Context-sensitive Facial Expressivity Prediction by Multimodal Hierarchical Bayesian Neural Networks

Ajjen Joshi[1], Soumya Ghosh[2], Sarah Gunnery[3],
Linda Tickle-Degnen[3], Stan Sclaroff[1], Margrit Betke[1]
[1]Boston University, [2]IBM T.J. Watson Research Center, [3]Tufts University

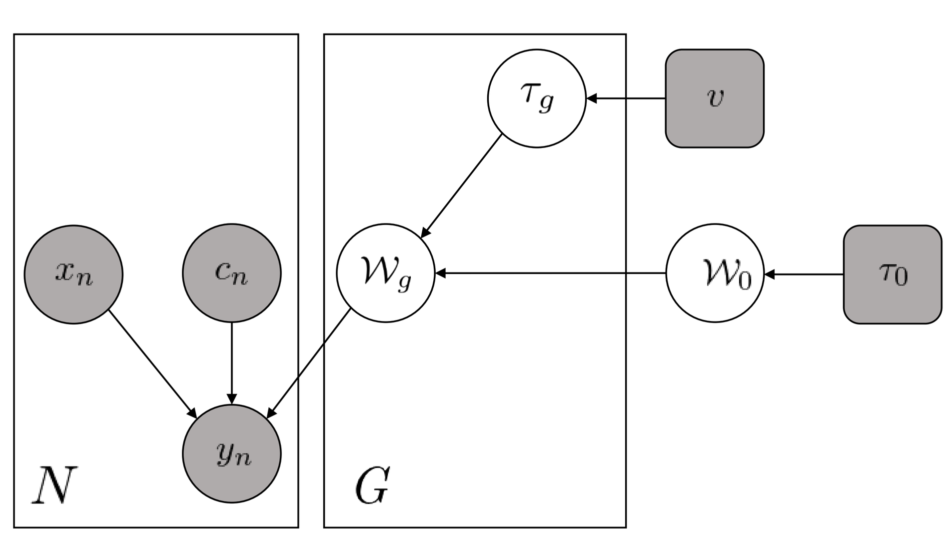
Overview

- We investigate whether contextual-information can be leveraged for the task of predicting facial expressivity in patients with Parkinson's Disease.
- We experiment with two notions of context: (1) *gender* and (2) *sentiment*.
- We train hierarchical Bayesian neural networks with multimodal feature representations.

Contributions

- We demonstrate the benefits of using a framework that adapts to contextual information.

Hierarchical Bayesian Neural Networks



- Given a dataset $\mathcal{D} = \{x_n, y_n\}_{n=1}^{N}$, each group is endowed with its own conditional distribution $p(y_n \mid z_n = g, f(x_n, \mathcal{W}_g))$.

$$p(\mathcal{W}_g \mid \mathcal{W}_0, \tau_g) = \prod_{l=1}^{L} \prod_{i=1}^{V_{l-1}} \prod_{j=1}^{V_l} \mathcal{N}(w_{ij,l}^{g} \mid w_{ij,l}^{0}, \tau_g^{-1})$$

$$p(\mathcal{W}_0 \mid \tau_0) = \prod_{l=1}^{L} \prod_{i=1}^{V_{l-1}} \prod_{j=1}^{V_l} \mathcal{N}(w_{ij,l}^{0} \mid 0, \tau_0^{-1})$$

$$p(\gamma_g \mid v) = \mathcal{N}(\gamma_g \mid 0, v); \quad \tau_g^{-1/2} = |\gamma_g|$$

- The joint distribution is given by:

$$p(\mathcal{W}_0, \mathcal{W}, \mathcal{T}, \mathbf{y} \mid \mathbf{x}, \mathbf{z}, \tau_0, v) = p(\mathcal{W}_0 \mid \tau_0^{-1}) \prod_{g=1}^{G} p(\gamma_g \mid v) p(\mathcal{W}_g \mid \mathcal{W}_0, \tau_g^{-1}) \prod_{n=1}^{N} \prod_{g=1}^{G} p(y_n \mid f(\mathcal{W}_g, x_n))^{\mathbf{1}[z_n = g]}$$

Inference

- We approximate the intractable posterior with a fully factorized variational approximation,

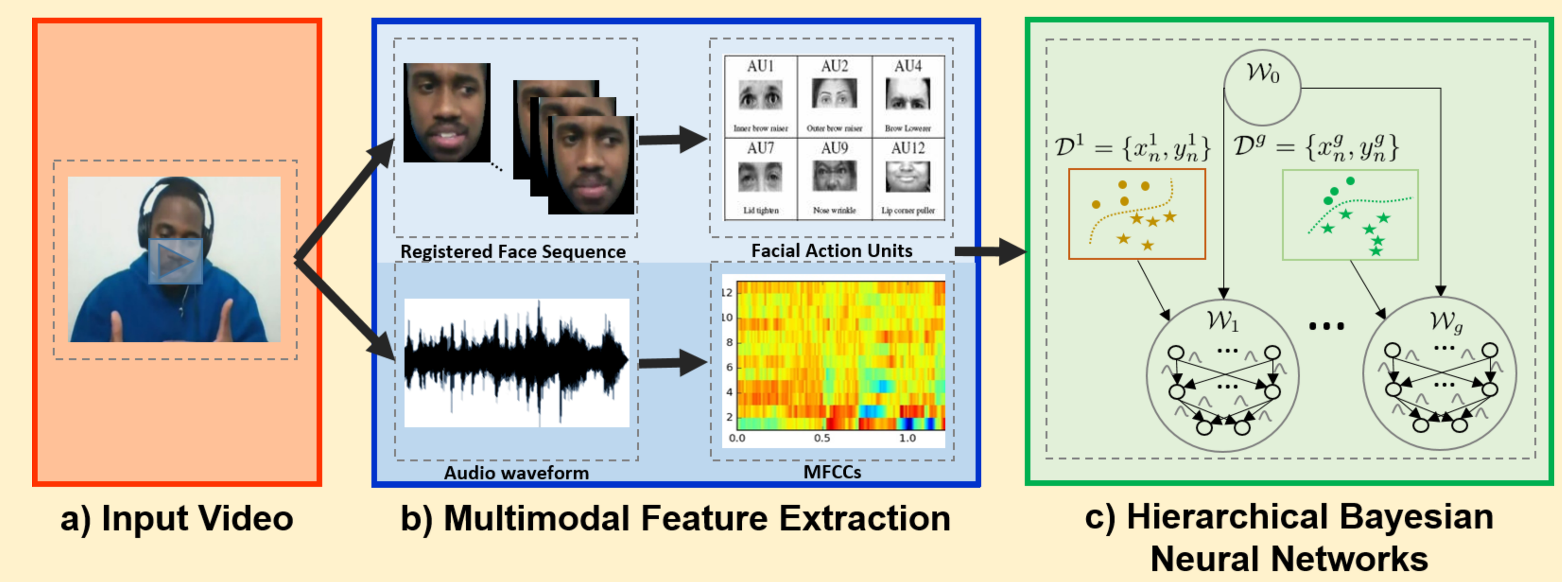
$$q(\mathcal{W}_0, \mathcal{W}, \mathcal{T} \mid \phi) = q(\mathcal{W}_0 | \phi_0) \prod_{g=1}^{G} q(\mathcal{W}_g | \phi_g) q(\tau_g^{-1/2} | \phi_{\tau_g})$$

- The Evidence Lower Bound (ELBO) is then maximized with respect to the variational parameters using variational Bayes.

$$\mathcal{L}(\phi) = \mathbb{E}_{q_\phi}[\ln p(\mathcal{W}_0, \mathcal{W}, \mathcal{T}, \mathbf{y} \mid \mathbf{x}, \mathbf{z}, \tau_0, v)] - \mathbb{E}_{q_\phi}[\ln q(\mathcal{W}_0, \mathcal{W}, \mathcal{T} \mid \phi)]$$

- In computing the Monte Carlo estimate of the gradients, we use the local reparameterization trick.

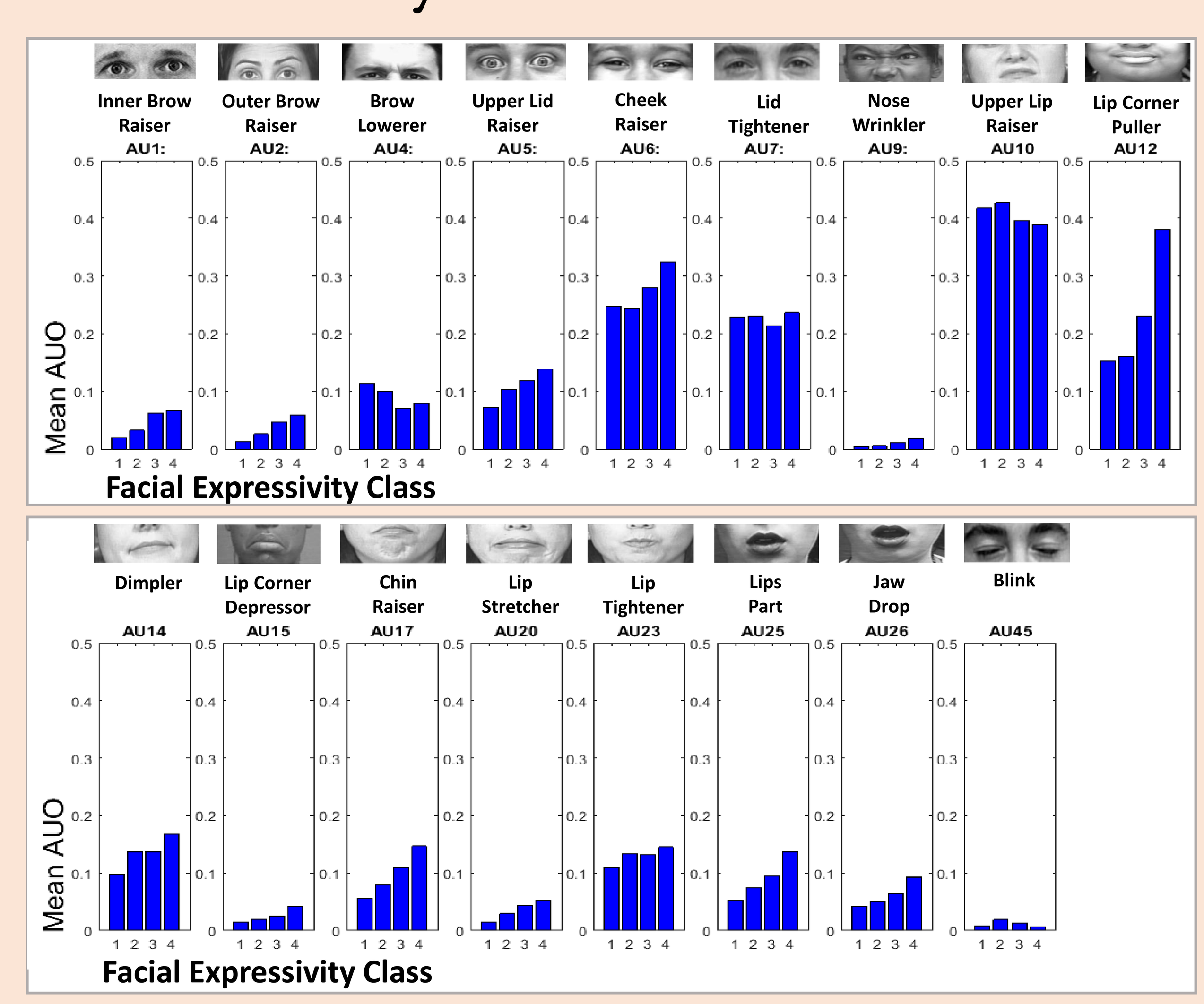
Model Pipeline



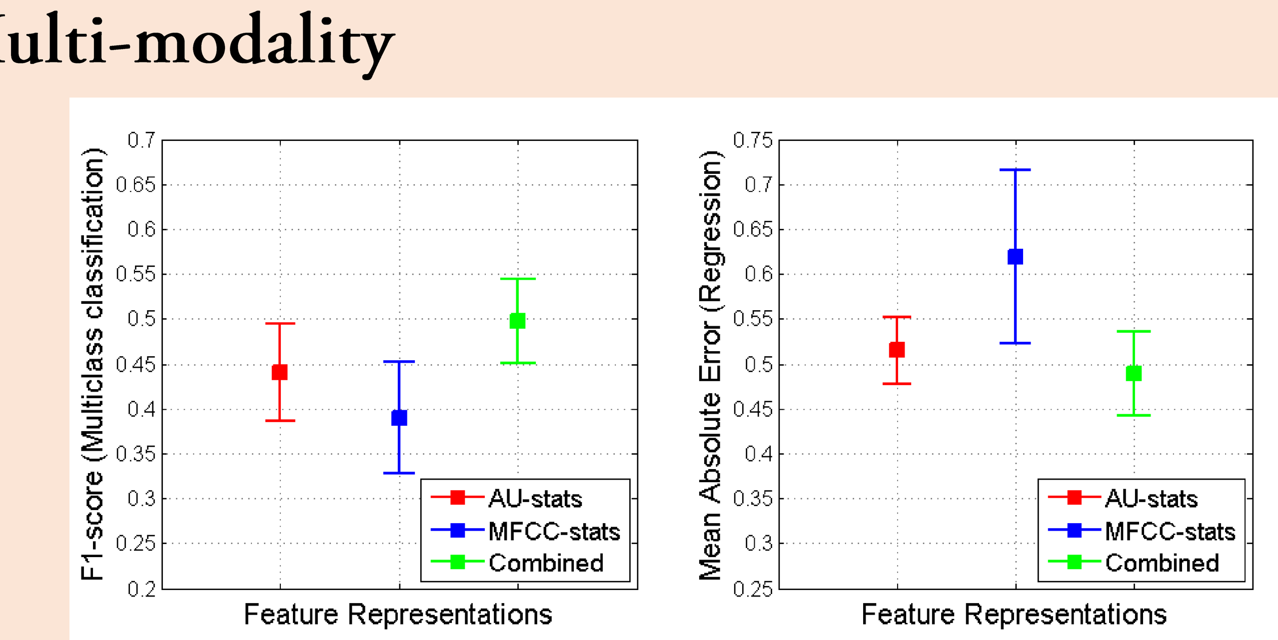
Results

- We test our method on a dataset of 772 short audio-video clips of patients with Parkinson's Disease using 9-fold cross validation.
- We divide the dataset into context-sensitive groups.
- For each video clip we extract:
  a) Action Unit stats (AU-stats) to capture visual features
  b) MFCC stats (MFCC-stats) to capture audio features

1. Action Unit Analysis



2. Multi-modality



3. Context-sensitive Models