# Unsupervised Speaker Cue Usage Detection in Public Speaking Videos

Anshul Gupta

Multimodal Perception Lab

IIIT Bangalore

# Nonverbal Cues in Public Speaking


# Nonverbal Cues in Public Speaking

Haider et al. (2017) highlight the importance extract high level features and correlate them to engagement

Chen et al (XX), XX et al (XX) use low level spatio-temporal cues such as XX

# Nonverbal Cues in Public Speaking

- Most work use low level spatio-temporal features

- Difficult to provide feedback, analyse in human-interpretable form

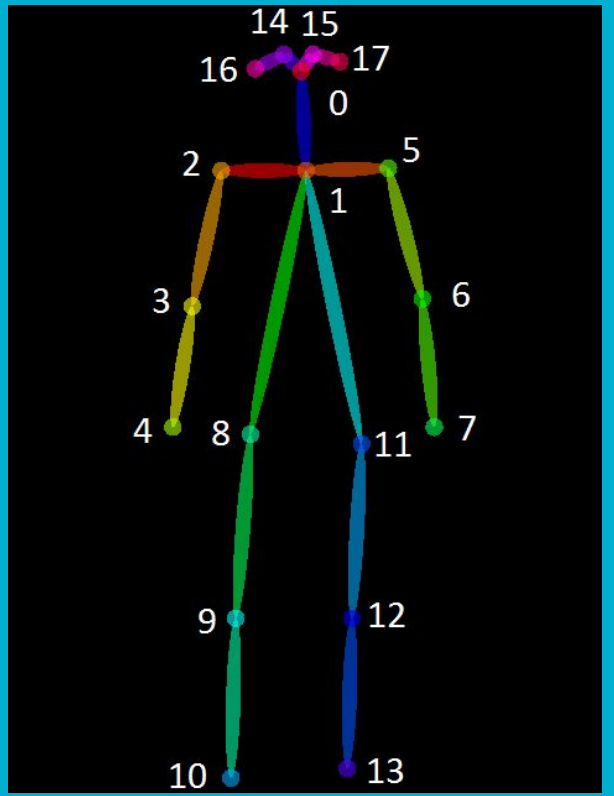
# Proposed Solution

- A set of higher level features that quantizes the amount of cue usage into 3 categories
- We test our approach on two datasets:
  - Classroom recorded videos
  - TED videos

# Proposed Solution

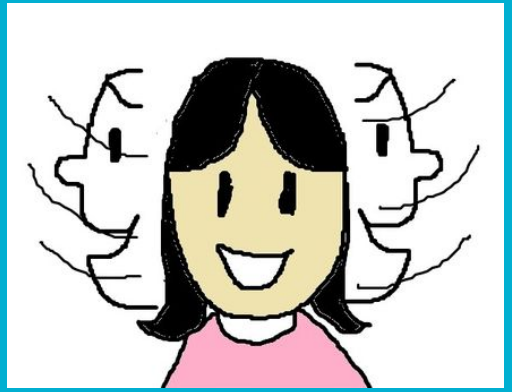
- Clip the last 1 minute of the video

- Get pose keypoints using the Openpose library

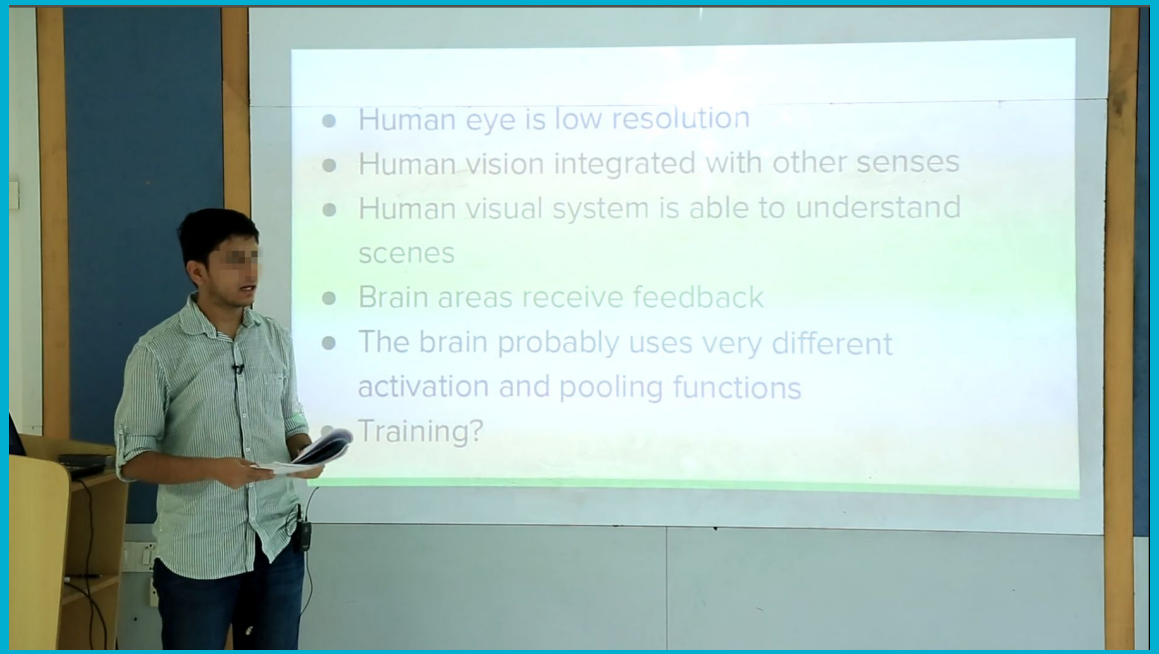


# Proposed Solution

- Compute low level features (std. dev, speed and acceleration)
- Compute mean of speed and acceleration across the clip
- Normalize features with respect to their max values
- Cluster into 3 groups



# Classroom Videos



# Classroom Videos

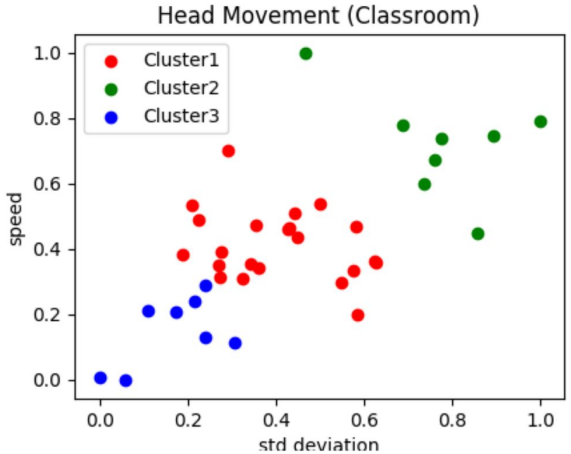
- For lateral head movement we use the nose X coordinate subtracted from the neck X coordinate
- For walking we use the hip X coordinate

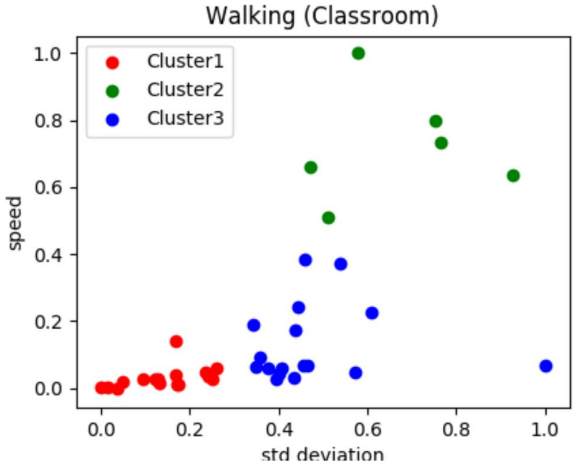
We sample 5 videos  each from from the cluster closest to the origin, and the cluster furthest from the origin to capture high and low usage of the cue

Binary classification problem

# Results

| Data | Accuracy |
| --- | --- |
| Head Movement | 80% |
| Walking | 60% |



# TED Videos

# TED Videos

Ted videos are complex due to a number of reasons:

- The camera following the speaker in close up shots
- Background being of a single colour
- Multiple cuts and zoom shots in the video
- Occlusion of the body with often only the upper body above the shoulders visible
- Multiple people in the frame such as audience members in pan shots

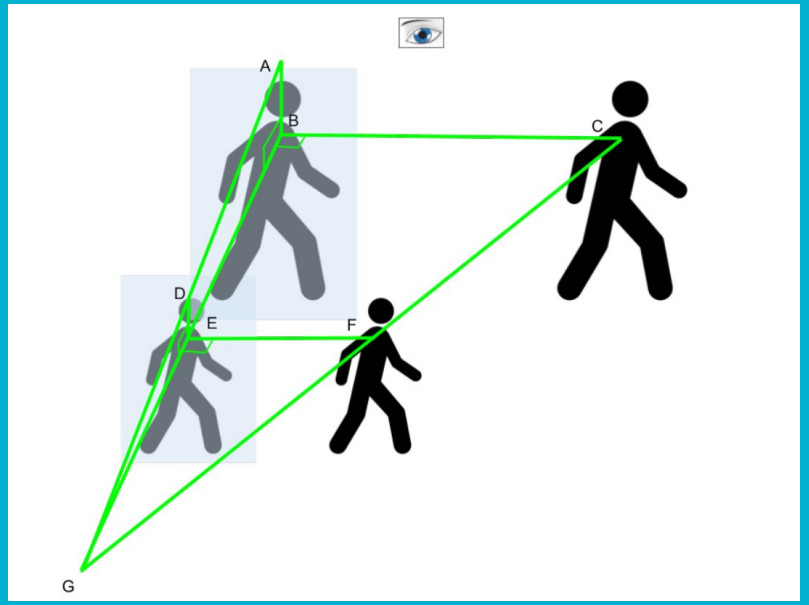
# TED Videos

Modifications to the algorithm:

- Extract a set of cuts using PySceneDetect library
- Compute low level features for a cut
- Scale each feature by an inverse zoom factor
- Average feature values across cuts

# Inverse Zoom Factor

- Inverse of the Y axis distance between the speaker’s nose and neck
- Calculated for each cut



# TED Videos

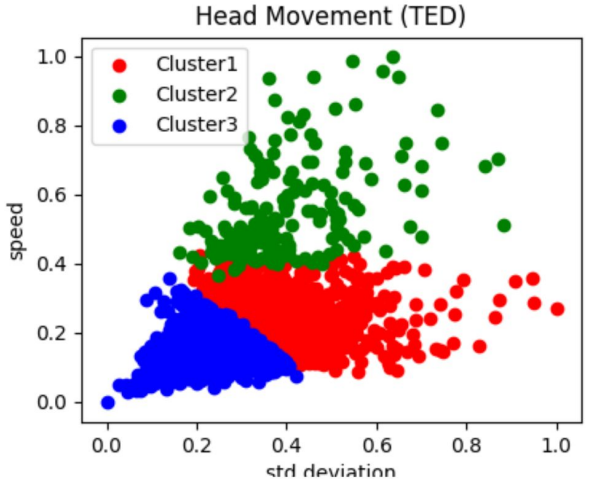
- For lateral head movement we use the nose X coordinate subtracted from the neck X coordinate
- For walking we use the neck X coordinate

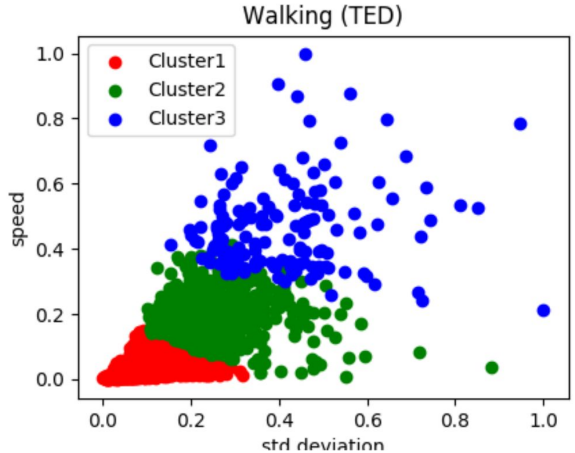
We sample 25 videos  each from from the cluster closest to the origin, and the cluster furthest from the origin to capture high and low usage of the cue

Binary classification problem

# Results

| Data | Accuracy |
| --- | --- |
| Head Movement | 50% |
| Walking | 58% |



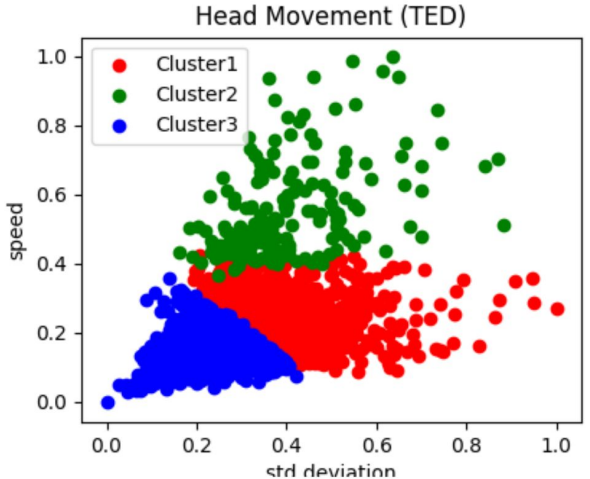



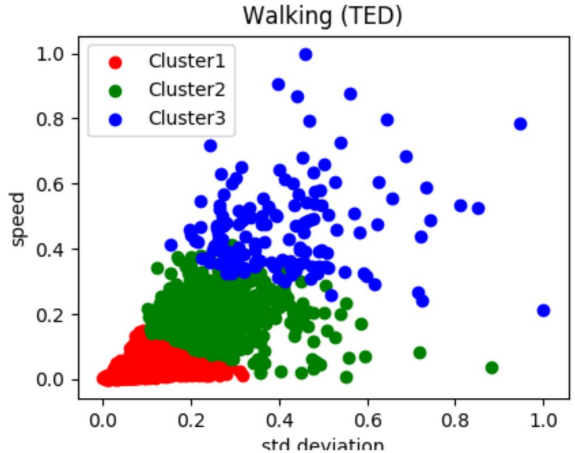
# Results

Low head cue usage identification accuracy:

# Results

| Data | Accuracy |
| --- | --- |
| Head Movement* | 76% |
| Walking | 58% |



# Summary

We have demonstrated a new approach to generating easily computable higher level features in public speaking videos that represent human interpretable ideas of walking and head movement

We hope that these features can be used in future studies to provide better feedback, and obtain more insights into results

# Future Work

To improve results:

- Track main speaker
- Increase number of annotators

Work on identifying “speaker styles” using human-interpretable, high level features

# Questions?

# Thank you!

# Prior work on feedback??

Primarily in the form of rating… (good, bad etc)

Some works correlate diff features to final rating, including gesture related features (Wortwein et al., 2015), but i) Supervised ii) cannot directly be used as a feedback mechanism..

# To Read

On action recognition

Geometric correction

# Regarding annotations??

What answer to give..