Ubiquitous and Mobile Computing
CS 403x: *Visage: A Face Interpretation Engine for Smartphone Applications*

Robert Esposito, Christopher Knapp, and Doruk Uzunoglu

Computer Science Dept.

Worcester Polytechnic Institute (WPI)

Introduction

- Mobile phone camera is a ubiquitous sensor like the microphone, accelerometer, etc.
  - However, it has not been exploited to nearly the same extent
- Emotion categories represented by universal facial expressions

Vision

- Set of sensing, tracking, and machine learning algorithms on smartphones
- Engine that interprets head poses and facial expressions to provide automatic feedback

Related Work

- Computer vision algorithms on mobile applications
  - SenseCam
  - MoVi
  - Recognizr
  - Google Goggles
- Mobile head pose trackers
  - PEYE

Design Considerations

- User mobility
  - Angle, motion, and light exposure level of mobile phones are unpredictable
- Limited phone resources
  - Video camera produces 50 times more data than microphone, 800 times more than accelerometer
  - Visage processes video streams in real time

System Architecture Design

- Sensing
- Preprocessing
- Tracking
- Inference

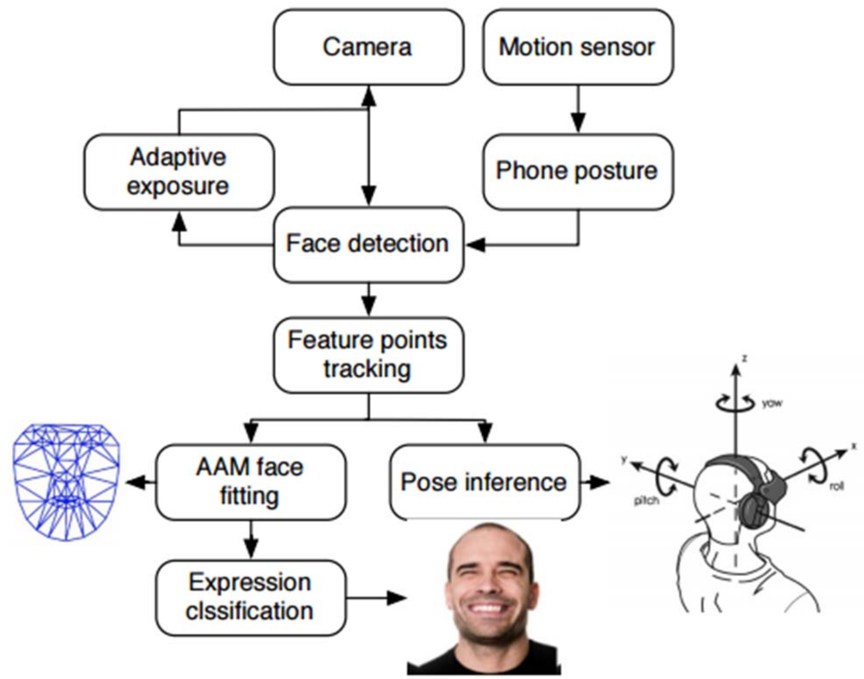


Fig. 1. Visage System Architecture

Preprocessing Stage

- Phone posture: Estimates gravity and motion using accelerometer and gyroscope readings
- Face detection with tilt compensation: Scans with decreasing window size until face detected
- Adaptive exposure: Corrects exposure level based on local lighting information within face region

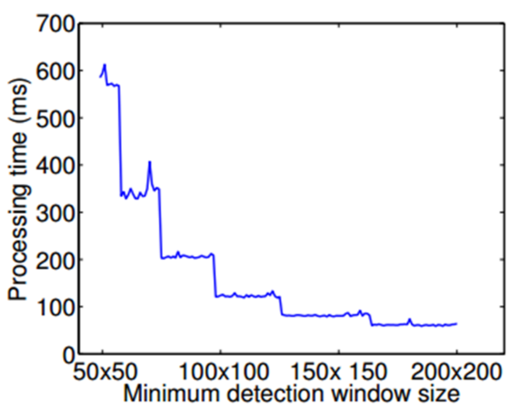


Fig. 2. Early termination scan scheme

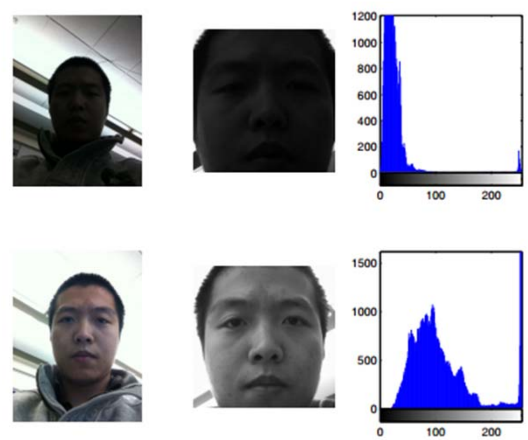


Fig. 3. Top: underexposed image, face region, and regional histogram; bottom: the image after adaptive exposure adjustment, face region, and regional histogram

Tracking Stage

- Feature points tracking
  - Identifies possible feature points on first frame and tracks their locations
- Pose estimation
  - POSIT algorithm estimates 3D geometry of user's head using 2D feature points

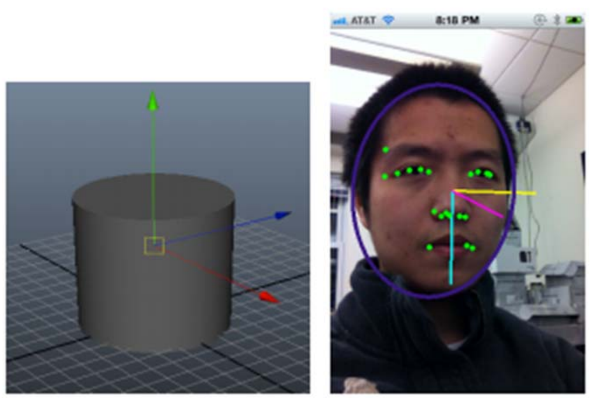


Fig. 4. (a) Cylinder model and (b) Tracking with pose estimation.

Inference Stage

- Active appearance model
  - Describes 2D image as triangular mesh of landmark points
  - Uses pixel color to enhance model accuracy
- Seven expression classes: angry, disgust, fear, happy, neutral, sad, surprise

Fig. 5. Visage expression classification on the iPhone 4

Implementation

- Prototyped on iPhone 4 in C and Objective C
- Downsampling images
  - Minimal performance penalty
- Drops oldest frames when processing cannot keep up with video stream

Resolution	Time (ms)
640 x 480	4090
480 x 360	2123
320 x 240	868
192 x 144	298
160 x 120	203
96 x 72	68
80 x 60	53

Table 1. Computational costs of face detection

CPU and Memory Benchmarks

Tasks	Avg. CPU usage	Avg. memory usage
GUI only	< 1%	3.18 MB
Pose estimation	58%	6.07 MB
Expression inference	29%	4.57 MB
Pose estimation & expression inference	68%	6.28 MB

Table 2. CPU and memory usage under various tasks

Component	Average processing time (ms)
Face detection	53
Feature points tracking	32
AAM fitting	92
Facial expression classification	3

Table 3. Processing time benchmarks

Accuracy of Head Pose Estimation

- Tested with angles from -90 to 90 degrees in increments of 15 degrees
- Mean absolute error = 5.51° ± 1.9°

Fig. 6. Images captured by the front-facing camera assuming varying phone tilted angles from -90 $\sim$ 90 degrees, separated by an angle of 15 degrees. The red boxes indicate the detection results. The first row is detected by the standard Adaboost face detector. The second row is detected by Visage's detector.

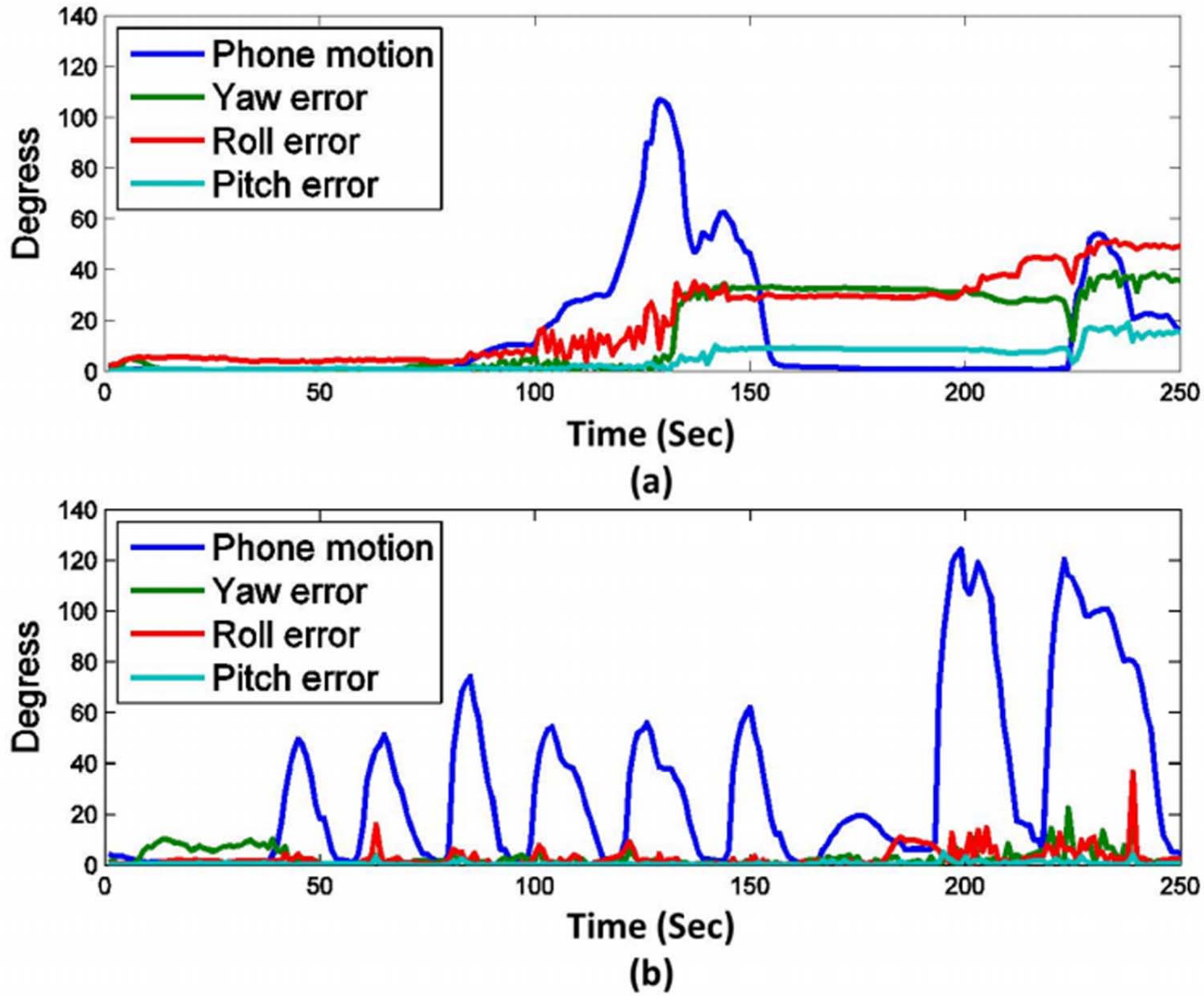


Fig. 7. Phone motion and head pose estimation errors (a) without motion-based reinitialization, and (b) with motion-based reinitialization

Accuracy of Facial Expression Classification

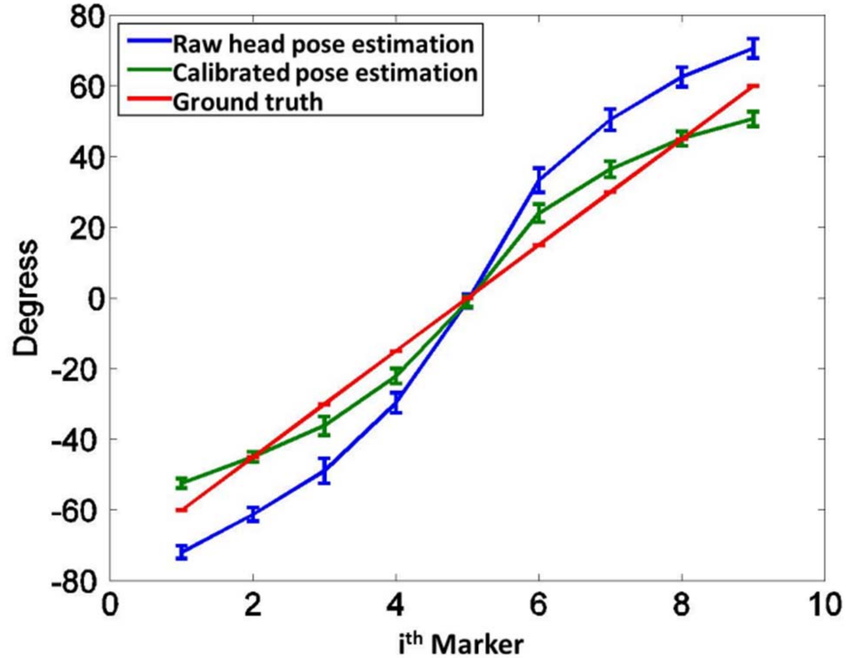


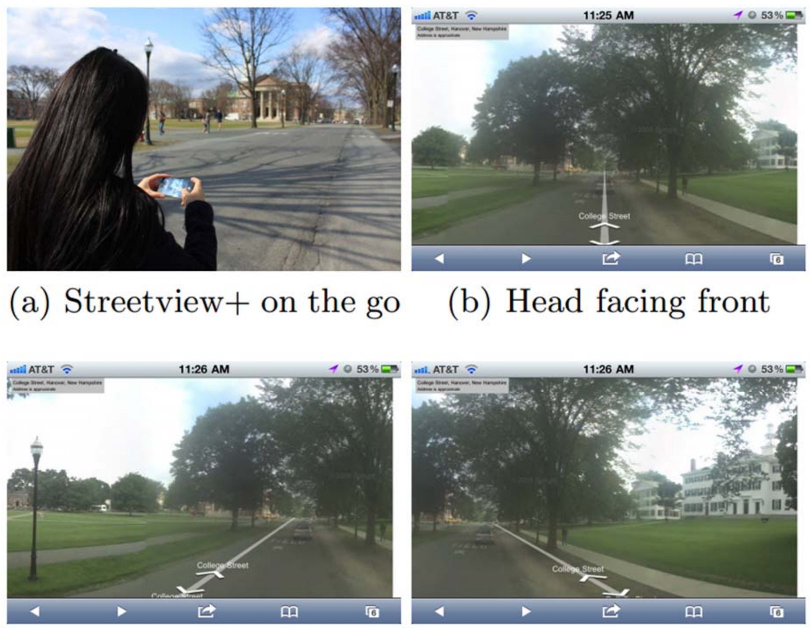
Fig. 8. Head pose estimation error

Expressions	Anger	Disgust	Fear	Happy	Neutral	Sadness	Surprise
Accuracy (%)	82.16	79.68	83.57	90.30	89.93	73.24	87.52

Table 4. Facial expression classification accuracy using the JAFFE dataset

Applications: Streetview+

- App tracks user's head rotation and GPS location to provide a panorama view from Google Streetview
- 12-15 frames per second



(a) Streetview+ on the go (b) Head facing front

(c) Head facing left (d) Head facing right

Fig. 9. Steetview+ enhanced with awareness of user head rotation

Applications: Mood Profiler

- Tracks user's mood in the background while they use other apps
- One classification per second

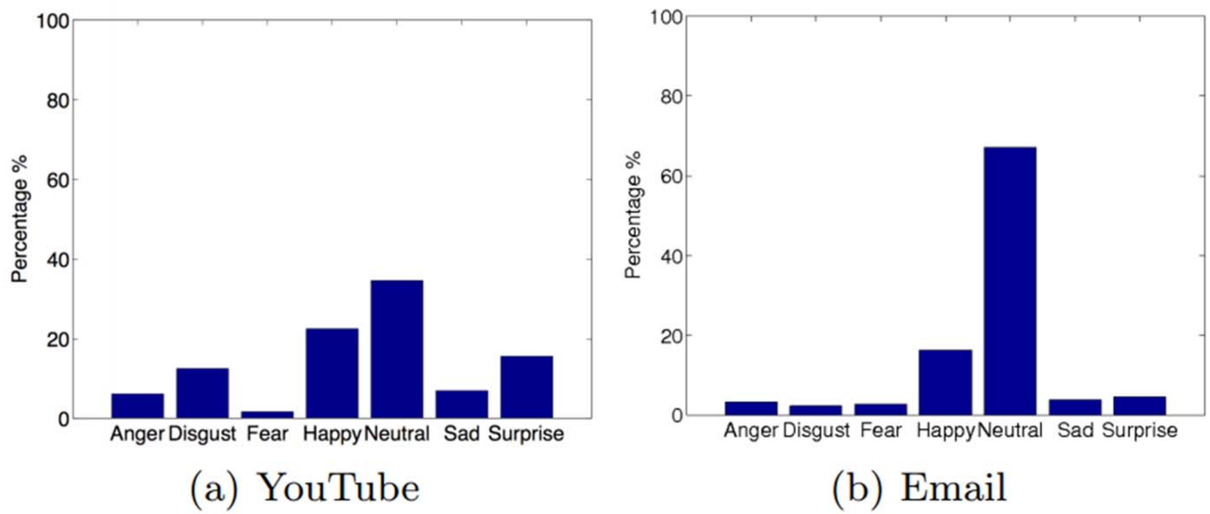


Fig. 10. Expression histogram when using (a) the YouTube mobile application and (b) an email client.

Conclusions

- Visage is able to track head position and facial expressions
- Carries out all sensing and classification tasks on the phone without server
- Relatively low computational cost

References

- *Visage: A Face Interpretation Engine for Smartphone Applications* http://www.cs.dartmouth.edu/~campbell/visage.pdf