Supplementary Material

1 TRAINING ALGORITHM SUMMARY

Due to space limitations the summary of our training algorithm for the Variational HCRF–DPM could not fit in Sect. 3.2, so we present it here. Equation numbers reflect equations in this supplementary material.

Algorithm 1 Model Training for Variational HCRF–DPM

```
Initialize s_{x,1}, s_{x,2}, s_{y,1}, s_{y,2}, s_{e,1}, s_{e,2}
Randomly initialize \alpha_x, \alpha_y, \alpha_e, \theta, \tau
Initialize nbItrs, nbVarItrs
itr = 0
converged = FALSE

while (\sim converged) and (itr < nbItrs) do
    varConverged = FALSE
    while (\sim varConverged) and (varItr < nbVarItrs) do
        \{Phase 1: Optimize variational parameters \tau\}
        Calculate \forall t q(s_t|X, y, s_{t-1}) by using (38)–(44)
        Compute approximate marginals \( q(s_t = h_k|i), q(s_t = h_k|y), \) and \( q(s_t = h_k,y,s_{t-1} = h_k|y) \) by using a forward–backward algorithm.
        Hyperparameter posterior sampling for \alpha_x, \alpha_y, \alpha_e by using (48)
        Calculate Kullback-Liebler divergence \( KL(varItr) \) by using (4)
        Update \tau by using (28)–(33)
        varConverged = \frac{KL(varItr) - KL(varItr-1)}{KL(varItr)} < \epsilon
        varItr = varItr + 1
    end while
    \{Phase 2: Optimize parameters \theta\}
    Gradient descent to find \theta(iteration) by using a quasi–Newton method and an Armijo backtracking line search with projected gradients to keep \theta non–negative
    converged = \sum (|\theta(itr) - \theta(itr - 1)|) < \epsilon'
itr = itr + 1
end while
```

2 CALCULATION OF VARIATIONAL UPDATES

This section is meant to complement Sect. 3.2.1: Optimization of Variational Parameters \( \tau \).

Since we have defined an approximate model distribution, we can approximate the necessary quantities \( q(s_t), q(s_t|y), q(s_t|s_{t-1}), q(s_t,s_{t-1},y) \), \( \log q(\tau_x), \log q(\tau_y), \log q(\tau_e) \). These approximations, as one can see later in this section, depend solely on our variational parameters \( \tau \). We calculate these by minimizing the Kullback-Liebler divergence (KL) between approximate and actual joint distributions of our model using a coordinate descent algorithm:

\[
KL[q(y,s,\tau'|X) \parallel p(y,s,\tau'|X)] = \int_\pi \sum_{y,s} q(y,s|X)q(\tau') \log \frac{q(y,s|X)q(\tau')}{p(y,s|X)p(\tau')} d\tau' = (1)
\]

\[
\int_\pi \sum_{y,s} q(y,s|X)q(\tau') \log \frac{Z(X)q(y,s|X)q(\tau')}{\mathcal{F}(y,s|X)p(\tau')} d\tau' = (2)
\]

Since the normalization factor \( Z(X) = \sum_{y,s} \mathcal{F}(y,s,X) \) is a constant for a given observation sequence, the Kullback–Liebler divergence becomes:

\[
KL[q||p] = \log Z(X) - \langle \log \mathcal{F}(y,x,\tau',X)p(\tau') \rangle_{q(y,s,\tau'|X)} + \langle \log q(y,s|X)q(\tau') \rangle_{q(y,s,\tau'|X)} = (3)
\]

where \langle \cdot \rangle_q is the expectation of \cdot with respect to \( q \). Thus, the energy of the configuration of our random variables \( y, s, \) and \( \tau' \) is \( \log \mathcal{F}(y,x,\tau',X)p(\tau') \) and the free energy of the variational distribution:

\[
\mathcal{L}(q) = -\langle \log \mathcal{F}(y,x,\tau',X)p(\tau') \rangle_{q(y,s,\tau'|X)} + \langle \log q(y,s|X)q(\tau') \rangle_{q(y,s,\tau'|X)} = (4)
\]

Since \( \log Z(X) \) is constant for a given observation sequence, minimizing the free energy \( \mathcal{L}(q) \) minimizes the KL divergence. And since \( KL[q||p] \) is positive, the free energy \( \mathcal{L}(q) \) \( \geq -\log Z(X) \). Therefore KL is minimized at 0 when \( \mathcal{L}(q) = \log Z(X) \).

We will obtain the variational updates for the two groups of latent variables \( q(y,s|X) \) and \( q(\tau') \) by setting the partial derivative with respect to each group of \( \mathcal{L}(q) \) to 0 and solving for the approximate distribution of each group of latent variables.
2.1 Updates for $q(\pi')$

\[
\frac{\partial L(q)}{\partial q(\pi')} = 0
\]

\[
\frac{\partial}{\partial q(\pi')} \int_\pi \sum_{y,s} q(y,s|X)q(\pi') \log(F(y,s,\pi',X)p(\pi')) - \sum_{y,s} q(y,s|X)q(\pi') \log(q(y,s|X)) \, d\pi' = 0
\]

(one solution):

\[
\frac{\partial}{\partial q(\pi')} \int_\pi \sum_{y,s} q(y,s|X)q(\pi') \log(F(y,s,\pi',X)p(\pi')) - \sum_{y,s} q(y,s|X)q(\pi') \log(q(y,s|X)) - \sum_{y,s} q(y,s|X)q(\pi') \log(q(\pi')) = 0
\]

As a reminder, for the HCRF–DPM:

\[
\log F(y,s,\pi',X,\theta) = \sum_{t=1}^T \sum_{i=1}^d \theta_x(s_t,i) f_t(i) \log \pi_x(s_t|i) + \theta_y(s_t,y) \log \pi_y(s_t|y) \sum_{t=2}^T \theta_c(s_t,s_{t-1},y) \log \pi_c(s_t,y|s_{t-1})
\]

Therefore, the expectation of $\log F$ is:

\[
\langle \log F(y,s,\pi',X,\theta) \rangle_{q(y,s|X)} = \sum_{t=1}^T \sum_{i=1}^d \theta_x(s_t,i) f_t(i) \langle \log \pi_x(s_t|i) \rangle_{q(y,s|X)} + \theta_y(s_t,y) \langle \log \pi_y(s_t|y) \rangle_{q(y,s|X)} + \sum_{t=2}^T \theta_c(s_t,s_{t-1},y) \langle \log \pi_c(s_t,y|s_{t-1}) \rangle_{q(y,s|X)}
\]

Let's consider only $\pi_x$

\[
\pi_x(s_t|i) = \prod_{k=1}^\infty \pi_x^t(s_t = k|i) I_{[s_t = k|i]} (1 - \pi_x^t(s_t > k|i)) I_{[s_t > k|i]}
\]

Since $\forall i, t \langle \log(1 - \pi_x^t(s_t = L|i)) \rangle_q = 0, q(y,s_t > L|X) = 0,$ and $q(y,s|X)$ factorizes as follows:

\[
q(y,s|X) = \prod_{i=1}^d q(s_t|i)q(s_t|y) \prod_{t=2}^T q(s_t|i)q(s_t|y)q(s_t, y|s_{t-1})
\]

we have

\[
\langle \log \pi_x(s_t|i) \rangle_{q(y,s|X)} = \sum_{k=1}^L q(s_t = k|i) \langle \log \pi_x^t(s_t = k|i) \rangle_q + q(s_t > k|i) \langle \log(1 - \pi_x^t(s_t > k|i)) \rangle_q
\]

\[
\langle \log \pi_x(s_t|i) \rangle_{q(y,s|X)} = \sum_{k=1}^L q(s_t = k|i) \log \pi_x^t(s_t = k|i) + q(s_t > k|i) \log(1 - \pi_x^t(s_t > k|i))
\]

Therefore, from (16):

\[
q(\pi_x^t(k|i)) \propto (1 - \pi_x^t(s_t = L|i))^{\alpha_x-1} \pi_x^t(k|i) \prod_{i=1}^T f_t[i] \theta_x(k,i) q(s_t = k|i) (1 - \pi_x^t(s_t > k|i)) \sum_{i=1}^d f_t[i] \theta_x(k,i) q(s_t > k|i)
\]

\[
\propto Beta \left( \sum_t f_t[i] \theta_x(k,i) q(s_t = k|i) + 1, \sum_t f_t[i] \theta_x(k,i) q(s_t > k|i) + \alpha_x^x \right)
\]

Similarly, we compute all our updates:

\[
q(\pi_y^t(k|y)) \propto Beta \left( \sum_t \theta_y(k,y) q(s_t = k|y) + 1, \sum_t \theta_y(k,y) q(s_t > k|y) + \alpha_y^y \right)
\]
2.2 Updates for $q(y, s | X)$

$$\frac{\partial L(q)}{\partial q(y, s | X)} = 0$$

$$\frac{\partial}{\partial q(\pi')} \int_\pi q(y, s | X) q(\pi') \log(F(y, s, \pi', X)p(\pi')) - \int_\pi q(y, s | X) q(\pi') \log(q(y, s | X)) d\pi' = 0$$

(one solution:)

$$\int_\pi q(\pi') \log(F(y, s, \pi', X)p(\pi')) - \frac{q(y, s | X) q(\pi') \log(q(y, s | X))}{q(\pi')} d\pi' = 0$$

$$\log(q(y, s | X)) = \left( \log \frac{F(y, s, \pi', X)p(\pi')}{q(\pi')} - 1 \right)_{q(\pi')}$$

From (18):

$$\log q(y, s_t = k | s_{t-1}, X) = \sum_i f_t(i) \theta_x(k, i) \left( \log \pi_x'(s_t = k|i) \right)_{q(\pi')} + \sum_{j=k+1}^L \langle \log(1 - \pi_x'(s_t = j|i)) \rangle_{q(\pi')} + \theta_x(k, y) \left( \langle \log \pi_y'(s_t = k|y) \rangle_{q(\pi')} \right) + \sum_{j=k+1}^L \langle \log(1 - \pi_y'(s_t = j|y)) \rangle_{q(\pi')} + \theta_x(k, k') \left( \langle \log \pi_x'(s_t = k,y|s_{t-1} = k') \rangle_{q(\pi')} + \langle \log \pi_x'(s_t = k|y) \pi_x'(s_t = k,y|s_{t-1} = k') \rangle_{q(\pi')} \right) + \langle \log(1 - \pi_x'(s_t = j|y) \pi_y'(s_t = k|y) \pi_x'(s_t = k,y|s_{t-1} = k')) \rangle_{q(\pi')}$$

$$\log q(y, s_t = k | s_{t-1}, X) \propto \sum_i f_t(i) \theta_x(k, i) \left( \log \pi_x'(s_t = k|i) \right)_{q(\pi')} + \sum_{j=k+1}^L \langle \log(1 - \pi_x'(s_t = j|i)) \rangle_{q(\pi')} + \theta_x(k, y) \left( \langle \log \pi_y'(s_t = k|y) \rangle_{q(\pi')} \right) + \sum_{j=k+1}^L \langle \log(1 - \pi_y'(s_t = j|y)) \rangle_{q(\pi')} + \theta_x(k, k') \left( \langle \log \pi_x'(s_t = k,y|s_{t-1} = k') \rangle_{q(\pi')} + \langle \log \pi_x'(s_t = k|y) \pi_x'(s_t = k,y|s_{t-1} = k') \rangle_{q(\pi')} \right) + \langle \log(1 - \pi_x'(s_t = j|y) \pi_y'(s_t = k|y) \pi_x'(s_t = k,y|s_{t-1} = k')) \rangle_{q(\pi')}$$

Since all $\pi'$ follow a Beta distribution, the expectations above are known. Since all $\pi'$ follow a Beta distribution, the expectations above are known.

$$\langle \log \pi_x'(s_t = h_k|i) \rangle = \Psi(\tau_{x,1}(k, i)) - \Psi(\tau_{x,1}(k, i) + \tau_{x,2}(k, i))$$

$$\langle \log(1 - \pi_x'(s_t = h_k|i)) \rangle = \Psi(\tau_{x,2}(k, i)) - \Psi(\tau_{x,2}(k, i) + \tau_{x,2}(k, i))$$

$$\langle \log \pi_y'(s_t = h_k|y) \rangle = \Psi(\tau_{y,1}(k, y)) - \Psi(\tau_{y,1}(k, y) + \tau_{y,2}(k, y))$$

$$\langle \log(1 - \pi_y'(s_t = h_k|y)) \rangle = \Psi(\tau_{y,2}(k, y)) - \Psi(\tau_{y,2}(k, y) + \tau_{y,2}(k, y))$$

$$\langle \log \pi_x'(s_t = h_k,y|h_{k'}) \rangle = \Psi(\tau_{x,1}(k, y,k')) - \Psi(\tau_{x,1}(k, y,k') + \tau_{x,2}(k, k'))$$

$$\langle \log(1 - \pi_x'(s_t = h_k,y|h_{k'}) \rangle = \Psi(\tau_{x,2}(k, y,k')) - \Psi(\tau_{x,2}(k, y,k') + \tau_{x,2}(k, k'))$$

$$\langle \log \pi_y'(s_t = h_k,y|h_{k'}) \rangle = \Psi(\tau_{y,1}(k, y,k')) - \Psi(\tau_{y,1}(k, y,k') + \tau_{y,2}(k, k'))$$

$$\langle \log(1 - \pi_y'(s_t = h_k,y|h_{k'}) \rangle = \Psi(\tau_{y,2}(k, y,k')) - \Psi(\tau_{y,2}(k, y,k') + \tau_{y,2}(k, k'))$$

where $\Psi(\cdot)$ is the digamma function.
2.3 Posterior sampling for DP scaling parameters

The scaling parameters $\alpha_x, \alpha_y, \alpha_z$ can have a significant effect on our HCRF–DPM model, as they control the growth of the used hidden states. It is suggested in [1] that for DPMs one should place a $\text{Gamma}(s_1, s_2)$ prior on these parameters and integrate over them. Since our model uses a number of DPMs, we include posterior updates for these scaling parameters as part of our variational coordinate descent algorithm. In this work, we use a different scaling parameter for each DPM, but with a common prior. The variational distribution for the scaling parameter $\alpha_{x,i}$ corresponding to the DPM for feature $i$ is

$$q(\alpha_{x,i}) = \text{Gamma}(w_{1,x}, w_{2,x,i})$$ (45)

where

$$w_{1,x} = s_{1,x} + L - 1$$ (46)

$$w_{2,x,i} = s_{2,x} - \sum_{k=1}^{L-1} (\log(1 - \pi'_x(k,i)))_q$$ (47)

and we replace the $\alpha_x$ values in (29) with the respective expectation:

$$\langle \alpha_{x,i} \rangle_q = \frac{w_{1,x}}{w_{2,x,i}}$$ (48)

The posterior updates for the rest of the scaling parameters are obtained in a similar fashion and so they are omitted for brevity.

3 Detailed Description of the Experimental Datasets

3.1 Synthetic Dataset Generation

The transition matrices of the Hidden Markov Models (HMM) generating our sequences, as well as the details of their Gaussian states are shown in Tables 1-3.

<table>
<thead>
<tr>
<th>HMM-1</th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>0.4</td>
<td>0.4</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>S2</td>
<td>0.1</td>
<td>0.4</td>
<td>0.4</td>
<td>0.1</td>
</tr>
<tr>
<td>S3</td>
<td>0.1</td>
<td>0.1</td>
<td>0.4</td>
<td>0.4</td>
</tr>
<tr>
<td>S4</td>
<td>0.4</td>
<td>0.1</td>
<td>0.1</td>
<td>0.4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>HMM-2</th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>0.1</td>
<td>0.7</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>S2</td>
<td>0.1</td>
<td>0.1</td>
<td>0.7</td>
<td>0.1</td>
</tr>
<tr>
<td>S3</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>0.7</td>
</tr>
<tr>
<td>S4</td>
<td>0.7</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
</tr>
</tbody>
</table>

3.2 Audiovisual Analysis of Human Behavior

The audiovisual dataset of spontaneous agreement and disagreement comprises of 53 episodes of agreement, 94 episodes of disagreement, and 130 neutral episodes of neither agreement or disagreement. These episodes feature 28 participants and they occur over a total of 11 real political debates from The Canal9 Database of Political Debates [2]. As the debates were filmed with multiple cameras, and edited live to one feed, the episodes selected for the dataset were only the ones that were contained within one personal, close-up shot of the speaker. We used automatically extracted prosodic features (F0 and Energy), based on previous work on agreement and disagreement classification, and manually annotated visual features, the hand and head gestures hypothesized relevant according to literature. The 2 prosodic features used were F0 and Energy, and the 9 gestures used in our experiments are the ‘Head Nod’, ‘Head Shake’, ‘Forefinger Raise’, ‘Forefinger Raise–Like’, ‘Forefinger Wag’, ‘Hand Wag’, ‘Hand Chop’, ‘Hands Scissor’, and ‘Shoulder Shrug’. (see [3] for details).

We encoded each gesture in a binary manner, based on its presence at each of the 5,700 total number of video frames, with each sequence ranging from 30 to 120 frames. The prosodic features were extracted with the publicly available software package OpenEar [4].

The database of pain we used was the UNBC-McMaster [5], which features 25 subjects–patients spontaneously expressing various levels of elicited pain in a total of 200 video sequences. The database was coded for, among others, pain level per sequence by expert observers on a 6–point scale from 0 (no pain) to 5 (extreme pain). Furthermore, each of the 48,398 video frames in the database was coded for each of the observable facial muscle movements–Action Units (AUs) according to the Facial Action Coding System (FACS) [6] by expert FACS coders. In our experiments we encoded each of the possible 45 AUs in a binary manner, based on their presence. We labeled sequences coded with 0 as ‘no pain’, sequences coded with 1–2 as ‘moderate pain’, and those coded as 3–5 as ‘strong pain’. For our experiments, we compared the finite HCRFs to our HCRF–DPM based on the F1 measure they achieved in each of the classification problems at hand.

4 Further Experiments

In an attempt to clearly show how a variational HCRF–DPM functions differently from a finite HCRF with the same number of “allowed” hidden states, we show in figure 1 a comparison of the learned potentials of an HCRF with 50 hidden states for the 2–label (dis)agreement recognition problem, next to

1. Publicly available at http://canal9-db.sspnet.eu/
Fig. 1. Hinton Diagrams of a variational HCRF-DPM with $L = 50$ for ADA2 vs a finite HCRF with 50 hidden states: the nonparametric prior on $\pi$ induces sparsity for our HCRF–DPM model.
the learned equivalent potentials of an HCRF–DPM with an upper bound of hidden states set to $L = 50$. One can clearly see that an HCRF uses all 50 states roughly equally, whereas the learned potentials for HCRF–DPM are a lot more sparse with only a few number of hidden states used, due to the nonparametric prior on the $\pi$-quantities.

REFERENCES