Hand Gesture Recognition using Depth Images

Project Members:

Niharika Gajam - 130001013
Gunjan Patil - 130002013

Project Mentor:

Dr. Surya Prakash.

Contents

1. Abstract:

To recognize the hand gestures signifying numbers from 1 to 5 from the depth images and their respective depth values using Finger Earth Mover's Distance (FEMD).

2. Theory:

The color images are of size 640x480. The depth values vary from 0 to around 9000.

2.1. Finger Earth Mover's Distance:

Earth Mover's Distance (EMD) is a general and flexible metric to measure the distance between signatures or histograms. EMD is a measure of the distance between two probability distributions. It is named after a physical analogy that is drawn from the process of moving piles of earth spread around one set of locations into another set of holes in the same space. The locations of earth piles and holes denotes the mean of each cluster in the signatures, the size of each earth pile or hole is the weight of cluster, and the ground distance between a pile and a hole is the amount of work needed to move a unit of earth. To use this transportation problem as a distance measure, i.e., a measure of dissimilarity, one seeks the least cost transportation — the movement of earth that requires the least amount of work.

Different from the EMD-based algorithm, which considers each local feature as a cluster, we consider the input hand as a signature with each finger (the global feature) as a cluster. And we add penalty on empty holes to alleviate partial matches on global features. The detailed explanation of FEMD can be found in section

3. Procedure:

The major steps involved are as follows:

3.1. Hand Segmentation:

3.1.1. Hand Detection: The images are such that the hand in the image is the foremost object and the user is seen to wear a black band. Thus, using a particular thresholding range of depth values, a tentative hand region from the color image is obtained as shown in Figure 1. The thresholding range of 600 – 900 gives desired hand regions for gestures 1 to 5. To extract a more precise hand shape, the black belt is first located.

Fig. 1 Hand region detected

3.1.2. Method to Locate the black belt: Since the belt consists of black pixels, we could find only the black pixels in the hand region. However, there may be other black pixels excluding that of the belt. To achieve the belt region exclusively masking with the help of a 3x3 box filter is done on the hand region. The pixels for which the masking gives a value less than 0.1 will determine a black pixel. This process of masking results in locating the black belt as shown in Figure 2.

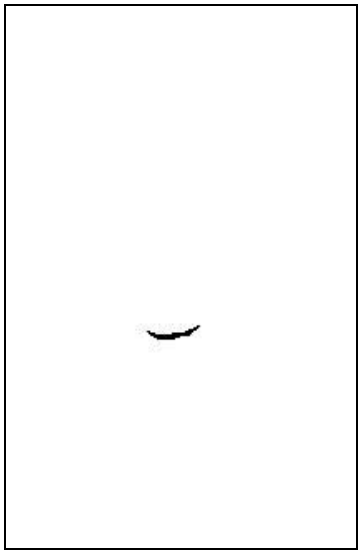

Fig. 2 Cropped belt

3.1.3. Method to extract a precise hand shape - The minimum and maximum x and y coordinates of the extracted belt in Figure 2 are calculated. These coordinates say, (xmin, ymin)

and (xmax, ymax) are used to form a diagonal line which will help exclude the part below the belt. The result of this method is shown in Figure 3.

Fig. 3 Precise hand detected

3.2. Shape representation:

To obtain the time series curve of the resulting colored hand shape, an initial point and a center point are to be determined first. The initial point is obtained as (xmin, ymin). To determine the center point, the obtained hand shape is converted to a binary image as shown in Figure 4. The Euclidean distance transform of this binary image yields image shown in Figure 5. The brightest point in Figure 5 represents the center point. The initial point (red point) and the center point (blue point) are shown in Figure 6. Furthermore, edge detection is applied on Figure 4 so as to get the exact boundary of the hand and also, to obtain the time series curve.

In our time-series representation, the horizontal axis denotes the angle between each contour vertex and the initial point relative to the center point, normalized by 360. The vertical axis denotes the Euclidean distance between the contour vertices and the center point, normalized by the radius of the maximal inscribed circle. Thus, the angle and distance from the center point is measured for all the contour vertices and plotted as shown in Figure 7.

Fig 4. Binary image of Hand shape to implement distance transform.

Fig 5. Distance transform of hand shape in Figure 4.

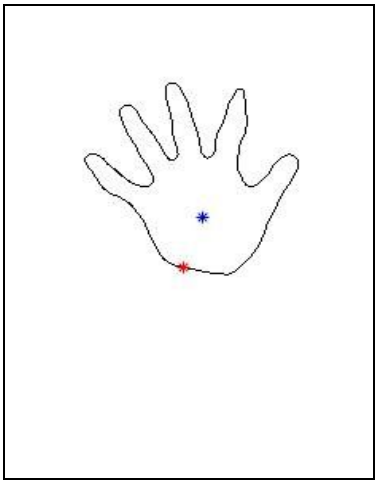

Fig 6. Boundary of hand shape with the center and initial points.

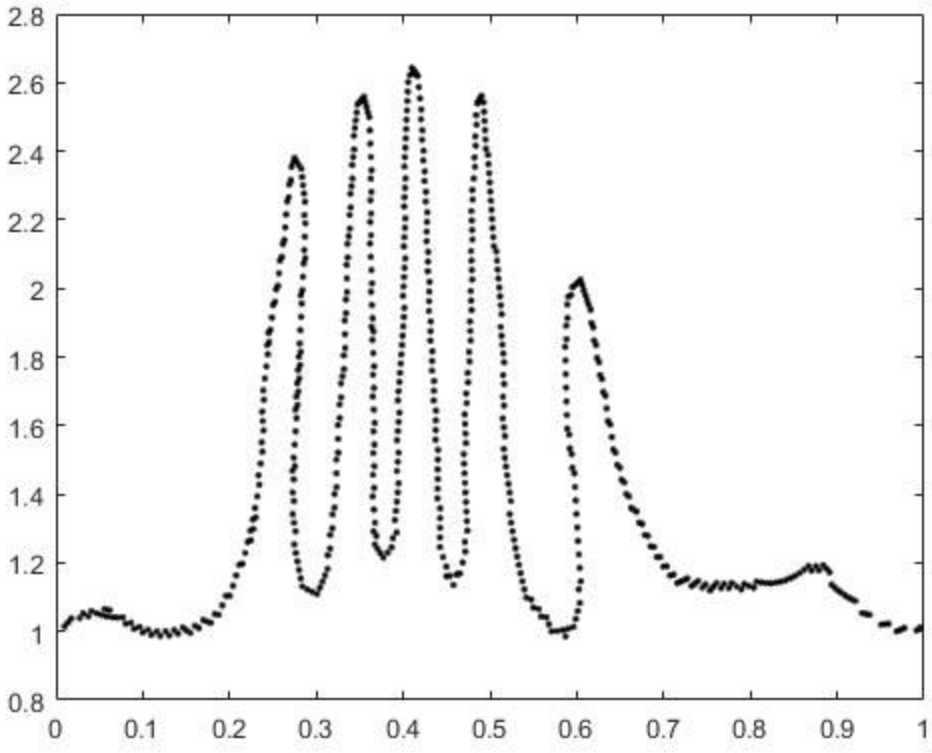

Fig 7. Time Series Curve Representation

3.3. Finger Detection:

Before measure the FEMD distance between two hand shapes, we have to obtain the finger clusters in their time-series curves, namely to detect the fingers from the hand shapes. The finger segments for each hand shape are detected by thresholding decomposition. An appropriate threshold in the distance values i.e. on y axis is chosen so that all the fingers segments are obtained. In our case, the threshold is chosen as 1.6. Thus, the points in the time series curve whose distance value is above 1.6 are separated. The separated finger segments are as shown in Figure 8.

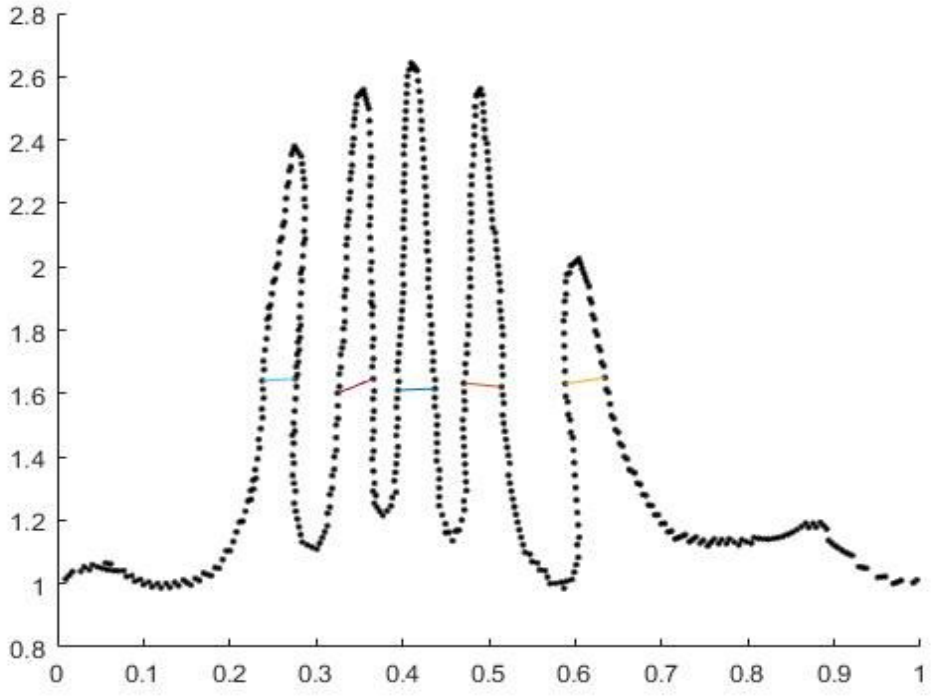

Fig 8. Time Series plot showing finger segments separated by different lines.

3.4. Finger Earth Mover's Distance (FEMD):

Each input hand is considered as a signature and each finger is considered as a cluster.

Formally, let R= { (r_1, w_{r1}), ..., (r_m, w_{rm})} be the first hand signature with m clusters, where ri is the cluster representative and w_{ri} is the weight of the cluster; T= { (t_1, w_{t1}),..., (t_n, w_{tn}) } is the second hand signature with n clusters.We define each cluster of a signature as the finger segment of the time-series curve: the representative of each cluster r_i is defined as the angle interval between the endpoints of each segment, $r_i=[r_{ia}; r_{ib}]$, where $0 \leq r_{ia} < r_{ib} \leq 1$; and the weight of a cluster, $w_{ri} \in (0; 1)$, is defined as the normalized area within the finger segment.

D=[d_{ij}] is the ground distance matrix of signature R and T, where dij is the ground distance from cluster r_i to t_j . d_{ij} is defined as the minimum moving distance for interval [r_{ia}; r_{ib}] to totally overlap with [t_{ja}; t_{jb}], i.e.:

$$d_{ij} = \begin{cases} 0; & r_i \text{ totally overlap with } t_j ; \\ \min(|r_{ia} - t_{ja}|; |r_{ib} - t_{jb}|); & \text{otherwise:} \end{cases}$$

For two signatures, R and T, their FEMD distance is defined as the least work needed to move the earth piles plus the penalty on the empty hole that is not filled with earth:

$$FEMD(R,T) = \beta E_{move} + (1-\beta)E_{empty},$$

$$= \beta \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} d_{ij} f_{ij}}{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}} + (1-\beta)\left|\sum_{i=1}^{m} w_{ri} - \sum_{j=1}^{n} w_{tj}\right|$$

where $\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}$ is the normalization factor, f_{ij} is the flow from cluster r_i to cluster t_j, which constitutes the flow matrix F. Parameter β modulates the importance between the first and the second terms. As we can see, E_{empty}, d_{ij} are constants given two signatures. To compute the FEMD, we need to compute the flow matrix F. We follow the definition of the flow matrix F in EMD, which is defined by minimizing the work needed to move all the earth piles.

4. Database:

The database comprises of 7 or more images for each gesture i.e. 1 to 5. The total number of images in the database is 37. For all the images in the database, all the steps up to finding FEMD are performed and the values namely, area of the finger segments, endpoints of each segment and gesture value for each image are stored.

5. Testing the dataset:

For each of the gestures two images with their depth values are selected randomly and steps 1,2,3 are performed and accordingly we get the values of the area of segments, end points of each segment, (we also have gesture value here, but we use it to check the accuracy of the algorithm). Now each of these test image and every entry in the database (except the ones which got selected for testing) are sent as parameters to the FEMD function. For each test image, the entry in the database which yields a minimum FEMD value is its gesture value. Now, it is checked with its actual gesture value . Accordingly, the recall of the algorithm is calculated as:

Recall = (number of images correctly recalled)/(total number of test images recalled) .

The above procedure is performed iteratively for 10 times and the recall of the algorithm is calculated .

6 .Results :

Serial no	Iteration 1 recall	Iteration 2 recall	Iteration 3 recall	Iteration 4 recall	Iteration 5 recall	Total accuracy/recall
1	80	90	70	90	80	82
2	80	90	80	70	50	74
3	90	90	90	90	100	92
4	90	70	80	80	90	82
5	80	70	90	100	100	88
6	90	70	100	80	70	82
7	80	80	80	70	80	78
8	80	70	90	90	90	84
9	70	90	90	90	90	86
10	90	80	80	70	80	80

7. Conclusion :

Using the novel method of FEMD , even the gestures in images with cluttered background are recognised with an average accuracy of 82.8 % . The Accuracy can be increased by implementing convex decomposition method for finger detection