Towards Wearable Cognitive Assistance

Mahadev Satyanarayanan
School of Computer Science
Carnegie Mellon University

Joint work with: Brandon Amos, Zhuo Chen, Benjamin Gilbert, Keryong Ha, Jan Harkes, Martial Hebert, Wenlu Hu, Roberta Klatzky, Padmanabhan Pillai (Intel), Dan Siewiorek

http://www.istc-cc.cmu.edu/
A Unique Moment in Time

Convergence of Advances in 3 Independent Arenas

- Algorithmic Advances
- Cloud-Mobile Convergence
- Wearable Hardware
Watson (2011)
Face Recognition

LFW Dataset

- Human cropped (97.5%)
- DeepFace-ensemble (97.35%)
- DeepFace-single (97.00%)
- TL Joint Baysian (96.33%)
- High-dimensional LBP (95.17%)
- Tom-vs-Pete + Attribute (93.30%)
- combined Joint Baysian (92.42%)

Speed: 330 milliseconds per image on single-core Intel 2.2 GHz CPU

YTF Dataset

- DeepFace-single (91.4%)
- VSOF+OSS(Adaboost) (79.7%)
- STFRD+PMML (79.5%)
- APEM+FUSION (79.1%)
- MBGS(mean) LBP (78.9%)
- MBGS(mean) FPLBP (78.9%)

“DeepFace: Closing the Gap to Human-Level Performance in Face Verification”
Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, Lior Wolf
Proceedings of CVPR 2014 (IEEE Conference on Computer Vision and Pattern Recognition), Columbus, OH, June 2014
Natural Language Translation

![Bar chart showing BLEU scores for different language translation systems.](chart.png)

- **Human Scoring Range**: Between 0.3859 and 0.7447
- **BLEU Scores**:
  - Google Chinese (‘06 NIST): 0.3859
  - Google Arabic (‘05 NIST): 0.5137
  - Systran Spanish: 0.5551
  - SDL Spanish: 0.5610
  - Google Spanish ‘08 top lang: 0.7289
  - CBMT Spanish: 0.7447

*Based on same Spanish test set* (slides from Carbonell, 2008)

---

Convergence: Wearable Cognitive Assistance

Entirely new genre of applications

Combine mobile and cloud with real-time cognitive engines

scene analysis, object/person recognition, speech recognition, language translation, planning, navigation, question-answering technology, voice synthesis, …

Seamlessly integrated into inner loop of human perception and cognition
Why?

*(more use cases later in talk)*

**Cognitive decline**

- traumatic brain injury
  (accidents, war, sports injuries, …)
- Alzheimer’s disease
- survivors of stroke
- mild cognitive impairment
- …

**Inability to**

- recognize people
- recognize objects
- interpret text and signs
- remember daily routines
- …

**Just in the United States**

- over 20M Americans affected
- heavy burden on caregivers
- est. savings of $1B+ annually
  by 1-month delay in nursing home admission

**Extrapolate to global scale!**
Glass Offers Hope

Glass-based *assistive system*

- real-time scene interpretation (vision and other sensors)
- offer helpful (audio) hints to user when appropriate

- Recognize faces and guides Ron to greet.
- Read text from signs.
- Translate when necessary.
- Reminder for daily routines.
- Water me!
- Please take me out!
Crisp Interactive Response

Humans are amazingly fast, accurate and robust

- face detection under hostile conditions \( < 700 \text{ ms} \) (low lighting, distorted optics)
- face recognition \( 370 \text{ ms} - 620 \text{ ms} \)
- is this sound from a human? \( 4 \text{ ms} \)
- VR head tracking \( < 16 \text{ ms} \) (2004 NASA study, Ellis et al)

Not enough to just match humans

- we need to be “superhuman”
- allow enough time budget for additional cognitive processing

Safe goal: \( E2E \text{ Latency} < \text{“few tens of ms”} \)
Conquering Latency

**Choice 1:** standalone apps

**Choice 2:** offload to cloud
- RTT is too long
- optimal Amazon site ~74 ms

**Choice 3:** offload to cloudlet
- “data center in a box”
- “bring cloud closer”
- 1-hop Wi-Fi access
- typical RTT < 10 ms

<table>
<thead>
<tr>
<th>Metric</th>
<th>Standalone</th>
<th>With Offload</th>
</tr>
</thead>
<tbody>
<tr>
<td>Per-image speed (s)</td>
<td>10.49 (0.23)</td>
<td>1.28 (0.12)</td>
</tr>
<tr>
<td>Per-image energy (J)</td>
<td>12.84 (0.36)</td>
<td>1.14 (0.11)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Year</th>
<th>Processor</th>
<th>Speed</th>
<th>Typical Server</th>
<th>Typical Handheld or Wearable Device</th>
<th>Speed</th>
</tr>
</thead>
<tbody>
<tr>
<td>1997</td>
<td>Pentium® II</td>
<td>266 MHz</td>
<td>Palm Pilot</td>
<td>16 MHz</td>
<td></td>
</tr>
<tr>
<td>2002</td>
<td>Itanium®</td>
<td>1 GHz</td>
<td>Blackberry 5810</td>
<td>133 MHz</td>
<td></td>
</tr>
<tr>
<td>2007</td>
<td>Intel® Core™ 2</td>
<td>9.6 GHz (4 cores)</td>
<td>Apple iPhone</td>
<td>412 MHz</td>
<td></td>
</tr>
<tr>
<td>2011</td>
<td>Intel® Xeon® X5</td>
<td>32 GHz (2x6 cores)</td>
<td>Samsung Galaxy S2</td>
<td>2.4 GHz (2 cores)</td>
<td></td>
</tr>
<tr>
<td>2013</td>
<td>Intel® Xeon® E5</td>
<td>64 GHz (2x12 cores)</td>
<td>Samsung Galaxy S4</td>
<td>6.4 GHz (4 cores)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Google Glass OMAP 4430</td>
<td>2.4 GHz (2 cores)</td>
<td></td>
</tr>
</tbody>
</table>

Source: adapted from [Flinn 2013]
Bring the Cloud Closer

Create a Small Cloudlet Nearby

- **Aircraft cloudlet**
- **Car cloudlet**
- **Glass**
- **Smartphone**
- **Coffee Shop Cloudlet**

cloudlet ≡ “data center in a box”
- compute cluster
- “good” Internet access
- no battery limitations
- VM-based multi-tenancy
Micro Data Centers Exist Today

Used as private clouds today
Need to be re-purposed as cloudlets i.e., as “2nd-tier infrastructure”
Today’s Unmodified Cloud

Internet

Like a CDN for computation
Loosely Coupled Architecture

Many human tasks involve distinct cognitive capabilities
  • e.g. human conversation
  • multiple independent sensor channels
    analyzed in parallel, combined in real-time
  • strong evidence that brains use coarse-grain parallelism

Leverage off-the-shelf building blocks
  • face / object recognition, OCR, speech-to-text, language translation …
  • use these to catalyze our new class of applications

Diverse programming languages, optimizing compilers, OSes
Gabriel Architecture


- Wearable device
- Wireless connection
- Sensor control
- Video/Acc/GPS/... sensor streams
- User assistance
- Control VM
  - Context Inference
  - Device Comm
  - PubSub
  - UPnP
- User Guidance VM
- Cloudlet
  - Cognitive VMs
    - Face recognition
    - Object Recognition (MOPED)
    - Object Recognition (STF)
    - OCR
    - Motion classifier
    - Activity Inference
    - Augmented Reality
      - ... many other cognitive VMs here...
Don’t Transmit Hopeless Frames

Input parameter is token count
(small integer, e.g. 1)

Glass Device

sensor

drop

tokens

a

b

a > b?

yes

no

∑ sum → sensor data ← counts

Control VM

MAX

Drop Count

Cognitive VM

Processed Items Count
Many Results in the Paper
(just a sampling here)

Cloudlets are essential

E.g. Face recognition (cloudlet versus Amazon):

![Graph showing response time comparison between Cloudlet and Cloud](image)
<table>
<thead>
<tr>
<th>Application</th>
<th>Cloudlet (Joule/query)</th>
<th>Cloud (Joule/query)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Face</td>
<td>0.48</td>
<td>0.82</td>
</tr>
<tr>
<td>AR</td>
<td>0.19</td>
<td>0.32</td>
</tr>
<tr>
<td>OCR(open)</td>
<td>2.01</td>
<td>3.09</td>
</tr>
<tr>
<td>OCR(comm)</td>
<td>1.77</td>
<td>2.41</td>
</tr>
</tbody>
</table>

Figure 14: Energy Consumption on Google Glass
## Performance Summary

*(many more results in paper)*

<table>
<thead>
<tr>
<th>Cognitive Engine</th>
<th>Response time (ms)</th>
<th>Glass Life</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1% 10% 50% 90% 99%</td>
<td>~1 hour</td>
</tr>
<tr>
<td>Face Recognition</td>
<td>196 389 659 929 1175</td>
<td></td>
</tr>
<tr>
<td>Object (MOPED)</td>
<td>877 962 1207 1647 2118</td>
<td></td>
</tr>
<tr>
<td>Object (STF)</td>
<td>4202 4371 4609 5055 5684</td>
<td></td>
</tr>
<tr>
<td>OCR (Open)</td>
<td>29 41 87 147 511</td>
<td></td>
</tr>
<tr>
<td>OCR (Comm)</td>
<td>394 435 522 653 1021</td>
<td></td>
</tr>
<tr>
<td>Motion Classifier</td>
<td>126 152 199 260 649</td>
<td></td>
</tr>
<tr>
<td>Augmented Reality</td>
<td>48 72 126 192 498</td>
<td></td>
</tr>
</tbody>
</table>

1. Today’s cognitive engines are slow *(none of medians are “few tens of ms”)*
2. Huge inter-frame variability *(content-specific)*
3. Faster engines/frames not hurt by slower ones *(token-based flow control)*
Beyond Cognitive Disabilities
Task-specific Assistance

Example: cooking

passive recipe display

versus active guidance

“Wait, the oil is not hot enough”
Inspiration: GPS Navigation Systems

Turn by turn guidance
  • Ability to detect and recover
  • Minimally distracting to user

Uses only one type of sensor: location from GPS

*Can we generalize this metaphor?*
Many Use Cases …

- Assembly instructions
- Medical training
- Industrial troubleshooting
- Strengthening willpower
Prediction

Wearable Cognitive Assistance will be the “killer app” of mobile computing in the next decade

(and we will not get there without cloudlets)
Thank You!