Estimating the Similarity of the Objects Using Feature Vectors

Hempriya Bali  
Department of Computer Science and IT  
University of Jammu  
Jammu, India  
E-mail: hempriyabali@gmail.com

Pawanesh Abrol  
Department of Computer Science and IT  
University of Jammu  
Jammu, India  
E-mail: pawanesh.abrol@gmail.com

Abstract: To understand similarity between the two given objects is an essential part of many applications like content based image retrieval, shape based image retrieval, text classification and clustering etc. The similarity measure thus can help us in identifying the most similar object images according to the problem at hand. Some similarity measure depend upon the features of the object image, the orientation of the image, the intensity of light that it is subjected to because of which certain limitations are faced. A lot of research is going on to evaluate the similarities between the two given object images. There are many techniques proposed by different researchers that calculate variations on different parameters. However different techniques form different bases and different methodologies for estimating the similarities. It is very essential to understand the methodologies of different distance measuring techniques in order to apply the same for various applications. Some methodologies use pixel based estimation, some use feature based and some require essential pre-processing of the images. The main aim of this manuscript is to examine and compare the different similarity measuring techniques. In our work, it was found that Euclidean and Manhattan distance work on the relative position of the pixel, Jaccard distance measures similarity between the objects that are exactly similar in size and in a binary or gray format and Mahalanobis distance is feature based. The analysis of results showed that Euclidean, Manhattan and Jaccard distance are easy in computation but are variant to object rotation, whereas Mahalanobis distance is invariant to object rotation but requires high computation time.

Keywords: Similarity Measure, Mahalanobis Distance, Jaccard Distance, Manhattan Distance, Euclidean Distance and Feature Vector

I. INTRODUCTION

Image Processing is the processing of the digital images using computer algorithms to acquire some useful information from it. Digital Image Processing has been widely used in areas like: Image sharpening and restoration, Computer/Robot Vision, Remote sensing, Pattern recognition, Video processing etc. Pattern recognition is used for detecting the chosen object from an image. It is used in recognition of handwriting and recognition of images etc. A simple approach of similarity measure using distance measure techniques can be used to measure similarity present between various objects. A central problem in estimating the similarity of the objects in Image processing and Computer Vision is determining the distance between the objects. Various features on the basis of which object recognition can be done are shape, texture, color, brightness etc. Out of these it is the shape of the object that is the most predominantly used feature for shape similarity in object recognition hence, it helps in recognizing objects more dexterously.

Normally, to determine similarity through the naked eye between the objects is easier for humans. Human brain can easily detect similar objects that project their images on to the retina of the eye as we change our position with regard to the object. So far, there have been various techniques developed over the years for recognizing objects. One such approach is the shape context which makes use of a descriptor at each point on the image for solving the correspondences between them [1,2], chord context is another descriptor that considers the length of chords that are parallel and equidistant in a shape and builds their histogram in each direction [3]. Recently researchers have widely used hidden markov model (HMM) [4], fuzzy cellular automata [8] to enhance the edges of the objects, neural network [5] and similarity measure [6] in the field of shape recognition. However, shape matching & object recognition using neural network and HMM requires dataset which are very large in size and it is not always practically possible in many applications to have such large datasets.

Similarity measure works on the principle that small distances between two objects would result in high similarity and large distances between two objects would result in less similarity [6].Studies have demonstrated that it is the distance that plays a vital role in estimating the similarity between the objects based on their feature vectors [7]. A feature vector is defined as a vector containing multiple features extracted from an image. It can also be described as a vector that contains information describing the important characteristics of an object. The features may represent a pixel or an object of an image. The features forming the feature vector depend upon the need of application.
Similarity measure using distance measure techniques has been widely used in pattern recognition, shape retrieval, shape matching, hand written digit-character recognition, information retrieval, text classification and clustering, transformation of 3D surface interpolation [28] and thresholding [29]. There are various distance measure techniques and each of them comes with their own strengths and weaknesses.

Keeping these points in mind the following objectives have been set for this study which are listed as follows:

- To study and understand the various methods for different distance measuring techniques.
- Experimental analysis and framework of techniques using distance measure
- Comparative analysis of select distance measuring techniques i.e. Mahalanobis, Jaccard, Manhattan, Euclidean distance (on objects like regular and irregular shapes) and also when subjected to rotation.
- Suggestions and inferences based on the experimental analysis for appropriate applications of distance measuring techniques.

This structure of this paper is organized as follows: The following section II explains and discusses the related work and the various distance measures employed for measuring similarity of objects. Section III throws light on to the four distance measure techniques that are being used for estimating the similarity of the objects. It addresses the details and the mathematical model of the distance measuring techniques listed in section III. The proposed methodology is presented in Section IV. In Section V, the proposed methodology is implemented. Section VI shows the experimental results of comparative study and finally Section VII concludes and remarks about some of the aspects analyzed in this paper.

II. LITERATURE REVIEW

Measuring similarity can prove to be difficult. Similarity is the quantity that determines the strength of relationship that occurs between the objects or features observed. Distance measure dissimilarity. Various types of distances and similarity have been reported and each has its own characteristics and properties contributing to their strengths and weakness. Hence, with increasing distances objects are more far apart from each other in their feature space meaning that distance can also be expressed as an absolute value of dissimilarity. In general distance is defined as a quantitative variable satisfying at least three of the following conditions -

\[ d_{ij} \geq 0 \] meaning that distance is always positive or zero; \( d_{ij} = 0 \) distance is zero if and only if it is compared to itself; \( d_{ij} = d_{ji} \) distance is symmetry; \( d_{ij} = d_{ir} + d_{jr} \) distance satisfies triangular inequality and is it called metric if it obeys all the above four conditions. The relationship that occurs between similarity and dissimilarity is as given below:

\[ \bar{d}_{ij} = 1 - d_{ij} \] (1)

i.e. when similarity is one (i.e. exactly similar), the dissimilarity is zero and when the similarity is zero (i.e. very different), the dissimilarity is one [23]. Recently many distance measure techniques have been reported that are used for evaluating the similarity that is present between the objects.

So the researcher must be careful in selecting the best suitable distance measure which will fulfill their needs and applications. Previous researchers who bought about some improvement on the following distance function: Chord distance, Cosine distance, Mahalanobis distance, Trigonometric distance and Jaccard distance observed objects of many forms like photography, hand sketches and drawing etc. Distance measure techniques include Euclidean distance [10], Mahalanobis distance [13], Chord distance [4], Cosine distance [11], Trigonometric distance [12], Jaccard distance [13], Manhattan distance [15, 18], Hausdorff distance [15] and others. Among these, the Euclidean distance measure is the simplest and commonly used technique. This computes the distance between two points in the Euclidean space [9]. Image Euclidean Distance (IMED) which takes into account the spatial relationships between pixels [10]. Cosine distance is a measure of similarity between two vectors of an inner product space that measures the cosine of the angle between them [11]. Trignometric distance [12] that normalize distance of two points in image similarity. Jaccard distance [13] which compares two objects in its binary format. Manhattan distance is the sum of the horizontal and vertical components between two points taken along a grid [14]. Hausdorff distance, measures how far two subsets of a metric space are from each other. It turns the set of non-empty compact subsets of a metric space into a metric space [15]. Mahalanobis distance is a measure of the distance between a point P and a distribution D, introduced by P.C. Mahalanobis in 1936. It is a multi-dimensional generalization of the idea of measuring how many standard deviations away P is from the mean of [16]. Hence from the study it can be noted that different distance measuring techniques can be used in several applications and it can be determined that distance measuring techniques can work by introducing different sizes of objects, degree of rotation and transformations. On the basis of this the following distance measures stated in section III were considered.
III. MATHEMATICAL MODEL OF DISTANCE MEASURE TECHNIQUES

Shape is the most dominant feature of an object as it consists of lines, contours, curves, and vertices and it is normally presented by discrete set of points or set of pixel values sampled from region or internal and external contour of objects [6, 1, 18]. Similarity measure is the most commonly used technique to measure the amount of resemblance that occurs between the objects. Whereas dissimilarity refers to distance. There are many distance measures that can be used to measure the similarity and dissimilarity between the objects. Below we have explained in detail the distance measures that we have considered in our study.

A. Mahalanobis Distance

Mahalanobis distance between two points \(x = (x_1, x_2, x_3, \ldots, x_N)^T\) and \(y = (y_1, y_2, y_3, \ldots, y_N)^T\) is defined as:

\[
d_{MH}(x, y) = \sqrt{(\bar{x} - \bar{y})^T \Sigma^{-1} (\bar{x} - \bar{y})}
\]

where \(\bar{x}\) and \(\bar{y}\) are the mean of two groups of data. \((\bar{x} - \bar{y})^T\) is the transpose of mean difference. \(\Sigma^{-1}\) is the inverse of the covariance matrix of group. [13]

Mahalanobis distance (or "generalized squared inter point distance" for its squared value) can also be defined as a dissimilarity measure between two random vectors \(\bar{x}\) and \(\bar{y}\) of the same distribution with the covariance matrix S:

\[
d(\bar{x}, \bar{y}) = (\bar{x} - \bar{y})^T S^{-1} (\bar{x} - \bar{y})
\]

If the covariance matrix is the identity matrix, the Mahalanobis distance reduces to the Euclidean distance. If the covariance matrix is diagonal, then the resulting distance measure is called a normalized Euclidean distance

\[
d(\bar{x}, \bar{y}) = \sqrt{\sum_{i=1}^{N} \frac{(x_i - y_i)^2}{s_i}}
\]

where \(s_i\) is the variance of the \(x_i\) and \(y_i\) over the sample set[ 27]

The Mahalanobis distance is the quadratic multiplication of the mean difference and inverse of pooled covariance matrix [13]. It measures the dissimilarity between the mean of two groups of data of the same distribution with the covariance matrix. A variance-covariance matrix is formed after the multiplication in which the diagonal elements represent the variance and the non-diagonal as covariance. Hence it provides a way of measuring how similar some set of conditions is to a known set of conditions [26]. Covariance can be described as the measure of strength by which two random variables will move in the same direction and variance is the measure of variability in the set of data. If there is positive covariance between the variables it means, that they move in the same direction while negative covariance indicate that they move opposite to each other. If the variables are independent of each other then the value of covariance is zero. Euclidean distance is a special case of Mahalanobis distance when the variance is equal. Mahalanobis distance is efficient in classifying two groups of data yet it shows weakness of high computation time.

B. Jaccard Distance

Another distance measure which requires less computation time is the Jaccard distance. Jaccard distance is one minus the Jaccard similarity which is the ratio of the sizes of intersection and union of finite sample sets given by the formula as stated below:

\[
J(A, B) = \frac{A \cap B}{A \cup B}
\]

\[
= \frac{|A| + |B| - |A \cap B|}{|A \cup B|}
\]

where A and B are the two finite sample sets. If both A and B are empty, then \(J(A, B) = 1\).

The Jaccard distance is the complementary of the Jaccard similarity and is given by the following formula:

\[
d_J(A, B) = 1 - J(A, B) = \frac{|A \cup B| - |A \cap B|}{|A \cup B|}
\]

This is used to measure similarity between images in their binary format and does not require large sets of data for training and testing purpose. It measures the asymmetric information on the binary variables, the comparison between two vector components. Jaccard distance can only be computed between the images that are exactly similar in size and in binary or gray format. The value of Jaccard Distance obtained lies in the range of 0 - 1, 0 being the most similar and 1 being completely different. It also shows variations in the results when object transformation is applied and to overcome these short coming a lot of preprocessing task is performed on the image before computation [6].

C. Euclidean Distance

The Euclidean distance is distance between two points \(p\) and \(q\) in 2-D Euclidean space given by the formula:

\[
d(p, q) = d(q, p) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \ldots + (q_n - p_n)^2}
\]

where \(n\) is the dimensionality of the space.
where n is the no. of dimensions and p\textsubscript{i} and q\textsubscript{i} are respectively, the k\textsuperscript{th} attributes (components) of data objects p and q [10]. It is the simplest of all the distance measure because of easy formulation and less computation time but the results obtained cannot be reliable as it shows less accuracy [6]. The value of the outcomes so obtained will be a value which is either zero or greater than zero, as distance cannot be a negative value.

**D. Manhattan Distance**

The Manhattan Distance computes the sum of difference in each dimension of two vectors in n dimensional vector space. It is the sum of the absolute difference of their corresponding components. It is also called the L\textsubscript{1} distance. If \( u = (x_1, x_2, x_3, \ldots, x_n) \) and \( v = (y_1, y_2, y_3, \ldots, y_n) \) are two vectors in the n dimensional hyper plane then the Manhattan Distance MD (u, v) between two vectors u, v is given by:

\[
\text{MD}(u, v) = |x_1 - y_1| + |x_2 - y_2| + \cdots + |x_n - y_n|
\]

The outcome so obtained is interpreted as:

\[
\text{MD}(u, v) = \begin{cases} 
0, & \text{i.e. similarity is 1} \\
> 0, & \text{dissimilar by a value of } d(p, q)
\end{cases}
\]

This distance between two points in a grid is calculated by measuring the horizontal and vertical components measured along the grid lines [14, 17].

**IV. PROPOSED METHODOLOGY**

This section gives an overview of the methodology as shown in the Figure 1 given below:

The basic steps involved are:

1. Read the images Image 1 and Image 2 as shown in the Fig.1.
2. Extract/identify the feature vector:
   Extracting the various feature vectors from the image, by using which we can determine the similarity
3. Applying Mahalanobis distance measure :
   Using the obtained feature vectors we will apply the Mahalanobis algorithm on it.
4. Analyze and Compare the Results:
   Finally analyze the results obtained and compare them with the other distance measure techniques mentioned above in section III. It helps us in notifying the degree of similarity between the objects.

**V. IMPLEMENTATION**

In this study, the experimental analysis has been conducted and evaluated on four different distance measuring techniques. The present work has been implemented and tested on Matlab R2015a. For the analysis, images have been classified into two different sets i.e. regular and irregular shapes. All the images were manually drawn and were taken to be of same dimension i.e. 150×150. The first step that was performed on to these shapes was the extraction of features of each shape. These features would be used to build a feature vector for every different object taken. The features that were considered were the R, G, B components, area, no. of pixels constituting the boundary, centroid of the shape, and size of the image. The R,G,B component was determined by finding the R,G,B value of each pixel constituting the image and then taking the mean of the R,G,B components separately. The area of the shape was calculated by converting the image to black and white and then counting the no. of on pixels in the image. The feature, no. of pixels forming the boundary and the centroid of the shape was calculated by using the technique of chain codes. The chain code technique comes under the boundary or contour based shape representation technique [21,22] in which the object is represented in terms of their external characteristics (boundary). Chain codes represent the boundary of an object by using the 4 or 8 connectivity of the segments. Hence forming straight connected sequences of line segments with directions on each segment [22]. Now using these feature vectors of each object we will find find the similarity and dissimilarity that exists between them using the distance measure techniques mentioned above in section III.
VI. EXPERIMENTAL RESULTS AND OBSERVATIONS

The presented work has been applied on to the images. Two sets of data were created. One consisting of regular shapes and the other consisting of irregular shapes. The data sets of regular and irregular shapes of same dimensions are first compared to itself. Mahalanobis, Jaccard, Manhattan, and Euclidean distance measure techniques are applied on to the images. All these techniques give zero value on totally similar objects meaning that there exists no dissimilarity, hence the similarity is one. Outcomes show variation when objects were changed on the basis of size and orientation. Hence all the distance measure techniques were able to identify that the two images are exactly alike.

Table I.

<table>
<thead>
<tr>
<th>Img 1</th>
<th>Img 2</th>
<th>MH</th>
<th>MN</th>
<th>EC</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="Image" /></td>
<td><img src="image2" alt="Image" /></td>
<td>0.3433</td>
<td>45090</td>
<td>462.08</td>
</tr>
<tr>
<td><img src="image3" alt="Image" /></td>
<td><img src="image4" alt="Image" /></td>
<td>0.3493</td>
<td>36843</td>
<td>387.18</td>
</tr>
<tr>
<td><img src="image5" alt="Image" /></td>
<td><img src="image6" alt="Image" /></td>
<td>0.3475</td>
<td>32010</td>
<td>327.96</td>
</tr>
<tr>
<td><img src="image7" alt="Image" /></td>
<td><img src="image8" alt="Image" /></td>
<td>0.3482</td>
<td>43228</td>
<td>392.33</td>
</tr>
<tr>
<td><img src="image9" alt="Image" /></td>
<td><img src="image10" alt="Image" /></td>
<td>0.3456</td>
<td>43138</td>
<td>365.84</td>
</tr>
</tbody>
</table>

In Table I, the data set of regular shapes has been compared with the same regular shapes that are slightly changed by reducing their size to approximately half of their original size. By proportionally reducing the size of the objects the results show variation as it is evident from Table I for three distance measuring techniques. Here, MH means Mahalanobis, MN means Manhattan and EC means Euclidean distance. From Table I to Table IV we observed that the value of Jaccard distance obtained are zero for all the images meaning that high similarity exists between the two images, which is not possible in every case. This happens because of the multiple features present in the feature vector. Jaccard distance takes these non-zero values of the features as one forming a binary vector. Using this binary vector the Jaccard distance obtained is zero using the formula stated above in Section III. Hence feature vector cannot be measured to use Jaccard distance.

Jaccard distance can only compare two images when present in their binary format and are of the same size as shown in Table V.

In Table II the images being compared are completely different. The values of Mahalanobis distance in Table I are greater than the values in Table II this is because the differences between the various features of the two images in Table I are greater compared to the features of Table II.

Table II

<table>
<thead>
<tr>
<th>Img 1</th>
<th>Img 2</th>
<th>MH</th>
<th>MN</th>
<th>EC</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image11" alt="Image" /></td>
<td><img src="image12" alt="Image" /></td>
<td>0.0080</td>
<td>3109476</td>
<td>2058.2</td>
</tr>
<tr>
<td><img src="image13" alt="Image" /></td>
<td><img src="image14" alt="Image" /></td>
<td>0.1108</td>
<td>2763869</td>
<td>2539.6</td>
</tr>
<tr>
<td><img src="image15" alt="Image" /></td>
<td><img src="image16" alt="Image" /></td>
<td>0.0104</td>
<td>1101860</td>
<td>2194.9</td>
</tr>
<tr>
<td><img src="image17" alt="Image" /></td>
<td><img src="image18" alt="Image" /></td>
<td>0.0341</td>
<td>2671333</td>
<td>2250.2</td>
</tr>
<tr>
<td><img src="image19" alt="Image" /></td>
<td><img src="image20" alt="Image" /></td>
<td>0.0159</td>
<td>2227841</td>
<td>2758.6</td>
</tr>
</tbody>
</table>

The Mahalanobis distance calculated in Table I and Table II is done by taking into consideration the feature vector. It gives us an insight of the correlation and variance that exists between the variable as discussed above in Section III. Whereas Euclidean distance has been calculated by finding the distance between the color pixel values of the two images. Similarly Manhattan distance is computed between the color pixel intensity of the two images by using the equation (9) stated above in Section III. From Table I and Table II it is observed that the values of Euclidean and Manhattan distance vary largely. In Table I the values for Euclidean and Manhattan distance are smaller than the values in Table II. Hence the images offer higher similarity in the former than the later table.

Data set of irregular shapes have been compared with the same shape by bringing slight variation in it over the distance measure techniques as shown Table III.

In Table IV completely different irregular shapes are being compared. From Table III and Table IV it can be seen that the Mahalanobis distance value for Table IV is greater for all the cases than in Table III, depicting that the measure of compared.
From Table III and Table IV it can be seen that similarity is higher in Table III than in Table IV. Also Euclidean and Manhattan distance values show high similarity in Table III than in Table IV.

In Table V, an image is compared with itself which is rotated at various angles. These distance measure techniques were applied on these three cases of regular and irregular shapes. The values obtained for Mahalanobis distance is zero, the value for Jaccard distance lies between 0-1. Euclidean and Manhattan distance show a value greater than zero. Hence Mahalanobis distance can be used for searching and retrieval of images from the database even if they are rotated by any angle because Mahalanobis distance shows zero dissimilarity hence similarity is one meaning that the two images are exactly alike. This is given by the equation (3) stated above Jaccard distance may be used in such applications where the extent of transformation is to be ascertained as it is evident from Table V. The value of Jaccard shows variation when it is transformed otherwise it remains zero. Euclidean and Manhattan distance are evaluated on the basis of the relative position of the pixels.

### VII. CONCLUSION

This paper presents the comparative study of estimating the similarity that exists between the objects using distance measuring techniques like Mahalanobis, Jaccard, Manhattan and Euclidean distance by calculating the distance that exists between the two images. The proposed methodology has been tested on regular and irregular shapes. The experimental results show that Mahalanobis distance is feature based i.e. it takes into account the correlation that exists between the variables whereas the other distance measures like Jaccard, Manhattan and Euclidean distance do not offer any correlation between the variables. Euclidean and Manhattan are easy to compute but offer lowest degree of accuracy. Jaccard outshines Euclidean and Manhattan distance due to its simplicity and high accuracy but works efficiently on binary images. The drawback of Jaccard is that it involves a lot of preprocessing task prior to finding similarity between the objects. Mahalanobis distance works well even on rotation of objects and hence Mahalanobis distance is invariant to object rotation whereas Jaccard, Manhattan and Euclidean distance measuring techniques are variant to object rotation hence
adding to their list of disadvantages. The drawback of Mahalanobis distance is that it offers high computation time. Euclidean and Manhattan do not always provide with the best matching results compared to other distance measures. Future work, may involve evaluation of four distance measure techniques on large image dataset. Futher, the analysis may also be done for evaluating the efficiency and applicability in online and real time environment.

REFERENCES


[24] De Maesschalck, Roy; Jouan-Rimbaud, Delphine; and Massart, Désiré L. (2000); The Mahalanobis distance, Chemometrics and Intelligent Laboratory Systems 50:1–18

[25] Jump up Gnanadesikan, Ramanathan; and Kettenring, John R. (1972); Robust estimates, residuals, and outlier detection with multireponse data, Biometrics 28:81–124


