Small Project F12

Visual Reconstruction of Played Music

Computer vision solution to transcribe piano performances from silent video

By Ignacio Bascuñán - 10984708

Politecnico di Milano
Computer Science and Engineering
099993 - Image Analysis And Computer Vision

Project Overview

Challenge

Detect which piano keys are pressed from silent video

Approach

Classical CV pipeline using keyboard rectification and fingertip tracking

Output

Transcribed key presses converted to MIDI files

Keyboard Calibration Process

Reference Frame

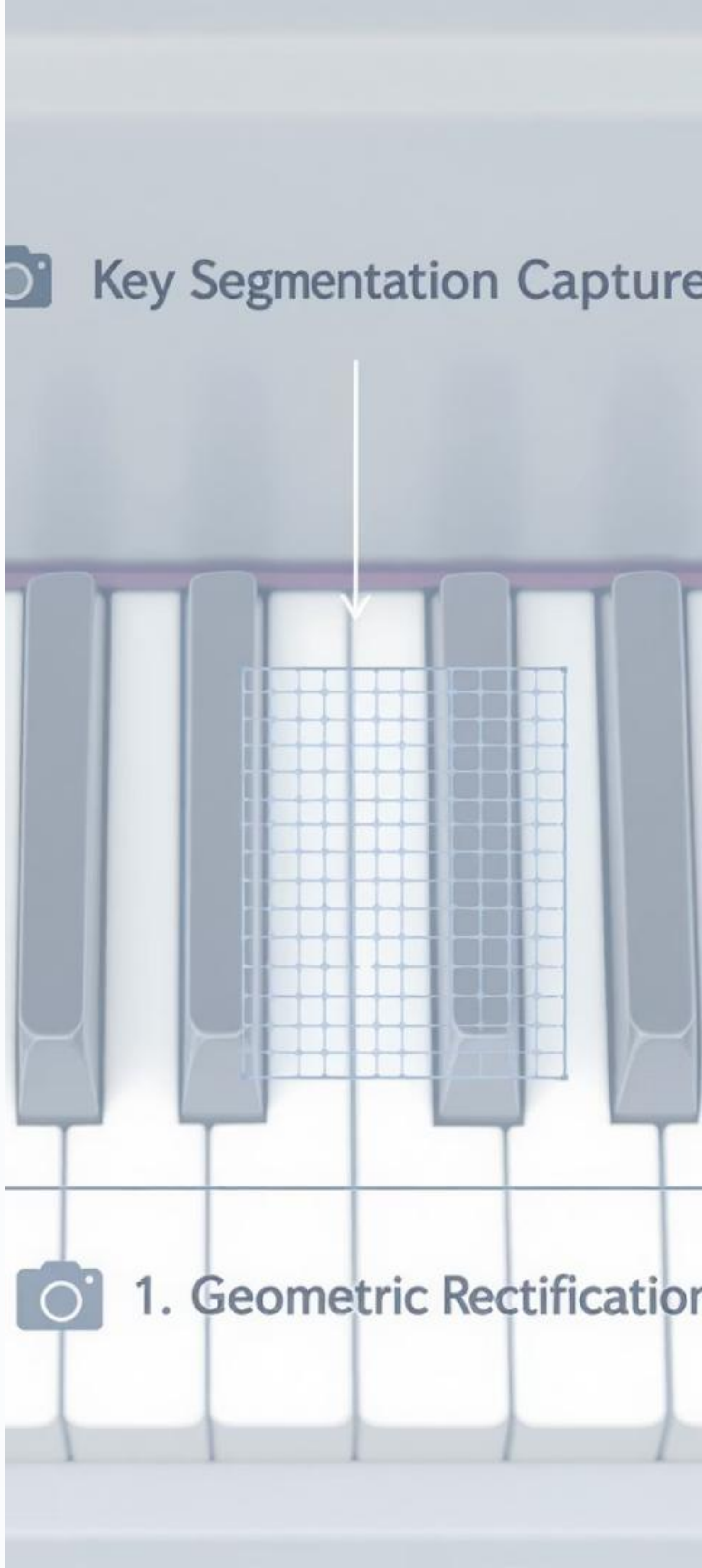
Capture still image without hands

Geometric Rectification

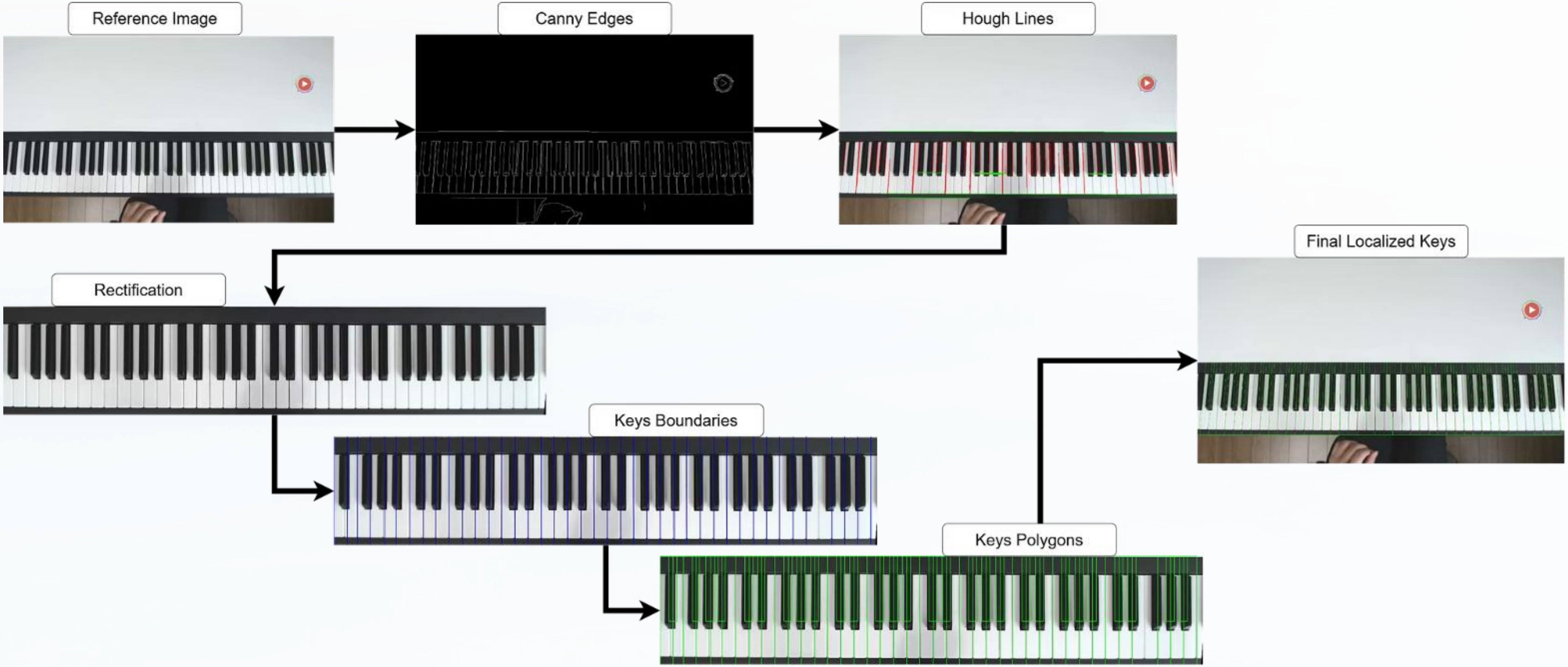
Extract boundaries using Canny and Hough transforms

Key Segmentation

Map key polygons in rectified domain



Keyboard Calibration Process



Key-Press Detection Pipeline

Hand Localization

MediaPipe tracks fingertips in real-time

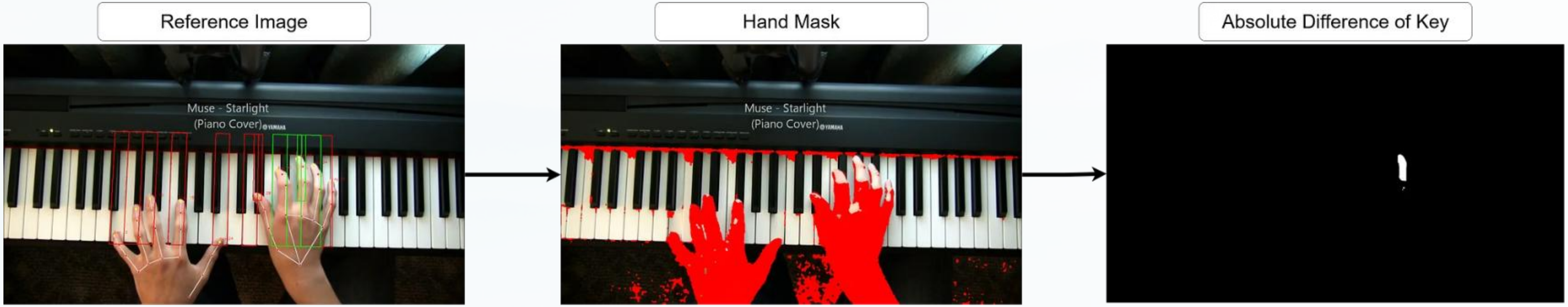
Candidate Mapping

Identify keys potentially touched by fingertips

Photometric Validation

Compare pixel differences to confirm press

Key-Press Detection Pipeline



Reference Image

Hand Mask

Absolute Difference of Key

Implementation Details

Tech Stack

- Python 3.10
- OpenCV 4.11
- MediaPipe 0.10

Key Components

- Homography mapping
- Background subtraction
- MIDI synthesis

Performance

- ~200ms per frame
- Real-time hand tracking

Experimental Results

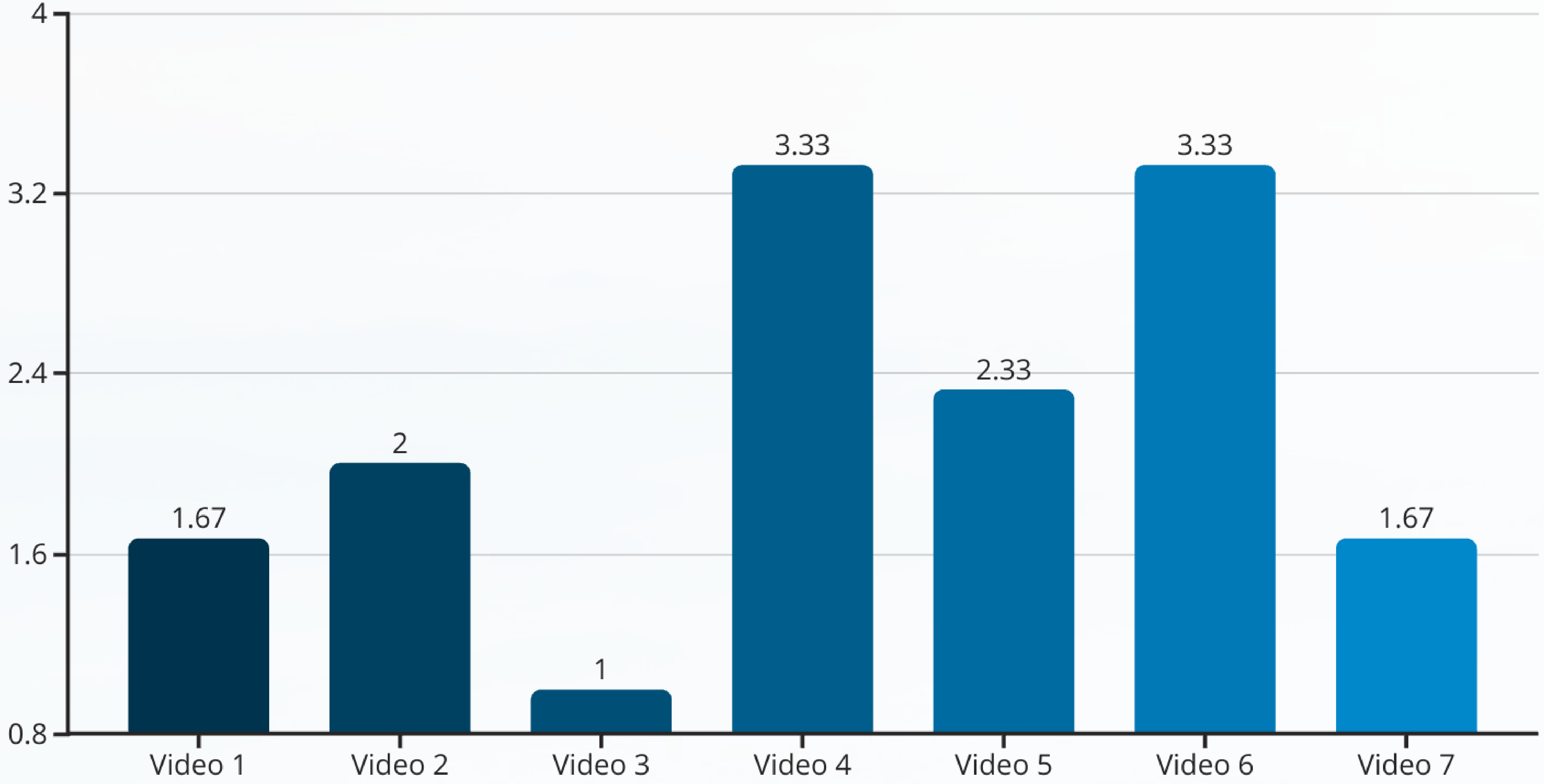
7 different videos for test cases

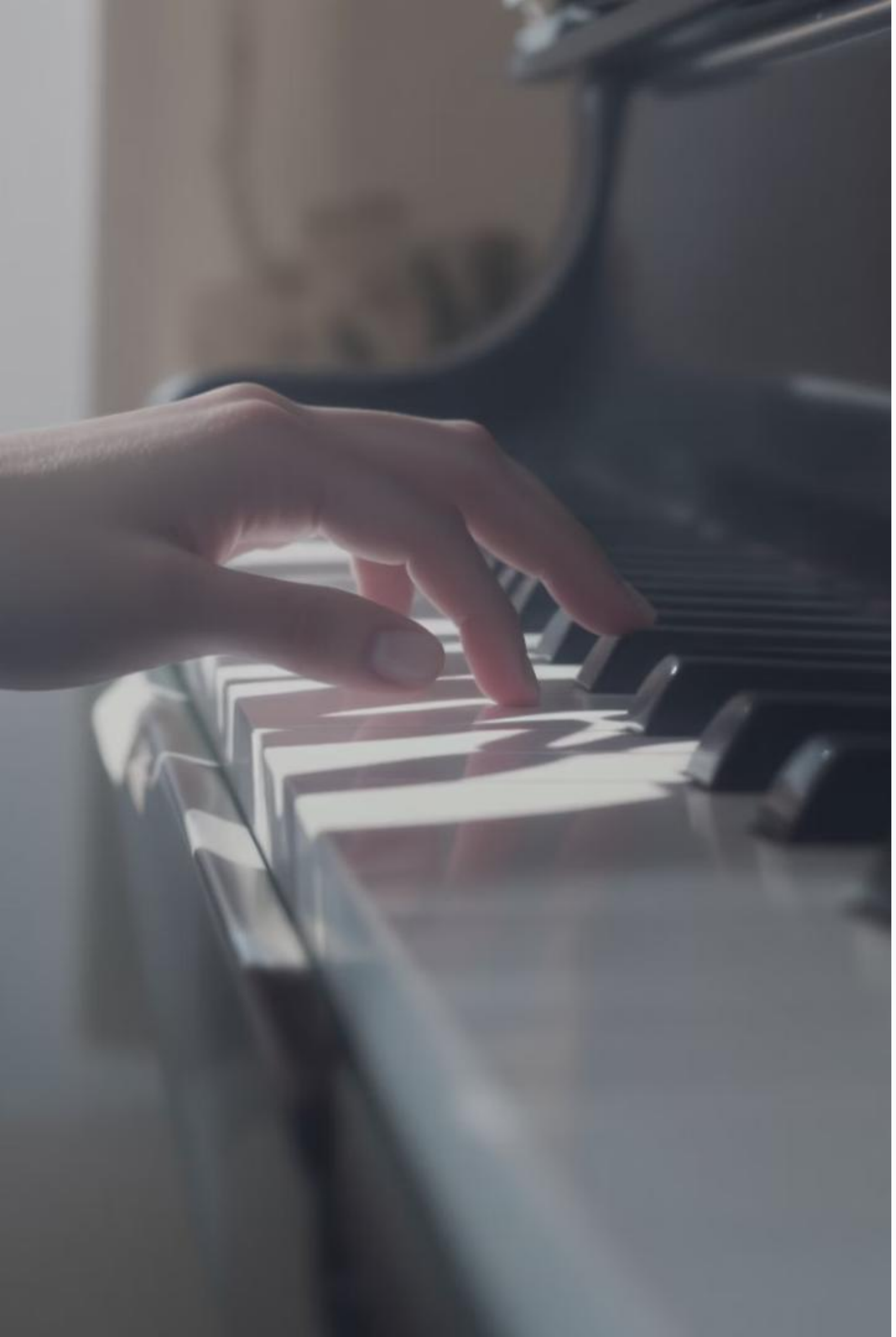
3 human testers

Evaluation Scale

1. Impossible to recognize
2. Some notes and rhythms seem to match the original song
3. The song is recognizable
4. The song not only is recognizable, but is pleasant to hear
5. Just some mistakes were made between the original song and the simulated output

Experimental Results (1 - 5)





Key Limitations

Lighting Issues

Sensitivity to reflections and shadows

Occlusion Problems

Cannot resolve overlapping fingertips

Black Key Detection

Poor accuracy due to minimal shadow contrast

Parameter Sensitivity

Requires manual calibration per video

Future Directions

Temporal Networks

Train on synthetic renderings

Multi-cue Fusion

Incorporate optical flow analysis

Adaptive Calibration

Auto-tune parameters per video

Self-supervised Learning

Fine-tune to individual performers

Demo: Youtube Video

Original Input

Produced Output

https://www.youtube.com/watch?v=P0kUMfg-dHE