Emotionally Intelligent HCI and Robotics

Zhengyou Zhang
Distinguished Scientist
Director, Tencent Robotics X
zhengyou@tencent.com
Evolution of Human-Computer Interaction
Virtual Receptionist
Example: Magic Mirror

Microsoft has developed a mirror that can read your emotions.

- Recognize and greet users
- Read their emotions
- Display the weather, time and other information.


https://github.com/MicrosoftEdge/magic-mirror-demo
Motivations

• Understanding human emotion is the holy grail for human-computer interaction
  • By combining it with speech and gesture understanding, computers will one day be able to communicate with human naturally

• Wide array of applications

Affect-aware personal assistant/companion devices
Autism intervention
Honest signal
Affect-aware game development
Emotion Intelligence

• EI is vital to make machine behave more like human

• Four branches of EI (Salovey and Mayer)
  • Perceiving emotions – understanding nonverbal signals
  • Reasoning with emotions – using emotions to prioritize attention
  • Understanding emotions – interpret the cause of emotion
  • Managing emotions – regulating and responding to emotions

• Computer EI is still in its infancy
Emotion API for MCS

(Microsoft Cognitive Services)

Applications
- Personal assistant
- Customer service
- Autism intervention
- Meeting dynamics
- Adaptive gaming
- Advertisement/marketing
- ...

Emotion APIs
- Image
- Video
- Depth
- Audio
- Text

Input

Other sensors (heart rate, gps, eda, etc.)
Basic Emotions

Neutral

Happiness

Surprise

Sadness

Angry

Contempt

Disgust

Fear
Facial Action Coding System (FACS)

<table>
<thead>
<tr>
<th>Upper Face Action Units</th>
<th>AU 1</th>
<th>AU 2</th>
<th>AU 4</th>
<th>AU 5</th>
<th>AU 6</th>
<th>AU 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inner Brow Raiser</td>
<td></td>
<td>Outer Brow Raiser</td>
<td>Brow Lowerer</td>
<td>Upper Lid Raiser</td>
<td>Cheek Raiser</td>
<td>Lid Tightener</td>
</tr>
<tr>
<td>*AU 41</td>
<td>*AU 42</td>
<td>*AU 43</td>
<td>AU 44</td>
<td>AU 45</td>
<td>AU 46</td>
<td></td>
</tr>
<tr>
<td>Lid Droop</td>
<td></td>
<td>Slit</td>
<td>Eyes Closed</td>
<td>Squint</td>
<td>Blink</td>
<td>Wink</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Lower Face Action Units</th>
<th>AU 9</th>
<th>AU 10</th>
<th>AU 11</th>
<th>AU 12</th>
<th>AU 13</th>
<th>AU 14</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nose Wrinkler</td>
<td></td>
<td>Upper Lip Raiser</td>
<td>Nasolabial Deepener</td>
<td>Lip Corner Puller</td>
<td>Cheek Puffer</td>
<td>Dimpler</td>
</tr>
<tr>
<td>AU 15</td>
<td>AU 16</td>
<td>AU 17</td>
<td>AU 18</td>
<td>AU 20</td>
<td>AU 22</td>
<td></td>
</tr>
<tr>
<td>Lip Corner Depressor</td>
<td></td>
<td>Lower Lip Depressor</td>
<td>Chin Raiser</td>
<td>Lip Puckerer</td>
<td>Lip Stretcher</td>
<td>Lip Funneler</td>
</tr>
<tr>
<td>AU 23</td>
<td>AU 24</td>
<td>*AU 25</td>
<td>*AU 26</td>
<td>*AU 27</td>
<td>AU 28</td>
<td></td>
</tr>
<tr>
<td>Lip Tightener</td>
<td></td>
<td>Lip Pressor</td>
<td>Lips Part</td>
<td>Jaw Drop</td>
<td>Mouth Stretch</td>
<td>Lip Suck</td>
</tr>
</tbody>
</table>
Mapping Between Basic Emotions & FACS

Neutral
1) Neutral Baseline
2) Expressionless face

Happiness
1) Muscles around the eyes tightened
2) “Crows feet” wrinkles around the eyes
3) Cheeks raised
4) Lip corners raised diagonally (smile)

Surprise
1) Eyebrow raised
2) Eyelids pulled up (eyes widened)
3) Mouth Open
Basic Emotion Types

Sadness
1) Inner corners of eyebrow raised
2) Eyelids loose
3) Lip corners pulled down (frown)

Anger
1) Eyebrows pulled down and together
2) Upper & lower lids pulled up (glaring)
3) Margin of lips pulled in
4) Lips may be tightened (optional)

Contempt
1) Eyes neutral
2) Lip corner raised and pulled back on only one side of the face
Basic Emotion Types

Disgust

1) Eyebrows pulled down
2) Nose wrinkled
3) Upper lip raised
4) Lips loose

Fear

1) Eyebrows pulled up and together
2) Upper eyelids pulled up
3) Mouth stretched
20 Years Ago ...

Dataset: 10 women; 213 images

Features: 34 points; Gabor wavelets
Gabor Wavelets

\[ \Psi(k, x) = \frac{k^2}{\sigma^2} \exp \left( -\frac{k^2 x^2}{2\sigma^2} \right) \left[ \exp(i k \cdot x) - \exp\left( -\frac{\sigma^2}{2} \right) \right] \]
Two-Layer Perceptron

outputs

hidden units

inputs

image

preprocessing

geometric positions

Gabor wavelet coefficients

bias

bias

$y_1, y_2, y_3, y_4, y_5, y_6, y_7$

$z_1, z_m, z_1', z_m'$

$x_1, x_d, x_1', x_d'$

$x_0$
Result

- 90.1% with 7 hidden units
- 79.5% agreement among 60 non experts
Examples

| Label: Surprise | NN outputs | | Label: Happiness | NN outputs | | Label: Anger | NN outputs |
|----------------|------------|-----------------|---------------|-----------------|-----------------|---------------|
|                | Neu. 0.000 | Hap. 0.000 | Sad. 0.000 | Sur. 1.000 | Ang. 0.000 | Dis. 0.000 | Fear 0.000 | Neu. 0.000 | Hap. 0.000 | Sad. 0.000 | Sur. 0.000 | Ang. 0.000 | Dis. 0.000 | Fear 0.000 |
|                | Neu. 0.122 | Hap. 0.720 | Sad. 0.000 | Sur. 0.000 | Ang. 0.000 | Dis. 0.000 | Fear 0.158 |

<table>
<thead>
<tr>
<th>Label: Disgust</th>
<th>NN outputs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neu. 0.001</td>
<td>Hap. 0.000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Label: Fear</th>
<th>NN outputs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neu. 0.002</td>
<td>Hap. 0.000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Label: Sadness</th>
<th>NN outputs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neu. 0.099</td>
<td>Hap. 0.000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Label: Happiness</th>
<th>NN outputs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neu. 0.822</td>
<td>Hap. 0.017</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Label: Fear</th>
<th>NN outputs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neu. 0.010</td>
<td>Hap. 0.000</td>
</tr>
</tbody>
</table>
Sensitivity Analysis

$$G^{(p)} = \left\| \frac{\partial E}{\partial x^{(p)}} \right\| = \sqrt{\sum_i \left( \frac{\partial E}{\partial x_i^{(p)}} \right)^2}$$

The importance is indicated by the size of a point
Data Reduction

Deleted 12 less informative points: 16, 17, 18, 19, 20, 21, 23, 24, 30, 13, 29, and 3.
Significance of Image Scales

$k=1$ (lowest res),

..., 

5 (highest res)
Now, emotions in the wild ...
Major Challenges

Input data is unrestricted
- A stress test for face detection/alignment
- Huge variation in pose, lighting, skin tone, etc.
- Few existing work on emotion recognition in the wild

People don’t always agree on the corresponding emotion
- A single face can exhibit multiple emotions
- For subtle emotion, still image doesn’t have enough information (no temporal info)

Most data available online are biased toward happy and neutral
Our Approach

• A deep learning based approach

• With various data augmentation
  • Affine transform, lighting augmentation, multi-cropping, voting, etc.
Start with FER 2013
   Web crawled + human labeling
   48x48 image resolution
   28709 training examples
   3589 validation examples
   3590 test examples
   Very noisy data

Train DCNN
   Without data augmentation
       65.07%
   With data augmentation
       71.73%
A Glimpse on Fear Category

Correct Prediction

Wrong Prediction

surprise  sad  angry  sad  happy  angry  sad  sad  happy  angry
happy  neutral  disgust  surprise  sad  happy  angry  surprise  sad  angry

happy  neutral  disgust  surprise  sad  happy  angry  surprise  sad  angry
Data Collection

Crawled ~4.5m images with emotional keywords
   166 emotional adjectives
   230 celebrity names, 100 popular first names, 166 people related words

Face detection
Active learning
   Use DCNN to select confusing facial images for tagging
Self-paced learning
   Use DCNN to expand training data based on classification results
   Randomly sampled
   Biased towards rare emotion types
Tagging: FACS vs. Basic Emotions

Use FACS (Facial Action Coding System)
- More accurate and less subjective.
- Easy expand to more emotions.
- Cons: Expensive and require a certified tagger.

Appearance based emotion
- Cheap and doesn’t require a certified tagger.
- Cons: Very noisy.
Crowd Source Tagging

Taggers are forced to choose one emotion out of 8, or tag the face image as “unknown”
We started with at least 2 taggers agree and up to 5 taggers.
  Quality was very bad specially with subtle emotions.
We retagged all our data with 10 taggers.
  Quality improved drastically (detailed next).
How many taggers do we need?

Diagram on percentage of agreement with final majority label
Old versus new label
Emotion Probability Distribution

Majority Voting (MV)
   Each face is associated with one emotion, the one that has the majority vote.

Multi-Label Learning (ML)
   All emotions above certain threshold are treated as valid emotion.

Probabilistic Drawing (PLD)
   During training draw the target emotion according to its probability.

Cross-entropy (CEL)
   Learn the actual probability distribution.
## Emotion Probability Distribution

### Training result

<table>
<thead>
<tr>
<th>Schemes</th>
<th>Trials 1</th>
<th>Trials 2</th>
<th>Trials 3</th>
<th>Trials 4</th>
<th>Trials 5</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>MV</td>
<td>83.60%</td>
<td>84.89%</td>
<td>83.15%</td>
<td>83.39%</td>
<td>84.23%</td>
<td>83.85±0.63%</td>
</tr>
<tr>
<td>ML</td>
<td>83.69%</td>
<td>83.63%</td>
<td>83.63%</td>
<td>84.62%</td>
<td>84.08%</td>
<td>83.97±0.36%</td>
</tr>
<tr>
<td>PLD</td>
<td>85.43%</td>
<td>84.65%</td>
<td>85.34%</td>
<td>85.01%</td>
<td>84.50%</td>
<td>84.99±0.37%</td>
</tr>
<tr>
<td>CEL</td>
<td>85.01%</td>
<td>84.59%</td>
<td>84.32%</td>
<td>84.80%</td>
<td>84.86%</td>
<td>84.72±0.24%</td>
</tr>
</tbody>
</table>
## Final Data Set FER + In-house

<table>
<thead>
<tr>
<th>Expression</th>
<th>Train</th>
<th>Valid</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neutral</td>
<td>55,180</td>
<td>1,151</td>
<td>4,396</td>
</tr>
<tr>
<td>Happiness</td>
<td>26,270</td>
<td>904</td>
<td>1,801</td>
</tr>
<tr>
<td>Surprise</td>
<td>15,421</td>
<td>422</td>
<td>725</td>
</tr>
<tr>
<td>Sad</td>
<td>11,221</td>
<td>418</td>
<td>308</td>
</tr>
<tr>
<td>Angry</td>
<td>14,063</td>
<td>305</td>
<td>843</td>
</tr>
<tr>
<td>Disgust</td>
<td>3,372</td>
<td>19</td>
<td>87</td>
</tr>
<tr>
<td>Fear</td>
<td>5,442</td>
<td>92</td>
<td>198</td>
</tr>
<tr>
<td>Contempt</td>
<td>5,329</td>
<td>24</td>
<td>26</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>136,298</td>
<td>3,335</td>
<td>8,384</td>
</tr>
</tbody>
</table>
Performance

- Nov 2015 release
  - 80.98% on test set
- Nov 2015 network on better tags
  - 83.66% on test set
- Our latest deeper network
  - 85.26% on test set

<table>
<thead>
<tr>
<th></th>
<th>Neu</th>
<th>Hap</th>
<th>Sur</th>
<th>Sad</th>
<th>Ang</th>
<th>Dis</th>
<th>Fea</th>
<th>Con</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neu</td>
<td>88.81%</td>
<td>3.09%</td>
<td>4.03%</td>
<td>2.00%</td>
<td>1.23%</td>
<td>0.11%</td>
<td>0.66%</td>
<td>0.07%</td>
</tr>
<tr>
<td>Hap</td>
<td>10.83%</td>
<td>82.62%</td>
<td>2.94%</td>
<td>1.05%</td>
<td>2.28%</td>
<td>0.06%</td>
<td>0.11%</td>
<td>0.11%</td>
</tr>
<tr>
<td>Sur</td>
<td>7.31%</td>
<td>1.66%</td>
<td>83.03%</td>
<td>0.14%</td>
<td>2.48%</td>
<td>0.00%</td>
<td>4.97%</td>
<td>0.41%</td>
</tr>
<tr>
<td>Sad</td>
<td>18.83%</td>
<td>1.30%</td>
<td>0.97%</td>
<td>70.45%</td>
<td>2.27%</td>
<td>0.97%</td>
<td>5.19%</td>
<td>0.00%</td>
</tr>
<tr>
<td>Ang</td>
<td>6.05%</td>
<td>1.66%</td>
<td>2.02%</td>
<td>0.95%</td>
<td>88.14%</td>
<td>0.36%</td>
<td>0.59%</td>
<td>0.24%</td>
</tr>
<tr>
<td>Dis</td>
<td>18.39%</td>
<td>2.30%</td>
<td>3.45%</td>
<td>2.30%</td>
<td>19.54%</td>
<td>51.72%</td>
<td>0.00%</td>
<td>2.30%</td>
</tr>
<tr>
<td>Fea</td>
<td>4.04%</td>
<td>1.01%</td>
<td>19.19%</td>
<td>4.04%</td>
<td>3.03%</td>
<td>0.00%</td>
<td>68.69%</td>
<td>0.00%</td>
</tr>
<tr>
<td>Con</td>
<td>26.92%</td>
<td>3.85%</td>
<td>3.85%</td>
<td>3.85%</td>
<td>7.69%</td>
<td>3.85%</td>
<td>0.00%</td>
<td>50.00%</td>
</tr>
</tbody>
</table>
Some Examples

"… Kim Kardashian, with a peculiar 12 percent tint of potential happiness."
“Pretty awesome that it detected the underlying emotion.”

“The face of pure happiness.”

“that oddness in her smile is contempt apparently, makes sense”
More Examples

"According to Microsoft's Emotion API, Sidney Crosby was rather angry about scoring the goal that won Canada a gold medal at 2010's Olympics in Vancouver."

“Microsoft thinks Sad Keanu is only 0.01831 sad”
Emotion from Video
I want to try it ...

Emotion API

- Image and video
- Detects happiness, sadness, surprise, anger, fear, contempt, disgust or neutral.
- Returns score for each emotion, sum to one
- REST API, samples in
  - Curl, C#, Java, JavaScript, Object C, PHP, Python, Ruby

Augmented Emotion Dataset and Code

- [https://github.com/Microsoft/FERPlus](https://github.com/Microsoft/FERPlus)
- Relabeled FER 2013 data set
  - 10 labels per image
- CNTK source code for training and testing in Python
  - Same code we use for creating Emotion API
References


Acknowledgment

Cha Zhang
Emad Barsoum
Anna Roth
Chris Thrasher
Cristian Canton Ferrer
Oliver Whyte
Tencent Robotics X
Why Tencent Robotics X?

**Technology Trends**
- Calculation: From PC, smartphones to various forms of intelligent terminals
- Sensors: From single form, limited quantity to everywhere, multimodal, personalized

**Company Strategy**
- Tencent is transforming from a product运营 company to a technology company
- Robots as future technology, Tencent needs to accumulate from now on

**Social Responsibility**
- Ma Huateng said at Tencent’s 18th birthday, Tencent has come of age and needs to assume more social responsibilities
- One urgent problem in Chinese society is aging; developing robots is a natural choice for addressing aging
- Responding to the Industrial 4.0 era, helping enterprises from digitalization, intelligence to intelligent manufacturing, robots are a necessary basis

**Human-centric**
- Tencent’s products: QQ, WeChat, games, are all human-centric
- Developing robot technology is the延续 of human-centric
Tencent Robotics X 使命

创建人与机器和谐互助的未来
create the future of harmonious human-robot cooperation

增强人的智力，augment human intelligence
关怀人的情感，care human emotion state
发挥体能潜力，enable physical potential
人口老龄化和老年人的抚养、陪护已成为重要的社会问题

社会责任：机器人帮助解决养老问题

### 人口老龄化和老年人的抚养、陪护

#### 1990年-2050年

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>65岁以上人口比例</td>
<td>5.60%</td>
<td>6.20%</td>
<td>7.00%</td>
<td>7.70%</td>
<td>8.90%</td>
<td>10.50%</td>
<td>11.98%</td>
<td>16.68%</td>
<td>24.41%</td>
</tr>
<tr>
<td>老年抚养比例</td>
<td>8.30%</td>
<td>9.20%</td>
<td>9.90%</td>
<td>10.70%</td>
<td>11.90%</td>
<td>14.30%</td>
<td>16.90%</td>
<td>22.00%</td>
<td>27.90%</td>
</tr>
</tbody>
</table>
老年人陪护机器人

生活便捷
老年人在体力活动方面需要得到帮助

情感陪伴
老年人需要陪伴和与家人方便的沟通

健康安全管理
老年人需尽量避免意外和疾病的突发
机器人技术的未来方向

现有机器人技术方向
- 智能交互
- 仿人的形态
- 系统控制

未来技术发展方向
- 学习进化
- 情感陪伴
- 环境智能
- 灵巧操控
研究方向

- Perception
- Control/Balance
- Mechatronics
- Design

- Mobility: Navigation, Motion Planning, Obstacle Avoidance
- Human-Robot Interaction
- Flexible Manipulation/Grasping
- Developmental/Reinforcement/Continuous/Lifelong Learning
WE ARE
HIRING!

You're
Hired!
Thanks! Questions?

zhengyou@tencent.com