fixed in the experiments reported in the next section.

4 Experimental Results

We evaluate CDBN-MODEC and compare its performance and efficiency with several state-of-the-art video pose estimation approaches, including both *online* and *offline* methods. In general, *offline* methods perform better than *online* ones, as they can use information in all available video frames. In comparison, *online* methods generally run faster because they only process incoming frames as they are captured.

Dataset. We evaluate the performance of CDBN-MODEC using two video pose datasets. [4]

- **VideoPose2** [17] consists of video clips from popular TV shows *Friends* and *Lost*. Every other frames of the original video sequences are selected in this dataset, resulting in videos with an average frame rate of 10 FPS. There are 44 clips of 2-3 seconds in length, with a total of 1,286 frames. To be able to compare with existing works as in [20], we use 26 clips to train the mode transition model (Section 3.3), and report performance on the remaining 18 clips.
- **HFR.** We collect a new high frame rate (HFR) upper-body pose dataset using Microsoft Kinect camera. [5] This dataset consists of subjects performing a class of actions (*e.g.*, hand-on-hip, touching face, arm crossing) that are suitable for behavior recognition. There are 18 clips of 30 FPS with per-frame poses manually labeled as the position of head, shoulder, elbow, and wrist. We use 12 clips to train the mode transition model (Section 3.3) and report performance on the remaining 6 clips.

Evaluation Metric. We use *percentage of predicted parts* (PCP) [16] to evaluate the accuracy of pose estimation. PCP measures of the percentage of N frames that have estimated poses close to the labeled poses, which is defined as:

$$\text{PCP}_i(r) = \frac{100}{N} \sum_{t=1}^{n} \mathbf{1}\left(\frac{\|\tilde{\mathbf{y}}_i^t - \mathbf{y}_i^t\|_2}{h_t/100} \leq r \right). \tag{9}$$

[4] We omit the evaluation on the Pose-in-the-Wild dataset published in [3], due to its low frame rate and the missing training set for evaluating the method of [3] on it. See the comparison of [3] to our method in the experiments.

[5] We will make public the HFR dataset and results with this paper. Depth information is available however not used in this paper.

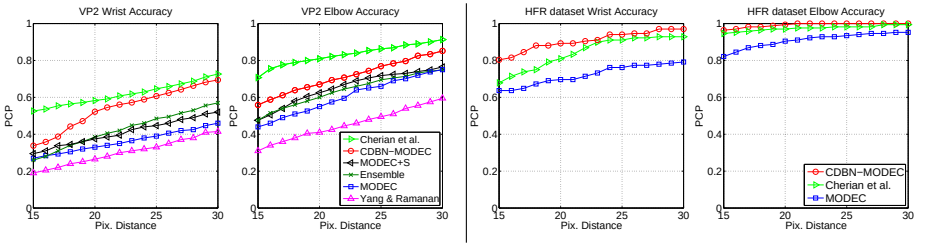

Figure 2: *(Left) Comparison of CDBN-MODEC with several state-of-the-art of online pose estimation methods on the VideoPose2 dataset including Ensemble [18], MODEC [16], MODEC+S [20], Yang & Ramanan [21] and a state-of-the-art offline approach from Cherian et al. [3]. (Right) Comparison of CDBN-MODEC with MODEC [16] and Cherian et al. [3] on the HFR dataset.*

In PCP, index i indicates joints, *i.e.* elbow and wrist; $\tilde{\mathbf{y}}_i^t$ is the location of the corresponding joints of labeled poses at frame t; $\mathbf{y}_i^t$ is the estimated locations of the joints. The error between ground truth and estimated poses, $\|\tilde{\mathbf{y}}_i^t - \mathbf{y}_i^t\|_2$ is normalized by the height of the bounding box of the detected torso h_t, and then rescaled to 100 pixels. We increase the count at a frame if the normalized error is less than r, using the indicator function $\mathbf{1}(*)$.

On the VideoPose2 dataset, we compare CDBN-MODEC with mainstream pose estimation methods in all categories: (1) single image pose estimation methods including MODEC [16] and Yang & Ramanan [21], (2) *online* video pose methods including MODEC+S [20], and (3) *offline* method including the Ensemble [18] and Cherian *et.al.* [3].

On the HFR dataset, we compare CDBN-MODEC with (1) MODEC [16] applied to individual frames of the video and (2) the *offline* method of Cherian *et.al.* [3]. The two methods with public available code are selected because MODEC represents best running efficiency of existing methods, and Cherian *et.al.* [3] is the state-of-art *offline* method.

Results. To provide a more comprehensive performance metric, we report the $\text{PCP}_i(r)$ for all experiments with r varying from 15 to 30 presented as a plotting curve in Fig.2. The area under this curve (AUC) is reported as a measure of overall performance of each method. On the VideoPose2 dataset, CDBN-MODEC outperforms all other *online* and single image methods [16, 18, 20, 21], and is slightly inferior to the *offline* method of Cherian *et.al.* To ensure fair comparison, we use the same training/testing dataset as in [20], where 26 clips are used for training and the remaining 18 clips for testing. The scores for four of these methods [16, 18, 20, 21] reported in [20] are used for comparison. For the *offline* method of Cherian *et.al.* [3], we report the score using their publicly available code [6] on the VideoPose2 dataset. Our performance gain is due to the effective modeling of both the between-frame and mode-to-mode temporal correlations.

On the HFR dataset, our CDBN-MODEC as an *online* method slightly outperforms the *offline* state-of-the-art method of Cherian *et.al.* [3], where our method gains considerable acceleration in running time (see Fig.4). This is mainly because the HFR dataset include videos of high frame rates. There are frames which are roughly identical, which is not the case in VideoPose2, where only keyframes with significant motions are retained. As the method of [3] relies on optical flow in inference, these static frames poses problems as they lead to weak optical flows. On the other hand, CDBN-MODEC does not directly use optical flow. This explains both of its robustness and improvements in running time.

Some qualitative results on the HFR dataset are presented in Fig.3. The three rows corresponding to pose estimation from two video frames of two different videos using frame-wise

[6] https://lear.inrialpes.fr/research/posesinthewild/

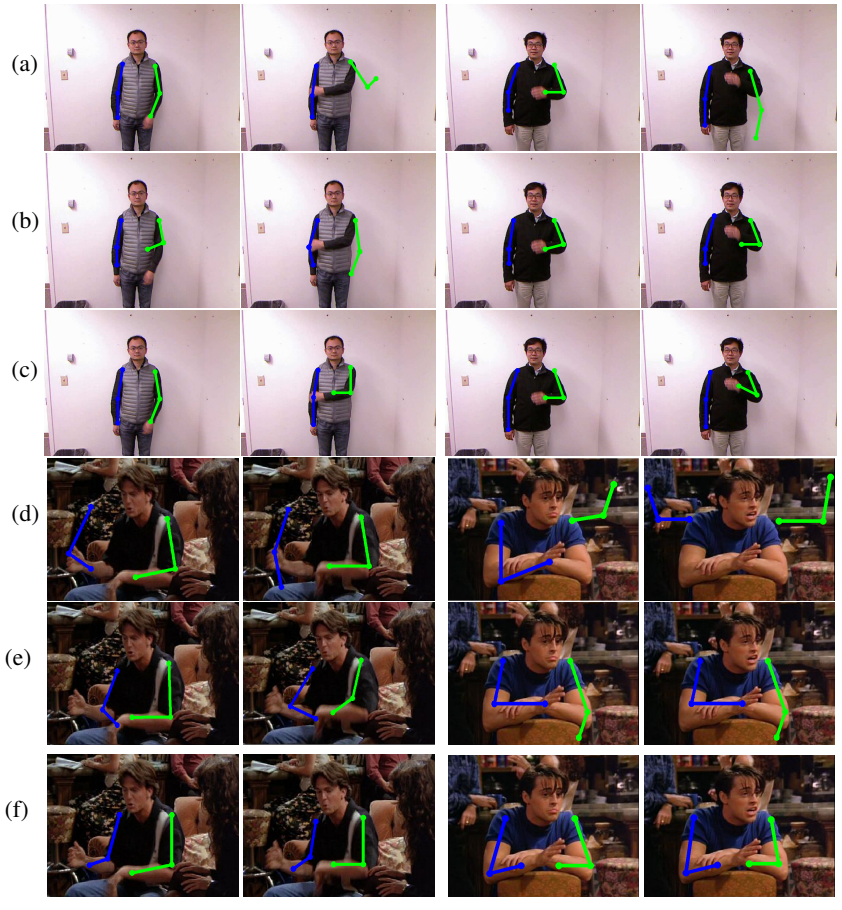

Figure 3: *Estimated upper body poses from frames of two videos in the HFR (a,b,c) and VideoPose2 (d,e,f) datasets using the MODEC method [17] applied to individual frame (***a,d***), the method of Cherian* et.al. *[3] (***b,e***), and CDBN-MODEC (***c,f***). Details are given in the text.*

MODEC, the *offline* video pose estimation method of [3], and CDBN-MODEC, respectively. These results demonstrate differences in effectiveness of these methods. MODEC applied to individual frames cannot take advantage of temporal correlations between frames. As such, there are gross errors in the examples. The *offline* pose estimation method of [3] is more robust, because it detects the whole upper body (i.e., head, torso, and arms) as well as the pose as they are fitted in a single model. In contrast, our CDBN-MODEC achieves more reliable estimations using accurate modes.

Running Time. In Fig.4 we compare the running time versus accuracy of MODEC [17], MODEC+S [20], CDBN-MODEC, and Cherian *et.al*. [3]. [7] We compare two implementations of our CDBN-MODEC: one in MATLAB and the other in C++. The C++ implementation contains further optimization using (1) multi-thread programming and (2) adapting levels of HOG pyramids in the inference. Our C++ implementation generally takes 320ms for a pose estimation after parallelization, where the computation of HOG pyramids from [5] takes about 200ms, and the rest of the steps including feature convolution in the refined

[7] The comparison against Ensemble [13] and Yang & Ramanan [21] is omitted, since these two methods run slower than MODEC [17] with inferior performance.

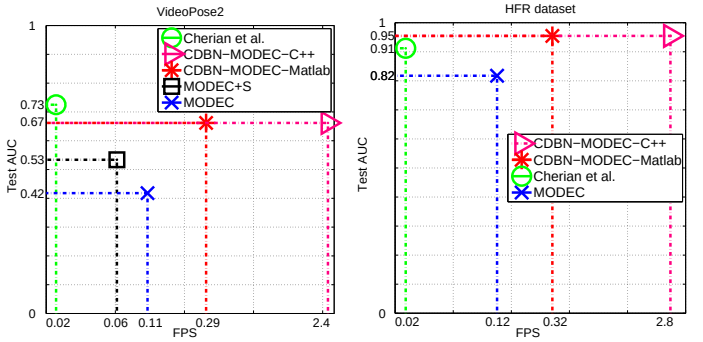

Figure 4: *Comparison of running time versus accuracy. See text for details.*

HOG layers, inference, back tracking, and the CDBN filtering take only about 120ms. All reported running time of methods that we have source code (*i.e.*, CDBN-MODEC, MODEC, and Cherian *et.al.*) are based on a machine with a 3.4GHz processor with 8 cores and 8GB memory. The running time of other methods are taken from [20], which is based on a machine with a 3.0GHz processor with 16 cores. The accuracy is measured with the average AUC values of elbow and wrist curve on each of the two test datasets. Note that the C++ implementation of CDBN-MODEC achieves a 10-fold acceleration in running time with no performance loss comparing to the MATLAB implementation, while the MATLAB implementation is already in general significantly faster than other compared methods, with improved or comparable estimation accuracy.

5 Conclusion

In this work, we describe a fast online pose estimation method based on the CDBN-MODEC model. The proposed algorithm presents outstanding pose estimation performance on both accuracy and running speed by two complementary system of Conditional Dynamic Bayesian Network and Multimodal Decomposable Model. We collect a new high frame rate upper body pose dataset that better reflects practical scenarios calling for fast *online* video pose estimation. When evaluated on this dataset and the VideoPose2 benchmark dataset, our method outperforms the state-of-the-art *online* methods on VideoPose2 datasets and show comparable performance to the state-of-the-art *offline* methods on the HFR dataset with significant running time acceleration.

There are a few directions we would like to further improve the current work. First, we can further take advantage of the flexibility of the CDBN model and combine it with more effective single image pose estimation method, such as those based on deep neural networks [2, 19]. Second, in our current method, we separate the effect of pose $\mathbf{y}_t$ and mode z_t in the pose mode prediction. We believe more accurate prediction can be obtained by considering them jointly and use a learned predictor from labeled data.

Acknowledgement. This project is partially supported by the following project: US NSF CAREER Award IIS-0953373 and US NSF Research Grant CCF-1319800 (X. Wang and S. Lyu) and NSF of China Research Grant No. 61472388 (H. Qi).

References

[1] Lubomir Bourdev and Jitendra Malik. Poselets: Body Part detectors trained using 3d human pose annotations. In *International Conference on Computer Vision (ICCV)*, pages 1365–1372, 2009.

[2] Xianjie Chen and Alan L. Yuille. Articulated Pose Estimation by a Graphical Model with Image Dependent Pairwise Relations. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1736–1744, 2014.

[3] 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.

[4] Navneet Dalal and Bill Triggs. Histograms of Oriented Gradients for Human Detection. In *Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 886–893, June 2005.

[5] Marcin Eichner and Vittorio Ferrari. Appearance Sharing for Collective Human Pose Estimation. In *Asian Conference on Computer Vision (ACCV)*, pages 138–151, 2013.

[6] P. F. Felzenszwalb, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2008.

[7] Pedro F. Felzenszwalb and Daniel P. Huttenlocher. Pictorial Structures for Object Recognition. In *International Journal of Computer Vision (IJCV)*, volume 61, pages 55–79, 2005.

[8] Vittorio Ferrari, Manuel Marin-jimenez, and Andrew Zisserman. 2D Human Pose Estimation in TV Shows. In *Statistical and Geometrical Approaches to Visual Motion Analysis, Volume 5604*, pages 128–147, 2009.

[9] Katerina Fragkiadaki, Han Hu, and Jianbo Shi. Pose from Flow and Flow from Pose. In *Computer Vision and Pattern Recognition (CVPR)*, pages 2059–2066, 2013.

[10] Arjun Jain, Jonathan Tompson, Yann Lecun, and Christoph Bregler. MoDeep: A Deep Learning Framework Using Motion Features for Human Pose Estimation. *LNCS, Asian Conference on Computer Vision (ACCV)*, 9004:302–315, 2014.

[11] Sam Johnson and Mark Everingham. Clustered Pose and Nonlinear Appearance Models for Human Pose Estimation. In *British Machine Vision Conference (BMVC)*, pages 12.1–12.11, 2010.

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

[13] Taegyu Lim, Seunghoon Hong, Bohyung Han, and Joon Hee Han. Joint Segmentation and Pose Tracking of Human in Natural Videos. In *International Conference on Computer Vision (ICCV)*, pages 833–840, 2013.

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

[15] Deva Ramanan, David A. Forsyth, and Andrew Zisserman. Strike a Pose: Tracking People by Finding Stylized Poses. In *Computer Vision and Pattern Recognition (CVPR)*, pages 271–278, 2005.

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

[17] Benjamin Sapp, Alexander Toshev, and Ben Taskar. Cascaded Models for Articulated Pose Estimation. In *European Conference on Computer Vision (ECCV)*, pages 406–420, 2010.

[18] Benjamin Sapp, David Weiss, and Ben Taskar. Parsing Human Motion with Stretchable Models. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1281–1288, 2011.

[19] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. *Computer Vision and Pattern Recognition (CVPR)*, pages 1653–1660, 2013.

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

[21] Yi Yang and Deva Ramanan. Articulated Pose Estimation with Flexible Mixtures-of-Parts. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1385–1392, 2011.

[22] Silvia Zuffi, Javier Romero, Cordelia Schmid, and Michael J. Black. Estimating Human Pose with Flowing Puppets. In *International Conference on Computer Vision (ICCV)*, pages 3312–3319, 2013.