NOV 22ND 2024

Montgomery AI

Revised Project Proposal

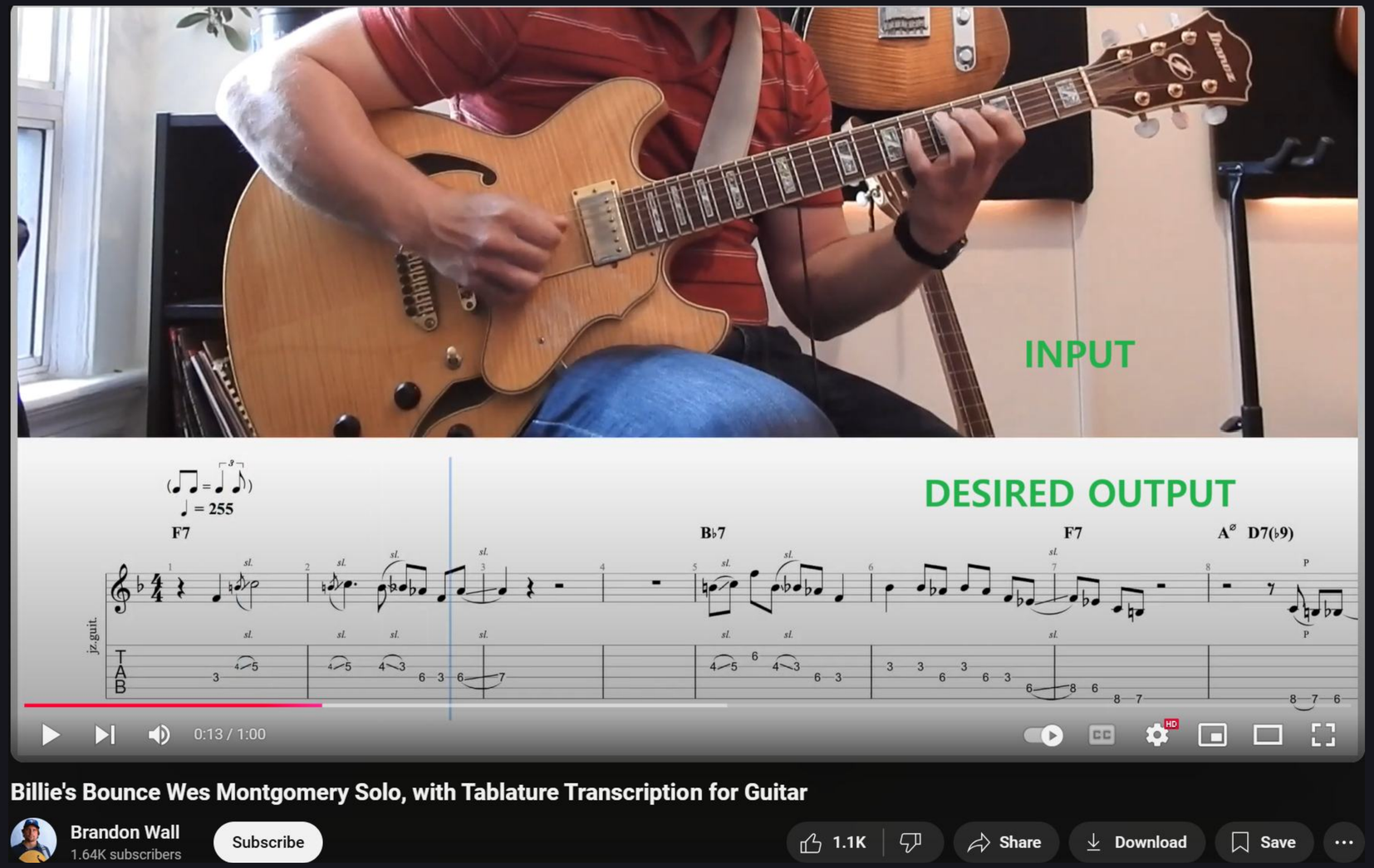
Sean Kim / EE P 596 Computer Vision

Montgomery AI is an multi-stage system that automatically transcribe guitar tab scores from images/videos of guitar performances

Learning a skill alone is not easy, especially in music. Can AI help?

More people are trying to learn instruments alone but the challenge of replicating what they see can be overwhelming. Automatic transcription can save time for learners.

OVERVIEW



Do this automatically for ANY guitar playing video on YouTube

Generate a tab score for Sweet Child O' Mine by Guns and Roses intro riff

Here's a simplified guitar tab for the iconic intro riff of "Sweet Child O' Mine" by Guns N' Roses:

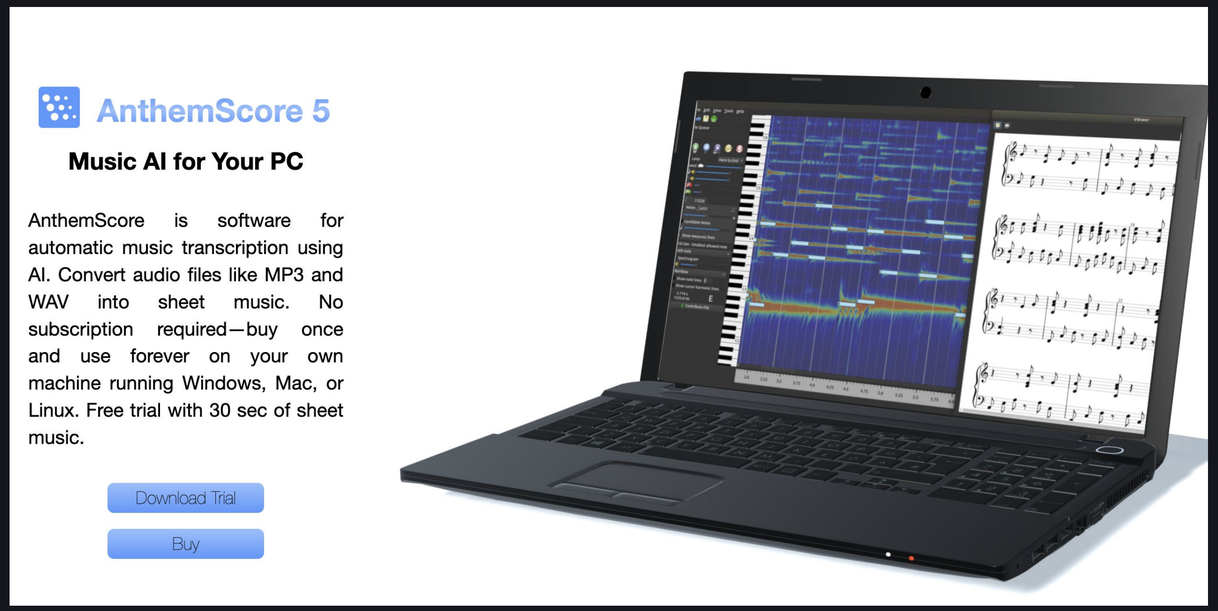
plaintext Copy code

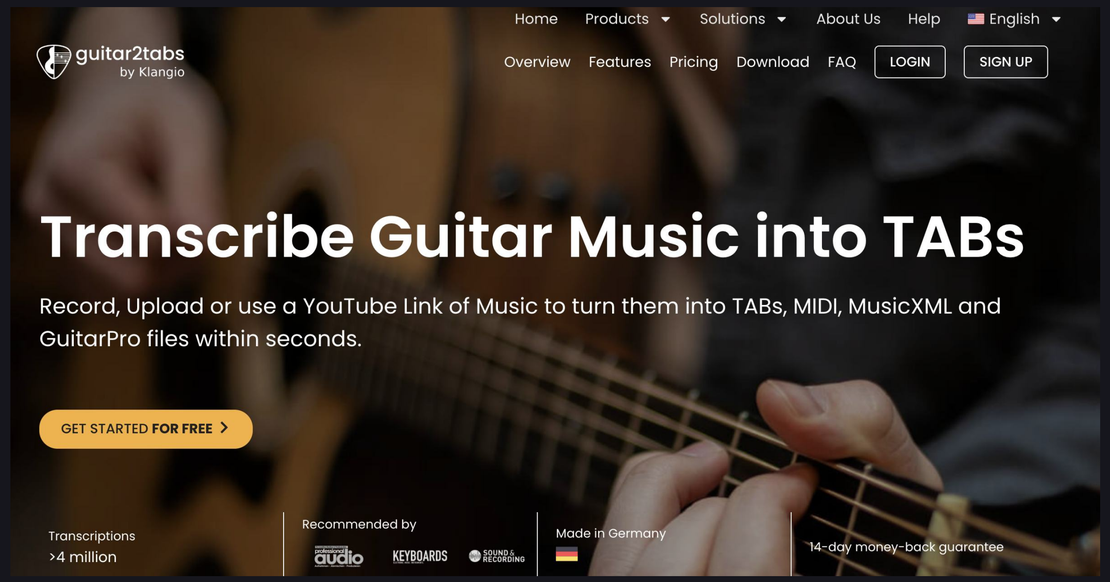
```
e|--------------------------15----14-------------------------|
B|----15----14----15-------------------15----14----15----14--|
G|-------14----14----14-------------------14----14----14-----|
D|-----------------------------------------------------------|
A|-----------------------------------------------------------|
E|-----------------------------------------------------------|

e|--------------------------15----14-------------------------|
B|----15----14----15-------------------15----14----15----17--|
G|-------14----14----14-------------------14----14----14-----|
D|-----------------------------------------------------------|
A|-----------------------------------------------------------|
E|-----------------------------------------------------------|
```

Text-based LLM hallucinate often when generating a score

(Some notes/numbering are correct but the order is completely off)

PRIOR ART



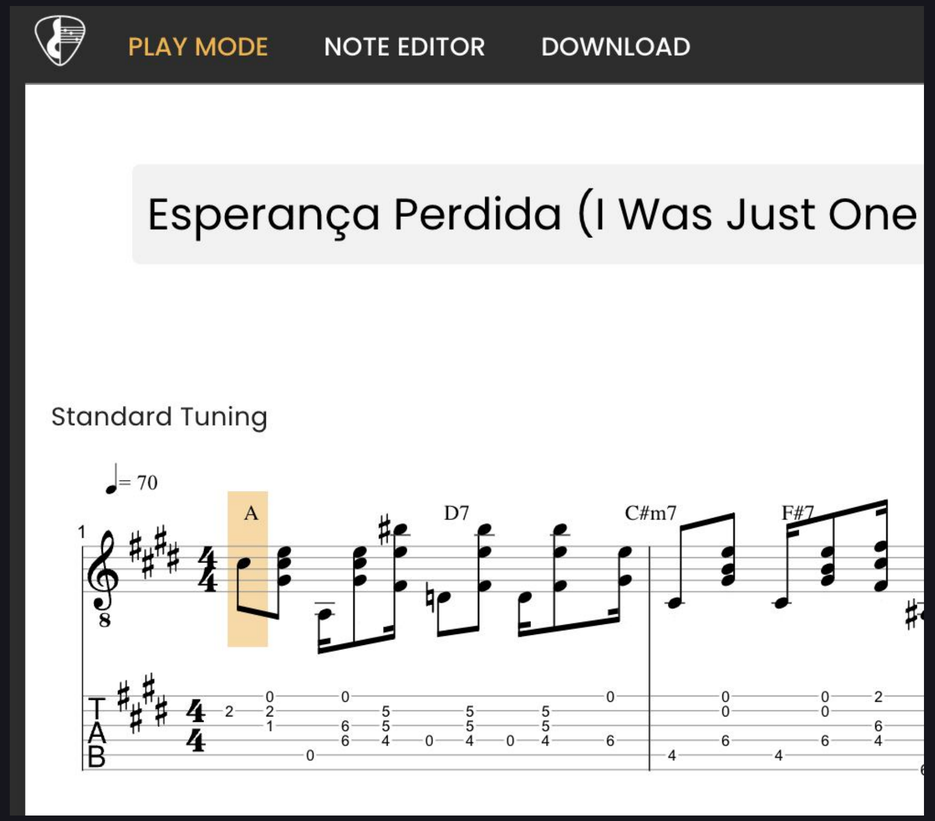


Existing Automatic transcription software are mostly audio based

Why audio-based doesn’t work

INPUT

OUTPUT



Notes are correct, but the finger position is wrong!

BASELINE

Compare the accuracy against-

1. Audio-based software: Guitar2tabs
2. Text-based LLM: ChatGPT (Ask to generate a tab score for a song)
3. Image-based LLM: LLaVa (Ask to transcribe a score given an image)

DATASET

1. Training: N/A (using pretrained)
2. Test set: Manually collect 100 images from transcribed videos on Youtube
3. (Maybe) Contact authors of relevant research papers (e.g. 3D Guitar Fingering Assessing System Based on CNN-Hand Pose Estimation...) and get the data

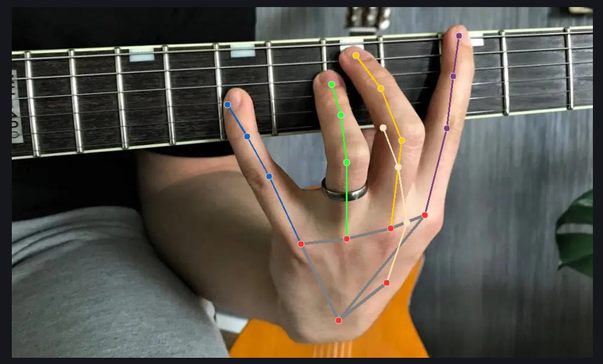
Proposed Methods

1. Identify fret-board using **segmentation**
2. Identify the edges inside the fret-board using **canny edge detection**
3. Identify the orientation of the fingers using **Pose Estimation**
4. Use **projective geometry** to calculate possible notes played
5. Process audio input and output a note being played on which string

Proposed Methods (Diagram)

input

Segment fretboard



Pose estimation

Canny Edge Detector

Audio cue (pitch)

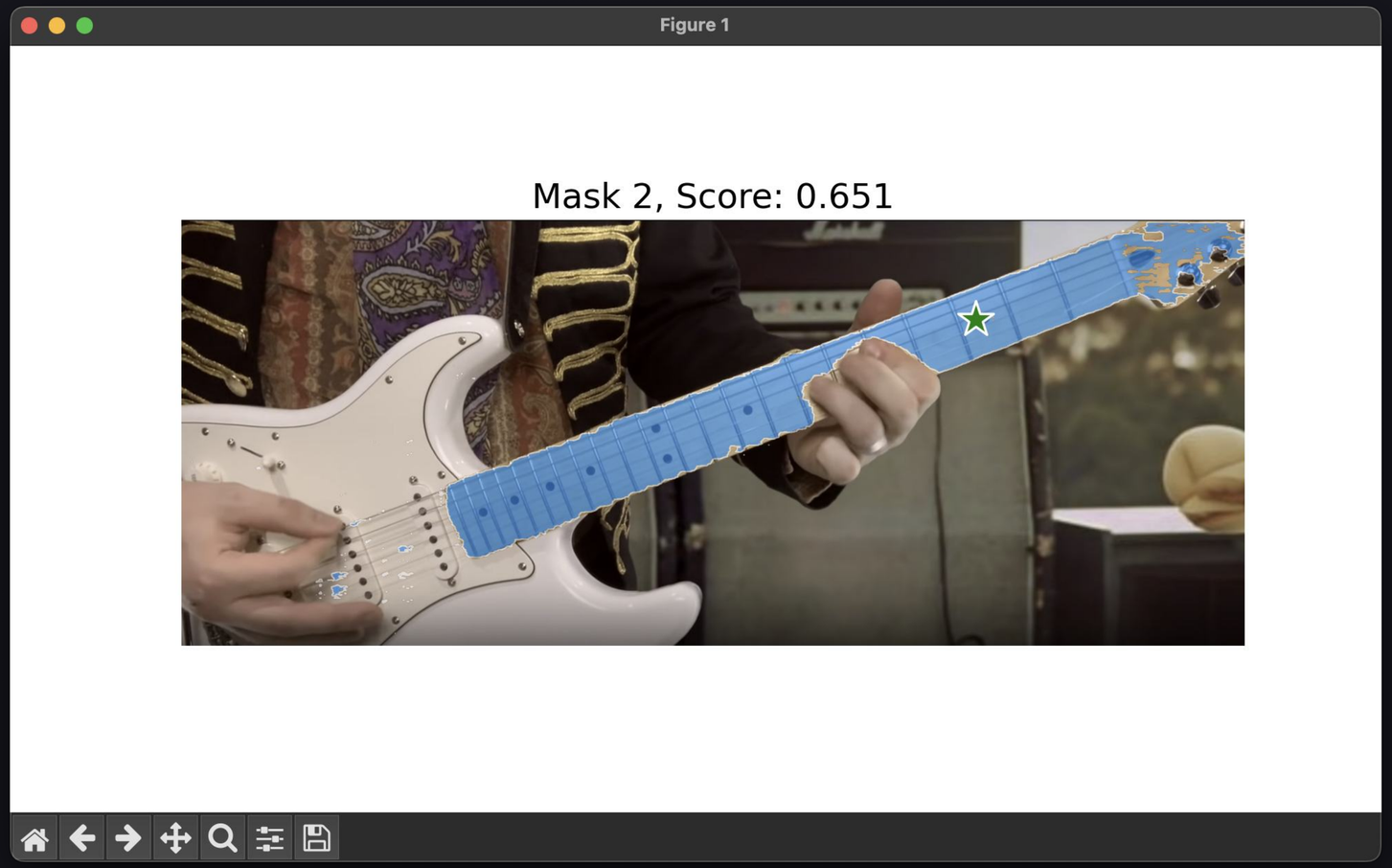
Possible Finger positioning

Tab Score

Milestones (by input type)

1. Image (Single Note)
2. Image (Multiple Notes)
3. Video (Single Note)
4. Video (Multiple Notes)
5. Image (Non-Orthogonal camera)
6. Video (Non-Orthogonal camera + Moving Guitar)

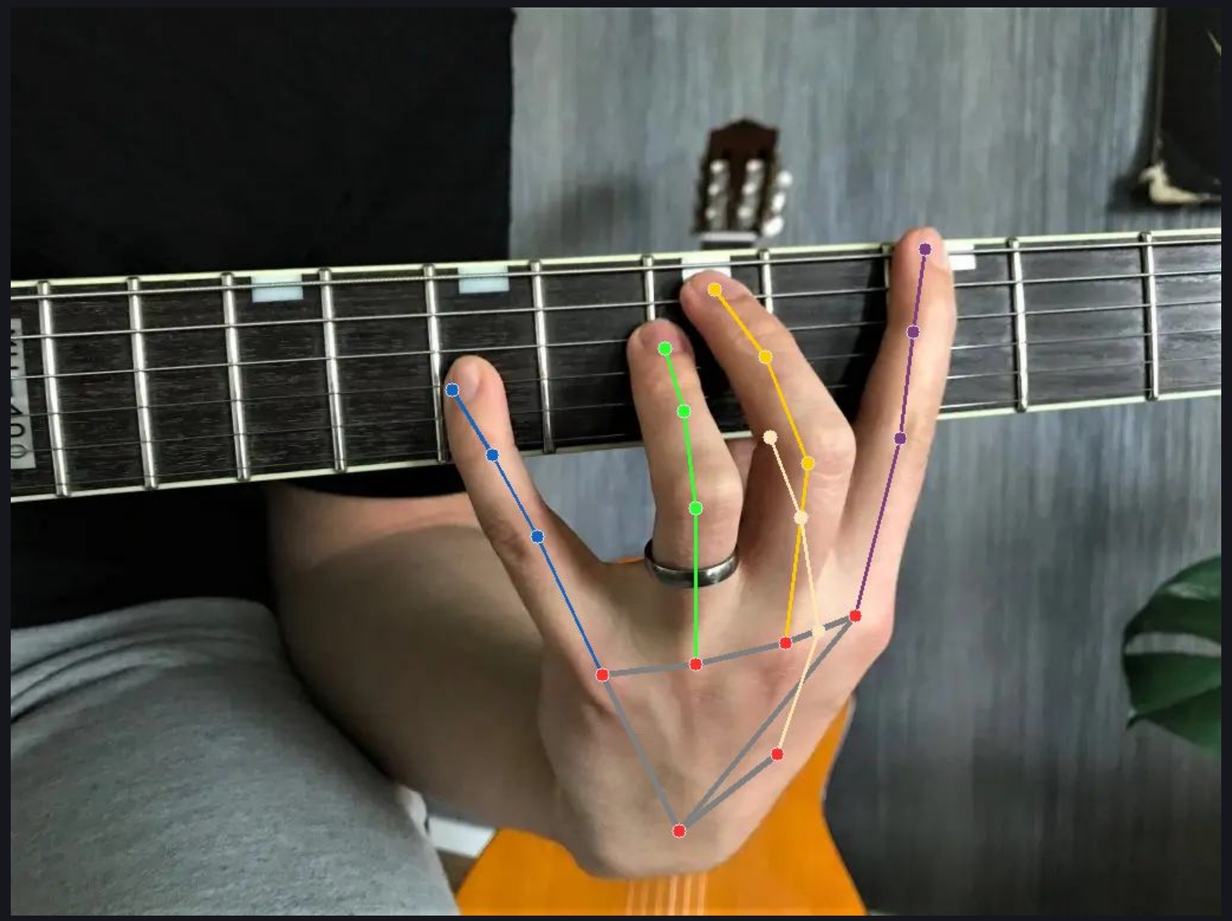
1. Fret-board Segmentation



Facebook’s Segment Anything 2.1

Works reliably. Reasonable inference time on Macbook M1 Chip

3. Hand Pose-Estimation



Google’s MediaPipe

Doesn’t detect hand sometimes. Need to investigate

AGENDA

Future Plan

1. Refresher on problem statement
2. Update on metrics
3. Review design proposal
4. Align on GTM

Challenges / Future Plan

1. Choosing best tools: Does Segment-Anything & Media Pipe give the best result? Is the inference time on Macbook M1 chip reasonable?
2. Choosing the correct mask: Segmentation gives multiple masks for a single point. How can we programmatically get the “fret-board” mask?
3. Canny edge detector: would a simple canny edge detector be enough to identify the fret board bars?
4. Pose estimation customization: would training MediaPipe further on a known hand orientation (e.g. hand orientation for a playing C chord) help?
5. Audio cue: Would pitch information be enough to make decision on which fingers are actually playing the note?