

Pattern Recognition 35 (2002) 1355-1370



www.elsevier.com/locate/patcog

On measuring the distance between histograms

Sung-Hyuk Cha^{a,*}, Sargur N. Srihari^b

^a School of Computer Science and Information Systems, Pace University, 861 Bedford Road,
Pleasantville, NY 10570–9913, USA
^b Center of Excellence for Document Analysis and Recognition (CEDAR), State University of New York at Buffalo,
520 Lee Entrance, Suite 202, Amherst, Buffalo, NY, USA

Received 29 June 2000; accepted 4 June 2001

Abstract

A distance measure between two histograms has applications in feature selection, image indexing and retrieval, pattern classification and clustering, etc. We propose a distance between sets of measurement values as a measure of dissimilarity of two histograms. The proposed measure has the advantage over the traditional distance measures regarding the overlap between two distributions; it takes the similarity of the non-overlapping parts into account as well as that of overlapping parts. We consider three versions of the univariate histogram, corresponding to whether the type of measurement is nominal, ordinal, and modulo and their computational time complexities are $\Theta(b)$, $\Theta(b)$ and $O(b^2)$ for each type of measurements, respectively, where b is the number of levels in histograms. © 2002 Pattern Recognition Society. Published by Elsevier Science Ltd. All rights reserved.

Keywords: Distance measure; Histogram; Nominal; Modulo; Ordinal

1. Introduction

A histogram representation of a sample set of a population with respect to a measurement represents the frequency of quantized values of that measurement among the samples. Finding the distance, or similarity, between two histograms is an important issue in pattern classification and clustering [1–3]. A number of measures for computing the distance have been proposed and used.

There are two methodologies in histogram distance measures: vector and probabilistic. In the vector approach, a histogram is treated as a fixed-dimensional vector. Hence standard vector norms such as *city block*, *Euclidean* or *intersection* can be used as distance measures. Vector measures between univariate histograms have been used in image indexing and retrieval [4–7].

E-mail address: scha@pace.edu (S.-H. Cha).

The probabilistic approach is based on the fact that a histogram of a measurement provides the basis for an empirical estimate of the probability density function (pdf) [8]. Computing the distance between two pdfs can be regarded as the same as computing the Bayes (or minimum misclassification) probability. This is equivalent to measuring the overlap between two pdfs as the distance. There is much literature regarding the distance between pdfs, an early one being the Bhattacharyya distance or B-distance measure between statistical populations [9]. The *B-distance*, which is a value between 0 and 1 provides bounds on the Bayes misclassification probability. An approach closely related to the B-distance was proposed by Matusita [3,10]. Kullback and Leibler [11] generalized Shannon's concept of probabilistic uncertainty or "entropy" [12] and introduced the "K-L distance" [1,2] measure that is the minimum cross entropy (see [13] for an extensive bibliography on estimation of misclassification).

The viewpoint of regarding the overlap (or intersection) between two histograms as the distance has the disadvantage that it does not take into account the similarity

^{*} Corresponding author. Tel.: +1-914-773-3440; fax: +1-914-773-3533.

of the non-overlapping parts of the two distributions. For this reason, we present a new definition of the distance for each type of histograms. The new measure uses the notion of the *minimum difference of pair assignments* or *MDPA* in short. We also describe the efficient algorithms to compute the distance for each type.

Rubner and Guibas showed a robust distance measure that overcomes this non-overlapping parts problem called an earth mover's distance [14]. It is the minimal amount of work that must be performed to transform one distribution into the other by moving "distribution mass". As the EMD is quite similar to the newly introduced distance measure, we give the comparisons between them. In their work, the EMD was computed by the linear optimization algorithm based on a solution to the transportation problem. The transportation problem is a special linear programming problem to distribute any commodity from any group of supply centers, called sources, to any group of receiving centers, called destinations, in such a way as to minimize the total distribution cost [15]. This solution has higher complexity than our O(b), O(b), and $O(b^2)$ solutions for nominal, ordinal and modulo type histograms distance measures where b is the number of levels.

The subsequent sections are constructed as follows. In Section 2, histograms are defined with respect to three measurement types. In Section 3, we define a new definition of distance measure using the notion of the minimum difference of pair assignments. In Section 4, we examine conventional definitions of distance between two histograms and give examples that show the inadequacy of them when they are used for certain types of measurements. Section 5 is dedicated to the description and analysis of each algorithm to compute the distance between each type of histograms. In Section 6, we address the character similarity problem with the proposed measure after extracting the *gradient direction* histograms. Finally, we conclude with the emphasis of the advantage of the MDPA as a distance measure between histograms and other conventional definitions are inadequate for ordinal or modulo type histograms.

2. Histogram definition

We will use the following notations and symbols. Let x be a measurement, or feature, which can have one of b values contained in the set, $X = \{x_0, x_1, \dots, x_{b-1}\}$. Consider a set of n elements whose measurements of the value of x are: $A = \{a_1, a_2, \dots, a_n\}$ where $a_j \in X$. The histogram of the set A along measurement x is H(x, A) which is an ordered list (or b-dimensional vector) consisting of the number of occurrences of the discrete values of x among the a_i . As we are interested only in comparing the histograms of the same measurement x, x will be used in place of x without loss of generality. If x if x if x is the following place of x without loss of generality. If x if x is the following place of x is the following place of x and x is the following place of x is the following

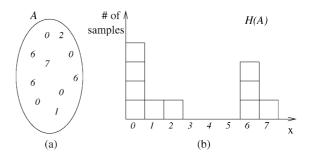


Fig. 1. (a) Measurements corresponding to a set of elements A and (b) its histogram H(A).

 $0 \le i \le b-1$, denotes the number of elements of A that have value x_i , then $H(A) = [H_0(A), H_1(A), \dots, H_{b-1}(A)]$, where

$$H_i(A) = \sum_{i=1}^n c_{ij} \quad \text{where } c_{ij} = \begin{cases} 1 & \text{if } a_j = x_i, \\ 0 & \text{otherwise.} \end{cases}$$
 (1)

If $P_i(A)$ denotes the probability of samples in the *i*th value or bin, then $P_i(A) = H_i(A)/n$.

As illustrated in Fig. 1, a histogram, H(A) is shown for a set of elements for n = 10 and b = 8, with $A = \{1,0,7,6,0,0,2,6,6,0\}$, H(A) = [4,1,1,0,0,0,3,1] and P(A) = [0.4,0.1,0.1,0,0,0,0.3,0.1] If the ordering of the elements in a sample set A is unimportant, then H(A) is a lossless representation of A in that A can be reconstructed from H(A).

2.1. Types of measurements

We consider three types of measurements: nominal, ordinal and modulo. According to the measurements, we consider three types of histograms. In a nominal measurement, each value of the measurement is named, e.g., the make of an automobile can take values such as GM, Ford, Toyota, Hyundai, etc. An example of a nominal type histogram is one that consists of the numbers of each automobile make in a parking lot. In an ordinal measurement, the values are ordered, e.g., the weight of an automobile can be quantized into 10 integer values between 0 and 9 tons. Most measurements are of the ordinal type, e.g., year, height, width or weight of automobiles or grey scale level in grey images. Finally, in modulo measurement, measurement values form a ring due to the arithmetic modulo operation, e.g., the compass direction of an automobile that can take eight values, N, NE, E, SE, S, SW, W, NW, form a ring under the operation of changing direction by 45°. The modulo type histograms are obtained along the angular values such as directions or "hue" in color images.

2.2. Permutability of levels

The measurement values are called *levels* when they are used in a histogram to index the sample values, e.g., grey level. In a nominal type histogram, the levels can be of any order and permuted freely as there is no particular ordering among themselves. In contrast, there exists an ordering among the levels in both ordinal and modulo type histograms.

In finding the distance between two histograms of nominal type measurements, the ordering of levels should not affect the outcome as long as the two histograms maintain the same ordering in their levels. For instance, the distance between two histograms of automobile make should not change whether the ordering is *GM*, *Ford*, *Toyota*, *Hyundai* or it is *Ford*, *Hyundai*, *GM*, *Toyota*. This *shuffling invariance* property is satisfied by the existing methods of distance measure, such as *city block*, *Euclidean*, *intersection*, *Bhattacharyya*, *Matusita* and *K–L distances*, because they are sums of individual distances of each level and due to the commutative law, the distances do not change when levels are permuted among themselves.

On the contrary, the *shuffling invariance* property is not desirable in the distance between the histograms of ordinal or modulo type measurements. Levels cannot be permuted by definition of ordering in levels. Consider the following histograms of ordinal measurement type where b = 8, the range = [0, 7] and n = 5:

$$H(D) = [0 5 0 0 0 0 0 0],$$

$$H(E) = [0 \ 0 \ 5 \ 0 \ 0 \ 0 \ 0],$$

$$H(F) = [0\ 0\ 0\ 0\ 0\ 0\ 5].$$

The distance between H(D) and H(E) must be smaller than that between H(D) and H(F) if histograms are ordinal in measurement type whereas they are the same in nominal measurement type. We will present the universal definition of distance that satisfies both *shuffling invariance* and *shuffling dependence* properties for nominal and other type measurements, respectively.

2.3. Difference between quantized measurement levels

Given a set of samples together with values of measurements (or attributes) made on the samples, and where the measurement is quantized into a discrete set of values (levels), a histogram represents the frequency of each discrete measurement. Corresponding to three types of measurements: nominal, ordinal and modulo, we define three measures of difference between two measurement levels, $x, x' \in X$ as follows:

nominal:
$$d_{\text{nom}}(x, x') = \begin{cases} 0 & \text{if } x = x', \\ 1 & \text{otherwise.} \end{cases}$$
 (2)

ordinal:
$$d_{\text{ord}}(x, x') = |x - x'|$$
. (3)

modulo:
$$d_{\text{mod}}(x, x') = \begin{cases} |x - x'| & \text{if } |x - x'| \leqslant \frac{b}{2}, \\ b - |x - x'| & \text{otherwise.} \end{cases}$$
(4)

The distance value between two nominal measurement sample values is either match or mismatch as shown in Eq. (2) and thus levels are permutable. Levels are totally ordered and non-permutable in ordinal measurement element values. The distance between two ordinal measurement values is the absolute difference between them as shown in Eq. (3). Finally, levels form a ring and non-permutable in modulo measurement element values. The distance between them is the interior difference as shown in Eq. (4). For example, for an angular measurement between 0° to 360° , $d_{\text{mod}}(350^{\circ}, 1^{\circ}) = 11^{\circ} \neq 349^{\circ}$.

The three measures in Eqs. (2)–(4) satisfy the following necessary properties of a metric:

$$d(x,x) = 0$$
 reflexivity, (5)

$$d(x, x') \ge 0$$
 non-negativity, (6)

$$d(x,x') = d(x',x)$$
 commutativity, (7)

$$d(x,x'') \le d(x,x') + d(x',x'')$$
 triangle ineq. (8)

Since they are straight-forward facts, we omit the proofs except for the triangle inequality of d_{mod} .

Fact 1. d_{mod} has triangle inequality property: $d_{\text{mod}}(x, x'') \leq d_{\text{mod}}(x, x') + d_{\text{mod}}(x', x'')$.

Proof. Let θ_1 be the interior angle between x and x'' and θ_2 and θ_3 be interior angles between x and x' and between x' and x'', respectively. There are four cases as shown in Fig. 2. Case (a) is such that x' lies between x and x''. Since $\theta_1 = \theta_2 + \theta_3$, $d_{\text{mod}}(x, x'') = d_{\text{mod}}(x, x') + d_{\text{mod}}(x', x'')$. Case (b) is such that $\theta_2 + \theta_3$ is the exterior angle between x and x''. As an exterior angle is always greater than or equal to their interior angle, $\theta_1 \leq \theta_2 + \theta_3$. Both cases (c) and (d) are such that either θ_2 or θ_3 covers θ_1 . There is no other case. Clearly, $\theta_1 \leq \theta_2 + \theta_3$. Therefore, $d_{\text{mod}}(x, x'') \leq d_{\text{mod}}(x, x') + d_{\text{mod}}(x', x'')$. \square

3. A new distance measure

The distance between any two histograms can be expressed in terms of the distances of element measurement values. Given two sets of n elements, A and B, we consider the problem as one of finding the minimum difference of pair assignments between two sets. The problem is to determine the best one-to-one assignment between two sets such that the sum of all differences between two individual elements in a pair is minimized. Given n

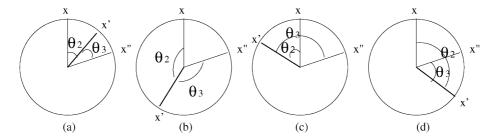


Fig. 2. 4 cases of 3 modulo measurement values.

elements $a_i \in A$ and n elements $b_j \in B$, we define the *Minimum difference of pair assignments* as

Minimum difference of pair assignments:

$$D(A,B) = \min_{A,B} \left(\sum_{i,j=0}^{n-1} d(a_i, b_j) \right),$$
 (9)

where D and d are designated as D_{nom} and d_{nom} , D_{ord} and d_{ord} , and D_{mod} and d_{mod} for nominal, ordinal and modulo measurements, respectively. \sum is a usual arithmetic summation in all three cases. The more similar the two histograms are, the smaller the value D(A,B) is. We are interested only in the value D(A,B) rather than the assignments.

As H(A) is a lossless representation of A, we define the distance measure between histograms, D(H(A), H(B)) = D(A, B) given in Eq. (9). Also, we shall use D(A, B) as a short form of the distance between two histograms, D(H(A), H(B)). First, we need to show that the proposed measure is indeed a metric so that it can be useful as a distance measure.

3.1. Metric property

We show that the new distance measure, D(A,B) satisfies conditions for being a metric: non-negativity, reflexivity, commutativity and triangle inequality. Since $d(a_i,b_j)$ is a metric, it follows that D(A,B) is also a metric that can be used to compare two histograms.

Fact 2. D(A,B) has non-negativity property: $D(A,B) \ge 0$.

Proof. D(A,B) is nothing but the sum of $d(a_i,b_j)$ and each $d(a_i,b_j)$ has non-negativity property. Therefore, D(A,B) also has the non-negativity by definition. \square

Fact 3. D(A,A) has reflexivity property: D(A,A) = 0.

Proof. Since $d(a_i, a_i) = 0$ is true, $\sum_{i=1}^n d(a_i, a_i) = 0$. Due to the non-negativity, 0 is the minimum bound of

the measure: $D(A,A) \ge 0$. Therefore, $D(A,A) = \sum_{i=1}^{n} d(a_i,a_i) = 0$ by definition. \square

Fact 4. D(A,B) has commutativity property: D(A,B) = D(B,A).

Proof. Let comm $(\sum_{i=1}^n d(a_i,b_i)) = \sum_{i=1}^n d(b_i,a_i)$ and vice versa. Due to the commutativity in $d(a_i,b_i)$, $\sum_{i=1}^n d(a_i,b_i) = \text{comm}(\sum_{i=1}^n d(a_i,b_i))$. Now suppose $D(A,B) \neq D(B,A)$ and $D(A,B) = \sum_{i=1}^n d(a_i,b_i)$. This means that $D(B,A) \neq \sum_{i=1}^n d(b_i,a_i)$. Let the assignment for D(B,A) be α . Then $\alpha < \sum_{i=1}^n d(b_i,a_i)$. Since $\alpha = \text{comm}(\alpha)$, $\text{comm}(\alpha) < \sum_{i=1}^n d(a_i,b_i)$. It contradicts that D(A,B) is the minimum difference of pair assignments. Therefore, D(A,B) = D(B,A) by contradiction.

Fact 5. *D* satisfies the triangle inequality property: $D(A, C) \leq D(A, B) + D(B, C)$.

Proof. Let the assignments $a_i \rightarrow b_i$ be the assignments of D(A,B). Let $b_i \rightarrow c_i$ be the assignments of D(B,C). Then a_i is assigned to c_i by $a_i \rightarrow b_i \rightarrow c_i$. However, $\sum_{i=1}^n d(a_i,c_i)$ is not necessarily the minimum and it may not be D(A,C). Thus, $D(A,C) \leqslant \sum_{i=1}^n d(a_i,c_i)$. Now from Eq. (8), the following equation can be drawn: $\sum_{i=1}^n d(a_i,c_i) \leqslant \sum_{i=1}^n d(a_i,b_i) + \sum_{i=1}^n d(b_i,c_i)$. $D(A,C) \leqslant \sum_{i=1}^n d(a_i,c_i) \leqslant D(A,B) + D(B,C)$. Therefore, $D(A,C) \leqslant D(A,B) + D(B,C)$.

3.2. Univariate case: example

Consider three sets of sample measurements with b = 8 and n = 10 as follows:

 $A = \{0, 0, 0, 0, 1, 2, 6, 6, 6, 7\}, B = \{0, 1, 1, 1, 1, 2, 6, 6, 6, 7\}$ and $C = \{0, 0, 1, 2, 6, 6, 6, 7, 7, 7\}$. The corresponding three univariate histograms are H(A) = [4, 1, 1, 0, 0, 0, 3, 1], H(B) = [1, 4, 1, 0, 0, 0, 3, 1] and H(C) = [2, 1, 1, 0, 0, 0, 3, 3]. We will use these three univariate histograms throughout the rest of this paper. Fig. 3 illustrates the minimum difference of pair assignments where $D_{\text{nom}}(A, C) = 2, D_{\text{ord}}(A, C) = 14$ and $D_{\text{mod}}(A, C) = 2$.

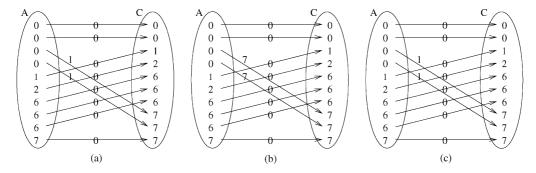


Fig. 3. Distances between H(A) and H(C); (a) Nominal: $D_{\text{nom}}[A, C] = 2$. (b) Ordinal: $D_{\text{ord}}[A, C] = 14$. (c) Modulo: $D_{\text{mod}}[A, C] = 2$.

3.3. Normalization

The numbers of collected samples for different classes are not always the same in practice. Thus, we provide a general definition for histogram distributions with arbitrary sample size. Let $N = CM(n_A, n_B)$ be the *common multiple* of n_A and n_B where n_A and n_B are the numbers of samples in integer in set A and B. One common multiple is $N = n_A \times n_B$. Now we can obtain the new histograms $H^N(A)$ and $H^N(B)$ with the same size of samples by following Eqs. (10) and (11) on each level.

$$H_l^N(A) = n_B H_l(A), \tag{10}$$

$$H_l^N(B) = n_A H_l(B). \tag{11}$$

The normalized distance is defined as follows:

$$D^{N}(H(A), H(B)) = \frac{D(H^{N}(A), H^{N}(B))}{N}$$

$$= \frac{\sum_{i=0}^{b-1} |\sum_{j=0}^{i} (H_{j}^{N}(A) - H_{j}^{N}(B))|}{N}.$$
(12)

Output values in Eq. (12) are real numbers while those in Eq. (26) are integer values. Eq. (12) is the general and normalized form of Eq. (26) as all metric properties are preserved.

Lemma 6. Let N_1 and N_2 be two common multiples. The normalized distances by any common multiple are the same

$$\frac{D(H^{N_1}(A), H^{N_1}(B))}{N_1} = \frac{D(H^{N_2}(A), H^{N_2}(B))}{N_2}.$$

Proof. Consider the least common multiple, $N_0 = \text{LCM}(n_A, n_B)$. Then all other common multiples, α are $N_\alpha = cN_0$ where c is a positive integer. The histogram $H^{N_\alpha}(A)$ can be viewed as $H^{N_\alpha/c}(A)$. The minimum distance of $D(H^{N_\alpha}(A), H^{N_\alpha}(B))$ is equivalent to $c \times D(H^{N_\alpha/c}(A), H^{N_\alpha/c}(B))$, because we can consider c

samples as one unit.

$$\begin{split} D^{N_{\alpha}}(H(A), H(B)) &= \frac{D(H^{N_{\alpha}}(A), H^{N_{\alpha}}(B))}{N_{\alpha}} \\ &= \frac{c \times D(H^{N_{0}}(A), H^{N_{0}}(B))}{N_{\alpha}} \\ &= \frac{D(H^{N_{0}}(A), H^{N_{0}}(B))}{N_{0}} \\ &= D^{N_{0}}(H(A), H(B)). \quad \Box \end{split}$$

In order to show the triangle inequality property, consider multiple histograms, H(A), H(B) and H(C) with different sizes.

Theorem 1. $D^{N_{AB}}(H(A), H(B)) \leq D^{N_{AC}}(H(A), H(C)) + D^{N_{CB}}(H(C), H(B)).$

Proof. The matrix of normalized distances by individual two histograms is equivalent to the matrix of distances of $H^N(A)$, $H^N(B)$ and $H^N(C)$ where $N = CM(n_A, n_B, n_C)$ by the Lemma 6. Therefore, $D^N(H(A), H(B)) \leq D^N(H(A), H(C)) + D^N(H(C), H(B))$. By definition of distance, $D^N(H(A), H(B))$ is the minimum distance between two histograms. Suppose $D^N(H(A), H(B)) > D^N(H(A), H(C)) + D^N(H(C), H(B))$, then $D^N(H(A), H(B))$ is not the minimum distance and it contradicts the definition. Therefore, the normalized distance holds the triangle inequality. \square

3.4. Comparison with earth mover's distance

The earth mover's distance or simply (EMD) was proposed by Rubner and Guibas [14]. Given two distributions, one can be seen as a mass of earth properly spread in space, the other as a collection of holes in that same space. They assumed that there is at least as much earth as needed to fill all the holes to capacity by switching what they call earth and what they call holes if necessary. In short, the EMD is defined as minimum over all

flows $f_{i,j}$ of $\sum_{i,j=1}^{n,m} f_{i,j}d(x_i, y_j)$ where |X| = n, |Y| = m. And then, they formulated a linear constrained optimization problem and proceeded to compute the *EMD* based on a solution to the old *transportation problem*.

Once suitable constraints are imposed on the set of all possible flows, the *minimum difference of pair assignment* of two distribution problem can be formalized and solved by a special case solution of the *EMD*. The *MDPA* can be realized as a special univariate histogram case of the *EMD*. The *EMD* considers more general ground distances that it works for multivariate histograms as well as weighted clusters [14]. However, The histogram distance utilizing the *MDPA* has been studied independently from Rubner's *EMD* and we give much efficient algorithms to compute the *MDPA* for three types of univariate histograms described in Section 5 as they are commonly encountered in problems such as histogram-based image indexing and feature selection.

4. Conventional definitions

There are several definitions of distance (or similarity) between histograms, based on vectors, probabilities, and clusters. Ten such distances defined in the literature are given below and denoted as D_1 – D_{10} . Their inadequacy when they are used to compute the distance between certain type histograms is considered.

4.1. List of definitions

A histogram is treated as a *b*-dimensional vector, and hence the standard vector norms can be used as distances between two histograms as follows:

City block (L_1 -norm):

$$D_1(A,B) = \sum_{i=0}^{b-1} |H_i(A) - H_i(B)|.$$
 (13)

Euclidean (L_2 -norm):

$$D_2(A,B) = \sqrt{\sum_{i=0}^{b-1} (H_i(A) - H_i(B))^2}.$$
 (14)

Another approach is a normalized similarity measure, S(A, B) based on the intersection between two histograms [7]

Intersection:
$$S(A,B) = \frac{\sum_{i=0}^{b-1} \min(H_i(A), H_i(B))}{\sum_{i=0}^{b-1} H_i(A)}$$

$$= \frac{1}{n} \sum_{i=0}^{b-1} \min(H_i(A), H_i(B)). \quad (15)$$

Intersection (15) of two histograms is the same as Bayes P_e , the minimum misclassification (or error) probability,

which is computed as the overlap between two PDF's, P(A) and P(B) [8]. To compute this as a distance measure, we will convert S(A,B) using the inverse operation: $n \times (1 - S(A,B))$:

Non-Intersection:

$$D_3(A,B) = n - \sum_{i=0}^{b-1} \min(H_i(A), H_i(B)).$$
 (16)

Measures D_1 – D_3 are widely used for histogram-based image indexing and retrieval [5,7].

The following lemma states that distance measures D_1 and D_3 are closely related when the size of two sets are equal. It suggests an alternative algorithm for $D_{\text{nom}}(A, B)$ later in Section 5.1.

Lemma 7. $D_1 = 2 \times D_3$, provided |A| = |B| = n.

Proof. Since for two integers $H_i(A)$ and $H_i(B)$, $\min(H_i(A), H_i(B)) = (H_i(A) + H_i(B) - |H_i(A) - H_i(B)|)/2$, it follows that $\sum_{i=0}^{b-1} \min(H_i(A), H_i(B)) = (2n - \sum_{i=0}^{b-1} |H_i(A) - H_i(B)|)/2$. By rearranging the equation, $\sum_{i=0}^{b-1} |H_i(A) - H_i(B)| = 2n - 2\sum_{i=0}^{b-1} \min(H_i(A), H_i(B))$. Thus, $D_1 = 2 \times D_3$. \square

Discrete versions of distance between probability density functions are also useful as distances between histograms as follows:

K-L distance:

$$D_4(A,B) = \sum_{i=0}^{b-1} P_i(B) \log \frac{P_i(B)}{P_i(A)}.$$
 (17)

Bhattacharyya distance:

$$D_5(A,B) = -\log \sum_{i=0}^{b-1} \sqrt{P_i(A)P_i(B)}.$$
 (18)

Matusita distance:

$$D_6(A,B) = \sqrt{\sum_{i=0}^{b-1} (\sqrt{P_i(A)} - \sqrt{P_i(B)})^2}.$$
 (19)

Note that K–L distance is not a true metric, rather it is the relative entropy.

Distance measures between clusters [8] can be regarded as distances between histograms as follows:

Nearest-neighbor:
$$D_7(A,B) = \min_{a_i \in A, b_j \in B} d(a_i,b_j)$$
. (20)

Furthest-neighbor:
$$D_8(A,B) = \max_{a_i \in A, b_j \in B} d(a_i,b_j)$$
. (21)

Mean distance: $D_9(A, B) = d(m_A, m_B)$,

where m_A and m_B are means of A and B. (22)

Table 1 Comparisons of distance measures D_1 – D_6 and D_{ord}

	A, B	A, C	B, C	$\operatorname{argmin}(D_x(A,\cdot))$
$\overline{D_1}$	6	4	6	\overline{C}
D_2	4.24	2.83	3.74	C
D_3	3	2	3	C
D_4	0.42	0.19	0.33	C
D_5	0.11	0.04	0.09	C
D_6	0.45	0.30	0.41	C
D_{ord}	3	14	13	В

Average distance:
$$D_{10}(A,B) = \frac{1}{n^2} \sum_{a_i \in A} \sum_{b_j \in B} d(a_i,b_j),$$
 (23)

where $d(a_i, b_j)$ can be defined according to measurement type.

4.2. Analysis of distance measures in various measurement types

Distance D_1 – D_6 are always the same regardless of the type of measurements. Each bin along the level is viewed as an individual independent feature because the correlation between these bins is not considered in computing the distance between two histograms. We shall examine definitions D_1 – D_{10} when they are applied to each type of measurements.

4.2.1. Ordinal

To show the inadequacy of D_1 – D_6 , consider the following example. Let x represent the length of fish in a pond. Let A, B and C in Section 3.2 represent samples drawn from three ponds. We are interested in determining the statistical similarity of fish in each of the three ponds. Note that length is an ordinal measurement. We wish to find the histogram most similar to H(A). H(A) and H(B) have more baby fish whereas H(C) has more adult fish. Three fish out of 10 in the group A differ by 1 in each from the group B whereas two fish differ by 7 in from the group C. Based on the distance between sets, H(A) is closer to H(B) rather than to H(C). Definitions D_1 – D_6 are excellent in counting the number of mismatches but do not consider the difference (inches in fish example) of each mismatch.

The distance values D_1 – D_6 and D_{ord} computed by enumerated definitions are shown in Table 1.

The smallest distance value between two histograms indicates the closest histogram pair. Note that only D_{ord} returns H(B) as the histogram closest to H(A) whereas D_1 – D_6 return H(C) as the closest.

The inadequacy of definitions D_1 – D_6 on ordinal type histograms can be explained by the following "shuffling

Table 2 Comparisons of distance measures D_7 – D_{10} when used in ordinal measurement types

	A,B	A,C	В,С	$\operatorname{argmin}(D_x(A,\cdot))$				
$\overline{D_7}$	0	0	0	tie				
D_8	7	7	7	tie				
D_9	0.3	1.4	1.1	B				
$D_7 \\ D_8 \\ D_9 \\ D_{10}$	3.02	3.36	3.18	В				

invariance" property. A distance measure between histograms is "shuffling invariant" if and only if the distance does not change when levels, $\{x_0, x_1, \ldots, x_{b-1}\}$ in histograms are permuted or reordered. Measures $D_1 - D_6$ have this property of "shuffling invariance". They are sums of individual distances of each level and due to the commutative law, the distances do not change when levels are permuted among themselves.

In case of ordinal type histograms, the levels cannot be shuffled because of the correlation among levels. If the resulting matrices are not affected by shuffling, the definition of distance is not suitable for ordinal or modulo type histograms. The extreme example in Section 2.2 convinces us why the conventional definitions are inappropriate for the ordinal type histograms as definitions D_1 – D_6 fail to tell which histograms are more similar than others.

Now consider the distance measures between clusters as shown in Table 2. In the given example of three histograms, D_7 and D_8 return 0 and 7 for all cases and thus do not discriminate the distances D(A,B), D(A,C) and D(B,C). The *mean* method, D_9 , does return B as the closest one to the set A in ordinal measurement case. However, this method has a disadvantage that it does not discriminate multi-modal histograms. The mean value can be equal although one histogram is unimodal and the other is bimodal.

Finally, the average method, D_{10} , is quite compatible with the MDPA measure. However, its resulting matrix does not have the reflexivity property that is $D(A,A) \neq 0$ but = 3.08. Suppose that there is a set D that is identical to the set A. This measure does not return D as the closest set. Another disadvantage of this method is its high complexity, $O(n^2)$ whereas the MDPA measure in Eq. (9) is much quicker and it is discussed in the following section.

4.2.2. Nominal

Now suppose the measurement type is nominal. The distance measures D_1 – D_6 return the exactly same matrix as the ordinal measurement type as given in Table 1. This is one disadvantage of D_1 – D_6 . Table 3 shows the comparisons of the MDPA measure with D_7 – D_{10} when they are used for the nominal type measurement. It is quite compatible with all measures from D_1 – D_6 and it

Table 3 Comparisons of distance measures D_7 – D_{10} and D_{nom} when used in nominal measurement types

	A,B	A, C	B, C	$\operatorname{argmin}(D_x(A,\cdot))$
$\overline{D_{\mathrm{nom}}}$	3	2	3	C
D_7	0	0	0	tie
D_8	1	1	1	tie
D_8 D_9				Not applicable
D_{10}	0.81	0.78	0.81	C

Table 4 Comparisons of distance measures D_7 – D_{10} and D_{mod} when used in modulo measurement types

	A, B	A, C	B, C	$\operatorname{argmin}(D_{x}(A,\cdot))$
$\overline{D_{\mathrm{mod}}}$	3	2	5	\overline{C}
D_7	0	0	0	tie
D_8	4	4	4	tie
D_8 D_9				Not applicable
D_{10}	1.58	1.44	1.62	C

is exactly the same as D_3 and $D_1/2$. According to the definition of the difference between quantized measurement levels given for the nominal type measurement in Eq. (2), some of distances between nominal type clusters are computed and shown in Table 3. Again, D_7 and D_8 are meaningless. Also, *Mean* cannot be defined for the nominal type measurement. D_{10} is criticized in the same way as in the ordinal type case.

4.2.3. Modulo

Finally, consider modulo measurement type. Again, the distance measures from D_1 – D_6 return the exactly same matrix as the ordinal measurement type. Table 3 shows the comparisons of the MDPA measure with D_7 – D_{10} when they are used for the modulo type measurement. According to the definition of the difference between quantized measurement levels given for the modulo type measurement in Eq. (4), some of distances between modulo type clusters are shown in Table 4. Again, D_7 and D_8 are meaningless. Also, *Mean* cannot be defined for the modulo type measurement; what is the mean of 0° and 180° ? Is it 90° , 270° or none? Note that D_{10} has the desirable property that it varies depending on the type of measurements. However, again D_{10} is criticized in the same way as in other measurement type cases.

5. Algorithms

A naive way to solve the distance between histograms, that is the minimum difference of pair assignments, can be exponential in time as there are n! number of possible assignments. Rubner and Guibas used the well-known transportation simplex algorithm to solve this linear optimization problem [14], yet it is too slow. In this section, we introduce efficient algorithms for univariate histograms for each type of measurement variable: nominal $\Theta(b)$, ordinal $\Theta(b)$ and modulo $O(b^2)$ insofar as histograms are given. Note that the number of levels, b is usually much smaller than the size of sample, e.g., the grey level (256) vs. the size of the image.

For the nominal type histograms, the half of the city block distance shown later in Eq. (13) as a distance is equivalent to the minimum difference of pair assignments in Eq. (9). For ordinal and modulo type histograms, the measure D(H(A), H(B)) can be realized as the necessary cell movements to transform one histogram into the target histogram as shown in Fig. 4. The minimum cost of moving cells within a histogram to make the same configuration as the target one is equivalent to the minimum difference of pair assignments. There needs a few steps of moving cells if two histograms have similar distribution.

5.1. Nominal type histogram

A distance between nominal type histograms is the number of elements that do not overlap or intersect, which is equivalent to Eq. (16). Hence, the definition shown in Eq. (9) becomes

$$D_{\text{nom}}(A, B) = \min_{A, B} \left(\sum_{i,j=0}^{n-1} d_{\text{nom}}(a_i, b_j) \right)$$
$$= n - \sum_{i=0}^{b-1} \min(H_i(A), H_i(B)). \tag{24}$$

The algorithm for Eq. (24) is straightforward and we will not discuss it in detail here. As an alternative algorithm, one can solve this problem using the City Block Distance in Eq. (13) as discussed in Lemma 7.

$$D_{\text{nom}}(A,B) = \frac{\sum_{i=0}^{b-1} |H_i(A) - H_i(B)|}{2}.$$
 (25)

In either equation, the computational time complexity is $\Theta(b)$.

5.2. Ordinal type histogram

A ordinal type histogram is a histogram whose level, *x* increases linearly. Many histograms fall into this category such as grey level, height, weight, length, temperature, and so forth. Earlier work on ordinal type histogram, motivated to expedite the image template matching problem, has been introduced briefly [16,17]. In this section,

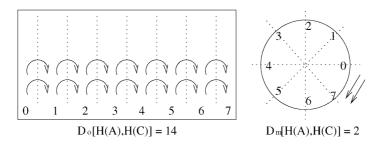


Fig. 4. Arrow representation of $D_{\text{ord}}(H(A), H(B))$ and $D_{\text{mod}}(H(A), H(B))$.

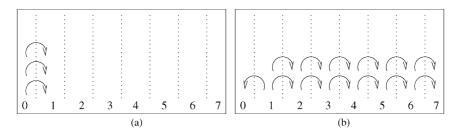


Fig. 5. Arrow representation; (a) D[H(X), H(Y)]. (b) D[H(Y), H(Z)].

we present the detailed description and analysis of the algorithm to compute the distance between two ordinal type histograms.

As discussed earlier, a histogram H(A) can be transformed into H(B) by moving elements to left or right and the total of all necessary minimum movements is the distance between them. There are two operations. Suppose a cell or element s belongs to a bin l. One operation is *Move left(s)*. This operation results that the cell s belongs to a bin l-1 and the cost to do so is 1. This operation is impossible for cells in the left-most bin. Another operation is *Move right*(s). Similarly, after the operation, s belongs to the bin l+1 and the cost is 1. The same restriction applies to the right-most bin. These operations are expressed in the arrow representation of two histograms as shown in Fig. 5. Fig. 5(a) shows the minimum number of cell operations required to transform of H(A) into H(B). The total number of arrows is the distance. It is the shortest movement and there is no other way to move cells in shorter steps to build the target histogram. In general, the number of arrows to transform H(A) into H(B) is equivalent to the cost of assigning elements in A to those in B and the minimum number of arrows necessary to transform them, hence D(H(A)), H(B)) is equivalent to D(A,B) in Eq. (9).

There are only one type of directional arrows along the border line between two levels or bins in order for the arrow representation to be the minimum. If there is a border line containing both directional arrows, they can be cancelled out without affecting the transform H(A) into H(B). Cancelling out operation reduces the total costs by two. This means that the arrow representation with mixed directions on a border line is not the minimum cost configuration.

The distance in histograms, that is the minimum number of necessary arrows in the arrow representation, is defined as follows for ordinal type histograms:

$$D_{\text{ord}}(H(A), H(B)) = \sum_{i=0}^{b-1} \left| \sum_{j=0}^{i} (H_j(A) - H_j(B)) \right|.$$
 (26)

It is the sum of absolute values of prefix sum of difference for each level. Therefore, the algorithm for finding the minimum distance between two histograms consists of three steps. The first step is to obtain the differences for each level. The second step is to calculate the prefix sum of the differences for each level. Finally, the absolute values of the prefix sums are added. The following pseudo code shows the exact steps.

Algorithm 1. Distance-ordinal-histogram(int*A, int*B)

```
    prefixsum = 0
    h_dist = 0
    for i = 0 to b - 1
    prefixsum + = A[i] - B[i]
    h_dist + = |prefixsum|
    return(h_dist)
```

For the example of H(A) and H(C), Algorithm 1 performs the following calculations:

The lines (1) and (2) represent the histogram H(A) and H(C), respectively, and the line (3) is the difference between elements in (1) and (2) on each level. The line (4) is the prefix sum of the elements in line (3). Note that the last element in the prefix sum list is always 0 since both histograms are of same size. The final step is adding the absolute value of each element in the prefix sum list, which is 14.

Both time and space complexities are $\Theta(b)$. The algorithm requires only two integer variables and two arrays for histograms.

5.2.1. Correctness

The following lemma is crucial since it will serve as a stepping stone to support the algorithm. Suppose that we have successfully constructed the arrow representation of the histograms such that the distance is the minimum.

Lemma 8. Let A_l denote the number of arrows from the bin l to l+1. It is positive if arrows are heading to right, or negative otherwise.

$$A_{l} = \sum_{i=0}^{l} H_{i}(A) - \sum_{i=0}^{l} H_{i}(B) = \sum_{i=l+1}^{b-1} H_{i}(B) - \sum_{i=l+1}^{b-1} H_{i}(A).$$

Proof. Consider two sub-histograms, $H_{0..l}(A)$ and $H_{0..l}(B)$ where bins are 0 to l. After transforming, population of $H_{0..l}(B) + A_l$ must be equal to that of $H_{0..l}(A)$. Suppose $A_l \neq \sum_{i=0}^l H_i(A) - \sum_{i=0}^l H_i(B)$. Then there is no way to transform $H_{0..l}(A)$ to $H_{0..l}(B) + A_l$. By contradiction

$$A_{l} = \sum_{i=0}^{l} H_{i}(A) - \sum_{i=0}^{l} H_{i}(B).$$
 (27)

Now the total population is $n = \sum_{i=0}^{l} H_i(A) + \sum_{i=l+1}^{b-1} H_i(A)$ and $\sum_{i=0}^{l} H_i(A) = n - \sum_{i=l+1}^{b-1} H_i(A)$. Similarly, $\sum_{i=0}^{l} H_i(B) = n - \sum_{i=l+1}^{b-1} H_i(B)$. Replacing the terms in (27), $A_l = \sum_{i=l+1}^{b-1} H_i(B) - \sum_{i=l+1}^{b-1} H_i(A)$. \square

The lemma implies that A_l is the difference of populations between two sub-histograms in the left-hand side of the border line of the bin l and l + 1.

Theorem 2. Algorithm 1 correctly finds the minimum distance between two histograms.

Proof. As Lemma 8 is true for all levels, the minimum distance is $\sum_{i=0}^{b-1} |A_i| = \sum_{i=0}^{b-1} |\sum_{j=0}^{i} H_j(A) - \sum_{j=0}^{i} H_j(B)|$. This is equivalent to the Eq. (26): $\sum_{i=0}^{b-1} |\sum_{j=0}^{i} (H_j(A) - H_j(B))|$. \square

5.3. Modulo type histogram

One major difference in an modulo type histogram is that the first bin and the last bin are considered to be adjacent to each other, and hence it forms a closed circle, due to the nature of the data type. Transforming such an modulo type histogram should allow cells to move from the first bin to the last bin or vice versa at a cost of a single movement. This results in a different distance value in modulo type histograms from the one in ordinal type histograms. The same histograms H(A), H(B) and H(C) are now treated as modulo type histograms and redrawn as shown in Fig. 6. The number inside of each slice represents the level of a bin. Table 4 indicates that the two histograms H(A) and H(C) are the closest pair and D(H(A), H(C)) = 2 is achieved by moving two cells from bin 0 in H(A) to bin 7 clockwise. Clearly, the difference in measurement type necessitates a new algorithm to find the distance between modulo type histograms. In this section, we modify Algorithm 1 to construct the algorithm for distance between modulo type histograms [18].

5.3.1. Properties

Before embarking on the new algorithm, it is important to discuss the properties of the arrow representation of the distance between two modulo type histograms. Consider another modulo type histograms, H(D) and H(E)as shown in Fig. 7. Blocks or cells can move to clockwise or counter-clockwise directions. Each cell movement to the next level in either direction costs 1. The minimum cost required to build the target histogram from a given histogram is the distance. Again, an intuitively appealing way to explain the distance is to use the arrow representation of two histograms as shown in Fig. 7. If one establishes an arbitrary one-to-one mapping for the cells between two histograms, one can transform H(D)into H(E) by moving cells in H(D) to the corresponding position in H(E). For the example in Fig. 7(a), the arrows α , β and γ indicate the path from the cell 0 in H(D)to the cell 0 in H(E). There are n! number of ways to transform in this manner. Among these ways, there exists a minimum distance whose number of movements is the lowest. Some element movements are illustrated as an arrow representation in Fig. 7. As a matter of fact, an modulo representation that satisfies the following properties gives the minimum configuration of D(H(D), H(E)).

Property 1. Arrows must be one directional on each border line.

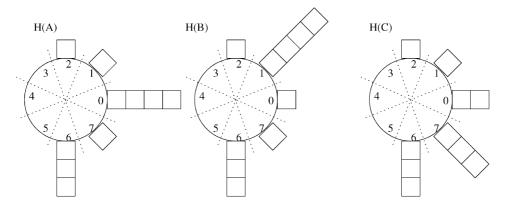


Fig. 6. Modulo representation of H(A), H(B) and H(C).

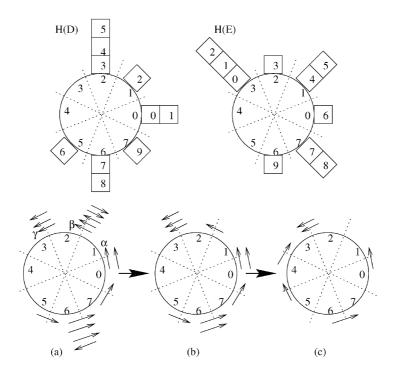


Fig. 7. Modulo histograms and angular arrow representation.

Property 2. The number of border lines of one direction cannot exceed b/2.

Property 3. No more reduction occurs when either basic operations are applied.

As discussed before in ordinal type histogram case, if there is a border line that has both directional arrows, they are cancelled out. These movements are redundant ones. The configuration in Fig. 7(a) becomes one in Fig. 7(b) by cancelling out the opposite arrows on each border line. By the property 1, there exists no border line of mixed directional arrows and each border line has either clockwise or counter-clockwise directional arrows. Suppose that the number of border lines of one direction is b' > b/2 and the first number of arrows were k, then after adding the circle, the number of arrows becomes k + b. By cancelling out the opposite directions on the same border line, we have k + b - 2b' < k. Therefore, if the number of border lines of one direction exceeds

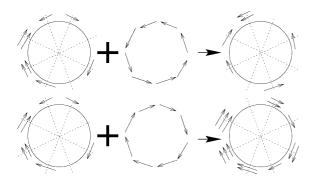


Fig. 8. Two basic operations.

b/2, this is not a configuration of the minimum distance. From these properties, two important basic operations are derived as shown in Fig. 8. A complete circle in the chain of same directional arrows can be added in any direction and then the opposite arrows on the same line are cancelled out.

Lemma 9. Let A_l denote the number of counter-clockwise arrows from bin l-1 to l. A_l is positive if arrows are counter-clockwise and negative otherwise. $H_l(A) = H_l(B) + A_l - A_{l+1}$.

Proof. Similar to Lemma 8. After transforming, population of cells on each level of both histograms must be the same. \Box

Since Lemma 9 is true for all levels, affecting cells in one bin means affecting all other bins as a chain reaction. Hence, there are only two possible operations to affect changes as shown in Fig. 8. The arrow representation of the minimum distance value is always constructible by the combination of these two basic operations. Consider a non-minimum distance arrow representation. By applying one of two operations of adding a clockwise or counter-clockwise circle, the lower number of arrows is achieved whereas it is unchanged or increased if the arrow representation has the minimum distance. Although first two properties are sufficient for the arrow representation of the minimum distance, the third property alone is also sufficient since it admits the other properties.

5.3.2. D_{mod} algorithm

An algorithm to compute the distance between modulo type histograms in $O(b^2)$ is presented. It gets an initial arrow representation from Algorithm 1 and then use two basic operations to derive the minimum distance arrow representation that guarantees all properties discuss in the previous section.

Algorithm 2. Distance-modulo-histogram(int *A, int *B)

```
prefixsum[0] = A[0] - B[0]
    for \forall i \text{ prefixsum}[i] = \text{prefixsum}[i-1] +
     A[i] - B[i]
     h_{\text{dist}} = \sum_{i=0}^{b-1} |\text{prefixsum}[i]|
 4
     for(;;)
 5
         d = min(positive prefixsum[i])
         for \forall i \text{ temp[i]} = prefixsum[i] - d
         h_{\text{dist2}} = \sum_{i=0}^{b-1} |\text{temp}[i]|
 7
 8
         if h\_dist2 < h\_dist
 9
          h_dist = h_dist2
10
          for \forall i \text{ prefixsum}[i] = \text{temp}[i]
11
       else break:
12 for(;;)
13
       d = max(negative prefixsum[i])
       for \ \forall i \ temp[i] = prefixsum[i] - d
14
       h_{\text{dist2}} = \sum_{i=0}^{b-1} |\text{temp[i]}|
15
16
       if h_{-}dist2 < h_{-}dist
17
          h_dist = h_dist2
18
          for \forall i \text{ prefixsum}[i] = \text{temp}[i]
19
       else break;
20 return(h_dist)
```

The algorithm is explained using the example shown in Fig. 7 along with the following calculations:

2	1	3	0	0	1	2	1				(1)
1	2	1	3	0	0	1	2			• • •	(2)
1	-1	2	-3	0	1	1	-1				(3)
1	0	2	-1	-1	0	1	0	\Rightarrow	6		(4)
1	0	2	-1	-1	0	1	0	\Rightarrow	6		(5)
0	-1	1	-2	-2	-1	0	-1	\Rightarrow	8		(6)

The line (3) is the difference between (1) and (2) (steps 1 and 2 in Algorithm 2). The line (4) is the initial arrow representation that is the prefix sum of the difference and the sum of the absolute value of these numbers (step 3). Note that steps 1–3 are exactly the same as Algorithm 1 that guarantees the property 1.

To ensure the Properties 2 and 3, two basic operations in Fig. 8 are applied repeatedly. First, circles of clockwise arrows are added to the current arrow representation until there is no more reduction on the total number of arrows (steps 4–11). The line (5) is the result of these steps. Next, circles of the counter-clockwise arrows are added in the similar manner (steps 12–19). The line (6) is the result of adding a circle and the resulting value is greater than the previous one. Therefore, the distance is 6.

5.3.3. Correctness

The correctness of the algorithm is asserted by the following theorem.

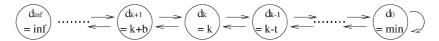


Fig. 9. Relation between valid arrow representations.

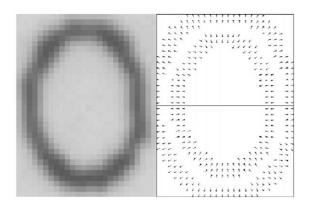


Fig. 10. Gradient direction map.

Theorem 3. Algorithm 2 correctly finds the minimum distance between two modulo histograms.

Proof. The arrow representation of minimum distance can be achieved from any arbitrary valid arrow representation by a combination of two basic operations. Fig 9 illustrates the relation between valid arrow representations. The arrows indicate one of the basic operations and the opposite arrow represents the other basic operation. All valid representations are related as a string and the distance value can increase infinitely. There exists only one minima among valid arrow representations. In order to reach to the minima, first test for the one of the two operations whether it gives higher or lower distance value. If the distance reduces, keep applying the operation until no more reduction occurs. Otherwise, check for the other operation in similar manner. Algorithm 2 first computes an arrow representation by Algorithm 1 and then applies the clockwise operation repeatedly until no more reduction occurs and then the counter-clockwise operation similarly. This guarantees the Property 3. Therefore, Algorithm 2 is correct. \square

Algorithm 2 runs in $O(b^2)$ time. The lines 1–3 is $\Theta(b)$. The lines 4–11 takes $O(b^2)$ because each iteration removes at least one positive number in the list and there can be up to b-1 number of positive numbers in the arrow representation. Similarly, the lines 12–19 takes $O(b^2)$.

Theorem 4. The worst-case time complexity of Algorithm 2 is $O(b^2)$.

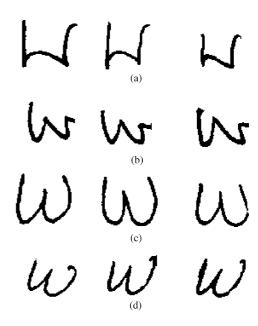


Fig. 11. Sample *W*'s; (a) Author "A". (b) Author "B". (c) Author "C". (d) Author "D".

Proof. Here is a worst-case example of two modulo histograms with 30 elements and 10 bins.

The distance is 52. The worst case is that the size of either positive or negative consecutive numbers is b-1. Each iteration reduces the size by 1 or 2. Therefore, the running time is $O(b^2)$. \square

The space required for the algorithm is O(b).

6. Experiment on character writer identification

We show the experimental results of the character writer identification problem using the earlier defined *MDPA* measure on *Gradient directions* histograms.

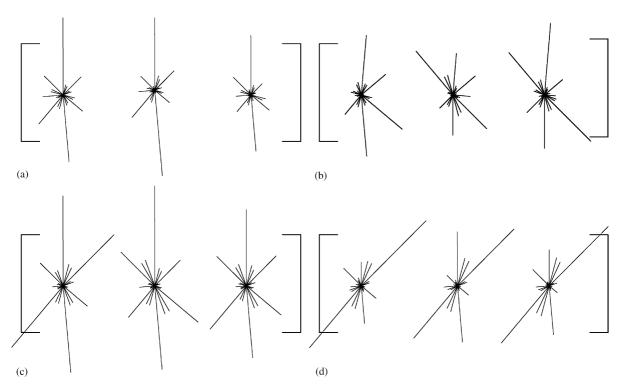


Fig. 12. Angular representation of gradient direction histograms for sample W's in Fig. 11; (a) Modulo histograms of author "A". (b) Modulo histograms of author "B". (c) Modulo histograms of author "C". (d) Modulo histograms of author "D".

We consider this modulo type histogram features of a character as one of character level image signatures for identification.

Gradient direction features are computed by the following *Sobel edge detection mask* operators [19] where I(i, j) represents the image of a character.

$$\underbrace{\begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix}}_{\text{Row mask}} \underbrace{\begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}}_{\text{Column mask}}$$

$$S_x(i,j) = I(i-1,j+1) + 2I(i,j+1) + I(i+1,j+1)$$

- $I(i-1,j-1) - 2I(i,j-1) - I(i+1,j-1),$

$$S_y(i,j) = I(i-1,j-1) + 2I(i-1,j) + I(i-1,j+1)$$

- $I(i+1,j-1) - 2I(i+1,j) - I(i+1,j+1),$

magnitude =
$$\sqrt{S_x^2(i,j) + S_y^2(i,j)}$$
,
direction = $\tan^{-1} \frac{S_y(i,j)}{S_x(i,j)}$. (28)

A sample of the gradient direction maps of a character image is shown in Fig. 10

Three character "W" 's per author are extracted and their gradient direction histograms are computed. There are four writers, $\{A, B, C, D\}$ and the corresponding angular representation of gradient direction histograms are shown in Figs. 11 and 12, respectively.

Table 5 shows the distance matrix. When two-dimensional information is represented in the one-dimensional histogram, certain information is lost. Therefore, while it is true that the two histograms from the similar character images tend to be similar, the reverse statement is not always true that two images with the similar histograms tend to be similar. For example, the second sample histogram from author "A" is similar to the first sample from author "C" although their characters are dissimilar. Yet, this histogram distance information are one of the most effective features [21] to determine the similarity of two letters and we claim that distances in Table 5 tend to be small if they were written by a same author. The writer verification system using the dichotomy model with 1000 writers with 3 samples each writer can be found in [20] and the detailed report on the performance of the angular histogram measure on the gradient direction histogram can be found in [18,21].

Table 5			
D_{mod}^{N} matrix of gradient	direction	histograms	of writers

	A-1	A-2	A-3	B-1	B-2	B-3	C-1	C-2	C-3	D-1	D-2	D-3
A-1	0.00	0.99	0.86	1.85	2.24	2.14	1.54	1.59	1.22	3.15	2.94	3.43
A-2	0.99	0.00	1.41	2.20	2.67	2.64	1.39	1.58	1.17	2.68	2.44	2.91
A-3	0.86	1.41	0.00	1.26	1.56	1.34	2.06	1.19	1.35	3.79	3.69	4.22
B-1	1.85	2.20	1.26	0.00	1.00	0.87	2.75	1.26	1.43	4.65	4.54	5.07
B-2	2.24	2.67	1.56	1.00	0.00	0.82	3.23	1.42	1.81	5.12	5.01	5.55
B-3	2.14	2.64	1.34	0.87	0.82	0.00	3.22	1.41	1.80	5.11	5.00	5.54
C-1	1.54	1.39	2.06	2.75	3.23	3.22	0.00	2.12	1.41	2.19	1.92	2.48
C-2	1.59	1.58	1.19	1.26	1.42	1.41	2.12	0.00	0.78	4.04	3.91	4.44
C-3	1.22	1.17	1.35	1.43	1.81	1.80	1.41	0.78	0.00	3.34	3.20	3.74
D-1	3.15	2.68	3.79	4.65	5.12	5.11	2.19	4.04	3.34	0.00	1.06	0.85
D-2	2.94	2.44	3.69	4.54	5.01	5.00	1.92	3.91	3.20	1.06	0.00	1.20
D-3	3.43	2.91	4.22	5.07	5.55	5.54	2.48	4.44	3.74	0.85	1.20	0.00

7. Conclusions and future work

We have criticized inadequacy of the way that existing definitions, D_1 – D_6 are used for ordinal and modulo type histograms. We considered three types of histograms characterized by their measurement type: nominal, ordinal and modulo. Different algorithms are designed to compute the histogram distance for each type of histograms; Eq. (24), Algorithms 1 and 2, correspondingly. Their computational time complexities are $\Theta(b)$, $\Theta(b)$ and $O(b^2)$, respectively, insofar as the histograms are given. These algorithms are based on one concept of distance between sets that is the problem of minimum difference of pair assignments that is similar to the EMD.

We introduced the problem of minimum difference of pair assignments to grasp the concept of the distance between two histograms. Extending the suggested Algorithms 1 and 2 facilitates the solution to this problem in $\Theta(n + b)$ and $O(n + b^2)$ time for ordinal and modulo type univariate data, respectively, reflecting the time complexity of $\Theta(n)$ to build the histogram.

Albeit the histograms that we dealt with in this paper are one-dimensional arrays (univariate), there can be any dimensional ones and measuring the distance between multivariate histograms in Eq. (9) can be useful in many applications [22]. Multivariate histograms is based on several measurement variables. Only difference is how $d(a_i,b_j)$ is defined. For example, if all variables are ordinal, the *Euclidean* norm in vector space, denoted by $||\cdot||$, is often used for the difference between quantized measurement levels

$$d_{\text{ord}}(a_i, b_i) = ||a_i - b_i||. \tag{29}$$

Various distance measures [1] such as *Minkowski* or *Tanimoto* can be used in place of the *Euclidean* norm.

However, designing the algorithms to compute the distance between multivariate histograms is non-trivial because there are n! number of possible assignments and because of high dimensionality. Although one can solve it by using the solution to the transportation problem, faster algorithms are on demands. We leave them as open problems to readers and the future work.

References

- R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, 2nd Edition, Wiley, New York, 2000.
- [2] J.E. Shore, R.M. Gray, Minimum cross-entropy pattern classification and cluster analysis, Trans. Pattern Anal. Mach. Intell. 4 (1) (1982) 11–17.
- [3] Chihau Chen, Statistical pattern recognition, Hayden Book Co., Rochelle Park, NJ, 1973.
- [4] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, P. Yanker, Query by image and video content: the qbic, Computer 28 (9) (1995) 23–32.
- [5] G. Pass, R. Zabih, J. Miller, Comparing images using color coherence, in: ACM International Multimedia Conference, ACM, New York, 1996, pp. 65–73.
- [6] H.S. Sawhney, J.L. Hafner, Efficient color histogram indexing, in: International Conference on Image Processing, Vol. 1, IEEE, New York, 1994, pp. 66–70.
- [7] M.J. Swain, D.H. Ballard, Color indexing, Int. J. Comput. Vision 7 (1) (1991) 11–32.
- [8] R.O. Duda, P.E. Hart, Pattern Classification and Scene Analysis, 1st Edition, Wiley, New York, 1973.
- [9] T. Kailath, The divergence and bhattacharyya distance measures in signal selection, IEEE Trans. Commun. Technol. COM-15 (1) (1967) 52–60.
- [10] K. Matusita, Decision rules, based on the distance, for problems of fit, two samples, and estimation, Ann. Math. Statist. 26 (1955) 631–640.

- [11] S. Kullback, R.A. Leibler, On information and sufficiency, Ann. Math. Statist. 22 (1951) 79–86.
- [12] C.E. Shannon, A mathematical theory of communication, Bell System Tech. J. 27 (1948) 379–423, 623–656.
- [13] G.T. Toussaint, Bibliography on estimation of misclassification, IEEE Trans. Inform. Theory 20 (4) (1974) 472–479
- [14] C.T. Yossi Rubner, L.J. Guibas, A metric for distributions with applications to image database, in: International Conference on Computer Vision, IEEE, New York, 1998, pp. 59-66.
- [15] F.S. Hillier, G.J. Lieberman, Introduction to Mathematical Programming, 2nd Edition, McGraw-Hill, New York, 1995.
- [16] Sung-Hyuk Cha, Fast image template and dictionary matching algorithms, in: Proceedings of ACCV '98 Lecture notes in Computer Science—Computer Vision, Vol. 1351, Springer, Berlin, January 1998, pp. 370–377.
- [17] Sung-Hyuk Cha, Efficient algorithms for image template and dictionary matching, J. Math. Imaging Vision 12 (1) (2000) 81–90.

- [18] Sung-Hyuk Cha, S.N. Srihari, Distance between histograms of angular measurements and its application to handwritten character similarity, in: Proceedings of 15th ICPR, IEEE CS Press, New York, 2000, pp. 21–24.
- [19] R.C. Gonzalez, R.E. Woods, Digital Image Processing, 3rd Edition, Addison-Wesley, 1992.
- [20] Sung-Hyuk Cha, S.N. Srihari, Writer identification: statistical analysis and dichotomizer, in: Proceedings of SS&SPR 2000, Lecture notes in Computer Science— Advances in Pattern Recognition, Vol. 1876, Springer, Berlin, September 2000, pp. 123–132.
- [21] Sung-Hyuk Cha, S.N. Srihari, Multiple feature integration for writer verification, in: Proceedings of Seventh IWFHR 2000, September 2000, pp. 333–342.
- [22] S.P. Michael Verman, A. Rosenfeld, A distance metric for multidimensional histograms, Comput. Vision, Graphics, Image Process. 32 (3) (1985) 328–336.

About the Author—SUNG-HYUK CHA received his B.S. and M.S. in Computer Science from Rutgers, the State University of New Jersey in 1994 and 1996, respectively, and Ph.D. in Computer Science & Engineering from the State University of New York (SUNY) at Buffalo in 2001. He is presently an Assistant Professor of Computer Science at Pace University in Westchester. From 1996 to 1998, he was working in the area of medical information and imaging systems at Information Technology R& D Center, Samsung SDS. During his Ph.D. years from 1998 to 2001, he was affiliated with the Center of Excellence for Document Analysis and Recognition. His main interests include *computer vision*, *data mining*, *pattern matching* & *recognition*.

About the Author—SARGUR N. SRIHARI received a B.E. in Electrical Communication Engineering from the Indian Institute of Science, Bangalore, in 1970 and a Ph.D. in Computer and Information Science from The Ohio State University in 1976. He is presently a University Distinguished Professor of the State University of New York (SUNY) at Buffalo. He was acting Chairman of the Computer Science Department from 1987 to 1988. He is presently a faculty member in the Computer Science and Engineering Department. He is the founding director of CEDAR, the Center of Excellence for Document Analysis and Recognition, located at SUNY Buffalo. The Center's work is in developing computer systems for the reading of text (handwritten and machine-printed) and document understanding. The Center has had several hundred individuals participate in its programs since its founding. The work of CEDAR has led to systems deployed for the United States Postal Service, Australia Post, and the Internal Revenue Service. Dr. Srihari is a Fellow of the IEEE and a Fellow of the International Association of Pattern Recognition (IAPR). He recently received the Distinguished Alumnus Award from The Ohio State University. He is an editor-in-chief of the International Journal of Document Analysis and Recognition and was/is General Chairman of several international conferences: the Third International Workshop on the Frontiers of Handwriting Recognition, Buffalo, NY, in 1993, the Fifth International Conference on Document Analysis and Recognition in Bangalore, India, in 1999, and the Eighth International Workshop on the Frontiers of Handwriting Recognition, at Niagara-on-the-Lake, Canada, in 2002.