Health Behavior Informatics

Louis-Philippe (LP) Morency

CMU Multimodal Communication and Machine Learning Laboratory [MultiComp Lab]

PhD students: Chaitanya Ahuja, AmirAli BagherZadeh, Volkan Cirik, Sayan Ghosh, and Alexandria Vail
Master students: Torsten Wörtwein and Paul Liang
Postdoctoral researcher: Elif Bozkurt
Research programmer: Chirag Raman
Project manager: Laura Alford
MultiComp Lab

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

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

**Vocal**
- 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
A Central Challenge:
Modeling Human Communication Dynamics

- Behavioral
- Multimodal
- Interpersonal
- Societal

50 shades of “yeah”
Broad Applicability

Medical
- Depression and PTSD with MIT, BBN and Cogito
- Suicide prevention with Cincinnati Hospital
- Autistic children with Yale University

Online Education
- Group learning analytics with Stanford and UCSD
- Virtual Learning Peer with CMU
- Public speaking training with USC

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
Lecture Outline

Mental Health Behavior Indicators
- Health behavior infomatics
- Technologies for Healthcare Decision Support
- Behavioral Indicators of depression and PTSD
- Suicidal ideation behavior indicators

Multimodal Machine Learning
- Core challenges in Multimodal ML
- Learning multimodal representations
- Multimodal prediction of psychological distress
- SimSensei: virtual interviewer
Technologies for Health Behavior Informatics

OR

Patient

Clinician

SimSensei

MultiSense

Report
Behavioral Indicators of Psychological Distress

Distress Assessment Interview Corpus

<table>
<thead>
<tr>
<th>Not-distressed</th>
<th>Distressed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smile</td>
<td></td>
</tr>
<tr>
<td>Tense Voice</td>
<td></td>
</tr>
<tr>
<td>Open Posture</td>
<td></td>
</tr>
<tr>
<td>Emotional Expressiveness</td>
<td></td>
</tr>
</tbody>
</table>

Not-Distressed
Distressed

Carnegie Mellon University

Language Technologies Institute
Psychological Distress: Study protocol

Semi-structured interviews (~20 minutes)

- Three interview Phases
  - Rapport building
  - Intimate/clinical questions (positive and negative valence)
  - Cool-down phase
- Self-Assessment Clinical Scales
  - Depression (PHQ-9) and Post-traumatic stress disorder (PCL-C)
  - Highly correlated with accepted clinical diagnosis (sensitivity & specificity > .80)


## Demographics

<table>
<thead>
<tr>
<th>Demographic</th>
<th>Category</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Age</strong></td>
<td>Mean</td>
<td>47.93</td>
</tr>
<tr>
<td></td>
<td>Median</td>
<td>51</td>
</tr>
<tr>
<td><strong>Gender</strong></td>
<td>Male</td>
<td>65.64%</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>34.36%</td>
</tr>
<tr>
<td><strong>Education</strong></td>
<td>HS/GED</td>
<td>13.54%</td>
</tr>
<tr>
<td></td>
<td>Some College</td>
<td>37.91%</td>
</tr>
<tr>
<td></td>
<td>2 yr college</td>
<td>13.54%</td>
</tr>
<tr>
<td></td>
<td>4 yr college</td>
<td>26.71%</td>
</tr>
<tr>
<td></td>
<td>Post Graduate Degree</td>
<td>8.30%</td>
</tr>
<tr>
<td><strong>Race</strong></td>
<td>African American</td>
<td>36.38%</td>
</tr>
<tr>
<td></td>
<td>Asian</td>
<td>4.47%</td>
</tr>
<tr>
<td></td>
<td>White/Caucasian</td>
<td>43.77%</td>
</tr>
<tr>
<td></td>
<td>Hispanic</td>
<td>13.62%</td>
</tr>
<tr>
<td></td>
<td>Native American</td>
<td>1.75%</td>
</tr>
<tr>
<td><strong>Military</strong></td>
<td>44.32%</td>
<td></td>
</tr>
<tr>
<td><strong>Branch</strong></td>
<td>Army</td>
<td>44.31%</td>
</tr>
<tr>
<td></td>
<td>Navy</td>
<td>23.57%</td>
</tr>
<tr>
<td></td>
<td>Marine Corps</td>
<td>19.11%</td>
</tr>
<tr>
<td></td>
<td>Air Force</td>
<td>11.79%</td>
</tr>
<tr>
<td></td>
<td>Coast Guard</td>
<td>1.22%</td>
</tr>
</tbody>
</table>
Co-morbidity

Especially high correlations between clinical severity (0.80)

Psychological Distress Indicators

[IEEE FG 2013 - best paper award]

- **Joy – Facial expr.**
  - Distress No-distress

- **Sad – Facial expr.**
  - Distress No-distress

- **Vertical eye gaze**
  - Distress No-distress

- **Smile intensity**
  - Distress No-distress

- **Hand self-adaptor**
  - Distress No-distress

- **Legs fidgeting**
  - Distress No-distress

- **Voice energy std.**
  - Distress No-distress

- **Voice quality (NAQ)**
  - Distress No-distress

**DAIC**

Language Technologies Institute
Indicators with different trends on both genders [ACII 2013]

AU4 (frown) Intensity

- **MEN**: G = 0.5057
- **WOMEN**: G = -0.9156

Disgust Intensity

- **MEN**: G = 0.8407
- **WOMEN**: G = -1.2158

Cross over interaction!
Indicators with similar trend on both genders [ACII 2013]

**Head Rotation Variation**
- MEN: PTSD, G = -0.7426
-非PTSD, G = -0.586
- WOMEN: PTSD, G = -0.5664
-非PTSD, G = -0.6852

**Emotional Variation**

- Gender effect!
Behavior Indicators of Suicidal Ideation [2010]

Experiment

- Nonverbal indicators of suicidal ideations
- 60 subjects
  - 30 non-suicidal adolescents
  - 30 suicidal adolescents
    - 17 non-repeater
    - 13 repeater

Source: CDC [2010]
Behavior Indicators of Suicidal Ideation [2010]

### Ubiquitous questions

<table>
<thead>
<tr>
<th>ID</th>
<th>Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1</td>
<td>Does it hurt emotionally?</td>
</tr>
<tr>
<td>Q2</td>
<td>Do you have any fear?</td>
</tr>
<tr>
<td>Q3</td>
<td>Are you angry?</td>
</tr>
<tr>
<td>Q4</td>
<td>Do you have any secrets?</td>
</tr>
<tr>
<td>Q5</td>
<td>Do you have hope?</td>
</tr>
</tbody>
</table>

### Non-Ubiquitous questions

<table>
<thead>
<tr>
<th>ID</th>
<th>Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q6</td>
<td>Have you ever had anybody, either in your family or a friend that you know of that may have tried suicide or committed suicide?</td>
</tr>
<tr>
<td>Q7</td>
<td>How many hours in a day do you think you spend on the internet?</td>
</tr>
<tr>
<td>Q8</td>
<td>How close or connected do you feel to your family?</td>
</tr>
<tr>
<td>Q9</td>
<td>How close or connected do you feel to your peers?</td>
</tr>
<tr>
<td>Q10</td>
<td>Do you attend a religious service on a regular basis?</td>
</tr>
<tr>
<td>Q11</td>
<td>Do you have any past history of like physical abuse or neglect or anything like that?</td>
</tr>
<tr>
<td>Q12</td>
<td>Do you have any access to firearms?</td>
</tr>
<tr>
<td>Q13</td>
<td>On a school night how much do you think you sleep?</td>
</tr>
<tr>
<td>Q14</td>
<td>And what about on a weekend when you don’t have school?</td>
</tr>
<tr>
<td>Q15</td>
<td>Do you feel you currently have insomnia where you have problems sleeping?</td>
</tr>
<tr>
<td>Q16</td>
<td>Have there been any recent changes in your sleep habits?</td>
</tr>
</tbody>
</table>

The eleven open-ended questions which were asked in the second segment of each interview defining the non-UQ case.
Verbal Behavior Indicators [TAFFC 2016]

Use of first person singular pronouns

- Non-Suicidal
- Repeaters
- Non-Repeaters

Use of negative emotion terms

- Non-Suicidal
- Repeaters
- Non-Repeaters
Vocal Behavior Indicator: Voice Quality [TAFFC 2016]

[Box plots showing normalized amplitude quotient (NAQ) and quasi open quotient (QQQ) for Non-Suicidal, Repeaters, and Non-Repeaters groups.]
Vocal Behavior Indicator: Vowel Space [ICASSP 2015]

A

Formant 1 (Hz)

Formant 2 (Hz)

Vowel Space Ratio 0.348

Reference

Depressed Subject

B

Formant 1 (Hz)

Formant 2 (Hz)

Vowel Space Ratio 0.677

Reference

Non-Depressed Subject
Behavior Indicators with Schizophrenic Patients

- 18 semi-structured interviews (5-8 minutes per session)

<table>
<thead>
<tr>
<th>Q1</th>
<th>Q2</th>
<th>Q3</th>
<th>Q4</th>
<th>Q5</th>
<th>Q6</th>
<th>Q7</th>
<th>Q8</th>
<th>Q9</th>
<th>Q10</th>
<th>Q11</th>
<th>Q12</th>
<th>Q13</th>
<th>END</th>
</tr>
</thead>
<tbody>
<tr>
<td>What brought you into the hospital?</td>
<td>Has anything been on your mind lately?</td>
<td>What has the team been helping you with?</td>
<td>Would you say they are doing a good job?</td>
<td>What are your goals while you are here?</td>
<td>How have people been treating you?</td>
<td>How is the food?</td>
<td>How has your mood been/how have your spirits been?</td>
<td>How is your thinking?</td>
<td>How is your energy?</td>
<td>How is your sleep?</td>
<td>How is your self-confidence?</td>
<td>Are there any changes that you observe?</td>
<td></td>
</tr>
</tbody>
</table>

- Clinical scales:
  - Positive And Negative Syndrome Scale of Schizophrenia (PANSS)
  - Brief Psychiatric Rating Scale (BPRS)
  - The Montgomery-Asberg Depression Rating Scale (MADRS)
Behavior Indicators with Schizophrenic Patients

- Schizophrenia
  - Unusual thought symptom

Facial expressivity

- With doctor

Gestures when alone
Multimodal Machine Learning

We saw the yellow dog

Disorders
- Depression
- Distress
- Autism

Social
- Leadership
- Empathy
- Engagement

Emotion
- Sentiment
- Persuasion
- Frustration
Core Challenges in Multimodal Machine Learning

Representation
Alignment
Fusion
Translation
Co-Learning

Multimodal Machine Learning: A Survey and Taxonomy
By Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency

https://arxiv.org/abs/1705.09406

- 5 core challenges
- 37 taxonomic classes
- 253 referenced citations
Input Modalities

Language

Visual

Acoustic

Fusion
Prediction

Translation

Big dog on the beach

Co-Learning

Alignment

Representation

1 2

𝑡 1 2 3 𝑡 𝑛

𝑡 2 3 𝑡 𝑛
Multimodal Representations

Multimodal
• Audio
• Visual
• Verbal

Representation

Verbal
We saw the yellow dog

Visual

Acoustic

Common Representation
(Multimodal Space)
Multimodal Representations

"I like it!"

Joyful tone

"Wow!"

Tensed voice

Common Representation
(Multimodal Space)
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
Multimodal Vector Space Arithmetic

[Kiros et al., Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models, 2014]
Multimodal Sentiment Analysis

MOSI dataset (Zadeh et al, 2016)

- 2199 subjective video segments
- Sentiment intensity annotations
- 3 modalities: text, video, audio

Multimodal joint representation:

\[ h_m = f(W \cdot [h_x, h_y, h_z]) \]
Speaker’s behaviors | Sentiment Intensity
--- | ---
“This movie is sick” | ? → Ambiguous!
“This movie is fair” | + → Unimodal cues
Smile | +
Loud voice | ? → Ambiguous!

Unimodal

“This movie is sick” | + + → Resolves ambiguity (bimodal interaction)
Smile | + +
Frown | -
Loud voice | ? → Still Ambiguous!

Bimodal

“This movie is sick” | ++ ++ → Different trimodal interactions!
Smile | ++ ++
Loud voice | +

Trimodal

“This movie is sick” | ++ ++
Smile | ++ ++
Loud voice | +
“This movie is fair” | + +
Multimodal Tensor Fusion Network (TFN)

Explicitly models unimodal, bimodal and trimodal interactions!

[Zadeh, Jones and Morency, EMNLP 2017]
# Experimental Results – MOSI Dataset

## Multimodal Baseline Evaluation

<table>
<thead>
<tr>
<th>Baseline</th>
<th>Binary Acc(%)</th>
<th>Binary F1</th>
<th>5-class Acc(%)</th>
<th>5-class MAE</th>
<th>Regression r</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>50.2</td>
<td>48.7</td>
<td>23.9</td>
<td>1.88</td>
<td>-</td>
</tr>
<tr>
<td>C-MKL</td>
<td>73.1</td>
<td>75.2</td>
<td>35.3</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>SAL-CNN</td>
<td>73.0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>SVM-MD</td>
<td>71.6</td>
<td>72.3</td>
<td>32.0</td>
<td>1.10</td>
<td>0.53</td>
</tr>
<tr>
<td>RE</td>
<td>71.4</td>
<td>72.1</td>
<td>31.9</td>
<td>1.11</td>
<td>0.51</td>
</tr>
<tr>
<td><strong>TFN</strong></td>
<td><strong>77.1</strong></td>
<td><strong>77.9</strong></td>
<td><strong>42.0</strong></td>
<td><strong>0.87</strong></td>
<td><strong>0.70</strong></td>
</tr>
<tr>
<td>Human</td>
<td>85.7</td>
<td>87.5</td>
<td>53.9</td>
<td>0.71</td>
<td>0.82</td>
</tr>
</tbody>
</table>

Δ\(^{SOTA}\)  

|       | ↑ 4.0 | ↑ 2.7 | ↑ 6.7 | ↓ 0.23 | ↑ 0.17 |

## Baseline Evaluation

<table>
<thead>
<tr>
<th>Baseline</th>
<th>Binary Acc(%)</th>
<th>Binary F1</th>
<th>5-class Acc(%)</th>
<th>5-class MAE</th>
<th>Regression r</th>
</tr>
</thead>
<tbody>
<tr>
<td>TFN(_{language})</td>
<td>74.8</td>
<td>75.6</td>
<td>38.5</td>
<td>0.99</td>
<td>0.61</td>
</tr>
<tr>
<td>TFN(_{visual})</td>
<td>66.8</td>
<td>70.4</td>
<td>30.4</td>
<td>1.13</td>
<td>0.48</td>
</tr>
<tr>
<td>TFN(_{acoustic})</td>
<td>65.1</td>
<td>67.3</td>
<td>27.5</td>
<td>1.23</td>
<td>0.36</td>
</tr>
<tr>
<td>TFN(_{bimodal})</td>
<td>75.2</td>
<td>76.0</td>
<td>39.6</td>
<td>0.92</td>
<td>0.65</td>
</tr>
<tr>
<td>TFN(_{trimodal})</td>
<td>74.5</td>
<td>75.0</td>
<td>38.9</td>
<td>0.93</td>
<td>0.65</td>
</tr>
<tr>
<td>TFN(_{notrimodal})</td>
<td>75.3</td>
<td>76.2</td>
<td>39.7</td>
<td>0.919</td>
<td>0.66</td>
</tr>
<tr>
<td><strong>TFN</strong></td>
<td><strong>77.1</strong></td>
<td><strong>77.9</strong></td>
<td><strong>42.0</strong></td>
<td><strong>0.87</strong></td>
<td><strong>0.70</strong></td>
</tr>
<tr>
<td>TFN(_{early})</td>
<td>75.2</td>
<td>76.2</td>
<td>39.0</td>
<td>0.96</td>
<td>0.63</td>
</tr>
</tbody>
</table>

**Improvement over State-Of-The-Art**
# Taxonomy of Multimodal Research

<table>
<thead>
<tr>
<th><strong>Representation</strong></th>
<th><strong>Fusion</strong></th>
<th><strong>Alignment</strong></th>
<th><strong>Co-learning</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Joint</td>
<td>Model agnostic</td>
<td>Explicit</td>
<td>Model-based</td>
</tr>
<tr>
<td>Neural networks</td>
<td>Early fusion</td>
<td>Unsupervised</td>
<td>Grammar-based</td>
</tr>
<tr>
<td>Graphical models</td>
<td>Late fusion</td>
<td>Supervised</td>
<td>Encoder-decoder</td>
</tr>
<tr>
<td>Sequential</td>
<td>Hybrid fusion</td>
<td></td>
<td>Online prediction</td>
</tr>
<tr>
<td>Coordinated</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Similarity</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Structured</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Translation</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Example-based</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Retrieval</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Combination</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Neural networks</th>
<th>Model-based</th>
<th>Co-learning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grammar-based</td>
<td>Parallel data</td>
<td>Co-training</td>
</tr>
<tr>
<td>Encoder-decoder</td>
<td>Non-parallel data</td>
<td>Transfer learning</td>
</tr>
<tr>
<td>Online prediction</td>
<td>Hybrid data</td>
<td>Bridging</td>
</tr>
<tr>
<td>Graphical models</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Neural networks</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Co-learning</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Multimodal Machine Learning

Verbal

We saw the yellow dog

Visual

Multimodal Probabilistic Learning

Vocal

Disorders
- Depression
- Distress
- Autism

Social
- Leadership
- Empathy
- Engagement

Emotion
- Sentiment
- Persuasion
- Frustration
Automatic Distress Level Prediction
MultiSense Live Demonstration

Carnegie Mellon University

FPS: 24

Confidence: 92%

Behavior Indicators
- Depression
- Post-Traumatic
- Negative expressions (male)
- Negative expressions (female)
- Head rotation variation
- Schizophrenia
- Expressivity
- Suicidality
- Pitch variability

MultiSense
Powered by OpenFace
Powered by COVAREP
OpenFace: An Open-source Facial Behavior Analysis Toolkit

- State-of-the-art performance in all tasks
- Real time performance (without a GPU)
- Open Source (both training and testing)
- Cross-platform (Linux and Windows)

Tadas Baltrusaitis
baltrus@cs.cmu.edu
https://github.com/TadasBaltrusaitis/OpenFace

[WACV 2016]
Why are we creating an AI agent (SimSensei)? [AAMAS 2014]

- Compare responses when participants believe the avatar is controlled by a human or by an AI
  - Computer-framed (N=77)
  - Human-framed (N=77) interactions
Why are we creating an AI agent (SimSensei)?

[Fear of evaluation]

- Computer: 18
- Human: 19

[Impression Management]

- Computer: 52
- Human: 62

[Intensity of sadness displayed]

- Computer: 0.04
- Human: 0.14

[Note: * indicates a significant difference between the computer and human ratings.]

[AAMAS 2014]
Technologies for Health Behavior Informatics

Behavioral

Human Communication Dynamics

Multimodal

- Audio
- Visual
- Verbal

Interpersonal

• Audio
• Visual
• Verbal

SimSensei
Suicide prevention with Cincinnati Hospital

TeleCoach
Negotiation outcomes

Virtual rapport

Look-to-talk
Social influence

Group learning analytics with Stanford and UCSD

OVAT

Cicero

Mel
MERCI !