SRI International

Modeling Natural Human Behaviors and Interactions

Presented by Behjat Siddiquie (behjat.siddiquie@sri.com)

Team: Saad Khan, Amir Tamrakar, Mohamed Amer, Sam Shipman, David Salter, Jeff Lubin, Ajay Divakaran and Harpreet Sawhney

SRI International

Holistic Assessment of Behavior – Multimodal Sensing

Voice:
Calm

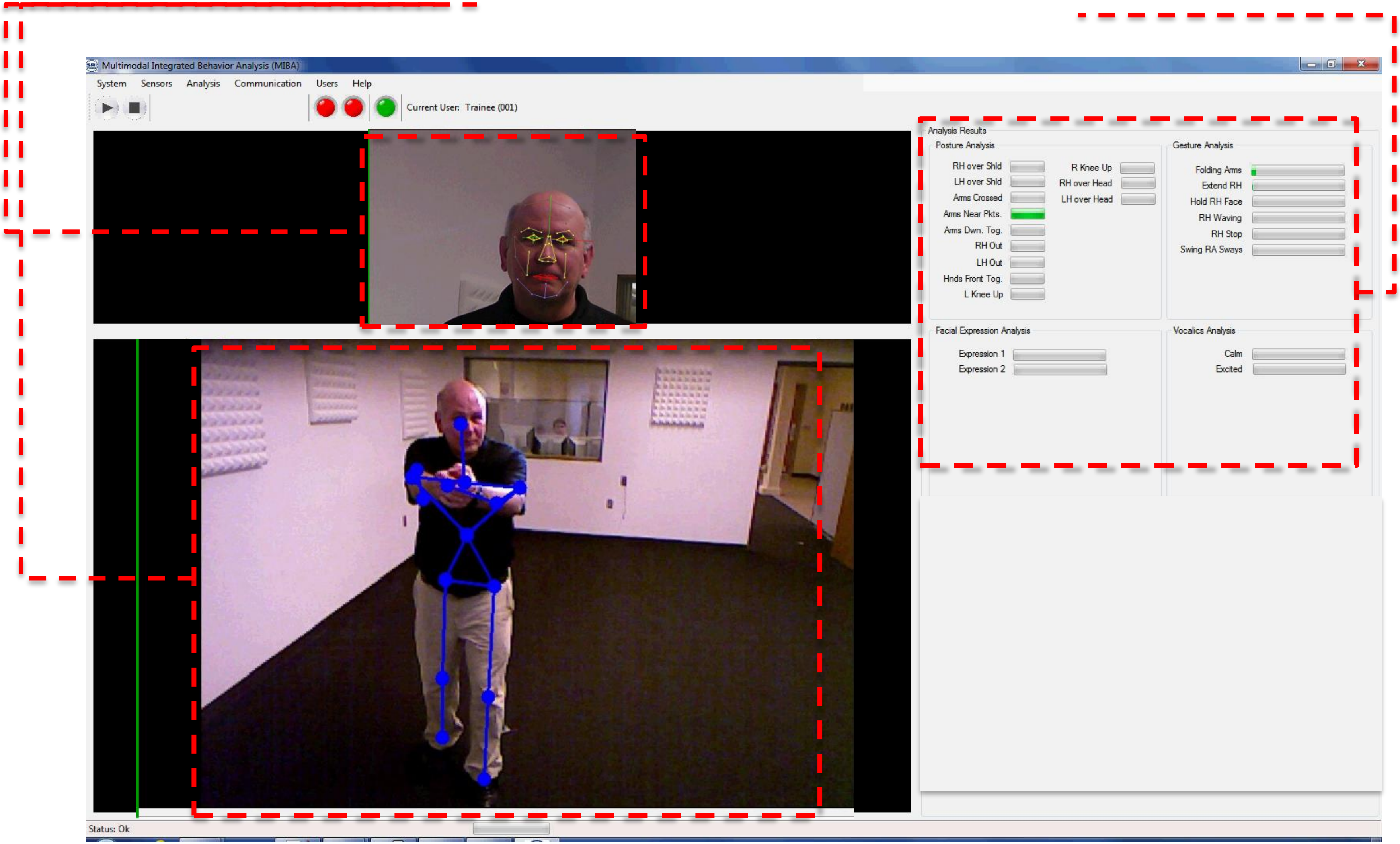
Facial Gesture:
Smiling

Body Posture:
Relaxed

Overall State:
Calm

- Need to combine multiple cues to arrive at holistic assessment of user state
  - Body Pose, Gestures, Facial Expressions, Speech Tone, Keywords-> NLU
- Provides contextual effects to produce predictions of behavior at Gottman's "construct" level of behavioral classification.

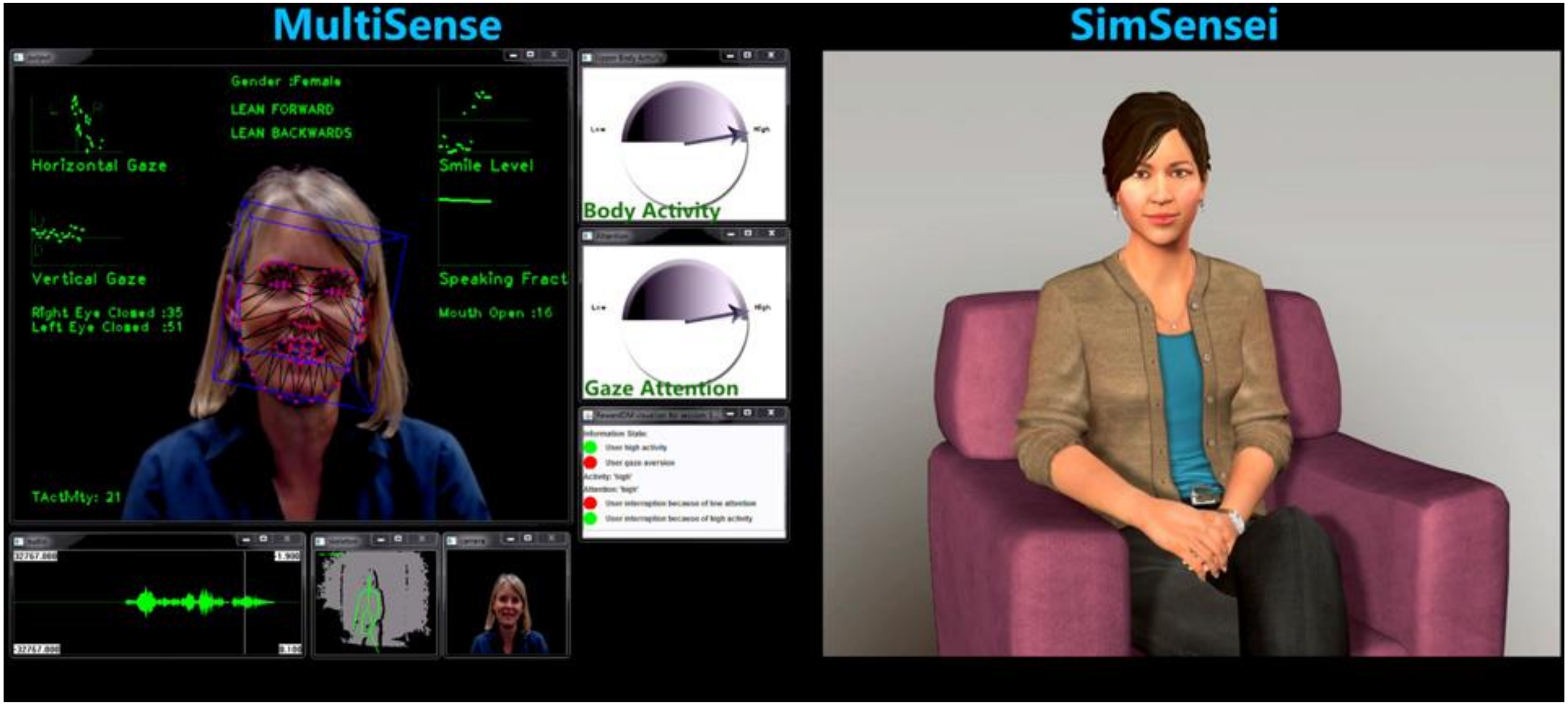
Training that Blends Tactical and Soft Skills?



Law Enforcement – Domestic Violence Scene – Training Video

Assist Doctors and Therapists?

- Virtual doctor's visits…



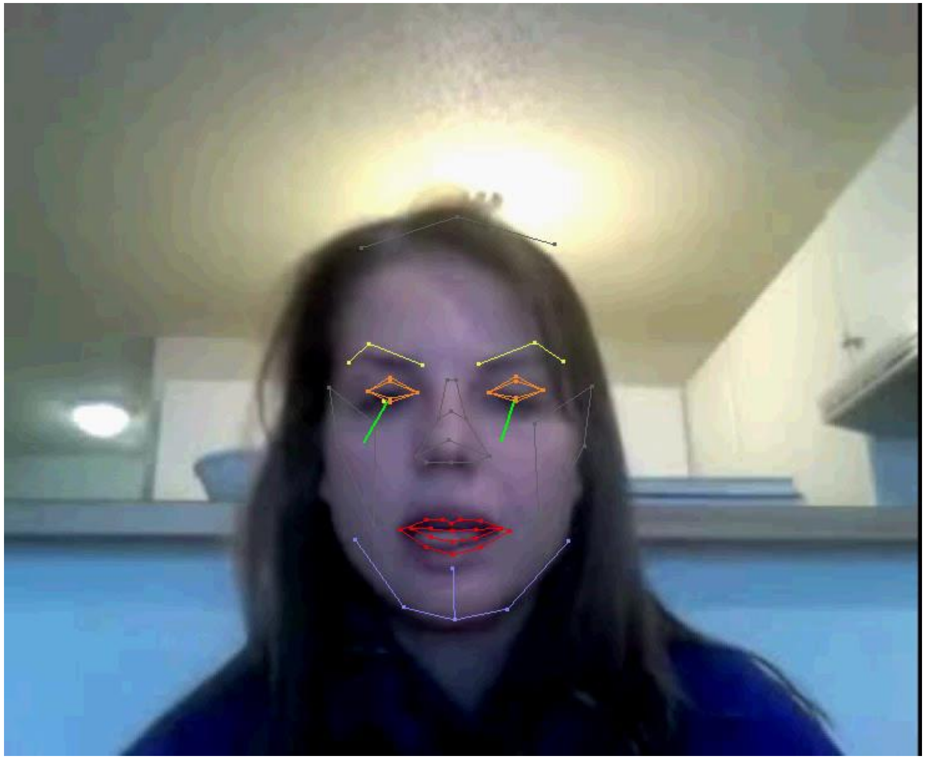
- SimSensei a "virtual therapist", ICT (Rizzo and Morency)

Automated Interviews?

Facial Expression:
Smiling, Positive Affect

Head Pose:
Nods/Shakes

Posture:
Leaning forward



Gaze: Averted,
Not looking
directly into
camera

Speech Tone:
Calm, Engaged

Affective and Cognitive State Analysis

Score

Overview

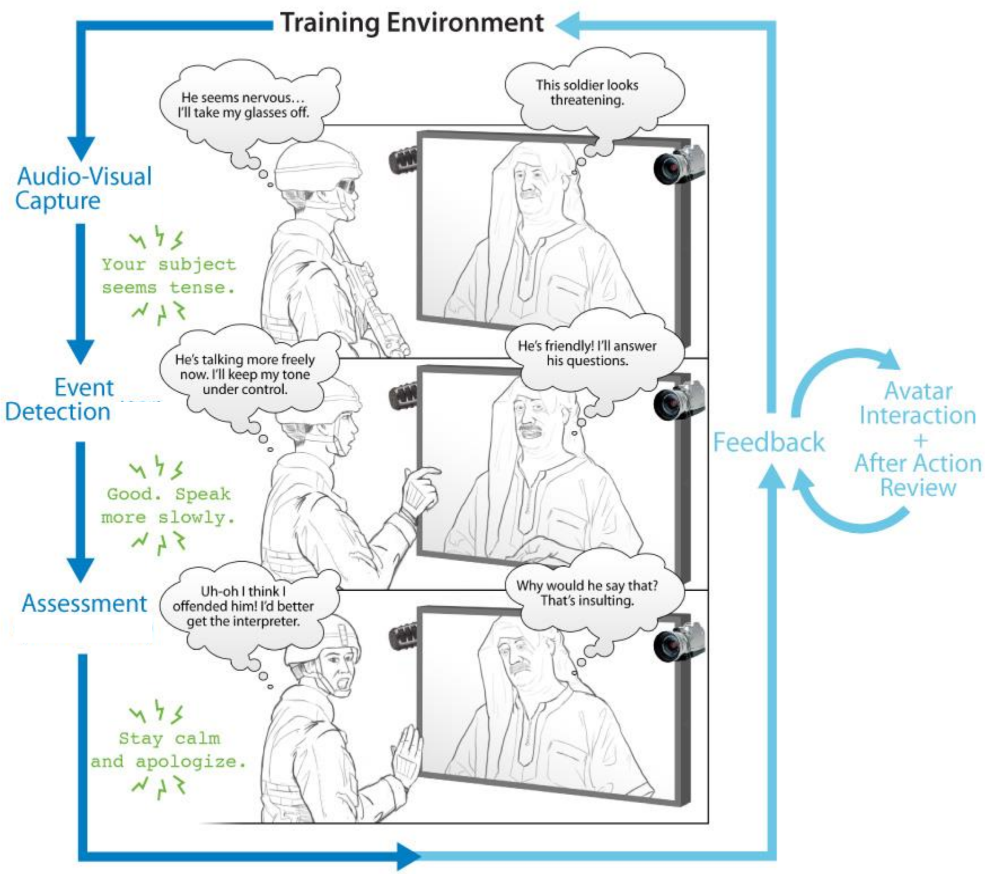
- Who we work with?
- What are we building?
- How we’re building it?
- How it’s used?
- Where do we go from here?

Who Do We Work With?

- Sociologists, Psychometricians, Ethnographers, SMEs
  - Goal: Study human behavior and how to impart pedagogy
- Computer Scientists
  - Goal: Develop the technology to implement the social training in a natural human machine interaction
- Social Psychologists
  - Goal: Study and evaluate impact of simulation tools and training

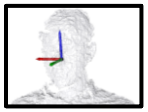
What we are building:
Multi-modal **I**ntegrated **B**ehavior **A**nalytics (MIBA)

- Interactive Game-Like Setup with Fluid Interactions
- Lifelike interactions
  - Real-time sensing of trainee behavior
  - Enable Real-time response of virtual characters
- Sensing of Trainee Behavior
  - Action Recognition – Gestures, Poses, Gaze, etc. – Large repertoire
  - Detection of prosody – speech tone etc.
  - Strong focus on non-verbal interaction to ensure culture general training
  - Interaction modeling
    - – Interaction between virtual character and trainee
    - – Conversation Modeling



Event Recognition at Multiple Time Resolutions

Head Pose

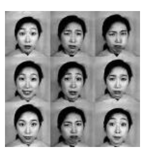


Gaze

Facial Muscles

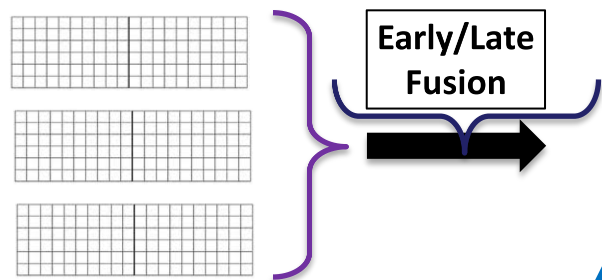
Body Pose

Facial Gestures

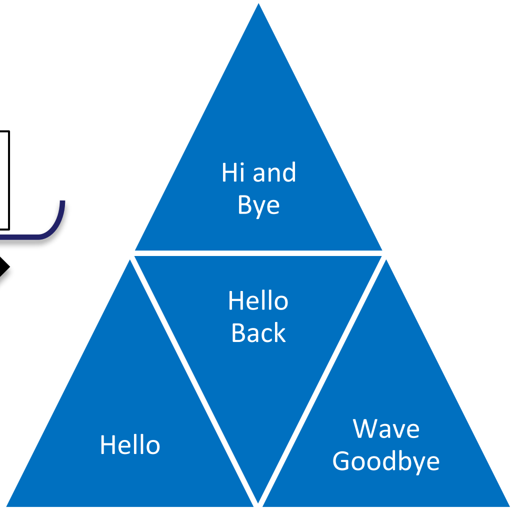


Paralinguistics

Baseline Approach



Multiple Temporal Scales



Multi-Resolution Event Classifier

Temporal Dynamics
Interaction Dynamics
Context

Full Body Affect: Gestures and Postures

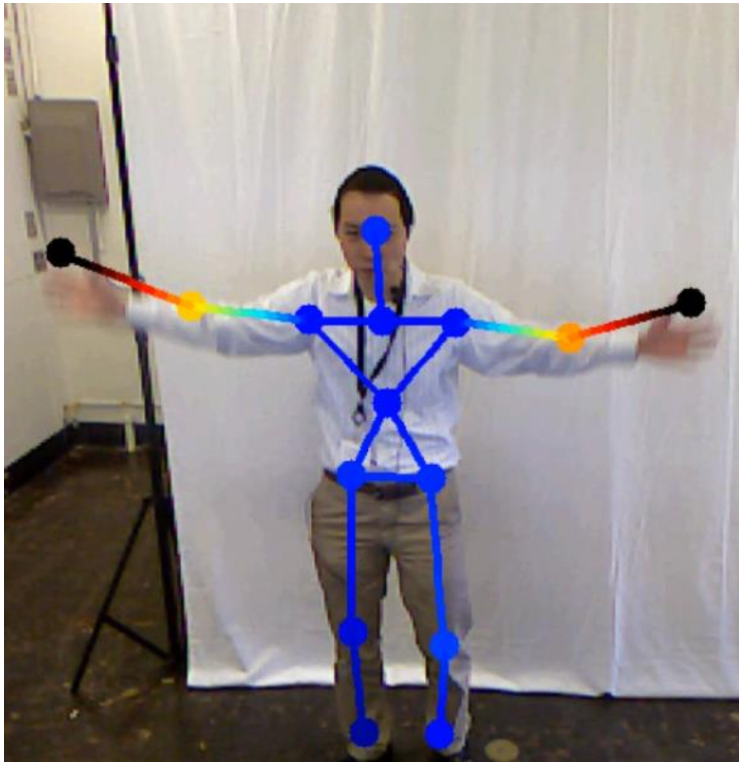
Interpreting Body Language

Full Body Affect Recognition

- The body is an important modality for expressing/recognizing affect complementing Facial Expressions and Vocalics
  - Some evidence that body posture is the influencing factor when the affective information displayed by body posture and facial expression are incongruent.
- Two kinds of information available
  - Static Pose (e.g., arms stretched out, head bent back, etc.)
  - Dynamics (e.g., smooth slow motions vs. jerky fast movements)
- Ideally should be independent of the actions performed and subject idiosyncrasies
- Public datasets for full body affect:
  - UCLIC, GEMEP, FABO, IEMOCAP

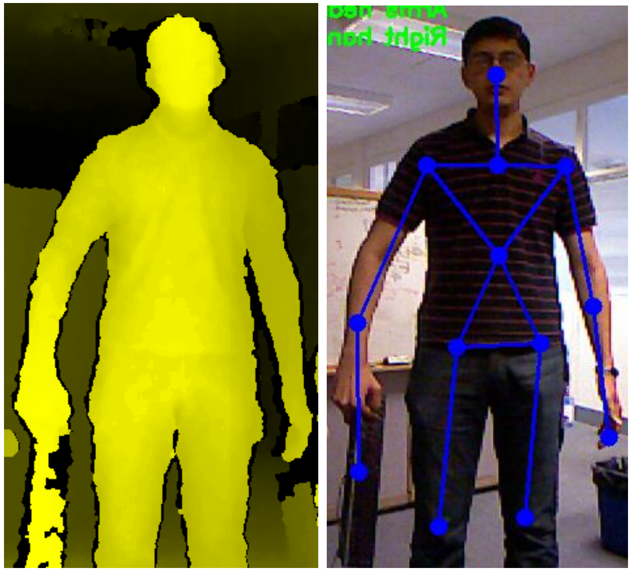
Elements of Interest

- **Specific Gestures**
  - Greeting, pointing, beckoning etc.
- **Head Posture** :
  - Bent backwards/forwards/upright/tilted
- **Arms:**
  - Raised/outstretched frontal or sideways/down/crossed/ elbows bent /arms at the side of the trunk
- **Shoulders**:
  - Lifted, slumped forward
- **Torso**:
  - Abdominal twist/straight, bowed trunk
- **Legs / Stance**:
  - Straight legs/ knees slightly bent/ stepping forward (triangular stance)
- **Motion:**
  - Smooth controlled motion / somewhat fast jerky motion / Large fluid slow motions /
- **Muscular States:**
  - Tense / Relaxed / Firm



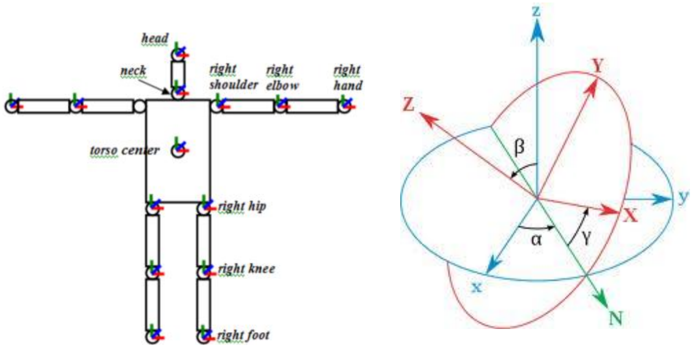
Skeletal Representation and Feature Set

- Articulated human model tracked with 15 joint locations using Microsoft **Kinect**
- Use Neck joint as reference
- Feature vector representing pose:
  - At any time frame *j*, the 3D locations of 14 joint locations relative to Neck joint
  - $V_j = \{\{v^1_{j,x}, v^1_{j,y}, v^1_{j,z}\}, \ldots, \{v^{14}_{j,x}, v^{14}_{j,y}, v^{14}_{j,z}\}\}$

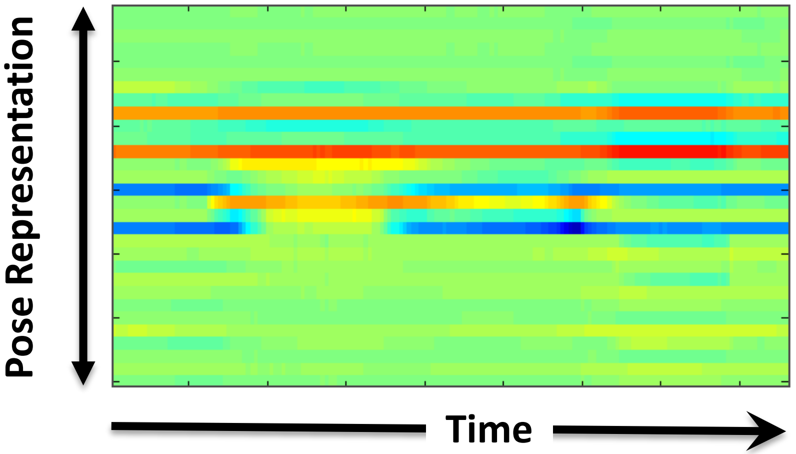


Articulated 3D Human Model

- Exemplars for training classifiers on pose and gesture models

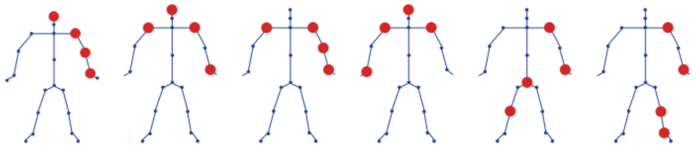


Features: Joint locations and angles



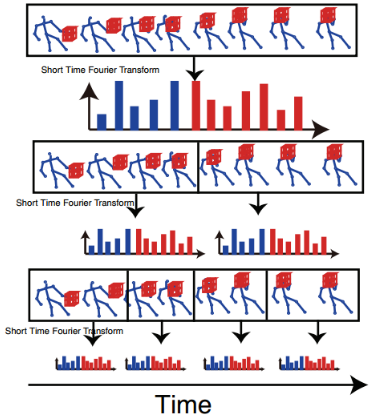
“Gesturelets”

- **Significance**: Individual limb/body-part motion is modeled



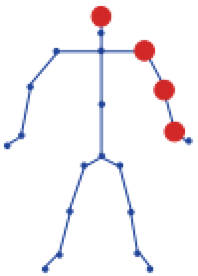
Yuan et al. CVPR 2012

- Novel Contributions
  - **Gesturelet ensemble**
    - Combination of discriminative Gesturelets (joint sets) for superior recognition
  - Pairwise Distance/Velocity Features
    - Pairwise Distance between each pair of joints
    - Pairwise velocity between each pair of joints
    - Normalized w.r.t. body size
    - Invariant to initial body orientation and absolute body position
  - Temporal Structure Representation
    - Models the Temporal Variation of Actions

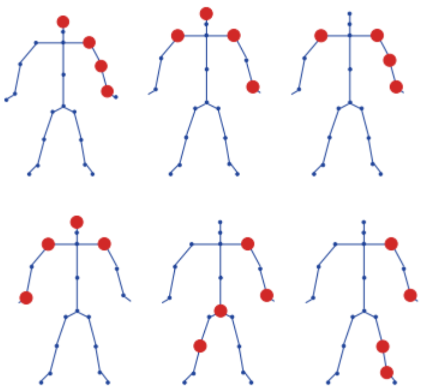


Gesturelets cont'd

- Gesturelet
  - Conjunctive Structure on base features
    - Base features are Fourier Temporal Pyramid representations of single joints
    - Represents the behavior of a set of joints

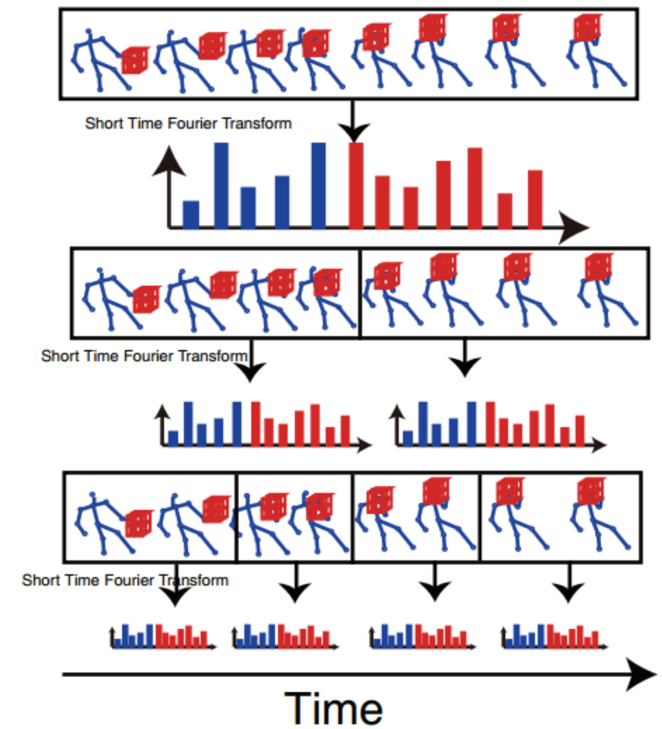


- Discriminative Gesturelets
  - Each action is characterized by interactions of a combination of a subset of joints
    - Determine a set of Gesturelets that recognize each action with high precision and recall
    - Set of discriminative Gesturelets learnt to represent each action and to capture the intra-class variance
  - Gesturelet Mining
    - Enormous number of possible Gesturelets
    - Greedy approach for mining a set of discriminative Gesturelets
- Gesturelet Ensemble
  - Combine the set of discriminative Gesturelets
    - PLS based dimensionality reduction for real time performance
    - SVM based classsifier to learn a model for the set of Gesturelets



Gesturelets cont'd

- Temporal Structure Representation
  - Temporal Variation of Actions
    - People perform actions at different speeds
    - Different segments of the same action are performed at variable rates
  - Fourier temporal pyramid
    - Robust to temporal variation of actions
    - Approach
      - Recursively partition action into a temporal pyramid
      - Apply Short Fourier Transform to each dimension
    - Advantages
      - Discards high frequency coefficients that often contain noise
      - Pyramidal structure makes it invariant to temporal misalignment



Training Dataset Statistics

- 2200 action instances
- 10 action classes – checkpoint scenario
- 20 Actors (expanding)

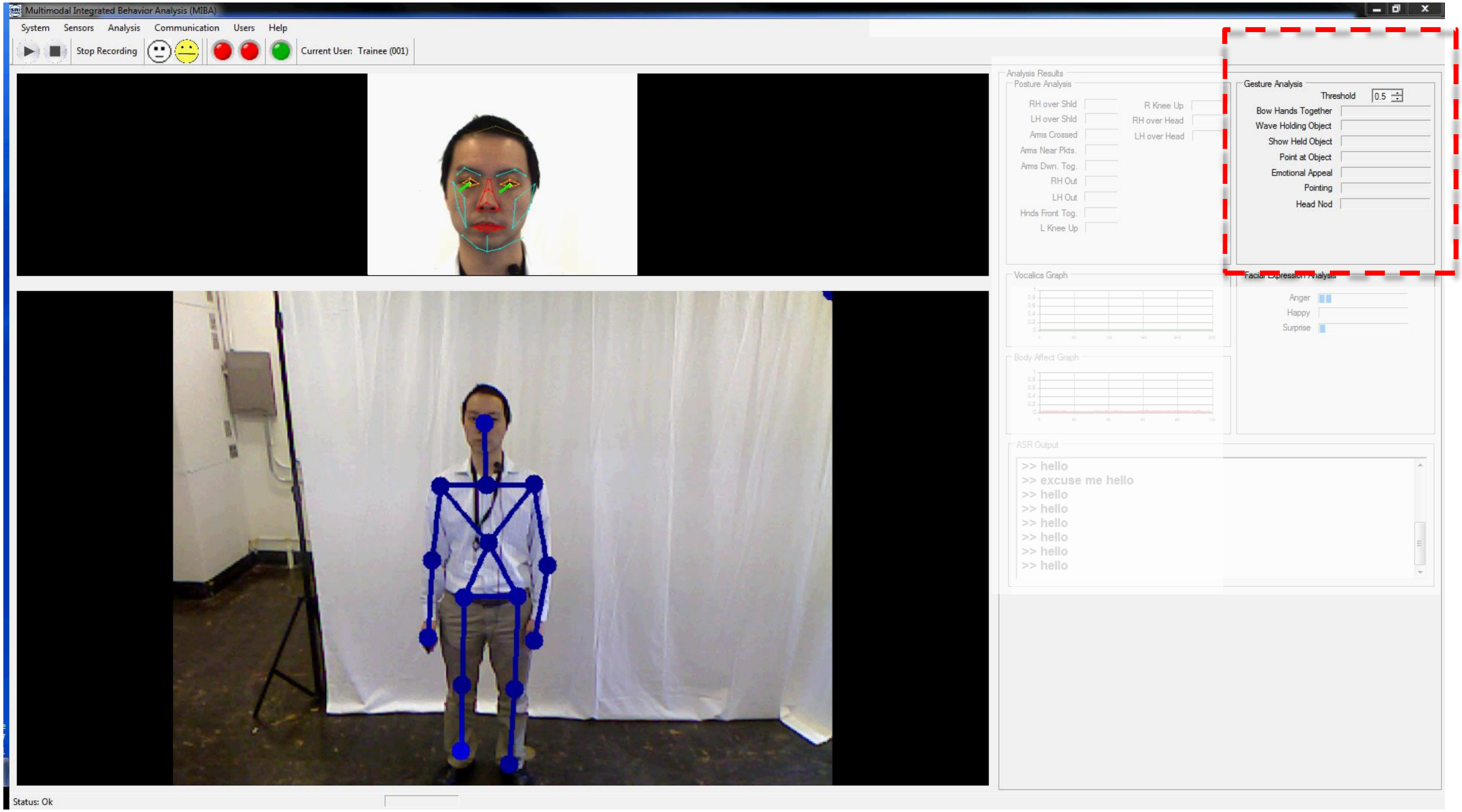
- Actions
  - 10 classes – checkpoint scenario
    - Folding Arms (208)
    - Right Hand Forward (215)
    - Head Nod (217)
    - Right Hand to Face (216)
    - Both Hands Extended (213)
    - Right Hand Wave (221)
    - Stop (213)
    - Swinging Arms Sideways (223)
    - Beckon-1 (200)
    - Beckon-2 (200)

Results

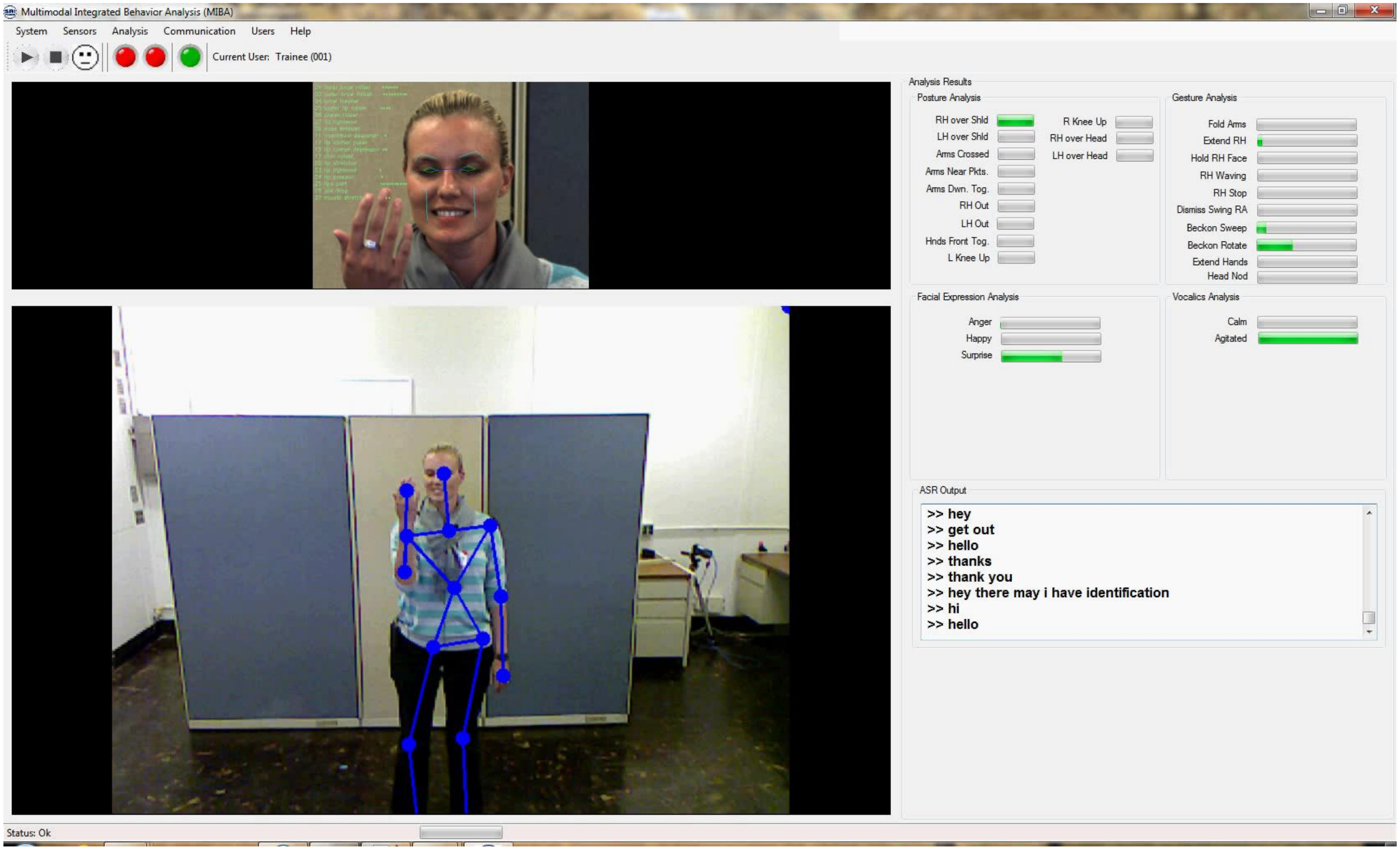
- Leave One Person Out
  - Train on 19 and test on the 20th

	FA	RF	RFS	RHF	RHF S	RH W	S	SAS	N	Neg
FA	**0.98**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.02
RF	0.0	**0.98**	0.01	0.0	0.0	0.0	0.0	0.01	0.0	0.0
RFS	0.0	0.03	**0.93**	0.0	0.0	0.0	0.03	0.0	0.0	0.0
RHF	0.0	0.01	0.0	**0.95**	0.02	0.01	0.0	0.0	0.0	0.0
RHFS	0.0	0.0	0.0	0.03	**0.93**	0.0	0.03	0.0	0.0	0.0
RHW	0.0	0.0	0.0	0.0	0.01	**0.80**	0.19	0.0	0.0	0.0
S	0.0	0.0	0.02	0.0	0.0	0.29	**.68**	0.0	0.0	0.0
SAS	0.0	0.0	0.01	0.0	0.0	0.01	0.0	**0.90**	0.0	0.07
N	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	**0.99**	0.01
Neg	0.01	0.0	0.0	0.0	0.0	0.01	0.02	0.02	0.14	**0.78**

Gestures-MIBA

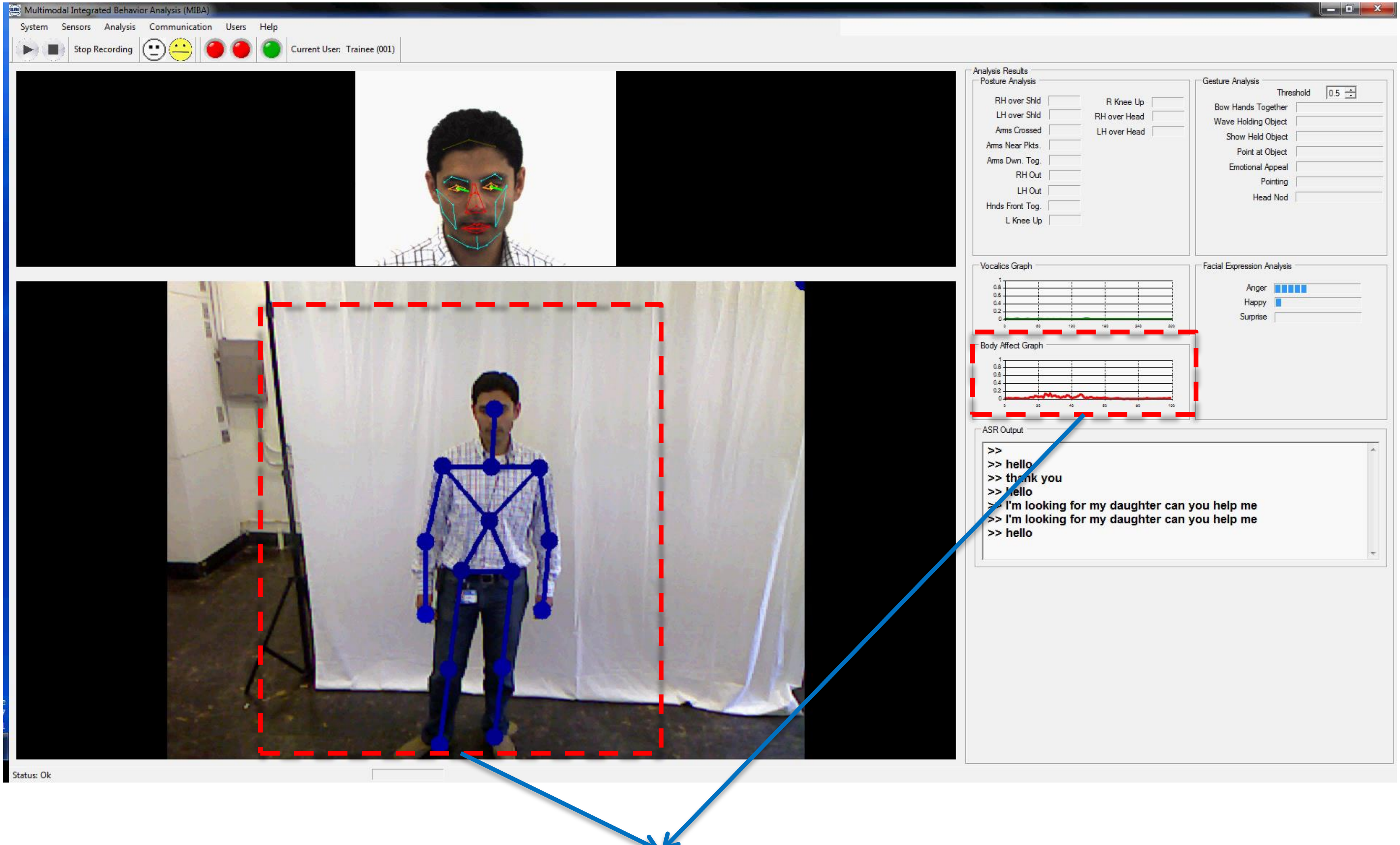


Demo video



Multimodal Integrated Behavior Analysis (MIBA)
System Sensors Analysis Communication Users Help
Current User: Trainee (001)
Analysis Results
Posture Analysis
RH over Shld
LH over Shld
Arms Crossed
Arms Near Pkts.
Arms Dwn. Tog.
RH Out
LH Out
Hnds Front Tog.
L Knee Up
R Knee Up
RH over Head
LH over Head
Gesture Analysis
Fold Arms
Extend RH
Hold RH Face
RH Waving
RH Stop
Dismiss Swing RA
Beckon Sweep
Beckon Rotate
Extend Hands
Head Nod
Facial Expression Analysis
Anger
Happy
Surprise
Vocalics Analysis
Calm
Agitated
ASR Output
>> hey
>> get out
>> hello
>> thanks
>> thank you
>> hey there may i have identification
>> hi
>> hello
Status: Ok

Real-time Visualization of Body Affect and Motion Hotspotting



Body affect time series data

SRI International

Facial Expression Recognition

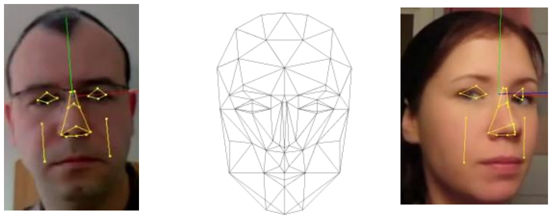
Challenges to address

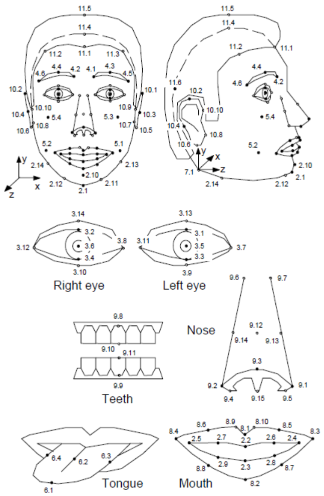
- Dynamic nature of facial expressions
  - Variations in intensity
- Spontaneity of facial gestures
- What are the relevant set of facial responses for the application?
  - Both posed and spontaneous
  - What do they control?
- When is system performance good enough?
  - Algorithm speed vs. accuracy
  - Sensor resolution and sensitivity requirements

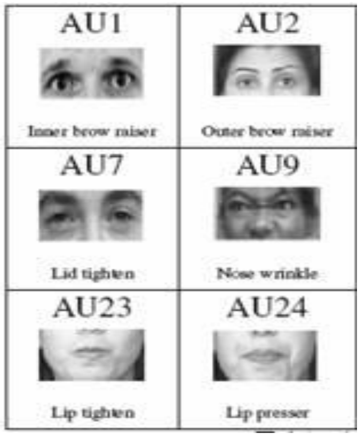
- **Intuition:** Sequence based approach that models dynamics will allow both:
  - Spontaneous expressions
  - Intensity

Expression Recognition Basic Components

- Head Tracking and Face Normalization
  - Jointly estimated by Kinect and Vision
- Facial Feature Extraction
  - Low Level Facial Feature Extraction (fiducial points)
  - Mid-level Features (Action Units: clusters of fiducial points)
- Emotional State Classification
  - Ground Truth Issues: Posed vs. spontaneous; gesture vs. reaction
  - Integration with body and speech tone analyses



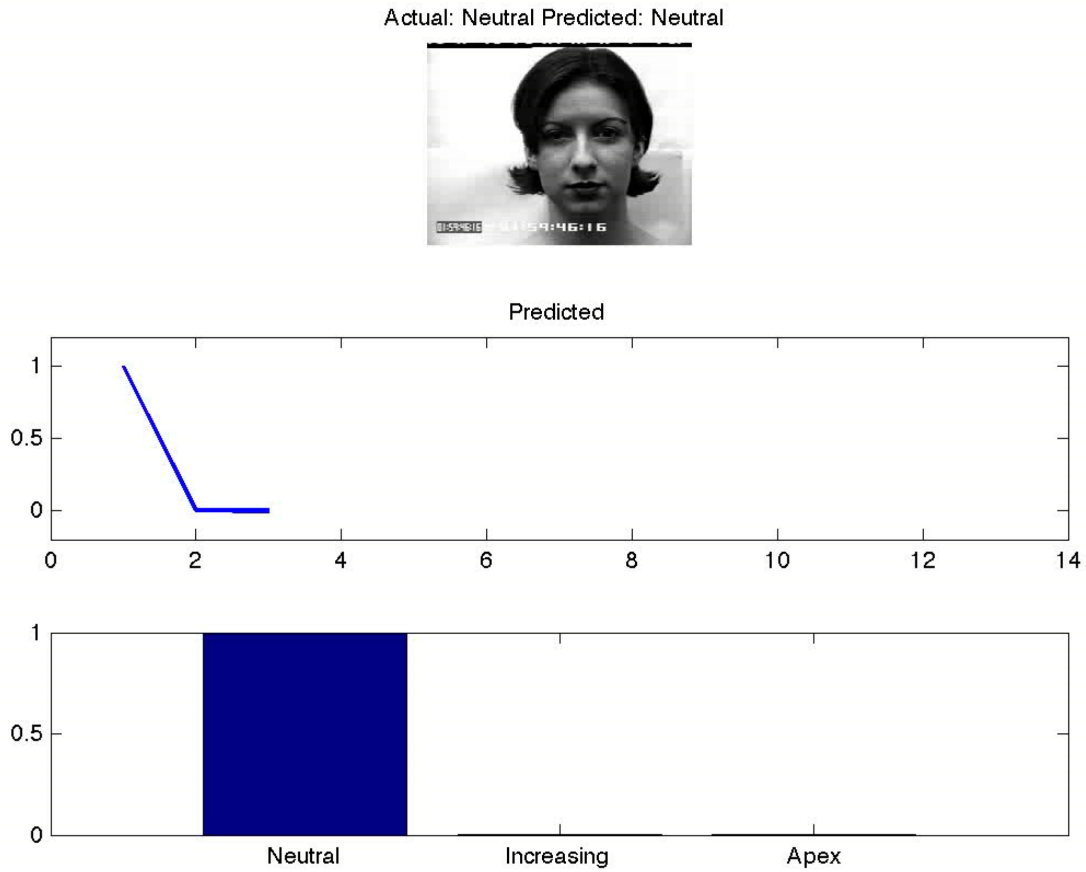




Our Approach

- Divide life-span of a facial expression into three phases: Onset, Increasing and Apex
- The increasing phase basically contains the real dynamics of the expression
- Conditional Random Fields (**CRFs**) to model temporal dynamics
  - CRFs are specifically designed and trained to maximize performance of sequence labeling. They model the *conditional distribution* P(*Q* | *O)*
  - CRFs also easily allow adding additional features without making independence assumptions.

Happiness: Neutral-Increasing-Apex



Empirical Results

CRF Results

	Precision	Recall	F1
An	79.487	72.093	75.61
Co	65	72.222	68.421
Di	82.258	94.444	87.931
Fe	85	70.833	77.273
Ha	94.03	95.455	94.737
Sa	78.571	84.615	81.481
Su	97.26	91.026	94.04

Gaze Detection

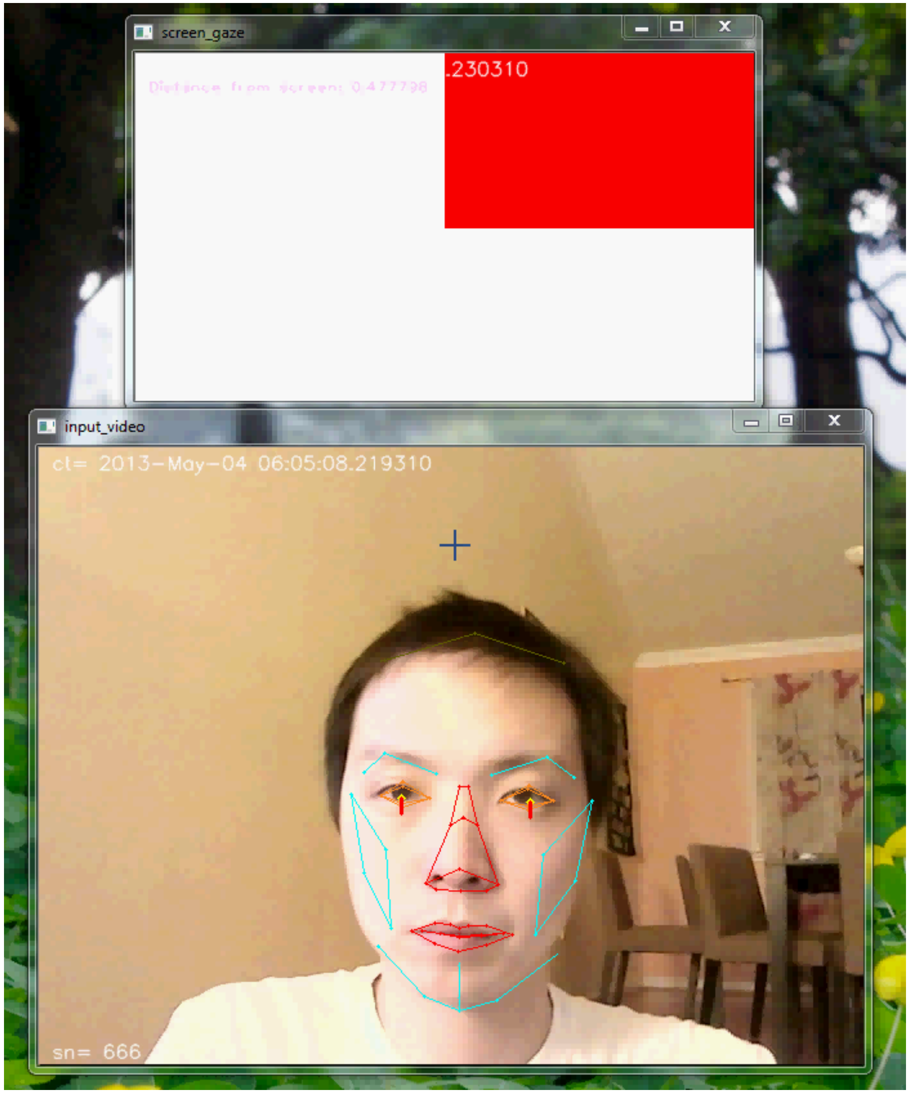
Tracking Gaze in Human Interactions

- Real-time gaze tracking *from monocular camera (no IR)*
  - Gaze vector
  - Gaze consistency
- Calculates the gaze vector and *where the learner is looking on the screen*
- Currently quadrant-level accuracy, but future improvements planned

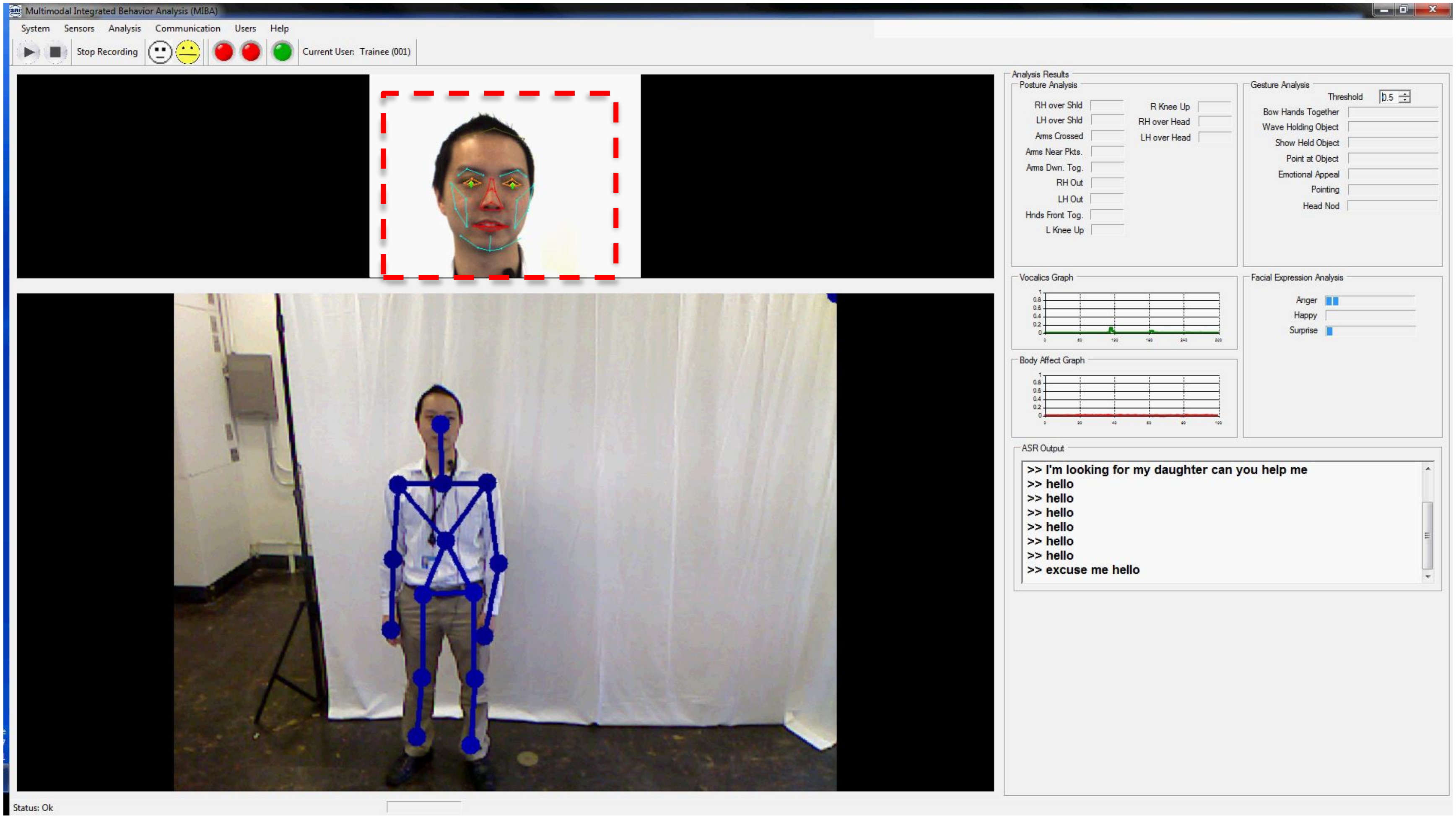
Gaze Constancy

- Infers if someone is:
  - Staring intensely (red)
  - Normal gaze (green)
  - Surveying the scene (blue)

- Learning relevance of gaze to important learner affects:
  - Nervousness
  - Attentiveness



Gaze Analysis Integrated in MIBA



SRI International

Paralinguistics

Challenging Problem!

- Challenges
  - Differences in emotion are subtle – even to Humans
  - Spontaneous vs acted – Same Challenge as Gestures
  - Collecting labeled data is a challenge
    - Agreement on affect
    - Proper segmentation

Angry	Excited	Fear	Frustrated	Happy	Sad	Surprise

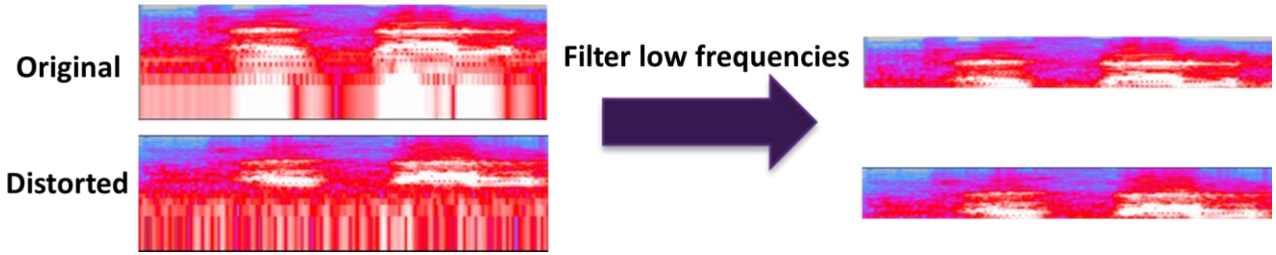
Seattle Police Database

- **Dataset Details**
  - Size
    - 7 classes
      - Neutral, Authoritative, Extremely Agitated, Agitated, Slightly Agitated, Indignant, Placating
    - 1723 samples
  - Annotation
    - Same person annotated twice independently
    - 1158 valid samples
      - N (421), A (74), EA (34), A (226), S A (391), I (12) , P (0)
  - Misc.
    - Audio Segments chosen from a full recording

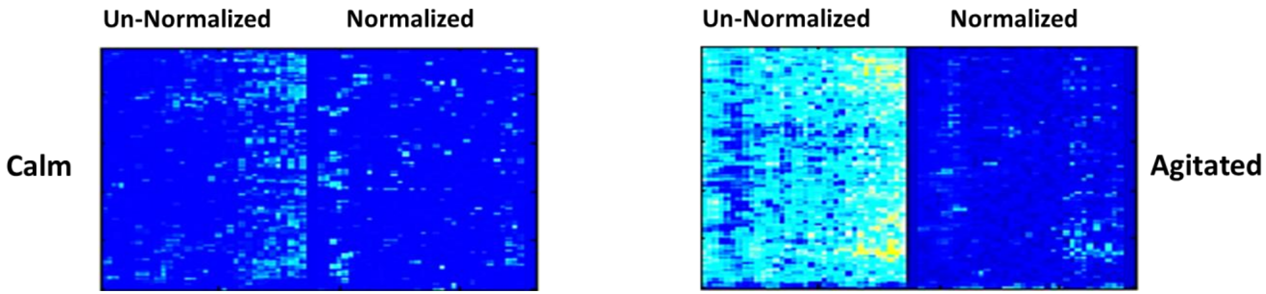
Neutral	Authoritative	Sl. Agitated	Agitated	Ex. Agitated	Indignant

Technical Approach

- **Spectrogram**
  - Filter low frequency bands: invariance to noise



- **Concatenate Normalized and Un-normalized spectrograms:**
  - Volume dependent and independent features



Preliminary Results

- **SRI-Rutgers**

	Agitated	Calm
Agitated	**0.932**	0.068
Calm	0.052	**0.948**

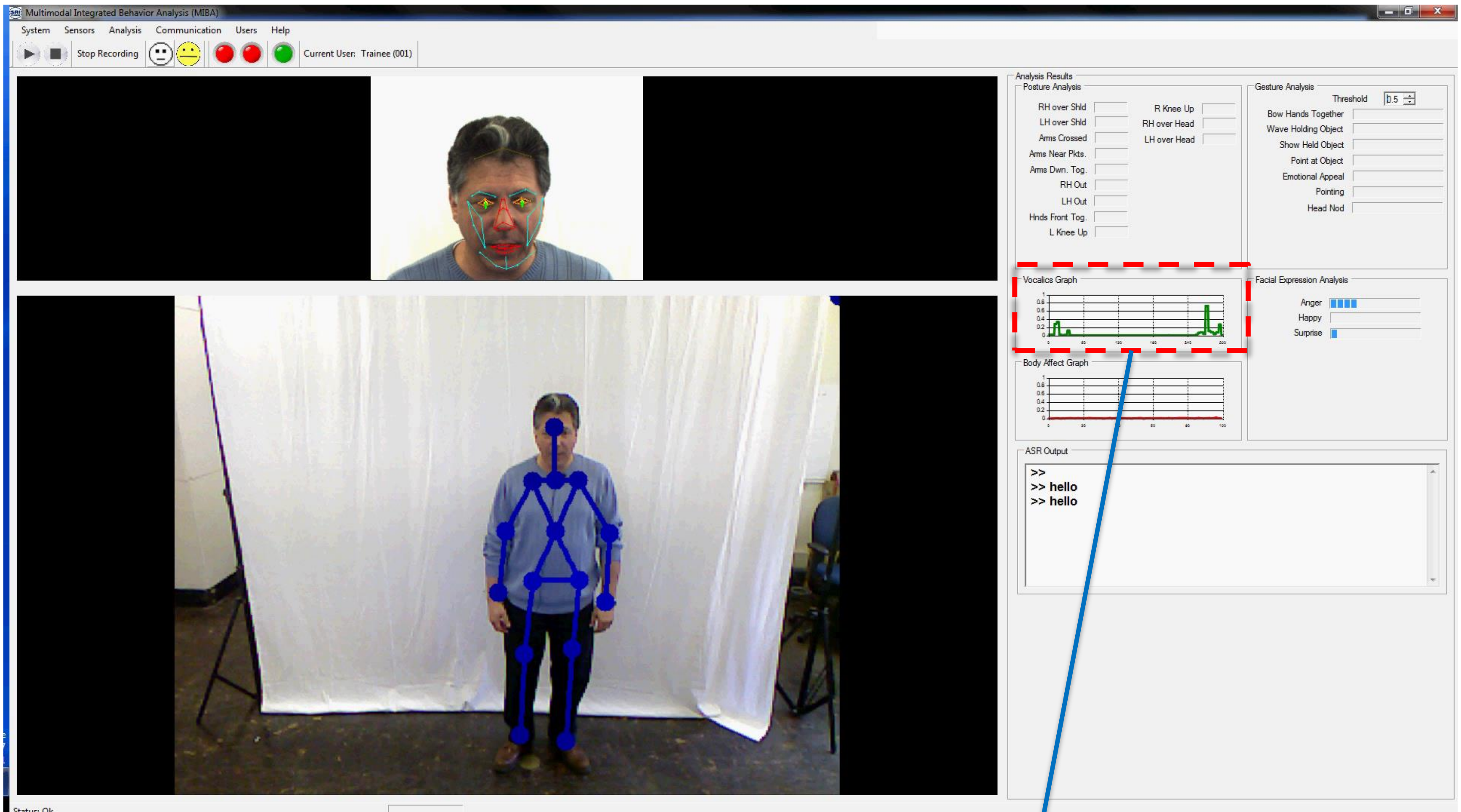
- **Seattle Police Department**
  - Unable to detect Authoritative speech
  - For other categories N, SA, A, EA
    - Performance variation as expected

neutral

agitated

	Acc	mAcc
N-Au-(SA+A+EA)	70.94	50.40
N-(SA+A+EA)	77.15	75.60
N-(A+EA)	89.57	88.48
N-SA	74.01	73.99
N-A	90.26	88.62
N-EA	98.68	93.88
N-Au	85.45	53.00
N-Au-(A+EA)	79.34	58.34

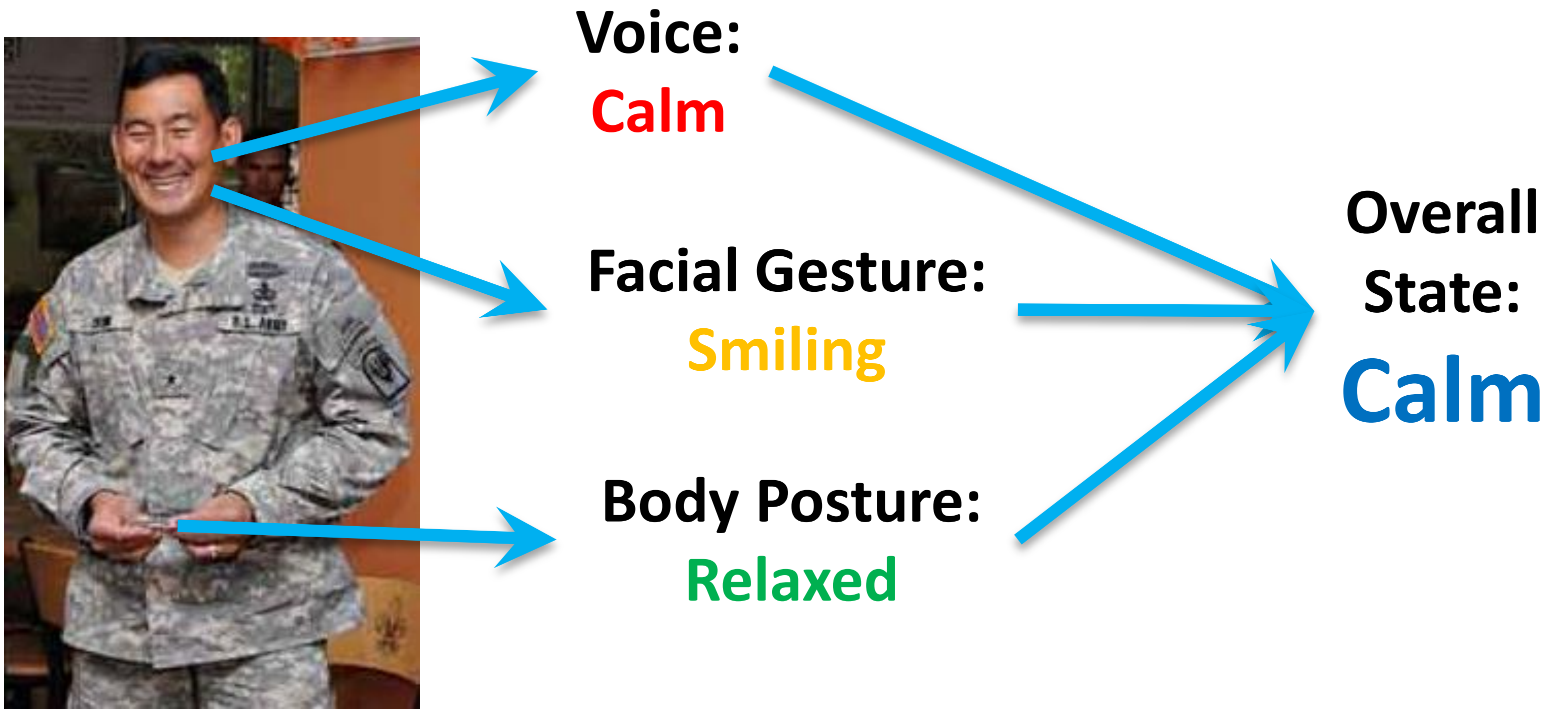
Visualization of Paralinguistics



Paralinguistics time-series data

Multimodal Affect Estimation

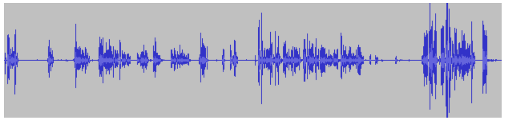
Holistic Assessment of Behavior – Multimodal Sensing



- Need to combine multiple cues to arrive at holistic assessment of user state
  - Body Pose, Gestures, Facial Expressions, Speech Tone, Keywords-> NLU
  - Starting simple with state on one scale
    - Threat level (agitated vs. calm to start)
- Provides contextual effects to produce predictions of behavior at Gottman's "construct" level of behavioral classification.

Individual Affect Modeling – Training Datasets

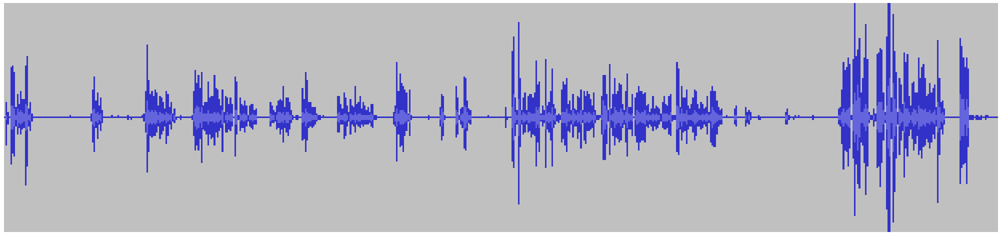
- **AVEC 2011 dataset**
  - Audio Visual Emotion Challenge
    - **Aim:** compare machine learning methods for audio, visual and audio-visual emotion analysis.
  - Dataset Details
    - Elicited Emotions: participants talk to emotionally stereotyped characters.
    - Over 8 hours of audio and video data.



  - Binary Labels
    - **Activation(Arousal):** is the individual's global feeling of dynamism or lethargy.
    - **Expectation (Anticipation):** subsumes various concepts that can be separated as expecting, anticipating, being taken unaware.
    - **Power(Dominance):** dimension subsumes two related concepts, power and control.
    - **Valence:** is an individual's overall sense of "weal or woe": Does it appear that on balance, the person rated feels positive or negative about the things, people, or situations at the focus of his/her emotional state?

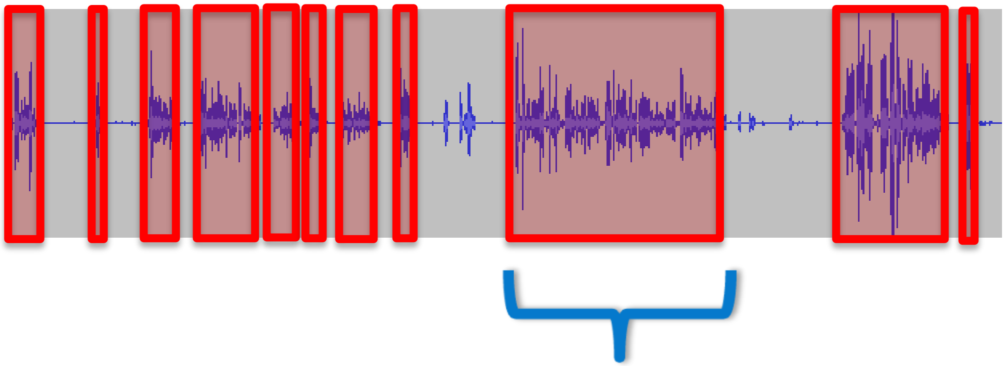
Feature Representation: Audio

- **Audio based Affect Recognition**
  - Features
    - Energy
    - Spectra
    - Voicing
    - Derivatives of energy/spectral features



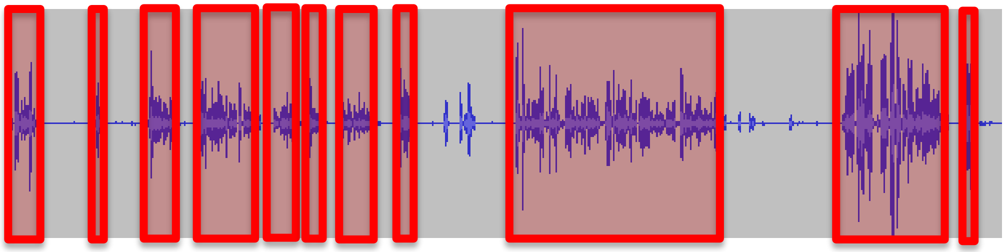
Feature Representation: Audio

- **Audio based Affect Recognition**
  - Features
    - Energy
    - Spectra
    - Voicing
    - Derivatives of energy/spectral features
  - Representation
    - Word Segmentation
    - Functionals over each feature



Feature Representation: Audio

- **Audio based Affect Recognition**
  - Features
    - Energy
    - Spectra
    - Voicing
    - Derivatives of energy/spectral features
  - Representation
    - Word Segmentation
    - Functionals over each feature
  - Dimensionality Reduction
    - Partial Least Squares
    - Supervised (Class Aware) Dimensionality Reduction

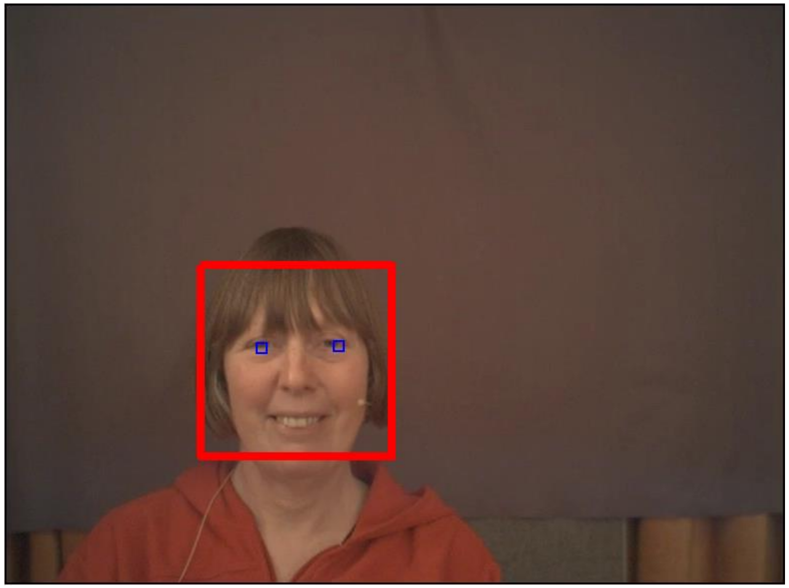


$$[\mathrm{cov}(t_i, u_i)]^2 = \max_{|w_{xi}|=1, |w_{yi}|=1} [\mathrm{cov}(Xw_{xi}, Yw_{yi})]^2$$

$$W_x = \{w_{x1}, w_{x2}, \ldots, w_{xp}\} \qquad W_y = \{w_{y1}, w_{y2}, \ldots, w_{yp}\}$$

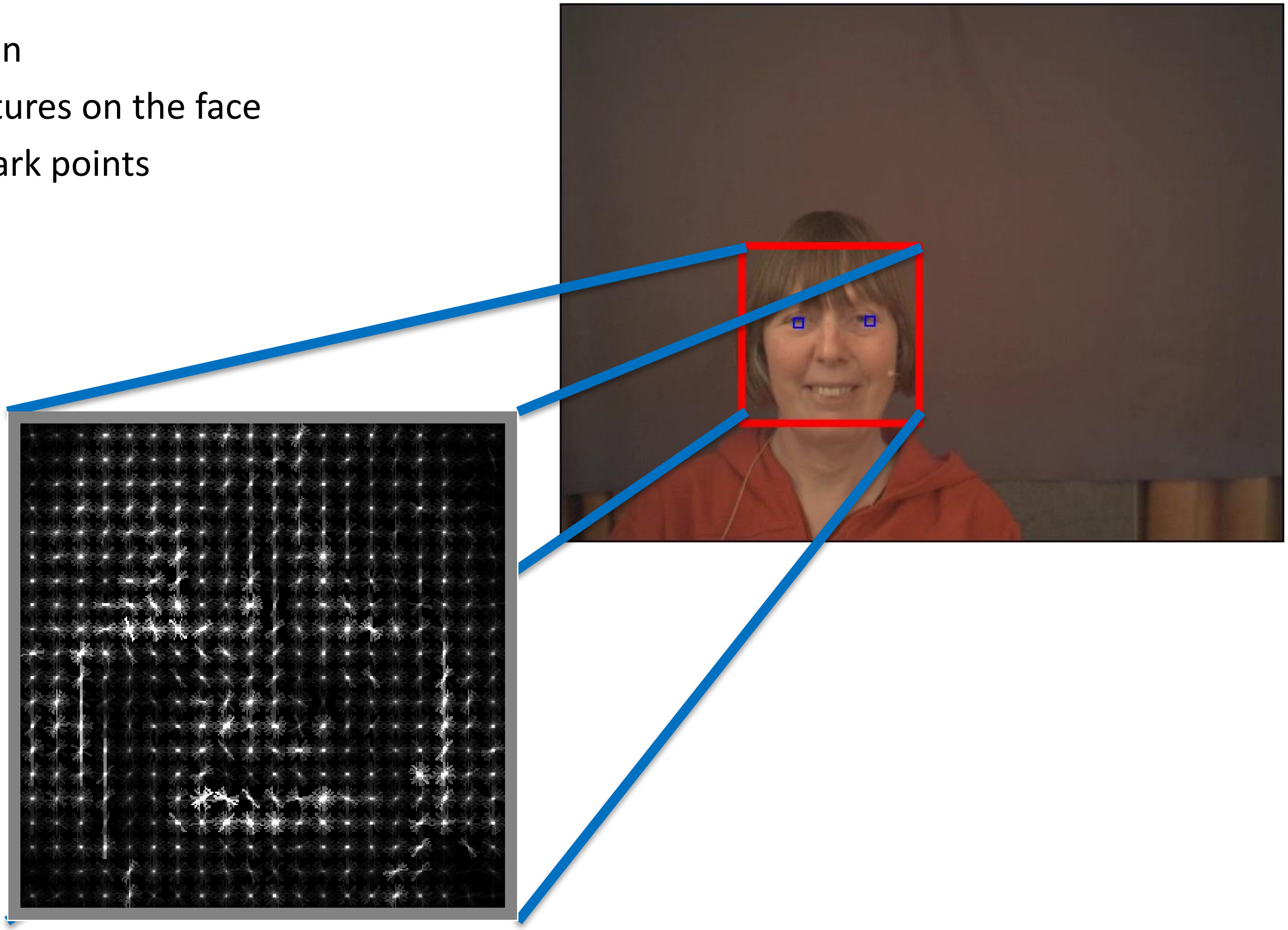
Feature Representation: Video

- **Video based Affect Recognition**
  - Features
    - Face Detection
    - LBP/HOG features on the face
    - Facial Landmark points



Feature Representation: Video

- **Video based Affect Recognition**
  - Features
    - Face Detection
    - LBP/HOG features on the face
    - Facial Landmark points

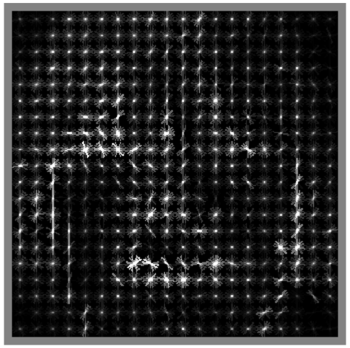


Feature Representation: Video

- **Video based Affect Recognition**
  - Features
    - Face Detection
    - LBP/HOG features on the face
    - Facial Landmark points

Feature Representation: Video

- **Video based Affect Recognition**
  - Features
    - Face Detection
    - LBP/HOG features on the face
    - Facial Landmark points
  - Representation
    - Framewise
  - Dimensionality Reduction
    - Partial Least Squares

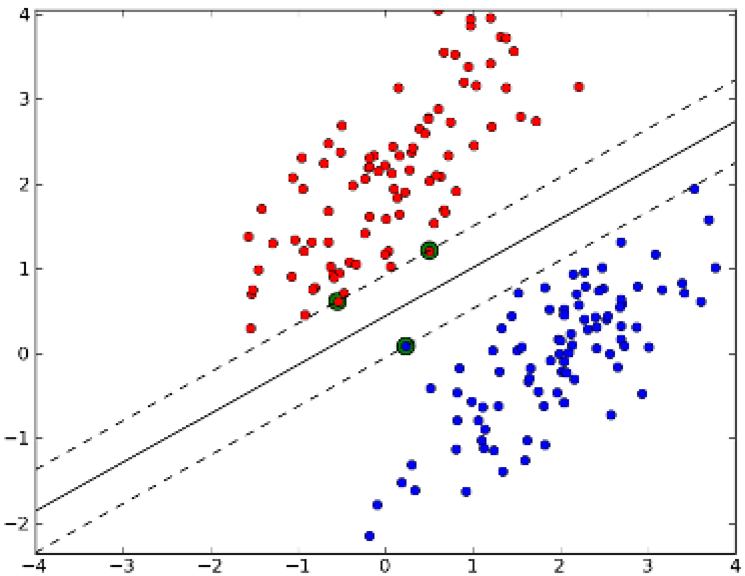


$$[\mathrm{cov}(t_i, u_i)]^2 = \max_{|w_{xi}|=1, |w_{yi}|=1} [\mathrm{cov}(Xw_{xi}, Yw_{yi})]^2$$

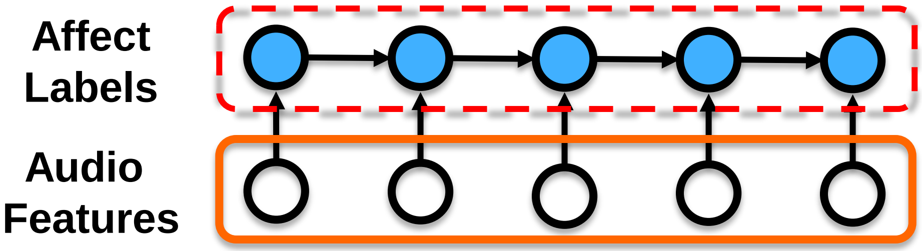
$$W_x = \{w_{x1}, w_{x2}, \ldots, w_{xp}\} \; W_y = \{w_{y1}, w_{y2}, \ldots, w_{yp}\}$$

Classifier Options

- **Audio based Affect Recognition**
  - Classifier
    - Static vs Dynamic Classifiers

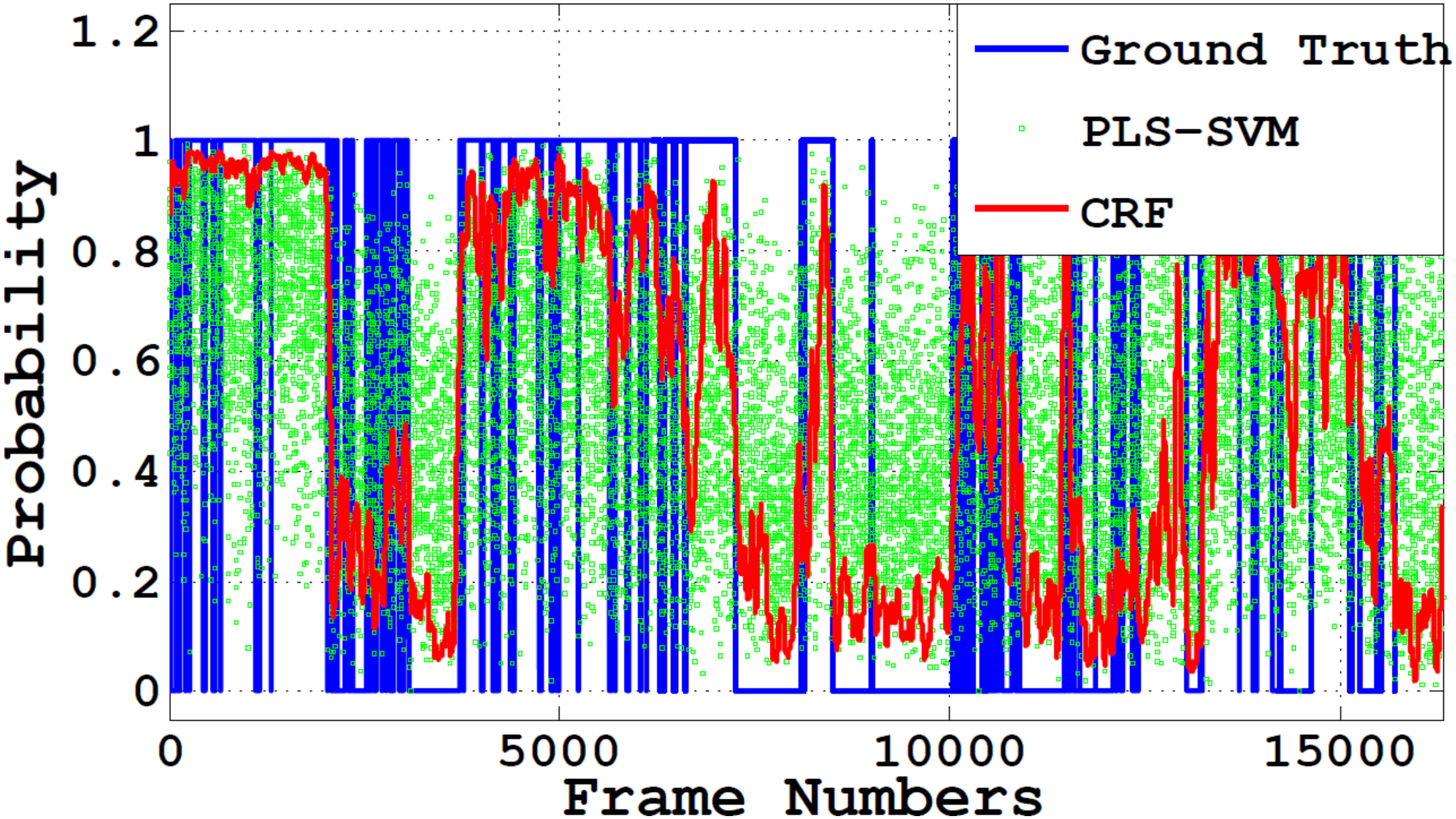


SVM



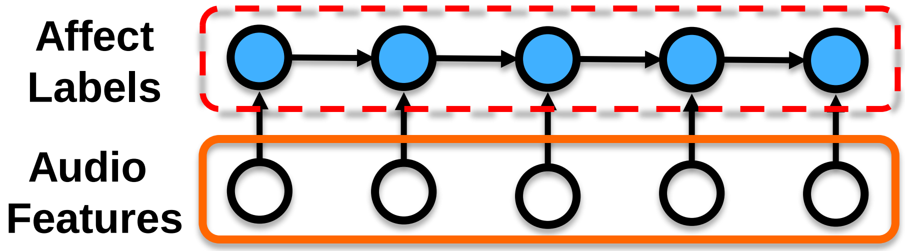
CRF

Modeling Temporal Dynamics with CRFs

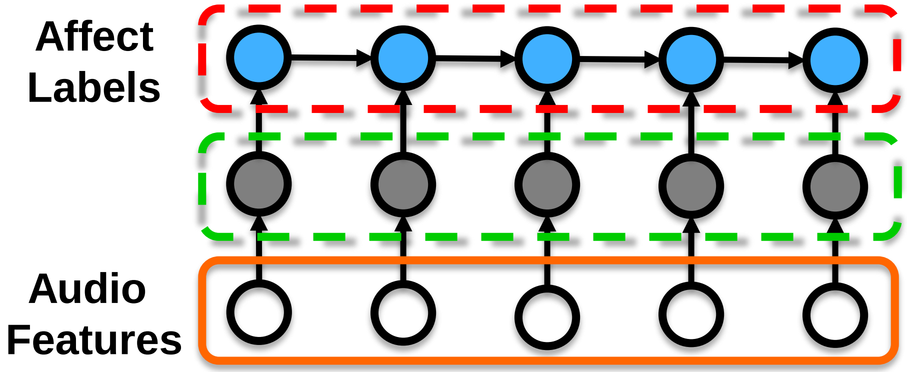


Adding Hidden Layers in Graphical Model

- **Audio based Affect Recognition**
  - Classifier
    - Static vs Dynamic Classifiers
    - CRF vs HMMs
    - Hidden CRFs



CRF



Hidden State CRF

Evaluating Impact: Audio

- **Audio based Affect Recognition**
  - Results

	A	E	P	V	mean
raw-SVM	63.7	63.2	65.6	58.1	**62.65**
PLS-SVM	64.6	66.6	66.2	61.9	**64.81**
PLS-CRF	76.9	65.5	68.7	61.7	**68.20**
PLS-HCRF	73.4	65.5	68.7	70.0	**69.42**

$$[\mathrm{cov}(t_i, u_i)]^2 = \max_{|w_{xi}|=1, |w_{yi}|=1} [\mathrm{cov}(Xw_{xi}, Yw_{yi})]^2$$

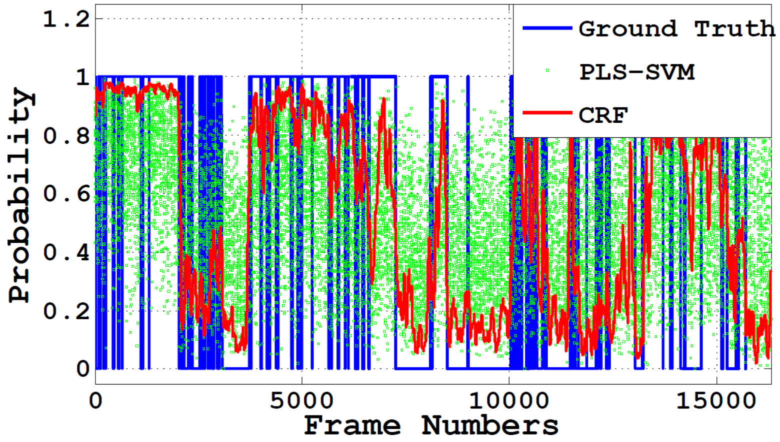
$$W_x = \{w_{x1}, w_{x2}, \ldots, w_{xp}\} \qquad W_y = \{w_{y1}, w_{y2}, \ldots, w_{yp}\}$$

PLS

Evaluating Impact: Audio

- **Audio based Affect Recognition**
  - Results

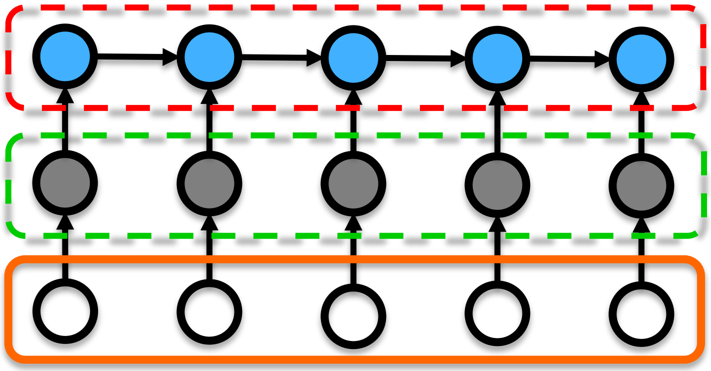
	A	E	P	V	mean
raw-SVM	63.7	63.2	65.6	58.1	**62.65**
PLS-SVM	64.6	66.6	66.2	61.9	**64.81**
PLS-CRF	76.9	65.5	68.7	61.7	**68.20**
PLS-HCRF	73.4	65.5	68.7	70.0	**69.42**



Evaluating Impact: Audio

- **Audio based Affect Recognition**
  - Results

	A	E	P	V	mean
raw-SVM	63.7	63.2	65.6	58.1	**62.65**
PLS-SVM	64.6	66.6	66.2	61.9	**64.81**
PLS-CRF	76.9	65.5	68.7	61.7	**68.20**
PLS-HCRF	73.4	65.5	68.7	70.0	**69.42**



HCRF

Evaluating Impact: Video

- **Video based Affect Recognition**
  - Results

	A	E	P	V	mean
raw-SVM	60.2	58.3	56.0	63.6	**59.52**
PLS-SVM	68.1	57.3	55.4	68.9	**62.43**
PLS-CRF	69.5	59.1	55.3	68.8	**63.17**
PLS-HCRF	70.1	59.5	55.4	68.8	**63.45**

$$[\mathrm{cov}(t_i, u_i)]^2 = \max_{|w_{xi}|=1, |w_{yi}|=1} [\mathrm{cov}(X w_{xi}, Y w_{yi})]^2$$

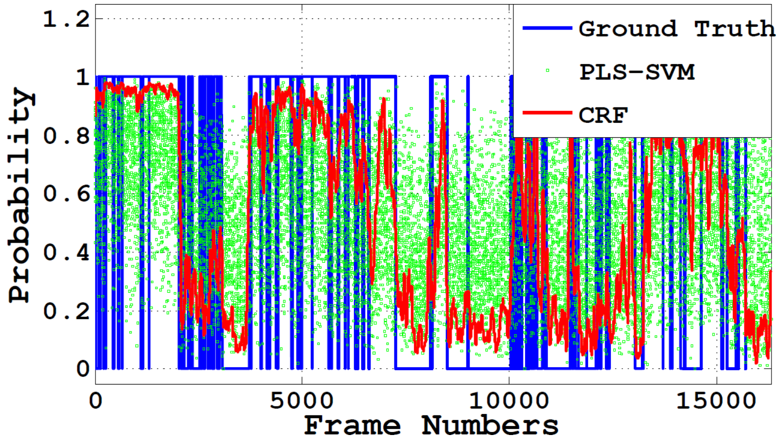
$$W_x = \{w_{x1}, w_{x2}, \ldots, w_{xp}\} \qquad W_y = \{w_{y1}, w_{y2}, \ldots, w_{yp}\}$$

PLS

Evaluating Impact: Video

- **Video based Affect Recognition**
  - Results

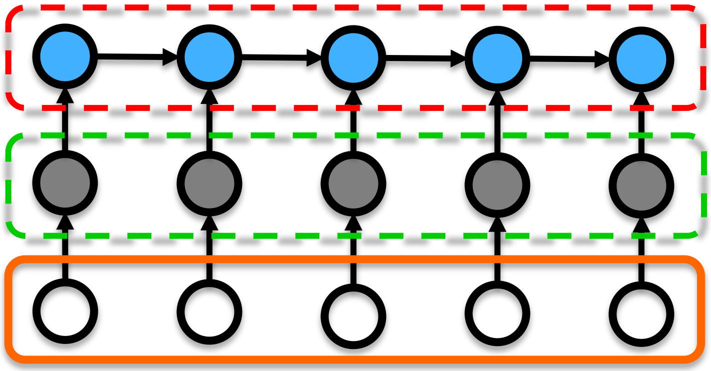
	A	E	P	V	mean
raw-SVM	60.2	58.3	56.0	63.6	**59.52**
PLS-SVM	68.1	57.3	55.4	68.9	**62.43**
PLS-CRF	69.5	59.1	55.3	68.8	**63.17**
PLS-HCRF	70.1	59.5	55.4	68.8	**63.45**



Evaluating Impact: Video

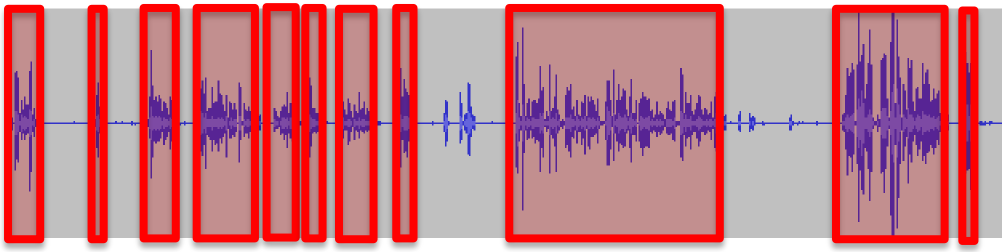
- **Video based Affect Recognition**
  - Results

	A	E	P	V	mean
raw-SVM	60.2	58.3	56.0	63.6	**59.52**
PLS-SVM	68.1	57.3	55.4	68.9	**62.43**
PLS-CRF	69.5	59.1	55.3	68.8	**63.17**
PLS-HCRF	70.1	59.5	55.4	68.8	**63.45**

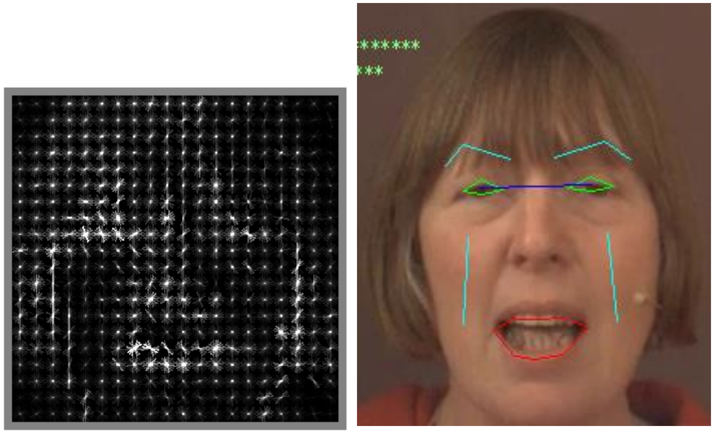


HCRF

Multi-modal Fusion



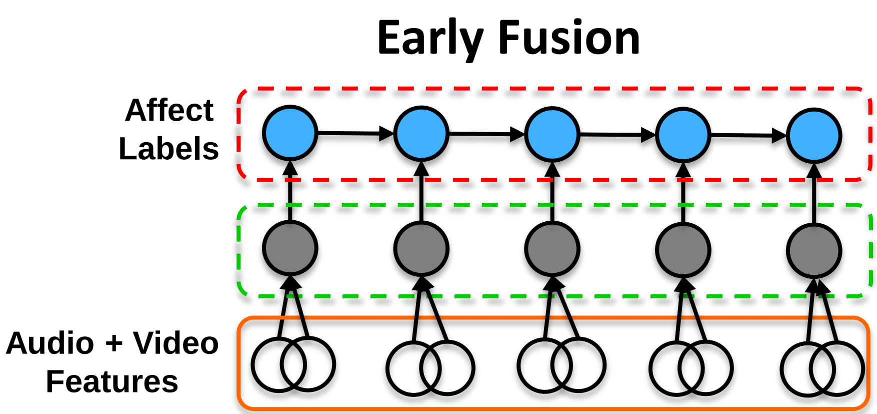
Audio Features

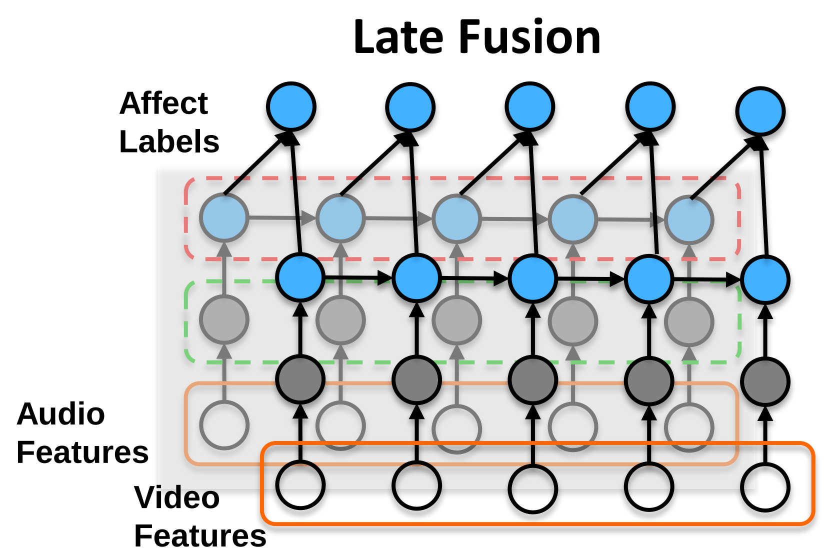


Video Features

Multi-modal Fusion – Traditional Options

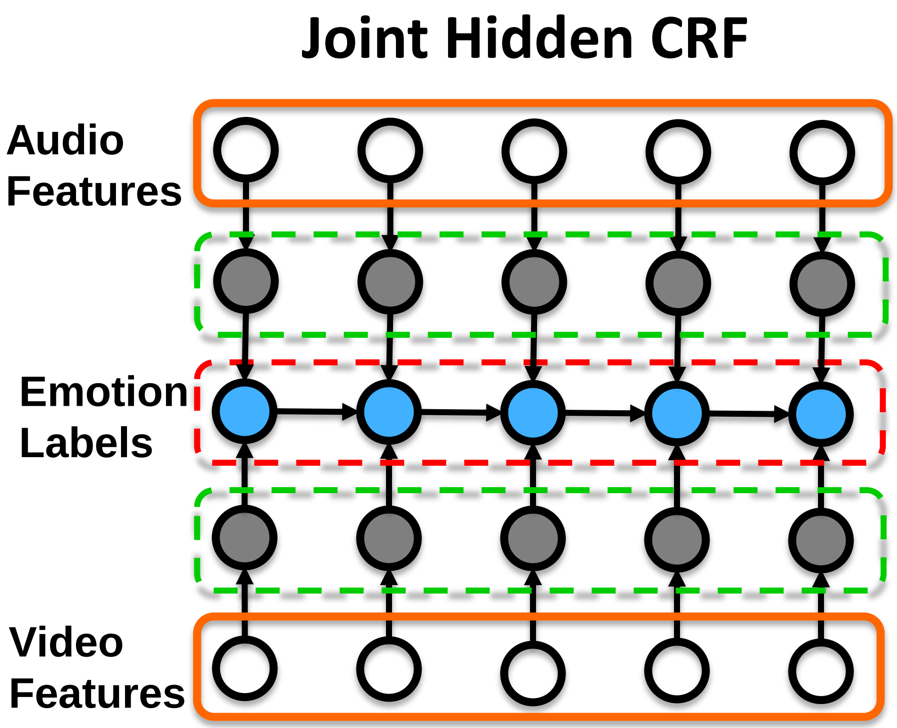
- **Audio-Visual Affect Recognition**
  - Early Fusion
    - Fuse inputs (features)
  - Late Fusion
    - Fuse outputs (decision values)





Our Model – Joint Hidden CRFs (JHCRFs)

- **Audio-Visual Affect Recognition**
  - Joint Hidden Conditional Random Fields
    - Information fused in a joint manner

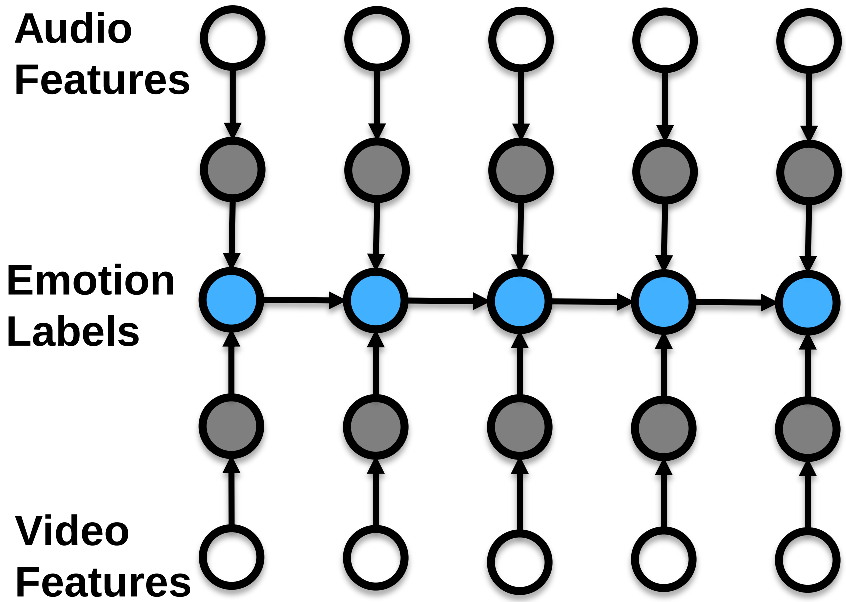


Our Model – Joint Hidden CRFs (JHCRFs)

- **Audio-Visual Affect Recognition**
  - Joint Hidden Conditional Random Fields
    - Information fused in a joint manner

$$p(W|X,\theta) = \frac{1}{Z(X,\theta)} \sum_H \exp(\Psi(X,H,W;\theta))$$

Joint Hidden CRF



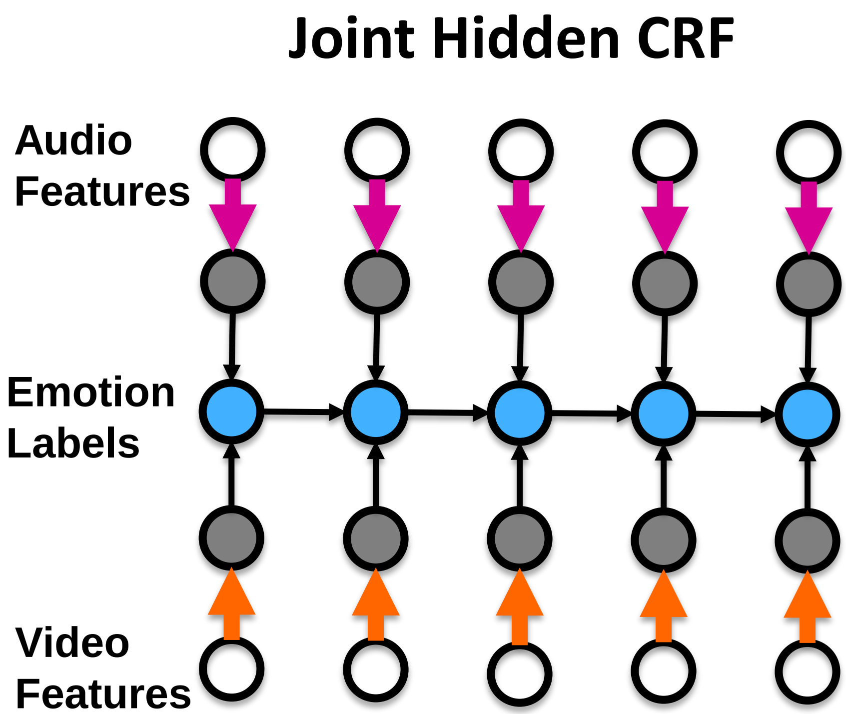
$$\Psi(X,H,W;\theta) = \sum_j \theta_i^{t^1} T_j^1(w_{i-1}, w_i, X, Y, i)$$
$$+ \sum_j \theta_j^{t^2} T_j^2(h_i^x, w_i, X, i) + \sum_j \theta_j^{t^3} T_j^3(h_i^y, w_i, Y, i)$$
$$+ \sum_k \theta_k^{s^1} S_k^1(h_i^x, X, i) + \sum_k \theta_k^{s^2} S_k^2(h_i^y, Y, i)$$

Our Model – Joint Hidden CRFs (JHCRFs)

- **Audio-Visual Affect Recognition**
  - Joint Hidden Conditional Random Fields
    - Information fused in a joint manner

$$p(W|X,\theta) = \frac{1}{Z(X,\theta)} \sum_H \exp(\Psi(X,H,W;\theta))$$

$$\Psi(X,H,W;\theta) = \sum_j \theta_i^{t^1} T_j^1(w_{i-1}, w_i, X, Y, i) + \sum_j \theta_j^{t^2} T_j^2(h_i^x, w_i, X, i) + \sum_j \theta_j^{t^3} T_j^3(h_i^y, w_i, Y, i) + \boxed{\sum_k \theta_k^{s^1} S_k^1(h_i^x, X, i)} + \boxed{\sum_k \theta_k^{s^2} S_k^2(h_i^y, Y, i)}$$

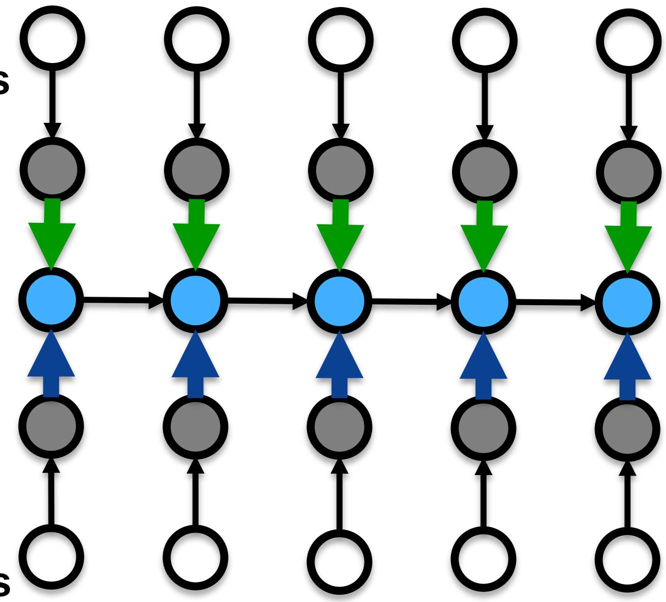


Our Model – Joint Hidden CRFs (JHCRFs)

- **Audio-Visual Affect Recognition**
  - Joint Hidden Conditional Random Fields
    - Information fused in a joint manner

$$p(W|X,\theta) = \frac{1}{Z(X,\theta)} \sum_{H} \exp(\Psi(X,H,W;\theta))$$

Joint Hidden CRF

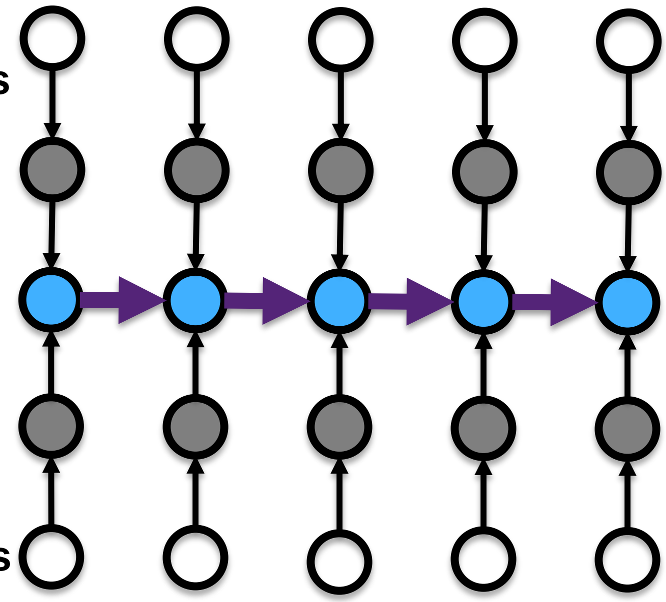


$$\Psi(X,H,W;\theta) = \sum_j \theta_i^{t^1} T_j^1(w_{i-1}, w_i, X, Y, i)$$
$$+ \boxed{\sum_j \theta_j^{t^2} T_j^2(h_i^x, w_i, X, i)} + \boxed{\sum_j \theta_j^{t^3} T_j^3(h_i^y, w_i, Y, i)}$$
$$+ \sum_k \theta_k^{s^1} S_k^1(h_i^x, X, i) + \sum_k \theta_k^{s^2} S_k^2(h_i^y, Y, i)$$

Our Model – Joint Hidden CRFs (JHCRFs)

- **Audio-Visual Affect Recognition**
  - Joint Hidden Conditional Random Fields
    - Information fused in a joint manner

Joint Hidden CRF



$$p(W|X,\theta) = \frac{1}{Z(X,\theta)} \sum_{H} \exp(\Psi(X,H,W;\theta))$$

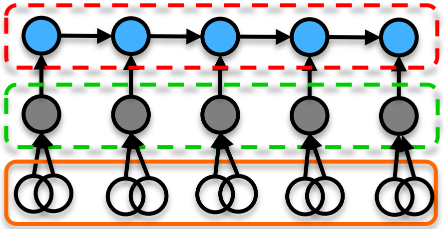
$$\Psi(X,H,W;\theta) = \boxed{\sum_j \theta_i^{t^1} T_j^1(w_{i-1}, w_i, X, Y, i)}$$
$$+ \sum_j \theta_j^{t^2} T_j^2(h_i^x, w_i, X, i) + \sum_j \theta_j^{t^3} T_j^3(h_i^y, w_i, Y, i)$$
$$+ \sum_k \theta_k^{s^1} S_k^1(h_i^x, X, i) + \sum_k \theta_k^{s^2} S_k^2(h_i^y, Y, i)$$

Comparing Performance

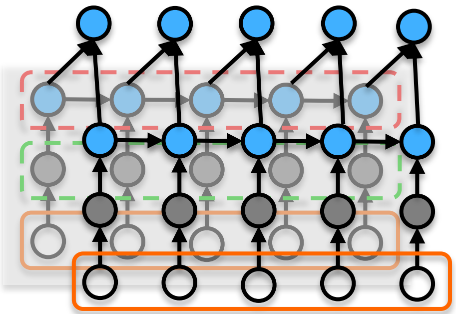
- **Audio Visual Emotion Recognition**
  - Classifier
    - Late Fusion
    - Early Fusion
  - JHCRF – To appear at ICME 2013 – Best Reported Results

	A	E	P	V	mean
Audio-SVM	64.6	66.6	66.2	61.9	**64.81**
Video-SVM	68.1	57.3	55.4	68.9	**62.43**
AudioVisual-SVM	67.5	65.8	65.8	70.4	**67.37**
Audio-HCRF	73.4	65.5	68.7	70.0	**69.42**
Video-HCRF	69.5	59.1	55.3	68.8	**63.17**
AudioVisual-JHCRF	75.7	66.3	69.1	76.3	**71.85**

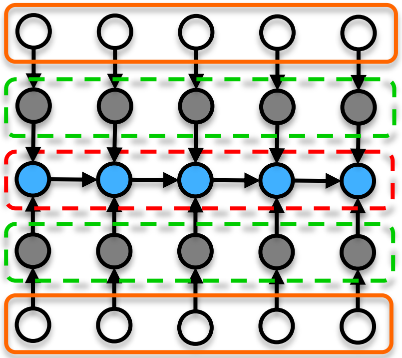
Early Fusion



Late Fusion



jHCRF

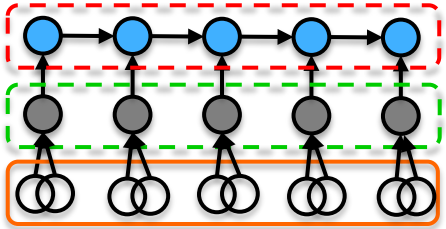


Comparing Performance

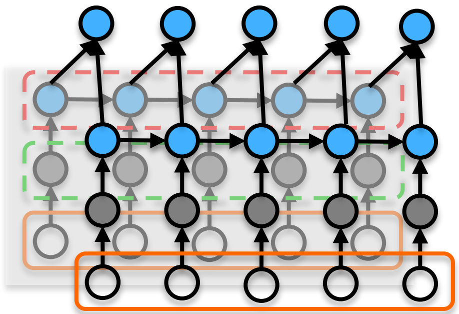
- **Audio Visual Emotion Recognition**
  - Classifier
    - Late Fusion
    - Early Fusion
  - JHCRF – To appear at ICME 2013 – Best Reported Results

	A	E	P	V	mean
Audio-SVM	64.6	66.6	66.2	61.9	**64.81**
Video-SVM	68.1	57.3	55.4	68.9	**62.43**
AudioVisual-SVM	67.5	65.8	65.8	70.4	**67.37**
Audio-HCRF	73.4	65.5	68.7	70.0	**69.42**
Video-HCRF	69.5	59.1	55.3	68.8	**63.17**
AudioVisual-JHCRF	75.7	66.3	69.1	76.3	**71.85**

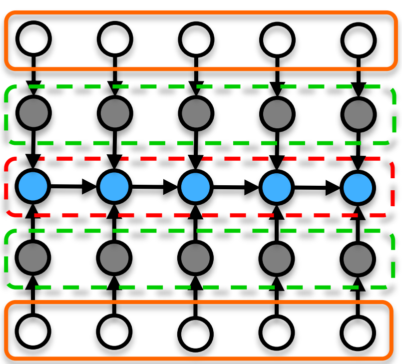
Early Fusion



Late Fusion



jHCRF

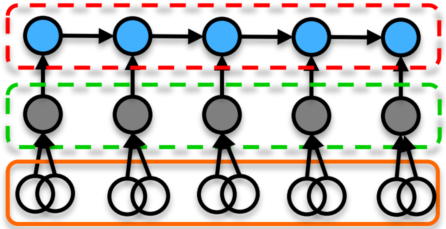


Comparing Performance

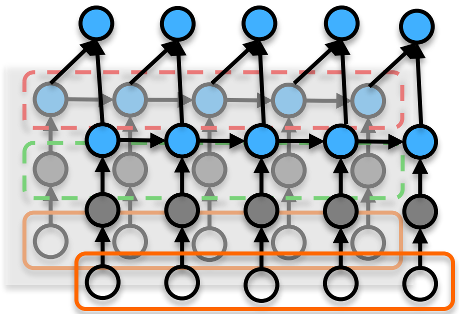
- **Audio Visual Emotion Recognition**
  - Classifier
    - Late Fusion
    - Early Fusion
  - JHCRF

	A	E	P	V	mean
HCRF (Early Fusion)	70.8	57.6	66.2	74.6	**67.29**
HCRF (Late Fusion)	70.5	66.5	65.6	77.1	**69.90**
JHCRF	75.7	66.3	69.1	76.3	**71.85**

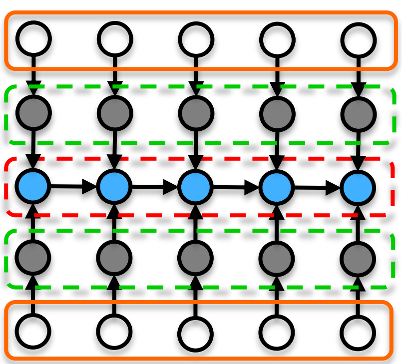
Early Fusion



Late Fusion



JHCRF

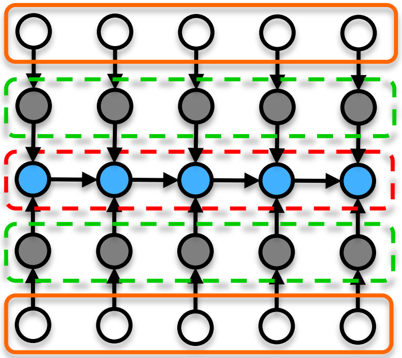


Deep Learning

- **Discriminative Model**

$$\underbrace{p(\mathbf{y}_t|\mathbf{v}_t,\mathbf{h}_t)}_{\text{Discriminative}}$$

JHCRF



Deep Learning

- **Generative Model**

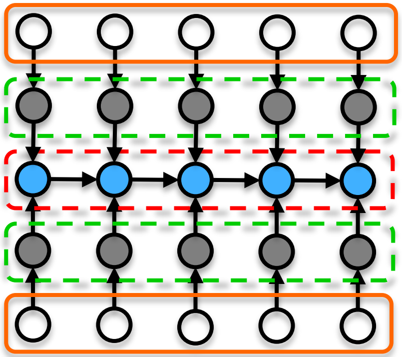
$$\underbrace{p(\mathbf{v}_t, \mathbf{h}_t | \mathbf{v}_{<t})}_{\text{Generative}}$$

Deep Learning

- **Hybrid Model**

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

JHCRF

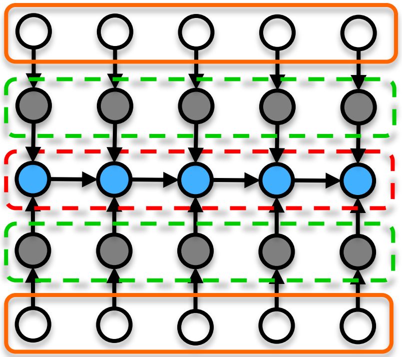


Deep Learning

- **Hybrid Model**

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

JHCRF

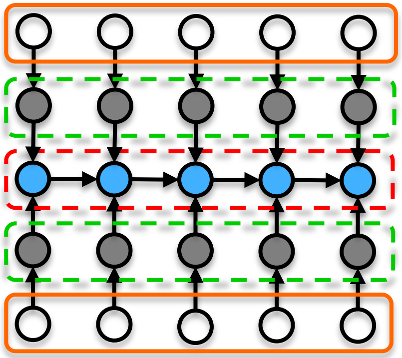


Deep Learning

- **Hybrid Model**

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

JHCRF



Deep Learning

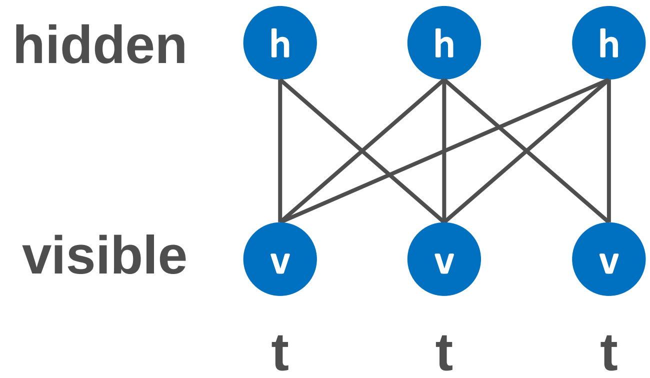
- **Restricted Boltzmann Machines**

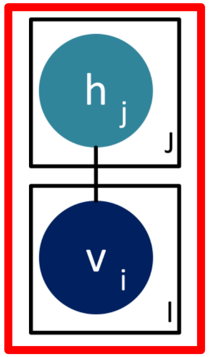
$$p(h_j = 1|\boldsymbol{v}) = f(b_j + \sum_i v_i w_{ij}),$$

$$p(v_i|\boldsymbol{h}) = \mathcal{N}(c_i + \sum_j h_j w_{ij}, 1),$$

$$\Delta w_{ij} \propto \langle v_i h_j \rangle_{\text{data}} - \langle v_i h_j \rangle_{\text{recon}}$$

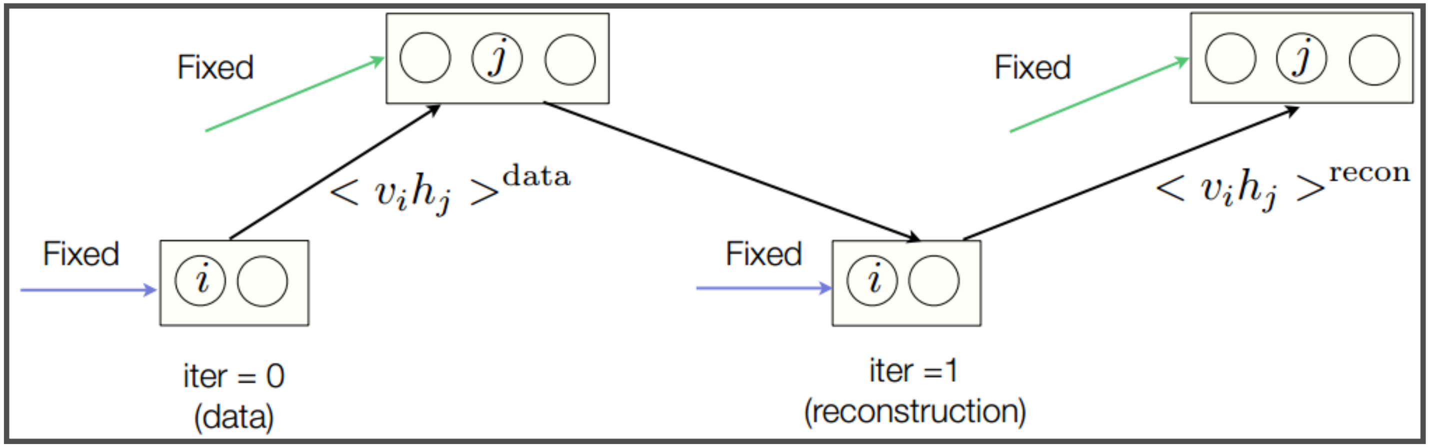
$$-\log p(\boldsymbol{v}, \boldsymbol{h}) = \sum_i \frac{(v_i - c_i)^2}{2\sigma_i^2} - \sum_j b_j h_j - \sum_{i,j} \frac{v_i}{\sigma_i} h_j w_{ij} + \text{const},$$





Deep Learning

- **Restricted Boltzmann Machines**



Graham Taylor

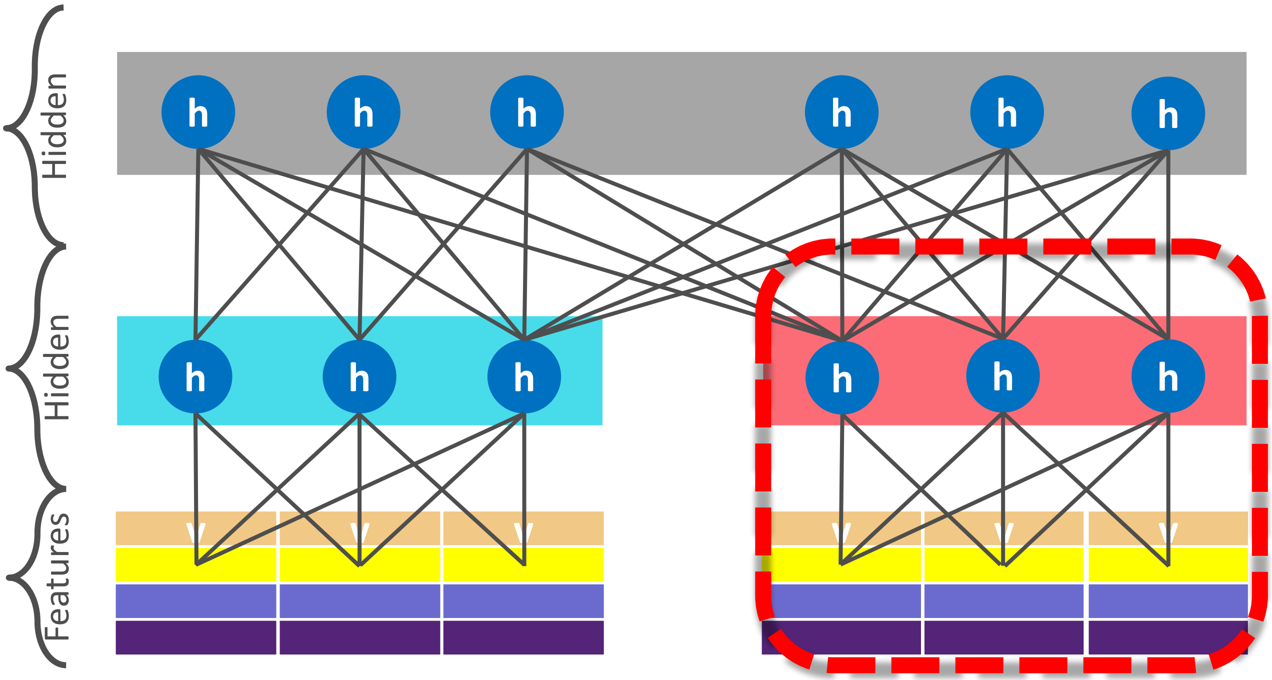
$$p(h_j = 1|\boldsymbol{v}) = f(b_j + \sum_i v_i w_{ij}),$$

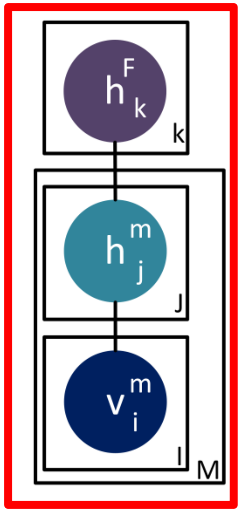
$$p(v_i|\boldsymbol{h}) = \mathcal{N}(c_i + \sum_j h_j w_{ij}, 1),$$

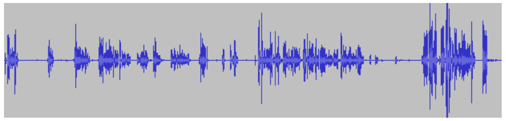
$$\Delta w_{ij} \propto \langle v_i h_j \rangle_{\mathrm{data}} - \langle v_i h_j \rangle_{\mathrm{recon}}$$

$$-\log p(\boldsymbol{v}, \boldsymbol{h}) = \sum_i \frac{(v_i - c_i)^2}{2\sigma_i^2} - \sum_j b_j h_j - \sum_{i,j} \frac{v_i}{\sigma_i} h_j w_{ij} + \mathrm{const},$$

Individual Affect Modeling:
Multimodal Generative (DBM)



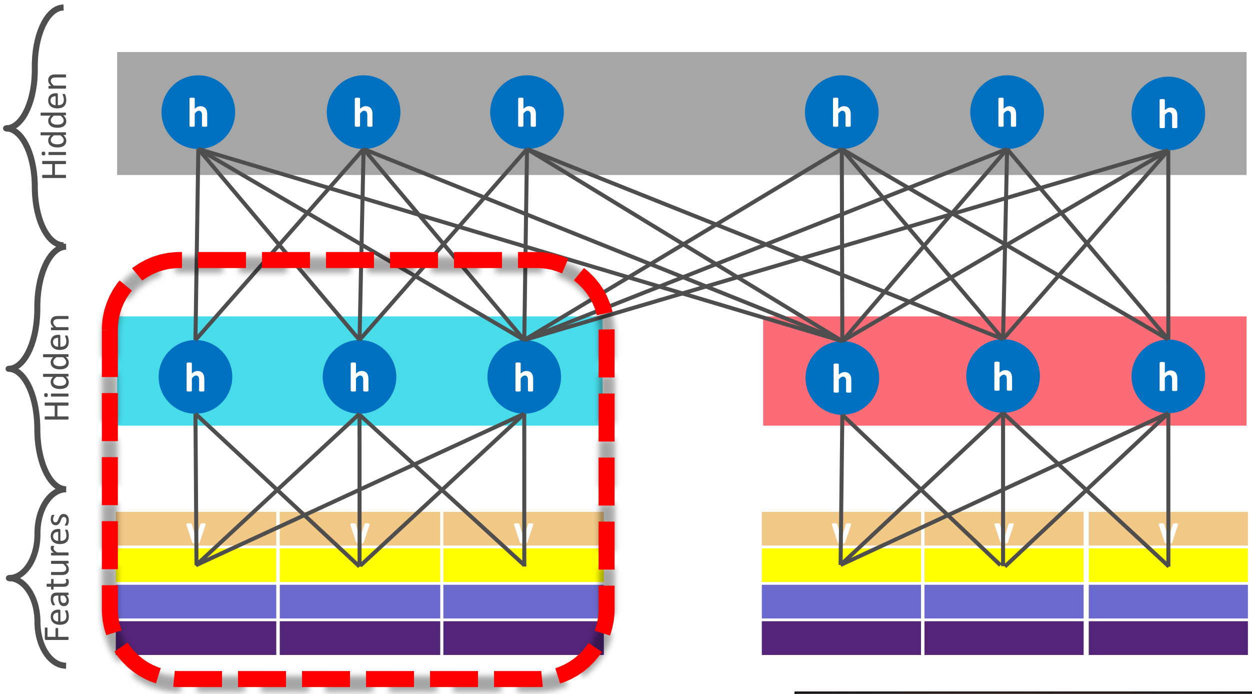


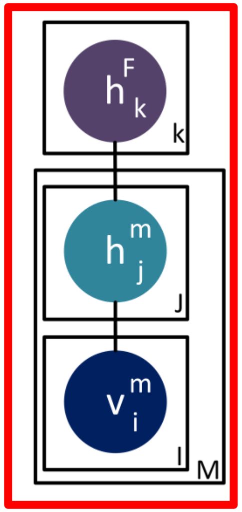


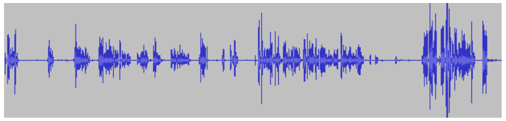
(tags/image) N. Srivastava ICML2012

(spectrogram/frame) J. Ngiam et al. NIPS2010

Individual Affect Modeling:
Multimodal Generative (DBM)



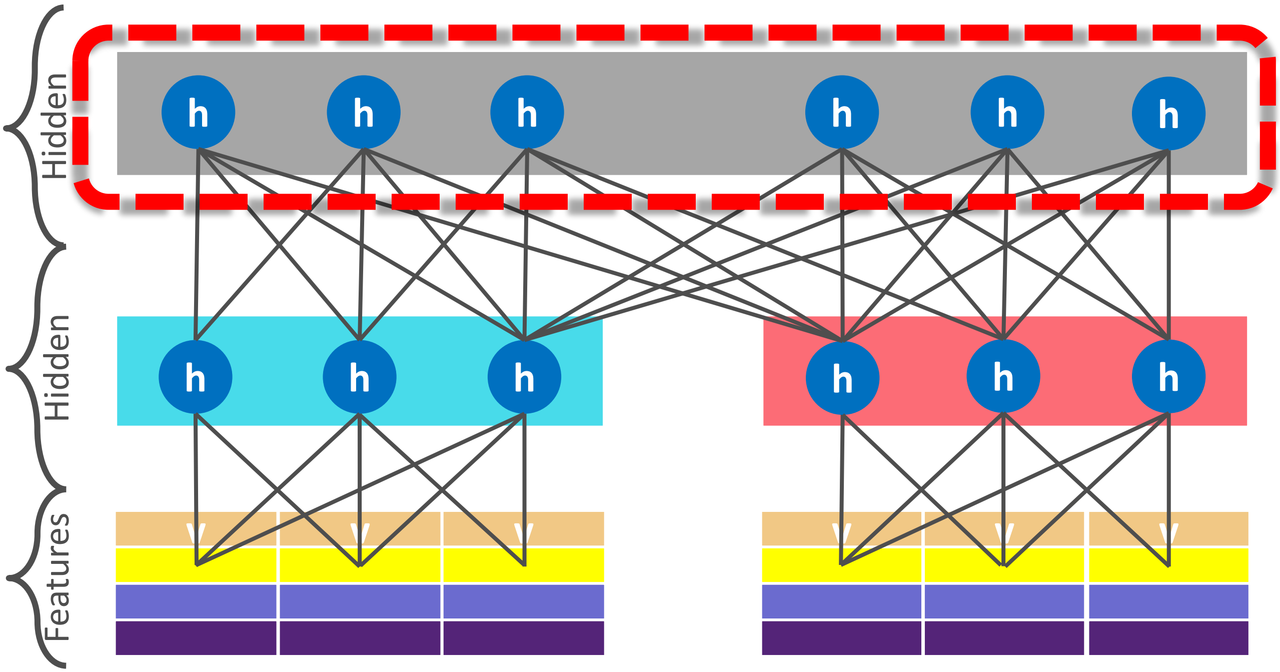


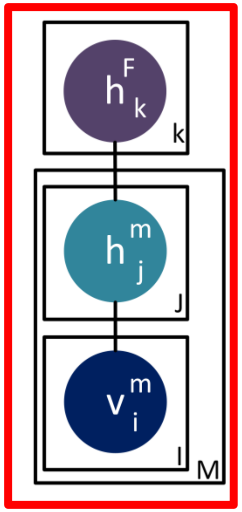


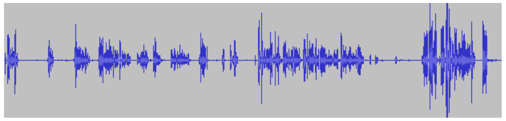
(tags/image) N. Srivastava ICML2012

(spectrogram/frame) J. Ngiam et al. NIPS2010

Individual Affect Modeling:
Multimodal Generative (DBM)





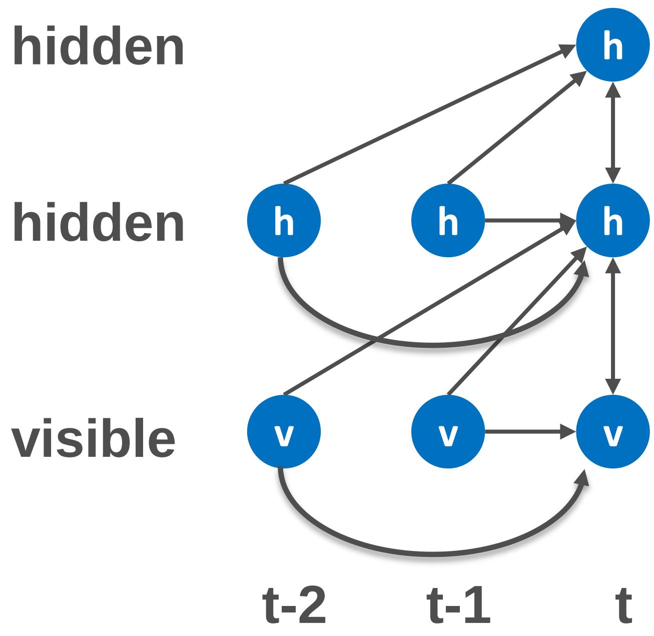


(tags/image) N. Srivastava ICML2012

(spectrogram/frame) J. Ngiam et al. NIPS2010

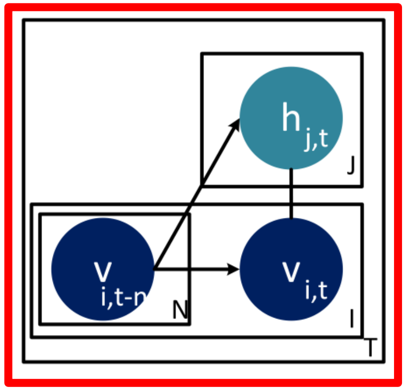
Deep Learning

- Temporal Deep Boltzmann Machines were first introduced by (G. Taylor, G. Hinton, S. Roweis NIPS2007)
- Each visible node has auto-regressive relations from previous time instances.
- Hidden nodes have both
  - Auto regressive connections from past frames hidden layers.
  - Connection from the past frames visible layers.



$$\Delta d_{ij}^{(t-q)} \propto v_i^{t-q} \left(\langle h_j^t \rangle_{\text{data}} - \langle h_j^t \rangle_{\text{recon}} \right)$$

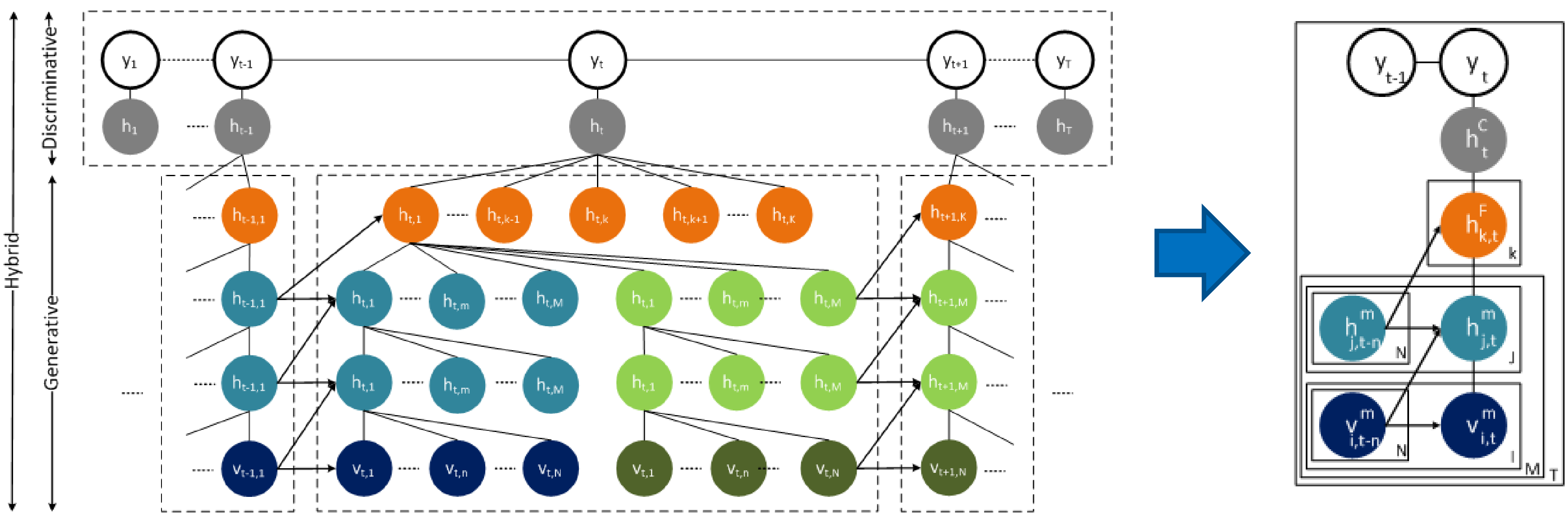
$$\Delta a_{ki}^{(t-q)} \propto v_k^{t-q} \left(v_i^t - \langle v_i^t \rangle_{\text{recon}} \right)$$



Deep Learning

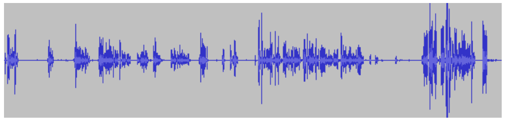
- **Hybrid Model**

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

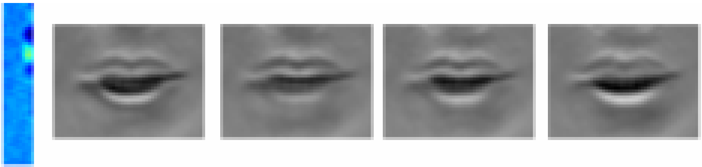


Deep Learning

- **AVEC** is an audio-visual dataset for single person affect analysis. The dataset consists of 31 sequences for training and 32 sequences for testing. The dataset provides pre-extracted set of features (MFCC/LBP).



- **AVLetters** consists of 10 speakers uttering the letters A to Z, three times each. The dataset is divided into two sets, 2/3 of the sequences for training and 1/3 for testing. The dataset provides pre-extracted 60x80 patches of lip regions along with audio features (MFCC features of 483 dimensions).



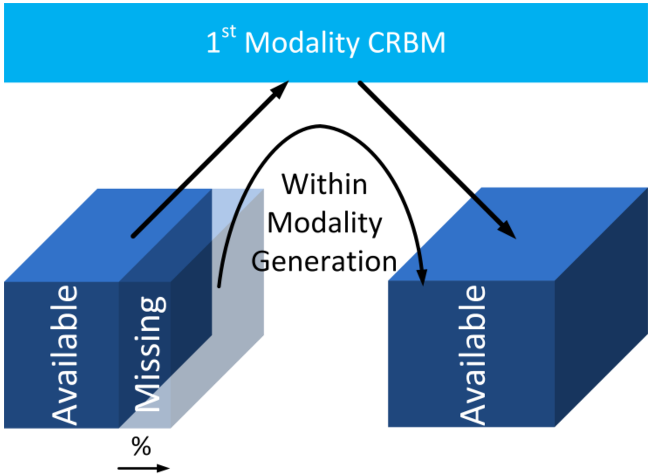
Deep Learning

Model/Dataset	AVE-A	AVE-V	AVE-AV	AVL-A	AVL-V	AVL-AV
SVM-RAW	64.8	62.4	67.4	55.8	46.2	58.5
CRF-RAW	68.2	63.2	69.9	57.0	**52.3**	58.8
HCRF-RAW	**69.4**	**63.5**	**69.9**	**58.4**	51.9	**60.0**
SVM-RBM	63.2	66.6	**67.6**	58.4	64.4	59.2
CRF-RBM	64.6	66.5	66.4	61.8	**64.6**	59.6
HCRF-RBM	**67.2**	**66.8**	67.2	**62.6**	61.5	**60.8**
SVM-CRBM	65.8	66.9	68.2	61.2	62.6	**63.8**
CRF-CRBM	67.7	68.2	69.1	**63.4**	**63.8**	63.1
HCRF-CRBM	**68.8**	**69.1**	**69.5**	63.1	61.8	61.9

Deep Learning

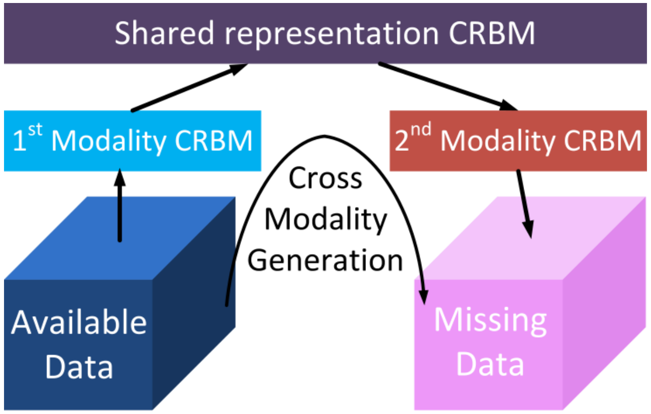
- Within Modality

Model/Dataset	AVE-A	AVE-V	AVL-A	AVL-V
SVM-RBM (0%)	63.2	66.6	58.4	64.4
SVM-CRBM (0%)	65.8	66.9	61.2	62.6
SVM-RBM (10%)	48.6	46.5	50.7	54.5
SVM-CRBM (10%)	54.9	52.1	53.6	48.2
SVM-RBM (30%)	35.5	31.2	39.2	32.1
SVM-CRBM (30%)	42.7	40.2	45.8	41.6



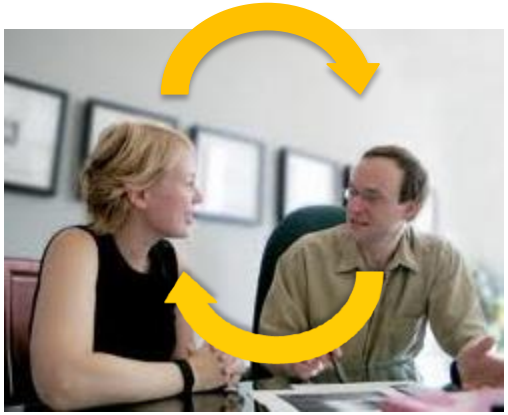
- Cross Modality

Model/Dataset	AVE-A||V	AVE-V||A	AVL-A||V	AVL-V||A
SVM-RBM	31.2	28.2	27.3	25.1
SVM-CRBM	40.4	32.1	29.6	26.5



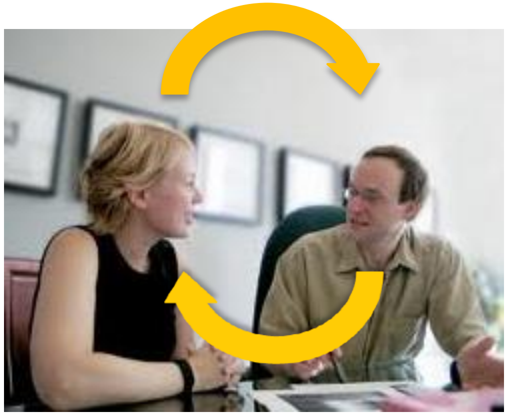
Multiple Person Affect Modeling

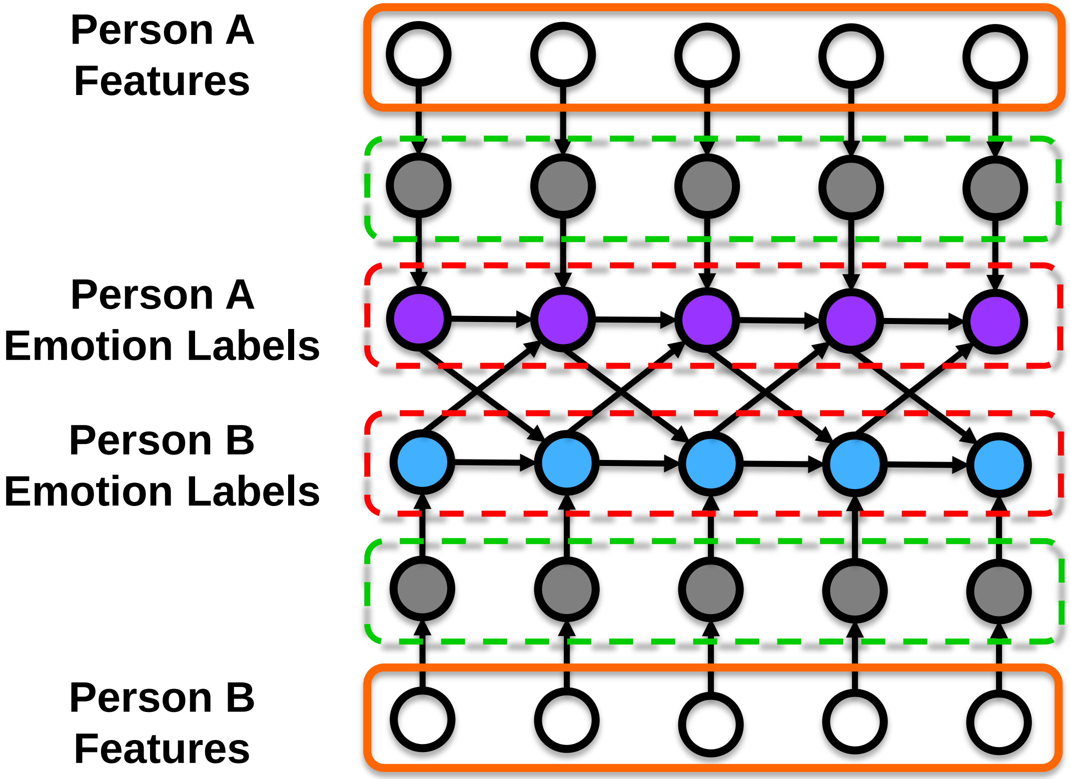
- **Interaction Dynamics**



Multiple Person Affect Modeling

- **Interaction Dynamics**





Where do we go from here?

- Still a ways to go before we can sense human behavior as well as other humans do
  - Micro expressions
  - Free flow gestures in-situ
  - Subtle variations in speech tone, inflections, emphasis
  - Cognitive and psychological states
- Simulations that people believe in
  - Blur the line between what's real and what's virtual
  - Augmented Reality
  - Mirroring behavior
- Evaluation and human factors understanding is key
  - Critical to understand the impact of pedagogy
  - Customization for individuals

Thank you!

- **Q?**