Project Kojak - Using Computer Vision and Gesture Detection to Control a Smart Home

Background

There is a huge interest in computer vision and the field is very cutting edge. The development has been rapid, and has resulted in a number of open source projects that I could build upon and create a great working product. Those libraries include OpenCV, Yolo, Pillow, Keras, and others.
https://www.kaggle.com/gti-upm/leapgestrecog
https://www.kaggle.com/ranjeetjain3/deep-learning-using-sign-langugage

These datasets are of hand gestures - the first is of gestures, specifically, and the second is of American Sign Language signs. I'm most interested in the gesture database, but I include the ASL dataset as a backup since it is already ordered and arranged like the original MNIST dataset. I would default back to using this one if I ran short on time.

Process

I'd start by training a neural network on the data set, and my MVP would be a model that could accurately predict a particular gesture from the data set. I'd like to be able to do this in real time, though, which would take some additional work. So, this would include bringing in a library like Yolo to accurately identify a person and their "hand."

Hurdles/Benchmarks

1. Train NN on MNIST model, get it working with some reliability
2. Determine which library to use to be able to take input from a webcam, process it and include it as input for the model
3. Use the library to accurately identify the background, so it can be zeroed/blacked out
4. Possibly would need to make the model inputs (dataset and webcam images) black & white so they can correspond nicely with the inputs
5. Find a reliable way to process the real time images, enter them as input in the model, and get back a result in a relatively small amount of time.
6. Use the output of the model's prediction to ring up an API and turn on lights or music

Concerns/Known Unknowns

It's obviously a big project, and I don't have much experience with computer vision to date. Still, there are a lot of tutorials online, and I'm not the first one to have done it - there are Kaggle Kernels that I can read if I get stuck, etc.

It could be difficult to make the video input less specific about things like:

- Where the hand in the frame is located
- Size of the hand
- Distance from the camera
- etc.

It could be difficult and expensive to train the model from scratch - I'll look into whether there are pre-trained models or not. If not, I've always got AWS and Google Cloud credits.

Overall, though, I'm excited about the project and learning about the tools that are used in the field.