Dynamical Binary Latent Variable Models for 3D Human Pose Tracking Supplementary Material

Graham W. Taylor
New York University
New York, USA
gwtaylor@cs.nyu.edu

Leonid Sigal
Disney Research
Pittsburgh, USA
lsigal@disneyresearch.com

David J. Fleet and Geoffrey E. Hinton
University of Toronto
Toronto, Canada
{fleet,hinton}@cs.toronto.edu

1. Details of imCRBM weight updates

In this section we describe the *positive and negative phase* statistics that were referenced in the Appendix of the submitted paper, and derive the weight updates for each set of parameters in the imCRBM.

Note that W_k, A_k and B_k can be thought of as either (1) the k^{th} slice of weight tensors W,A and B, respectively; or (2) the weights of the k^{th} component CRBM in the mixture. Similarly, $\mathbf{c}_k$, and $\mathbf{d}_k$ can be thought of as either (1) the k^{th} column of weight matrices C and D respectively; or (2) the biases of the k^{th} component CRBM in the mixture. Both views are equivalent.

Contrastive divergence learning consists of (1) a positive phase in which we hold the visible variables fixed to a training vector, $\mathbf{x}_t^+$, and sample the latent variables $\mathbf{q}_t, \mathbf{z}_t$; and (2) a negative phase, in which we alternate between reconstructing (by sampling) the visible variables given latent variables and sampling the latent variables given the reconstructed visible variables. The positive phase yields k^+ (i.e. $\mathbf{q}_{kt}^+ = 1$) and $\mathbf{z}_t^+$ which are our sampled latent variables conditional on the training data $\mathbf{x}_t^+$. The negative phase yields $\mathbf{x}_t^-$ which are the values of the visible variables after M steps of alternating sampling and k^- and $\mathbf{z}_t^-$ which are our sampled latent variables conditional on $\mathbf{x}_t^+$. Note that all distributions which we draw from are conditional on the history $\mathbf{x}_{h_t}$ which remains constant throughout the entire learning algorithm.

The imCRBM has two types of learnable parameters: weights and offsets (biases). The weights capture pairwise interactions between variables and therefore the statistics that comprise the updates are outer products. The statistics for the offset parameters are simply average activities. Reflecting the pairwise connectivity of the model, the positive phase statistics are:

$$W_k^+ = \mathbf{x}_t^+ \mathbf{z}_t^{+T} \tag{1}$$

$$A_k^+ = \mathbf{x}_t^+ \mathbf{x}_{h_t}^T \tag{2}$$

$$B_k^+ = \mathbf{z}_t^+ \mathbf{x}_{h_t}^T \tag{3}$$

$$\mathbf{c}_k^+ = \mathbf{x}_t^+ \tag{4}$$

$$\mathbf{d}_k^+ = \mathbf{z}_t^+ \tag{5}$$

The negative phase statistics have the same form:

$$W_k^- = \mathbf{x}_t^- \mathbf{z}_t^{-T} \tag{6}$$

$$A_k^- = \mathbf{x}_t^- \mathbf{x}_{h_t}^T \tag{7}$$

$$B_k^- = \mathbf{z}_t^- \mathbf{x}_{h_t}^T \tag{8}$$

$$\mathbf{c}_k^- = \mathbf{x}_t^- \tag{9}$$

$$\mathbf{d}_k^- = \mathbf{z}_t^- \tag{10}$$

Repeating the positive and negative phase for a mini-batch of S training pairs of the form $\{\mathbf{x}_{h_t}, \mathbf{x}_t^+\}$ results in ten sets of statistics for each component k. The first five sets correspond to the U_k times k was chosen in the positive phase: $\{W_{k1}^+, \ldots, W_{kU_k}^+\}$,$\{A_{k1}^+, \ldots, A_{kU_k}^+\}$, $\{B_{k1}^+, \ldots, B_{kU_k}^+\}$, $\{\mathbf{c}_{k1}^+, \ldots, \mathbf{c}_{kU_k}^+\}$, $\{\mathbf{d}_{k1}^+, \ldots, \mathbf{d}_{kU_k}^+\}$. The next five sets correspond to the V_k times k was chosen in the negative phase: $\{W_{k1}^-, \ldots, W_{kV_k}^-\}$,$\{A_{k1}^-, \ldots, A_{kV_k}^-\}$, $\{B_{k1}^-, \ldots, B_{kV_k}^-\}$, $\{\mathbf{c}_{k1}^-, \ldots, \mathbf{c}_{kV_k}^-\}$, $\{\mathbf{d}_{k1}^-, \ldots, \mathbf{d}_{kV_k}^-\}$.

The weight updates for the k^{th} component CRBM are averages over the mini-batch:

$$\Delta W_k = \frac{\lambda}{S} \left(\sum_{u=1}^{U_k} W_{ku}^+ - \sum_{v=1}^{V_k} W_{kv}^- \right), \tag{11}$$

$$\Delta A_k = \frac{\lambda}{S} \left(\sum_{u=1}^{U_k} A_{ku}^+ - \sum_{v=1}^{V_k} A_{kv}^- \right), \tag{12}$$

$$\Delta B_k = \frac{\lambda}{S} \left(\sum_{u=1}^{U_k} B_{ku}^+ - \sum_{v=1}^{V_k} B_{kv}^- \right), \tag{13}$$

$$\Delta \mathbf{c}_k = \frac{\lambda}{S} \left(\sum_{u=1}^{U_k} \mathbf{c}_{ku}^+ - \sum_{v=1}^{V_k} \mathbf{c}_{kv}^- \right), \tag{14}$$

$$\Delta \mathbf{d}_k = \frac{\lambda}{S} \left(\sum_{u=1}^{U_k} \mathbf{d}_{ku}^+ - \sum_{v=1}^{V_k} \mathbf{d}_{kv}^- \right). \tag{15}$$

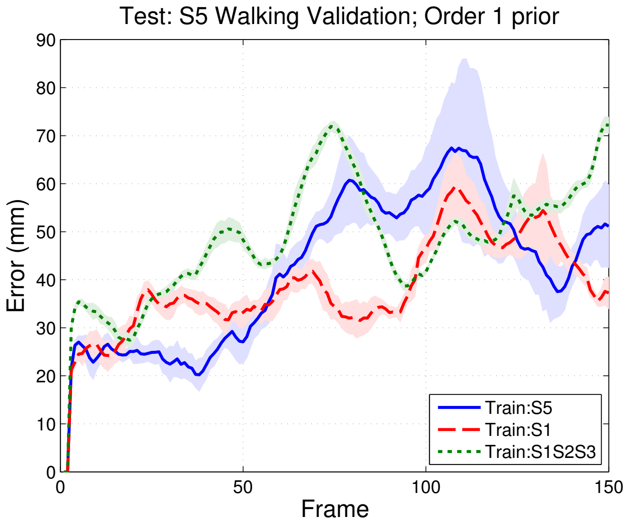

Figure 1. **Multi-view tracking.** Tracking of subject S5 using a 1st order prior model learned from S5, S1, and S1+S2+S3 training data. Results are averaged over 10 runs (per-frame standard deviation is shaded).

where λ is a learning rate which could be made parameter-specific.

Note that the U_k and V_k need not be the same for a given component, k, since k^+ and k^- are sampled. For the case of $K = 1$, $U_1 = V_1 = S$ and these weight updates reduce to the updates of a standard CRBM [2].

2. Multi-view tracking

Figure 1 shows the result of tracking subject S5 using three different first-order CRBM priors, each trained on a different dataset. The plot is the illustration of the condensed results reported in Table 1 of the submitted paper. It was not included in the main document due to space constraints. We see that the prior is able to generalize to subjects who are not included in the training set.

3. Monocular tracking

The analysis of performance of imCRBM-2L prior on the monocular ”combo” sequence for subject S3 is illustrated in Table 1. Similarly to other experiments in the paper we use 1000 particles (though similar results were achieved even with 200 particles). We run imCRBM-2L on each of the 3 views independently and report the performance averaged over 5 runs in each case. A single, representative, run with Camera 2 data was illustrated in Figure 5 of the main submission. Since with monocular observations we do not expect to resolve the depth reliably (due to depth ambiguity), as is standard in the literature, we report the *relative* average marker error. The average relative marker error corresponds to the marker distance with respect to the pelvis (see [1] for details). The performance is very encouraging, particularly considering that only one camera view is used and the motion contains transitions. Consequently, the baseline algorithm is not able to track the sequence resulting in a quick failure.

	imCRBM-2L
Camera 1	118.87 ± 33.12
Camera 2	84.26 ± 6.85
Camera 3	90.44 ± 7.64

Table 1. **Monocular tracking with transitions: quantitative performance.** Tracking of subject S3 “combo” sequence using a first-order imCRBM-2L prior model with 1000 particles. Results are averaged over 5 runs.

References

[1] L. Sigal, A. Balan, and M. Black. HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *IJCV*, 2010.

[2] G. Taylor, G. Hinton, and S. Roweis. Modeling human motion using binary latent variables. *NIPS*, 19, 2007.