Modeling Human Communication Dynamics
Louis-Philippe Morency

USC Multimodal Communication and Machine Learning Lab [MultiComp Lab]
PhD students: Derya Ozkan and Sunghyun Park
Master student: Moitreya Chatterjee
Post-doctoral researcher: Stefan Scherer
Research programmer: Giota Stratou
Project manager: Alesia Egan
Multimodal Communication and Machine Learning Lab

© Keith Schaffer
Multimodal Perception of User State: Applications

Medical

- Distress Indicators
  - with MIT and CognitoHealth

- Suicide prevention
  - with Cincinnati Hospital

Education

- Group learning analytics
  - with Stanford and UCSD

- Virtual Learning Peer
  - with CMU

Business

- Negotiation outcomes
  - with USC business school

- YouTube: Opinion mining
  - with UNT and UT Dallas

Disorders

- Depression
- PTSD
- Distress

Social

- Engagement
- Dominance
- Empathy

Affect

- Frustration
- Agreement
- Sentiment
Multimodal Perception of Distress Indicators
Multimodal Communicative Behaviors

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

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

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
## From Audio-Visual Signals to Perceived User State

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

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

### 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

### Disorders
- Depression
- PTSD
- Distress

### Social
- Engagement
- Dominance
- Empathy

### Affect
- Frustration
- Agreement
- Sentiment

---

USC Institute for Creative Technologies

University of Southern California
From Audio-Visual Signals to Perceived User State

**Audio signals**
- Voice pitch

**Visual signals**
- Head pose

**Perceived User State**
- Distress
- Engagement
- Sentiment

---

**Sensing**

**Understanding**
From Audio-Visual Signals to Perceived User State

Audio signals
- Voice pitch

Visual signals
- Head pose

Sensing

KINECT
- Low-cost depth sensor
- Articulated body tracking

GAVAM
[FG 2008, best paper award]

CLM-Z
[CVPR 2012]
- 3D head pose estimation
- Real-time facial feature tracker

Audio Toolbox
- Pitch, energy and speaking rate
- Automatic voice quality analysis

Audio signals
- Voice pitch

Visual signals
- Head pose
From Audio-Visual Signals to Perceived User State

**Audio signals**
- Voice pitch

**Visual signals**
- Head pose

**Perceived User State**
- Distress
- Engagement
- Sentiment

**Sensing**

**Behaviors**

**Multimodal**
- Audio
- Visual
- Verbal

**Interaction**

**Cultural**

**Human Communication Dynamics**
Computational Behavior Indicators

- Human in the loop:
  - Identify behaviors useful to human task
- Behavior quantification
  - Quantify changes in human behaviors
Visualization of Computational Behavior Indicators

- Transparent comparisons of observed behavioral indicators

- Analogous to medical lab result sheets
Detection and Computational Analysis of Psychological Signals (DCAPS)

Albert (Skip) Rizzo
Clinical Expert

Louis-Philippe Morency
Multimodal Perception

Jonathan Gratch
Evaluation

Arno Hartholt
Integration

Stacy Marsella
Animation

David Traum
Dialogue

+ 27 researchers, programmers, artists and clinicians

cogito
Visualization & Audio Analysis
Psychological Distress: Datasets

Aims:

- Study behaviors associated to psychological distress
  - Depression (PHQ-9)
  - Trait Anxiety (STAI)
  - PTSD (PCL-C/PCL-M)
- Examine how these cues may differ
  - Across different interaction settings
    - Face-to-face: expected to evoke strongest indicators
    - Computer-mediated (TeleCoach): intended use case
    - Human-computer (SimSensei): intended use case
  - Across different populations
    - General Los Angeles population (Craigslist)
    - Recent veterans (US Vets)
Multimodal Perception of Distress Indicators
Psychological Distress Indicators

- Distress
- Anxiety
- Depression
- PTSD

[IEEE FG 2013 – Best paper award]
Effect of Gender on Distress Indicators

- Distress
- Anxiety
- Depression
- PTSD

[ACII 2013]
Effect of Gender on Distress Indicators

- Distress
- Anxiety
- Depression
- PTSD

A) AU4(frown) Intensity

B) Disgust Intensity

[ACII 2013]
Suicide Prevention

- Nonverbal indicators of suicidal ideations
- Dataset: 30 suicidal adolescents/30 non-suicidal adolescents
- Suicidal teenagers use more breathy tones

Source: CDC

[ICASSP 2013]
MultiSense: Multimodal Perception Library

TOOLS\MODULES

TRANSFORMERS:
- Facetrackers
- Gavam
- CLM/CLMZ
- Okao
- Shore
- Real-time hCRF
- ActiveMQ –VHMessenger
- EmoVoice

PROVIDERS:
- Audio Capture
- Webcam Capture
- Mouse
- Kinect (Depth/Intensity/IR image/Skeleton)

CONSUMERS:
- Image Painter
- Signal Painter

*Each one exported as .dll

PML:
Perception Markup language

Sensing Layer

```xml
<person id="subjectA">
  <sensingLayer>
    <headPose>
      <position z="223" y="345" x="193" />
      <rotation rotZ="15" rotY="35" rotX="10" />
      <confidence>0.34<confidence/>
    </headPose>
    ...
  </sensingLayer>
</person>
```

Behavior Layer

```xml
<person id="subjectB">
  <behaviorLayer>
    <behavior>
      <type>attention</type>
      <level>high</level>
      <value>0.6</value>
      <confidence>0.46<confidence/>
    </behavior>
    ...
  </behaviorLayer>
</person>
```
Temporal Probabilistic Learning

Multimodal

- Audio
- Visual
- Verbal

I. Temporal dynamic

Naïve Bayes Classifier

\[ y_1 \rightarrow y_2 \rightarrow y_3 \rightarrow \cdots \rightarrow y_n \]

\[ x_1 \rightarrow x_2 \rightarrow x_3 \rightarrow \cdots \rightarrow x_n \]

Hidden Markov Model

\[ h_1 \rightarrow h_2 \rightarrow h_3 \rightarrow \cdots \rightarrow h_n \]

\[ x_1 \rightarrow x_2 \rightarrow x_3 \rightarrow \cdots \rightarrow x_n \]

Maximum Entropy Model & Support Vector Machine

\[ y_1 \rightarrow y_2 \rightarrow y_3 \rightarrow \cdots \rightarrow y_n \]

\[ x_1 \rightarrow x_2 \rightarrow x_3 \rightarrow \cdots \rightarrow x_n \]

Conditional Random Field

\[ y_1 \rightarrow y_2 \rightarrow y_3 \rightarrow \cdots \rightarrow y_n \]

\[ x_1 \rightarrow x_2 \rightarrow x_3 \rightarrow \cdots \rightarrow x_n \]

Caption

- \( y_i \): Labels (e.g., head-nod, other-gesture)
- \( x_i \): Observations (e.g., yaw, roll, pitch)

- Generative models
- Discriminative models
Temporal Probabilistic Learning

Multimodal

- Audio
- Visual
- Verbal

I. Temporal dynamic
II. Hidden substructure

### Hidden Conditional Random Field

![Hidden CRF Diagram]

Hidden Conditional Random Field

[CVPR 2006, PAMI 2007]

### Latent-Dynamic Conditional Random Field

![Latent-Dynamic CRF Diagram]

Latent-Dynamic Conditional Random Field

[CVPR 2007]

### Model Accuracy

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMM</td>
<td>84.2%</td>
</tr>
<tr>
<td>CRF</td>
<td>86.0%</td>
</tr>
<tr>
<td>HCRF (w=0)</td>
<td>91.6%</td>
</tr>
<tr>
<td>HCRF (w=1)</td>
<td>93.9%</td>
</tr>
</tbody>
</table>

### True Positive Rate vs. False Positive Rate

![ROC Curve]

- LDCRF
- HMM
- SVM
- CRF

I. Temporal dynamic
II. Hidden substructure
Latent-Dynamic Conditional Neural Field

**Multimodal**

- Audio
- Visual
- Verbal

I. Temporal dynamic
II. Hidden substructure
III. Nonlinear input fusion

AVEC dataset
(95 videos)

Audio-visual sub-challenge

<table>
<thead>
<tr>
<th></th>
<th>Audio only</th>
<th>Visual only</th>
<th>Early fusion</th>
<th>Late fusion</th>
<th>LDCNF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy (%)</td>
<td>75</td>
<td>65</td>
<td>75</td>
<td>85</td>
<td>90</td>
</tr>
</tbody>
</table>

Rapport dataset: LDCRF, LDCNF
Taskar dataset: LDCRF, LDCNF

**Extended Abstract**

**Introduction**

Latent-Dynamic Conditional Neural Field (LDCNF) is a novel approach that integrates dynamic and conditional neural networks to model temporal and non-linear relationships in multimodal data.

**Temporal Dynamic**

- **I. Temporal dynamic**
  - Taskar dataset
  - Rapport dataset

**Hidden Substructure**

- **II. Hidden substructure**
  - Early fusion
  - Late fusion

**Nonlinear Input Fusion**

- **III. Nonlinear input fusion**
  - Audio-visual sub-challenge
  - AVEC dataset

**Conclusion**

LDCNF exhibits superior performance compared to state-of-the-art methods in both audio-visual and visual-only tasks, demonstrating its potential for improving multimodal interaction analysis.
Multi-View Hidden Conditional Random Field

**Multimodal**
- Audio
- Visual
- Verbal

I. Temporal dynamic
II. Hidden substructure
III. Nonlinear input fusion
IV. Multi-stream models

Canal 9 debate dataset

Multi-View HCRF

Multi-View LDCRF

MV-HCRF

Audio • Visual • Verbal

CVPR 2012
Multi-View Hidden Conditional Random Field

**Multimodal**
- Audio
- Visual
- Verbal

I. Temporal dynamic
II. Hidden substructure
III. Nonlinear input fusion
IV. Multi-stream models
V. Multimodal synchrony

Canal 9 debate dataset

![Diagram of Multi-View HCRF and Multi-View LDCRF models]

![Graph showing performance comparison between HMM, CRF, HCRF, MV-HCRF, and MV-HCRF + KCCA]
Multimodal Machine Learning

Multimodal

- Audio
- Visual
- Verbal

I. Temporal dynamic in label sequence
II. Hidden substructure in label sequence
III. Nonlinear modeling of instantaneous input features
IV. Multi-stream modeling of hidden substructure
V. Synchrony in multimodal input streams
VI. Multi-label structure and correlation
VII. Symbol and signal integration
VIII. Uncertainty in behavior labels

HCRF Library:
ICT open-source machine learning library
http://hcrf.sf.net
SimSensei Virtual Human Interaction Loop

Physical World

Perception
- Social Cues
- Affective Cues
- Physical Cues

Understanding

Cognition
- Decision-making
- Dialog
- Emotion

Action
- Social Cues
- Affective Cues
- Physical Cues

Virtual World
SimSensei Virtual Human Interaction Loop

**Physical World**

---

**MultiSense**

---

**Flores**
Dialogue manager

---

**Cerebella + Smartbody**

---

USC Institute for Creative Technologies

University of Southern California
MultiSense + SimSensei: Video Demonstration
MultiSense + SimSensei: Video Demonstration

USC Institute for Creative Technologies
University of Southern California

DCAPS: Detection & Computational Analysis of Psychological Signals

3rd Interim Progress Video

Albert (Skip) Rizzo, PI
Louis-Philippe Morency, Co-PI

The work depicted here was sponsored by the U.S. Defense Advanced Research Projects Agency. Statements and opinions expressed do not necessarily reflect the position of the policy of the United States Government, and no official endorsement should be inferred.
Collaborations and Available Technologies

- **MultiSense**: standardized perception framework
  - Real-time facial tracking, articulated body tracking and auditory analysis
  - Modular and multi-threaded architecture for easy extension
    - [http://multicomp.ict.usc.edu/](http://multicomp.ict.usc.edu/)

- **hCRF library, machine learning library**
  - Matlab and C++ implementations of LDCRF, HCRF and CRF.
  - 2559 downloads during the last year
    - [http://hcrf.sf.net/](http://hcrf.sf.net/)

- **GAVAM + CLM-Z, real-time nonverbal behavior recognition**
  - Real-time head position and orientation estimation
  - 66 facial feature tracking with automatic initialization
    - [http://multicomp.ict.usc.edu/](http://multicomp.ict.usc.edu/)
Thank you!