During our previous iterations, we attempted to implement our project using Detectron2 and LSTM architectures. However, while using Google Colab as a runtime environment, we encountered multiple version/dependency incompatibility issues with PyTorch and OpenCV libraries.

After considering all other possibilities and options, we found that the CNN/RNN architecture would be the best fit for our problem statement.
In this architecture, CNNs are used in the first stage of the model to extract spatial features from the input frames.

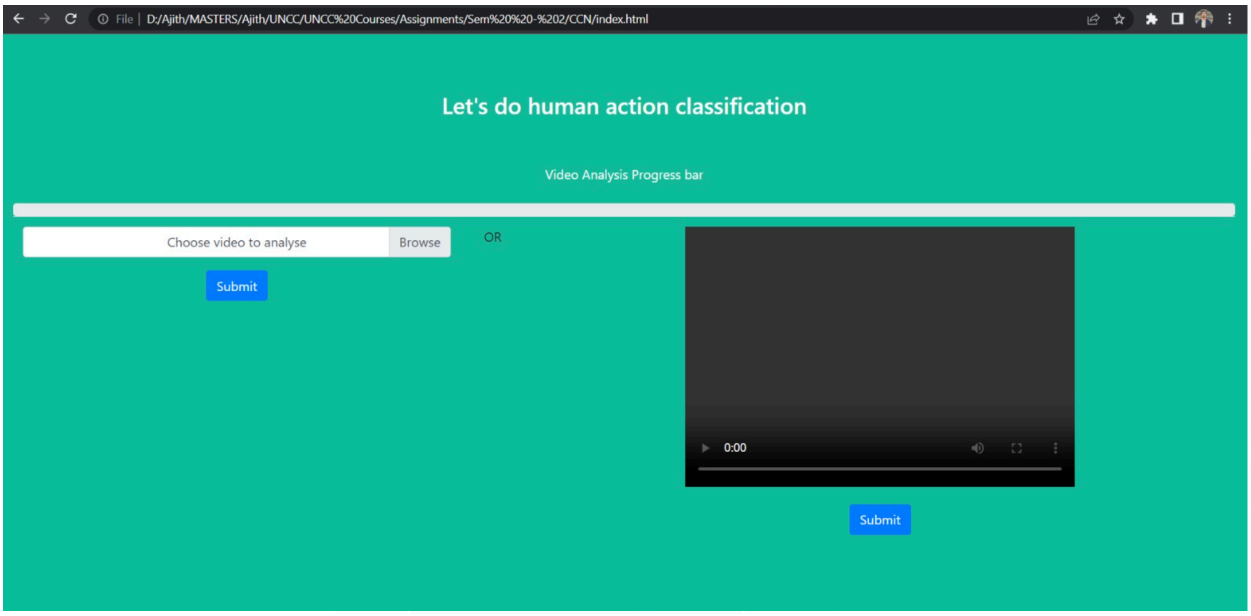
These spatial features are then fed into an RNN, which processes the features in a temporal sequence to capture the dynamics of the action being performed. The RNN is used to learn the temporal dependencies between the frames, which is important for recognizing actions that span multiple frames.
The CNN extracts features from each frame, and these features are then fed into an RNN. The RNN captures the temporal dependencies between the frames and makes the final prediction about the action being performed. The CRNN architecture is a powerful approach to action recognition because it can capture both spatial and temporal information from the input frames.

Currently, our model can recognize several actions, including headbutting, clapping, finger snapping, push-ups, fixing hair, trimming, tying a bow tie, stretching an arm, counting money, making a bed, and sign language interpretation. However, we are facing a challenge in the model's low confidence ratio while detecting these actions. To overcome this challenge, we plan to explore different architectures and hyperparameters to improve the model's accuracy.

We will also consider increasing the training data and applying data augmentation techniques to enhance the model's performance. In addition, we aim to implement a more robust user interface that includes features such as video playback controls and progress indicators.

Overall, we are determined to continue our efforts to enhance the accuracy of our model and provide a better user experience. With the CNN/RNN architecture and a thorough exploration of different techniques, we hope to overcome the current challenges and achieve our goals for this project.

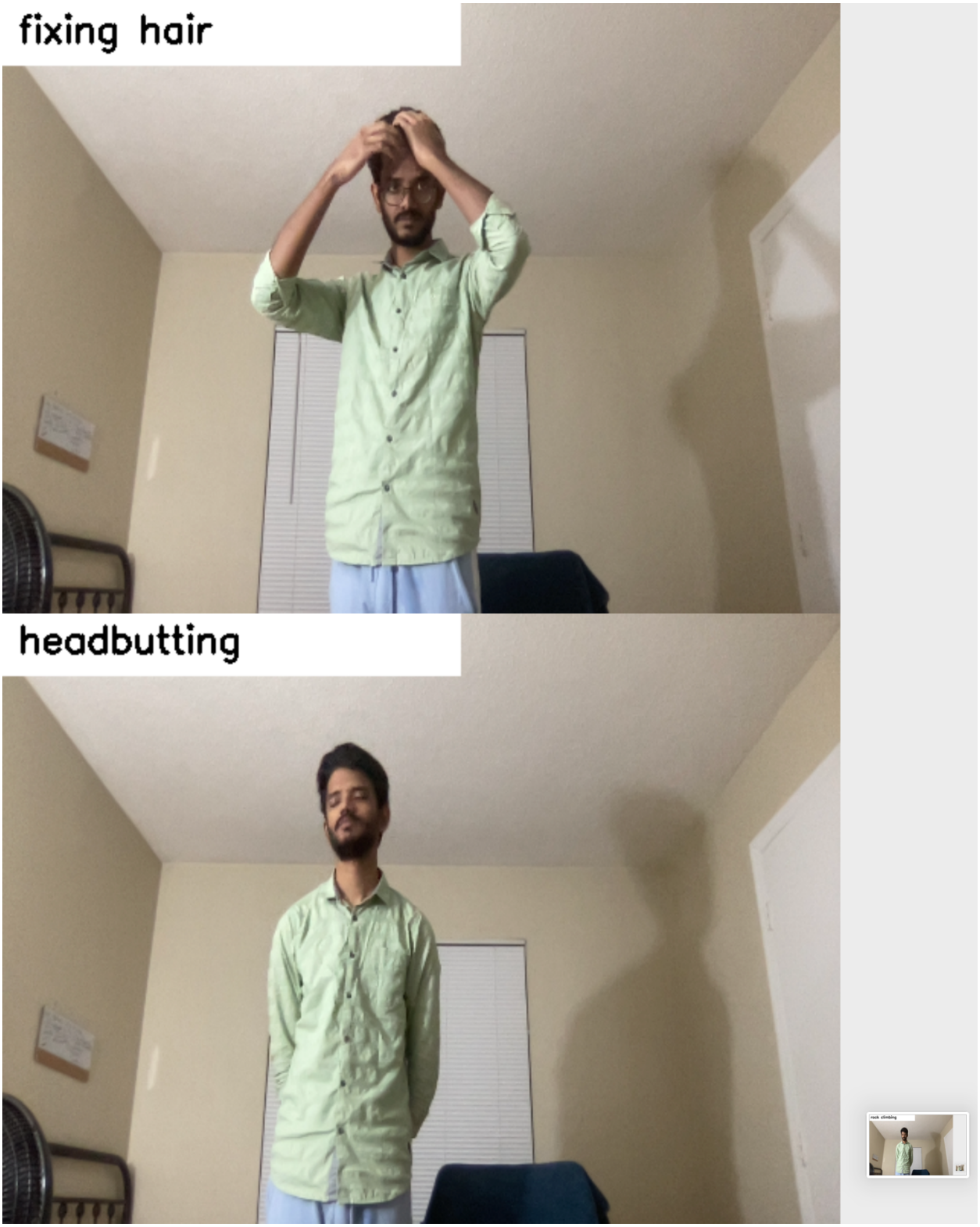
We have enhanced the UI by adding the progress bar to let the users know about the progress on action recognition.



Future action items:

- Increase the confidence ratio of the model by adding more data points.
- Integrate the model with the front end User Interface so that the user can upload videos directly.
- Add an extra functionality to the user interface so that the user can record videos directly from the device camera instead of uploading it.

Below are the snapshots of the results:



trimming or shaving beard

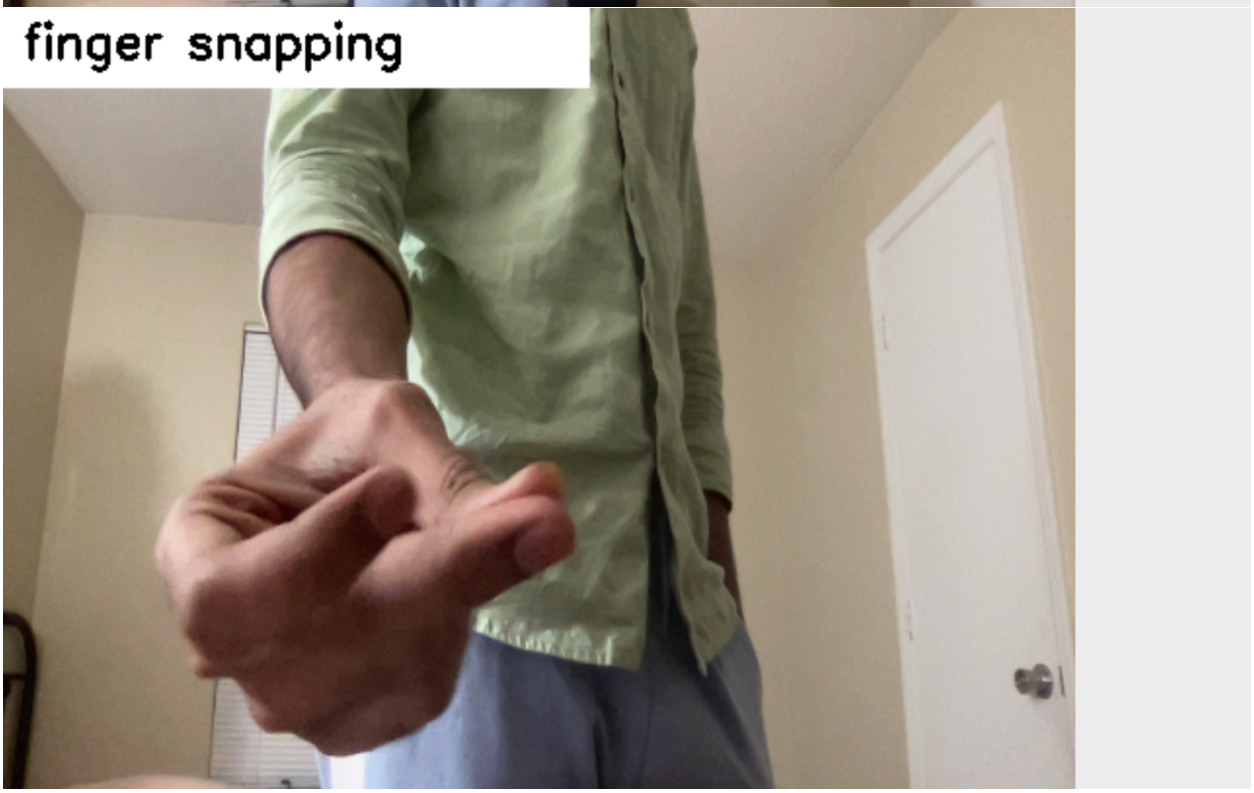
tying bow tie

stretching arm

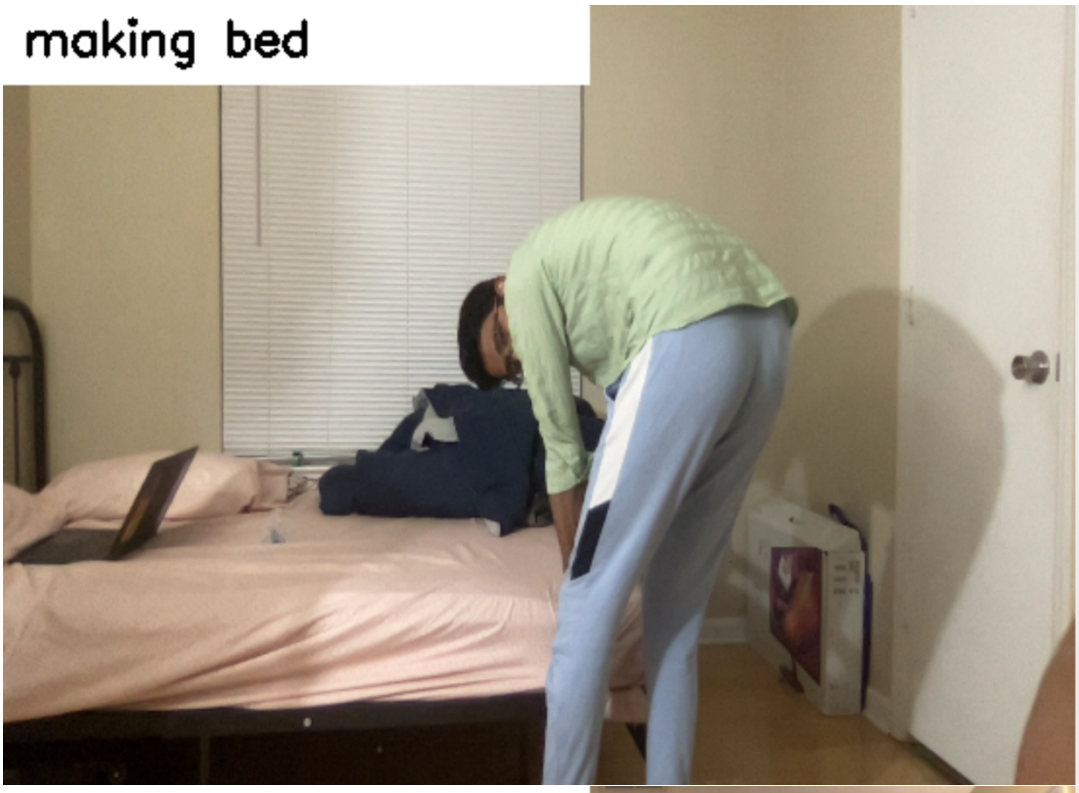
sign language interpreting

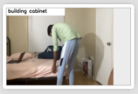
counting money

finger snapping

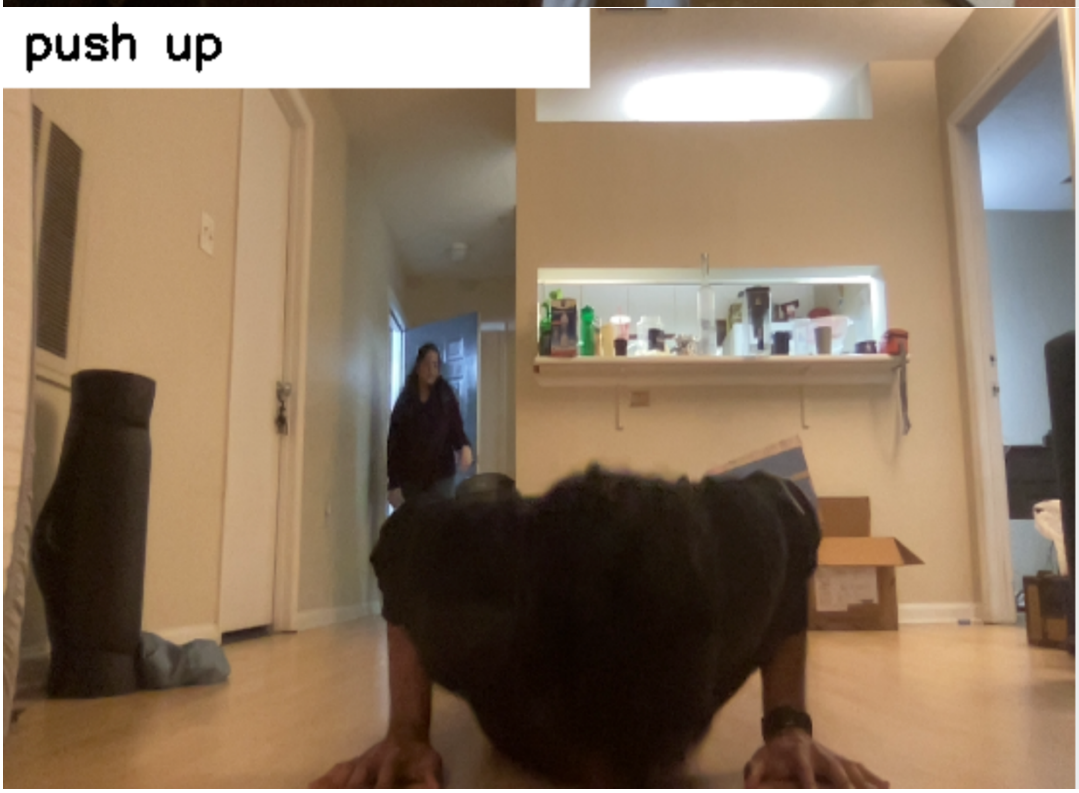


making bed





push up



squat