STEP 1.

How I captured imagery

I captured my images in the corner of my dorm room. The walls had the color of off-white. The only light source in the room is in the ceiling at the center of the room. Hence I chose to capture images from the corner since the corner had the least strong direct light. Strong direct light seemed to create noise in the binary image, which sometimes were picked up as a contour. I also wore dark long sleeve clothes. When I exposed my arm, the binary would show both my hand and my arm, and wearing dark long sleeves made the binary show only my hands. This is helpful because then I can compare just hand shape to hand shape, which made my accuracy increase. I used the webcam in my computer. The camera was about a meter away from the farthest part of the wall, with no object in between other than my hands.

Environment

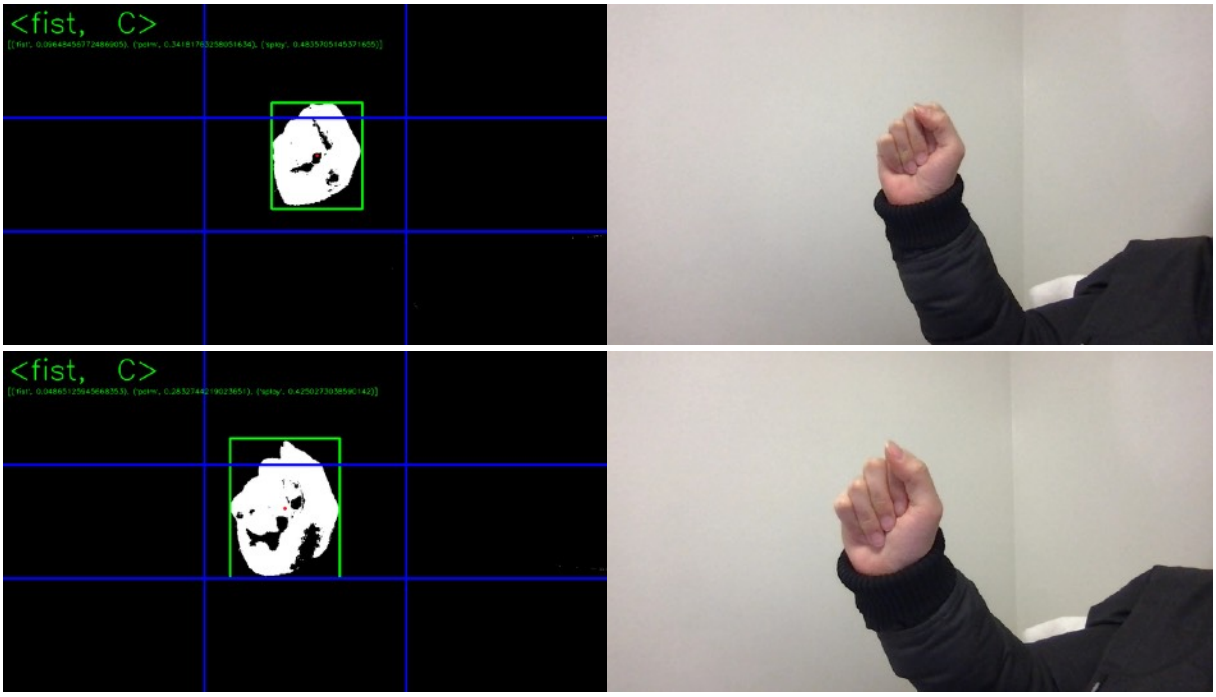
Hardware: MacBook Pro, 2018
OS: macOS Ventura 13.2
Language: Python (used PyCharm)
Packages:
- cv2(OpenCV): for taking, processing and saving images
- numpy: for setting skin tone range
- shutil: to cleanup folder before saving image files
- os: for accessing image folders

Library

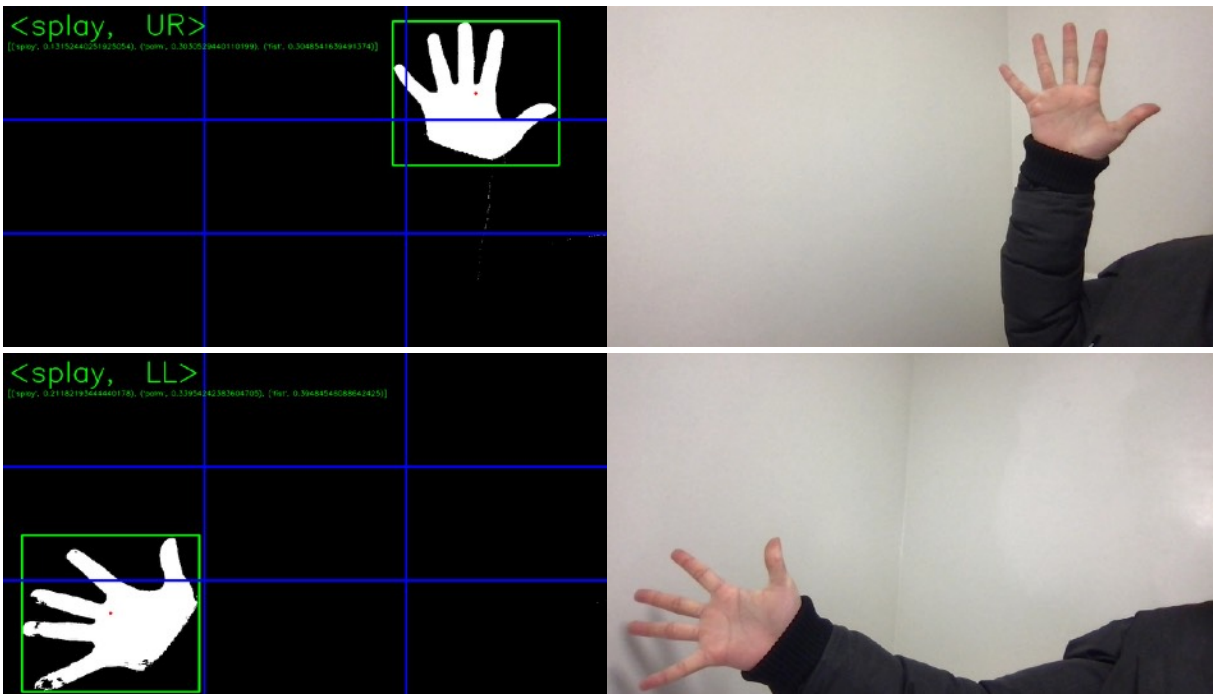
Library is named as Ref_Images (reference images).
There are 9 images in total: 3 for palm, 3 for fist and 3 for splay. It is in .jpeg format.
There are 3 for each shape because I intended to capture slight variations in rotation, light, hand position, etc. For example, one reference image for fist has thumb next to the fingers, and one reference image for fist has thumb on top of and crossing over the other fingers. Since the unlabeled image would be compared to all images in the reference image folder, this effort will help the system recognize different types of fists since different people make different fists.

STEP 2.

Two images of a centered fist that are accurately recognized and labeled as “fist, center” (true positive)

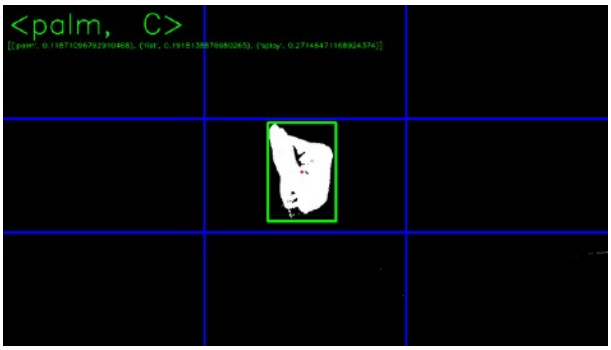


Two images of a cornered splay that are accurately recognized and labeled as “splay, upper right” [or some other corner] (true positive)



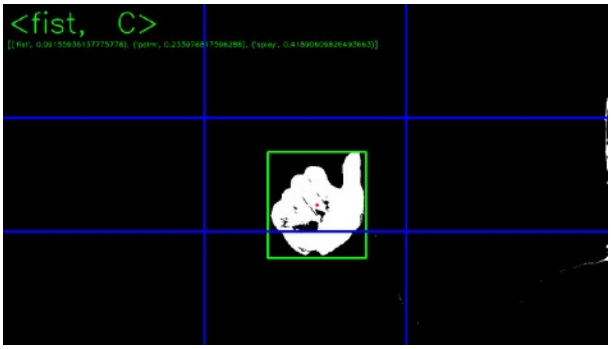
STEP 3.

One image of a centered fist that is not recognized as “fist, center” (false negative)



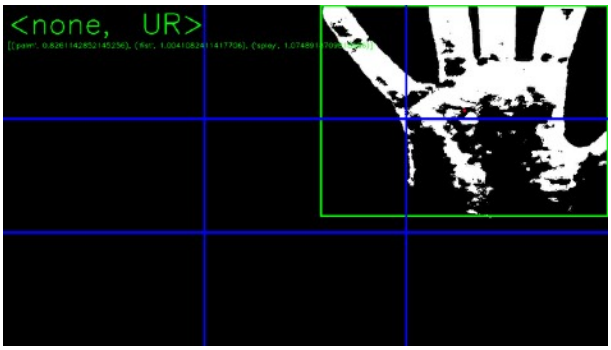
It is a tilted fist with only its side showing, hence the program cannot detect it as a fist.

One image that is incorrectly recognized as “fist, center” (false positive)



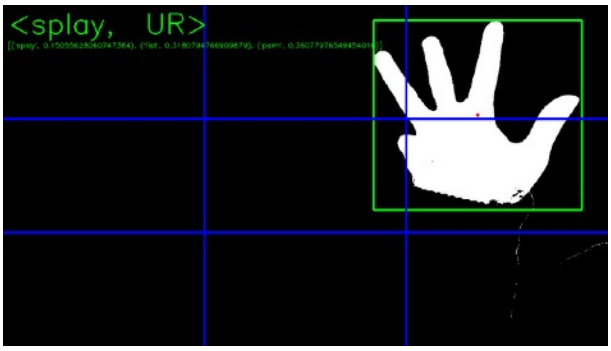
Although it has thumbs up, the general shape is closer to fist than palm or splay, the system classified it as a fist.

One image of a cornered splay that is incorrectly recognized as “splay, upper right” (false negative)



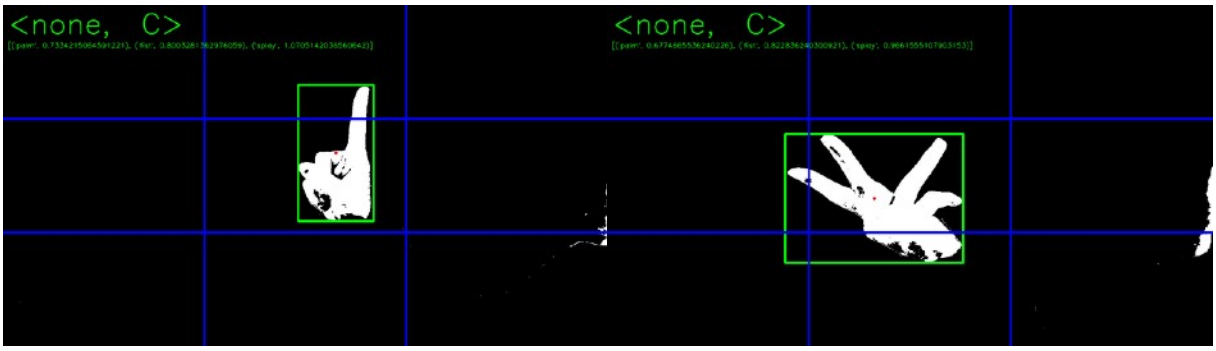
In this image, the hand is too close to the camera that the top of the hand got cut off. Hence, the system could not identify the hand shape and outputted none.

One image that is incorrectly recognized as “splay, upper right” (false positive)



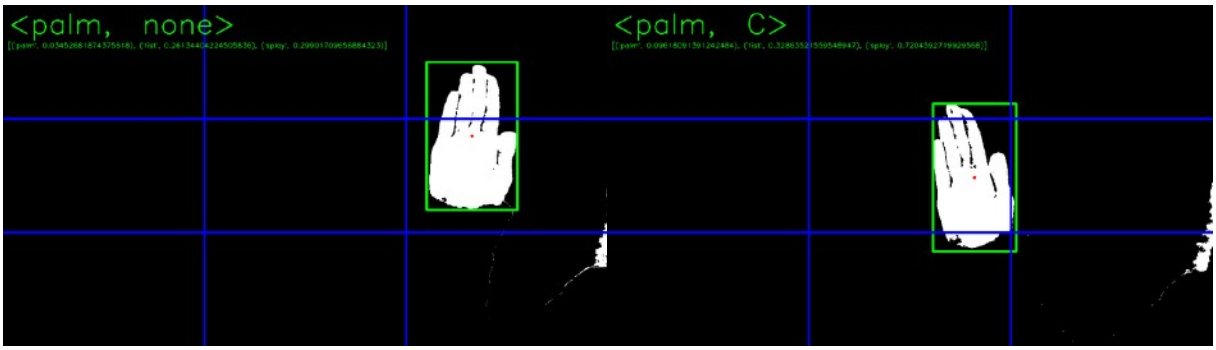
Since the general shape is a splay, the system outputted splay. However, the ground truth is that the pinky was folded, creating a non splay shape.

Two images that are accurately recognized as “unknown”, that is, neither “fist” nor “splay (true negative)

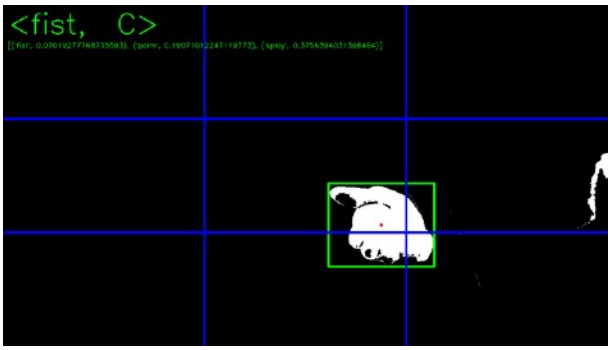


Second, extend the “what” vocabulary to include “palm”, which is midway between “fist” and “splay”. Show the following, where the “where” is not important:

Two images of a palm that are correctly recognized as “palm”. (true positive)

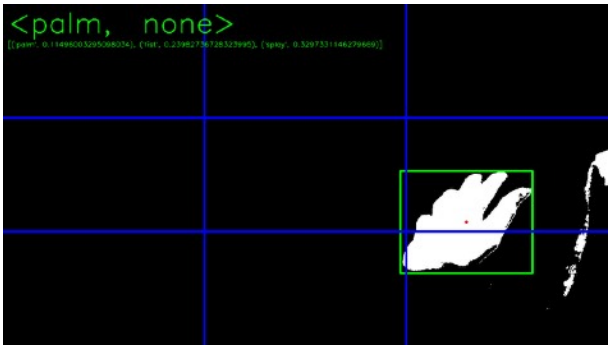


One image of a palm that is incorrectly recognized as “fist” (false negative)



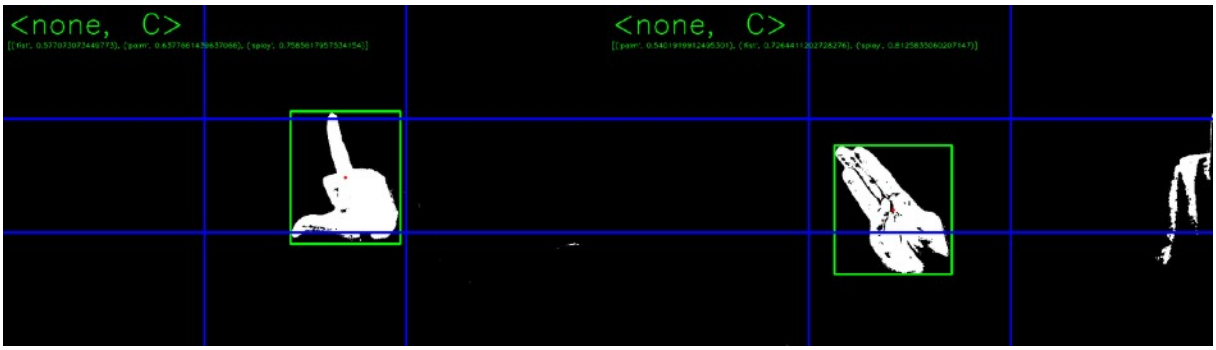
In this image, the palm is placed horizontally facing the ceiling, hence the binary image it makes has more of a shape of a fist, hence the system recognized it as a “fist.”

One image of “splay” that is incorrectly recognized as “palm” (false positive)



My hand is making a splay, but it is in an angle where the space between the fingers are not visible to the camera. Hence it looks like a shape closer to the palm in the binary image, so the system recognized it as a palm.

Two images that are accurately recognized as “unknown”, that is, neither “fist” nor “splay” nor “palm” (true negative)



STEP 4.

Determining the three sequences

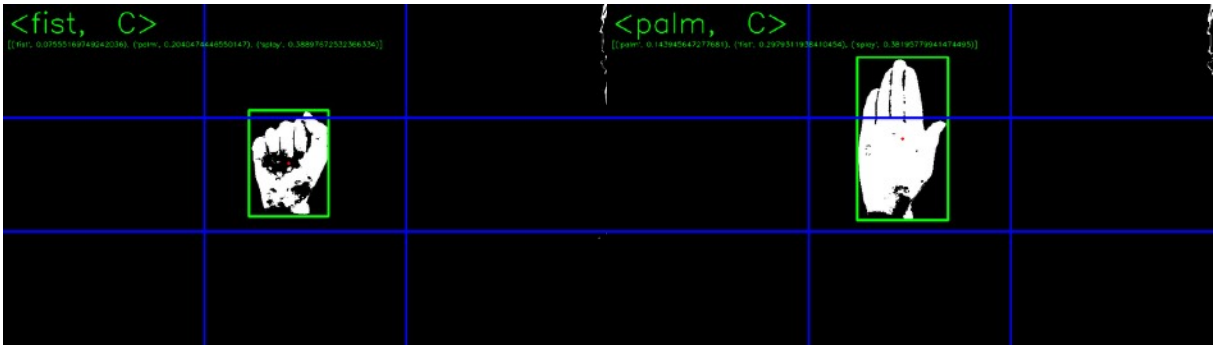
Easy: [("fist", "C"), ("palm", "C"), ("splay", "C")]
Good: [("fist", "UR"), ("palm", "LR"), ("splay", "UR")]
Hard: [("splay", "UL"), ("splay", "LL"), ("palm", “UL")]}

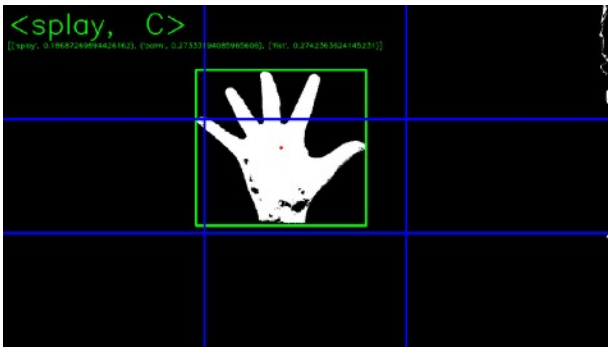
First, I found that out of all sections, center is the easiest for the system to recognize. It is easier for us to place our hand correctly in the center than to place it the corner. It is because when we try to place it in the corner, we worry about our hands going out of camera that we miss the grid. However, there is no such problem with center. Therefore easy sequence focused on hand location in the center.

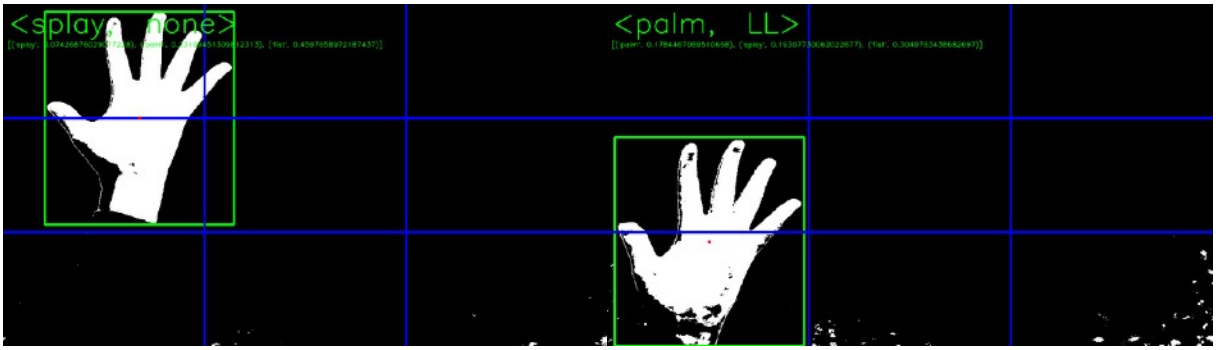
I also realized left is harder than right. This was because I am right handed and was using my right hand to take the pictures. Because I had to make sure that my face is out of the frame, it was easier to place my hand closer to myself, than to extend it to the other hand while making sure that my face is not in the camera, and also making sure that my hand gesture is clear. Therefore the good sequence had hand location in the right, since it is harder than center but easier than the left.

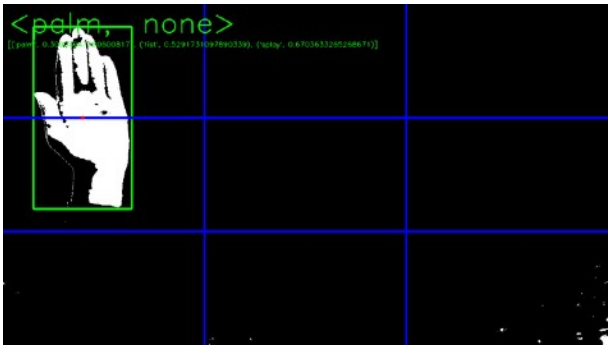
Lastly, out of the three shapes, splay had the most error rate since if I don’t extend my fingers apart enough, and if the spaces between my fingers are not shown clearly enough due to poor angle, the system would classify it as a palm. Therefore, the hard sequence consisted more of splay and left corners.

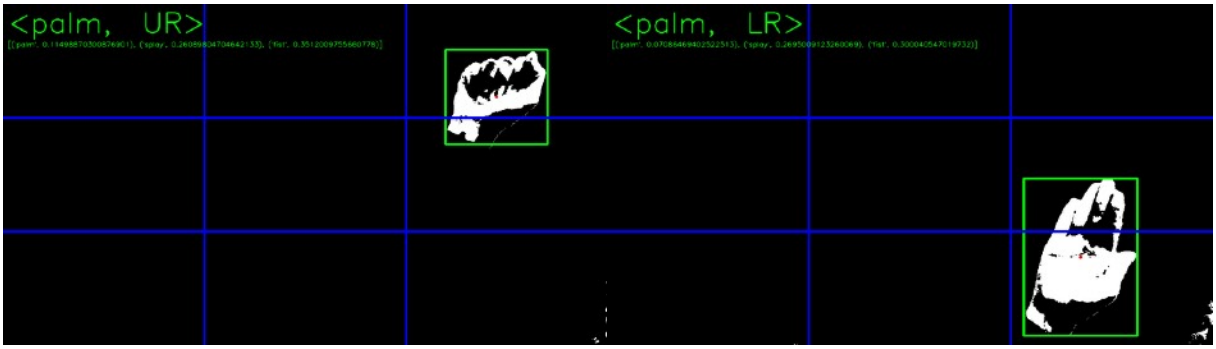
The (new) reduced resolution binary intermediates I gathered from me

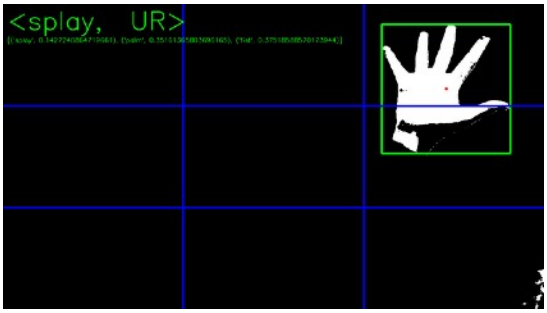




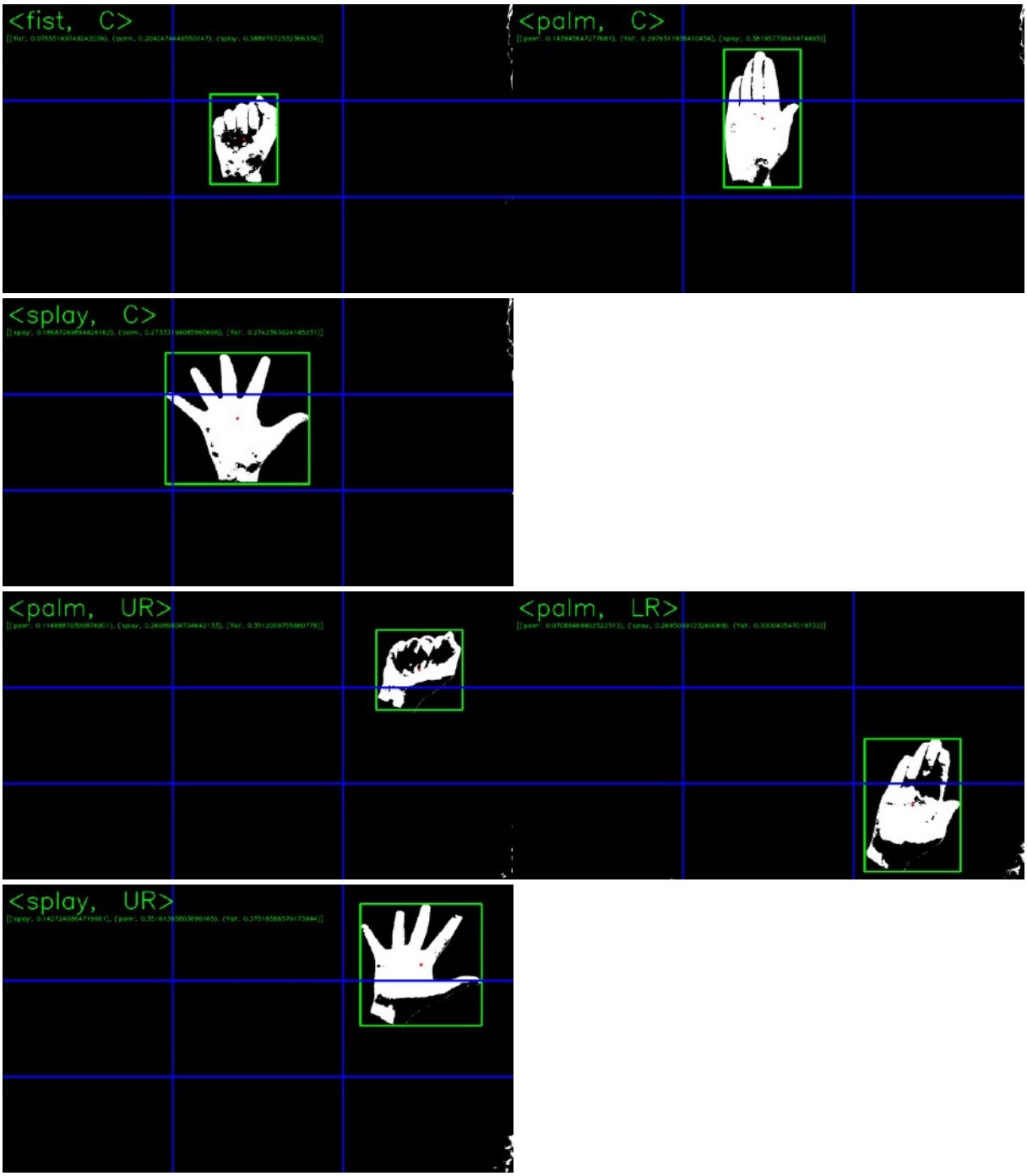

<splay, none>
<palm, LL>

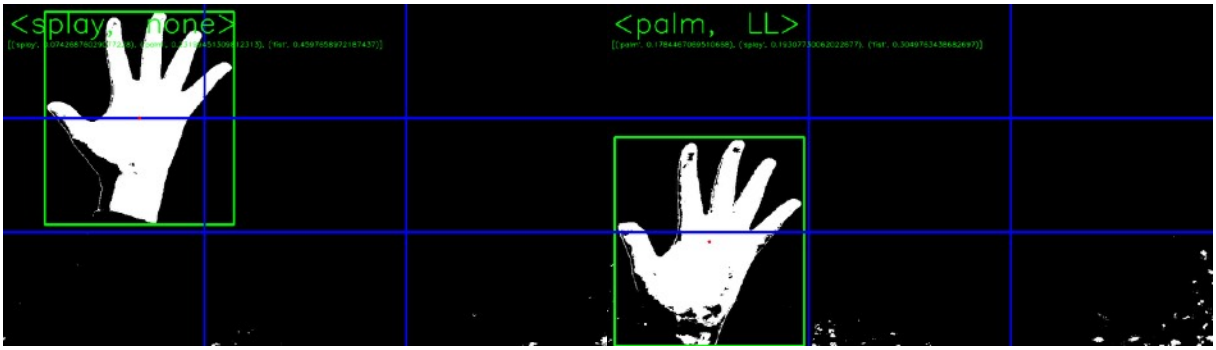

<palm, none>

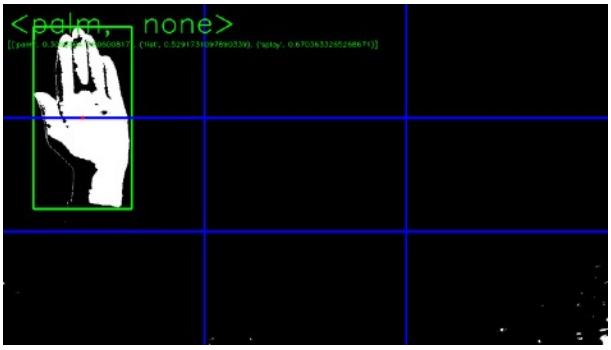

<palm, UR>
<palm, LR>


<splay, UR>

The (new) reduced resolution binary intermediates I gathered from my friend.







The system's overall success rate in terms of accuracy

I got 2/3 sequences right, 8/9 hand motion right, and 17/18 details right. The only thing I got wrong is that for the first motion of hard sequence, I needed <fist, UR>, but my image was recognized as <palm, UR>.

My friend got 1/3 sequences right, 7/9 hand motion right, 14/18 details right. For first hand motion in good sequence, she did <fist, UR>, but since her fingernails had shadow, the fist looked like there were a lot of empty space in the middle, and i was identified as a palm. Second motion in the difficult sequence was supposed to be splay, but since her fingers were not fully extended apart, the system recognized it as palm. The first and third images in the difficult sequence was recognized as none although both should have been UL. If you examine the image, you can see that the center fell right in the boundary between UL and UC. This is because her sleeves were slightly down, and since her wrist was included in the contour, the center was calculated to be slightly lower than it should have been.

Feedback on the system: ease of use, confusability of "what", and of "where", and any other comments that could be used to improve the system.

My friend had trouble adjusting her hand because the frame was mirrored, and she had to move right in order to move left in the frame and vice versa. My definition of splay was extremely separated fingers instead of casually separated fingers for better recognition. However, it was not very conventional and I realized her splay was more closer together than my splay. She also said that it was difficult having to click "s" to take a picture since she also had to adjust the hand gesture and get her body out of the frame.

What I would improve on

I think I would show the frame mirrored so that the user can adjust more easily. I would also develop a mechanism to cut out wrists if they are selected in the contour. I think I might be able to do it by creating a wrist shaped contour (rectangular) and deselecting it from the hand contour. I could also make the system take a picture with voice activation so that the user can more comfortably adjust and take a picture more easily.