Behavioral Analytics in Intelligent Training Systems

Saad Khan
skhan002@ets.org

Educational Testing Service
Introduction

• Core competencies to match 21st century needs:
  - Reading, Writing and Arithmetic
  - Collaboration, Problem Solving, Interpersonal Skills

• These competencies are often hard to assess in the traditional sense:
  - *Process* used to arrive at a conclusion more important than the end product e.g. in problem solving
  - *Non-Cognitive* behaviors e.g. motivation, self control
  - *Emotion and affective states* influence interpersonal interactions
Need: Assessing 21st Century Competencies *in-vivo*

- Measuring 21st century competencies in real-life authentic situations:
  - Student-teacher interaction in classrooms
  - Collaborative learning and team interactions
  - Communication competency and interpersonal skills etc.
- MOOCs, Online learning: *Flipped Classrooms*
  - Anywhere, anytime learning
  - *Gamification* of education
Challenges

- Lots of low-level sensor data (truly big-data):
  - Videos, images
  - Audio
  - Simulation log files etc.
- Unclear how to link with high-level competencies e.g. collaboration or communication ability
  - Constructs associated with these competencies (evidence) may not be directly/readily observable
  - A human rater brings years of experience and innate ability to process sensory (audio/visual) data to make holistic assessments
    - Can we automate this process to some degree?
## Hierarchical Processing and Analysis

<table>
<thead>
<tr>
<th>Multi-modal data</th>
<th>High-level interpretations</th>
<th>Mid-level representations</th>
<th>Low-level features</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Video</strong></td>
<td><strong>Communication</strong></td>
<td><strong>Affect, Non-verbal Behavior</strong></td>
<td><strong>Facial Expressions, Gestures, Speech Prosody</strong></td>
</tr>
<tr>
<td><strong>3D</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Audio</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **SVM**, **HMM**, **JHCRF**
- **HOG**, **Euler Angles**, **MFCC**
Tactical and Non-Kinetic/Soft Skills Training Systems

- Automatically analyze trainee performance as measured against certain objective characteristics
- Performance metrics that require understanding behavior of individuals as well as groups

Tactical/Cultural Training

Khan et al. HCI 2013
Key Elements of System

1. Multitude of sensors to capture low-level data

2. Automated trainee location tracking

3. Behavior analysis using a hierarchical representation approach

4. Trainee performance evaluation

Khan et al. HCI 2013
Behavior Analysis Architecture

1. Action Detection and Recognition
   - Low-level actions e.g. move to objective, run for cover, shoot

2. Events / States Detection and Recognition
   - High-level events (complex activities) e.g. React to IED, Patrol

3. Performance Evaluation
   - Assessment with respect to TTP (Techniques Tactics and Procedures)

Knowledge Base
TTP, Activity Taxonomy, Ontology

Sensor Data
E.g. Position, Pose & Trigger Data

Exercise Database

Khan et al. HCI 2013
Social Simulation Training

- Interactive Game-Like Setup with Fluid Interactions
  - Real-time sensing of trainee behavior
  - Enable Real-time response of virtual characters

- Sensing of Learner/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

DARPA SSIM
Khan and Divakaran et al.
Using Wearable Sensors: Sociometric badge

- “Honest Signals” research by Alex (Sandy) Pentland and others

Microphone to extract speech features
- Enthusiasm, energy, turn-taking

Accelerometer to measure body movement
- Gross body movements, mirroring (coarse)

Bluetooth to detect relative positions
- Distribution and formations of group

Infra-red sensor to detect line of sight
- Encounters, orientation, focus of attention
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
  - Representation
    - Word Segmentation
    - Functionals over each feature
  - Dimensionality Reduction
    - Partial Least Squares
    - Supervised (Class Aware) Dimensionality Reduction

\[
\begin{align*}
[cov(t_i, u_i)]^2 &= \max_{|w_x|=1, |w_y|=1} [cov(Xw_x, Yw_y)]^2 \\
W_x &= \{w_{x1}, w_{x2}, \ldots, w_{xp}\} \\
W_y &= \{w_{y1}, w_{y2}, \ldots, w_{yp}\}
\end{align*}
\]
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

\[
\begin{align*}
\left[ \text{cov}(t_i, u_i) \right]^2 &= \max_{|w_{xi}|=1,|w_{yi}|=1} \left[ \text{cov}(Xw_{xi}, Yw_{yi}) \right]^2 \\
W_x &= \{w_{x1}, w_{x2}, \ldots, w_{xp}\} \quad W_y = \{w_{y1}, w_{y2}, \ldots, w_{yp}\}
\end{align*}
\]
Classifier Options

- **Audio based Affect Recognition**
  - Classifier
  - Static vs Dynamic Classifiers
Multi-modal Fusion – Traditional Options

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

**Early Fusion**

**Late Fusion**

Audio + Video Features

Affect Labels

Audio Features

Video Features

Affect Labels
Joint Hidden Conditional Random Fields (JHCRFs)

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

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

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

- **Audio Visual Emotion Recognition**
  - Classifier
    - Late Fusion
    - Early Fusion
  - JHCRF ICME 2013, Best Reported Results on AVEC dataset

### Classification Accuracy on AVEC dataset

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>E</th>
<th>P</th>
<th>V</th>
<th>mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Audio-SVM</td>
<td>64.6</td>
<td>66.6</td>
<td>66.2</td>
<td>61.9</td>
<td>64.81</td>
</tr>
<tr>
<td>Video-SVM</td>
<td>68.1</td>
<td>57.3</td>
<td>55.4</td>
<td>68.9</td>
<td>62.43</td>
</tr>
<tr>
<td>AudioVisual-JHCRF</td>
<td>75.7</td>
<td>66.3</td>
<td>69.1</td>
<td>76.3</td>
<td>71.85</td>
</tr>
</tbody>
</table>
Assessing Communication Ability

Facial Expression:
Smiling, Positive Affect

Head Pose:
Nods/Shakes

Posture:
Leaning forward

Gaze:
Averted, Not looking directly into camera

Speech Tone:
Calm, Engaged

Multi-Attribute Ranking Model

Score

+ Recounting Vector
Studying influence of non-cognitive skills in collaborative tasks

Tetralogue: Educational Simulation

Collaborators/Students

Tracking:
- Facial Expressions
- Gaze
- Head nods

Multimodal Data
Concluding Remarks

• Innovative, agile, decentralized systems
• Behavioral analytics an important element
• New tools for cognitive and non-cognitive skills assessment
  – Multi-modal sensing, interfaces
• At the confluence of a multitude of disciplines
  – Computer Vision, Machine Learning, AI
  – Sociology, Psychometrics, Cognitive Science
Thank You!

Questions?