Multimodal Fusion using Dynamic Hybrid Models

Mohamed R. Amer, Behjat Siddiquie, Saad Khan, Ajay Divakaran and Harpreet Sawhney
SRI International, Princeton, NJ

WACV 14

Problem and Motivation

Goal

- Detect Multimodal events in time varying sequences.

Application

- Analysis of Human behaviors and emotions: Facial expressions, paralinguistics, eye gaze, hand gestures, head motion etc.
- Temporal dynamics within and across modalities is key to modeling and capturing affect.

Approach

- Staged hybrid model: exploits the strength of discriminative classifiers along with the representational power of generative models.

Staged-Dynamic Hybrid Model

Why Staged Hybrid Dynamic Model?

- (Staged) training each model separately, where the discriminative model trained on representations learned by the generative model.
- (Hybrid) exploiting the generative model's expressiveness and the discriminative model's classification power.
- (Dynamic) Modeling the temporal content of time varying data is important.

Generative – Multimodal Conditional Restricted Boltzmann Machines:

- Multimodal CRBMs consists of single CRBMs, and fusion CRBM.

$$p_{\rm G}(\mathbf{v}_t, \mathbf{h}_t | \mathbf{v}_{<t}) = \exp[-E_{\rm G}(\mathbf{v}_t, \mathbf{h}_t | \mathbf{v}_{<t})]/Z(\boldsymbol{\theta}_{\rm G})$$

$$Z(\boldsymbol{\theta}_{\rm G}) = \sum_{\mathbf{v},\mathbf{h}} \exp[-E_{\rm G}(\mathbf{v}_t, \mathbf{h}_t | \mathbf{v}_{<t})], \quad \boldsymbol{\theta}_{\rm G} = \{\mathbf{a}, \mathbf{b}, A, B, W\}$$

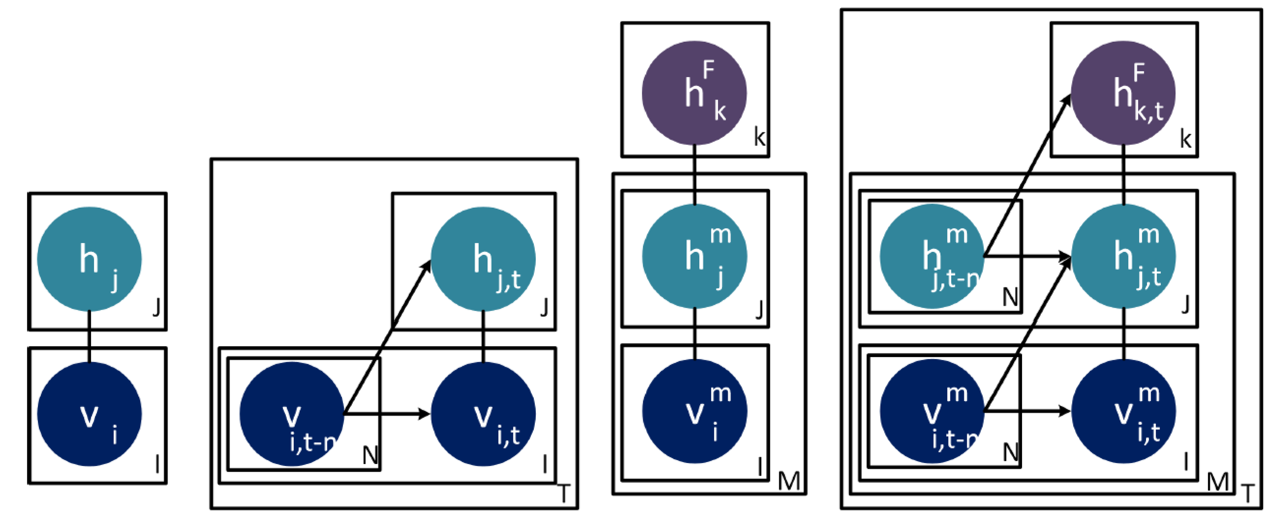
$$E_{\rm G}(\mathbf{v}_t, \mathbf{h}_t | \mathbf{v}_{<t}; \boldsymbol{\theta}_{\rm G}) = \sum_m E_{\rm S}(\mathbf{v}_t^m, \mathbf{h}_t^m | \mathbf{v}_{<t}^m) + E_{\rm F}(\mathbf{h}_t^{1,\ldots,M}, \mathbf{h}_t^F | \mathbf{h}_{<t}^{1,\ldots,M})$$

- Each of the single CRBMs captures the representation of one modality,

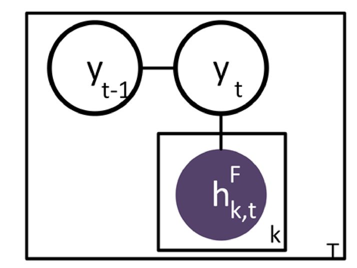
$$E_{\rm S}(\mathbf{v}_t^m, \mathbf{h}_t^m | \mathbf{v}_{<t}^m) = -\textstyle\sum_i (c_{i,t}^m - v_{i,t}^m)^2/2 - \sum_j d_{j,t}^m h_{j,t}^m - \sum_{i,j} v_{i,t}^m w_{i,j}^m h_{j,t}^m$$

- The fusion CRBM combines the representations learned from the different modalities.

$$E_{\rm F}(\mathbf{h}_t^{1,\ldots,M}, \mathbf{h}_t^F | \mathbf{h}_{<t}^{1,\ldots,M}) = -\sum_{i,m} c_{i,t}^m h_{i,t}^m - \sum_j d_{j,t}^F h_{j,t}^F - \sum_{i,j,m} h_{i,t}^m w_{i,j}^F h_{j,t}^F$$



Discriminative – Conditional Random Fields:



- The CRF operates on the features learned using MMCRBM.

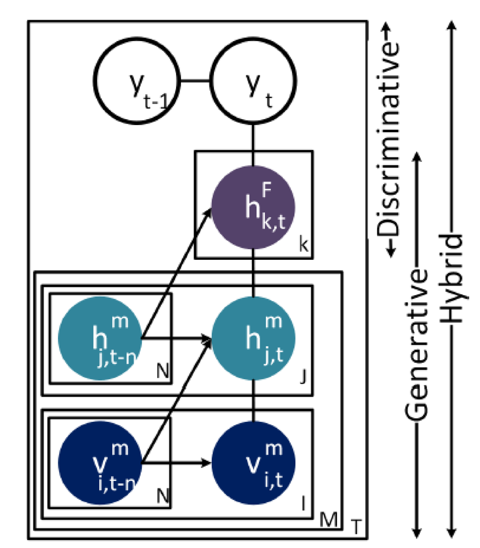
$$p_{\rm D}(\mathbf{y}_t | \mathbf{h}_t^F; \boldsymbol{\theta}_{\rm D}) = \exp[E_{\rm D}(\mathbf{y}_t | \mathbf{h}_t^F; \boldsymbol{\theta}_{\rm D})]/Z(\boldsymbol{\theta}_{\rm D}),$$

$$Z(\boldsymbol{\theta}_{\rm D}) = \sum_{\mathbf{y}} E_{\rm D}(\mathbf{y} | \mathbf{h}_t^F; \boldsymbol{\theta}_{\rm D}), \quad \boldsymbol{\theta}_{\rm D} = \{\boldsymbol{\omega}^1, \boldsymbol{\omega}^2\}$$

$$E_{\rm D}(\mathbf{y}_t | \mathbf{h}_t^F; \boldsymbol{\theta}_{\rm D}) = \sum_j \omega_j^1 f_j^1(\mathbf{y}_{t-1}, \mathbf{y}_t, \mathbf{h}_t^F) + \sum_k \omega_k^2 f_k^2(\mathbf{y}_t, \mathbf{h}_t^F)$$

Hybrid Dynamic Model

$$\underbrace{p(\mathbf{y}_t, \mathbf{v}_t, \mathbf{h}_t | \mathbf{v}_{<t})}_{\text{Hybrid}} = \underbrace{p_D(\mathbf{y}_t | \mathbf{v}_t, \mathbf{h}_t)}_{\text{Discriminative}} \cdot \underbrace{p_G(\mathbf{v}_t, \mathbf{h}_t | \mathbf{v}_{<t})}_{\text{Generative}}$$



Learning

Generative – Contrastive Divergence $\boldsymbol{\theta}_{\rm G} = \{\mathbf{a}, \mathbf{b}, A, B, W\}$

Discriminative – Max. Likelihood Estimation $\boldsymbol{\theta}_{\rm D} = \{\boldsymbol{\omega}^1, \boldsymbol{\omega}^2\}$

Inference (Bottom-Up)

- Activate each modality of the MMCRBM:

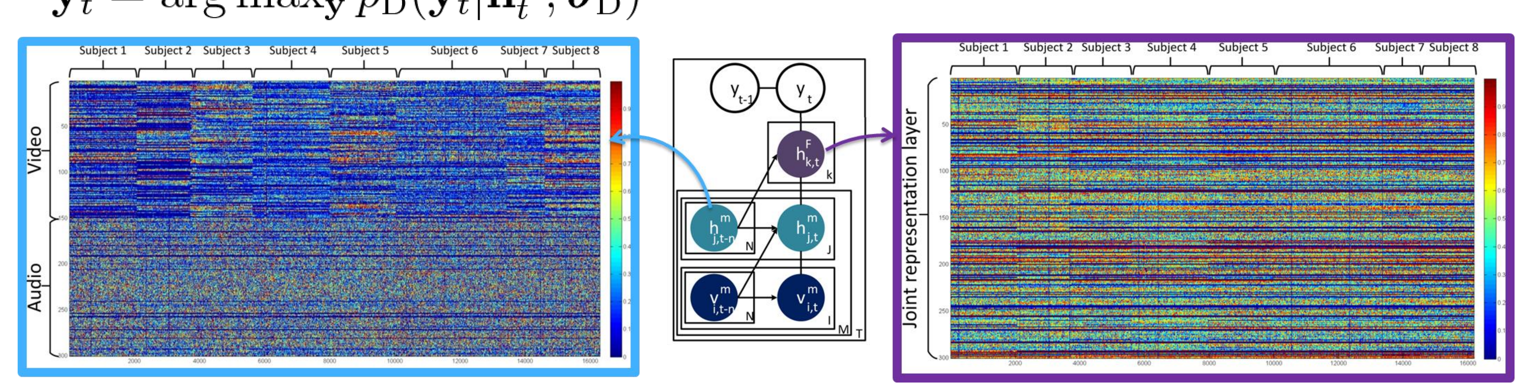
$$p_{\rm G}(h_j^m = 1 | \mathbf{v}^m, \mathbf{v}_{<t}^m) \sim \sigma(c_j^m + \textstyle\sum_i v_i^m w_{ij}^m)$$

- Activate the fusion layer:

$$p_{\rm G}(h_k^F = 1 | \mathbf{h}^{1,\ldots,M}, \mathbf{h}_{<t}^{1,\ldots,M}) \sim \sigma(c_k^F + \textstyle\sum_j h_j^{1,\ldots,M} w_{jk}^F)$$

- Fused features classified by the CRF

$$\mathbf{y}_t = \arg\max_{\mathbf{y}} p_{\rm D}(\mathbf{y}_t | \mathbf{h}_t^F; \boldsymbol{\theta}_{\rm D})$$



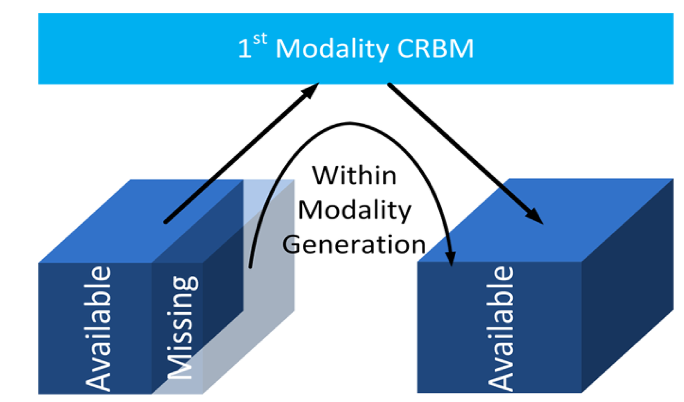
Results and Conclusions

Average Classification Accuracy

- Compare our approach against the relevant baselines and the state-of-the-art on the AVEC, AVLetters and CUAVE datasets.

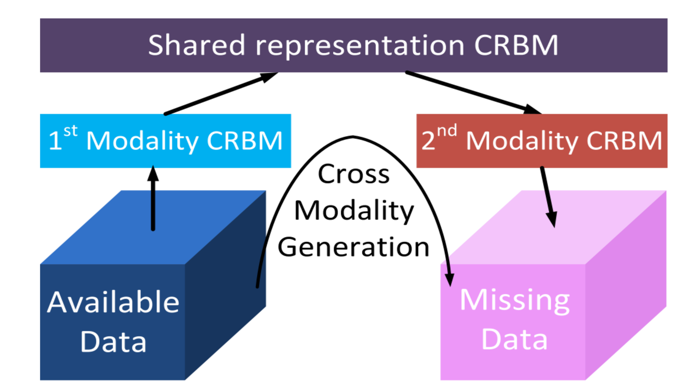
Model/Dataset	AVEC-A	AVEC-V	AVEC-AV	AVLetters-A	AVLetters-V	AVLetters-AV	CUAVE-A	CUAVE-V	CUAVE-AV
SVM-RAW	64.8	62.4	67.4	55.8	56.2	58.5	61.5	58.4	65.0
CRF-RAW	**68.1**	**63.5**	**69.9**	**58.4**	**59.3**	**60.0**	**64.3**	**62.0**	**66.8**
SVM-RBM	61.8	63.9	67.8	58.4	62.1	62.9	65.1	61.8	65.4
CRF-RBM	**67.6**	**65.4**	**68.3**	**62.6**	**64.6**	**63.8**	**67.6**	**65.2**	**68.6**
SVM-CRBM	65.8	66.9	68.2	61.2	62.6	64.8	65.3	64.6	66.7
CRF-CRBM	**69.2**	**70.1**	**70.8**	**66.9**	**64.8**	**67.1**	**67.9**	**66.3**	**69.1**

Missing Data: within Modality



Model/Dataset	AVEC-A	AVEC-V	AVLetters-A	AVLetters-V	CUAVE-A	CUAVE-V
SVM-RBM (0%)	61.8	63.9	58.4	62.1	65.1	61.8
SVM-CRBM (0%)	**65.8**	**66.9**	**61.2**	**62.6**	**65.3**	**64.6**
SVM-RBM (10%)	48.6	46.5	50.7	54.5	59.7	42.8
SVM-CRBM (10%)	**54.9**	**52.1**	**53.6**	**58.2**	**63.1**	**52.6**
SVM-RBM (30%)	35.5	31.2	39.2	32.1	36.1	31.9
SVM-CRBM (30%)	**42.7**	**40.2**	**45.8**	**41.6**	**43.7**	**41.2**

Missing Data: across Modalities



Model/Dataset	AVEC-A‖V	AVEC-V‖A	AVLetters-A‖V	AVLetters-V‖A	CUAVE-A‖V	CUAvE-V‖A
SVM-RBM	31.2	28.2	27.3	25.1	23.1	19.5
SVM-CRBM	**40.4**	**32.1**	**29.6**	**26.5**	**30.7**	**24.4**

Conclusions

- **Hybrid Dynamic Model:** effective for classifying sequential data from multiple heterogeneous modalities.
- **Generative Model (MMCRBM):** Models short-term temporal characteristics and learns a rich feature representation.
- **Discriminative Model (CRF):** Models long range temporal dynamics.

Acknowledgements

•Defense Advanced Research Projects Agency
•Army Research Office