The 13th IEEE Conference on Automatic Face and Gesture Recognition
2018 FG Xi'an, China | May 15-19, 2018

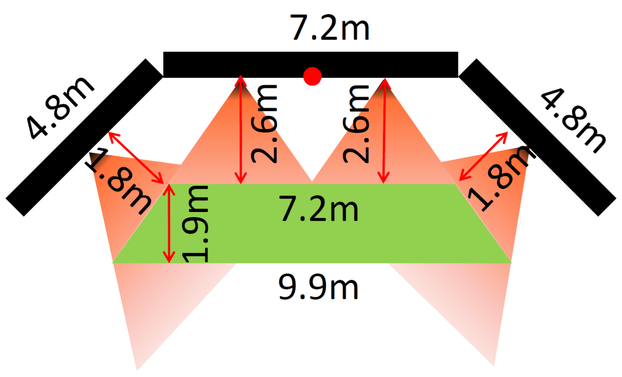
# An Immersive System with Multi-modal Human-computer Interaction

Rui Zhao[1], Kang Wang[1], Rahul Divekar[1], Robert Rouhani[1], Hui Su[1,2], Qiang Ji[1]

[1] RPI, [2] IBM T. J. Watson Center

## Overview

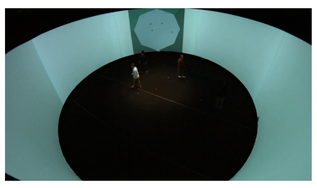
- An immersive system prototype that integrates face, gesture and speech recognition techniques to support multi-modal human-computer interaction capability.
- Understand users' intention by resolving query like 'who is doing what at where' in real-time.



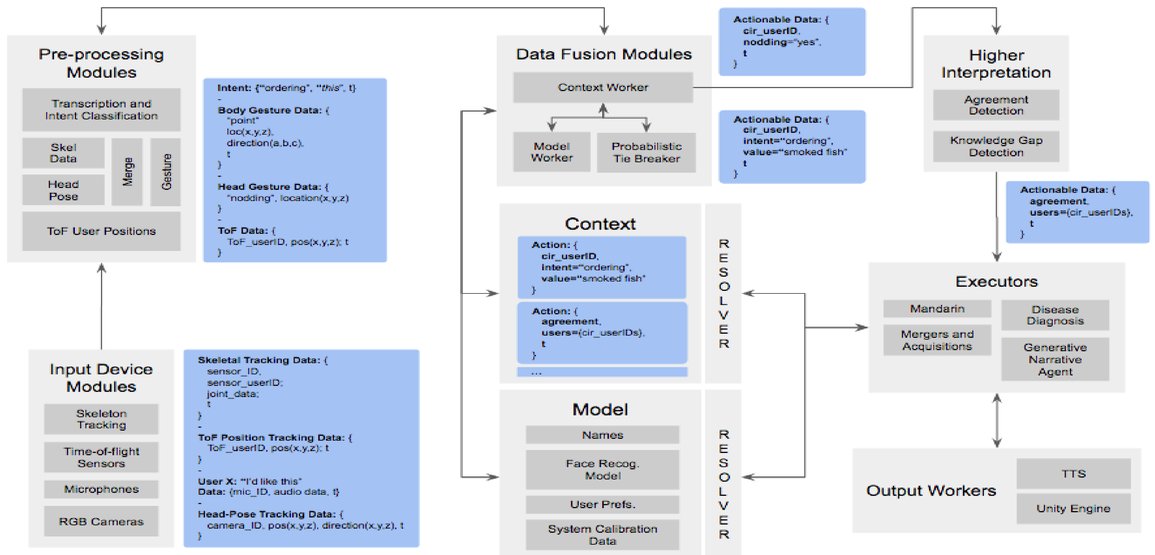

Studio @ EMPAC, RPI

3-screen Display Set-up

Panoramic Display Set-up



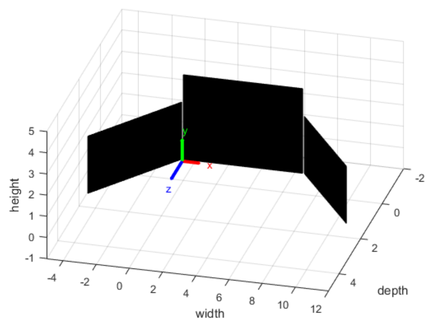
## System Architecture



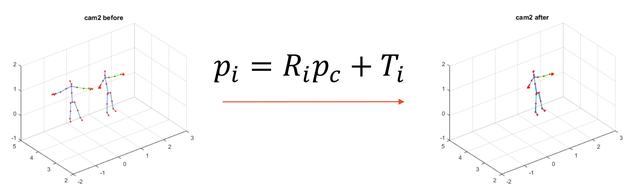

- Sensing Devices

Kinect Camera, Pan-Tilt-Zoom Camera, Webcam, Lapel Microphone

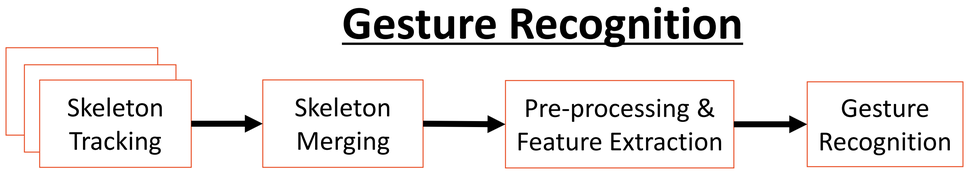
## System Calibration

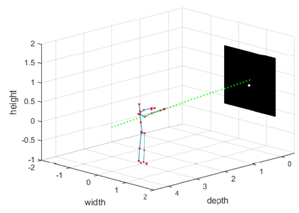


- Estimate camera pose against common room coordinate system




## Gesture Recognition



- List of gestures: Point, Wave, Clap, Swipe, Cross-arm, Zoom, Boxing, Throw, Shoot, Jumping Jack, etc.

Table II
CLASSIFICATION ACCURACY ON DIFFERENT DATASETS.

| Dataset | # of Action | SVM | HMM | SNV[36] |
|---|---|---|---|---|
| CIR | 6 | 90.6% | 91.2% | 99.0% |
| UTD | 27 | 87.0% | 89.5% | 90.9% |
| Real-time | | Yes | Yes | No |

## Face Recognition

Snapshots:

Robert Kang Rui

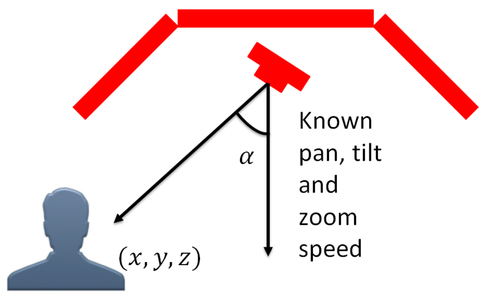
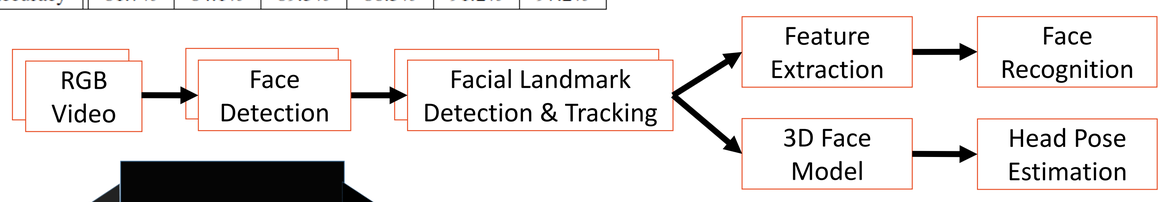
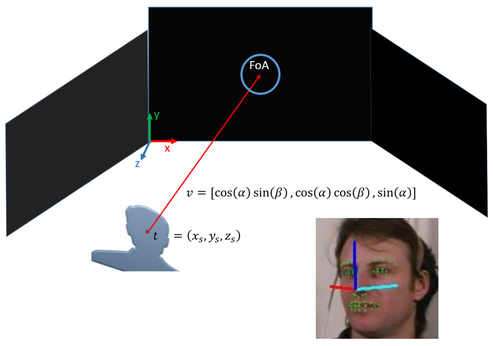



Table III
FACE RECOGNITION ACCURACY WITH DIFFERENT FEATURES

| Features | Eigen | Fisher | LBP | HoG | SIFT | All |
|---|---|---|---|---|---|---|
| Accuracy | 81.7% | 84.1% | 89.3% | 86.5% | 91.2% | 97.2% |



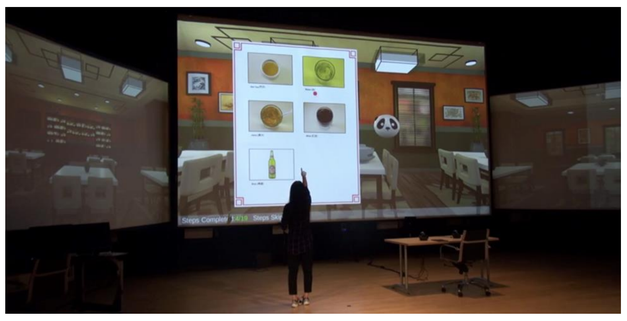




## Head Pose Estimation

Table IV
POSE ESTIMATION ERROR ON BENCHMARK DATASETS

| Dataset / Angle | Pitch | Yaw | Row | Average |
|---|---|---|---|---|
| CAS-PEAL [21] | 4.2 | 3.1 | 2.2 | 3.2 |
| Multi-Pie [22] | 5.4 | 3.3 | 1.9 | 3.5 |

## Use Case: Immersive Classroom

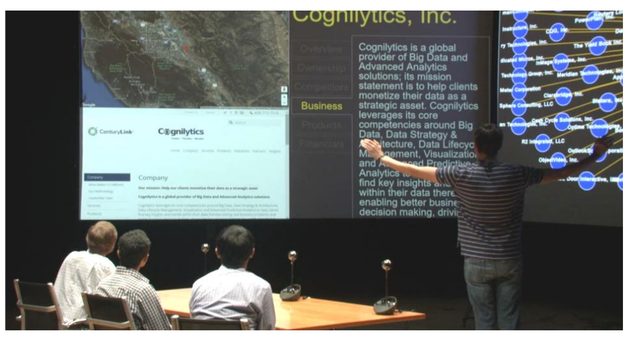


- Immersive Language Learning Experience through enabling natural multi-modal interaction
- User study shows benefit of the system

Table VI
AVERAGE HELPFULNESS SCORE. ON A SCALE OF 1-5, 5 BEING VERY HELPFUL AND 1 BEING VERY UNHELPFUL.

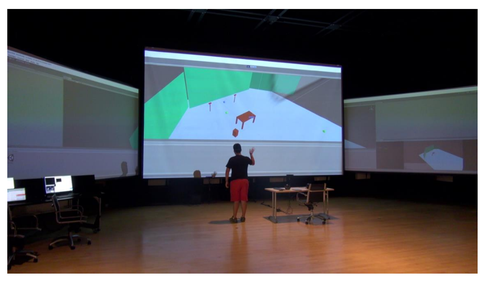
| Function | CIR | Real-world |
|---|---|---|
| Pointing to the items | 4.4 | 4.5 |
| Switching languages with speech | 3.8 | 4.3 |
| Combine gesturing and speech | 3.8 | 4.1 |

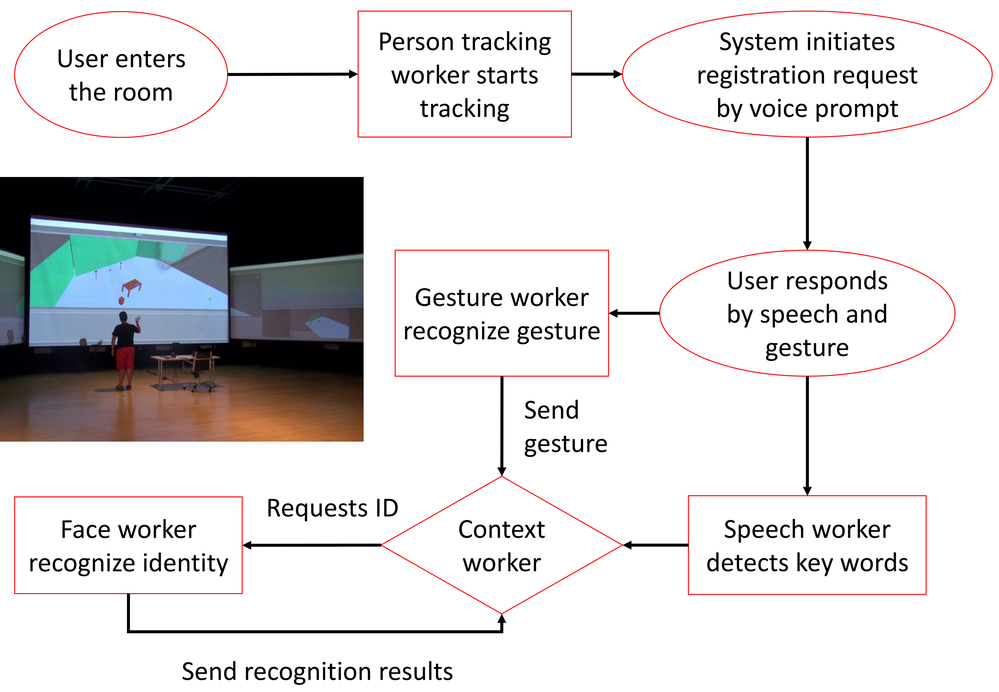
## Use Case: Meeting Assistance



- Assist information analysis and decision making through verbal and gesture interaction.
- Monitoring the progress through fusing multiple input modalities like speech, gesture and facial expression.
- Detect user attention through consensus of head pose.

## Use Case: User Registration






**Acknowledgement** This work is supported by Cognitive Immersive Systems Lab (CISL), a collaboration between IBM and RPI, and also a center in IBM's Artificial Intelligence Network.

**Contact author: zhaor@rpi.edu**