Final Project
Gesture based paint system
Visual Interface COMSW 4735 Spring 2015

Angus Ding ad3180
Ayaka Kume ak3682

May 13, 2015

Contents

1 Introduction

1.1 Purpose of this project

In this project, we implemented a system that allows the user to paint on the screen using intuitive gestures instead of the mouse. Instead of using any specific draw pad or other device, we will implement the system using a simple white paper, a web cam, and a light source. Without any tactile input, we are going to rely on visual inputs to determine the gesture and the position of user's hand in real-time. The challenging part about this project is how to define the natural gestures that human use to indicate the drawing on a blank paper, and how to recognize them using pure visual signal processing. Because the system only rely on the visual input, this project is perfectly suitable as an example of a visual interface.

1.2 Previous work

EnhancedDesk[1] is a two handed drawing system using infrared camera. Left hand and right hand have the different role. Isard, Michael, and John MacCormick implemented a vision based drawing package to demonstrate the hand tracking method[2].

1.3 Program features

The user will be given a device that consists of a white paper(or a whiteboard), a web cam that looks down from above, and a light source that projects light onto the paper from a non-perpendicular angle. The distance between the web cam and the paper, the paper and the light source, and the angle of the light source are all fixed. On the bottom of the paper will be some color blocks which represent a palette, and possibly some symbols which represent the drawing tools that the user can choose. The user can simply touch the color blocks to choose the color, and touch the tool symbols to choose the painting tool he or she wants to use. On the same time there will be a program on the computer screen which shows the canvas on which the user draws. To draw a picture, the user can use the most intuitive gesture – use the index finger like a pen. Touching the paper with the index finger means to draw, while moving the finger without touching the paper means to move the pen without drawing. This gesture, when the user touches the palette or the symbols instead of the empty area, means to select the color or the tool instead of drawing. For convenience, we will also define an erase gesture, which is a palm facing downwards with the four fingers stretching straight. This gesture is easy to use, and suitable for the semantic of erase, because it is the movement one will use to wipe something away from a surface.

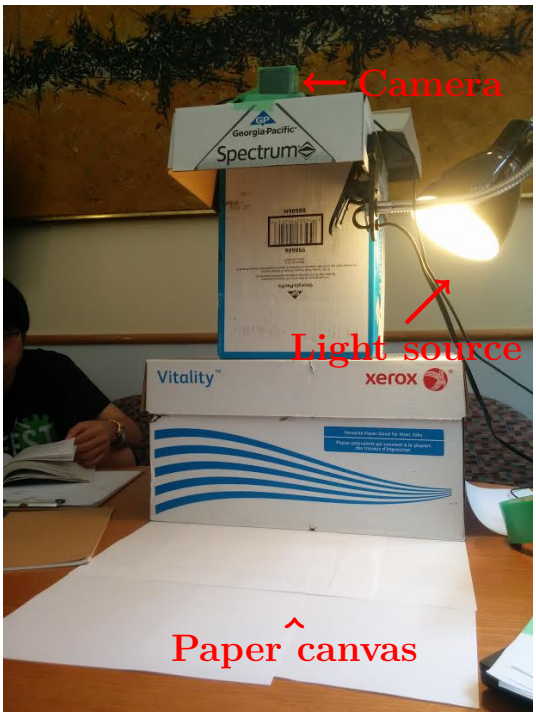


Figure 1: The input device

1.4 Domain engineering

Fig.1 shows the setting of our device. The height from the camera to the paper canvas is 54 cm. The camera looks down so that we can convert the coordinate of the fingertip to the canvas coordinate easily. The canvas on the paper is 19cm by 14cm. We set all of back ground as white in order to detect hands easily. We use the web cam, logicool carl zeiss tessar. We use Windows7 and python 2.7. As a light source, we use handy light. The user is assumed to be east Asian, because our training data contains only east Asian.

1.5 Division

Angus Ding (ad3180) implemented and wrote a report about posture detection, shadow detection and GUI part. Ayaka Kume (ak3682) implemented and wrote a report about hand detection, fingertip detection and evaluation part. We wrote introduction and discussion together.

2 Overview of the method

3 Hand detection

Because of our settings, there are only white background, shadow and the hand in an image. First, we detect hand using skin color. Then we detect wrist. Because there are only hand or wrist in the scene, we can get hand mask by erasing wrist region. Fig.2 shows the overview of the system. For wrist detection, we modify the method from [3]. All of the functions in this section is in major.py and hand_detection.py.

3.1 Skin color detection

Because of our settings, there are only white background, shadow and the hand in an image. So we detect the hand by color. We use both RGB and HSV value to detect our skin. Because both of our team mates are East Asian, we tried skin detection only for East Asian people. In particular, we define skin color pixel as:

- its Red value is larger than Blue value
- its Red value is larger than Green value
- its Value (HSV) is smaller than 73 %
- its Saturation is larger than 30 %

Also we mask where outside of the canvas.

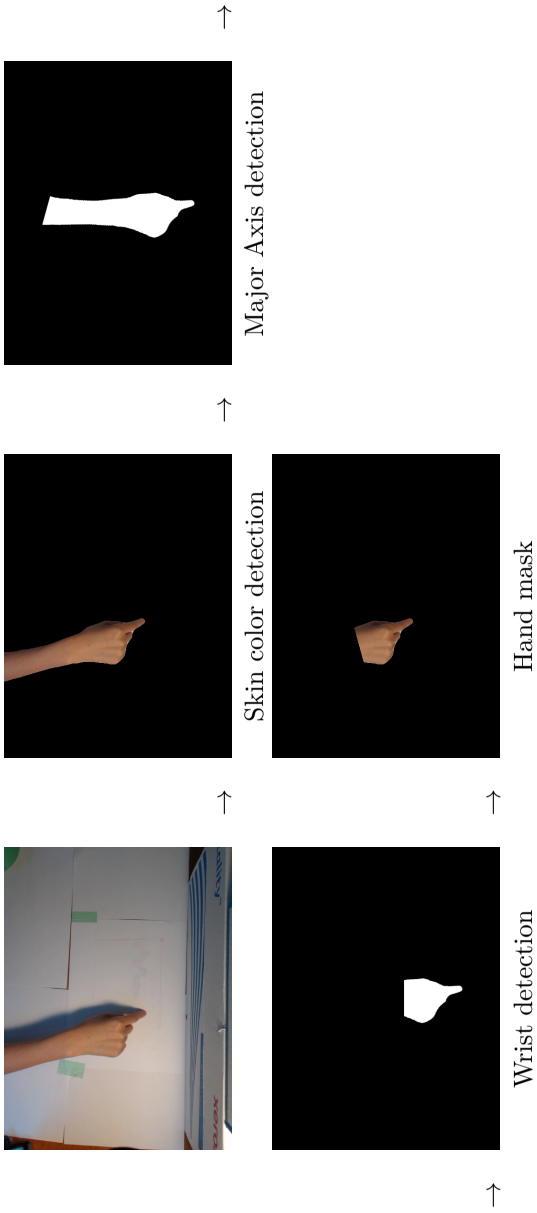


Figure 2: Overview of hand detection

3.2 Orientation of the hand

We want to detect the orientation of the hand in order to use training data, or detecting hand region. In our program, major axis.py has this function. We define the orientation of the hand as the angle of the axis of least second moment.

The input is the binary image of skin region. Axis of least second moment minimizes E, the sum of the distance from all points to the line. That is,

$$E = \int\int r^2 b(x,y)dxdy$$

where r is a distance from b(x,y) to the axis and b(x,y) = 1 when a pixel (x,y) belongs to the object, otherwise 0. Let the axis be $x\sin\theta - y\sin\theta + \rho = 0$. Distance of point (x,y) from axis is:

$$r = |x\sin\theta - y\cos\theta + \rho|$$

. Thus minimizing E means minimizing

$$E = \int\int (x\sin\theta - y\cos\theta + \rho)^2 b(x,y)dxdy$$

Because $\partial E/\partial\rho = 0$, we get

$$A(x_c\sin\theta - y_c\cos\theta + \rho) = 0$$

where A is an area of the object and (x_c, y_c) is center of the object. This means, the axis should pass the center point of the object. Then, we shift the coordinate system in order to set the center point as origin. That is, $x' = x - x_c, y' = y - y_c$. Because this line should pass the origin, the line can be represented as $x'\sin\theta - y'\cos\theta$ So,

$$E = a\sin^2\theta - b\sin\theta\cos\theta + c\cos^2\theta$$

. Where $a = \int\int (x')^2 b(x,y)dx'dy', b = 2\int\int x'y'b(x,y)dx'dy', c = \int\int (y')^2 b(x,y)dx'dy'$.

Because $\partial E/\partial\theta = 0$, we get

$$(a-c)\sin 2\theta - b\cos 2\theta = 0$$

Also, minimizing E means the second derivative is larger than 0. Using these information, the orientation $\theta = atan2(b, a-c)/2$.

Fig.3 shows the results of orientation detection and the rotated image. The first column is an original image. The second column is a translated image. First, the center point of the hand moves to the center point of the image. The light blue line is the axis. The blue dot is the center point. For the mask, the image is rotated by the angle of $-\theta$ and translated to the original position. The figure shows that this axis does not depends on the small fingertip movement. If the binary image of hand has enough amount of areas, this system can detect the angle of the hand. If the image of hand does not have enough amount of areas, for example, it can detect only a part of fingers, it cannot detect the angle of the hand correctly. However, because of our settings, we always can see enough amount of hand in the target area.

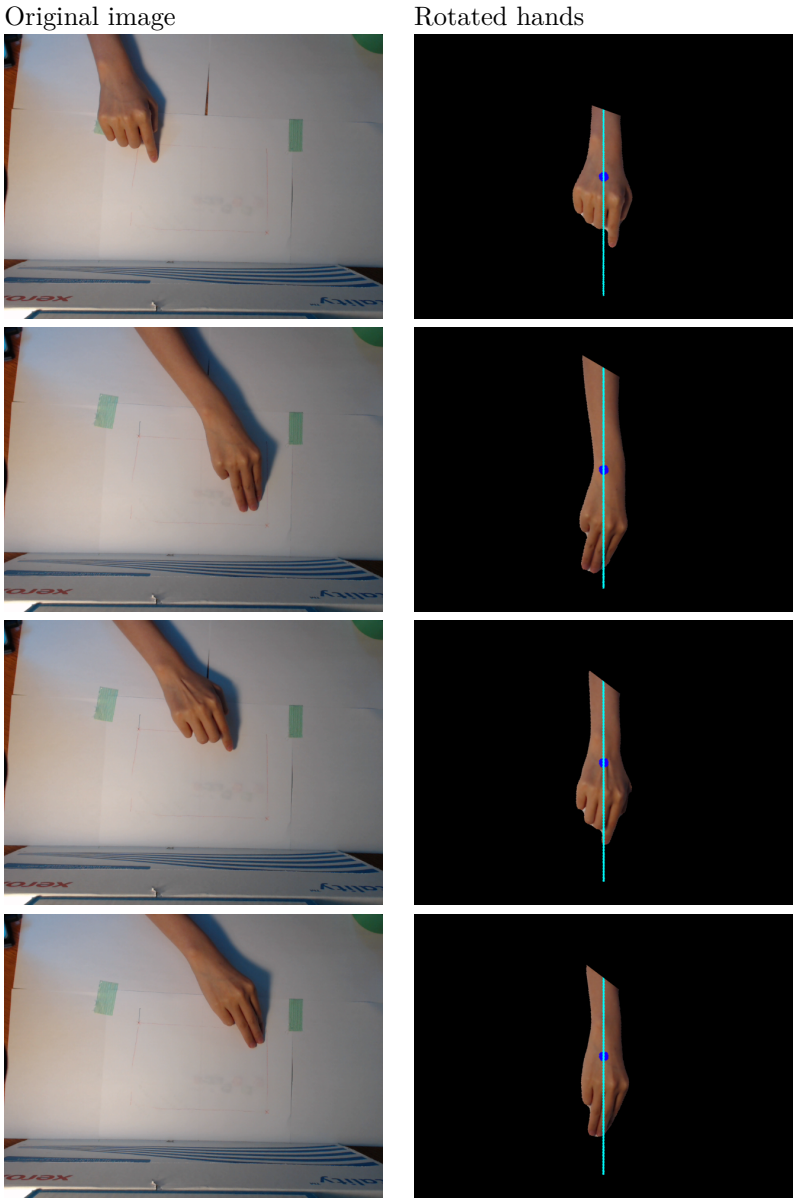


Figure 3: The results for orientation detection

Figure 4: Hand image Left: original skin color image Right: rotated skin color image

3.3 Hand detection

Then we have to extract only hand region. If there are long arms, we may not classify the gesture correctly. We modified the method introduced by [3]. The method is first detect the skin region by color (HSV), and then detect the wrist end. Wrist end is detected by the simple method as follows: Fig.4 shows the image of the hand. First, we calculate the number of the pixels on boundaries, up (between blue point and red point), down (between green point and yellow point), left (between blue point and green point), right (between red point and yellow point). We can assume the most largest among up, down, left and right is the wrist side. This is because the author of [3] and we assume hand is inside of the image, but human itself is not. Then, they detect the wrist end using intensity histogram. Intensity histogram is the sum of the number of the pixels on the row/cols. If wrist is up/ down, it calculate along rows. Otherwise, it calculate along columns. Assume the wrist is down. Let p be a point which is either on a left or right boundaries and the nearest from the down boundary. The author of [3] found that the slope on the histogram between b and the wrist end is highest among the slopes between b and other points which is nearer to down than b. b is the point which is the most left or right point so it is always near the most widest region.

However, the wrist detection does not work well because the paper assumes the hand gesture as spray hand and so the palm is always the widest. Like Fig.4, we cannot find appropriate b because the most left or right points are not palm, but finger and wrist. This is because palms are not fully opened so it is difficult to detect hand region as it is. So we rotate the image along the axis so that we can assume the wrist is always 'up' side and the most left or right part tend to be the palm region. Fig.4 can detect wrist by our method but not the previous method. This is because our method can detect the most widest point as b. We assume b is not too near to the wrist end. That is, if left or right is within 30 pixel from up, use another point. Fig.5 shows the example of the histogram. Red point in the histogram corresponds to the b's coordinate. Green point in the histogram corresponds to the wrist end's coordinate. The

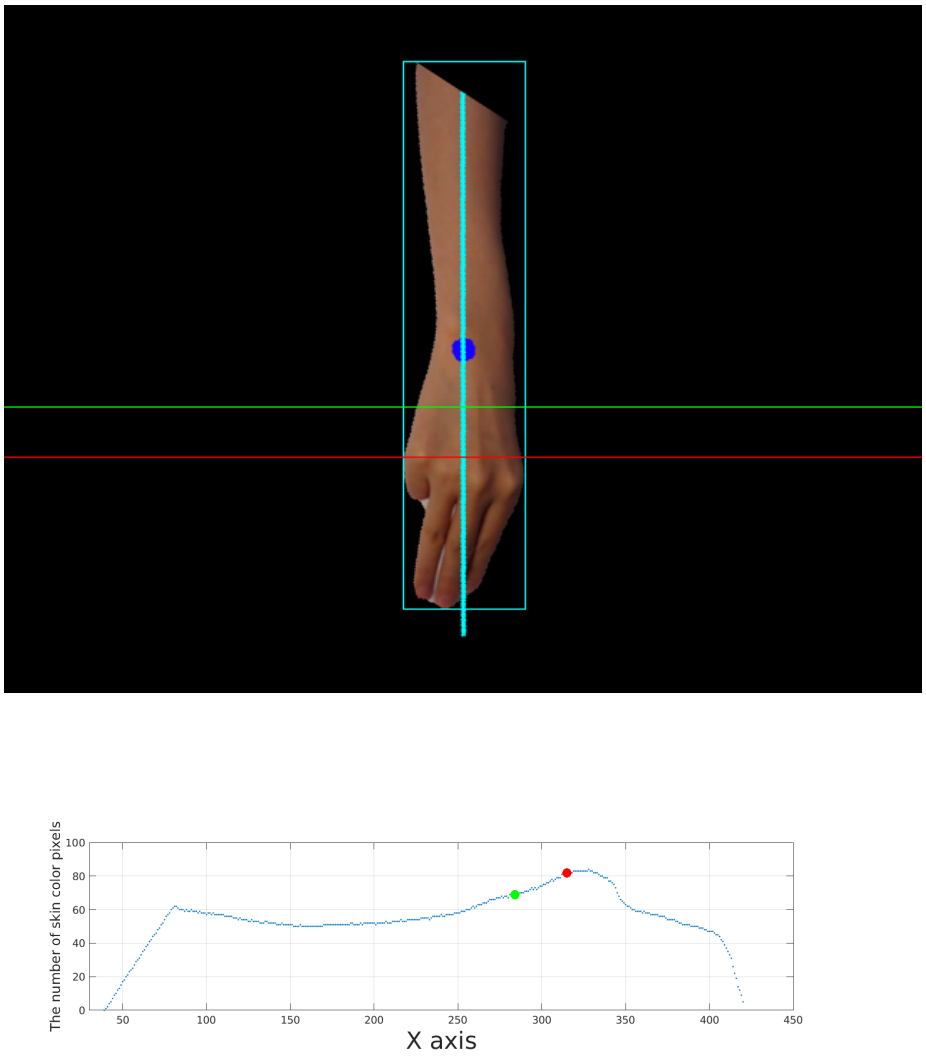


Figure 5: Hand image Top: Hand image. Green line is wrist position and red line is b. Bottom: histogram

red line and green line in the left image corresponds to the b and wrist end.

Taken together, our method is as follows.

1. Find skin color area
2. Find orientation of the skin area and rotate
3. Assume up side is wrist
4. Find wrist end and crop

Even though we improve the method, in case the skin detection fails and the hand become smaller of the palm is too small to be a widest length, in that case, we simply extract 1/4 of the all of the region.

3.4 Result

Fig.6 shows the results of the hand detection. As we can see, the hands are correctly detected.

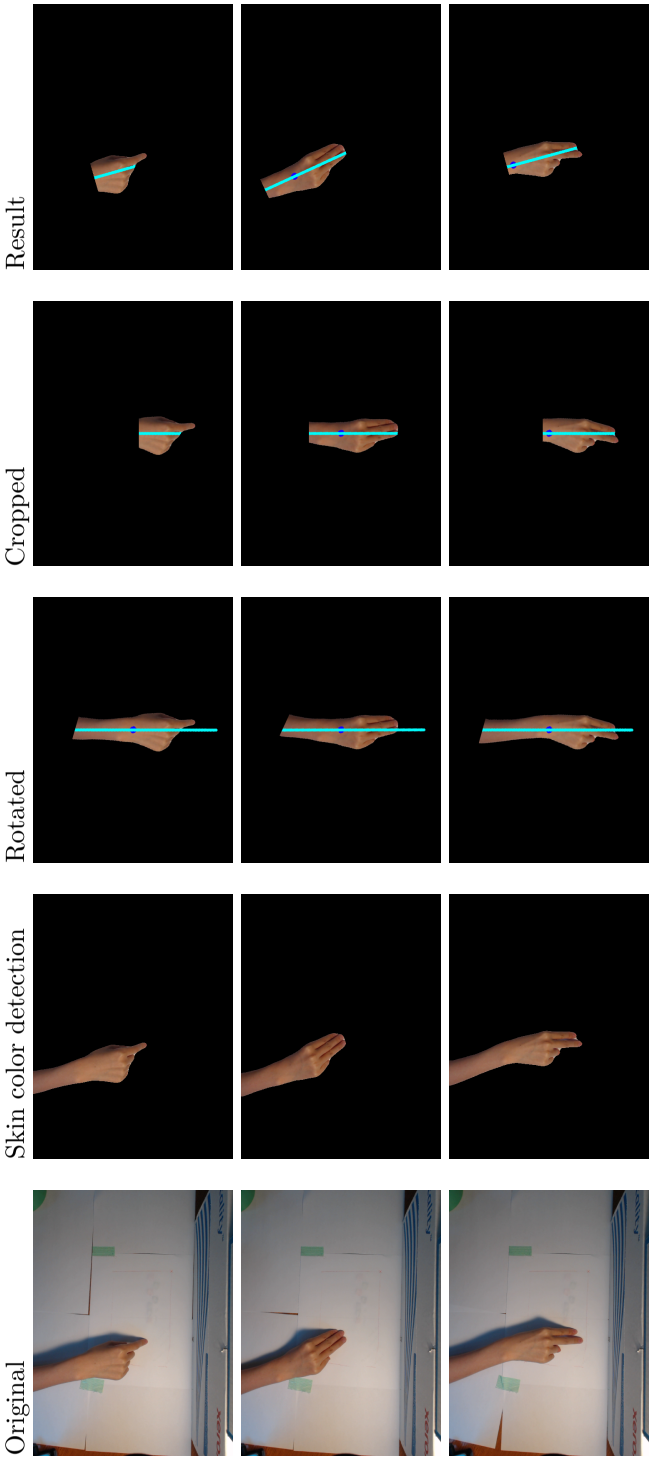


Figure 6: The results of hand detection

4 Posture recognition

Table 1: The result of three fingers' finger tip detection

The location of the result	
Index finger	0(0%)
Between index finger and middle finger	1(0%)
Middle finger	124 (87%)
Between middle finger and third finger	12(8%)
Third finger	6(4%)
Total	143

5 Fingertip detection

In order to draw points using finger information, we have to detect the location of the fingertip. If the gesture is 'draw (one finger)', we detect the fingertip of the index finger. If the gesture is 'erase (three fingers)' or 'move (two fingers)', we detect the fingertip of the middle finger.

We assume the fingertip is the further point from the center of the mass and also not the wrist side. Our environment allows us to assume the wrist side is always top of the image, we simply find the furthest point from the center of the mass among the bottom half of the mask.

This function is implemented in fingertip.py.

Fig. 7 shows the results of the finger tip detection. This method detects the finger tip correctly when the gesture is one finger or two fingers. When the gesture is three fingers, the method sometimes detect the location between middle finger and other fingers. Table 1 shows the result of detected finger tip when the gesture is 'three fingers'. We tried 143 frames of three fingers. The error ratio is 12 %. Fig. 8 shows the examples of failure. As we can see, it fails when the hand is not on the canvas. This would cause about 10 pixels error in video image and cause about 20 pixels error in canvas. This is about 3% of the average length of the canvas.

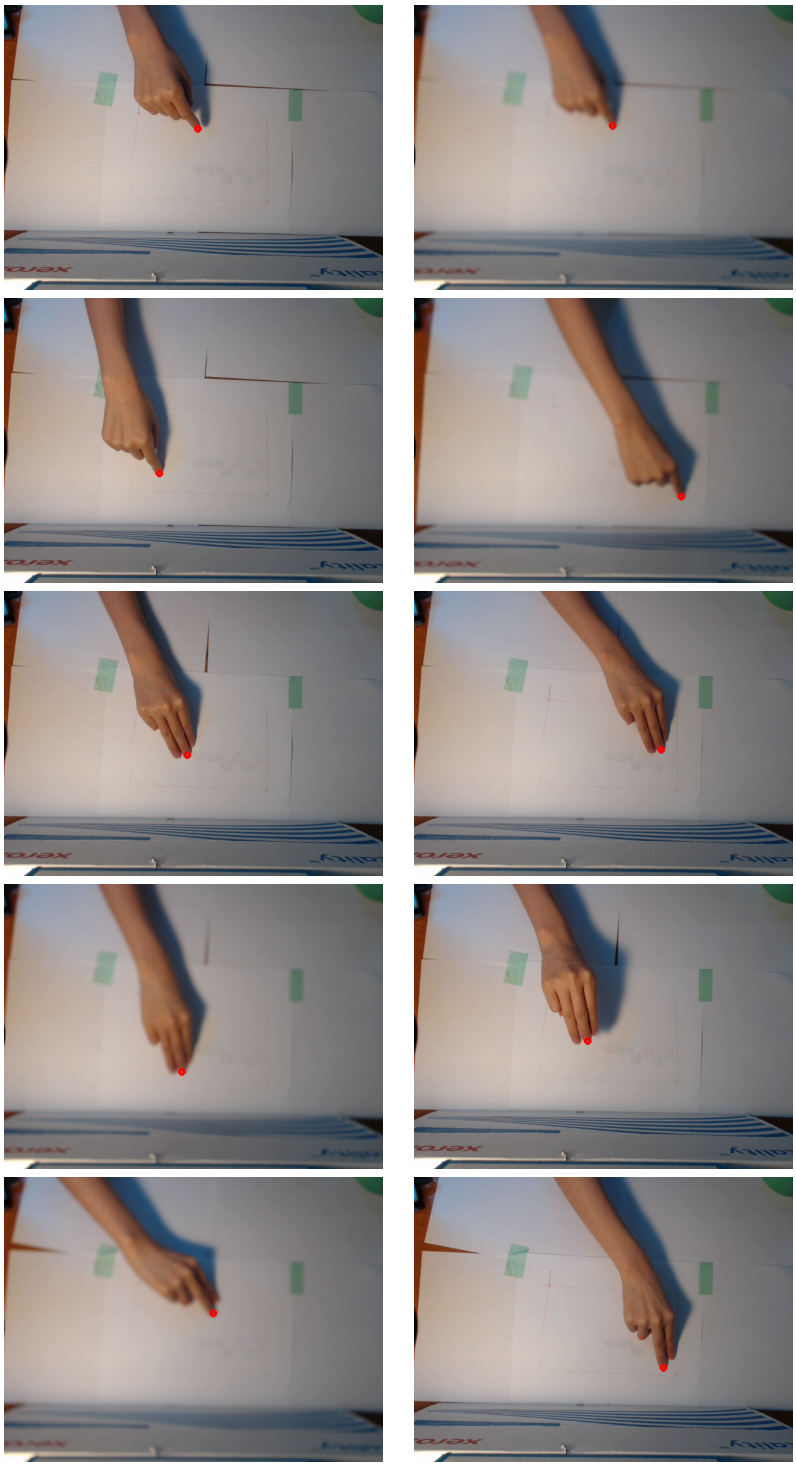


Figure 7: The results of finger detection

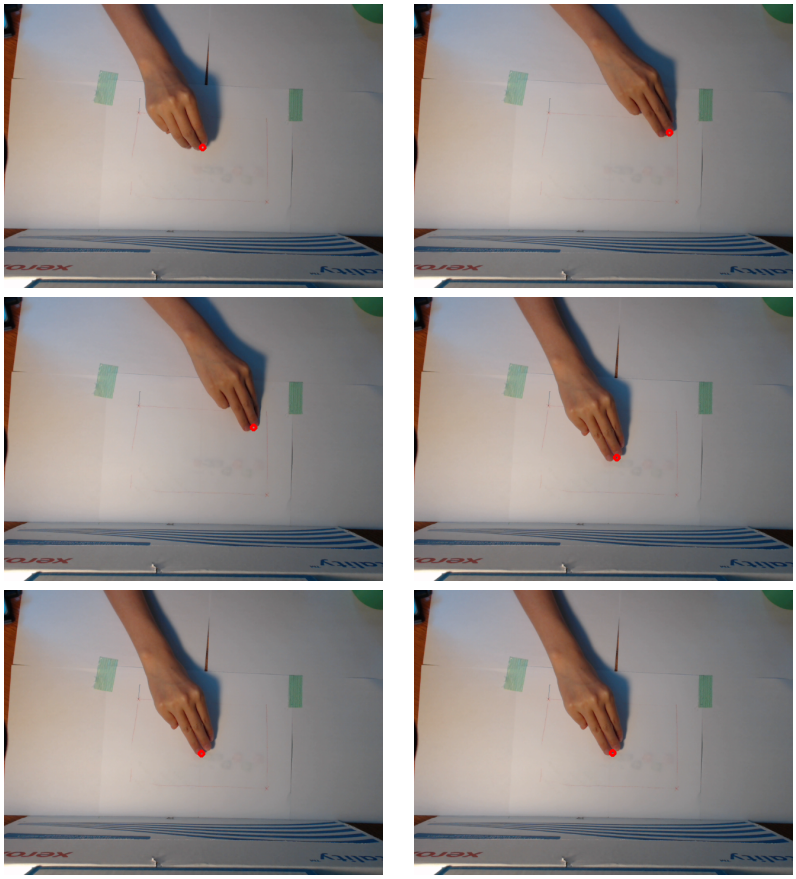


Figure 8: The errors

6 Determine if the finger is touching the paper

7 Evaluation

For evaluation, we measure time, error rate and the consistency of each trials. If time is fast, it means the system is easy to operate. If the average error is low, it means the system is accurate. If time and average error does not change for every trial, it means the system is stable.

We only evaluate about 'drawing' function because 'drawing' function is the most essential function of the system. We evaluate this system using three tasks. Fig. 9 shows three examples. The flow of the evaluation is as follows. At first, there is a white canvas and pointer. When the tester press 'r', it shows the sample to trace and starts timer. User starts drawing and when he finish drawing, the tester press 'q' and timer ends. The user do same task three times. Time is measured by this timer. The average error is measured by the difference between the example and what he draw. The consistency is measured by the standard deviation of time and average error.

To compare the results, we also tested the mouse as the input device. When we point a point in the canvas using a mouse, the pointer moves. When we drag with left click, it draws a line.

We used user as both of our team member.

7.1 The average error measurement

7.1.1 Points

We measure the distance between what the user draw and the nearest point of the sample. To make the measurement easy, we divide the canvas into 4 section. If a point the user draw is on upper left, we compare the distance with upper left point of the sample. Same as upper right, lower left and lower right.

7.1.2 Circle

Circle is the points which distance from the center point are radius. We know the center point and the radius of the sample circle. Thus we measure the absolute difference between radius and the actual difference from the center point for each points. Then we average them.

7.1.3 Line

We measure the distance between the actual line and sample line by comparing only 20 points of the lines. We equally sample 20 points in the lines along the vertical axis. And from the most highest points to the most lowest Points, we measure the difference of them. Fig.10 shows the example. Blue line is a sample line and red line is user's line. The points are sample points. Each points are measured with the points which are connected by the black line.

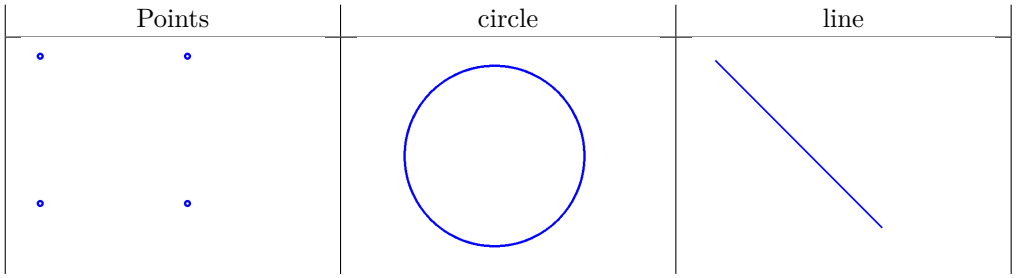


Figure 9: Three tasks

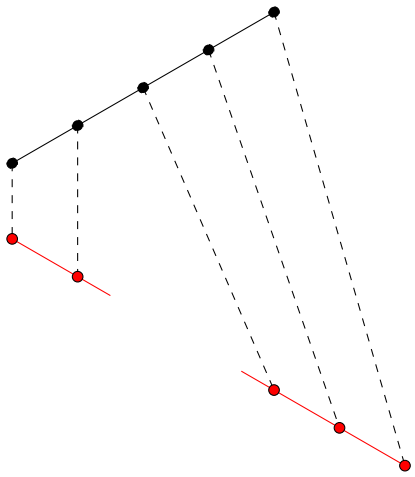


Figure 10: Line error measurement

7.2 Results

Table 2 shows the measurement results of the points task. Table 3 and Table 4 show the results of the circle task and the actual canvas image. Table 5 and Table 6 show the results of the line task and the actual canvas image. Table 7 shows the standard deviation of them. In all tasks, mouse device results are better than our input device. Among three tasks, for our device performance is better in circle task. We can see the time deviation is better than the error deviation.

Table 2: The result of the user study (points)

User	Device	Trial	Time[sec]	Error Pixels
A	VI	1	19.74	43.59
A	VI	2	16.64	58.73
A	VI	3	14.38	67.47
B	VI	1	23.06	49.28
B	VI	2	19.85	42.29
B	VI	3	22.29	66.24
A	Mouse	1	8.18	4.65
A	Mouse	2	6.28	7.80
A	Mouse	3	6.60	6.47
B	Mouse	1	6.35	3.78
B	Mouse	2	5.48	4.18
B	Mouse	3	5.29	4.55

Table 3: The result of the user study (circle)

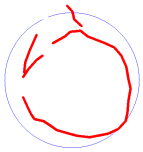
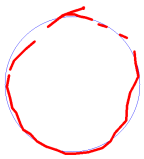
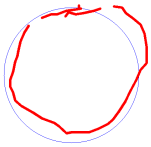
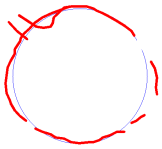
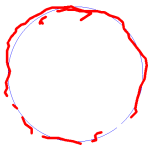
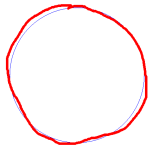
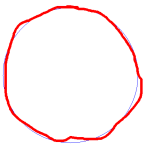
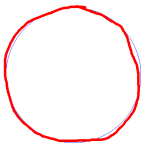
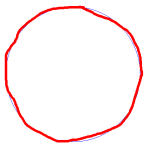
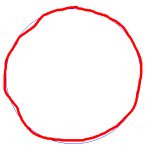
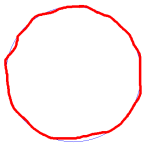
	Trial 1	Trial 2	Trial 3
User A (VI)			
User B (VI)			
User A (Mouse)			
User B (Mouse)			

Table 4: The result of the user study (circle)

User	Device	Trial	Time[sec]	Error Pixels
A	VI	1	13.31	38.25
A	VI	2	12.65	9.47
A	VI	3	13.16	24.21
B	VI	1	18.53	11.54
B	VI	2	22.82	12.97
B	VI	3	22.10	7.69
A	Mouse	1	7.90	6.73
A	Mouse	2	7.97	5.12
A	Mouse	3	7.79	4.95
B	Mouse	1	9.38	3.84
B	Mouse	2	9.58	4.27
B	Mouse	3	9.98	3.64

Table 5: The result of the user study (line)

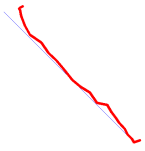
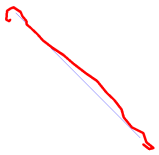
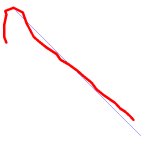
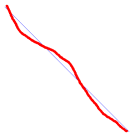
	Trial 1	Trial 2	Trial 3
User A (VI)			
User B (VI)			
User A (Mouse)			
User B (Mouse)			

Table 6: The result of the user study (line)

User	Device	Trial	Time[sec]	Error Pixels
A	VI	1	20.16	22.25
A	VI	2	14.75	76.46
A	VI	3	5.73	25.77
B	VI	1	7.76	39.92
B	VI	2	8.03	88.50
B	VI	3	10.03	58.27
A	Mouse	1	4.07	13.87
A	Mouse	2	5.19	6.53
A	Mouse	3	4.07	8.01
B	Mouse	1	3.98	8.99
B	Mouse	2	4.83	7.24
B	Mouse	3	4.61	7.71

Table 7: Standard Deviation of User study

Task	User	Device	StdDev of Time	StdDev of Error
Points	A	VI	2.20	9.87
Points	B	VI	1.37	10.06
Points	A	Mouse	0.83	1.29
Points	B	Mouse	0.46	0.31
Circle	A	VI	0.28	11.75
Circle	B	VI	1.88	2.23
Circle	A	Mouse	0.07	0.80
Circle	B	Mouse	0.25	0.26
Line	A	VI	5.95	24.77
Line	B	VI	1.01	20.03
Line	A	Mouse	0.53	3.17
Line	B	Mouse	0.36	0.74

References

[1] Xinlei Chen et al. "Two-handed drawing on augmented desk system". In: *Proceedings of the Working Conference on Advanced Visual Interfaces*. ACM. 2002, pp. 219–222.

[2] Michael Isard, John MacCormick, et al. *Hand tracking for vision-based drawing*. Tech. rep. Technical report, Visual Dynamics Group, Department of Engineering Science, University of Oxford, 2000.

[3] Jagdish Lal Raheja, Karen Das, and Ankit Chaudhary. "An efficient real time method of fingertip detection". In: *arXiv preprint arXiv:1108.0502* (2011).

Appendix:Code

```
import sys
import os.path
includespath = os.path.abspath('../includes')
sys.path.insert(0, includespath)
import posture
import cv2
import gui
import hand_detection
import fingertip
import hand_detection as hd
import numpy as np

def mouse_callback(event,x,y,flags,param):
    i,j = (y-1,x-1)
    if event == cv2.EVENT_LBUTTONUP:
        print i,j

def getinput(cap, pos_recognizer):
    ret, frame = cap.read()
    # Display the input stream only for debug purposes
    cv2.imshow('input',frame)
    # print "check point 1"
    label, hand_mask, theta, skin_mask = pos_recognizer.classify(frame)
    # print "label = ", label


    #cv2.imshow('debug', frame[])
    # frame_tmp = np.copy(frame)
    # frame_tmp[hand_mask==False] = 0
    # cv2.namedWindow('debug')
    # cv2.setMouseCallback('debug',mouse_callback)
    # cv2.imshow('debug', frame_tmp)

    if(label == posture.poses["UNKNOWN"]):
        print "posture = UNKNOWN"
    # print "check point 2"
    location, wrist_end = fingertip.find_fingertip(label, skin_mask)
    wrist_end = 'up'
    if(not location):
        return label, location, False
    # print "location= ", location
    # print "wrist_end = ", wrist_end
    # print "check point 3"
```

```
    touching = posture.isTouching(frame, label, location, wrist_end, hand_mask)
    return label, location, touching

def main():
    modelfilename = sys.argv[1]
    pos_recognizer = posture.PostureRecognizer.load(modelfilename)
    ui = gui.GUI()
    #Get the image and do the classification here
    cap = cv2.VideoCapture(1)

    while(True):
        # Capture frame-by-frame
        label, location, touching = getinput(cap, pos_recognizer)
        if(not location):
            if cv2.waitKey(20) == 27:
                break
            continue
        # print "touching=" ,touching
        #=======The grammar goes here=============
        print "label = ", label, "location", location, "touching", touching
        ui.handle_input(label, location, touching)
        cv2.imshow('Canvas', ui.get_screen())


        pressedKey = cv2.waitKey(60)
        if pressedKey == 27:
            break

if __name__ == '__main__':
    main()
```