MOMENT

Autonomous Video Conferencing System

NICHOLASH BEDI, NAVID MIA, ALEXANDER SHI, AILEEN YU

Problem

Conventional camera systems lack the capability to capture subjects outside of their fields of view and the intelligence to determine the importance between multiple points of interest.

For video calls, traditional cameras provide a bad user experience whenever there are multiple or moving subjects to focus on and capture.

55 Million+
Video calls per day on WhatsApp

17 Billion+
Video calls in 2017 on Messenger

2 Trillion+
Video call minutes over the last decade on Skype

Goals

Design a device that is capable of autonomously identifying and tracking points of interest (POI) for conference calls.

Locate a person based on sound

Track a person within frame

Switch between people of interest

Alternatives Designs

1

360° Camera
- Portable form-factor and simple mechanical design
- Image distortion

2

Single Rotating Camera
- Intelligent user-tracking and good presentation
- Poor performance for multiple users

Challenges

Mechanical Challenges

- Continuous 360° rotation requires rotation without wires twisting
- Shaft diameter limited due to hollow requirements to pass wires through, resulting in shafts prone to failure
- Bending in cantilever platforms due to the weight of the motor and drive gear

Electrical Challenges

- Number of wires limited to 12 connections due to the slip rings and limited space inside shafts
- Not possible to produce 50Ω controlled impedance through slip rings, resulting in lower USB signal integrity

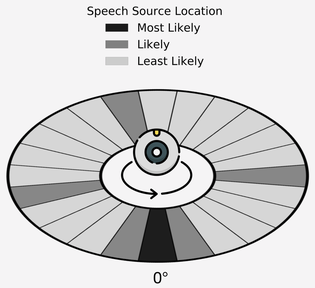
Software Challenges

- Audio speech detection accuracy
- Motor control smoothness for optimal video experience
- Avoiding overtraining of neural network

Technologies

Microphone Array

- Uses four microphones to triangulate noise location
- Processes the noise to determine if it is a person speaking

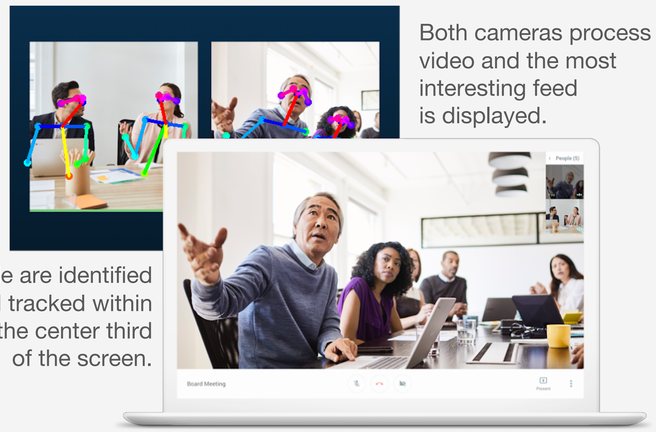


Speech Mapping

- Enables accurate detection of actual sources of speech
- Uses Bayesian statistics to map speech around the device (equation shown below)

$$logit\left(P(POI|noise)\right) = log\left(\frac{P(noise|POI)}{P(noise|\overline{POI})}\right) + logit(P(noise))$$

People Tracking and Focus Selection

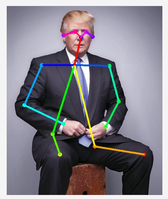


Both cameras process video and the most interesting feed is displayed.

People are identified and tracked within the center third of the screen.

Continuous 360° Rotation

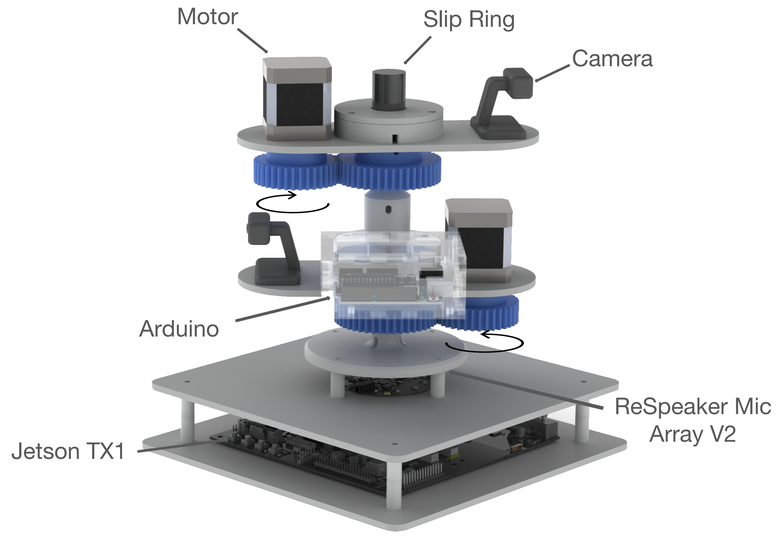
- Both cameras rotate independently with full 360° range
- Zero blind spots for the cameras
- Motor motion smoothing using a p-controller to reduce rotation speed as destination angle approaches



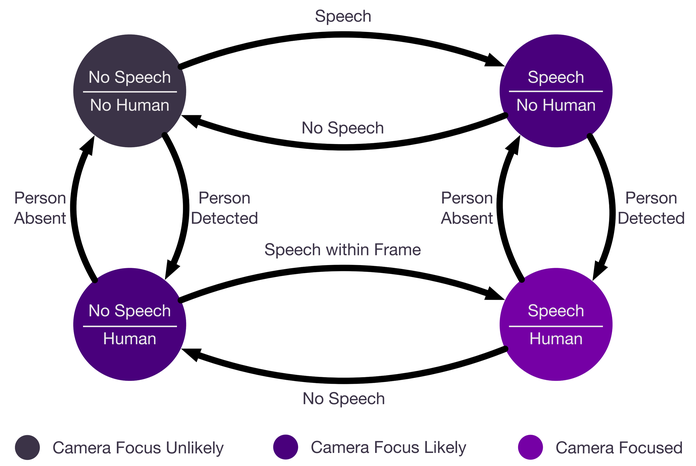
Neural Network

- Trained a model with outputs of OpenPose on 1200+ images to:
  - Determine if a person is sitting or standing
  - Determine if a person's hand is above their head
- Used as a call-to-action for the camera

Design Overview



Tracking Methodology



Results

Neural Network Test Accuracy	180° Rotation Time	Initial Tracking Time	Continuous Tracking Time	Speech Accuracy
88%	1.2 sec	2.8 sec	3.6 sec	93%

References

Special thanks to Prof. Bedi and Prof. Kennings, the supervisors for this project, for their assistance and advice.

Cantisano, T. (2016, January 12). Neowin. Retrieved December 3, 2018, from https://www.neowin.net/news/750-millionusers-and2-trillion-minutes-of-free-videocalls-skype-celebrates-10-years
Wakeel, N. (2017, 5 11). aaj. Retrieved 12 3, 2018, from https://www.aaj.tv/2017/05/fifty-fivemillion-whatsapp-video-calls-per-day-stats/
Welch, C. (2017, December 13). The Verge. Retrieved December 3, 2018, from https://www.theverge.com/2017/12/13/16772704/facebook-messenger-17-billionvideo-chats-2017
G Suite. "Google Cloud". Retrieved March 3, 2019 from https://gsuite.google.com/products/meet/