Robust Part-Based Hand Gesture Recognition Using Finger-Earth Mover’s Distance

Ashwini Kolhe  
Dept. of E&TC, JSPM’s RSCOE, Thathwade, Pune, India

Prof. R.R. Itkarkar  
Dept. of E&TC, AISSMS COE, Pune, India

Dr. Anilkumar V. Nandani  
Dept. of E&TC, BVB COE Hubli, India

DOI: 10.23956/ijarcsse/V7I7/0196

Abstract— Hand gesture recognition is of great importance for human-computer interaction (HCI), because of its extensive applications in virtual reality, sign language recognition, and computer games. Despite lots of previous work, traditional vision-based hand gesture recognition methods are still far from satisfactory for real-life applications. Because of the nature of optical sensing, the quality of the captured images is sensitive to lighting conditions and cluttered backgrounds, thus optical sensor based methods are usually unable to detect and track the hands robustly, which largely affects the performance of hand gesture recognition. Compared to the entire human body, the hand is a smaller object with more complex articulations and more easily affected by segmentation errors. It is thus a very challenging problem to recognize hand gestures. This work focuses on building a robust part-based hand gesture recognition system. To handle the noisy hand shapes obtained from digital camera, we propose a novel distance metric, Finger-Earth Mover’s Distance (FEMD), to measure the dissimilarity between hand shapes. As it only matches the finger parts while not the whole hand, it can better distinguish the hand gestures of slight differences. The experiments demonstrate that proposed hand gesture recognition system’s mean accuracy is 80.4% which is measured on 6 gesture database.

Keywords— Hand Gesture Recognition; Human Computer Interaction; Finger-Earth Movers Distance; Time series curve

I. INTRODUCTION

In recent years, human action recognition has drawn increasing attention of researchers, primarily due to its potential in areas such as video surveillance, robotics, human computer interaction, user interface design, and multimedia video retrieval. Despite lots of previous work, traditional vision-based hand gesture recognition methods [1]–[3],[4] are still far from satisfactory for real-life applications. Because of the nature of optical sensing, the quality of the captured images is sensitive to lighting conditions and cluttered backgrounds, thus optical sensor based methods are usually unable to detect and track the hands robustly, which largely affects the performance of hand gesture recognition. To enable a more robust hand gesture recognition, one effective way is to use other sensors to capture the hand gesture and motion, e.g., through the data glove [2]. Unlike optical sensors, such sensors are usually more reliable and are not affected by lighting conditions or cluttered backgrounds. However, as it requires the user to wear a data glove and sometimes requires calibration, it is inconvenient to use and may hinder the natural articulation of hand gesture. Also, such data gloves are usually more expensive than optical sensors, e.g., cameras. As a result, it is not a very popular way for hand gesture recognition.

Generalized gesture recognition system as shown in Fig. 1 consists as following steps: gesture image acquisition, feature extraction, classification.

![Gesture Image](image)

Fig.1 Generalized Hand gesture recognition system
Many vision-based hand gesture recognition approaches have been proposed in the literature [5]–[7], see [8]–[9] for more complete reviews. Vision-based hand gesture recognition methods can be classified into two categories. The first category is Statistics Learning based approaches: For a dynamic gesture, by treating it as the output of a stochastic process, the hand gesture recognition can be addressed based on statistical modeling, such as PCA, HMMs [3], [4], and more advanced particle filtering [12] and condensation algorithms [13]. The second category is Rule based approaches: Rule based approaches propose a set of pre-encoded rules between input features, which are applicable for both dynamic gestures and static gestures. When testing a hand gesture, a set of features are extracted and compared with the encoded rules, the gesture with the rule that best matches the test input is outputted as the recognized gesture [18].

Unfortunately, all existing hand gesture recognition methods have constraints on the user or the environment, which greatly hinders its widespread use in real-life applications. On one hand, to infer the pose of the palm and angles of the joints, many methods use colored markers to extract high-level features, such as the fingertip, joint locations or some anchor points on the palm [14]–[15]. On the other hand, some methods proposed to represent the hand region by edges or an ellipse [16]–[17] using skin color model. However, a common problem of the methods in these two categories is the inaccurate hand segmentation: none of these methods operates well in cluttered environments due to the sensitivity of colored markers and skin color model to the background. Besides, a few studies try to first fully reconstruct the 3D hand surfaces [8], [18]–[31]. Even though the 3D data provides valuable information that can handle problems like self-occlusion, an accurate, real time and robust 3D reconstruction is still very difficult. Furthermore, the high computational cost forbids its widespread adoption. Fortunately, recent development of depth sensors (e.g., Kinect sensor) provides a robust solution to hand segmentation. However, due to the low resolution and inaccuracy of the depth map, the obtained hand contour can be quite noisy. Classic shape recognition methods are not robust to severe distortions in hand shapes. For instance, contour-based recognition approaches, such as moments, are not robust when the contour is polluted by local distortions. Skeleton-based recognition methods [11] also suffer from contour distortions, because even little noise or slight variations in the contour often severely perturb the topology of its skeletal representation. Bai et al. proposed a skeleton pruning method in [10], which makes skeleton robust to contour noise. However, skeleton-based methods still cannot deal with the ambiguity problem as the second and the third shape have more similar skeletons than that of the first and the second shape. As for the correspondence-based shape recognition methods such as shape contexts [2] and inner-distance [1], they are not effective in solving the ambiguity either, because the correspondences of the second and the last hands have more similar contexts than the first and the second one do.

The paper is organized with six sections: The first section is an introduction and related work description of hand gesture recognition. The second section gives detail of the system methodology. The third section offers information foundation of FEMD algorithm. Section four describes about the template matching. Section five provides experimental results of gesture recognition and performance analysis. Last section concludes the work.

II. SYSTEM METHODOLOGY

Methodology implemented for hand gesture recognition using FEMD algorithm is shown in fig. 2 which consists of three main steps: hand gesture image capturing, hand gesture segmentation, feature extraction and gesture classification. In order to segment the hand shape, firstly we locate the hand position using the portable webcam.

Fig. 2 Block diagram of proposed system
The hand shape is generally of 100x100 pixel resolution, with possibly severe distortions. Color based hand segmentation is used to extract the hand region from the background. After detecting the hand shape, we represent it as a time-series curve. Such a shape representation has been successfully used for the classification and clustering of shapes. The time-series curve records the relative distance between each contour vertex and a center point. We define the center point as the point with the maximal distance after Distance Transform on the shape. In 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. The time series curve Dij is obtained by using equation (1). The time-series curve captures nice topological properties of the hand, such as the finger parts. For the general web camera, no depth information is available. Therefore hand part is segmented on the basis of color thresholding or contour detection.

\[ D_{ij} = \sqrt{(X_i - C_x)^2 + (Y_i - C_y)^2} \times 0.5 \]  

(1)

Where, Dij normalized time series distance and Cx and Cy are x and y coordinates of centroid of hand region.

A. Hand Segmentation

Normally the hand region is having more concentration of red color because human skin color has red tone. Therefore color based hand gesture segmentation is used to extract the hand region from the background. For the general web camera, no depth information is available. Therefore hand part is segmented on the basis of color thresholding or contour detection. While capturing the image the background is normally kept as dark colored and plain.

B. RGB to Gray Conversion

An RGB Image consists of 3 layers R, G, B as it is clearly see through its name. It’s a 3 dimensional matrix, for example, 3 consecutive pages in your book. Where grayscale image is of only 2 dimensions, and the values ranges between 0–255 (8-bit unsigned integers). Converting RGB to binary is often used in order to find a Region of Interest -- a portion of the image that is of interest for further processing. The intention is binary, "Yes, this pixel is of interest" or "No, this pixel is not of interest". Also it saves the computational cost and memory and adds accuracy to the algorithm.

C. Morphological Filtering

If we take close look to the segmented image, on the original gray scale image we find that the segmentation is not perfectly done. Background may have some 1s which is known as background noise and hand gesture mat have some 0s that is known is gesture noise. These errors can lead to a problem in contour detection of hand gesture so we need to remove these errors. A morphological filtering approach has been applied using sequence of dilation and erosion to obtain a smooth, closed, and complete contour of a gesture. In the morphological dilation and erosion we apply a rule on a binary image. The value of any given pixel of any given pixel in output image is obtained by allying set of rules on the neighbors in the input image. The dilation and erosion operation on a binary image A and with a structuring element B defined as follow.
D. Time Series Curve Representation
In order to segment the hand shape, firstly we locate the hand position using the Digital Camera. After detecting the hand shape, we represent it as a time-series curve. Such a shape representation has been successfully used for the classification and clustering of shapes. The time-series curve records the relative distance between each contour vertex and a center point. We define the center point as the point with the maximal distance after Distance Transform on the shape. In 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. The time-series curve captures nice topological properties of the hand, such as the finger parts.

![Fig. 6](image)

The time series curve is calculated using following formula,

\[ \text{Dist} = \sqrt{(X1 - C1)^2 + (Y1 - C2)^2} \]

Where, 
X1 and y1 are x and y poison of perimeter pixel.
C1 and c2 are x and y coordinates of cancroids.

III. FINGER EARTH MOVER’S DISTANCE ALGORITHM
Rubner et al. presented a general and flexible metric, called Earth Mover’s Distance (EMD), to measure the distance between signatures or histograms. EMD is widely used in many problems such as content-based image retrieval and pattern recognition. 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 location of earth pile and hole 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 costly transportation—the movement of earth that requires the least amount of work. EMD is applied to shape matching and contour retrieval, which represents the contour by a set of local descriptive features and computes the set of correspondences with minimum EMD costs between the local features. However, the existing EMD-based contour matching algorithms have two deficiencies when applied to hand gesture recognition:

\[
\text{FEMD} = \frac{\sum_{i=1}^{\text{row}} \sum_{j=1}^{\text{col}} \text{Dij} \times \text{Tdist} (W1 - W2)}{\sum_{i=1}^{\text{row}} \sum_{j=1}^{\text{col}} \text{Tdist}}
\]

Where,
Dij= ground distance of test and template histogram
Tdist=time series curve of test image
W1, W2 =are the weight of histogram

EMD measures the distance between two histograms, two probability distribution functions. Two hand shapes are differing mainly in global features not in local features. FEMD algorithm represents the input hand by global gestures. Ground distance matrix is similarity constraint between two images based on histogram of images:

\[
\text{Dij} = \frac{(F1 - F2)^2}{0.5}
\]

Where, F1 and F2 is histogram of testing and template image.

- Two hand shapes differ mainly in global features while not local features. Besides, the large number of local features slows down the speed of contour matching. Therefore, it is better to consider global features in contour matching.
- EMD allows for partial matching, i.e., a signature and its subset are considered to be the same in EMD measure. Clearly, they should be considered different. Our Finger-Earth Mover’s Distance (FEMD) can address these two deficiencies of the contour matching methods using EMD. Different from the EMD-based algorithm which considers each local feature as a cluster, we represent the input hand by global features (the finger clusters). And we add penalty on empty holes to alleviate partial matches on global features.
IV. TEMPLATE MATCHING

We use template matching for recognition, i.e., the input hand is recognized as the class with which it has the minimum dissimilarity distance:

\[ c = \arg \min_c \text{FEMD}(H, T_c), \]

\[ C = \arg \min \text{FEMD}(H, T_c), \]

Where H is the input hand; Tc is the template of class; FEMD (H, Tc) denotes the proposed Finger-Earth Mover’s Distance between the input hand and each template.

V. EXPERIMENTAL RESULTS

The system is implemented on Personal Computer with Core i3 2.64 GHz processor, 4 GB RAM and MATLAB R2013b software. This database consists of 20 hand gestures of 20 users (10 male and 10 female) as shown in Fig 3.

![Sample Hand gesture images](image)

Fig. 7 Sample Hand gesture images

![Original hand gesture image](image)

![time series curve representation](image)

Fig. 8 a) Original hand gesture image b) time series curve representation

![Experimental result of gesture 1](image)

Fig. 9 Experimental result of gesture 1

Table I gives the mean accuracy of Hand recognition procedure. It should be noted that FEMD algorithm leads to better result.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Mean Cross validation Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LBP</td>
<td>64.00 %</td>
</tr>
<tr>
<td>SVD</td>
<td>66.00 %</td>
</tr>
<tr>
<td>FEMD</td>
<td>80.40 %</td>
</tr>
</tbody>
</table>

VI. CONCLUSION

Hand gesture recognition for real-life applications is very challenging because of its requirements on the robustness, accuracy and efficiency. In this paper, we presented a robust part-based hand gesture recognition system. A novel distance metric, Finger-Earth Mover’s Distance (FEMD), is used for dissimilarity measure, which represents the hand shape as a signature with each finger part as a cluster and penalizes the empty finger-holes. Extensive experiments on a challenging 10-gesture dataset validate that our part-based hand gesture recognition system is accurate and efficient. One
major contribution of our paper is the distance metric based on part-based representation. Traditional distance measures such as shape contexts distance and path similarity is not robust to local distortions and shape variations, since their representations, i.e., shape contexts and skeleton, are not consistent in the case of hand variations or severe local distortions. The proposed FEMD distance metric is based on a part-based representation which represents a hand shape as a signature with each finger part as a cluster. Such a representation enables the computation on the global features, thus it is robust to local distortions. And it is robust to articulation, orientation, scale changes.

REFERENCES


