The Study of Different Similarity Measure Techniques in Recognition of Handwritten Characters

C. Naveena, V.N. Manjunath Aradhya
Dept. of Information Science
Dayananda Sagar College of Engineering
Bengaluru, India
naveena.cse@gmail.com
aradhya.mysore@gmail.com

S.K. Niranjan
Dept. of MCA
Sri Jayachamarajendra College of Engineering
Mysore, India
sriniranjan@gmail.com

ABSTRACT
In this paper, we compare the affect of four different similarity measure techniques namely Euclidean distance, Modified squared euclidean distance, Correlation distance and Angle distance for an unconstrained handwritten character recognition. The strength of these similarity measures are estimated between feature vectors with respect to the recognition performance of the Gabor-PCA method. Gabor filter is used to extract spatially localized features of character image. The dimensions of such Gabor feature vector is prohibitively high & in order to compress Gabor features we used PCA method. The experiments were performed using the database containing 22,600 samples of Kannada and English. From the analysis the better recognition accuracy were achieved using angle distance measure.

Categories and Subject Descriptors
I.4 [IMAGE PROCESSING AND COMPUTER VISION]; I.5 [PATTERN RECOGNITION]

Keywords
Handwritten Character Recognition (HCR); Gabor filter; PCA; Distance Measures

1. INTRODUCTION
Optical Character Recognition (OCR) is a very well suited problem in the vast area of pattern recognition. The first commercial OCR systems begin to appear in the early 1950s and soon they were being used in US postal service to sort mails. The accurate recognition of Latin-script, typewritten text is now considered largely a solved problem on the applications where clear image is available such as scanning of printed documents. The typical accuracy rates on these exceed 99%; total accuracy can only be acheived by human review. Other area including recognition of hand printing, cursive handwriting, and printed text in other scripts (especially those with very large number of characters) are the still subject area of an active research.

As per the literature survey many works have been observed on recognition of handwritten characters. Pal et al. [18] proposed a system for offline Oriya handwritten character recognition using curvature features. Curvature features of a character image is extracted using bi-quadratic interpolation method and quantized into 3 levels according to concave, linear and convex regions respectively. The 3 levels curvature features of the feature vector has higher dimensionality. To reduce this higher dimensionality feature vector PCA method is applied. Mane and Ragha [13] presented a handwritten character recognition using elastic image matching and PCA. The elastic image matching (EM) technique is based on the eigen-deformations, which are used for character recognition. The eigen-deformations are estimated by the principal component analysis of actual deformations, which are automatically collected by the elastic image matching technique. For the purpose of classification similarity measures techniques, euclidean and Mahalanobis similarity measures are used. Niranjan et al. [17] described that the FLD based unconstrained handwritten Kannada character recognition. The technique extracts features from FLD, two dimensional FLD (2D-FLD) and diagonal FLD. In order to classify the characters, different distance measure techniques are used. Leung and Leung [11] presented a recognition of handwritten Chinese characters by critical region analysis. The algorithm critical region analysis is used to distinguish one character from another similar character.

The critical regions are identified based on the output of the Fisher's discriminant.

Aradhya et al. [2] proposed that the robust unconstrained handwritten Digit recognition using Radon transform. In this technique radon function is used to represents an image as a collection of projections along various directions. The resultant feature vector is the input for the classification.
stage. A nearest neighbor classifier is used for the subsequent recognition purpose. Nemmour and Chibani [16] described that the handwritten Arabic word recognition based on Ridgelet transform and support vector machines. The technique describes the concept of Ridgelets in generating pertinent features of handwritten words. Mahmoud and Abu-Amara [12] presented that the use of radon transform in handwritten Arabic (Indian) numerals recognition. The algorithm is based on Radon-Fourier method, which is used for feature extraction and to represent Arabic digits.

Recently, Gabor filter is widely used in image processing, computer vision, neuroscience, psychophysics and pattern recognition applications because of its nature of less sensitivity to noises and small amount of translation, rotation and scaling properties. Also, The Gabor filter is one of the powerful tool for extracting spatially localized features. Based on this many works have been proposed on handwritten character recognition. Hamid et al.[6] proposed an Arabic character recognition using Gabor filters. Su and Wang [22] proposed a novel stroke extraction method for Chinese character recognition using Gabor filters. In this approach Gabor filter is used to break down an image of a character into different directional features. Hu et al [7] presented a recognition of gray characters using gabor filters. With these effective properties of the Gabor filter, in this paper we chose Gabor filter as a pre-processing step.

Similarity measure technique is one of the classification method in pattern recognition system. Normally, Euclidean distance is used for similarity measures. Cha [5] proposed a comprehensive survey on Distance/Similarity measures between probability density functions. Distance/similarity measures are fluid and promulgated differently because these measures depends on the measurement type or representation of the objects. The Probability density function (pdf) is one of the most popular function for reviewed and categorized in both syntactic and semantic relationships in pattern representation applications. We have motivated from the properties of different similarity measures in this article, to compare recognition performance of the four similarity measures namely, Euclidean distance, Modified squared euclidean distance, Correlation distance and Angle distance using Gabor-PCA features.

The organization of the paper is as follows: In Section 2, we describe the concept of proposed method and its advantages. In Section 3, explained about properties of various similarity measure techniques for classification. In Section 4, experimental results and analysis were shown in detail. Finally, describes the conclusion of our study.

2. PROPOSED METHOD

The proposed method consists of two stages. The first stage consists of Gabor filter & in the second stage, feature extraction is performed using PCA.

2.1 Gabor Filter

Spatial frequencies and their orientations are important characteristics of textures in images. The frequency characteristics of images can be analyzed using spectral decomposition methods like Fourier analysis. Fourier analysis has proven to be one of the most powerful tools in signal processing. However, a key problem with Fourier analysis is that spectral features from different parts of the image are mixed together. For the image analysis and pattern recognition application like character recognition, face recognition, etc., require spatially localized features. Gabor filters are a popular tool for this task of extracting spatially localized features.

Gabor filter works a bandpass filter for local spatial distribution, achieving an optimal resolution in both spatial and frequency domains. The 2D Gabor filter \( G(x, y, f, \theta) \) can be represented as a complex sinusoidal signal modulated by a Gaussian kernel function as follows [4].

\[
G(x, y, f, \theta) = \exp \left[ -\frac{1}{2} \left( \frac{x'}{\sigma_x'} \right)^2 + \left( \frac{y'}{\sigma_y'} \right)^2 \right] \exp(2 \pi f x')
\]  
(1)

\[
x' = x \cdot \sin \theta + y \cdot \cos \theta
\]  
(2)

\[
y' = x \cdot \cos \theta - y \cdot \sin \theta
\]  
(3)

where \( f \) is the frequency of the sinusoidal plane wave along the direction \( \theta \) from the x-axis and \( \sigma_x' \) & \( \sigma_y' \) are the space constants of the gaussian envelope along \( x \) and \( y \) axes respectively.

An even symmetric Gabor filter has the following general form in the spatial domain.

\[
G(x, y, f, \theta) = \exp \left[ -\frac{1}{2} \left( \frac{x'}{\sigma_x'} \right)^2 + \left( \frac{y'}{\sigma_y'} \right)^2 \right] \cos(2 \pi f x')
\]  
(4)

Design of the Gabor filter is a accomplished by tuning the filter with a specific band of spatial frequency and orientation by appropriately selecting the filter parameter. The spread of the filter \( \sigma_x' \), \( \sigma_y' \), radial frequency \( f \) and the orientation of the field \( \theta [8, 9] \). Hence in this work, The filtering is performed in the frequency \( f \) value is 4 and orientation parameters \( \theta \in \frac{\pi}{4}, \frac{3\pi}{4}, \frac{5\pi}{4}, \frac{7\pi}{4} \). The values of \( \sigma_x' \) and \( \sigma_y' \) were empirically determined and each is set to 2 and 4 respectively. The resultant Gabor filter images are as shown in the figures 1(b-e). These different orientation features are will be subsequently used by subspace methods. Which reduces the working dimension space significantly.

2.2 Subspace Methods

Subspace methods are important for pattern recognition systems to employ an effective feature extraction scheme to enhance separability between pattern classes which should maintain and enhance features of the input data that make distinct pattern classes separable [10]. In general, there exist a number of different feature extraction methods. The most common feature extraction methods are subspace analysis methods such as principal component analysis (PCA) [24,
Figure 1: The results of using the Gabor filtering process: (a) an original image; (b) horizontal feature image (0 deg orientation); (c) right-diagonal feature image (pi/4 deg orientation); (d) vertical feature image (pi/2 deg orientation); and (e) left-diagonal feature image (3pi/4 deg orientation).

2.3 Advantage of using combination of Gabor and PCA

Gabor filter satisfy the minimum space-bandwidth product per the uncertainty principle. Hence, it provide simultaneous optimal resolution in both the space and spatial-frequency domains. A central issue in applying Gabor filter to texture segmentation in the determination of the filter parameters [25]. In our experiment we varied the orientation parameter θ(explained in Section 2.). From this variation in the parameter θ we obtained the texture features according to orientation. This texture feature set dimension is likely high and it take more computation time. To overcome this problem, we used to PCA to reduce dimension and extracts most efficient features to represent a character image. In order to study the feature behavior property of Gabor-PCA and PCA, we selected 50 dominant features.Figure 2 and 3 shows the data sequence diagrams of Gabor-PCA and PCA methods.

Figure 2: Inter class data sequences obtained for Gabor-PCA and PCA methods

\[ Y = U^T X \] (6)

3. SIMILARITY MEASURES

To classify an unknown vector by calculation of its class-distance to predefined classes, which in turn are defined by the distance to their individual prototype vectors. Now we will briefly look at some commonly used distance measures. Depending on the application at hand, each of the distance measures has its pros and cons, and we will discuss their important properties in next subsections. There are several properties that most of the useful distance measures have:

1. Symmetry, \( d(x, y) = d(y, x) \).
2. Positive definitive, \( d(x, y) > 0 \) for \( x \neq y \) and \( d(x, x) = 0 \).
3. There is the triangle inequality which may sometimes be useful and which makes $d(x, y)$ a metric.
- $d(x, y) \leq d(x, z) + d(y, z)$

4. In addition the following properties are usually described in a distance measure.
- $d(x, y)$ should have a physically meaningful interpretation.
- $d(x, y)$ should be efficiently computable.

### 3.1 Euclidean distance & Modified Squared Euclidean distance

In mathematics, Euclidean distance refers to the distance between two points as measured in a straight line. This can be proven with repeated application of the Pythagorean Theorem. In a plane with $p_1$ at $(x_1, y_1)$ and $p_2$ at $(x_2, y_2)$, the Euclidean distance is given below

$$d(x, y) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$  \hspace{1cm} (7)

Euclidean distance is used to express the distance between any two points in any metric space is given below.

$$d(x, y) = \sqrt{\sum_{i=0}^{n} (x_i - y_i)^2}$$  \hspace{1cm} (8)

In classification, Euclidean distance is used to express the difference between user preferences. This difference is used as a measure of correlation between user preferences. The smaller the difference, the closer the correlation between user preferences. The greater the distance, the less correlation exists between user preferences. The main advantage of this distance measure is to compare the relationship between actual ratings. This means that the Euclidean distance is a fair measure of how similar ratings are for specific preferences or items. The main disadvantages of euclidean distance is, can’t determine the correlation between user profiles who have similar trends in tastes, but different ratings for some of the same items.

To improve the correlation property of the Euclidean distance, the equation 8 is modified into given below.

$$d(x, y) = \frac{\sum_{i=1}^{n} (x_i - y_i)^2}{\sum_{i=1}^{n} (x_i)^2 \sum_{i=1}^{n} (y_i)^2}$$  \hspace{1cm} (9)

The above equation called as a Modified Squared Euclidean distance.

### 3.2 Correlation distance

It is measure of extent to which two samples are linearly related. This correlation coefficient can be used calculate a distance between the samples using the formula

$$D = (1 - r)$$  \hspace{1cm} (10)

In general, sample $x_k$ has $n$ measurements, or features it can be written as $x_k = x_{1k}, x_{2k}, \ldots, x_{nk}$. There are many formulas that can be used calculate $r$, and the most detailed formula is given below.

$$r = \frac{n \sum_{i=1}^{n} x_{ik} x_{im} - (\sum_{i=1}^{n} x_{ik})(\sum_{i=1}^{n} x_{im})}{\sqrt{(n \sum_{i=1}^{n} x_{ik}^2 - (\sum_{i=1}^{n} x_{ik})^2)[n \sum_{i=1}^{n} x_{im}^2 - (\sum_{i=1}^{n} x_{im})^2]}}$$  \hspace{1cm} (11)

If a linear transformation is used on each sample such that the set of $n$ measurements become standard or z-scores, $x_k$ becomes $z_k$ and each $x_{ik}$ becomes $z_{ik}$ in this expression. This transformation to z-scores is such that the mean is zero and the standard deviation is one.

Therefore,

$$\sum_{i=1}^{n} z_{ik} = 0$$  \hspace{1cm} (12)

and

$$\sum_{i=1}^{n} z_{ik}^2 = 0$$  \hspace{1cm} (13)

which means that the expression for $r$ reduces to

$$r = \frac{\sum_{i=1}^{n} z_{ik} z_{im}}{n}$$  \hspace{1cm} (14)

Therefore, the correlation coefficient can be described as the dot-product of two standard vectors divided by the rank of the these vectors.

The problem with this analysis is that large changes can occur in the measurements when they are standardized if
the number of the measurements is small, and this is mainly due to setting the mean value of the measurements to zero. The probability of obtaining an $r$-value that is not representative of the true similarity or difference between samples decreases as the number of measured values increases, this probability is never zero. Therefore, any application of Pearson's correlation coefficient, whether determining the similarity or distance between samples, always has a non-zero probability of producing an incorrect value. This correlation coefficient is regularly used and the user should be aware that a problem may exist in the results.

3.3 Angle based Distance

It is also called as coefficient of correlation. The degree or level of correlation is measured with the help of correlation coefficient or coefficient of correlation. For population data, the correlation coefficient is denoted by the joint variation of $X$ and $Y$ is measured by the covariance of $X$ and $Y$. The covariance of $X$ and $Y$ denoted by $\text{Cov}(X, Y)$ is defined as:

$$\text{Cov}(X, Y) = E[X - E(X)Y - E(Y)]$$  \hspace{1cm} (15)

The $\text{Cov}(X, Y)$ may be positive, negative or zero. The covariance has the same units in which $X$ and $Y$ are measured. When $\text{Cov}(X, Y)$ is divided by $\sigma_X$ and $\sigma_Y$, we get the correlation coefficient $\rho$.

Thus, $\rho = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$, $\rho$ is free of the units of measurement. It is a pure number and lies between -1 and +1. If, $\rho = \pm 1$ it is called perfect correlation. If, $\rho = -1$ it is called perfect negative correlation. If there is no correlation between $X$ and $Y$, then $X$ and $Y$ are independent and. For sample data the correlation coefficient denoted by $r$ is a measure of strength of the linear relation between $X$ and $Y$ variables, where $r$ is a pure number and lies between -1 and +1. On the other hand Karl Pearson’s coefficient of correlation is:

$$r = \frac{\sum_{i=1}^{n}X_iY_i}{\sqrt{\sum_{i=1}^{n}X_i^2 \sum_{i=1}^{n}Y_i^2}}$$  \hspace{1cm} (16)

4. EXPERIMENTAL RESULTS

The effectiveness of four different similarity measures are tested with a dataset containing handwritten Kannada and English characters. Most of the available work on handwritten character recognition of Indian scripts are based on the small databases and collected in laboratory environments. Hence in this regards, we created a very large database comprising of Kannada and English characters of 22,600 samples [15]. The database has been created with the assistance of 100 writers of different streams such as school children, degree students, university students of various age group. The dataset holds 200+26=226 classes (Kannada (vowels, consonants and modifiers) and English lower Case is considered) where each class intern contains 100 samples written by individual writer. Some of the sample images of handwritten Kannada vowels, consonants and modifiers are shown in Figure 4(a) (b) and (c) respectively. Sample images of English is shown in Figure 4 (d).

We carried out the experiment in five stages. In each stage, we varied class size $n$ (where $n = 50, 100, 150, 200, 226$) and defined each class containing 100 samples. In every stage of our experiment, the system was trained by varying sample number 25,50,75 and remaining samples of each character class are used for testing. Each experiment is repeated 5 times by varying number of projection vectors $p$ (where $p = 10, 20, 30, 40, 50$). Since $p$, has a considerable impact on recognition accuracy, we chose the value that corresponds to best classification result on an image set. All our experiments were carried out on a PC machine with Core 2 duo processor, 2.2GHz CPU and 1GB RAM memory under Matlab 10.0 platform.

The effectiveness of the similarity measures techniques and proposed method are compared with well known existing PCA [24, 23], FLD [3] and Ridgelet-PCA [15] methods. The recognition accuracy of the proposed method and existing methods are shown in Table 1 to Table 5. The obtained recognition accuracy is based on 50,100,150,200,226 classes of 75 samples for training & 23 samples for testing. From the experimental results, it is clear that the proposed method has higher recognition accuracy when compared with existing methods. From the analysis of different stages of experiment, it is clear that the angle distance measure yields better recognition accuracy compared to other distance measure techniques.

4.1 Analysis on Angle based Distance

An angular range query is defined by $(Q, \alpha)$ where $Q$ is the query point $(q_1, q_2, ..., q_d)$ in a d-dimensional space and $\alpha$ denotes the angle that represents the range, and seeks all data in the cone whose axis $(\hat{O}Q)$ is the line defined by the origin $O$ and the query point $Q$ and whose apex or the vertex is on the origin, as illustrated in Figure 5. The angle between the axis and all the lines on the lateral surface of the cone is $\alpha$. All the feature vectors that are equally similar to the query vector are on an equivalence region which corresponds to a conic surface [1].

If a feature vector is represented as $X(x_1, x_2, ..., x_d)$, the cosine angle measure (a widely used similarity measure) is defined by the following formula:

$$\cos(\alpha) = \frac{\sum_{i=1}^{d} x_i q_i}{\|X\| \cdot \|Q\|}$$  \hspace{1cm} (17)
Without loss of generality, if we assume the query point to be normalized and also feature vectors are also normalized, then the equation becomes the inner product of the query with a feature vector in the domain. Similarly, angle based distance measure (Karl Pearson’s correlation coefficient (refer equation 16)) also defined as the inner-product of two vectors are standardized, i.e., the means of the new vectors are 0 and the standard deviations are 1. From this standardization property, angle based distance similarity measure give more recognition results compare to other distance measure.

5. CONCLUSION
In this paper, we studied the performance effectiveness of four different distance similarity measure techniques namely Euclidean distance, Modified squared euclidean distance, Correlation distance and Angular distance for an unconstrained handwritten character recognition. The strength of the four distance similarity measure techniques are determined by the Gabor-PCA. Angle distance measures leads to higher accuracy because of its data standardization property. The implementation of our proposed method is compared with well known PCA, FLD and Ridgelet methods and obtained better recognition accuracy because of its convergence property.

6. REFERENCES

Table 1: Recognition accuracy obtained for 50 class problem.

Table 2: Recognition accuracy obtained for 100 class problem.

Table 3: Recognition accuracy obtained for 150 class problem.

Table 4: Recognition accuracy obtained for 200 class problem.

Table 5: Recognition accuracy obtained for 226 class problem.


[10] T. Jan. Learning and Generalization using Kernel Based Subspace Model Approaches. Department of


