1-1-2005

Enhancing Binary Feature Vector Similarity Measures

Sung-Hyuk Cha
Pace University

Sungsoon Yoon
Pace University

Charles C. Tappert
Pace University

SCHOOL OF COMPUTER SCIENCE AND INFORMATION SYSTEMS

TECHNICAL REPORT

Number 210, January 2005

Enhancing Binary Feature Vector Similarity Measures

Sung-Hyuk Cha
Sungsoo Yoon
Charles C. Tappert

PACE
UNIVERSITY

Sung-Hyuk Cha is Assistant Professor of Computer Science at Pace University, based in Westchester. Dr. Cha holds baccalaureate and masters degrees from Rutgers University and a doctorate in computer science from the State University of New York at Buffalo. He joined the faculty at Pace University in September, 2001.

Dr. Cha's research interests are in the areas of distance measure and pattern matching algorithms, pattern recognition, image analysis, and machine intelligence and data mining.

Sungsoo Yoon came to Pace in 2004 as an Adjunct Professor in Computer Science. Dr. Yoon holds the Ph.D. in computer science, with supplemental study in psychology, from Yonsei University in Korea. His research interests include pattern recognition, image processing, computer vision based security (biometrics), bioinformatics, character recognition, document analysis and recognition, and cognitive science.

Before coming to Pace, Dr. Yoon was a researcher at the Biometrics Engineering Research Center (Korea) working on the development of technology for face recognition. Dr. Yoon has contributed over twenty journal articles and conference presentations.

Charles C. Tappert has a B.S. in Engineering Sciences from Swarthmore College, an M.S. and Ph.D. in Electrical Engineering from Cornell University, and was a Fulbright Scholar. He worked at IBM for 26 years, mostly at the T.J. Watson Research Center, on speech recognition and pen computing. He has over 100 publications: journal articles, book chapters, conference papers, patents, and technical disclosures. While at IBM, he taught part-time as an adjunct at Pace, SUNY Purchase, and North Carolina State University. He taught full-time at the U.S. Military Academy at West Point for seven years before joining Pace University in 2000 as Professor of Computer Science.

At Pace Dr. Tappert has been involved in the development of the Doctorate of Professional Studies in Computing, for which he is the Associate Program Chairperson. In addition, he has substantially enhanced the capstone seminar in software engineering of the Masters in Computer Science program with real projects for real customers. His research interests include pattern recognition, pen computing and voice applications, graphics, algorithms, artificial intelligence, human-computer interaction, and e-commerce.

Enhancing Binary Feature Vector Similarity Measures

Sung-Hyuk Cha, Sungsoo Yoon, and Charles C. Tappert

School of Computer Science and Information Systems, Pace University
861 Bedford Rd. Pleasantville, NY 10570 USA
scha@pace.edu, ssyoon@csai.yonsei.ac.kr,ctappert@pace.edu

Abstract

Similarity and dissimilarity measures play an important role in pattern classification and clustering. For a century, researchers have searched for a good measure. Here, we review, categorize, and evaluate various binary vector similarity/dissimilarity measures. One of the most contentious disputes in the similarity measure selection problem is whether the measure includes or excludes negative matches. While inner-product based similarity measures consider only positive matches, other conventional measures credit both positive and negative matches equally. Hence, we propose an enhanced similarity measure that gives variable credits and show that it is superior to conventional measures in iris biometric authentication and offline handwritten character recognition applications. Finally, the proposed similarity measure can be further boosted by applying weights and we demonstrate that it outperforms the weighted *Hamming* distance.

Key words: *Binary Similarity, Distance Metric, Nearest neighbor, Genetic Algorithm, Iris Biometric Verification, Handwriting Recognition*

1. Introduction

A common method for classifying an unknown input vector involves finding the top *k* similar vectors in a reference set. The *k*-nearest neighbor, or simply *k-nn*, has wide acceptance in pattern classification problems (see [1, 2] for extensive surveys). There are two important aspects in this approach. One is extracting important features from the pattern, and the other is selecting an appropriate similarity measure. Although there are many types of features, in this paper we consider only binary features and their similarity/dissimilarity measures.

Distance and similarity measures are encountered in various fields such as image retrieval [3], information retrieval, chemistry [4], ecology [5], psychology, and biological taxonomy [6]. There is a wealth of literature regarding the similarity measure selection problem dating as far back as 1908 (see [7, 8] for extensive lists of similarity measures). Conventional definitions of similarity or dissimilarity measures include *Hamming* [9], *inner-product, Tanimoto distance*, etc., and for a century, researchers have searched for a good measure. The most recent comparative works on similarity measures include Tubbs [10] who summarized seven conventional similarity measures for the template matching problem [10], and Zhang et al. who compared these seven measures for their recognition capability in handwriting identification [11]. Yet another distance measure that has been used for binary features extracted from offline character images can be found in [12, 13, 14]. Here, we categorize and review various similarity and distance measures.

The most favored distance measure is the *Hamming distance* when the featurs are binary. To further improve the performance, there are two approaches. First, weights can be applied to features [14] and optimized using techniques such as genetic algorithms [15, 16]. Another approach is to use a similarity measure that gives full credit to features present in both patterns, less credit to those not present in either pattern, and no credit to those present in only one of the patterns to be matched [14]. Both approaches have been reported to perform better than the simple Hamming distance approach. In this paper, we create a new measure that combines these two approaches, and we present experimental results that demonstrate its superiority over the other measures.

One of the most contentious problems in binary feature vector similarity measures is whether the measure includes or excludes the number of *negative matches*. When features are absent in both patterns, should we consider it as important as those present in both patterns? This problem has been argued in [7, 17] and several binary vector coefficients include the negative matches as well as the positive matches equally. Here, the proposed measure includes both positive and negative matches, and has weights that can be optimized to control the degree to which the positive and negative matches are considered.

To evaluate similarity measures for binary features, we chose two binary feature databases: an *iriscode* database and an offline handwritten character database. First, we consider the problem of iris biometric verification which often uses a distance or similarity between two samples of the same class and between samples of two different classes. Two patterns are categorized into one of only two classes – the patterns are either from the same class or from two different classes. Given two iris biometric samples, the feature distance between the two samples is classified as intra-person (identity) or inter-person (non-identity). The intra and inter-person distances form two distributions having some overlap with each other. Two types of errors, *False Accept Rate* (*FAR*) and *False Reject Rate* (*FRR*), are used to evaluate the various similarity measures.

In the offline handwritten character image database, *Gradient, Structural, and Concavity* binary features, or simply *GSC*, have been developed and utilized in character recognition [12]. While GSC features, which are binary in type, are considered significant ones, relatively little study has been conducted on selecting and designing a good similarity measure for these features. In this paper, we evaluate numerous similarity measures and determine the optimal one.

The subsequent sections are organized as follows. Section 2 enumerates many similarity measures and their weighted variations. Sections 3 and 4 evaluate similarity measures on *iriscode* and offline handwritten character databases, respectively. Finally, section 5 concludes the paper.

2. Preliminary

In this section, we give the definitions of conventional binary vector similarity and dissimilarity (distance) measures and then show how some of these measures can be refined with weights that can be optimized to enhance their discrimination capability.

2.1. Basic Binary Similarity Measures

Let x, y, and z be binary feature vectors of fixed length d, and let x_i denote the ith feature value which is either 0 or 1. One of the most popular measures in comparing two fixed-length bit patterns is the Hamming distance in eqn (1), which is the count of the bits that differ in the two patterns [9]. It is a simple geometrical L_1 distance, also known as Manhattan or city block distance, applied to d-dimensional binary space.

$$\begin{aligned} &D_{\text{Hamming}}(x,y) = x'\overline{y} + \overline{x}'y \\ &D_{\text{Hamming}}(x,y) = \sum_{i=1}^{d} |x_i - y_i| \\ &S_{\text{Hamming}}(x,y) = d - D_{\text{Hamming}}(x,y) = x'y + \overline{x}'\overline{y} \\ &S_{\text{Hamming}}(x,y) = \sum_{i=1}^{d} s_i \\ &\quad \text{where } s_i = \begin{cases} 1 & \text{if } x_i = y_i \\ 0 & \text{otherwise} \end{cases} \end{aligned} \tag{1}$$

The term $x'y$ denotes the positive matches, i.e., the number of 1 bits that match between x and y. The term $\overline{x}'\overline{y}$ is the negative matches, i.e., the number of 0 matching bits. The terms $x'\overline{y}$ and $\overline{x}'y$ denote the number of bit mismatches – the first where pattern x has 1 and pattern y has 0, and the second where pattern x has 0 and pattern y has 1.

Fact 1. The Hamming distance has been shown to be metric [6].

While the Hamming distance is the number of bits differing in the two patterns, the Hamming similarity is the number of identical bits in the two patterns. Sokal and Michener normalized the Hamming similarity as in eqn (2) [18], and an alternative normalized Hamming similarity is given by Rogers and Tanimoto in eqn (3) [19].

$$S_{\text{Sokal-Michener}}(x,y) = \frac{x'y + \overline{x}'\overline{y}}{d} \tag{2}$$

$$S_{\text{Rogers-Tanimoto}}(x,y) = \frac{x'y + \overline{x}'\overline{y}}{x'y + \overline{x}'\overline{y} + 2x'\overline{y} + 2\overline{x}'y} \tag{3}$$

The term $x'y$ is the inner product of two vectors, which yields a scalar, and it is sometimes called the scalar product or dot product. It can be converted to a distance by subtracting it from d, and this distance is clearly non-metric because of the reflexivity violation; $D_{\text{inner-product}}(x,y) = 0$ iff $x = y$ and $|x| = |y| = d$ and $D_{\text{inner-product}}(x,y) > 0$, otherwise.

Fact 2. Nonnegativity, symmetry, and triangle inequality are trivial and preserved in the inner product [20].

$$\begin{aligned} &S_{\text{inner-product}}(x,y) = x'y \\ &S_{\text{inner-product}}(x,y) = \sum_{i=1}^{d} s_i \\ &\quad \text{where } s_i = \begin{cases} 1 & \text{if } x_i = y_i = 1 \\ 0 & \text{otherwise} \end{cases} \end{aligned} \tag{4}$$

A normalized inner product is given in eqn (5) [6] and various alternative normalizations in eqns (6~9) [7,21-23].

$$S_{\text{normalized-inner-product}}(x,y) = \frac{x'y}{\|x\|\|y\|} = \frac{x'y}{\sqrt{x'xy'y}} \tag{5}$$

$$S_{\text{Russell-Rao}}(x,y) = \frac{x'y}{d} \tag{6}$$

$$S_{\text{Jaccard-Needham}}(x,y) = \frac{x'y}{x'y + x'\overline{y} + \overline{x}'y} \tag{7}$$

$$S_{\text{Dice}}(x,y) = \frac{x'y}{2x'y + x'\overline{y} + \overline{x}'y} \tag{8}$$

$$S_{\text{Kulzinsky}}(x,y) = \frac{x'y}{x'\overline{y} + \overline{x}'y} \tag{9}$$

The Jaccard, Dice, and Kulzinsky similarity measures differ in their ranges: the Jaccard measure ranges from 0 to 1, the Dice measure from 0 to ½, and the Kulzinsky measure from 0 to ∞. Eqns (7~9) can be generalized to eqn (10) which for $\sigma = 0$ becomes the Kulzinsky coefficient, for $\sigma = 1$ the Jaccard coefficient, and for $\sigma = 2$ the Dice coefficient.

$$S_{\text{Generalized Jaccard}}(x,y) = \frac{x'y}{\sigma x'y + x'\overline{y} + \overline{x}'y} \tag{10}$$

Another popular distance measure between binary feature vectors is the Tanimoto metric defined in eqn (11) [6] where n_x and n_y are the numbers of 1 bits in x and y, respectively, and $n_{x,y}$ is $x'y$.

$$\begin{aligned} D_{Tanimoto}(x,y) &= \frac{n_x + n_y - 2n_{x,y}}{n_x + n_y - n_{x,y}} \\ &= \frac{x'\overline{y} + \overline{x}'y}{x'\overline{y} + \overline{x}'y + x'y} \end{aligned} \tag{11}$$

The Tanimoto coefficient [6], defined in eqn (12), is another variation of the normalized inner product which is frequently encountered in the fields of information retrieval and biological taxonomy.

$$S_{Tanimoto}(x,y) = \frac{x'y}{x'x + y'y - x'y} \tag{12}$$

Most similarity measures are variations either of Hamming or of the inner-product. Generally, the former ones treat the presence, $x'y$, and the absence, $\overline{x}'\overline{y}$, of features equally while the later take only the presence, $x'y$, into account and exclude $\overline{x}'\overline{y}$. The decision to include or exclude the $\overline{x}'\overline{y}$ term is a difficult and contentious one [7, 17]. Prior to 1950 when the *Hamming* distance was introduced, the use of inner-product based similarity coefficients flourished. Sokal and Michener made a good argument to include the negative matches [7, 17, 18] but they used equal weights for both positive and negative matches.

Hence, we propose a new measure with variable credit for the $\overline{x}'\overline{y}$ term, eqn (13), where σ is the contribution factor, and $0 \le \sigma \ll \infty$. We call it the *azzoo* similarity measure because we can alter the credit for the zero-zero matches relative to that for the one-one matches (azzoo = alter zero zero one one).

$$\begin{aligned} S_{azzoo}(x,y) &= x'y + \sigma\overline{x}'\overline{y} \\ &= \sum_{i=1}^{d} x_i y_i + \sigma \sum_{i=1}^{d} (1-x_i)(1-y_i) \end{aligned} \tag{13}$$

$$S_{azzoo}(x,y) = \sum_{i=1}^{d} s_i$$

$$\text{where } s_i = \begin{cases} 1 \text{ if } x_i = y_i = 1 \\ \sigma \text{ if } x_i = y_i = 0 \\ 0 \text{ otherwise} \end{cases}$$

Note that for $\sigma = 0$, S_{azzoo} becomes the inner product, and for $\sigma = 1$, the Hamming similarity measure. Although S_{azzoo} requires finding the optimal σ factor, the experimental results in later sections show that it outperforms both the Hamming and inner-product similarity measures.

Originally, the half-credit similarity, $S_{00\text{-}11}$ was used in an offline handwriting recognition system [12] and it is the same as S_{azzoo} with $\sigma = 0.5$. It gives full credit to features present in both patterns, $x'y$, half credit to those not present in either pattern, $\overline{x}'\overline{y}$, and no credit to those present in only one of the patterns, $x'\overline{y}$ and $\overline{x}'y$, as defined in eqn (14) [12, 13, 14] and here we generalized the half-credit similarity to S_{azzoo}.

$$S_{00\text{-}11}(x,y) = x'y + \frac{\overline{x}'\overline{y}}{2} \tag{14}$$

The range of S_{azzoo} is $[0,d]$ if $0 \le \sigma \le 1$ and $[0,\sigma d]$ if $\sigma > 1$. Assuming $0 \le \sigma \le 1$, S_{azzoo} can be converted to a distance measure for metric property testing, eqn (15).

$$D_{azzoo}(x,y) = d - S_{azzoo}(x,y) = d - \left(x'y + \sigma\overline{x}'\overline{y}\right) \tag{15}$$

Nonnegativity and symmetry are trivial and preserved. Reflexivity is violated, however, because $D_{azzoo}(x,y) = 0$ iff $x = y$ and $|x| = |y| = d$ and $D_{azzoo}(x,y) \ne 0$ otherwise. Similarly,

$S_{azzoo}(x, y) = d$ iff $x = y$ and $|x|=|y|=d$ and $\sigma d \le S_{00-11}(x, y) < d$ if $x = y$ and $|x|<d$.

Theorem 1. The triangle inequality property is valid for $D_{azzoo}(x, y)$, i.e., $D_{azzoo}(x, y) + D_{azzoo}(y, z) \ge D_{azzoo}(x, z)$.

Proof

$\overline{x}'\overline{y} + \overline{y}'\overline{z} \ge \overline{x}'\overline{z}$ by Fact 1 line 1

$d - (x'y + \overline{x}'\overline{y}) + d - (y'z + \overline{y}'\overline{z}) \ge d - (x'z + \overline{x}'\overline{z})$ by Fact 2 line 2

Now, evaluate $D_{azzoo}(x, y) + D_{azzoo}(y, z) \ge D_{azzoo}(x, z)$.

$$d - (x'y + \sigma\overline{x}'\overline{y}) + d - (y'z + \sigma\overline{y}'\overline{z}) \ge d - (x'z + \sigma\overline{x}'\overline{z})$$

$$d - (x'y + \overline{x}'\overline{y}) + d - (y'z + \overline{y}'\overline{z}) + (1-\sigma)\overline{x}'\overline{y} + (1-\sigma)\overline{y}'\overline{z}$$
$$\ge d - (x'z + \overline{x}'\overline{z}) + (1-\sigma)\overline{x}'\overline{z}$$

Hence, the theorem is true by line 1 and line 2 □

Similarly, since $S_{azzoo}(x, y) = x'y + \sigma\overline{x}'\overline{y} = S_{\text{inner-product}}(x, y) + \sigma S_{\text{inner-product}}(\overline{x}, \overline{y})$, the properties of the azzoo similarity measure are similar to those of the inner-product measure.

Other popular similarity measures utilize coefficients of correlation and have been used frequently in both psychology and ecology studies [7]. The correlation similarity measure is given in eqn (16) and Yule and Kendall [24] suggested a similar coefficient given in eqn (17).

$$S_{\text{correlation}} = \frac{x'y \times \overline{x}'\overline{y} - x'\overline{y} \times \overline{x}'y}{\sqrt{(x'\overline{y} + x'y)(\overline{x}'y + \overline{x}'\overline{y})(x'y + \overline{x}'y)(\overline{x}'\overline{y} + x'\overline{y})}} \tag{16}$$

$$S_{\text{Yule}} = \frac{x'y \times \overline{x}'\overline{y} - x'\overline{y} \times \overline{x}'y}{x'y \times \overline{x}'\overline{y} + x'\overline{y} \times \overline{x}'y} \tag{17}$$

While Hamming based similarity measures are additive forms of the positive and negative matches, the correlation based measures are multiplicative forms. Nonetheless, contribution factors of positive and negative matches are considered equally important in correlation based similarity measures as well as Hamming based ones.

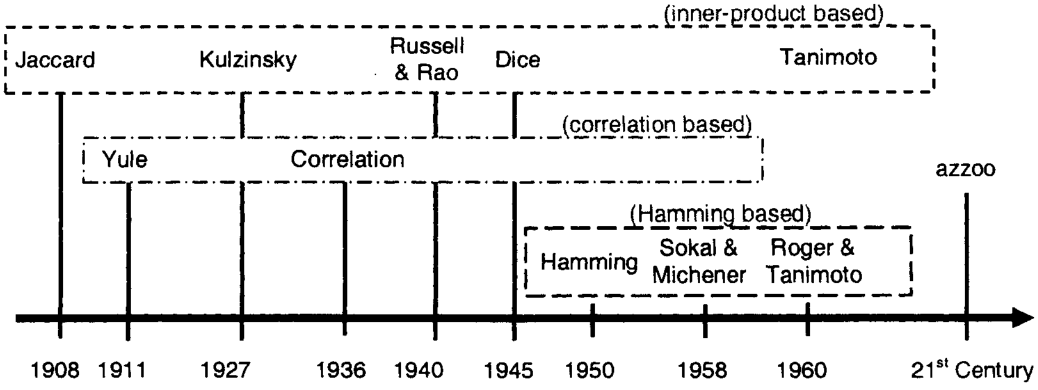


Figure 1. A chronological table for binary vector similarity measures.

Historically, all the measures enumerated above have had great value in their respective fields. In the following sections 3 and 4, we evaluate these measures in the applications of iris biometric authentication and offline handwriting character recognition. Figure 1 shows a chronological table for binary feature vector similarity measures in which these conventional measures are categorized into three major groups: inner-product, Hamming, and correlation based groups.

2.2. Binary Similarity Measures with Weights

To further improve their discrimination capability, weights can be applied to distance or similarity measures [14] and optimized using techniques such as genetic algorithms [15, 16]. When features have numeric values, a scaling problem arises. In order to mitigate this problem, one can combine the nonlinear accuracy weighting with the Minkowski distance concept as shown in eqn (18) where $P(C/i)$ is the probability of being correct when only feature i is used [25, 26].

$$D_{\text{weighted Minkowski}} = \sum_{i=1}^{d} \left[P(C/i)^{a} \left| x_i - y_i \right|^{r} \right] \tag{18}$$

When features are binary, one can still generalize eqn (18) to eqn (19) by setting $r = 1$ and $P(C/i)^{a} = w_i$.

$$\begin{aligned} D_{\text{weighted-Hamming}}(x, y) &= \sum_{i=1}^{d} w_i \left| x_i - y_i \right| \\ S_{\text{weighted-Hamming}}(x, y) &= \sum_{i=1}^{d} w_i (x_i y_i + \overline{x}_i \overline{y}_i) \end{aligned} \tag{19}$$

The weighted Hamming distance has been applied to numerous applications such as image template matching [27, 28] and object recognition [28]. The weighted Hamming distance provides an improvement over the simple Hamming distance for discriminating between similar images [27, 28]. This distance measure gives greater importance to error pixels which appear in close proximity to other error pixels. Error pixels which appear close together tend to correspond to structurally meaningful features. In [29], a slightly different weighted Hamming distance was introduced to optimize the distance measure for object detection by adding a null weight, w_0. Similarly, the inner product similarity measure can be optimized by applying weights as shown in eqn (20).

$$S_{\text{weighted-inner-product}}(x, y) = \sum_{i=1}^{d} w_i x_i y_i \tag{20}$$

Here, we claim that the performance can be further improved by optimizing the similarity measure rather than distance measure. Since the Hamming distance is the number of mismatches, the weights are applied to the mismatched bits, whereas in a similarity measure the weights are applied to the matching bits. As discussed in the earlier section, there are two kinds of matches: positive and negative matches. Although the Hamming similarity can be improved by applying the equal weights are applied to both positive and negative matches, we claim that if different weights are applied, the performance is further improved, and the proposed weighted 00-11 similarity measure is given in eqn

(21).

$$S_{weighed\text{-}00\text{-}11}(x,y)=\sum_{i=1}^{d} w_{\oplus i}x_i y_i+\sum_{i=1}^{d} w_{\ominus i}\bar{x}_i\bar{y}_i \qquad (21)$$

Note that if $w_\oplus$ and $w_\ominus$ are identical, $S_{weighed\text{-}00\text{-}11}=S_{\text{weighted-hamming}}$ and if $w_\ominus=0$, $S_{weighed\text{-}00\text{-}11}=S_{\text{weighted-inner-product}}$.

There are twice as many coefficients to optimize in this new similarity measure than in the weighted Hamming or inner product similarity measures. This is a multi-dimensional, space optimization problem, and one can use a genetic algorithm to determine the weights from training data. A genetic algorithm can be a general optimization method that searches a large space of candidate objects to find one that performs near optimal according to the fitness function [15, 16]. Genetic algorithms offer a number of advantages: they search from a set of solutions rather than from a single one, they are not derivative-based, and they explore and exploit the parameter space. For the weight adaptive model, we create a numerical optimization model that depends on a set of weights.

3. Similarity Measure Evaluation on Iris Biometric Verification

In order to evaluate the binary vector similarity measures, we consider an iris biometric database. Daugman proposed the degrees of freedom of iris mismatch score distribution as a measure of the individuality or uniqueness of an iris pattern [30]. The biometric verification problem is a simple *dichotomy* classification problem that places the input into one of only two categories – that is, given two randomly selected biometric samples, the problem is to determine whether the two samples belong to the same person or two to different people. Figure 2 depicts the biometric verification model.

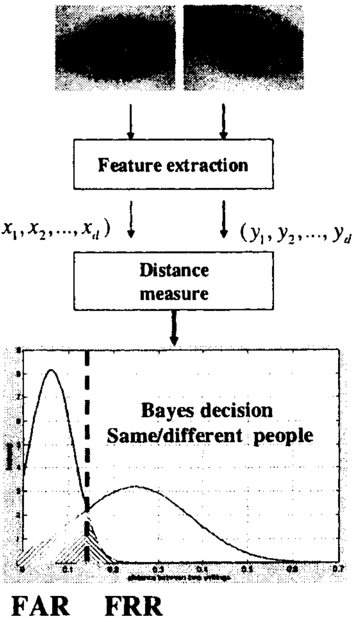


Figure. 2. The iris verification process

First, features are extracted from iris biometric image data x and y: $\{x_1, x_2, \cdots, x_d\}$ and $\{y_1, y_2, \cdots, y_d\}$. Let $c(x)$ denote the class (the person) to which x belongs. The iriscode, proposed by Daugman [30], is a 8x256 binary feature extracted from an iris image by applying a 2D Gabor wavelet filter, and Daugman used the Hamming distance. When a distance measure is applied, two distributions are generated. One distribution, called the intra distance (or within person) distribution, occurs when $c(x) = c(y)$. The other distribution, called the inter distance (or between two different people) distribution, occurs when $c(x) \neq c(y)$. By assuming the distributions are normal one can easily find the decision threshold to minimize the FAR (false accept rate) and the FRR (false reject rate).

3.1. Performance evaluation method

In Figure 2, FAR is the probability of error that one classifies two biometric data as coming from the same person even though they belong to two different people, and FRR is the probability of error that one classifies two biometric samples as coming from different people even though they belong to a same person. Note that when a distance measure is used, the intra distance distribution tends to be close to 0 whereas the inter distance distribution tends to be far from 0. Thus, FAR is usually the left side area of the decision boundary. When a similarity measure is used, on the other hand, FAR is the right-side area of the decision boundary because the larger value means that the two biometric samples are similar, as defined in eqns (22) and (23) where T is the threshold value for the Bayesian decision.

$$\mathrm{FAR} = \Pr\ (S(x,y) \geq T \mid c(x) \neq c(y)) \tag{22}$$

$$\mathrm{FRR} = \Pr\ (S(x,y) < T \mid c(x) = c(y)) \tag{23}$$

The overall performance is the number of correctly classified instances divided by the total testing database size.

3.2. Experimental results

In this section, we compare the experimental results obtained by using several similarity measures. From the iris biometric image database [31], we selected 10 left bare eye samples of 52 subjects. In order to test the described models, two sets of samples are required: intra-class distance and inter-class distance sets. The intra-class distance sample is acquired by randomly selecting two iris data from the same subject while the inter-class distance sample is obtained by randomly selecting two iris data from two different subjects. We prepared three sets of inter and intra distance data for training and three independent ones for testing, each of size 1000 (500 intra-class and 500 inter-class pairs).

The iris biometric verification models were trained on 500 distance or similarity values obtained from the intra- and inter-class sets. These scalar values form distributions and the mean and variance can be computed for each distribution. Assuming normal distributions, one can easily find the Bayes decision threshold. For testing, each scalar distance value is classified into the intra or inter person class by comparing to the threshold value. First, we used the Hamming distance as Daugman originally proposed [30] and Figure 3 shows the results on the database used in this experiment [31].

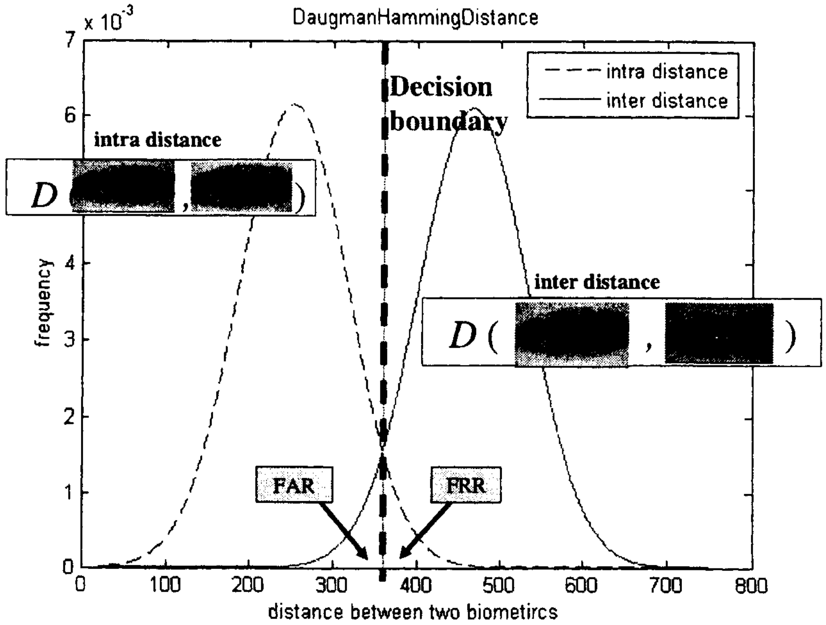
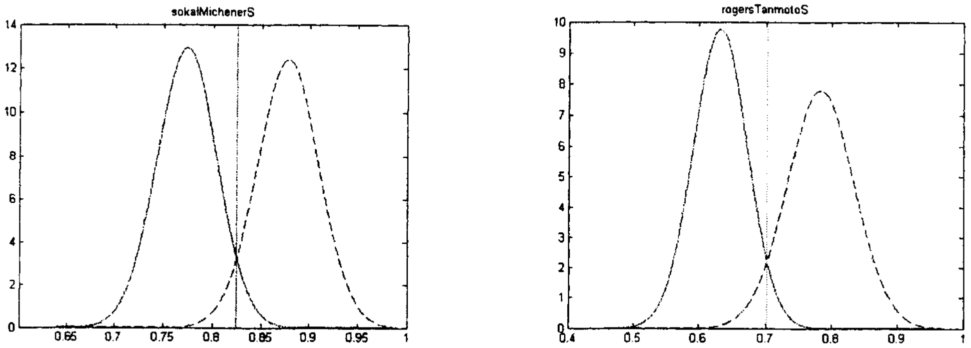
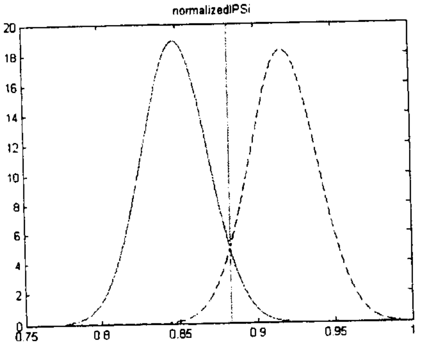
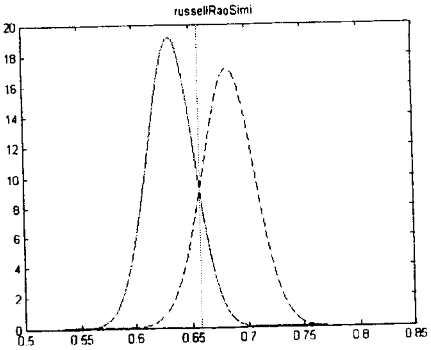


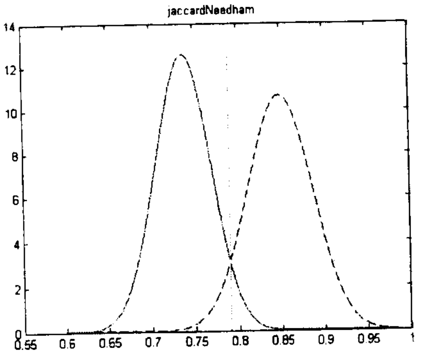
Figure 3. Intra and inter distance distribution using *Hamming* distance

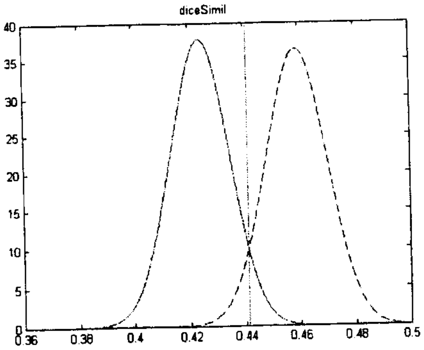
We then obtained results on the other similarity measures. Figure 4 depicts the intra and inter similarity distributions using various similarity measures and Table 1 shows the comparative results of the overall performances. Finally, Figure 5 shows the performance as a function of the contribution factor, σ, and highlights the relative performance of the inner product, Hamming, and azzoo measures. The S_{azzoo} with $\sigma = 1.175$ yields the best performance.

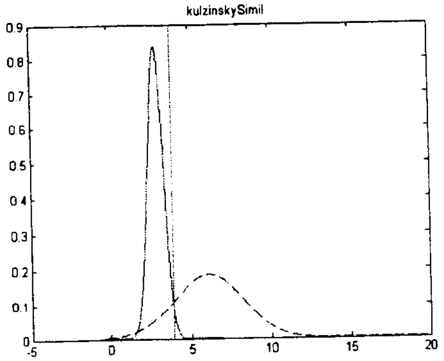


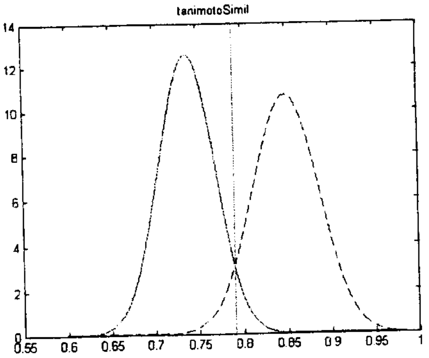

normalizedIPSi
20
18
16
14
12
10
8
6
4
2
0
0.75
0.8
0.85
0.9
0.95
1

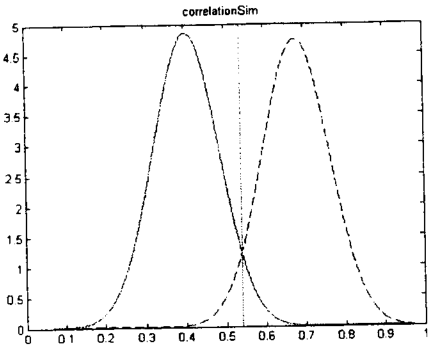

russellRaoSimi
20
18
16
14
12
10
8
6
4
2
0
0.5
0.55
0.6
0.65
0.7
0.75
0.8
0.85

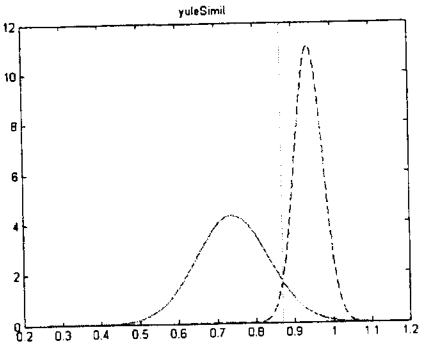

jaccardNeedham
14
12
10
8
6
4
2
0
0.55
0.6
0.65
0.7
0.75
0.8
0.85
0.9
0.95
1


diceSimil
40
35
30
25
20
15
10
5
0
0.36
0.38
0.4
0.42
0.44
0.46
0.48
0.5


kulzinskySimil
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-5
0
5
10
15
20


tanimotoSimil
14
12
10
8
6
4
2
0
0.55
0.6
0.65
0.7
0.75
0.8
0.85
0.9
0.95
1


correlationSim
5
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1


yuleSimil
12
10
8
6
4
2
0
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
1.1
1.2

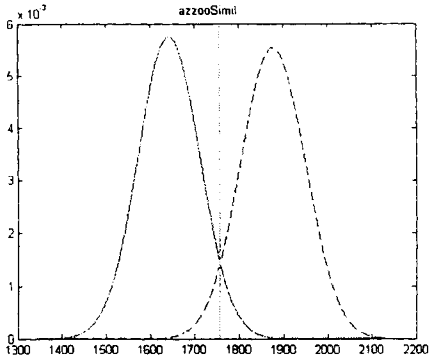


Figure 4. Intra and inter distance distributions for the various similarity measures.

Table 1. Performance evaluation of the similarity measures on the iris database.

	Data 1			Data 2			Data 3			Data 4			Total
Method	FAR	FRR	Rate	FAR	FRR	Rate	FAR	FRR	Rate	FAR	FRR	Rate	Rate
azzoo	5.0	4.4	95.3	6.4	3.2	95.2	7.6	3.2	94.6	4.4	3.8	95.9	95.3
normalized I.P.	4.8	4.8	95.2	6.0	4.4	94.8	7.4	3.6	94.5	5.0	4.4	95.3	95.0
SokalMichener	5.0	4.8	95.1	6.4	3.6	95.0	7.8	3.2	94.5	4.6	3.8	95.8	95.1
RogersTanmoto	5.0	4.8	95.1	6.4	3.6	95.0	7.8	3.2	94.5	4.6	3.8	95.8	95.1
RussellRao	12.8	12.2	87.5	11.8	10.4	88.9	11.4	8.6	90.0	11.2	10.4	89.2	88.9
JaccardNeedham	4.8	4.8	95.2	6.2	4.2	94.8	7.4	3.6	94.5	5.0	4.0	95.5	95.0
Dice	4.8	4.8	95.2	6.0	4.2	94.9	7.4	3.6	94.5	5.0	4.4	95.3	95.0
Kulzinsky	6.8	3.8	94.7	7.4	3.0	94.8	9.0	2.6	94.2	6.4	3.4	95.1	94.7
Tanimoto	4.8	4.8	95.2	6.2	4.2	94.8	7.4	3.6	94.5	5.0	4.0	95.5	95.0
correlation	5.4	4.6	95.0	6.4	3.8	94.9	7.8	3.2	94.5	4.2	3.6	96.1	95.1
Yule	4.8	4.8	95.2	5.6	4.8	94.8	7.8	3.0	94.6	4.0	3.8	96.1	95.2

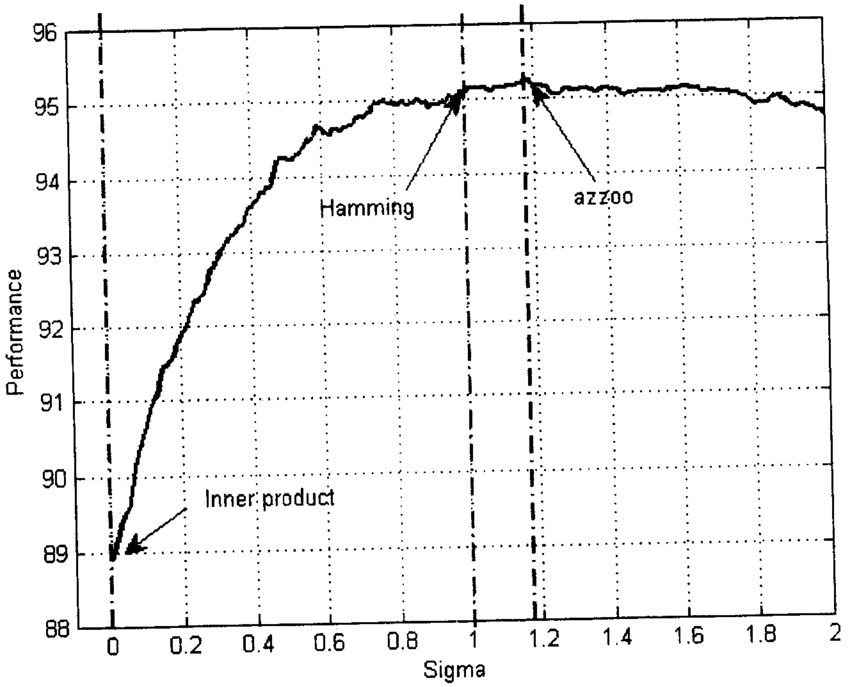


Figure 5. Performance vs. the contribution factor σ.

4. Similarity Measure Evaluation on Character Recognition

To further evaluate binary vector similarity measures, we consider an offline handwritten character image database.

4.1. Binary feature extraction

Among many features, the *GSC* (*Gradient, Structural, and Concavity*) feature set has been shown to have high accuracy in offline character recognition problems [12] based on the philosophy that feature sets can be designed to extract certain types of information from the image. These types are gradient, structural, and concavity information. Gradient features use the stroke shapes on a small scale, structural features use stroke trajectories on an intermediate scale, and concavity features use stroke relationships at long distances.

The input character image is a binarized and slant-normalized image. A bounding box is placed around the image and divided into a 4 x 4 grids, which is known as a quasi-multiresolution approach shown in Figure 6. For each grid region, all directional rules and various concavity features are checked, resulting in 192 Gradient, 192 Structural, and 128 Concavity features, for a total of 512 features as listed in Table 2. A sample vector for a character "*A*" is given in Figure 7. See [12] for a detailed description of the rules.

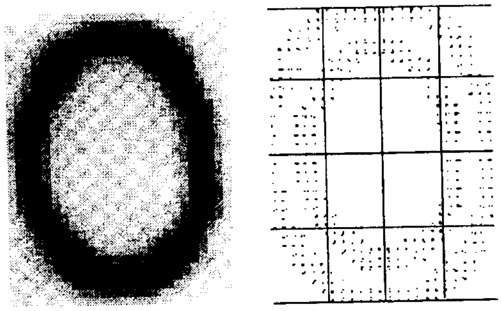
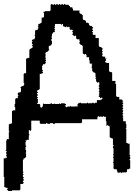


Figure 6. Character recognition 4×4 grid.

Table 2. GSC Features where $x = 0 \cdots 3$ and $y = 0 \cdots 3$.

	Gradient		Structural		Concavity Features	
Grid Pos.	*ID*	*Directional*	*ID*	*Rule**	*ID*	*Concavity*
(0,0) ⋮ (3,3)	G01-00 ⋮ G01-33	1° ~ 30° ⋮ 1° ~ 30°	S01-00 ⋮ S01-33	r_1 ⋮ r_1	C-CP-00 ⋮ C-CP-33	Coarse Pixel Density ⋮ Coarse Pixel Density
(x,y)	G02-xy	31° ~ 60°	S02-xy	r_2	C-HR-xy	horizontal run length
(x,y)	G03-xy	61° ~ 90°	S03-xy	r_3	C-VR-xy	vertical run length
(x,y)	G04-xy	91° ~ 120°	S04-xy	r_4	C-UC-xy	Upward concavity
(x,y)	G05-xy	121° ~ 150°	S05-xy	r_5	C-DC-xy	Downward concavity
(x,y)	G06-xy	151° ~ 180°	S06-xy	r_6	C-LC-xy	Left concavity
(x,y)	G07-xy	181° ~ 210°	S07-xy	r_7	C-RC-xy	Right concavity
(x,y)	G08-xy	211° ~ 240°	S08-xy	r_8	C-HC-xy	Hole concavity
(x,y)	G09-xy	241° ~ 270°	S09-xy	r_9		
(x,y)	G10-xy	271° ~ 300°	S10-xy	r_{10}		
(x,y)	G11-xy	301° ~ 330°	S11-xy	r_{11}		
(x,y)	G12-xy	331° ~ 360°	S12-xy	r_{12}		



```
Gradient      00000000001100000000011000011100000000111000000001100000110
(192bits)     00100000000011000000000000000111001100011111000001111100000000
              10010100000010001110011111001111110000001000000100000000000000
              000000000100000001001000
Structural    000000000000000000000011000011100010000100001000000010000000
(192bits)     0000001001010000000000001100001010011000011000000000000000100
              1000110011000000000000000110010100000000000000001100000000000
              000000000000000000010000
Concavity     1111011010011111011001100000011011110110100110010000001100
(128bits)     00011100000000000000000000000000000000000000000000011111111000
              00000000000000
```

Figure 7. A sample character and its GSC feature vector.

4.2. Experimental Results

The problem of offline handwritten character recognition is to classify an unknown handwritten character image as one of the 26 letters of the alphabet. There are 800 samples per letter of 512 binary feature vectors in the database: 400 samples per letter are used for the reference set and the remaining samples are divided into four sets of 100 samples per letter for testing and tuning purposes. The k-nearest neighbor (k-nn) approach is used, and table 3 shows the number of errors by each of the similarity measures tested. The numbers in parentheses are the errors after optimizing the weights in the weighted variations of the measures.

Among the numerous similarity measures without weights, azzoo performed the best. Figure 8 shows the performance as a function of the contribution factor, σ, and stresses the relative performance of the inner product, Hamming, and azzoo measures, clearly showing the superiority of the azzoo measure when $\sigma = 0.44$.

Table 3. Similarity measures and their errors on handwritten character recognition.

Category	Measure	Errors
	Azzoo	330 (312)
Hamming	Hamming	398 (331)
	Sokal-Michener	398
	Roger-Tanimoto	398
Inner Product	Inner-product	700 (616)
	Russell-Rao	700
	Normalized I.P.	343
	Jaccard-Needham	341
	Dice	341
	Kulzinsky	341
	Tanimoto	341
Correlation	Correlation	347
	Yule	474

Applying different weights further improves the performance. Note that the number of weights is $d = 512$ in the weighted Hamming distance and $2 \times d = 1024$ in the weighted azzoo. We use a genetic algorithm to determine the weights and the weighted azzoo significantly outperforms the weighted Hamming distance.

We previously conducted similar optimizing experiments where we used fewer weights [14]. Since the number of weights is enormous when each feature is given its own weight, in those experiments we simplified the features into three feature groups (gradient, structural, and concavity feature sets), and then used the additive model to combine these measures as follows:

$$S[x, y] = S'[x_g, y_g] + S''[x_s, y_s] + S'''[x_c, y_c] \tag{24}$$

where x_g, x_s, and x_c are the gradient, structural, and concavity subsets of x, so that $x = x_g \cup x_s \cup x_c$. First, we consider each individual set of GSC features. For $S'[x_g, y_g]$, $S''[x_s, y_s]$ and $S'''[x_c, y_c]$, we use the azzoo similarity measure.

Next we associate weights with each feature group as shown in eqn (25).

$$\begin{aligned} S_{w\text{-azzoo}}(x, y) = {} & w_{g11} x_g y_g + w_{g00} \overline{x}_g \overline{y}_g \\ & + w_{s11} x_s y_s + w_{s00} \overline{x}_s \overline{y}_s \\ & + w_{c11} x_c y_c + w_{c00} \overline{x}_c \overline{y}_c \end{aligned} \tag{25}$$

Here, w_{g11}, w_{g00}, w_{s11}, w_{s00}, w_{c11}, and w_{c00} are the weights for the Gradient, Structural, and Concavity feature groups. Optimizing these six coefficients, we found the number of errors to be 319 which, although an improvement over the azzoo measure without weights, is not as good as the 312 errors when the full set of weights is used.

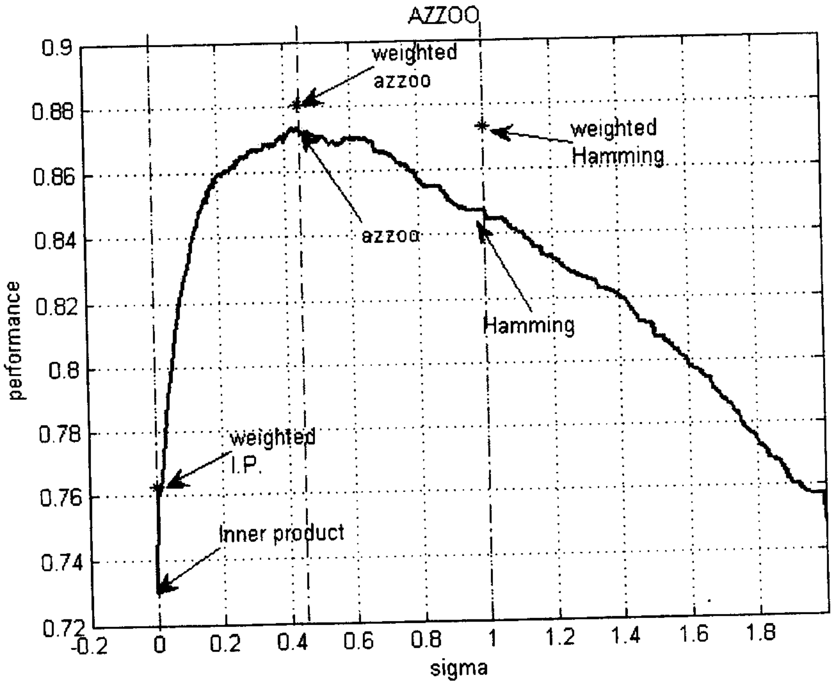


Figure 8. Performance vs. the contribution factor σ.

5. Finale

To conclude, we emphasize that selecting and designing a similarity measure is extremely important. First, we reviewed and categorized ten different binary feature vector similarity measures. Conventional similarity coefficients were categorized into three groups depending on their type: inner-product, Hamming, and correlation based similarity measures as depicted in Figure 9. The first major division is between the inner-product based similarity measures that consider positive matches only and those that credit both positive and negative matches. Next, those that consider both positive and negative matches are further categorized into additive forms and multiplicative forms (correlation based measures).

Patterns can be analyzed by either distance or similarity. Pattern classification or clustering using Hamming distance will have the identical results as those using Sokal and Michener's normalized Hamming similarity measure. From the point of view of distance, positive and negative matches are treated equally. By first converting the Hamming distance into a similarity measure, we derived another similarity measure that distinguishes the positive and negative matches, and we called it the '*azzoo*' similarity measure. In our version of a taxonomy, the azzoo similarity measure is under the additive form of similarity measures that take both positive and negative matches into accounts.

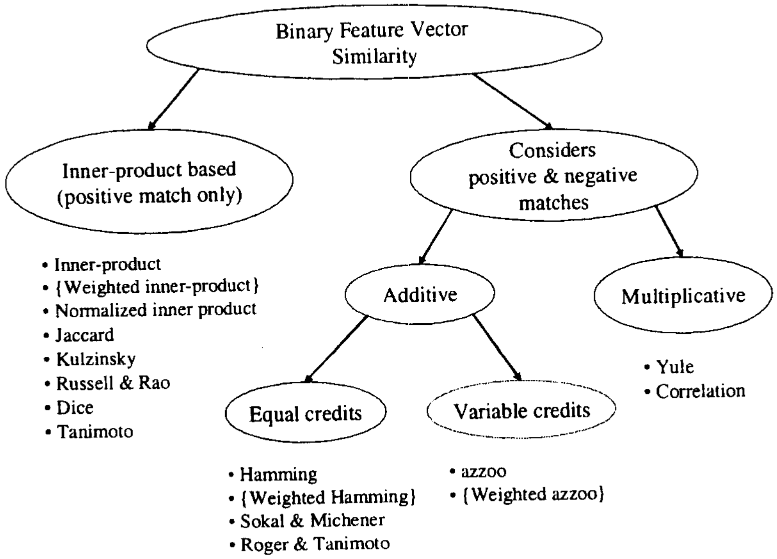


Figure 9. Taxonomy of Binary Feature Vector Similarity Measures

We showed that the azzoo measure outperforms all conventional measures in the applications of iris biometric verification and offline handwritten character recognition. While the azzoo measure is superior to the other measures, it is interesting to note that the value of the contributing factor σ can vary considerably depending on the application data – in this case the optimal value was 1.175 on the iris data and 0.44 on the handwriting data.

Moreover, we explored enhancing the similarity measures by applying weights that can be optimized to specific application data. While the weighted Hamming similarity measure gives identical weights to both positive and negative matches, we demonstrated that the weighted azzoo similarity measure that gives different weights to positive and negative matches can further improve the discrimination performance.

References

[1] B. V. Dasarathy, Visiting nearest neighbors - a survery of nearest neighbor pattern classification techniques, Proceedings of the International Conference on Cy- bernetics and Society, IEEE, September 1977, pp. 630–636.

[2] B. V. Dasarathy, Nearest neighbor pattern classification techniques, IEEE Computer Society Press, 1991

[3] J.R.Smith and S.-F. Chang, Automated binary texture feature sets for image retrieval, International Conf. Accoust., Speech, Signal processing, Atlantic, GA, May 1996

[4] P. Willett, J.M. Barnard, and G.M. Downs. Chemical similarity searching. *J Chem Inf Comput Sci* 1998. 38: 983-996.

[5] L. C. Cole, The measurement of partial interspecific association, *Ecology*, 1957 38:226-233.

[6] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. John Wiley & Sons, Inc., 2nd ed., 2000.

[7] P. H. A. Sneath and R. R. Sokal, Numerical Taxonomy, London: Freeman, 1973

[8] D. H. T. Clifford and W. Stephenson, An Introduction to Numerical Classification, New York: Academic, 1975

[9] R. V. Hamming, Error detecting and error correcting codes, Bell Sys. Tech. Journal, 29:147-160, 1950

[10] J. D. Tubbs, A note on binary template matching, Pattern Recognition, 1989, 22(4):359-365.

[11] B. Zhang and S. N. Srihari, Binary vector dissimilarities for handwriting identification, Proceedings of SPIE, Document Recognition and Retrieval X, 2003, p 15-166

[12] J. T. Favata and G. Srikantan, A multiple feature/resolution approach to handprinted digit and character recognition, International Journal of Imaging Systems and Technology, 1996, pp. 7:304–311.

[13] S.-H. Cha and S. N. Srihari, A fast nearest neighbour search algorithm by filtration, Pattern Recognition 35, P 515-525, 2000.

[14] S.-H. Cha and C. C. Tappert , Optimizing Binary Feature Vector Similarity Measure using Genetic Algorithm, ICDAR, Edinburgh, Scotland, 2003.

[15] M. Mitchell, *An introduction to genetic algorithms*, Cambridge, MA: MIT Press, 1996.

[16] L. Davis, *Handbook of genetic algorithms*. New York: Van Nostrand Reinhold, 1991

[17] G. Dunn and B. S. Everitt, An introduction to mathematical taxonomy, Cambridge University Press 1982

[18] R. R. Sokal and C. D. Michener, A statistical method for evaluating systematic relationships, University of Kansas Scientific Bulletin 38, 1409-1438, 1958

[19] D. J. Rogers and T. T. Tanimoto, A computer program for classifying plants, Science, 132:1115-1118, 1960.

[20] Paul R. Halmos, *An Introduction to Hilbert Spaces and the Theory of Spectral Multiplicity*, Chelsea Plublishing, 2nd ed, 1957

[21] P. F. Russell and T. R. Rao, On habitat and association of species of anopheline larvae in south-eastern Madras. J. Malar. Inst. India, 3:153-178, 1940

[22] P. Jaccard, Nouvelles recherches sur la distribution florale, Bulletin de la Societe Vaudoise de Science Naturelle, 44, 223-270, 1908

[23] L. R. Dice, Measures of the amount of ecologic association between species, *Ecology*, 26:297-302, 1945

[24] G. U. Yule and M.G.Kendall, An Introduction to the Theory of Statistics, 14th ed. Hafner, New York, pp701, 1950

[25] C. Gose, R. Johnsonbaugh, and S. Jost, Pattern Recognition and Image Analysis, Prentice Hall, Inc., p 172, 1996.

[26] B. F. Wu, Comparative performance evaluation of some techniques for ranking pattern recognition features, Ph.D. Dissertation, Bioengineering Program, University of Illinois at Chicago, 1977.

[27] W. Pratt, P. Capitant, W. Chen, E. Hamilton, and R. Willis, Combining symbol matching facsimile data compression system, Proceedings of the IEEE, 68:786-796, 1980.

[28] I. Witten, A. Moffat, and T. Bell. Managing Gigabytes: Compressing and Indexing Documents and Images. Van Nostrand Reinhold, 1994

[29] S. Mahamud and M. Hebert, The Optimal Distance Measure for Object Detection, *IEEE Computer Vision and Pattern Recognition*, Wisconsin, p 248-256, 2003.

[30] J.G. Daugman, High confidence visual recognition of persons by a test of statistical independence, IEEE Transactions on Pattern Analysis and Machine Intelligence, 15(11):1148-1161,1993.

[31] G. Kee, Y. Byun, K, Lee and Y. Lee, Improved Techniques for an Iris Recognition System with High Performance, Advances in Artificial Intelligence, LNCS Vol 2256, 2001, pp. 177-184