HEAD POSE, GESTURE, EYE TRACKING AND SCORE CORRELATION

by

Haoyu He

Supervisor: Professor John R. Kender
Columbia University
Dec 15, 2018

Table of Contents

List of Tables

List of Figures

1 Introduction

Based on previous research [3], this work we investigate how speaker head pose enhance hand gestures. We also show the correlation between speaker head pose and student's eye tracking. Our research visualization software used in previous work [4] is also updated.

2 Data Preprocessing

2.1 Video

In previous work, we have shown how to extract speaker's 2D skeleton from video [1, 2] and how to reconstruct 3D head pose from 2D skeleton [3]. In this work, we use the same videos and skeleton data. The two videos are: "Perspective" and "Bicycle". The speaker gestures in videos are labeled manually. During a time window there maybe two gestures happen at the same time. Table 1 shows the detail gestures information of each video. The gestures are extracted from excel sheet using python scripts (Appendix B). To get a clean data, the data that contains more than one gestures are removed. Only one gestures per time window are valid gestures.

Video	Short Name	Length [min:sec]	Gesture: Beats, Metaphoric, Deictic, Iconic	Valid Gestures: Beats, Metaphoric, Deictic, Iconic
Perspective	ps	4:56	62, 28, 17, 7	44, 18, 11, 4
Bicycle	bi	4:59	67, 20, 20, 4	32, 3, 2, 0
Total	-	-	129, 48, 37, 11	76, 21, 13, 4

Table 1: Video length and gestures number for each videos

The FPS of every video is 30. The resolution is 1920×1080. The upper body 2D skeleton is extracted using OpenPose [1, 2]. Only use 5 head points (index: 0, 14, 15, 16, 17) are used in our 3D head pose model. For each gesture happen at 0s, we extract data from -0.1s to +1s. The total size of data for each time window is $[5 \times 40 \times 2]$, where 5 is the 5 head points, 40 is the number of frames, 2 is the x, y coordinates.

2.2 Participant Data

There are total 60 students watched three videos under three different conditions. After watching video, student need to answer 6 questions. The answer and eye tracking of each

student are recorded.

Condition	Participant Number	Total Questions
SINGLE: only show slides	20	360
DUAL: show slides and audio	20	360
FULL: show slides, audio and speaker	20	360

Table 2: Participant number for different condition

3 Head Poses, Gesture and Scores Correlation

3.1 Classify Head Pose Types

In previous research [3], we are able to extract the speaker 3D head pose from 2D video. In this research, based on the 3D head position, we classify the speaker's head pose into 5 different types: look at camera, look away from slides, look down, look at slides and other. Table 3 shows the detail classification parameters. The x and y are the normalized horizontal and vertical rotation angle of speaker's head.

Type	x	y
slides	$x \leq -0.8$	$y > -0.8$
away_slides	$x \geq 0.8$	$y > -0.8$
camera	$-0.8 < x < 0.8$	$-0.8 < y < 0.8$
down	-	$y \leq -0.8$
other	-	$y \geq 0.8$

Table 3: Head pose type classification parameters

Figure 3.1 shows the count frequency of different head type for perspective video. We can see the camera type has highest counts, because the look at camera is the rest position.

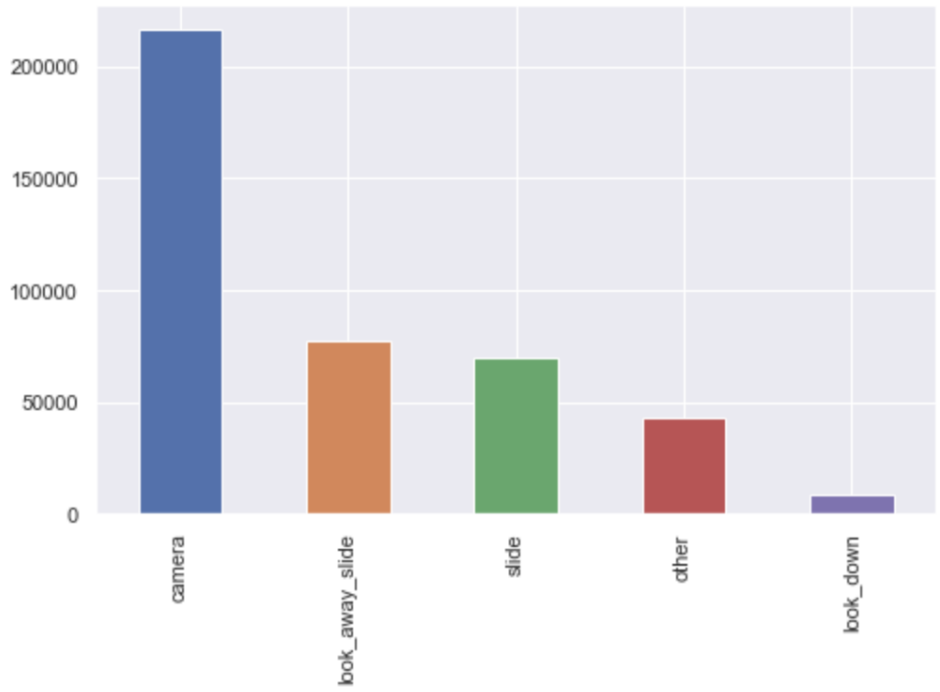

Figure 3.1: Head pose type frequency counts for perspective video. Window size: 40 frames, step size: 1 frame.

3.2 Head Pose, Gesture and Score Correlation

In previous study [3], we show that there are correlation between speaker's 3D head position and hand gesture. We also show that some gestures have positive correlation with scores [4]. In this research, we investigated how head pose can enhance the effect of hand gestures.

We first get the head pose type for each question. Figure 3.2 shows the counts for different head pose type and hand gesture pair. Here we classify hand gestures into one hand and two hand gestures.

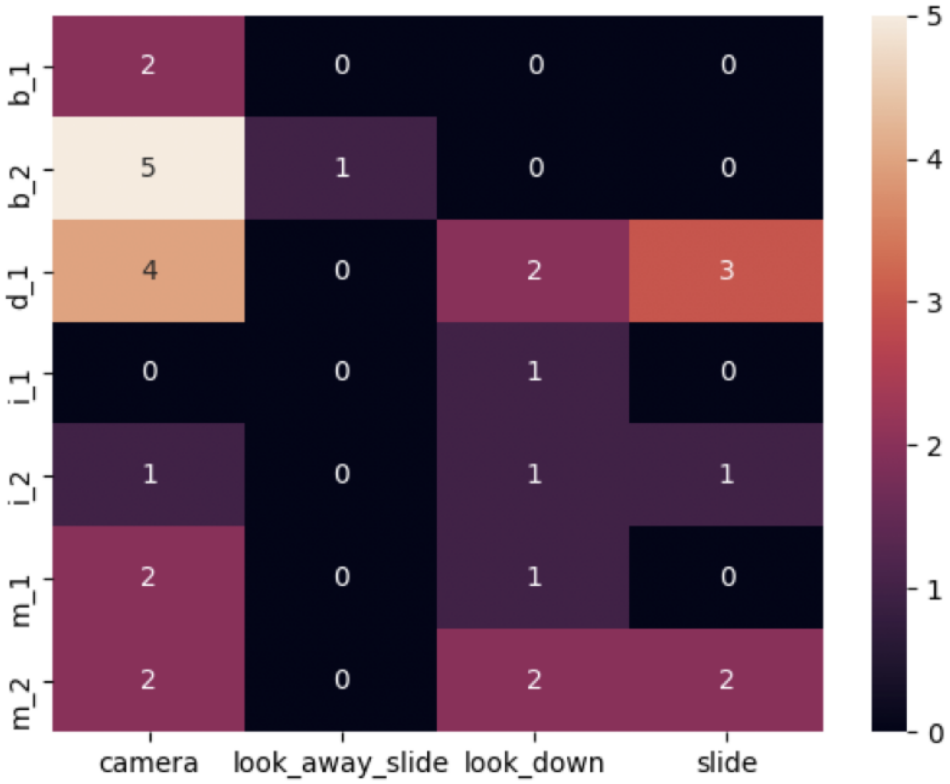

Figure 3.2: Counts of head pose type and head gesture pairs. d: deictic, i: iconic, m: metaphoric, b: beats gestures. m_2: means two hand gesture, m_1 means one hand gesture.

Then for each gesture pair $(hand\_gesture, head\_pose) = (i, j)$, we calculated the normalized score using following equation 3.2.1:

$$score(i,j) = \frac{\sum_{hand,head=i,j} score(i,j)}{\sum_{hand,head=i,j} 1} \tag{3.2.1}$$

Where $score(i,j)$ is the sum of score for all pair equal to (i,j). $\sum_{hand,head=i,j} 1$ is the total counts of the pair (i,j).

Figure 3.3 shows the score for FULL condition. To isolate the effect of speaker's gesture, we subtract the FULL condition score by DUAL condition score. Figure 3.4 shows the results.

Figure 3.3: Normalized score of FULL condition.

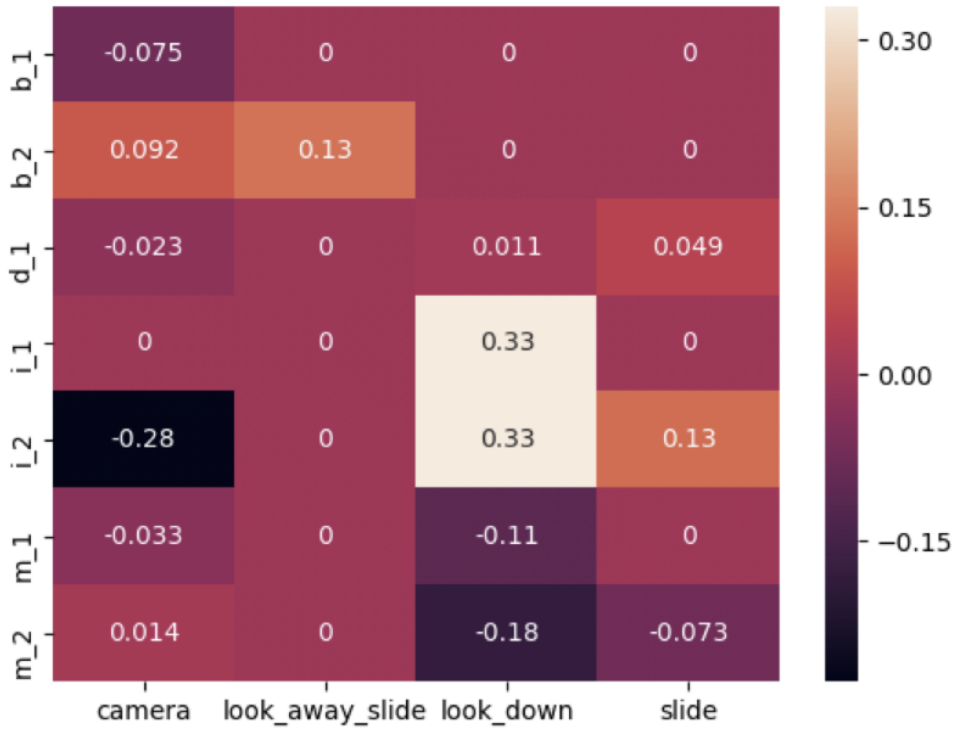

Figure 3.4: Normalized score of (FULL - DUAL) condition.

From the results, we can clearly see that the (iconic, look_down) pair has strong positive effects on the score. So look_down head pose enhanced the iconic gesture.

To compare the effects of head pose and hand gesture separately, Figure 3.5 shows the score sum. From the results we can see look down head pose has higher score. Look at camera has negative effect on score. For the hand gestures, iconic has positive effect on score. Metaphoric hand gesture has negative correlation with score.

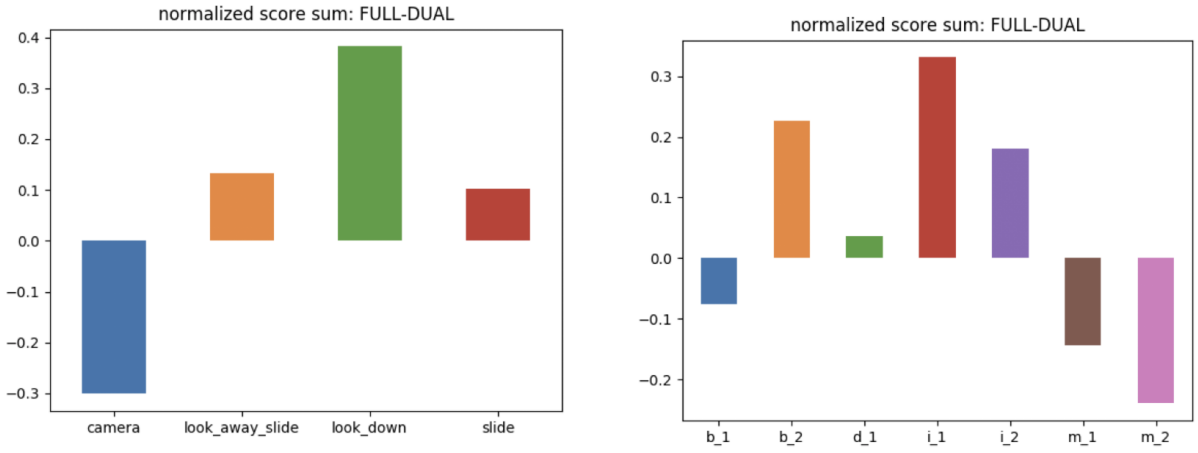

Figure 3.5: Left: score sum for head pose. Right: score sum of hand gesture.

4 Eye Tracking Analysis

4.1 Head Pose and Eye Tracking Correlation

In order to investigate the head pose effect on students' attention, we did research on the correlation between students' eye tracking and speaker's head pose. Figure 4.1 shows one student eye tracking data, we label it using speaker's head pose type.

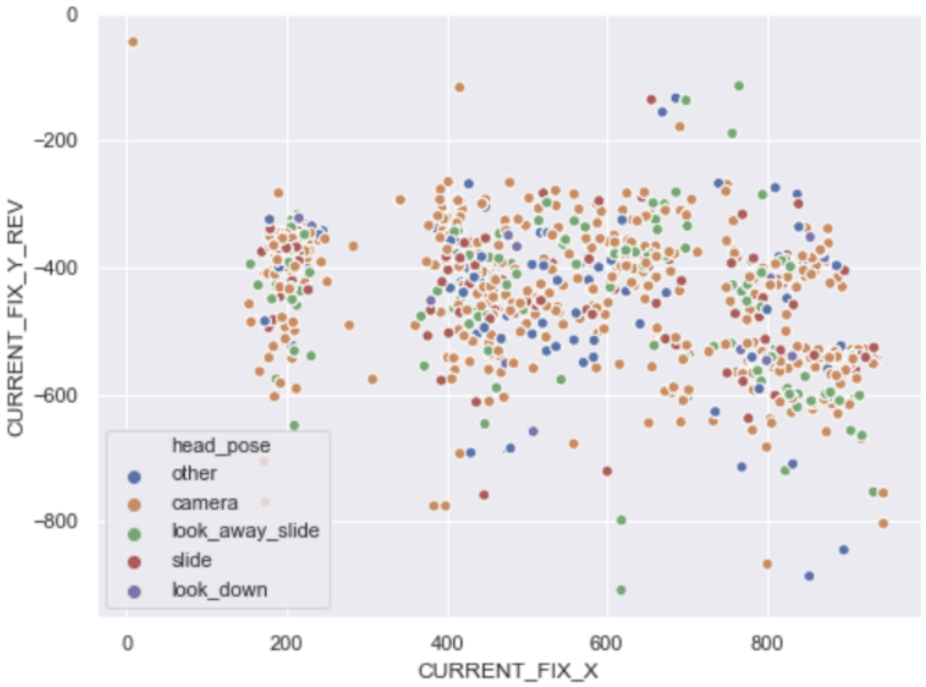

Figure 4.1: Example of eye tracking data. Data points are labeled using speaker's head pose.

To show the effect of head pose, we plot the eye tracking x distribution for each head pose (Figure 4.2). Where x-axis is the eye tracking x coordinate, y-axis is the eye x coordinate distribution for that head pose type.

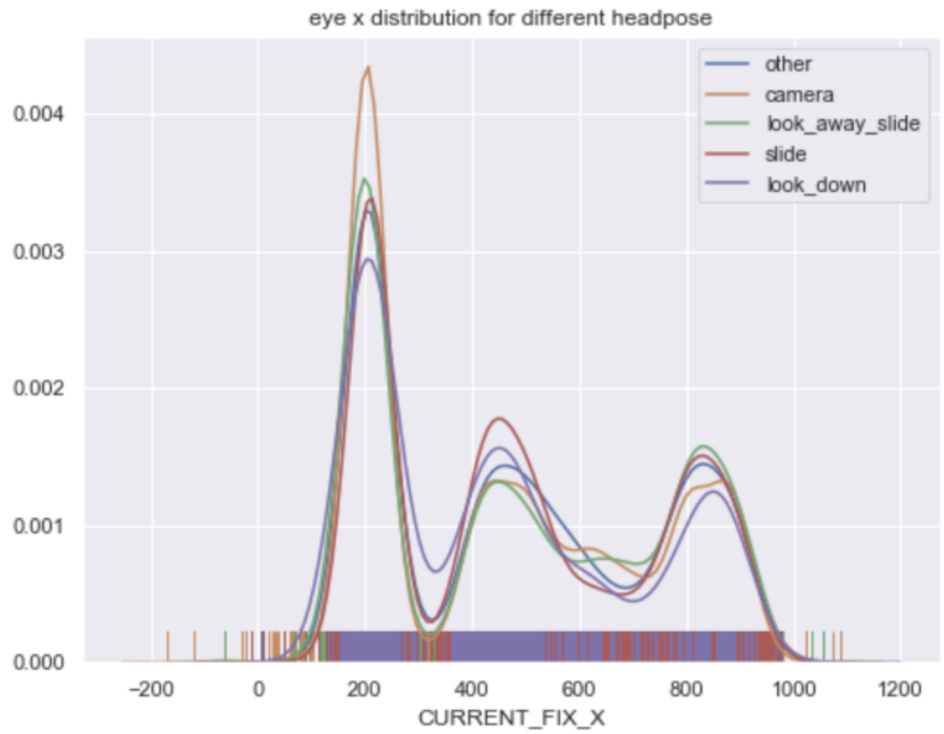

Figure 4.2: Distribution plot of eye tracking x data. Perspective video.

From the figure, we can see students focus on three sections: speaker (x=200), slides text (x=500) and slides image (x=900). By comparing the distribution line for different head poses, it shows the students attention (eye tracking) follows the speakers' head pose direction. When speaker look at camera, students also look at speaker. When speaker look at slides, student will also look at slides text. When speaker look away from slides, student will have lower attention on the slides text.

4.2 Eye Tracking and Score Correlation

We also investigated the correlation between eye tracking and scores. Figure 4.3 shows the distribution of eye tracking x for correct and wrong answer.

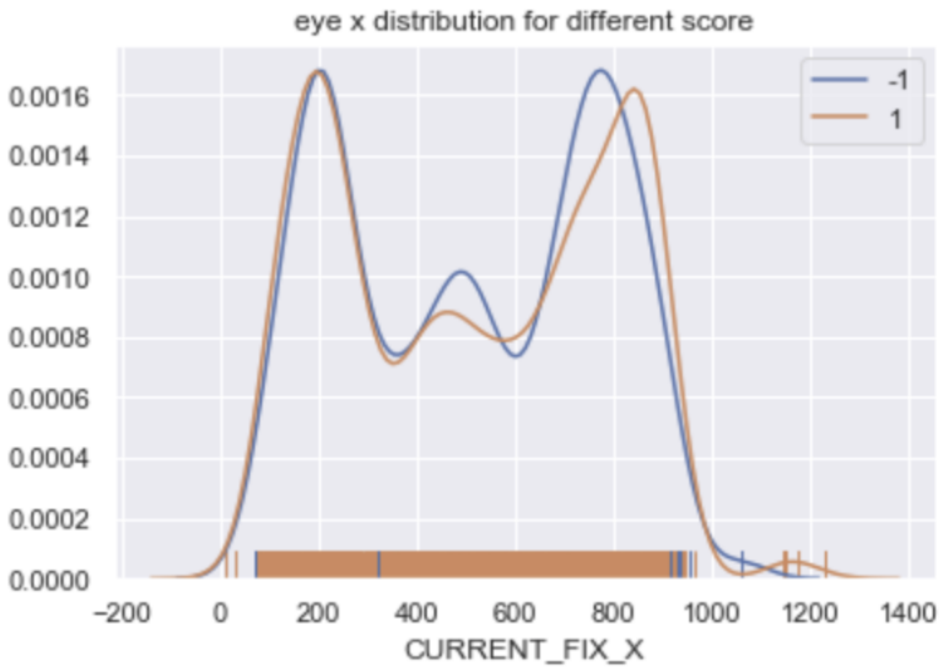

Figure 4.3: Distribution plot of eye tracking x data for correct (1) and wrong (-1) answer. Perspective video.

It shows that students will have relatively lower score if the them focus more on slides text. So slides text may distract students attention and students may missing some information in video.

4.3 Head Pose, Eye Tracking and Score Correlation

To investigate the correlation between speaker's head pose, student's eye tracking and student's score, for each head pose, we plot the distribution for wrong (-1) and correct (1) answer. Figure 4.4 shows the distribution of look at slide and look away from slide head pose. From the plots, we can see students get higher score if students follow the speakers' head pose direction.

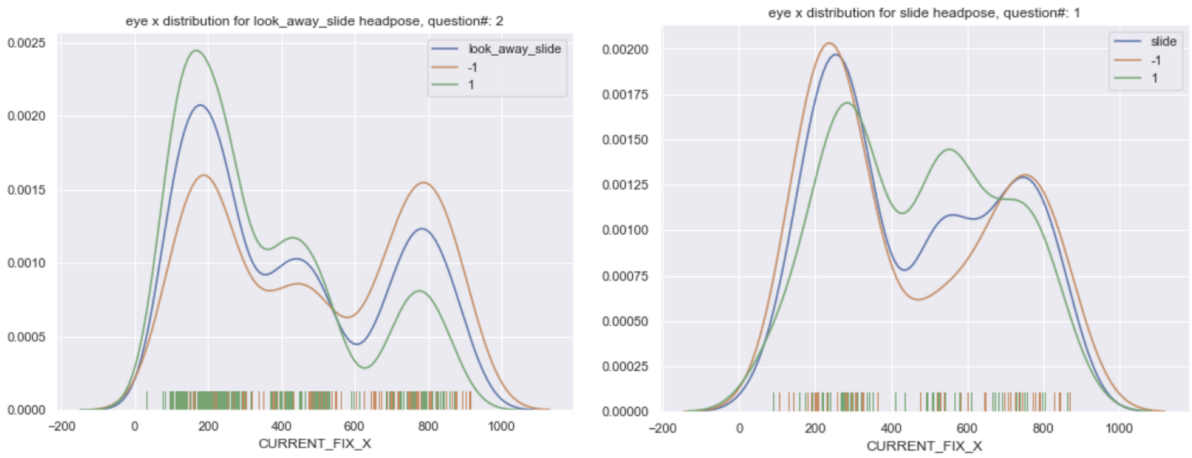

Figure 4.4: Perspective video. Distribution plot of eye tracking x data for slide (right) and away_slide (left) head pose type. Green and yellow line is the distribution of student who answer question correct and wrong respectively. Blue line is the average of correct and wrong. Student get correct answer if student follow the head pose direction.

5 Visualization Tool

We also developed a visualization software tool for researchers to explore different type of data. The code is Python based. Pyqt5 library is used for visualization. Based on previous version [4], the new version include dynamic plot for skeleton, eye tracking. The new version also let user to adjust parameters in real-time. Figure 5.1 shows the UI of the new software tool. Red circles are the eye tracking data. The larger the circle, the longer the student eye focus on that point.

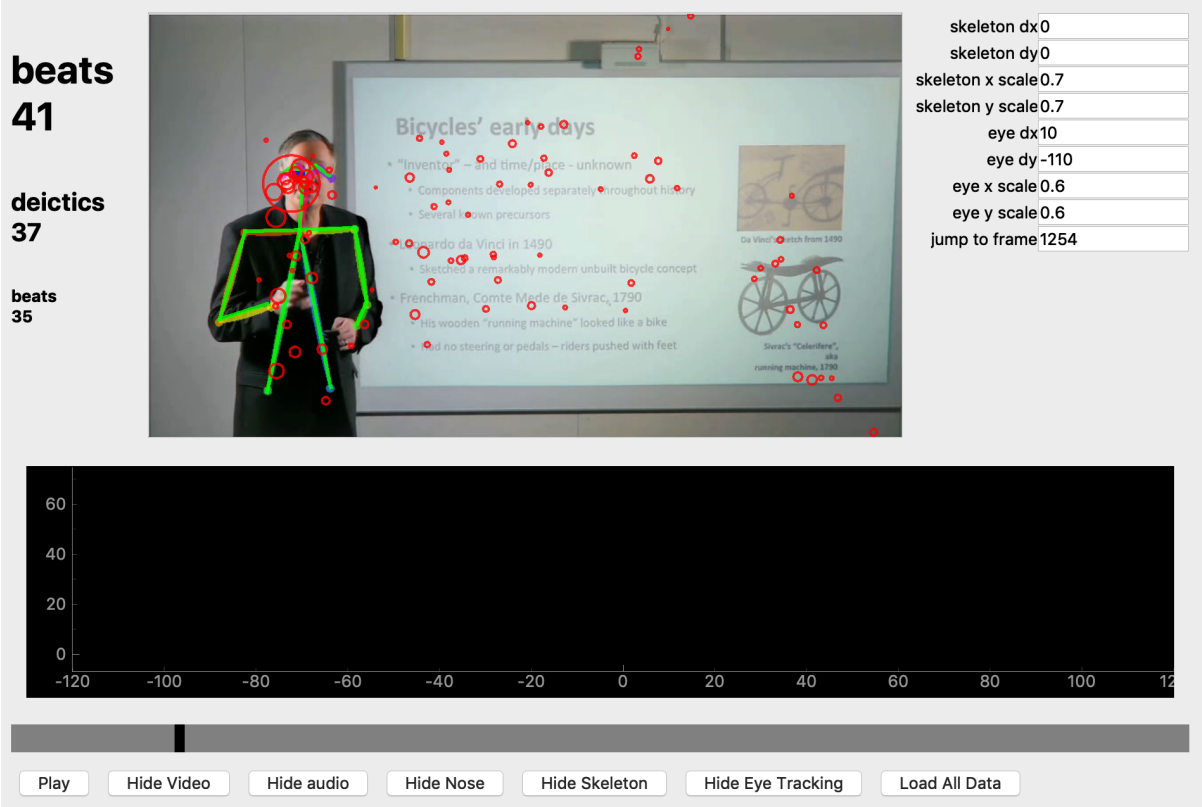

Figure 5.1: Visualization tool user interface.

5.1 Skeleton and Eye Tracking

The main update of the new version are skeleton and eye tracking plots. Both skeleton and eye tracking are dynamically plotted. User can show or hide plot by click the buttons on UI. Skeleton data and eye tracking data use different coordinate system. In order to synchronize the coordinate system, user can tune the coordinate parameters. For both skeleton and eye tracking, each has 4 different parameters: dx, dy, x scale and y scale. It will map the original coordinate (x, y) into new coordinate (x', y') using equation 5.1.1:

$$\begin{aligned} x' &= (x + dx) * x\_scale \\ y' &= (y + dy) * y\_scale \end{aligned} \tag{5.1.1}$$

By tunning the skeleton parameters, we find when parameters (dx, dy, x scale, y scale) = (0, 0, 0.7, 0.7), our dynamically plotted skeleton matches well with the speaker's skeleton in video. For eye tracking, (10, -100, 0.6, 0.6) is a good parameter for all three videos.

5.2 Usage Instructions

The visualization tool code is at GitHub repository:
`https://github.com/hyu2707/HeadPoseClustering/tree/master/python`
To use the visualization tool, user needs to install Python 3 and some other required python libraries first. Then follow those steps to run the tool:

1. Set data path: modify the data path in `visproj/qt5_structuralize/config.py`
2. Run `visproj/qt5_structuralize/main_widget.py` to bring up the display interface
3. Click "Load All Data" button to load data
4. Click "Play" button

User can click "Hide" button to hide certain plots. To jump to certain frame, user can input frame number at "jump to frame" input box or drag the slider.

6 Conclusions and Future Works

In this work, we show that head pose can enhance the effect of hand gesture. Especially look down head pose can enhance iconic gesture. Look at slides head pose also can enhance deictic and iconic gesture.
We also demonstrate that student eye will follow speaker's head pose direction. When speaker look at slide or look away from slides, student eye will also look at slides text or away from slides text. We also show student will get lower score if student spend relatively more time on slides text. By investigating the correlation between head pose, eye tracking and score, we find that student who follow speaker head pose will get higher score.
The visualization tool developed in [4] is also updated. Now the software tool can plot eye tracking and skeleton data in real-time. The x, y coordinate mapping function of eye tracking and skeleton can be easily adjusted in UI, which makes the data calibration process easier.
There are a few things can be investigated next based on current work:

1. More samples: From figure 3.2, we can see there are lots of (head pose, hand gesture) pairs have zero count. If we can get more samples of deictic, iconic and metaphoric gestures, I believe we can get more information about how head pose enhance the gesture.

2. Analyzing eye tracking of other two videos: Right now, we only used "Perspective" video in eye tracking analysis. If we can map the coordinate of other two video into the same coordinate as "Perspective" video, we can get more eye tracking data.

A Supplemental Information

All scripts and code used in this work can be found on GitHub:
`https://github.com/hyu2707/HeadPoseClusterin`

B Preprocessing - process_csv.py

```
import re as re
import matplotlib.pyplot as plt
from matplotlib.pyplot import cm
from datetime import datetime
import numpy as np

p_t = re.compile('(\d):(\d+).*')
p_ges_b = re.compile('.*\(.*(beats).*\).*')
p_ges_m = re.compile('.*\(.*(metaphoric).*\).*')
p_ges_d = re.compile('.*\(.*(deictic).*\).*')
p_ges_i = re.compile('.*\(.*(iconic).*\).*')

filename="ges_tr.csv"
#output_file="ges.csv"
f_out= open("counts_"+filename,"w+")
with open(filename) as f:
lis=[line.split() for line in f]        # create a list of lists
out_str = "time(s),beats,metaphoric,deictic,iconic\n"
f_out.write(out_str)
for i,x in enumerate(lis):              #print the list items
print("line{0} = {1}".format(i,x))
if(len(x)==0): continue
m_t = p_t.search(x[0])
cur_str = ",".join(x)
if (m_t is None):
print("skip:%s"%(cur_str))
continue

b_len, m_len, d_len, i_len = 0,0,0,0
```

```
m_b = p_ges_b.search(cur_str)
if(m_b is not None):
b_len = len(m_b.groups())
m_m = p_ges_m.search(cur_str)
if (m_m is not None):
m_len = len(m_m.groups())
m_d = p_ges_d.search(cur_str)
if (m_d is not None):
d_len = len(m_d.groups())
m_i = p_ges_i.search(cur_str)
if (m_i is not None):
i_len = len(m_i.groups())

t_min = int(m_t.group(1))
t_sec = int(m_t.group(2))
t = t_min*60 + t_sec

out_str = "%d, %d, %d, %d, %d\n"%(t, b_len, m_len, d_len, i_len)
f_out.write(out_str)
f_out.close()
```

References

[1] Ishan Manjani, Upper Body Poses and Gestures in Classroom Speaker Videos, Research Report, 2017.

[2] Ng, Chan Wah, and Surendra Ranganath. "Real-time gesture recognition system and application." Image and Vision computing 20.13-14 (2002): 993-1007.

[3] Haoyu He, "Speaker head pose and gesture correlation", Research Report, May 2018.

[4] Abhinav Sharma, "Visualization tool for project on what behaviours make speakers engaging to an audience", Research Report, May 2018.