Daniel Nkunga Scientific Notebook

Capstone Project: Using Artificial Intelligence to Read Lips
Instructor: Nicholas Seward

Notebook Start Date: August 22, 2023

Day 1: Tuesday, August 22, 2023

Daily Goal: Research face tracking programs, previous attempts at this project, potential languages for this project

Daily Introduction

Today was done in class. This is the first day of real research on this project. Mr. Sewards wants a minimum viable product as soon as possible and my project, especially the first method which is going to be by brute force, shouldn't take long to get up and running. All research today will be on how to start and run the program. Very little code if any will be written today. I would like to find a few potential facial recognition programs chosen, a desired language, and research done on a prior project done by the end of the day. Not in that order. Prior project research should probably come first.

Prior Attempts: Engadget

Source: Tarantola, Andrew. "AI Is Already Better at Lip Reading That We Are."
Engadget, 5 Oct. 2022,
www.engadget.com/ai-is-already-better-at-lip-reading-that-we-are-183016968.html.

Notes

- On average, humans can only guess ⅕ words correctly when reading lips
    - 2009 Study
- Forensic lip readers are able to determine 100 year old dialogue from silent videos with enough accuracy to even predict the accent the people are speaking in
- According to the CDC's Hearing Loss in Children Parent's Guide, a good speech reader might only catch 4-5 words in a twelve word sentence
- 2011 study in Oklahoma only saw about 10 percent accuracy in subjects
    - There is an additional quote here talking about outliers lying at 30 and 45 percent accuracy
- Today's state of the art machine systems are able to achieve over 95 percent accuracy
- Liopa: Lip reading app headed by Fabian Campbell-West
- Visemes: visual units of a person's speech such as their lip movements
    - Phonemes: audible units of a person's speech
- There are about three times as many phonemes as visemes
- Lipnet: Oxford Develops lip reading program developed by Yannis Assael
- Most visemes are hard to decipher without context
- The methods for understanding lip reading is generally the same despite the language

- Tonal languages are again harder to interpret but still possible to understand
- The difference in how humans read lips and how AI reads lips is stark because humans rely much more on context and getting the general meaning of the speech understood while AI focuses on turning lip movements into words
- Major obstacles in the lip reading industries are in a lack of intensive database
- Visual Speech Recognition (VSR) research is what this project will focus on
- Two approaches to building models: looking for the best possible architecture vs. using as much data to cover variations as possible
- Oxford-BBC Lip REading Sentences 2 is a database of thousands of spoken lines from the BBC
    - https://www.robots.ox.ac.uk/~vgg/data/lip_reading/lrs2.html
- Similar datasets exist for chinese, French, Russian, Spanish, and Cehzh

Summary

On average, when reading lips, humans can only guess with about 20 percent accuracy while AI can reach up to 95 percent accuracy. When reading lips, humans rely a lot more on context to get the general meaning of the sentence while AI tries its hardest to turn lip movements into words. Visual Speech Recognition (VSR) today is focused on trying to build the best architecture or using smaller databases. When using large databases, there are existing large databases of captioned speech to use.

Daily Conclusion

Today was productive but not a lot of goals were accomplished. Research was done by an article overviewing the VSR AI industry. It also went over some of the hiccups when it comes to training VSR and what methods are used to overcome them.

Day 2: Thursday, August 24, 2023

Daily Goal: Establish a MVP because Seward didn't like it and look into MediaPipe for Python face tracking

Daily Introduction

Today was done in class and the first day where we have the results of the first weekly assessment. Mr. Seward did not like the minimum viable project so we will need to rework it. Otherwise we will be looking into how MediaPipe works and trying to figure out.

Surrounding Research: MediaPipe

Source: Kukil. "Introduction to MediaPipe." *LearnOpenCV*, 1 Mar. 2022, learnopencv.com/introduction-to-mediapipe/.

Notes

- "Framework for building machine learning pipelines for processing time-series data like video, audio, etc"
- Google product that requires minimal resources compared to other Machine Learning frameworks
- Written in C++, Java, and Obj-C
    - Still can be written in Python
- Perception Pipeline is called Graph
- Consist of nodes called calculators (dots) connected by edges called streams (lines) where stream carry packages
- Media pipe solutions are pre-built examples based on specific models [should be trained]
    - Has Face Detection and Face Mesh
    - Solutions are available in C++, Python, JavaScript, Android, iOs, and Coral
        - Python has the third most amount of solutions
- Performance time optimization aer built in for the most part
- Its recommended getting started with OpenCV before getting started with MediaPipe

Summary

MediaPipe is a powerful machine learning framework developed by Google that has built in solutions for facial tracking. It can work with Python and generally seems kind of forward to run but its recommended you have a basic understanding of OpenCV before using it.

- This is the tutorial the article recommended for learning OpenCV

*MVP Pitch to Seward

- Research has already been done on how to use machine learning/AI to read lips and turn human lip movements into speech
- My project will be testing and comparing different ways to train an AI how to read lips (hopefully) in real time; if real time translation cannot be reached, it will be sued to caption videos
- MVP
    - Brute Force Method (BFM): give an AI hours of video feeds and scripts of what the people are saying and force it to come to its own conclusions on how to read lips

  - User Training Method (UTM): give an AI tracking of lips, eyebrows, and maybe a third thing [more surface level research is needed on how humans read lips] and have it learn to read lips with those inputs serving as its foundation
- Second Level MVP (these are pretty doable)
  - Training AI in the Human Way: using LLM, give the AI a word for context and train it to (more) predict what a person is saying based of their facial movements
    - This can be redone in the BFM and UTM
    - This would be great to display captioning YouTube videos based off their hashtags
  - Other methods based on the more ways regular humans and professionals read lips

Summary

Tl;dr, Seward's problem with the MVP was that it went too far. The MVP should be as simple as making an AI know when you're saying the word "the" or something. Full lip reading can come later as a further research of if you actually have time to take this project further, it shouldn't be the baseline.

Surround Research: Python MediaPipe Landmark Tutorials

Source: Sergio, director. *Facial Landmarks Detection | with Opencv, Mediapipe and Python*, Pysource, 14 May 2021, https://youtu.be/LF7Lgz4_lus?si=1P8Y78HrCubT9Y8h. Accessed 24 Aug. 2023.

Daily Conclusion

You need to downsize your MVP. The goal for now is to create an AI that is able to predict a few select words and not one that is able to predict the entirety of human language. MediaPipe is an AI training framework that has facial tracking built into it and would heavily speed up the process of training the AI. There are many tutorials on how to use MediaPipe but the one linked is only 20 minutes and has source code linked in the description so. . . :P