Language Technologies Institute

Carnegie Mellon University

Multimodal Machine Learning: Modeling Human Communication Dynamics

Louis-Philippe Morency

CMU Multimodal Communication and Machine Learning Laboratory [MultiComp Lab]

PhD students: **AmirAli BagherZadeh, Claire Chang, Sayan Ghosh, KangGeon Kim and Sunghyun Park**
Master student: **Liangke Liu**
Project manager: **Stefan Hadricky**

Multimodal Communication and Machine Learning Laboratory

Algorithms to analyze, recognize and predict human subtle communicative behaviors in social context.

Human Multimodal Communication

- "[…] music online. Guy Kewney is the editor of the technology website News Wireless. Hello, good morning to you."
- "Good morning."
- "Were you surprised by this verdict today?"
- "I'm very surprised to see this verdict to come on me. Because I was not expecting that. When I came they told me something else and I'm coming. And they told me something else. Big surprise any way."
- "A big surprise…"
- "Exactly."
- "Yeah yeah. With regard to the cost that is involved. Do you think more people will be downloading online?"
- "Actually if you can go everywhere, you gonna see people downloading through the internet and the websites. […]"

Human Multimodal Communication

Human Communicative Behaviors

Visual

- **Gestures**
  - Head gestures
  - Eye gestures
  - Arm gestures
- **Body language**
  - Body posture
  - Proxemics
- **Eye contact**
  - Head gaze
  - Eye gaze
- **Facial expressions**
  - FACS action units
  - Smile, frowning

Verbal

- **Lexicon**
  - Words
- **Syntax**
  - Part-of-speech
  - Dependencies
- **Pragmatics**
  - Discourse acts

Auditory

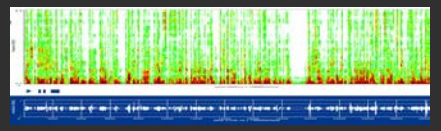
- **Prosody**
  - Intonation
  - Voice quality
- **Vocal expressions**
  - Laughter, moans

Enabling Sensing Technologies

- Low-cost sensor
- Body tracking
- Speech recognition

ANN_{OQ} [ICASSP 2013, Interspeech 2013]

- Robust voice quality estimation



GAVAM [FG 2008, best paper award]

- 2D/3D head pose estimation

CLM-Z [CVPR 2012]

- 3D facial feature tracker

CLNF [ICCV-W 2013, ECCV 2014]

- Robust facial landmark detection

Broad Applicability

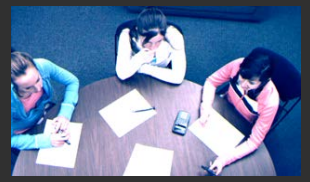
Medical

Depression and PTSD
with MIT, BBN and Cogito

Suicide prevention
with Cincinnati Hospital

Autistic children
with Yale University

Education



Group learning analytics
with Stanford and UCSD

Virtual Learning Peer
with CMU

Public speaking training
with Smartbody team

Online

Opinion mining
with Univ. of Michigan

Social influence
with EPFL

Negotiation outcomes
with Microsoft Research

Disorders
- Depression
- Distress
- Autism

Social
- Leadership
- Empathy
- Engagement

Emotion
- Sentiment
- Persuasion
- Frustration

Broad Applicability

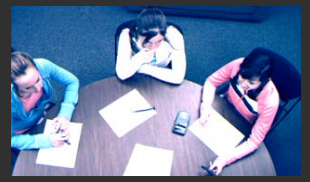
Medical

Depression and PTSD
with MIT, BBN and Cogito

Suicide prevention
with Cincinnati Hospital

Autistic children
with Yale University

Education



Group learning analytics
with Stanford and UCSD

Virtual Learning Peer
with CMU

Public speaking training
with Smartbody team

Online

Opinion mining
with Univ. of Michigan

Social influence
with EPFL

Negotiation outcomes
with Microsoft Research

Disorders
- Depression
- Distress
- Autism

Social
- Leadership
- Empathy
- Engagement

Emotion
- Sentiment
- Persuasion
- Frustration

Multi-Disciplinary Research Topic

Speech

- Audio-visual speech recognition
- Multimodal language processing

Language

Multimedia

- Cross-media event retrieval and tagging
- Image and video captioning

Vision

Affective

- Audio-visual emotion recognition
- Multimodal deep learning

Learning

Multi-Disciplinary Research Topic

Speech

Multimedia

Affective

- Audio-vi… speech recogn…
- Multim… langua… processi…

…dio-visual …tion …gnition

…modal …p learning

Language

Vision

Learning

A Central Challenge:

Modeling Human Communication Dynamics

Behavioral

Multi-Disciplinary Research Topic

Speech

- Audio-vi
  speech
  recogn

Multimedia

Affective

- dio-visual
  tion
  gnition

- Multim
  langua
  processi

- modal
  p learning

Language

Learning

A Central Challenge:

Modeling Human Communication Dynamics

Behavioral

50 shades of "*yeah*"

Multi-Disciplinary Research Topic

Speech

Multimedia

Affective

- Audio-vi… speech recogn…
- …dio-visual …tion …gnition
- Multim… langua… processi…
- …modal …p learning

Language

Vision

Learning

A Central Challenge:

Modeling Human Communication Dynamics

Behavioral

Multimodal

Interpersonal

Societal

Multi-Disciplinary Research Topic

Speech

Multimedia

Affective

- Audio-vi… speech recogn…
- …dio-visual …tion …gnition
- Multim… langua… processi…
- …modal …p learning

Language

Vision

Learning

A Central Challenge:

Modeling Human Communication Dynamics

Behavioral

Multimodal

Interpersonal

Societal

Multimodal Machine Learning

Verbal

We saw the yellow dog

Visual

Acoustic

Multimodal Probabilistic Learning

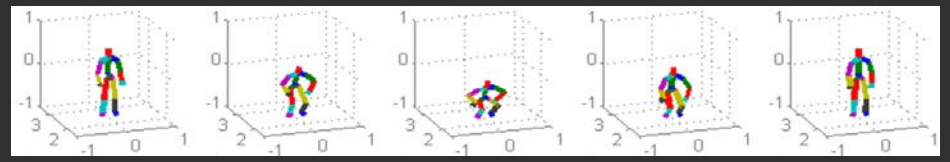
Disorders
- Depression
- Distress
- Autism

Social
- Leadership
- Empathy
- Engagement

Emotion
- Sentiment
- Persuasion
- Frustration

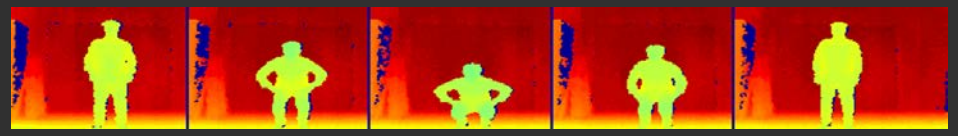
“Multimodal” Face and Gesture Recognition

Skeleton



Intensity

Depth



Multimodal Probabilistic Learning

Disorders
- Depression
- Distress
- Autism

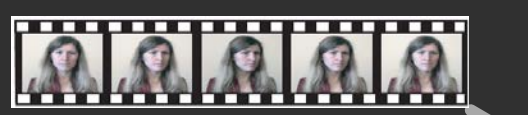
Social
- Leadership
- Empathy
- Engagement

Emotion
- Sentiment
- Persuasion
- Frustration

Unique Research Questions

Verbal

We saw the yellow dog



Visual

Acoustic

1. **Hidden temporal structure**
   - How to learn the hidden structure in spoken utterances, gestures or vocal patterns?
   - Can we automatically identify shared components in this hidden structure?
   - How can we model long-range dependencies in the hidden structure?
2. **Multimodal representation**
   - How to model the nonlinear relationships between multimodal features?
   - How to learn a joint representation from multiple modalities?
   - How to represent different granularities in unimodal or multimodal streams?
3. **Multimodal synchrony and complementarity**
   - How to improve recognition by taking advantage of multimodal complementarity?
   - How can we model the synchrony between multimodal streams?
4. **Multimodal behavior interpretation**
   - How to model the "fuzziness" in people interpretation of multimodal behaviors?
   - How to jointly learn multiple emotions and behaviors (multi-task learning)?

Talk Outline

Structure

Representation

Dynamics

Interpretation

1. **Hidden temporal structure**
   - How to learn the hidden structure in spoken utterances, gestures or vocal patterns?
   - Can we automatically identify shared components in this hidden structure?
   - How can we model long-range dependencies in the hidden structure?
2. **Multimodal representation**
   - How to model the nonlinear relationships between multimodal features?
   - How to learn a joint representation from multiple modalities?
   - How to represent different granularities in unimodal or multimodal streams?
3. **Multimodal synchrony and complementarity**
   - How to improve recognition by taking advantage of multimodal complementarity?
   - How can we model the synchrony between multimodal streams?
4. **Multimodal behavior interpretation**
   - How to model the "fuzziness" in people interpretation of multimodal behaviors?
   - How to jointly learn multiple emotions and behaviors (multi-task learning)?

Modeling Behavioral Dynamics

Behavioral

Structure

Verbal

Noun phrase Noun phrase

We saw the yellow dog

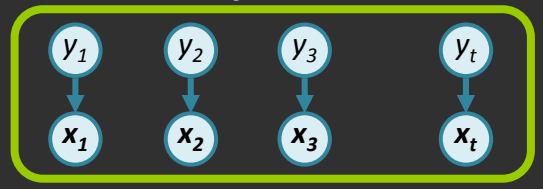
Visual

Head shake Nod

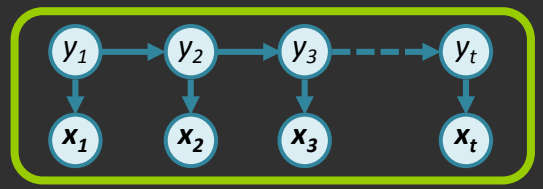
Acoustic

Tensed voice Breathy

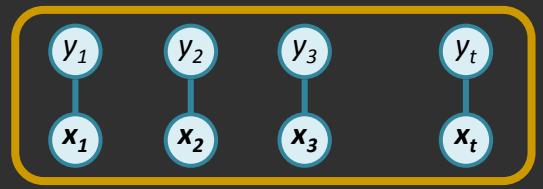
Naïve Bayes Classifier



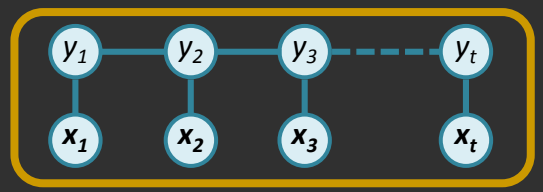
Hidden Markov Model



Maximum Entropy Model & Support Vector Machine



Conditional Random Field



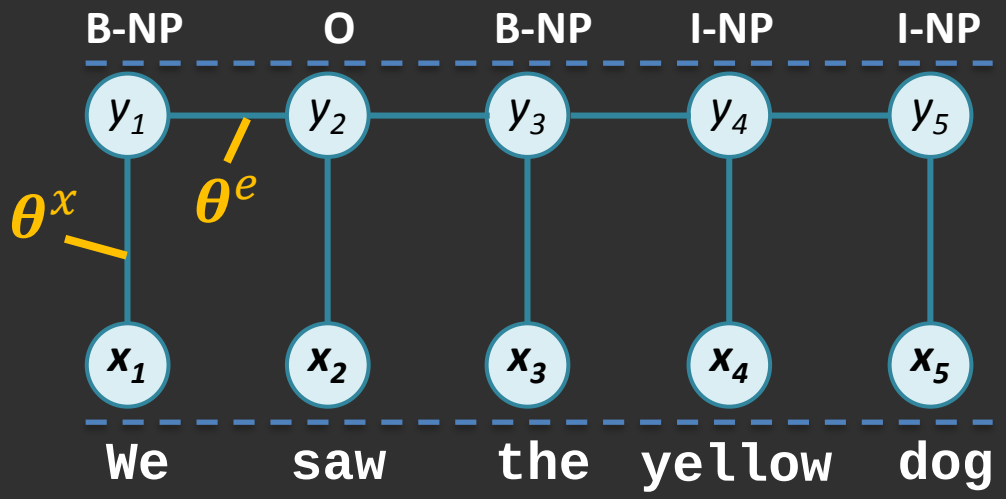
Conditional Random Field [Laferty et al., 2001]

Random variables **(e.g., noun phrase labels: {B-NP,I-NP,O})**

$$\boldsymbol{y} = \{y_1, y_2, y_3, \dots, y_t\} \quad \text{where } y_t \in \mathcal{Y}$$

Observations **(e.g., words)**

$$\boldsymbol{x} = \{\boldsymbol{x_1}, \boldsymbol{x_2}, \boldsymbol{x_3}, \dots, \boldsymbol{x_t}\} \quad \text{where } \boldsymbol{x_t} \in \mathbb{R}^d$$



$$p(\mathbf{y} \mid \mathbf{x}; \boldsymbol{\theta}) = \frac{1}{Z(\mathbf{x}; \boldsymbol{\theta})} \exp\left\{ \sum_t \boldsymbol{\theta}^x \cdot f^x(y_t, \mathbf{x_t}) + \sum_t \boldsymbol{\theta}^e \cdot f^e(y_t, y_{t-1}) \right\}$$

Singular potentials — Pairwise potentials

Regularization

$$\text{Solve} \quad \boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{\text{argmax}} \sum_{i=1}^{N} \log p(\mathbf{y}_i \mid \mathbf{x}_i; \boldsymbol{\theta}) - \lambda \, ||\boldsymbol{\theta}||^2$$

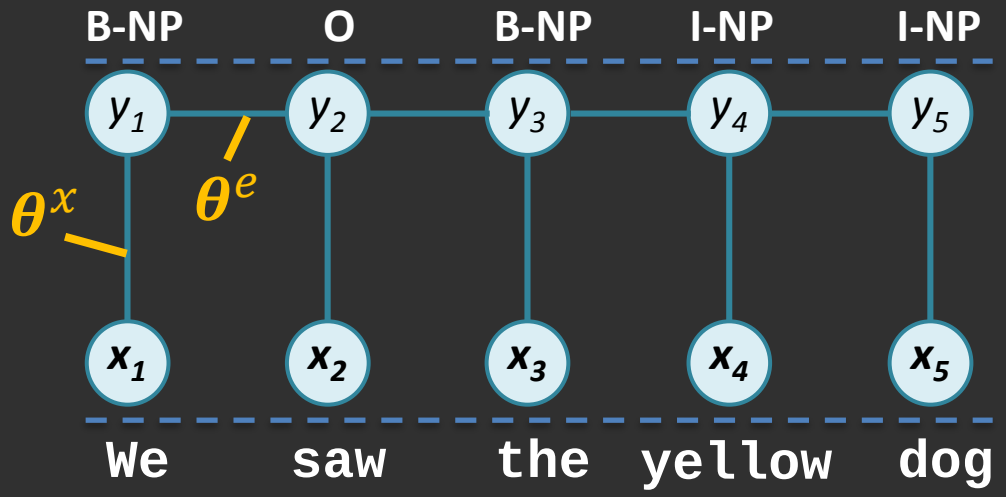
Conditional Random Field [Laferty et al., 2001]

Random variables **(e.g., noun phrase labels: {B-NP,I-NP,O})**

$$\boldsymbol{y} = \{y_1, y_2, y_3, \ldots, y_t\} \quad \text{where } y_t \in \mathcal{Y}$$

Observations **(e.g., words)**

$$\boldsymbol{x} = \{\boldsymbol{x_1}, \boldsymbol{x_2}, \boldsymbol{x_3}, \ldots, \boldsymbol{x_t}\} \quad \text{where } \boldsymbol{x_t} \in \mathbb{R}^d$$



$$p(\mathbf{y} \mid \mathbf{x}; \boldsymbol{\theta}) = \frac{1}{Z(\mathbf{x}; \boldsymbol{\theta})} \exp\left\{ \underbrace{\sum_t \boldsymbol{\theta}^x \cdot f^x(y_t, \mathbf{x_t})}_{\text{Singular potentials}} + \underbrace{\sum_t \boldsymbol{\theta}^e \cdot f^e(y_t, y_{t-1})}_{\text{Pairwise potentials}} \right\}$$

Regularization

$$\text{Solve} \quad \boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{\text{argmax}} \sum_{i=1}^{N} \log p(\mathbf{y}_i \mid \mathbf{x}_i; \boldsymbol{\theta}) \; \underbrace{-\lambda \, ||\boldsymbol{\theta}||^2}$$

- **But do not model hidden structure…**

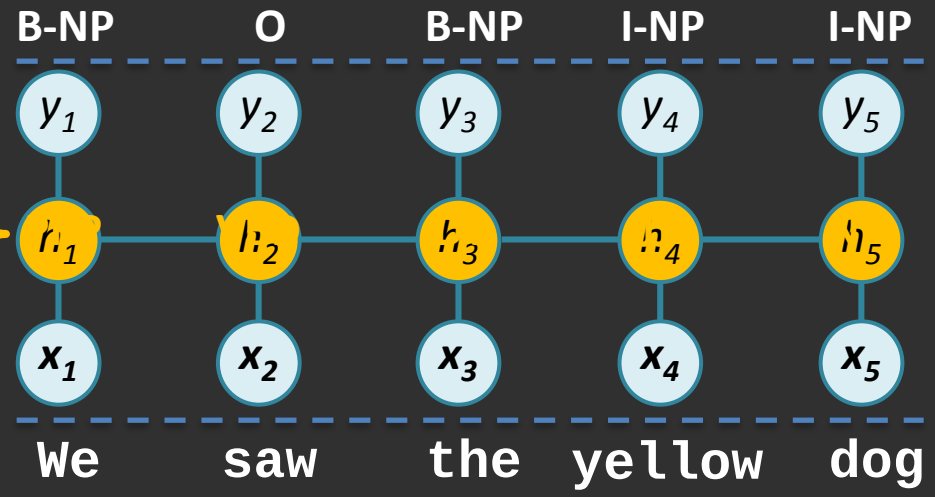
Latent-Dynamic Conditional Random Field [CVPR 2007, COLING 2008]

Latent variables **(e.g., POS tags)**

$$\boldsymbol{h} = \{h_1, h_2, h_3, \dots, h_t\} \quad \text{where } h_t \in \{\mathcal{H}_{y_t}\}$$

For example: $\mathcal{H} = \{\mathcal{H}_{\text{B-NP}}, \mathcal{H}_{\text{I-NP}}, \mathcal{H}_{\text{O}}\}$

$$\mathcal{H} = \{B_1, B_2, B_3, B_4, I_1, I_2, I_3, I_4, O_1, O_2, O_3, O_4\}$$



$$p(\boldsymbol{y}|\,\boldsymbol{x};\boldsymbol{\theta}) = \sum_{\boldsymbol{h}} p(\boldsymbol{y}|\boldsymbol{h};\boldsymbol{\theta})p(\boldsymbol{h}|\boldsymbol{x};\boldsymbol{\theta}) \qquad \text{where} \quad p(\boldsymbol{y}|\boldsymbol{h};\boldsymbol{\theta}) = \begin{cases} 1 & \text{if } \forall h_t \in \mathcal{H}_{y_t} \\ 0 & \text{otherwise} \end{cases}$$

$$= \sum_{\boldsymbol{h}:\forall h_t \in \mathcal{H}_{y_t}} p(\boldsymbol{h}|\boldsymbol{x};\boldsymbol{\theta}) = \sum_{\boldsymbol{h}:\forall h_t \in \mathcal{H}_{y_t}} \frac{1}{\mathcal{Z}(\mathbf{x};\boldsymbol{\theta})} \exp\left\{\sum_t \boldsymbol{\theta}^x \cdot f^x(h_t, \mathbf{x_t}) + \sum_t \boldsymbol{\theta}^e \cdot f^e(h_t, h_{t-1})\right\}$$

Same as CRF!

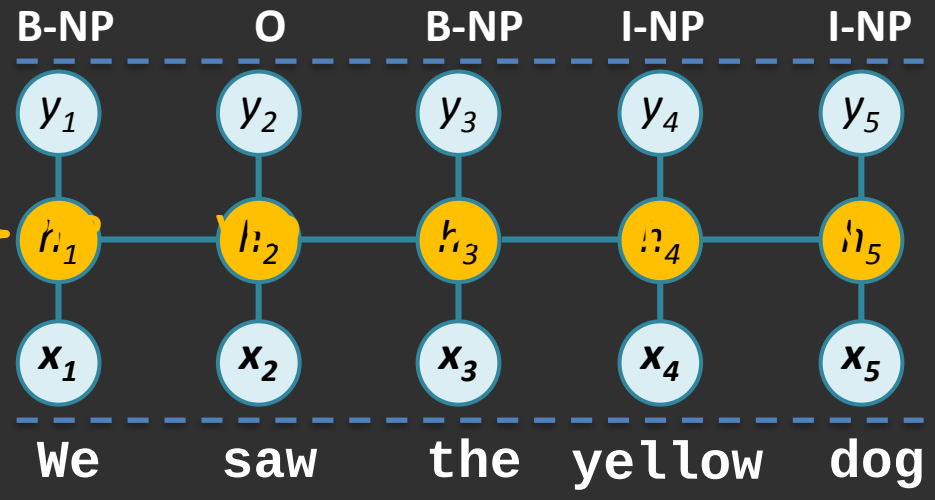
Latent-Dynamic Conditional Random Field [CVPR 2007, COLING 2008]

Latent variables (e.g., POS tags)

$\boldsymbol{h} = \{h_1, h_2, h_3, \ldots, h_t\}$ where $h_t \in \{\mathcal{H}_{y_t}\}$

For example: $\mathcal{H} = \{\mathcal{H}_{\text{B-NP}}, \mathcal{H}_{\text{I-NP}}, \mathcal{H}_{\text{O}}\}$

$\mathcal{H} = \{B_1, B_2, B_3, B_4, I_1, I_2, I_3, I_4, O_1, O_2, O_3, O_4\}$



$$p(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\theta}) = \sum_{\boldsymbol{h}} p(\boldsymbol{y}|\boldsymbol{h};\boldsymbol{\theta}) p(\boldsymbol{h}|\boldsymbol{x};\boldsymbol{\theta}) \quad \text{where} \quad p(\boldsymbol{y}|\boldsymbol{h};\boldsymbol{\theta}) = \begin{cases} 1 & \text{if } \forall h_t \in \mathcal{H}_{y_t} \\ 0 & \text{otherwise} \end{cases}$$

$$= \sum_{\boldsymbol{h}:\forall h_t \in \mathcal{H}_{y_t}} p(\boldsymbol{h}|\boldsymbol{x};\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{x})} \exp\left\{\sum_t \boldsymbol{\theta}^x \cdot f^x(h_t, \boldsymbol{x}_t) + \sum_t \boldsymbol{\theta}^e \cdot f^e(h_t, h_{t-1})\right\}$$

- Edge potential parameters θ^e encode:
  1) Intrinsic dynamics (within a label)
  2) Extrinsic dynamics (between labels)

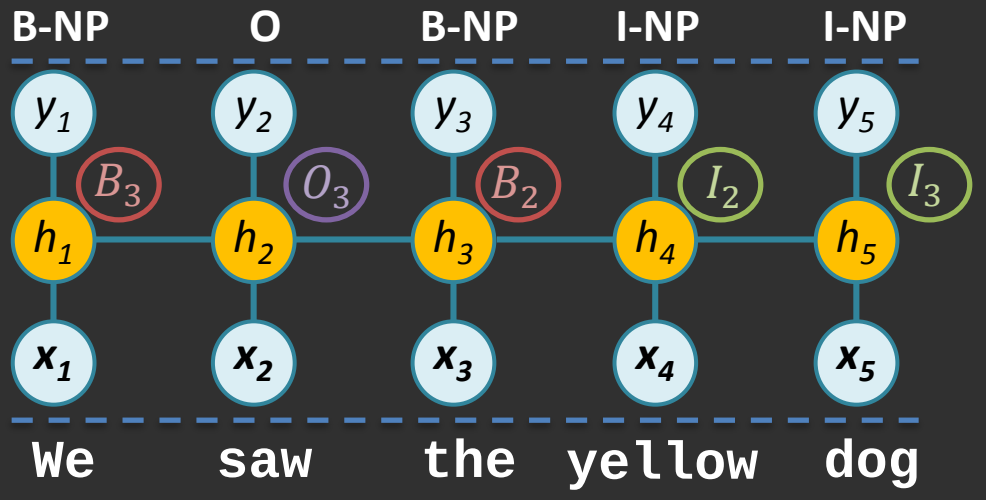
Latent-Dynamic Conditional Random Field [CVPR 2007, COLING 2008]

Latent variables **(e.g., POS tags)**

$$\boldsymbol{h} = \{h_1, h_2, h_3, \ldots, h_t\} \quad \text{where } h_t \in \{\mathcal{H}_{y_t}\}$$

For example: $\mathcal{H} = \{\mathcal{H}_{\mathrm{B-NP}}, \mathcal{H}_{\mathrm{I-NP}}, \mathcal{H}_{\mathrm{O}}\}$

$$\mathcal{H} = \{B_1, B_2, B_3, B_4, I_1, I_2, I_3, I_4, O_1, O_2, O_3, O_4\}$$



Experiment – Analyzing latent variables

- **Task:** Shallow parsing with CoNLL 2000 dataset
- **Input features:** word feature only
- **Output labels:** Noun phrase labels

1) Select hidden state a^* with highest marginal:

$$a^* = \arg\max_a p(\boldsymbol{h_t} = a | \boldsymbol{x}; \boldsymbol{\theta})$$

2) Compute relative frequency for each word

Label	State	Words	POS	Freq.
B	B_1	*That*	*WDT*	0.85
		who	*WP*	0.49
		Who	*WP*	0.33
	B_2	*any*	*DT*	1.00
		an	*DT*	1.00
		a	*DT*	0.98
	B_3	*They*	*PRP*	1.00
		we	*PRP*	1.00
		he	*PRP*	1.00
	B_4	*Nasdaq*	*NNP*	1.00
		Florida	*NNP*	0.99
		cities	*NNS*	0.99

Label	State	Words	POS	Freq.
O	O_1	*but*	*CC*	0.88
		by	*IN*	0.73
		or	*IN*	0.67
	O_2	*4.6*	*CD*	1.00
		1	*CD*	1.00
		1 1	*CD*	0.62
	O_3	*were*	*VBD*	0.94
		rose	*VBD*	0.93
		have	*VBP*	0.92
	O_4	*been*	*VBN*	0.97
		be	*VB*	0.94
		to	*TO*	0.92

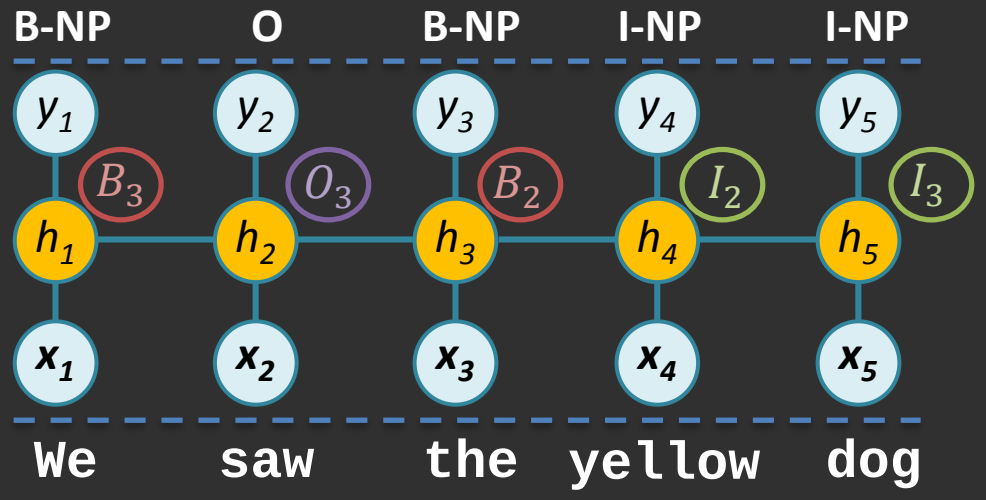
Latent-Dynamic Conditional Random Field [CVPR 2007, COLING 2008]

Latent variables **(e.g., POS tags)**

$$\boldsymbol{h} = \{h_1, h_2, h_3, \ldots, h_t\} \quad \text{where } h_t \in \{\mathcal{H}_{y_t}\}$$

For example: $\mathcal{H} = \{\mathcal{H}_{\mathrm{B-NP}}, \mathcal{H}_{\mathrm{I-NP}}, \mathcal{H}_{\mathrm{O}}\}$

$$\mathcal{H} = \{B_1, B_2, B_3, B_4, I_1, I_2, I_3, I_4, O_1, O_2, O_3, O_4\}$$



Experiment – Analyzing latent variables

- **Task:** Shallow parsing with CoNLL 2000 dataset
- **Input f**
- **Outpu**

1) Sele

2) Com

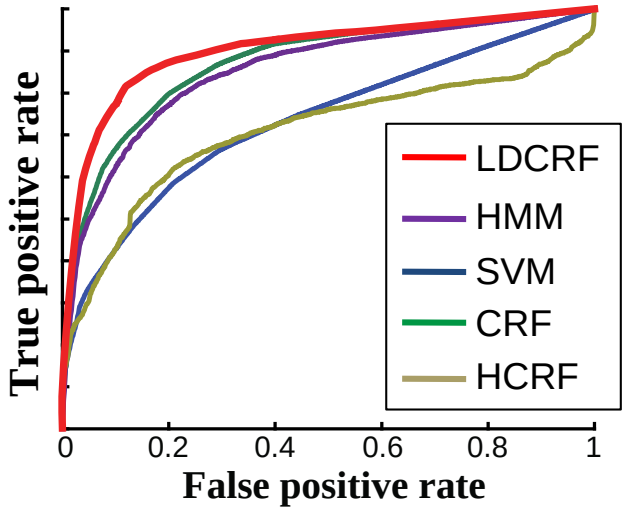
Label	State	Words	POS	Freq.
				0.88
				0.73
				0.67
				1.00
				1.00
				0.62
				0.94
				0.93
				0.92
				0.97
				0.94
				0.92

Experiments - Comparison with other models

- **Task:** Head and eye gesture recognition
- **Input:** Head and eye gaze direction
- **Output:** Head nods, shakes or gaze aversion

➢ Simultaneous recognition and segmentation

➢ Real-time inference performance



How to model only one sequence label?

Lan

ersity

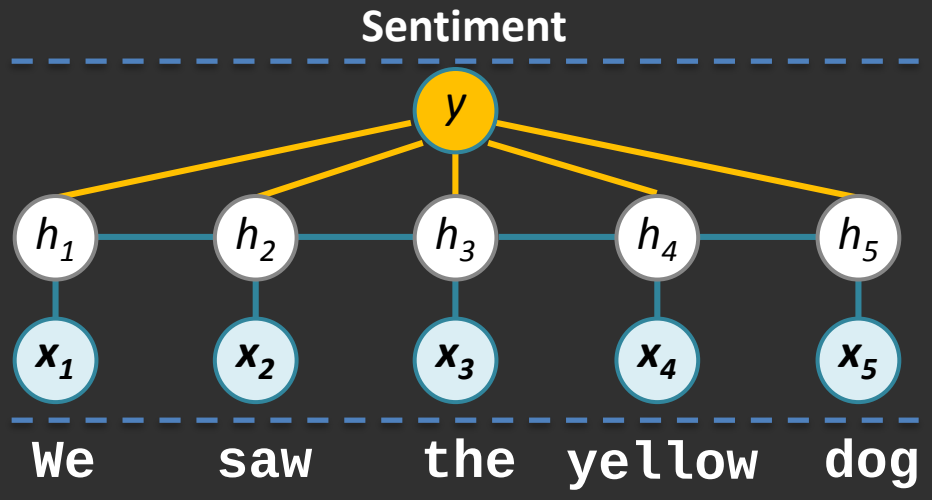
Hidden Conditional Random Field [PAMI 2007, CVPR 2006]

Sequence label:

$y \in \mathcal{Y}$ for example, $\mathcal{Y}$: {positive, negative}

Latent variables with shared hidden states:

$\boldsymbol{h} = \{h_1, h_2, h_3, \ldots, h_t\}$ where $h_t \in \mathcal{H}$



$$p(y, \boldsymbol{h} \mid \mathbf{x}; \boldsymbol{\theta}) = \frac{1}{\mathcal{Z}(\boldsymbol{x}; \boldsymbol{\theta})} \exp\left\{\sum_t \boldsymbol{\theta}^x \cdot f^x(h_t, \mathbf{x_t}) + \sum_t \underbrace{\boldsymbol{\theta}^e \cdot f^e(h_t, h_{t-1}, y)}_{\text{Different edge potentials for each label } y} + \sum_t \underbrace{\boldsymbol{\theta}^y \cdot f^y(y, \boldsymbol{h_t})}_{\text{Shared hidden states}}\right\}$$

$$p(y \mid \boldsymbol{x}; \boldsymbol{\theta}) = \sum_{\boldsymbol{h}} p(y, \boldsymbol{h} \mid \boldsymbol{x}; \boldsymbol{\theta})$$

Modeling Multimodal Dynamics

Multimodal

Representation

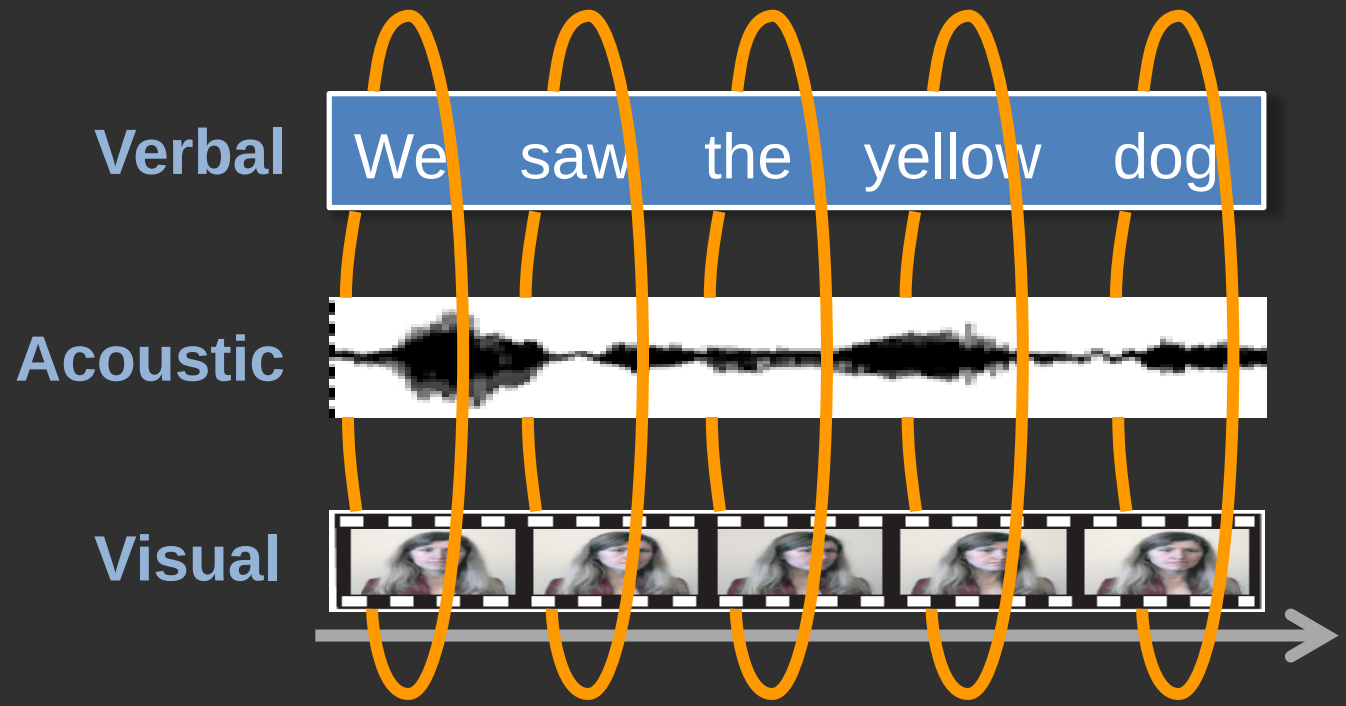
- Nonlinear

Dynamics

Interpretation

Multimodal representations

- How to model the nonlinear relationships between multimodal features?



Multimodal representation ≈ Instantaneous Fusion

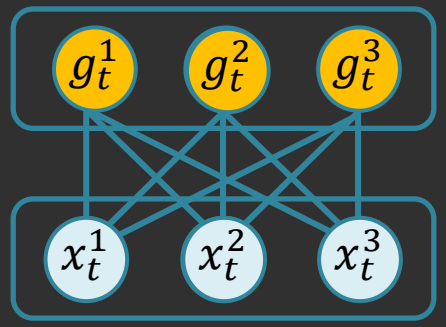
Latent-Dynamic Conditional Neural Fields [FG 2013]

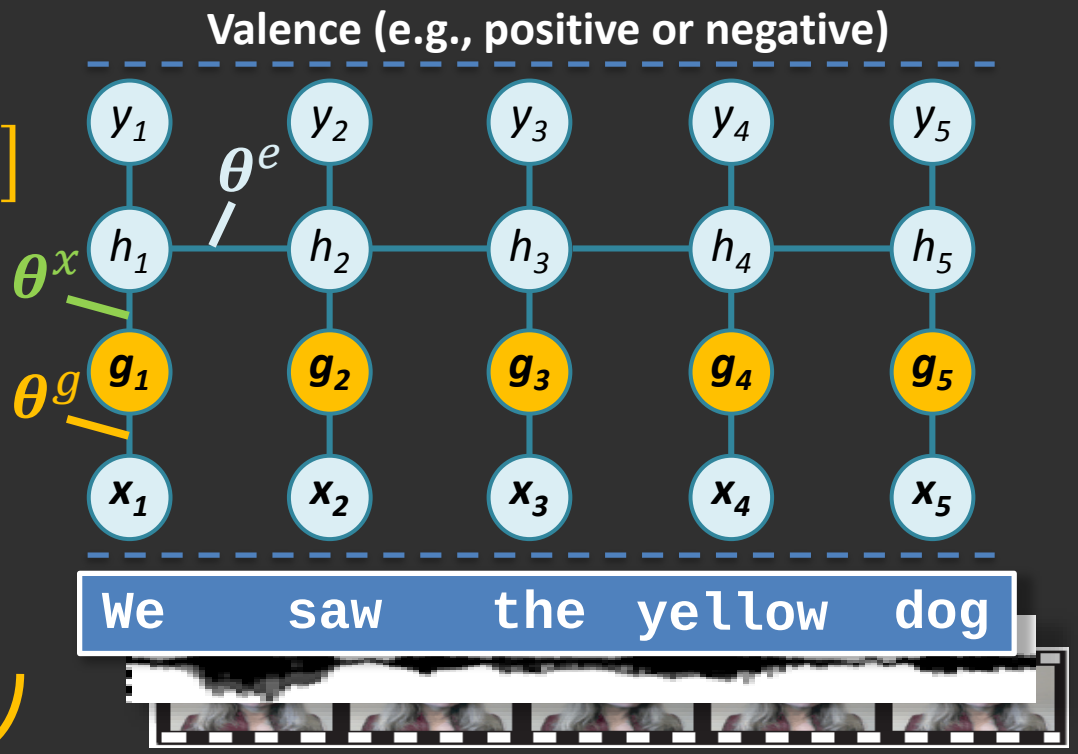
Gate functions:

$$\mathcal{G}(\boldsymbol{x_t}, \boldsymbol{\theta^g}) = [g_1(\boldsymbol{x_t} \cdot \boldsymbol{\theta_1^g}), g_2(\boldsymbol{x_t} \cdot \boldsymbol{\theta_2^g}), \dots, g_n(\boldsymbol{x_t} \cdot \boldsymbol{\theta_n^g})]$$

Neuron activation function:

$$g(z) = \frac{1}{(1 + \exp(-z))}$$





$$f^x(h_t, \mathbf{x}_t) = \mathbb{I}[h_t = h'] \cdot \mathcal{G}(\boldsymbol{x}_t, \boldsymbol{\theta^g})$$

$$p(\boldsymbol{y}|\, \boldsymbol{x}; \boldsymbol{\theta}) = \sum_{\boldsymbol{h}:\forall h_t \in \mathcal{H}_{y_t}} \frac{1}{\mathcal{Z}(\mathbf{x}; \boldsymbol{\theta})} \exp\left\{\sum_t \boldsymbol{\theta^x} \cdot f^x(h_t, \mathbf{x_t}) + \sum_t \boldsymbol{\theta^e} \cdot f^e(h_t, h_{t-1})\right\} + \lambda\, R(\theta^x)$$

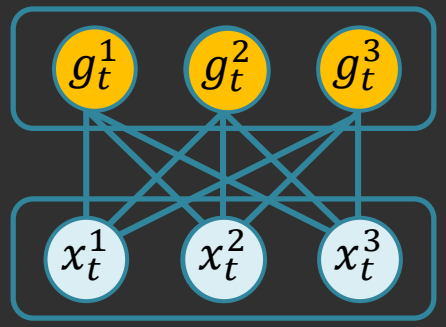
Latent-Dynamic Conditional Neural Fields [FG 2013]

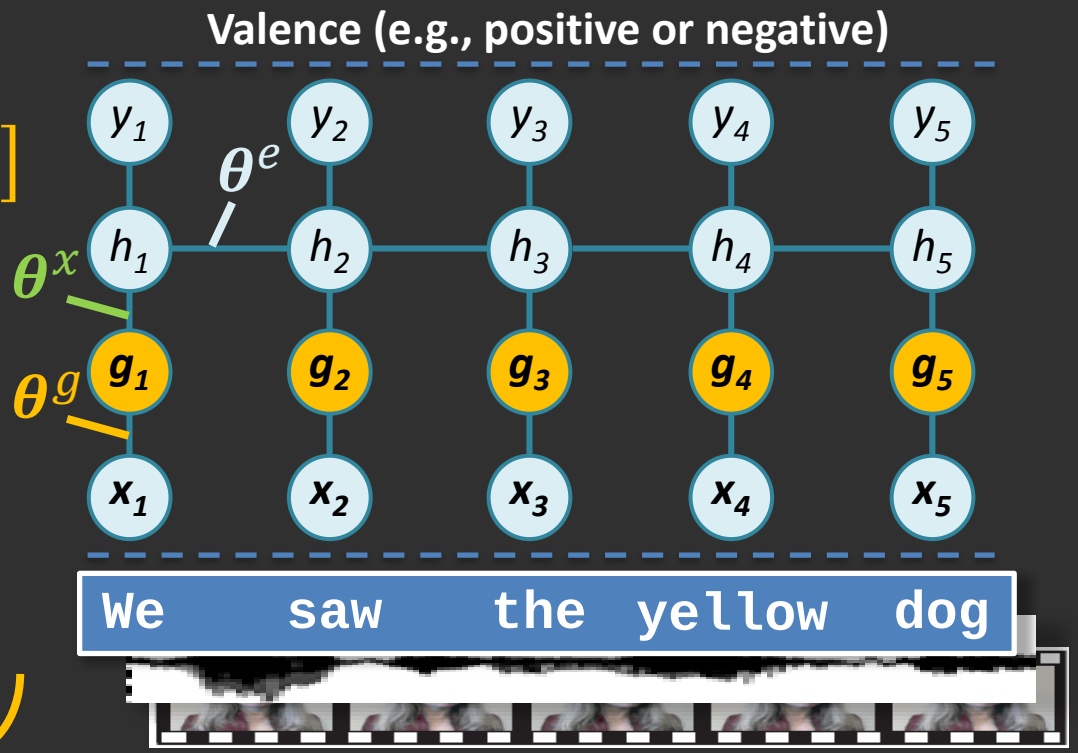
Gate functions:

$$\mathcal{G}(\boldsymbol{x_t}, \boldsymbol{\theta^g}) = [g_1(\boldsymbol{x_t} \cdot \boldsymbol{\theta_1^g}), g_2(\boldsymbol{x_t} \cdot \boldsymbol{\theta_2^g}), \dots, g_n(\boldsymbol{x_t} \cdot \boldsymbol{\theta_n^g})]$$

Neuron activation function:

$$g(z) = \frac{1}{(1 + \exp(-z))}$$





$$f^x(h_t, \mathbf{x}_t) = \mathbb{I}[h_t = h'] \cdot \mathcal{G}(\boldsymbol{x_t}, \boldsymbol{\theta^g})$$

$$p(\boldsymbol{y}|\, \boldsymbol{x}; \boldsymbol{\theta}) = \sum_{\boldsymbol{h}:\forall h_t \in \mathcal{H}_{y_t}} \frac{1}{\mathcal{Z}(\mathbf{x}; \boldsymbol{\theta})} \exp \left\{ \dots , h_{t-1}) \right\} + \lambda\, R(\theta^x)$$

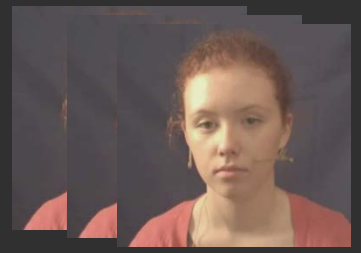
Sparsity constraint:

$$R(\theta^x) = \sum_{j=1}^{m} \sum_{k=j+1}^{m} \theta_{h_j}^x \cdot \theta_{h_k}^x$$

Latent-Dynamic Conditional Neural Fields [FG 2013]

Experiments

Dataset: Audio-Visual Emotion Challenge (AVEC)



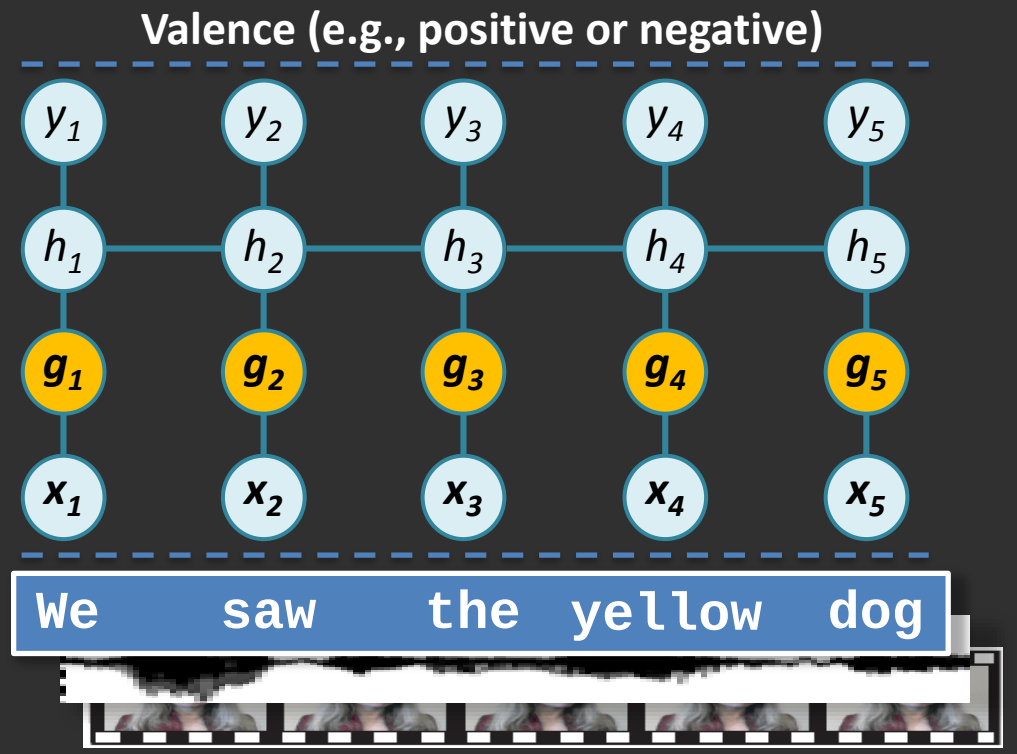
- 63 videos
- Continuous emotion labels
- 2+ coders per label

Input features:

- Visual: Smiles, gaze, head pose
- Acoustic: Pauses, pitch, energy, MFCC

Output labels:

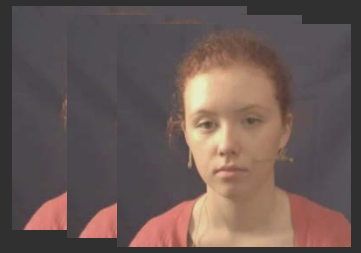
- Emotional arousal (discrete label)



Latent-Dynamic Conditional Neural Fields [FG 2013]

Experiments

Dataset: Audio-Visual Emotion Challenge (AVEC)

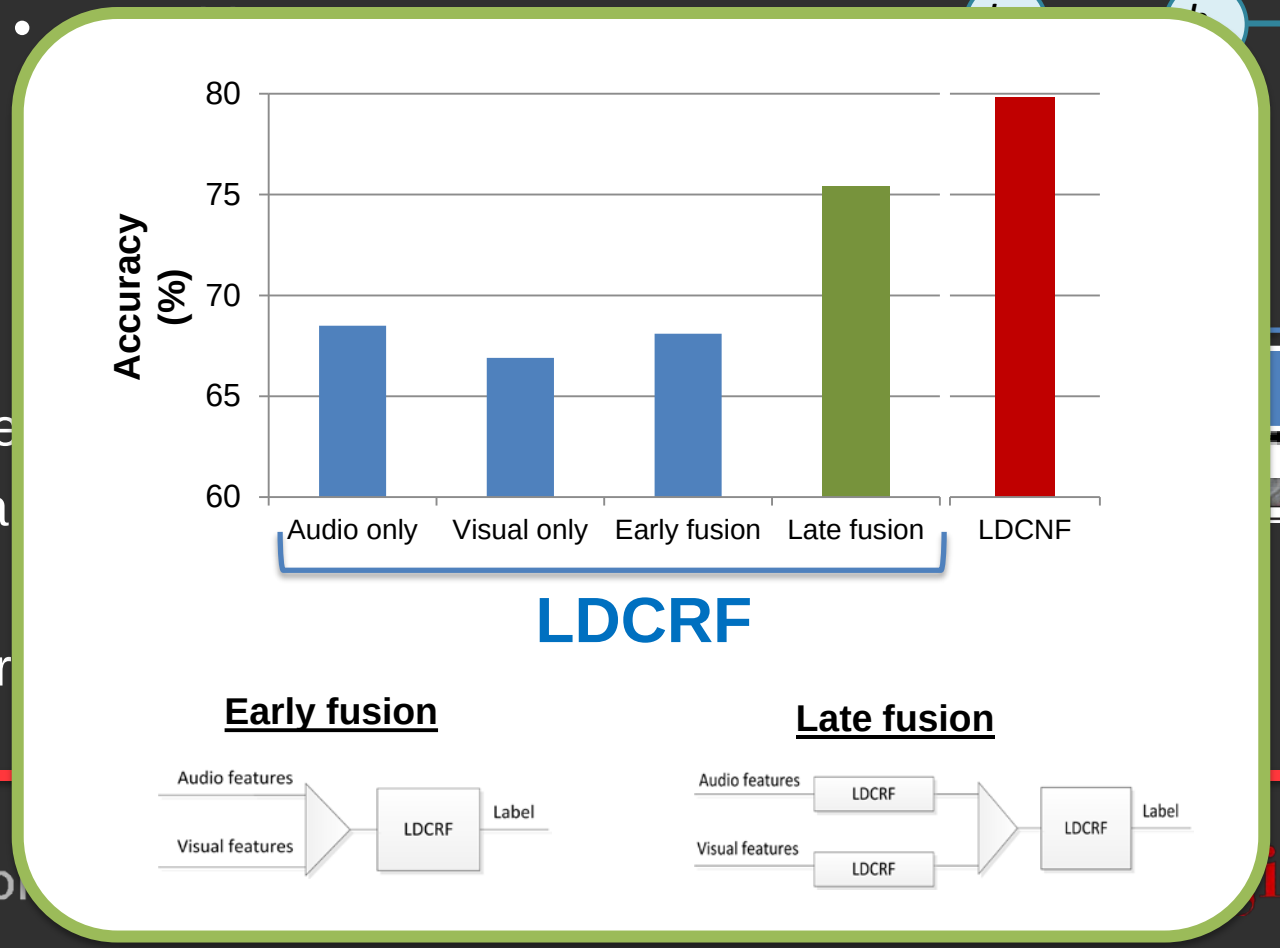


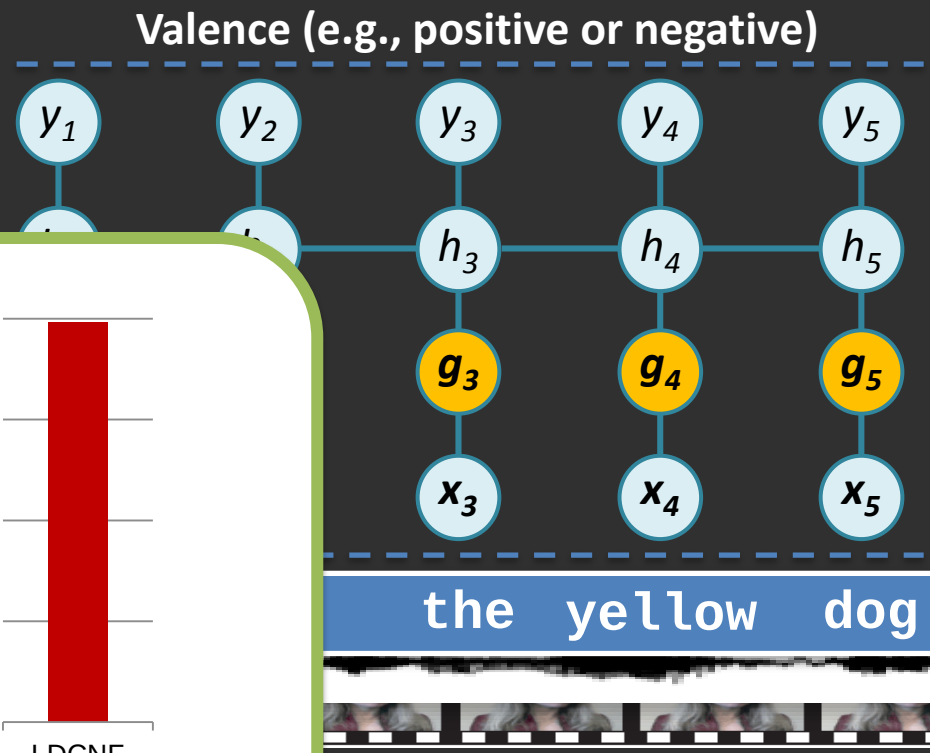
Input features:

- Visual: Smile
- Acoustic: Pa

Output labels:

- Emotional ar





Deep Conditional Neural Fields [IVA 2015-in review]

Multiple layers of gate functions:

$$\mathcal{G}^l(\boldsymbol{x_t}, \boldsymbol{\theta^g}) = [g_1^l(\boldsymbol{x_t} \cdot \boldsymbol{\theta_1^g}), g_2^l(\boldsymbol{x_t} \cdot \boldsymbol{\theta_2^g}), \dots, g_n^l(\boldsymbol{x_t} \cdot \boldsymbol{\theta_n^g})]$$

$$p(\mathbf{y} \mid \mathbf{x}; \boldsymbol{\theta}) \propto \exp\left\{\sum_t \boldsymbol{\theta}^x \cdot f^x(y_t, \mathbf{x_t}) + \sum_t \boldsymbol{\theta}^e \cdot f^e(y_t, y_{t-1})\right\}$$

$$f^x(y_t, \mathbf{x}_t) = \mathbb{I}[y_t = y'] \cdot \mathcal{G}(\boldsymbol{a_t^{m-1}}, \boldsymbol{\theta^g})$$

$$a^i = \mathcal{G}(\boldsymbol{a_t^{i-1}}, \boldsymbol{\theta^g}) \quad \text{for } i = 2 \dots m-1$$

Iterate

Deep Conditional Neural Fields [IVA 2015-in review]

Experiments

Dataset: Distress Assessment Interaction Dataset

- 15 videos
- 12 co-verbal gestures
- 2+ coders per label

Input features:

- Text: Word unigrams
- Acoustic: Pitch, energy, voice quality

Output labels:

- Visual gestures (for virtual human)

Deep Conditional Neural Fields [IVA 2015-in review]

Experiments

Dataset: Distress Assessment Interaction Dataset

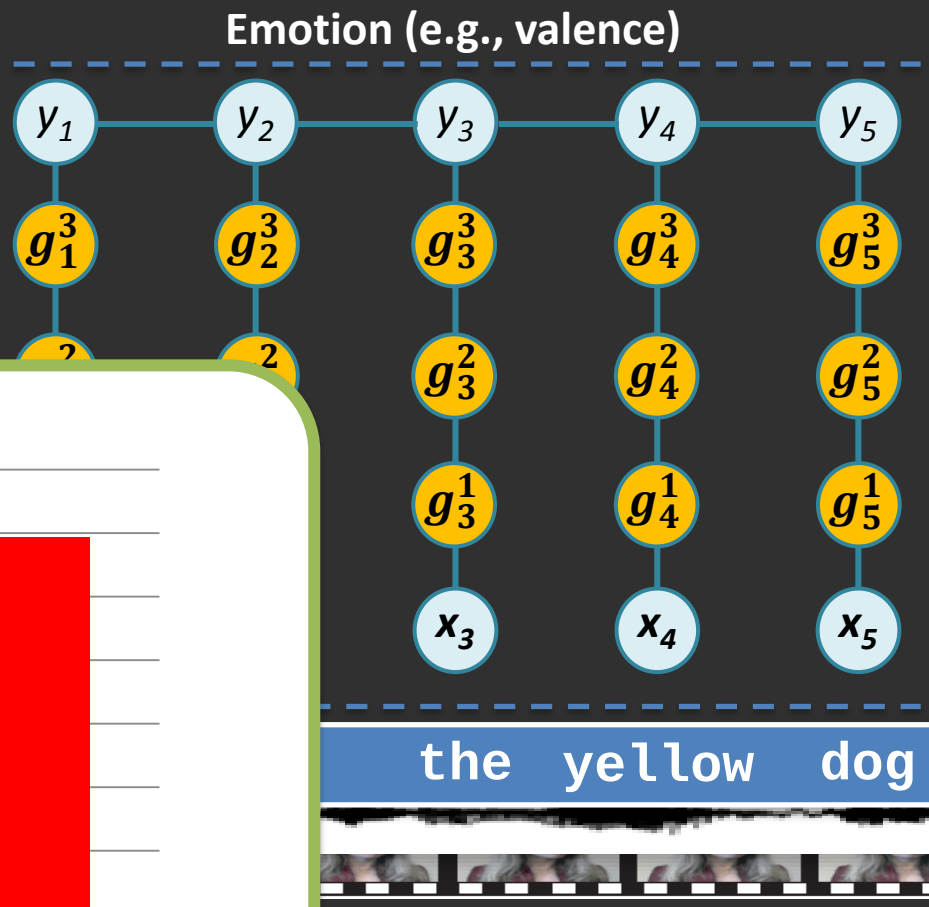
- 15 videos
- 12 co-verbal gestures

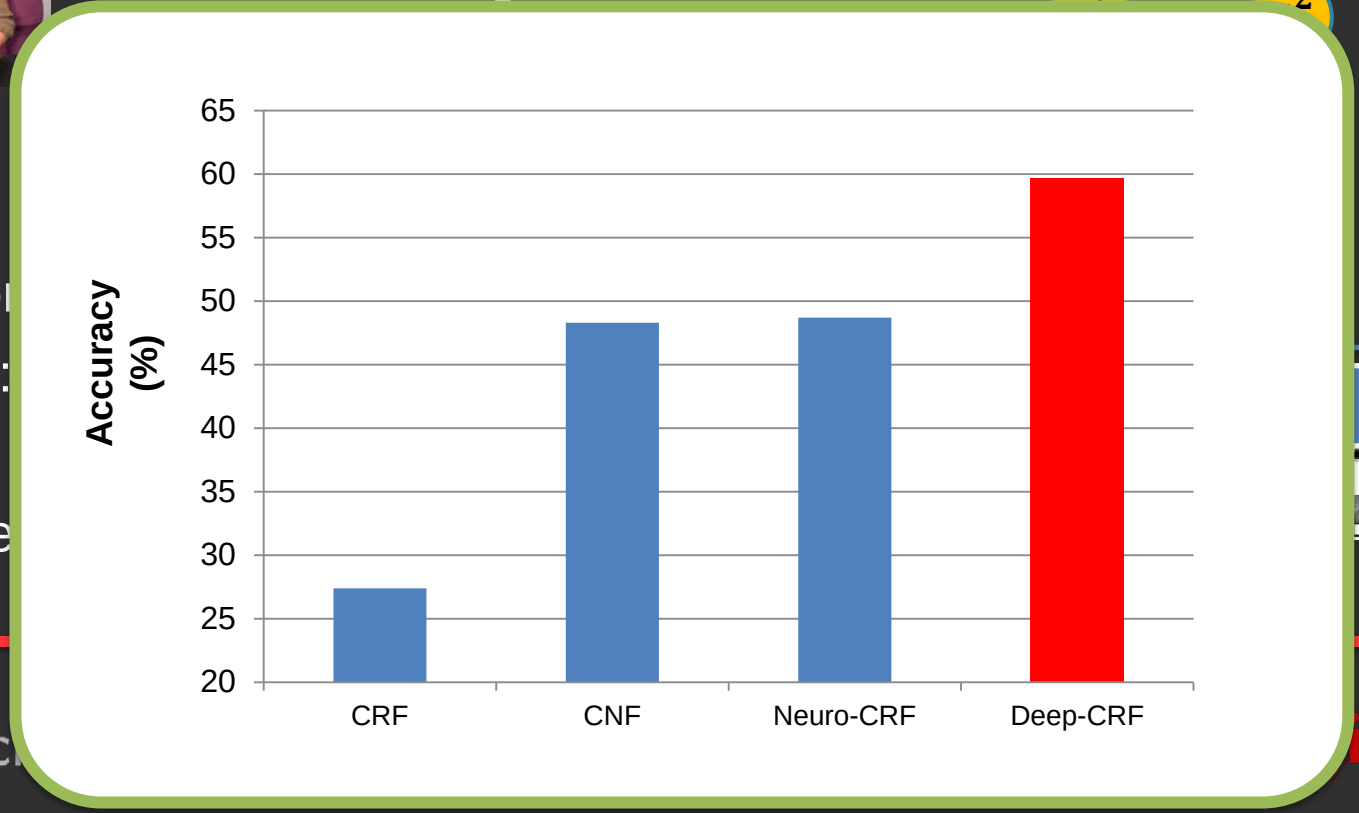
Input features:

- Text: Wo
- Acoustic:

Output labels:

- Visual ge





Joint Learning of Deep Multimodal Representations

Audio-visual speech recognition

[Ngiam et al., ICML 2011]

- Bimodal Deep Belief Network

Image captioning

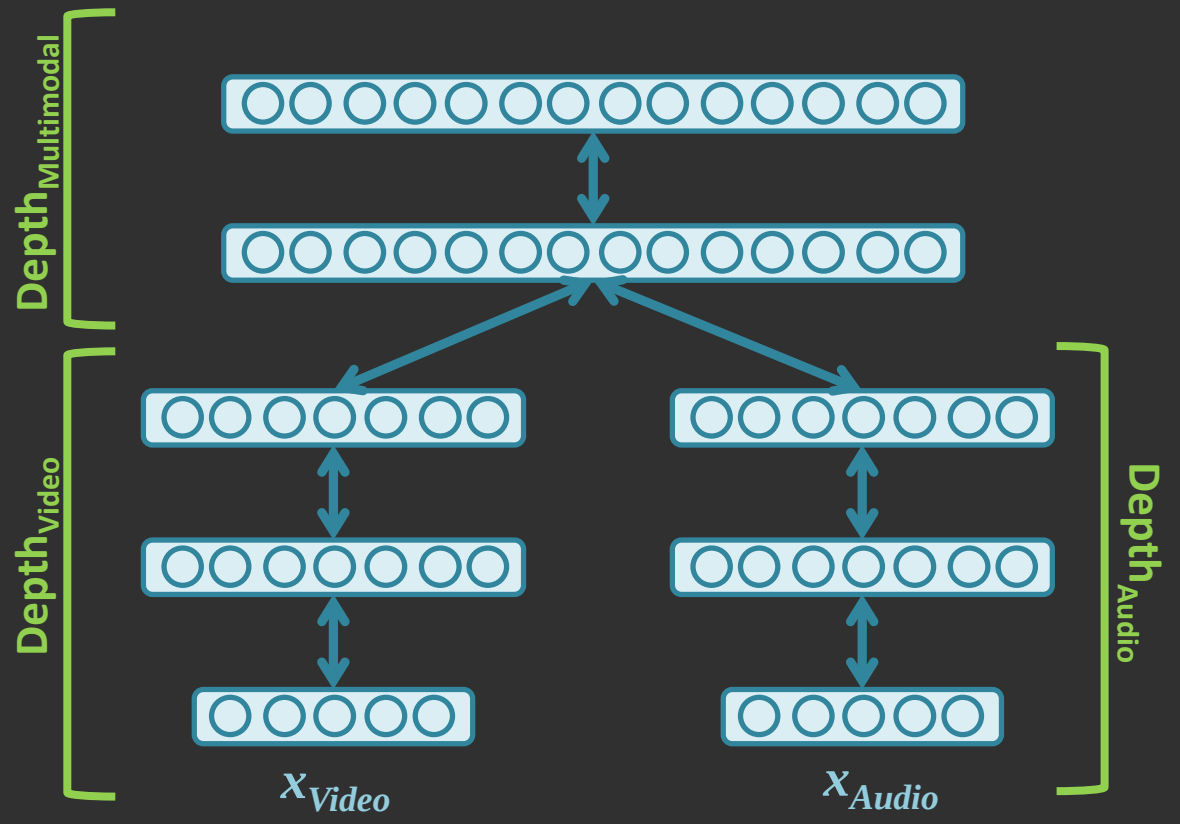
[Srivastava and Salahutdinov, NIPS 2012]

- Multimodal Deep Boltzmann Machine

Audio-visual emotion recognition

[Kim et al., ICASSP 2013]

- Deep Boltzmann Machine



Modeling Multimodal Dynamics

Multimodal

Representation

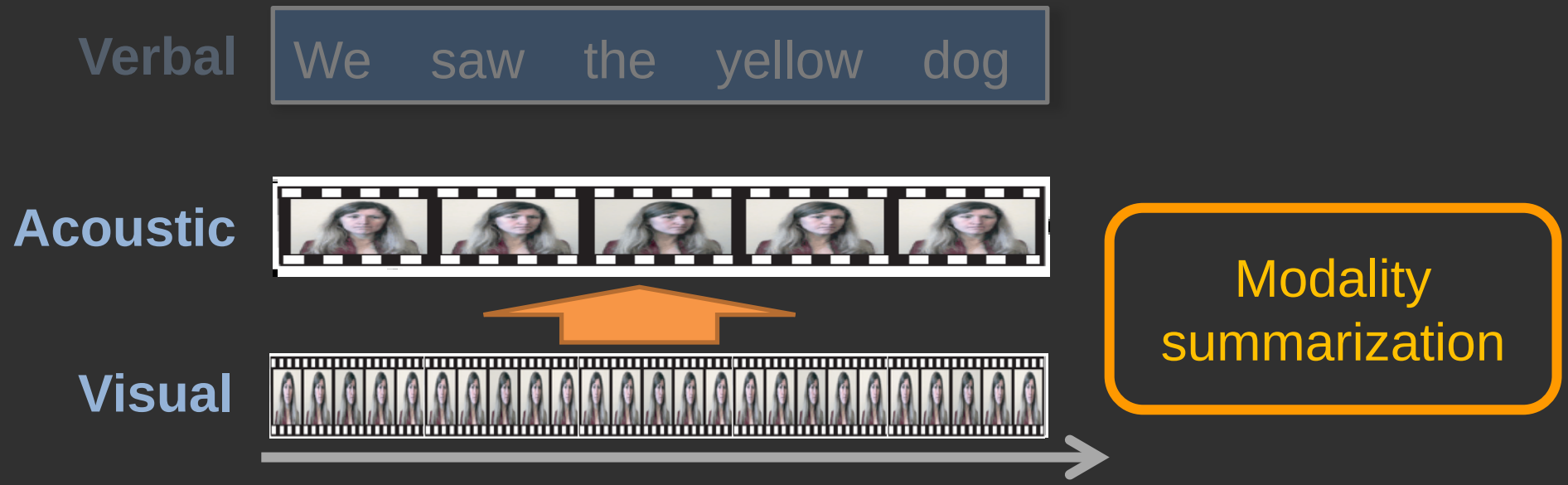
- Nonlinear
- Granularity

Dynamics

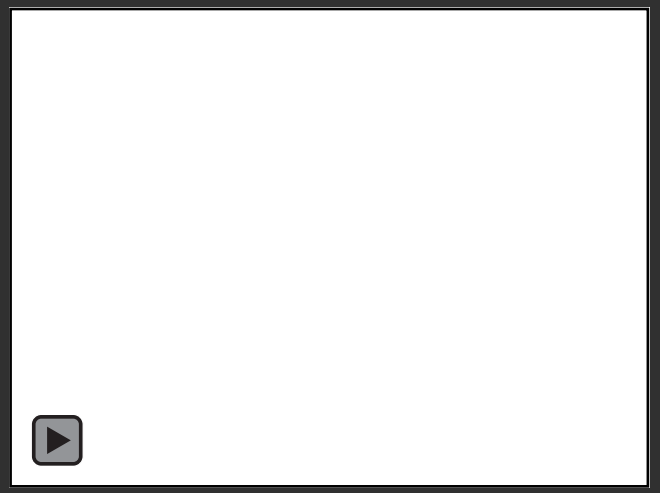
Interpretation

Multimodal representations

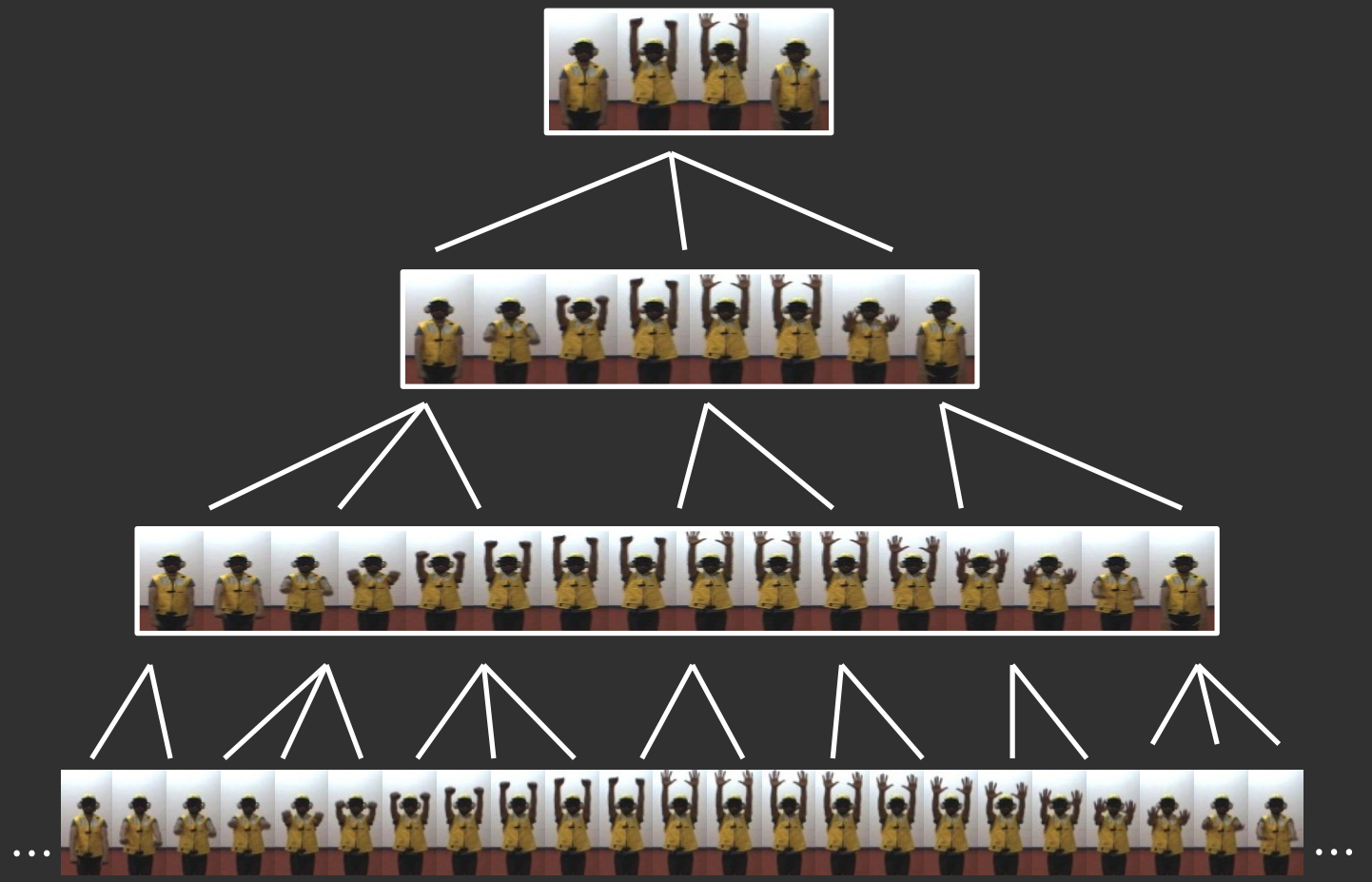
- How to model the nonlinear relationships between multimodal features?
- How to represent different granularities in unimodal or multimodal streams?



Hierarchical Sequence Summarization [CVPR 2013]



1. Arms up
2. Hands *closed → open*
3. Arms down



Hierarchical Sequence Summarization [CVPR 2013]

Sequence Summarization

I. Subsample at Fixed Time Interval

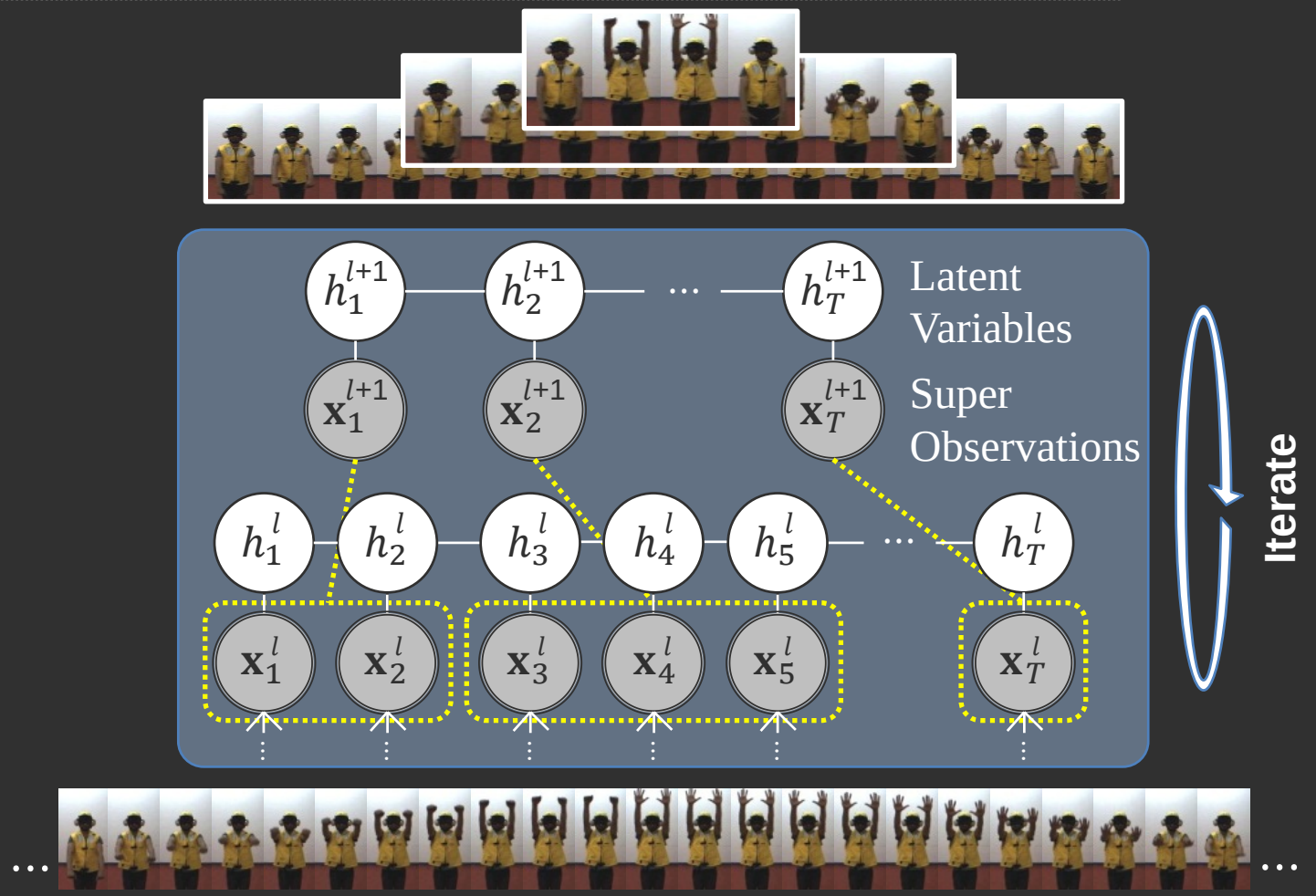
- Sensitive to frame rate
- Can't deal with speed variations

II. Image / Feature Similarity

- Sensitive to noise, scale, and range

III. Latent State Similarity

- ***Doesn't have above limitations***
- ***Nice probabilistic interpretation***



Hierarchical Sequence Summarization [CVPR 2013]

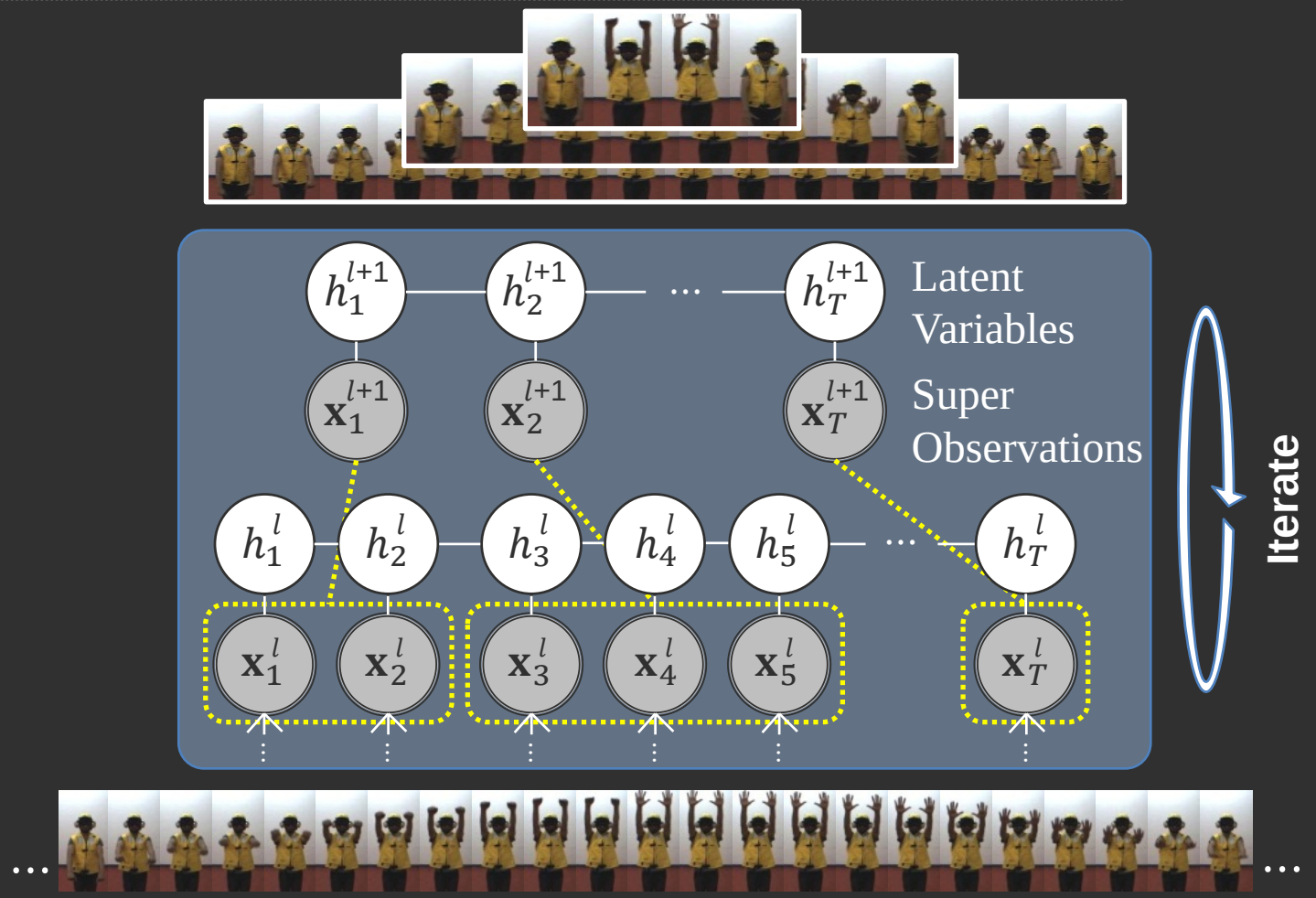
Experiments

NATOPS dataset: Classification of both arm and hand gestures from videos

- **Input features:** Automatically tracked arm and hand postures
- **Output labels:** 6 gesture classes

ArmGesture dataset: Classification of arm gestures from videos

- **Input features:** automatically tracked body postures
- **Output labels:** 6 gesture classes



Hierarchical Sequence Summarization [CVPR 2013]

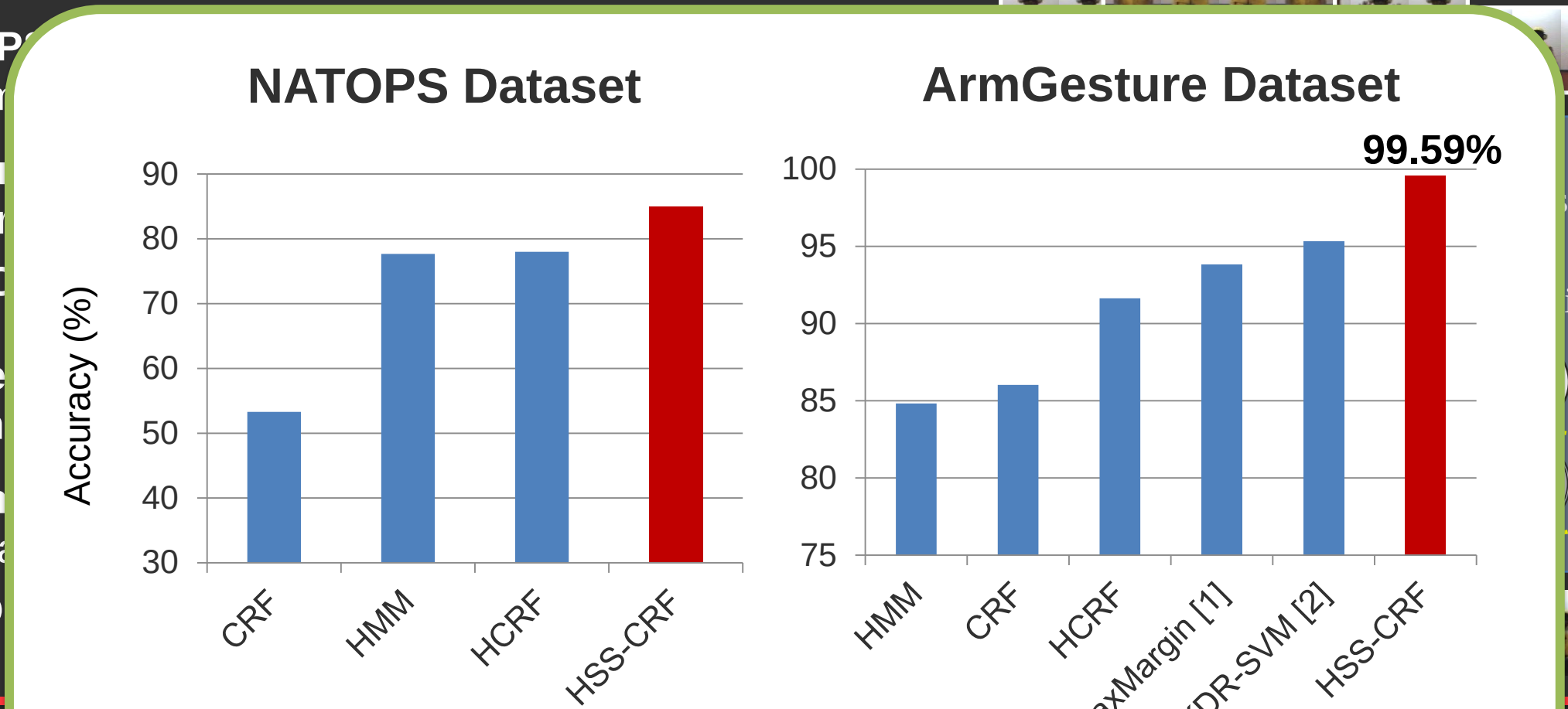
Experiments

NATOP
arm

- **I**
  tr
- **C**

ArmGe
arm

- **In**
  tra
- **O**



Automatic Gesture Summarization [CVPR 2013]

Modeling Multimodal Dynamics

Multimodal

Representation

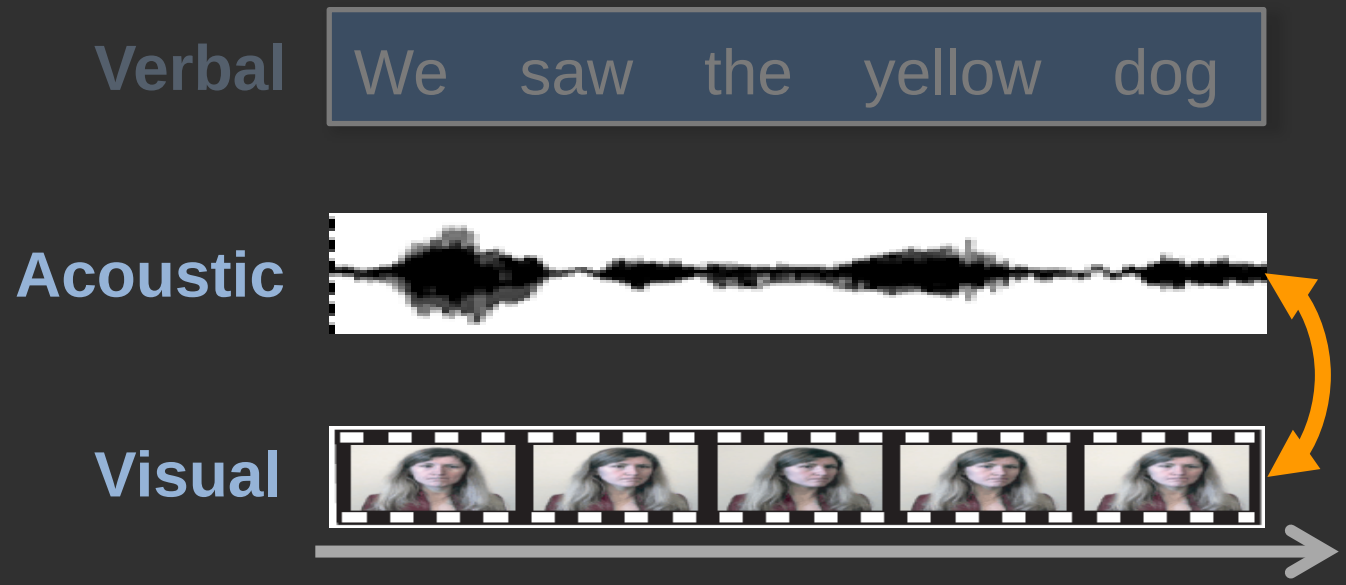
Dynamics

- Complementarity

Interpretation

Multimodal dynamics

- How to improve recognition by taking advantage of multimodal complementarity?



Modeling Multimodal Dynamics

Multimodal

Representation

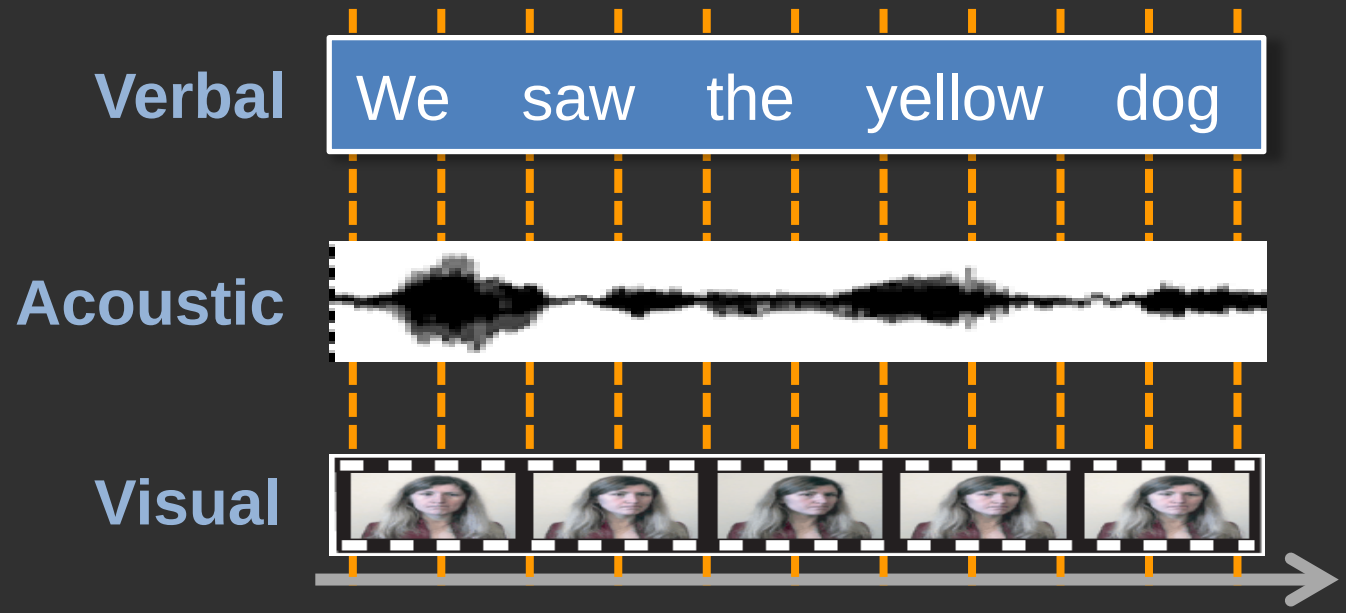
Dynamics

- Complementarity
- Synchrony

Interpretation

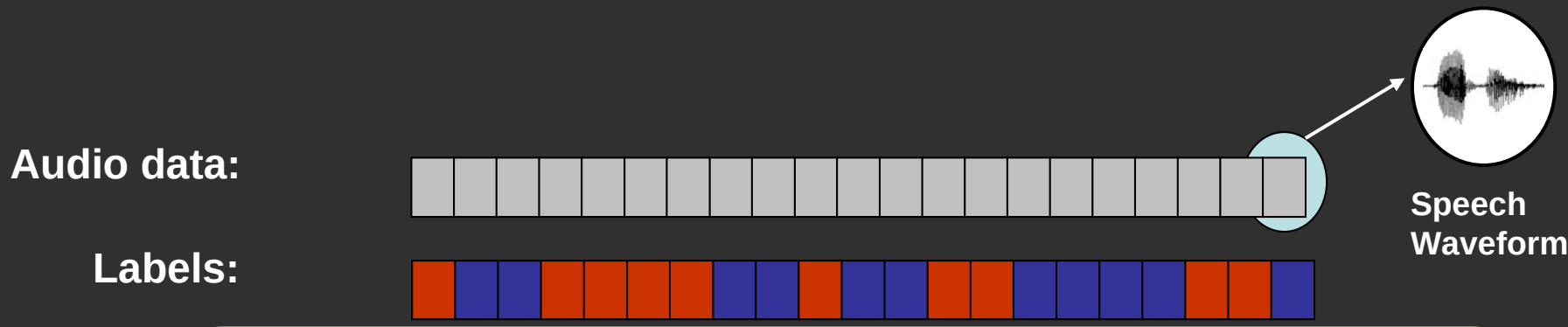
Multimodal dynamics

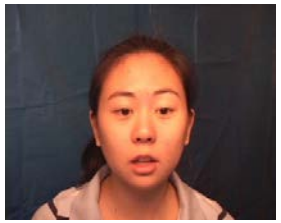
- How to improve recognition by taking advantage of multimodal complementarity?
- How can we model the synchrony between multimodal streams?



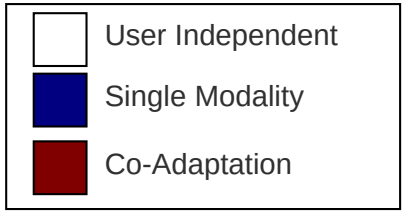
Multimodal Co-Adaptation [ICMI 2006]

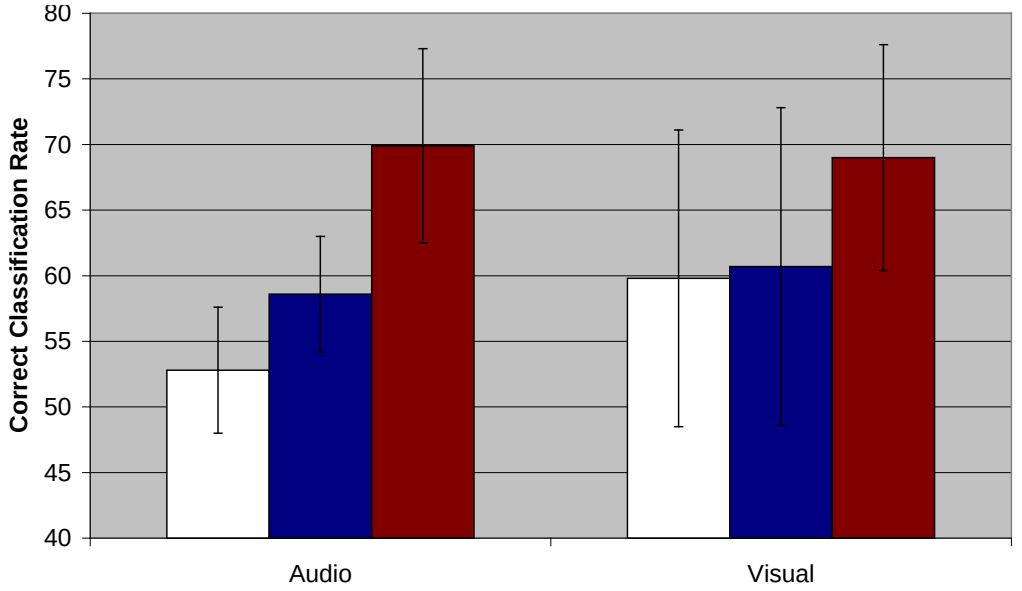
Multimodal Co-Adaptation [ICMI 2006]





Speech Unit Classification





Learning Multimodal Structure

Modality-***private*** structure

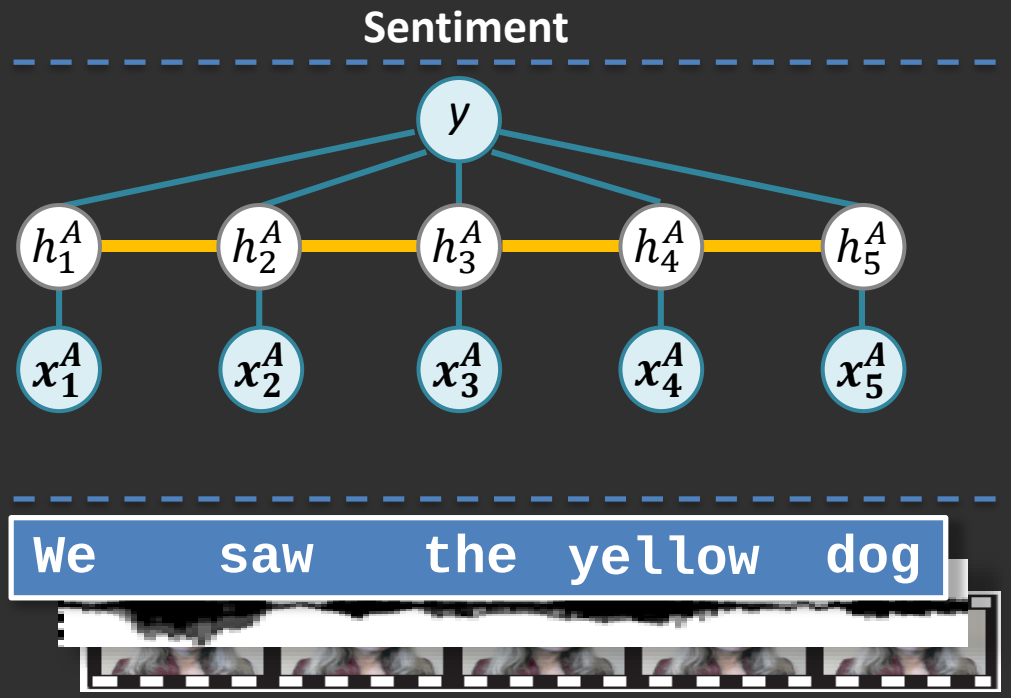
- Internal grouping of observations

Modality-***shared*** structure

- Interaction and synchrony

Early / Late fusion is inappropriate

- Early – strong modality can dominate
- Late – cross-modality dependency is discarded



Learning Multimodal Structure

Modality-***private*** structure

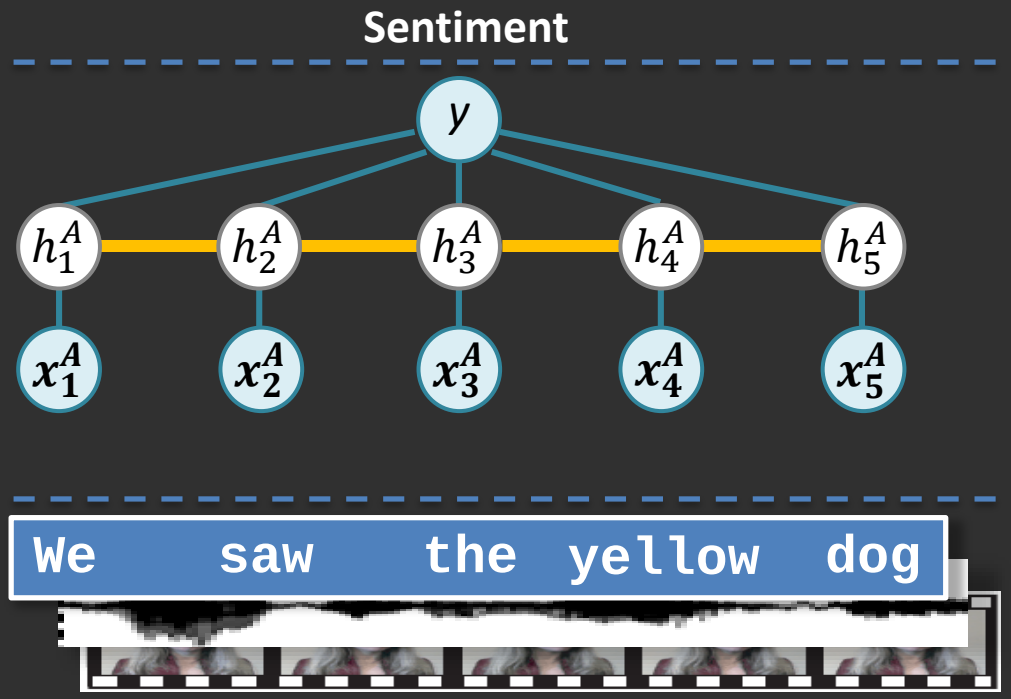
- Internal grouping of observations

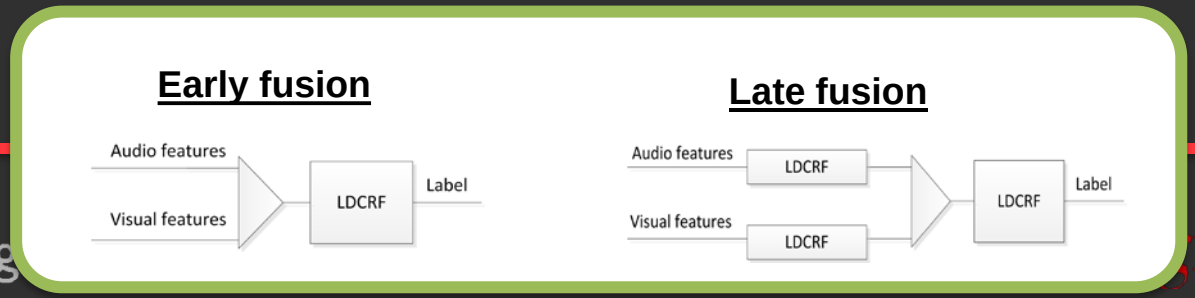
Modality-***shared*** structure

- Interaction and synchrony

Early / Late fusion is inappropriate

- Early – strong modality can dominate
- Late – cross-modality dependency is discarded





Multi-view Latent Variable Discriminative Models [CVPR 2012]

Modality-***private*** structure

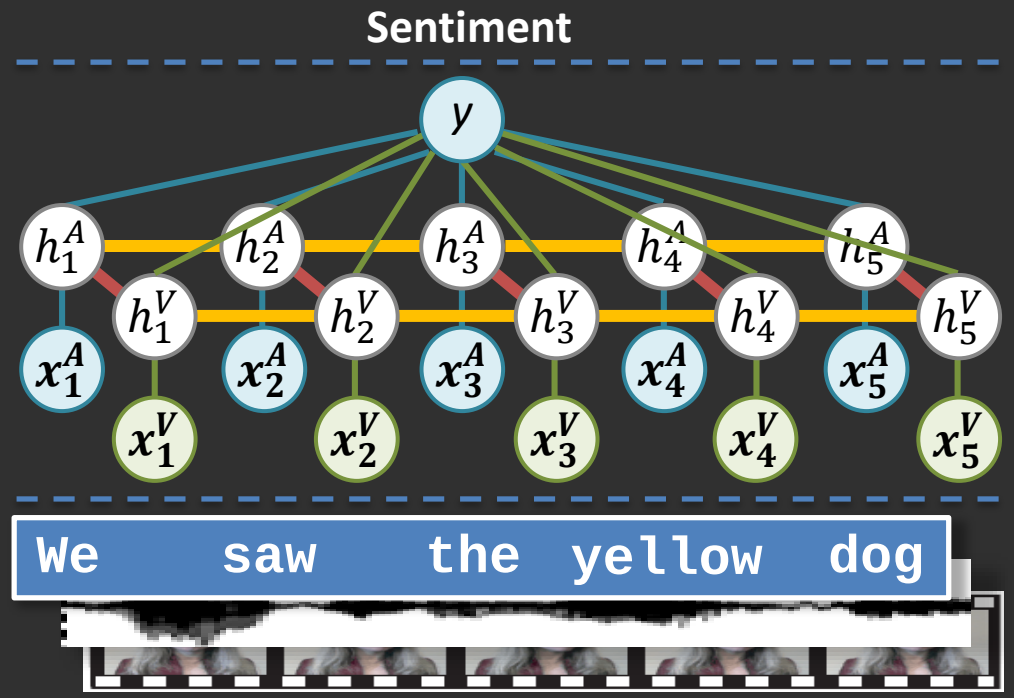
- Internal grouping of observations

Modality-***shared*** structure

- Interaction and synchrony

$$p(y|\, \boldsymbol{x}^{A}, \boldsymbol{x}^{V}; \boldsymbol{\theta}) = \sum_{\boldsymbol{h}^{A}, \boldsymbol{h}^{V}} p\left(y, \boldsymbol{h}^{A}, \boldsymbol{h}^{V} \middle| \boldsymbol{x}^{A}, \boldsymbol{x}^{V}; \boldsymbol{\theta}\right)$$

➢ Approximate inference using loopy-belief



Multi-view Latent Variable Discriminative Models [CVPR 2012]

Modality-***private*** structure

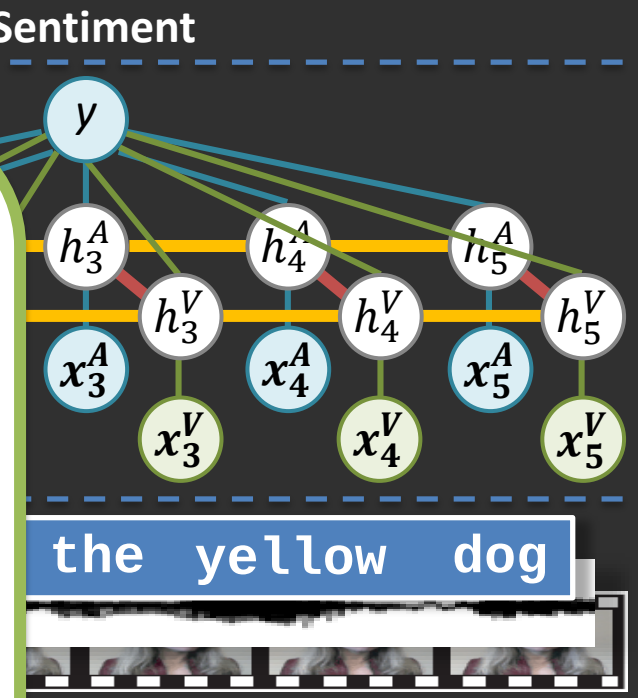
- Internal gro

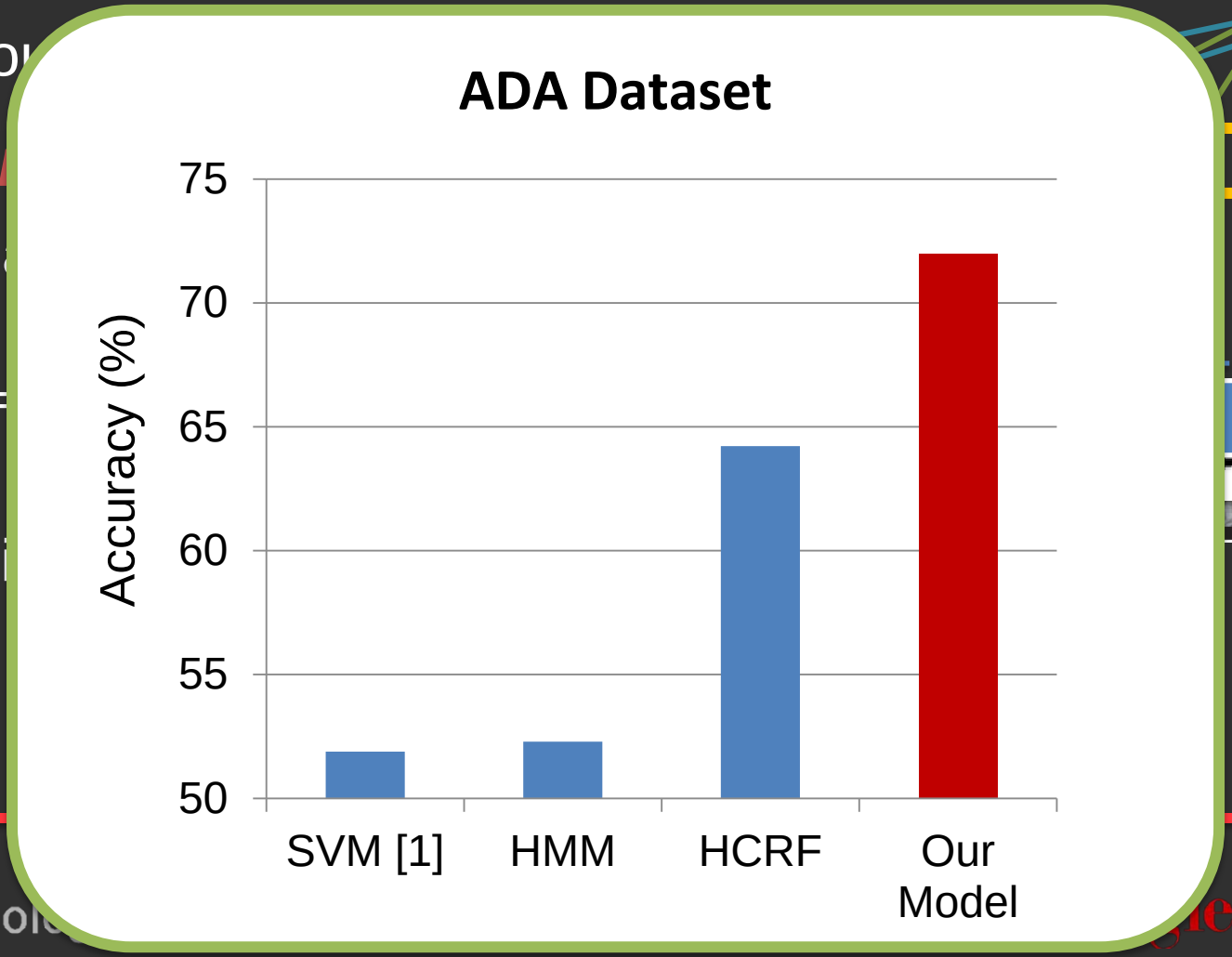
Modality-***shar***

- Interaction

$p(y|\, x^A, x^V; \theta)$ =

➢ Approximate i





Multimodal Behavior Interpretation

Multimodal

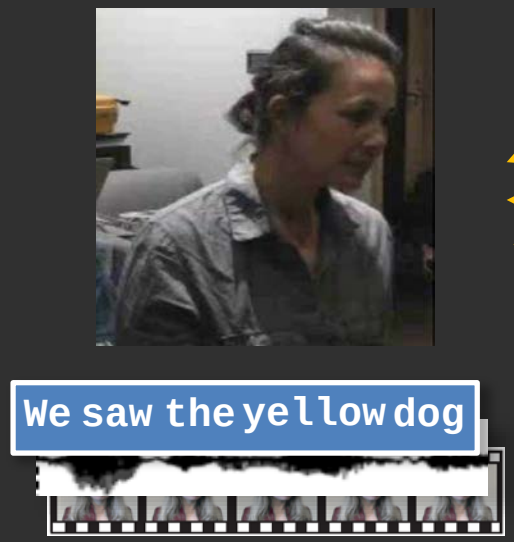
Representation

Dynamics

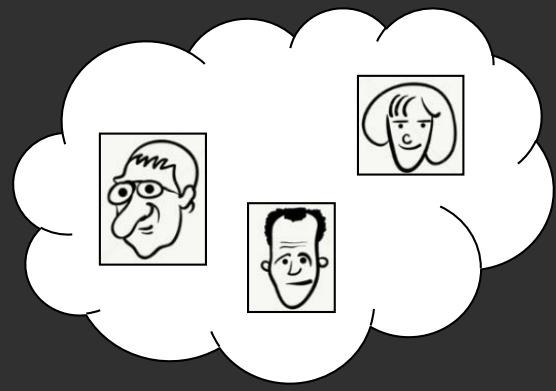
Interpretation

Multimodal Behavior interpretation

- How to model the "fuzziness" in people interpretation of multimodal behaviors?
- How to jointly learn multiple emotions and behaviors (multi-task learning)?



(e.g., perception of emotions)



Observers

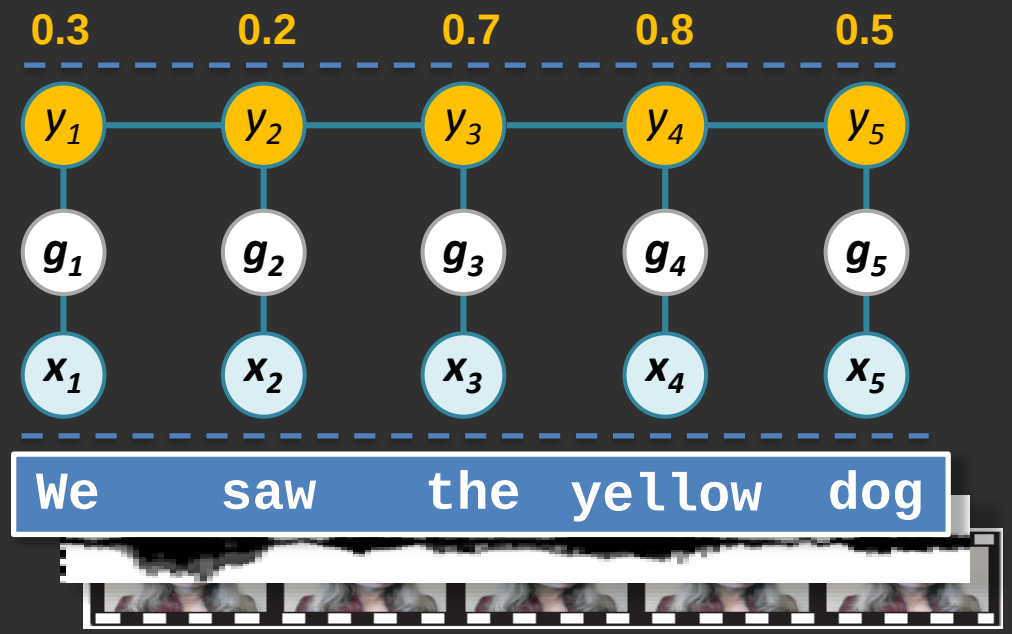
Continuous Conditional Neural Field [ECCV 2014]

Continuous output variables: **(e.g., continuous emotional label)**

$$\boldsymbol{y} = \{y_1, y_2, y_3, \cdots, y_t\} \quad \text{where } y_t \in \mathbb{R}$$

$$p(\boldsymbol{y}|\, \boldsymbol{x}; \boldsymbol{\theta}) = \frac{1}{\mathcal{Z}(\mathbf{x}; \boldsymbol{\theta})} \exp\left\{\sum_t \boldsymbol{\theta} \cdot F(y_t, y_{t-1}, \mathbf{x}_t, \boldsymbol{\theta}^g)\right\}$$

$$\mathcal{Z}(\mathbf{x}; \boldsymbol{\theta}) = \int_{-\infty}^{\infty} \exp\left\{\sum_t \boldsymbol{\theta} \cdot F(y_t, y_{t-1}, \mathbf{x}_t, \boldsymbol{\theta}^g)\right\} d\boldsymbol{y}$$



- How to solve this infinite integral?

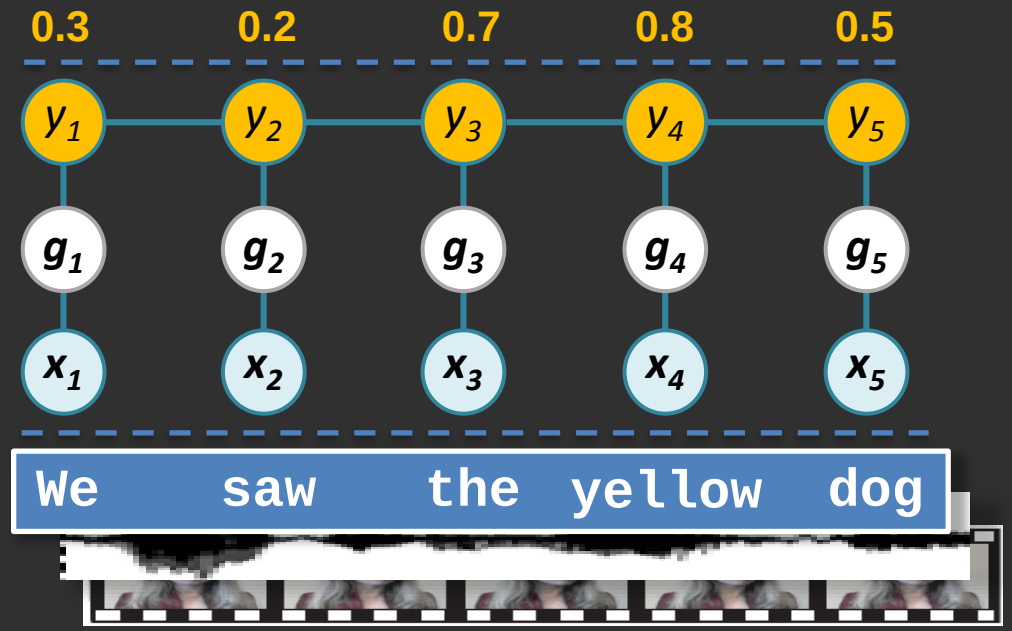
Continuous Conditional Neural Field [ECCV 2014]

Continuous output variables: **(e.g., continuous emotional label)**

$$\boldsymbol{y} = \{y_1, y_2, y_3, \cdots, y_t\} \quad \text{where } y_t \in \mathbb{R}$$

$$p(\boldsymbol{y}|\, \boldsymbol{x}; \boldsymbol{\theta}) = \frac{1}{\mathcal{Z}(\mathbf{x}; \boldsymbol{\theta})} \exp\left\{\sum_t \boldsymbol{\theta} \cdot F(y_t, y_{t-1}, \mathbf{x}_t, \boldsymbol{\theta}^g)\right\}$$

$$\mathcal{Z}(\mathbf{x}; \boldsymbol{\theta}) = \int_{-\infty}^{\infty} \exp\left\{\sum_t \boldsymbol{\theta} \cdot F(y_t, y_{t-1}, \mathbf{x}_t, \boldsymbol{\theta}^g)\right\} d\boldsymbol{y}$$



- How to solve this infinite integral?

Multivariate Gaussian integral:

$$\int_{-\infty}^{\infty} \exp\left\{\frac{1}{2}\boldsymbol{y}^T \Sigma^{-1} \boldsymbol{y} + \boldsymbol{y}\Sigma^{-1}\boldsymbol{\mu}\right\} d\boldsymbol{y} = \frac{(2\pi)^{n/2}}{|\Sigma^{-1}|^{1/2}} \exp\left(\frac{1}{2}\boldsymbol{\mu}\, \Sigma^{-1} \boldsymbol{\mu}\right)$$

[Radosavljevic et al., 2010]

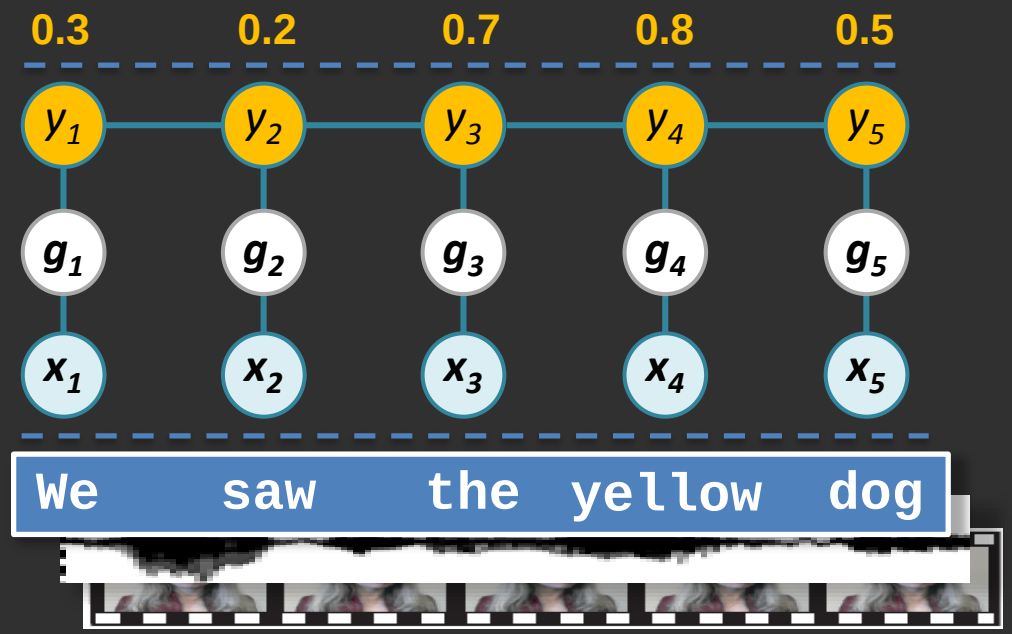
Continuous Conditional Neural Field [ECCV 2014]

Continuous output variables: **(e.g., continuous emotional label)**

$$\boldsymbol{y} = \{y_1, y_2, y_3, \cdots, y_t\} \quad \text{where } y_t \in \mathbb{R}$$

Multivariate Gaussian distribution:

$$p(\boldsymbol{y}|\,\boldsymbol{x};\boldsymbol{\theta}) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right)$$



Since CCNF can be viewed as a multivariate Gaussian, the prediction of y' is simply the mean value of distribution:

$$\boldsymbol{y}' = \arg\max_{\boldsymbol{y}}\left(P(\boldsymbol{y}|\boldsymbol{x})\right) = \boldsymbol{\mu}$$

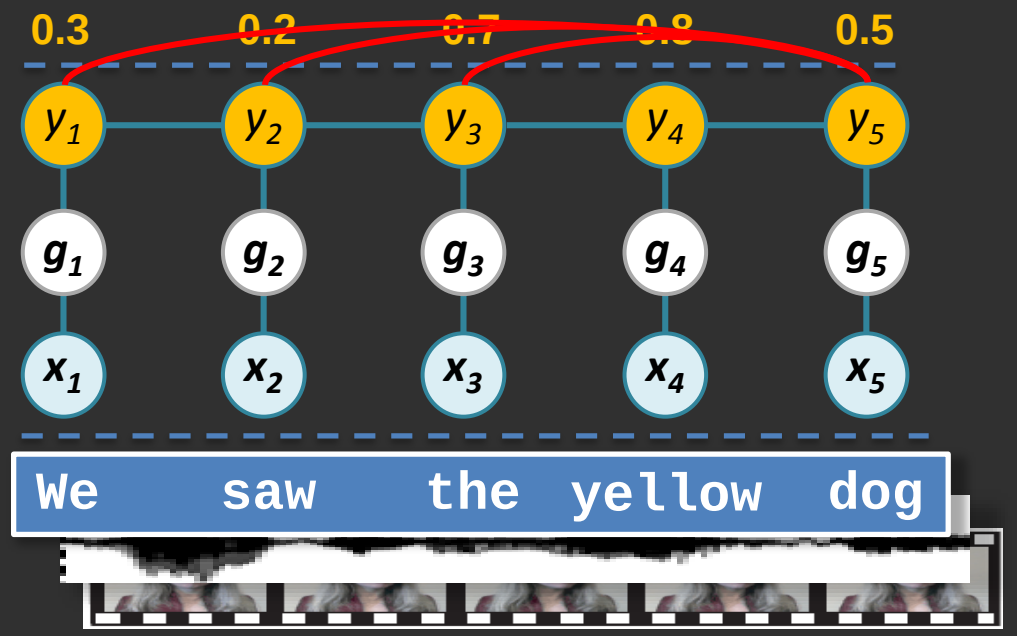
Continuous Conditional Neural Field [ECCV 2014]

Continuous output variables: **(e.g., continuous emotional label)**

$$\boldsymbol{y} = \{y_1, y_2, y_3, \cdots, y_t\} \quad \text{where } y_t \in \mathbb{R}$$

Multivariate Gaussian distribution:

$$p(\boldsymbol{y}|\,\boldsymbol{x};\boldsymbol{\theta}) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right)$$



- *k*-order potential functions:

$$f^{e_k}(y_t, y_{t-k}) = -\frac{1}{2}(y_t - y_{t-k})^2$$

- *Sparsity* potential functions:

$$f^{s_k}(y_t, y_{t-k}) = -\frac{1}{2}(y_t + y_{t-k})^2$$

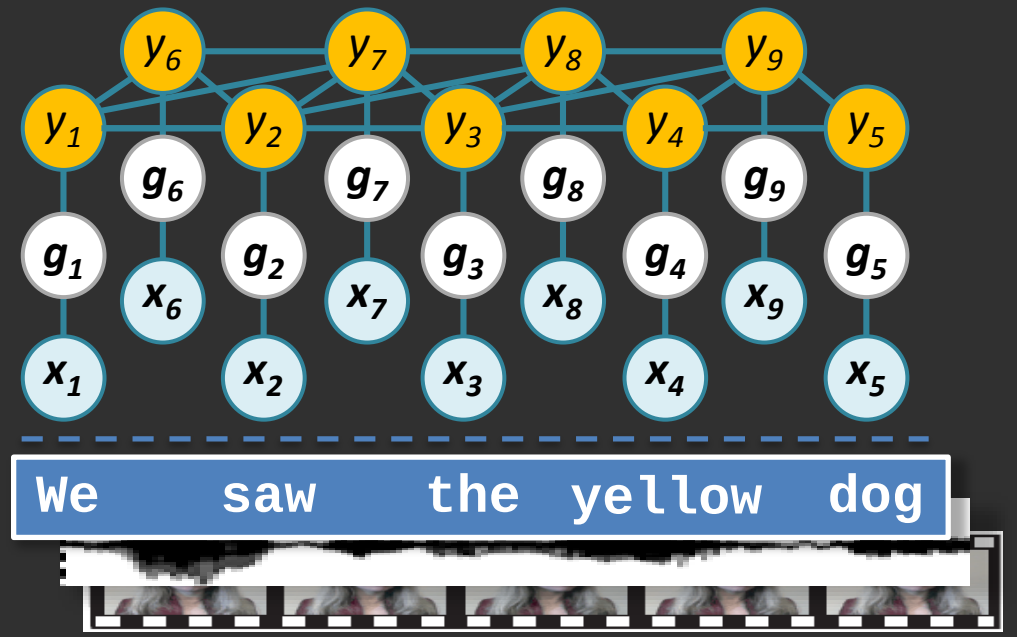
Continuous Conditional Neural Field [ECCV 2014]

Continuous output variables: **(e.g., continuous emotional label)**

$$\boldsymbol{y} = \{y_1, y_2, y_3, \cdots, y_t\} \quad \text{where } y_t \in \mathbb{R}$$

Multivariate Gaussian distribution:

$$p(\boldsymbol{y}|\, \boldsymbol{x}; \boldsymbol{\theta}) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^T \Sigma^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right)$$



- *k*-order potential functions:

$$f^{e_k}(y_t, y_{t-k}) = -\frac{1}{2}(y_t - y_{t-k})^2$$

- *Sparsity* potential functions:

$$f^{s_k}(y_t, y_{t-k}) = -\frac{1}{2}(y_t + y_{t-k})^2$$

- *Grid* potential functions:

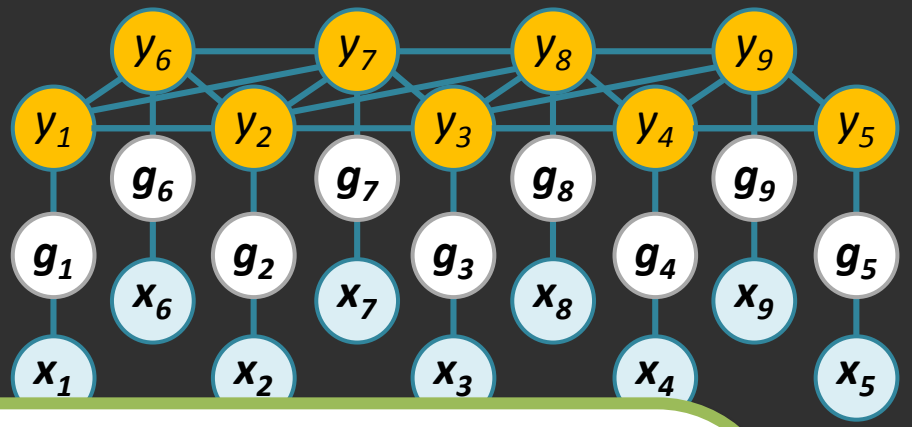
$$f^{2D}(y_i, y_j) = -\frac{1}{2} S_{ij}(y_i - y_j)^2$$

where $S_{i,j}$ specifies which nodes are connected.

Continuous Conditional Neural Field [ECCV 2014]

Continuous output variables: **(e.g., continuous emotional label)**

$$\boldsymbol{y} = \{y_1, y_2, y_3, \cdots, y_t\} \quad \text{where } y_t \in \mathbb{R}$$



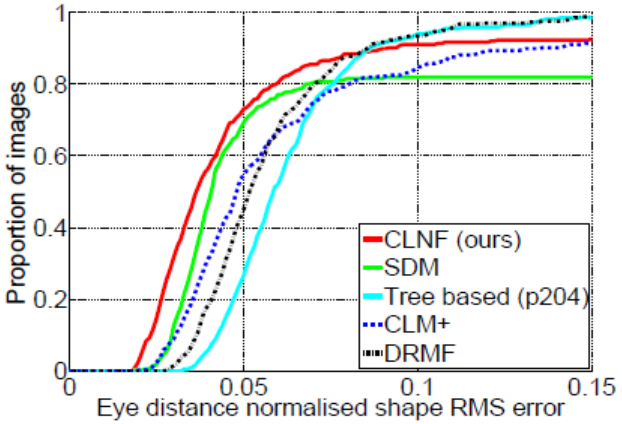
Multivariate Gaussian distribution:

$$p(\boldsymbol{y}|\,\boldsymbol{x};\boldsymbol{\theta}) = \frac{1}{n/\quad 1/} \exp\left(-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right)$$

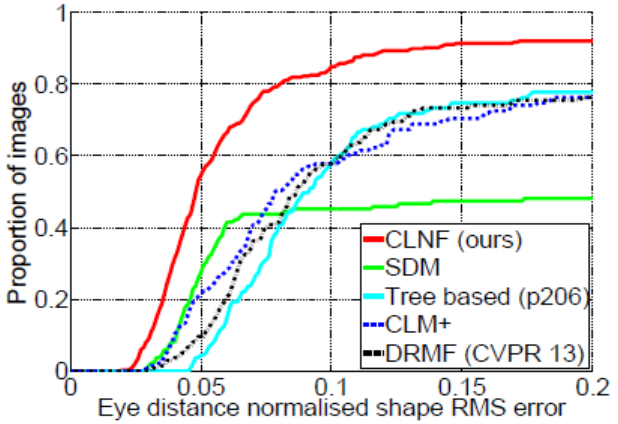
- *k*-o
  - *f*
- Sp
  - *j*

➢ **2D CCNF models can be used for real-time facial landmark detection:**

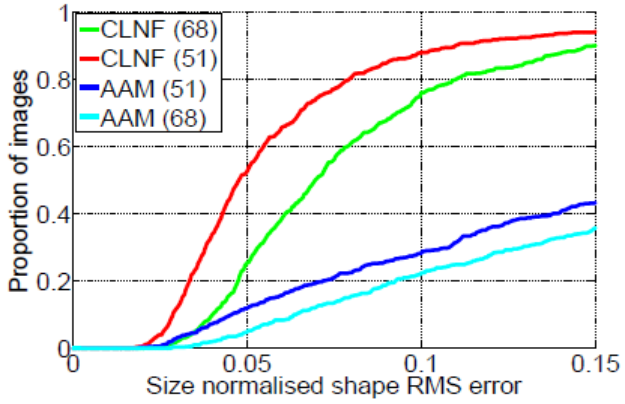
[ICCV workshop 2013, ECCV 2014]



(a) AFW



(b) IBUG



(c) 300W

d.

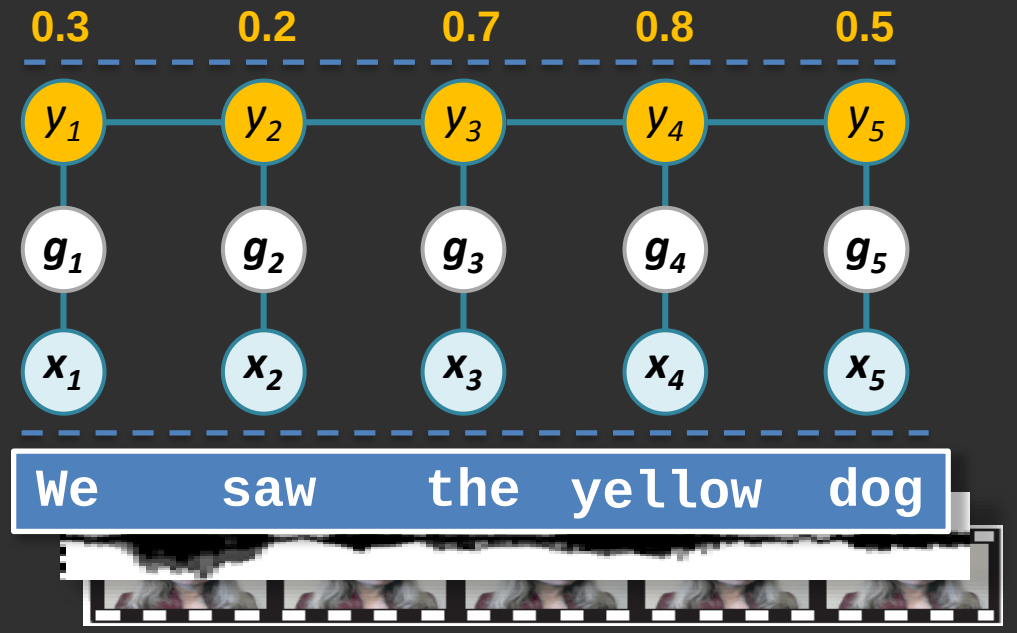
Lang ersity

Continuous Conditional Neural Field [ECCV 2014]

Experiments

DISFA dataset: Intensity of spontaneous facial expressions

- **Input features:** visual facial feature descriptors
- **Output labels:** 12 facial action units with 6-point ordinal scale



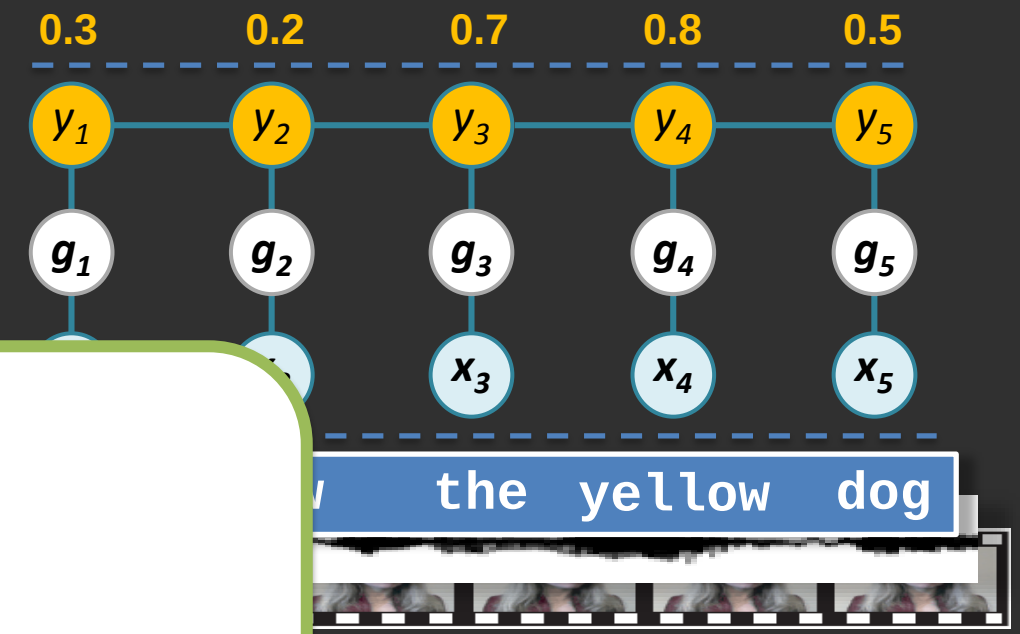
[1] Qin et al. 2009

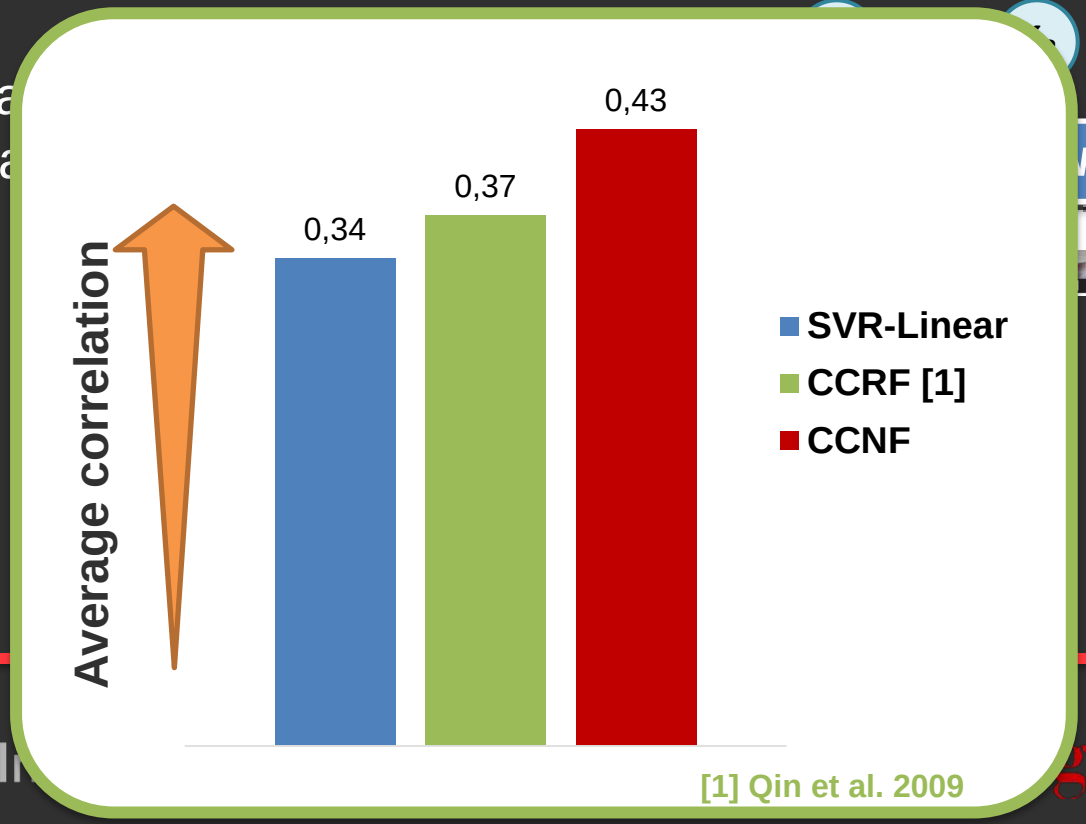
Continuous Conditional Neural Field [ECCV 2014]

Experiments

DISFA dataset: Intensity of spontaneous facial expressions

- **Input features:** visual facial feature descriptors
- **Output labels:** 12 facia
  units with 6-point ordina





Continuous Conditional Neural Field [ECCV 2014]

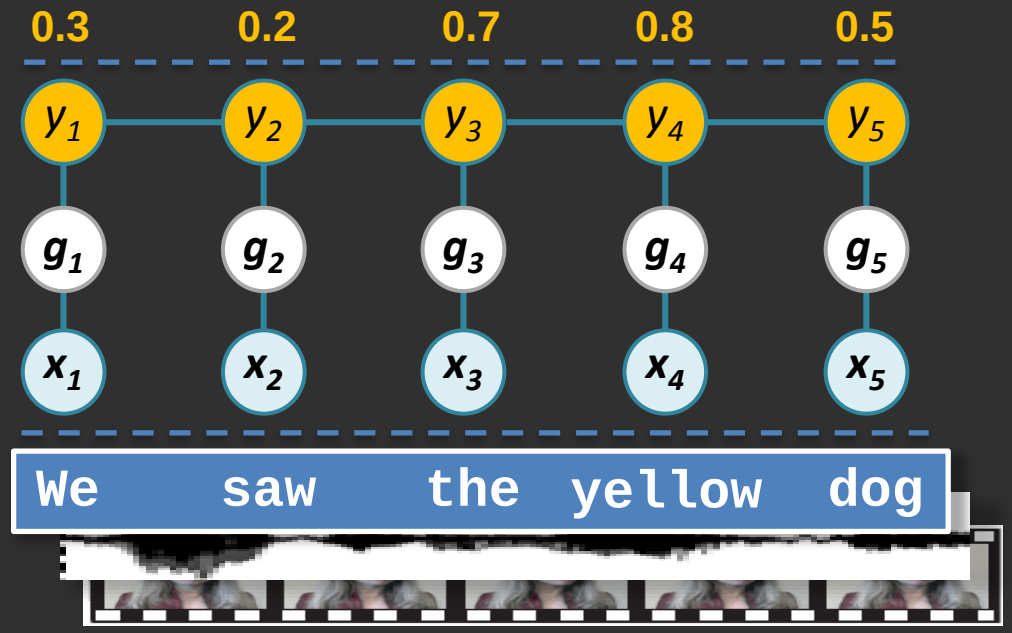
Experiments

DISFA dataset: Intensity of spontaneous facial expressions

- **Input features:** visual facial feature descriptors
- **Output labels:** 12 facial action units with 6-point ordinal scale

MTurk dataset: Emotional perception of music extracts

- **Input features:** Standard acoustic features extracted from music
- **Output labels:** 6 emotional dimensions (arousal and valence



[1] Qin et al. 2009

Continuous Conditional Neural Field [ECCV 2014]

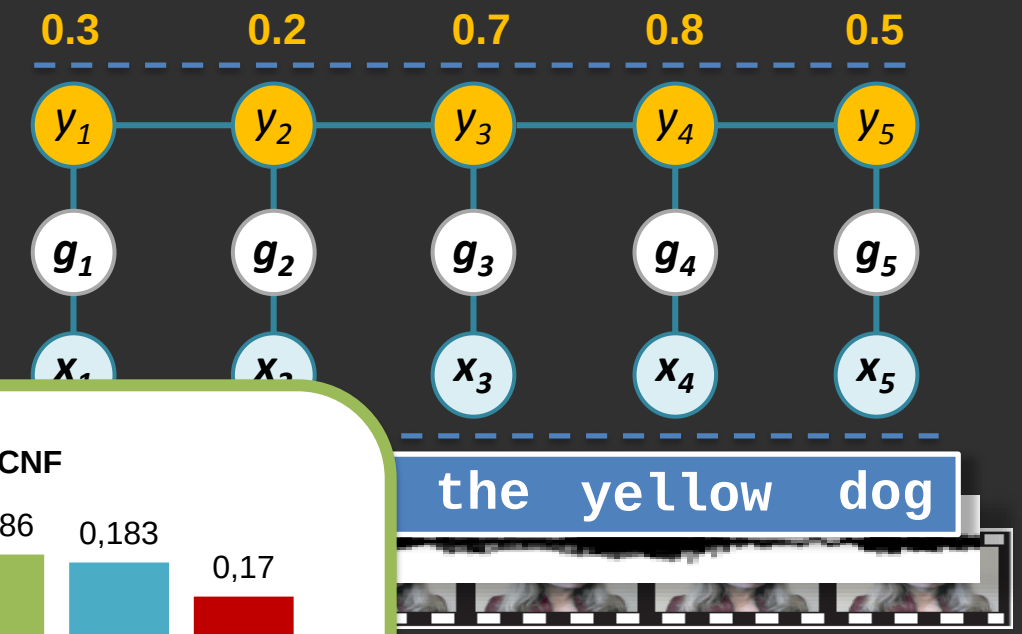
Experiments

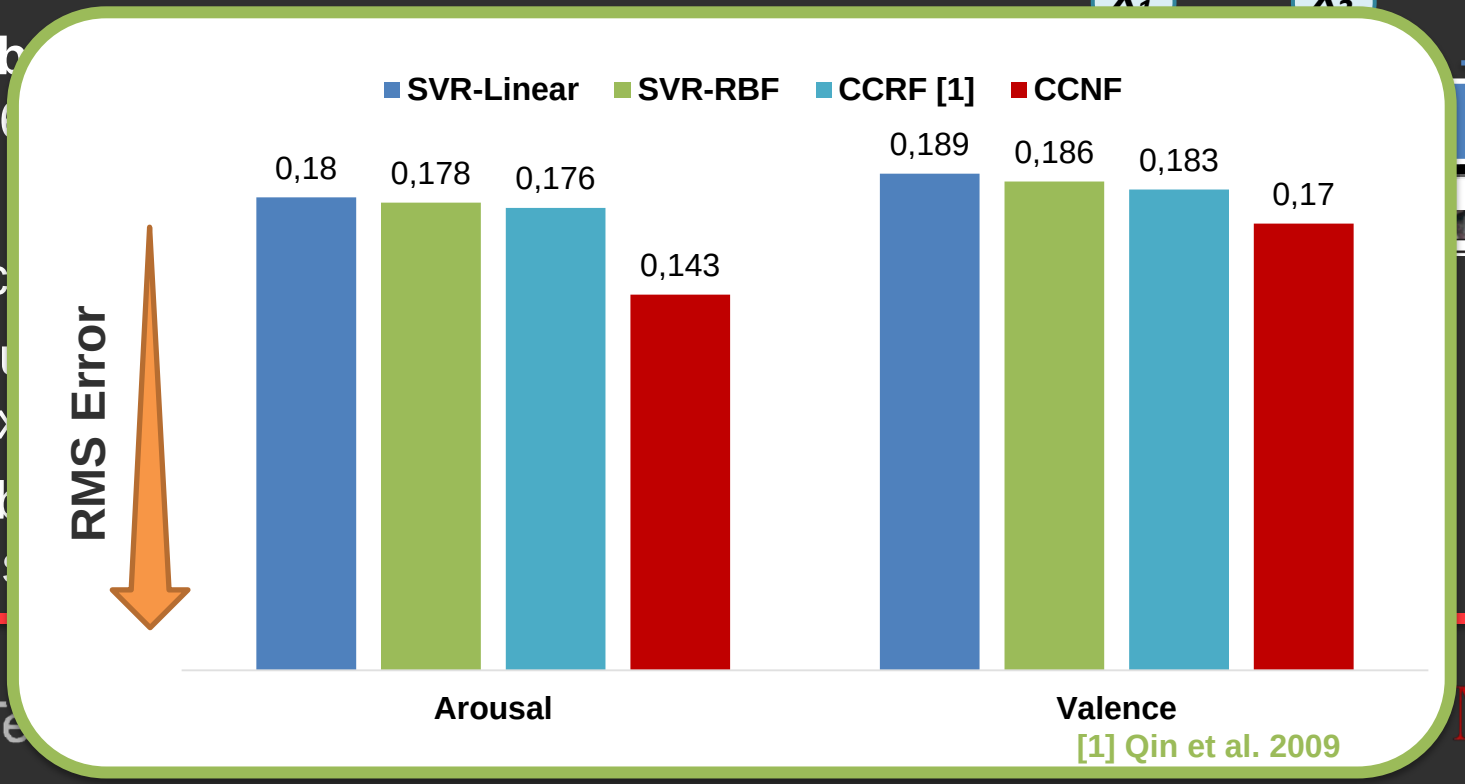
DISFA dataset: Intensity of spontaneous facial expressions

- **Input features:** visual facial feature descriptors
- **Output lab** units with

MTurk dataset: music extrac

- **Input feat** features ex
- **Output lab** dimension





Multimodal Machine Learning

1. **Latent temporal structure**
   - Latent-Dynamic Conditional Random Field (LDCRF) [COLING 2008, CVPR 2007]
   - Hidden Conditional Random Field (HCRF) [PAMI 2007, CVPR 2006]
   - Infinite Hidden Conditional Random Field (iHCRF) [ECML 2013, IEEE TNN 2012]
2. **Multimodal nonlinear representation**
   - (Deep) Latent-Dynamic Conditional Neural Field (LDCNF) [FG 2013, IVA 2015]
   - Hierarchical Sequence Summarization (HSS-CRF) [CVPR 2013]
3. **Multimodal synchrony and complentarity**
   - Multimodal co-training and co-adaptation (Co-Adapt) [ICMI 2006]
   - Multi-view Latent Variable Discriminative Models (MV-CRF) [CVPR 2012]
4. **Behavior prediction and interpretation**
   - Latent Mixture of Discriminative Experts (LMDE) [COLING 2010, ACL 2011]
   - Continuous Conditional Neural Field (CCNF) [ICCV-W 2013, ECCV 2014]
   - Multi-task Continuous Conditional Neural Field (MT-CCNF) [NIPS 2015 – in prep.]

Multimodal Machine Learning

1. **Latent temporal structure**
   - Latent-Dynamic Conditional Random Field (LDCRF) [COLING 2008, CVPR 2007]
   - Hidden Conditional Random Field (HCRF) [PAMI 2007, CVPR 2006]
   - Infinite Hidden Conditional Random Field (iHCRF) [ECML 2013, IEEE TNN 2012]
2. **Multimodal nonlinear representation**
   - (Deep) Latent-Dynamic Conditional Neural Field (LDCNF) [FG 2013, IVA 2015]
   - Hierarchical Sequence Summarization (HSS-CRF) [CVPR 2013]
3. **Multimodal synchrony and complentarity**
   - Multimodal co-training and co-adaptation (Co-Adapt) [ICMI 2006]
   - Multi-view Latent Variable Discriminative Models (MV-CRF) [CVPR 2012]
4. **Behavior prediction and interpretation**

HCRF Library

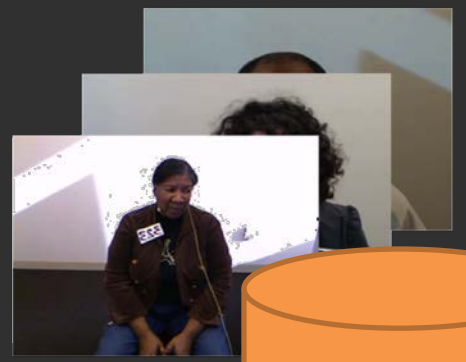
- **Open-source library for multimodal machine learning**

  http://hcrf.sf.net/

Health Behavior Informatics

Behavioral Indicators of Psychological Distress [AAMAS 2014]

Distress Assessment Interview Corpus



DAIC

Modeling Human Communication Dynamics

Human Communication Dynamics

Behavioral

Multimodal

- Audio
- Visual
- Verbal

Interpersonal

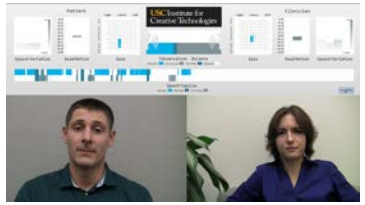
Suicide prevention
with Cincinnati Hospital

SimSensei

Cicero

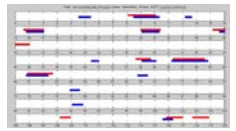
Negotiation outcomes

TeleCoach



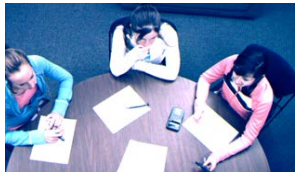
Virtual rapport

OVAT



Social influence

Look-to-talk



Group learning analytics
with Stanford and UCSD

Mel

Future Challenges for Multimodal Machine Learning

- Tri-modal deep representations
  - Integrating language, speech and vision
- Modeling long-term temporal contingency
  - Temporal models with selective memory
- Integrate interpersonal dynamics
  - Model the dyadic or small group interactions
- Interpretation based on social context
  - Model the cultural and societal influence

MERCI !

Learning from the Web

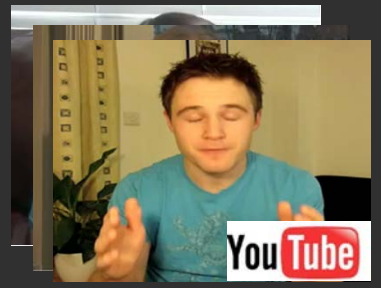
Multimodal

- Audio
- Visual
- Verbal

YouTube

- Multiplicity of expressions
- 10,000+ new videos per days
- Verbal, vocal and visual modalities
- Spontaneous and natural behaviors
- Limited motion range (fixed camera)

Multimodal Sentiment Analysis [ACL 2013, IEEE Intelligent Systems 2012]



?

- Positive
- Neutral
- Negative

47

Audio
- Pauses
- Pitch

Visual
- Smile
- Gaze

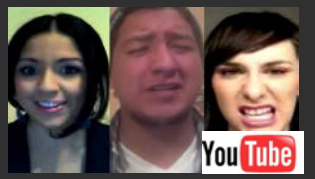
Verbal
- Polarized words

Extracted features

	Positive	Neutral	Negative
Smile intensity	+	−	−
Look at camera	+	−	−
Pause length	−	+	−
Voice pitch level	+	−	+
Polarized words	+		−

Utterance-level classification

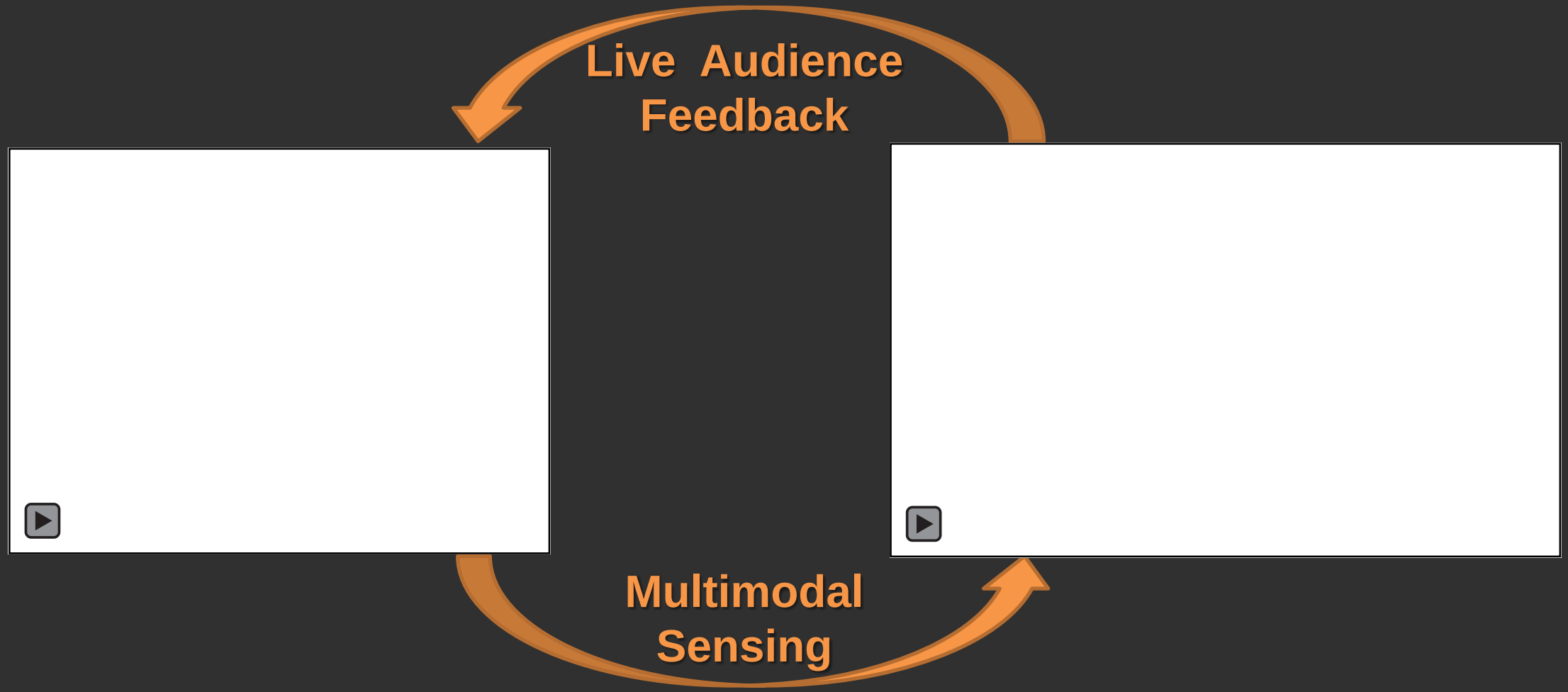
Modality	Accuracy
Baseline	55.93%
One modality at a time	
Linguistic	70.94%
Acoustic	64.85%
Visual	67.31%
Two modalities at a time	
Linguistic + Acoustic	72.88%
Linguistic + Visual	72.39%
Acoustic + Visual	68.86%
Three modalities at a time	
Linguistic+Acoustic+Visual	74.09%



Spanish videos

Modality	Accuracy
Text only	64.94%
Visual only	61.04%
Audio only	46.75%
Text-visual	73.68%
Text-audio	68.42%
Audio-visual	66.23%
Text-audio-visual	**75.00%**

Cicero: Multimodal Virtual Audience Platform for Public Speaking Training [IVA 2013]



Cicero: Multimodal Virtual Audience Platform for Public Speaking Training [IVA 2013]

Virtual audie

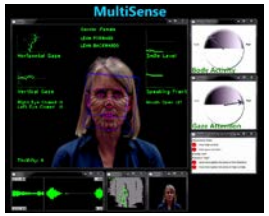
Correlations between expert assessed behavior and automatically computed behavior descriptors:

Source	Assessed behavior	Behavior descriptor	Spearman's ρ	p-value
Voice	Flow of speech	Num. pauses	-.469	.09
	Clear intonation	Avg. intensity	**.805**	.002
		Breathiness	**-.615**	.033
	Interrupted speech	Num. pause fillers	**.612**	.034
	Speaks too quietly	Avg. intensity	**-.842**	$< .001$
	Vocal variety	Std. f_0	**.709**	.010
		Spectral Stationarity	**-.586**	.045
Body	Paces too much	Leg movement	**.682**	.021
	Gestures to emphasize	Arm movement	**.710**	.014
	Gestures to much	Arm movement	.437	.179

Experts

Automatic assessment (SVR):

Correlation with experts: 0.66 ($p<0.05$)



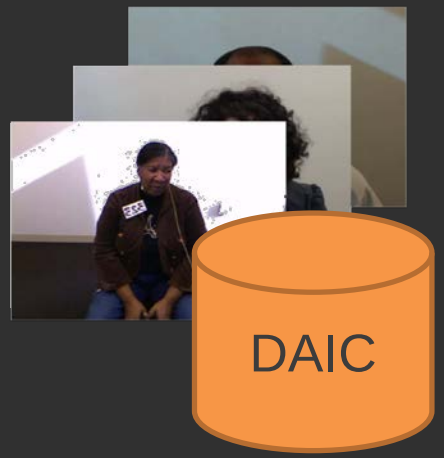
MultiSense

Language

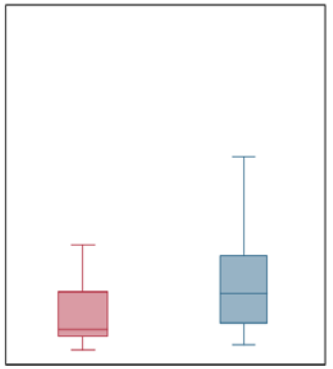
niversity

Psychological Distress Indicators [IEEE FG 2013 - best paper award]

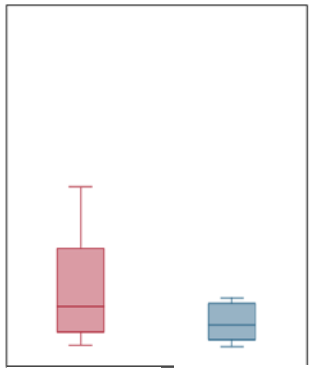
Behavioral



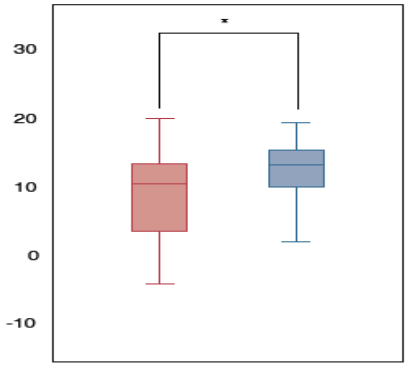
Joy – Facial expr.



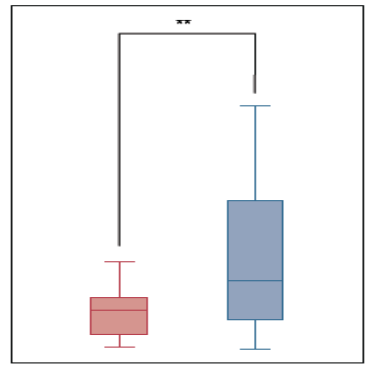
Sad – Facial expr.



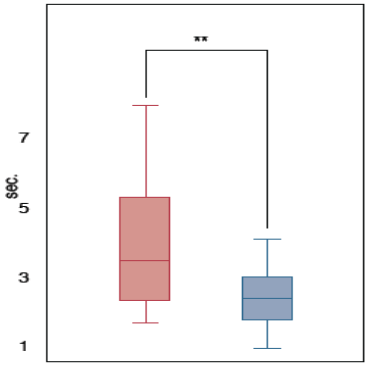
Vertical eye gaze



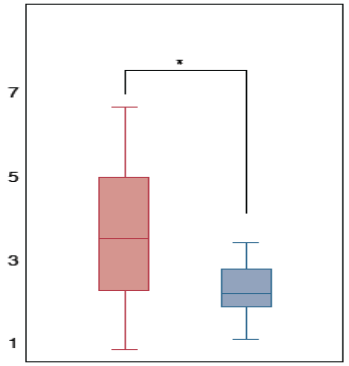
Smile intensity



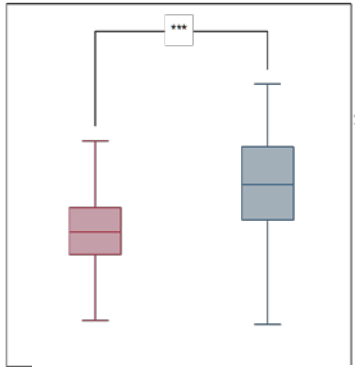
Hand self-adaptor



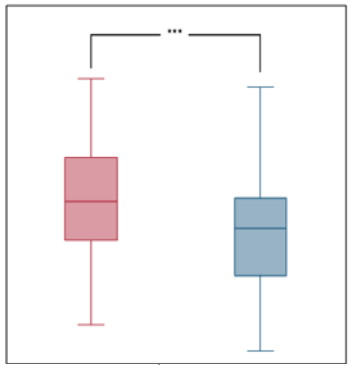
Legs fidgeting



Voice energy std.

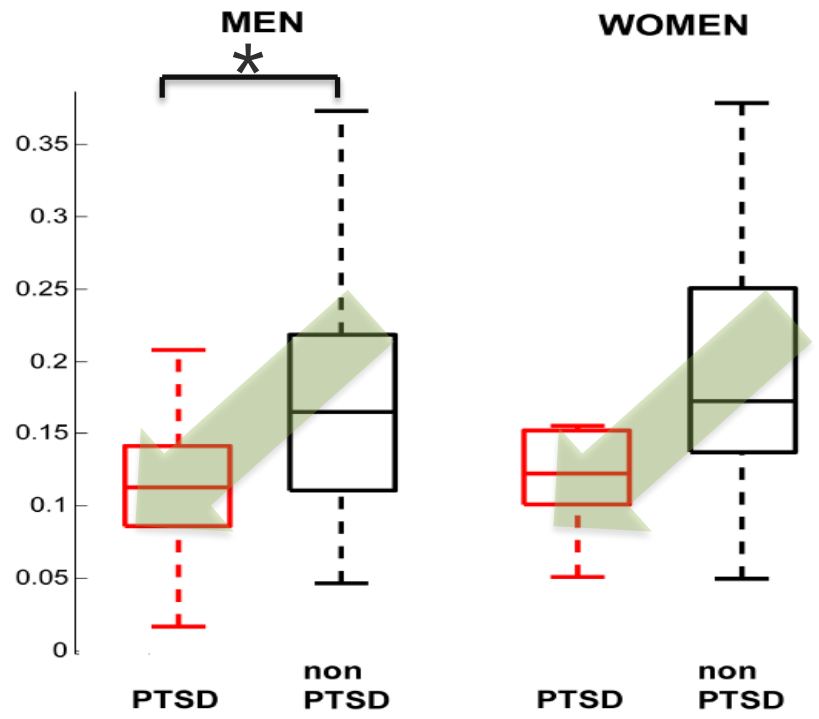


Voice quality (NAQ)



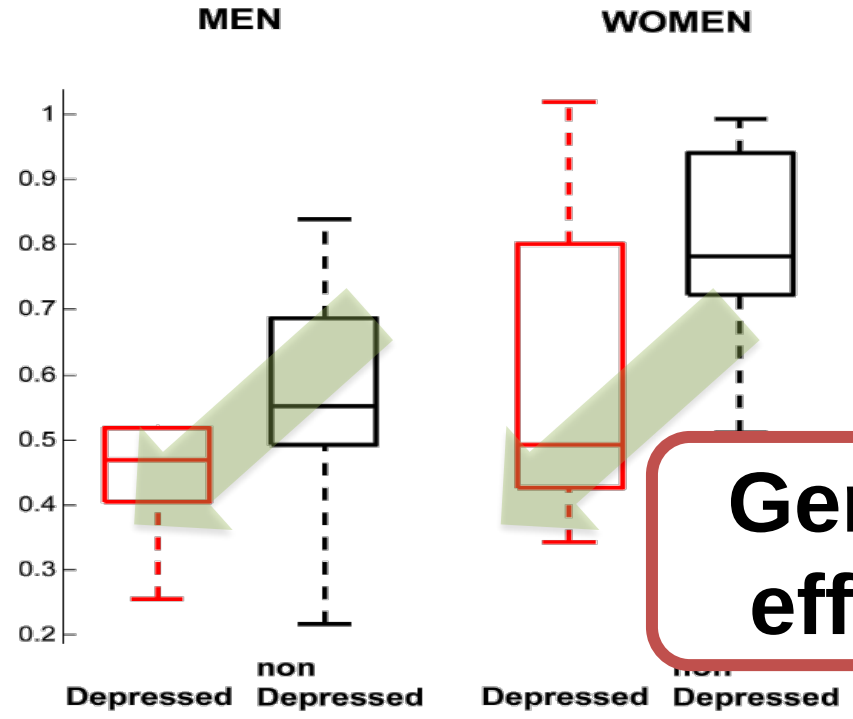
Indicators with similar trend on both genders [ACII 2013]

Head Rotation Variation



G = -0.7426 G = -0.586

Emotional Variation

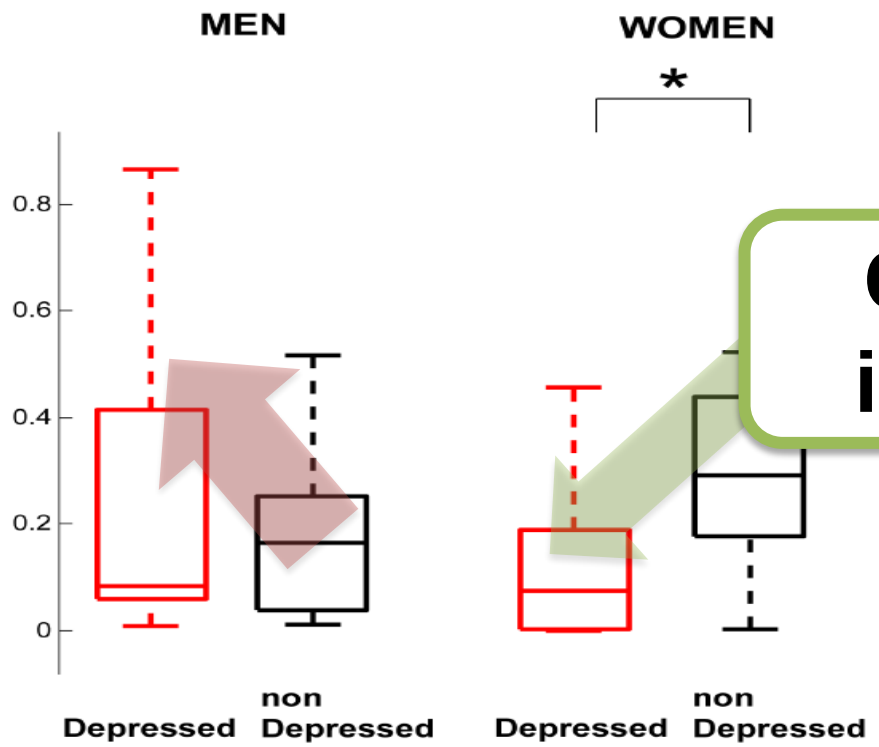


G = -0.5664 G = -0.6852

Gender effect!

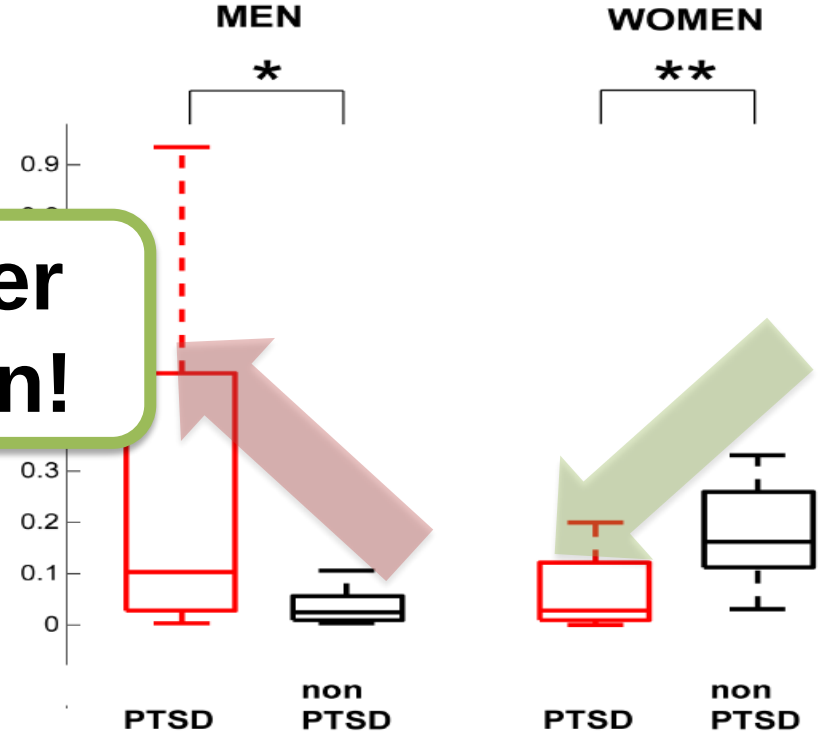
Indicators with different trends on both genders [ACII 2013]

AU4 (frown) Intensity



G = 0.5057 G = -0.9156

Disgust Intensity



G = 0.8407 G = -1.2158

Cross over interaction!