A Normalized Levenshtein Distance Metric

Li Yujian and Liu Bo

Abstract—Although a number of normalized edit distances presented so far may offer good performance in some applications, none of them can be regarded as a genuine metric between strings because they do not satisfy the triangle inequality. Given two strings X and Y over a finite alphabet, this paper defines a new normalized edit distance between X and Y as a simple function of their lengths ($|X|$ and $|Y|$) and the Generalized Levenshtein Distance (GLD) between them. The new distance can be easily computed through GLD with a complexity of $O(|X| \cdot |Y|)$ and it is a metric valued in [0, 1] under the condition that the weight function is a metric over the set of elementary edit operations with all costs of insertions/deletions having the same weight. Experiments using the AESA algorithm in handwritten digit recognition show that the new distance can generally provide similar results to some other normalized edit distances and may perform slightly better if the triangle inequality is violated in a particular data set.

Index Terms—Sequence comparison, Levenshtein distance, normalized edit distance, metric, AESA.

1 Introduction

Quantifying the similarity between strings is an important scientific problem that has attracted much interest because key information can be expressed by symbolic sequences in many applications such as text retrieval, signal processing, and computational biology. Though a number of pertinent distance measures and their applications have been proposed and discussed [1], [2], the Generalized Levenshtein Distance (GLD) is the most promising one to compare strings by various edit operations, usually including the deletion, insertion, and substitution of individual symbols [3, pp. 37-39]. This measure is often called the "edit distance" and can be defined as the minimum cost of transforming one string into another through a sequence of weighted edit operations. The GLD can be computed by algorithms presented in [4], [5], [6], [7] and has been applied to error correction, pattern recognition, etc., [1], [2], [8], [9]. However, GLD is not suitable for certain applications such as recognizing noisy subsequences [10] and skeletal images [11] since it lacks an appropriate normalization with respect to the lengths of the compared strings. It should be clear that two errors in a comparison of short strings are more critical than in a comparison of long strings. Therefore, it is necessary to normalize the GLD in some circumstances.

So far as we are aware, there are two well-known normalizing approaches for the weighted edit distance GLD, one based on the editing path lengths and the other on the string lengths [12]. Although both of them offer somewhat better performance over the GLD in a few practical situations, from a theoretical standpoint neither of them can fulfill the triangle inequality. On the other hand, normalized metrics for symmetric set difference and Euclidian distance do not apply to edit distance [13] nor do those metrics based on Lempel-Ziv complexity [14]. Until now, defining a normalized edit distance that can be regarded as a genuine metric between two strings has remained an unsolved problem. This communication presents a solution for defining such a metric as a simple function of the string lengths and the GLD.

- *L. Yujian is with the College of Computer Science and Technology, Beijing University of Technology, Pingleyuan 100, Chaoyang District, Beijing 100022, P.R. China and the Beijing Municipal Key Laboratory of Multimedia and Intelligent Software Technology. E-mail: liyujian@bjut.edu.cn.*
- *L. Bo is with the College of Computer Science and Technology, Beijing University of Technology, Pingleyuan 100, Chaoyang District, Beijing 100022, P.R. China. E-mail: liubo@emails.bjut.edu.cn.*

Manuscript received 27 Apr. 2006; revised 1 Aug. 2006; accepted 20 Nov. 2006; published online 18 Jan. 2007.
Recommended for acceptance by D. Lopresti.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-0327-0406.
Digital Object Identifier no. 10.1109/TPAMI.2007.1070.

2 Generalized Levenshtein Distance

In terms of notation, Σ is the alphabet and Σ^* is the set of strings over Σ. $\lambda \notin \Sigma$ is the null string. A string $X \in \Sigma^*$ is denoted as $X = \mathrm{x}_1\mathrm{x}_2 \ldots \mathrm{x}_n$, where x_i is the ith symbol of X. $X_{i...j}$ is referred to as the substring of X including the symbols from x_i to x_j, $1 \le i \le j \le n$, its length is defined as $|X_{i...j}| = j - i + 1$, and it is the null string $\lambda(|\lambda| = 0)$ if $i > j$. An elementary edit operation is a pair $(a, b) \neq (\lambda, \lambda)$, often written as $a \to b$, where both a and b are strings of lengths 0 or 1. The forms $\lambda \to a$, $a \to b$, and $b \to \lambda$, respectively, represent insertions, substitutions, and deletions that are the three types of elementary edit operations. $T_{X,Y} = T_1T_2 \ldots T_l$ is used to denote an edit transformation of X into Y that is a sequence of elementary edit operations transforming X into Y. If a weight function γ assigns to $a \to b$ a nonnegative real number $\gamma(a \to b)$, the weight of an edit transformation $T_{X,Y}$ can be computed by $\gamma(T_{X,Y}) = \sum_{i=1}^{l} \gamma(T_i)$.

Given $X, Y \in \Sigma^*$, the Generalized Levenshtein Distance (GLD) is then defined as

$$\mathrm{GLD}(X, Y) = \min\{\gamma(T_{X,Y})\}. \tag{1}$$

It has been shown that GLD is a metric over Σ^* if the following conditions are satisfied:

$$\forall a, b \in \Sigma \cup \{\lambda\},\ \gamma(a \to a) = 0,$$
$$\gamma(a \to b) > 0 \text{ if } a \neq b, \text{ and } \gamma(a \to b) = \gamma(b \to a).$$

If γ is a metric over the set of elementary edit operations, it was also seen in [11] that

$$\mathrm{GLD}(X, Y) = \min\{W(P_{X,Y})\}, \tag{2}$$

where $P_{X,Y}$ is an editing path between X and Y, $W(P_{X,Y}) = \sum_{k=1}^{L(P_{X,Y})} \gamma(X_{i_{k-1}+1...i_k} \to Y_{j_{k-1}+1...j_k})$ is the weight of $P_{X,Y}$. In fact, $P_{X,Y}$ is a sequence of points or ordered pairs of integers (i_k, j_k), $0 \le k \le L(P_{X,Y}) = l$ satisfying the following:

1. $0 \le i_k \le |X|; 0 \le j_k \le |Y|; (i_0, j_0) = (0, 0); (i_l, j_l) = (|X|, |Y|)$,
2. $\forall k \ge 1,\ 0 \le i_k - i_{k-1} \le 1;\ 0 \le j_k - j_{k-1} \le 1$, and
3. $i_k - i_{k-1} + j_k - j_{k-1} \ge 1$.

3 Normalized GLD

As mentioned above, there are two normalization techniques for GLD, which are denoted here by NED_1 and NED_2, respectively, and defined as:

$$\mathrm{NED}_1(X, Y) = \min\left\{\frac{W(P_{X,Y})}{L(P_{X,Y})}\right\}, \quad \mathrm{NED}_2(X, Y) = \min\left\{\frac{W(P_{X,Y}))}{|X| + |Y|}\right\},$$

where $L(P_{X,Y})$ is the number of elementary edit operations described by $P_{X,Y}$ [11], [15], [16], [17].

Although GLD is a metric over Σ^*, it has been demonstrated that neither NED_1 nor NED_2 can be regarded as a genuine metric [11], [12], [18]. How to define a normalized edit metric for two strings, up to now an unsolved problem, can be addressed as follows:

Definition 1. $\alpha = \max\{\gamma(a \to \lambda), \gamma(\lambda \to b), a, b \in \Sigma\}$.

Definition 2. *Given $X, Y \in \Sigma^*$, the Generalized Levenshtein Similarity (GLS) between X and Y is defined as*

$$\mathrm{GLS}(X,Y) = \frac{\alpha \cdot (|X| + |Y|) - \mathrm{GLD}(X,Y)}{2}.$$

It is not difficult to see that Theorem 1 holds for GLS.

Theorem 1. *If $\forall a \in \Sigma$, $\gamma(\lambda \to a) = \gamma(a \to \lambda) = \alpha$, and γ is a metric over the set of elementary edit operations, then $\forall X, Y, Z \in \Sigma^*$,*

1. $\mathrm{GLS}(X,X) = \alpha|X|$,
2. $\mathrm{GLS}(X,Y) = \mathrm{GLS}(Y,X)$,
3. $0 \le \mathrm{GLS}(X,Y) \le \min\{\mathrm{GLS}(X,X), \mathrm{GLS}(Y,Y)\} = \alpha \cdot \min\{|X|, |Y|\}$, *and*
4. $\mathrm{GLS}(Y,Y) + \mathrm{GLS}(X,Z) \ge \mathrm{GLS}(X,Y) + \mathrm{GLS}(Y,Z)$.

Proof.

1. Because $\mathrm{GLD}(X,X) = 0$, we directly get
$$\mathrm{GLS}(X,X) = \alpha|X|.$$
2. From $\mathrm{GLD}(X,Y) = \mathrm{GLD}(Y,X)$, we easily obtain $\mathrm{GLS}(X,Y) = \mathrm{GLS}(Y,X)$.
3. According to Lemma 4.1 in [11], we have $\max\{|X|, |Y|\} \le L(P_{X,Y}) \le |X| + |Y|$ so that
$$\mathrm{GLD}(X,Y) = \min\{W(P_{X,Y})\} = \min\left\{\sum_{k=1}^{L(P_{X,Y})} \gamma(X_{i_{k-1}+1\ldots i_k} \to Y_{j_{k-1}+1\ldots j_k})\right\} \le \min\left\{\sum_{k=1}^{L(P'_{X,Y})} \gamma(X_{i'_{k-1}+1\ldots i'_k} \to Y_{j'_{k-1}+1\ldots j'_k})\right\} = \alpha \cdot (|X| + |Y|),$$
where $\forall a, b \in \Sigma, \gamma(a,b) \le \gamma(a,\lambda) + \gamma(\lambda,b) = 2\alpha$ and $P'_{X,Y} = (i'_0, j'_0), (i'_1, j'_1), \ldots, (i'_{|X|+|Y|}, j'_{|X|+|Y|})$ is constructed from $P_{X,Y} = (i_0, j_0), (i_1, j_1), \ldots, (i_{L(P_{X,Y})}, j_{L(P_{X,Y})})$ by replacing (i_k, j_k) with "$(i_k, j_{k-1}), (i_k, j_k)$" if $i_k = i_{k-1} + 1$ and $j_k = j_{k-1} + 1 (k > 0)$,
$$\begin{aligned}\mathrm{GLD}(X,Y) &= \min\{W(P_{X,Y})\} \\ &= \min\left\{\sum_{k=1}^{L(P_{X,Y})} \gamma(X_{i_{k-1}+1\ldots i_k} \to Y_{j_{k-1}+1\ldots j_k})\right\} \\ &\ge \alpha \cdot (\|X| - |Y\|).\end{aligned}$$
(The number of insertions/deletions in $P_{X,Y}$ is at least $\|X| - |Y\|$).
Therefore,
$$\mathrm{GLS}(X,Y) = \frac{\alpha \cdot (|X| + |Y|) - \mathrm{GLD}(X,Y)}{2} \ge 0,$$
$$\begin{aligned}\mathrm{GLS}(X,Y) &= \frac{\alpha \cdot (|X| + |Y|) - \mathrm{GLD}(X,Y)}{2} \\ &\le \frac{\alpha \cdot (|X| + |Y|) - \alpha \cdot (\|X| - |Y\|)}{2} \\ &= \alpha \cdot \min\{|X|, |Y|\}.\end{aligned}$$
4. From the triangle inequality,
$$\mathrm{GLD}(X,Y) + \mathrm{GLD}(Y,Z) \ge \mathrm{GLD}(X,Z),$$
$$[\alpha(|X| + |Y|) - 2 \cdot \mathrm{GLS}(X,Y)] + [\alpha(|Y| + |Z|) - 2 \cdot \mathrm{GLS}(Y,Z)] \ge [\alpha(|X| + |Z|) - 2 \cdot \mathrm{GLS}(X,Z)],$$
$$\alpha|Y| + \mathrm{GLS}(X,Z) \ge \mathrm{GLS}(X,Y) + \mathrm{GLS}(Y,Z),$$
$$\mathrm{GLS}(Y,Y) + \mathrm{GLS}(X,Z) \ge \mathrm{GLS}(X,Y) + \mathrm{GLS}(Y,Z).$$

□

Definition 3. *Given $X, Y \in \Sigma^*$, the normalized GLD is defined as*

$$\begin{aligned}d_{N-GLD}(X,Y) &= \frac{2 \cdot \mathrm{GLD}(X,Y)}{\alpha \cdot (|X| + |Y|) + \mathrm{GLD}(X,Y)} \\ &= \frac{\mathrm{GLS}(X,X) + \mathrm{GLS}(Y,Y) - 2 \cdot \mathrm{GLS}(X,Y)}{\mathrm{GLS}(X,X) + \mathrm{GLS}(Y,Y) - \mathrm{GLS}(X,Y)},\end{aligned}$$

where $d_{N-GLD}(\lambda, \lambda) = 0$.

Obviously, d_{N-GLD} can be easily computed through GLD with a complexity of $O(|X| \cdot |Y|)$. Although d_{N-GLD} is not always shown to be a metric over Σ^* for an arbitrary weight function, Theorem 2, inspired by the work reported in [19], can be shown to hold.

Theorem 2. *If $\forall a \in \Sigma, \gamma(\lambda \to a) = \gamma(a \to \lambda) = \alpha$ and γ is a metric over the set of elementary edit operations, then d_{N-GLD} is a metric over Σ^* whose value is in [0,1].*

Proof. According to Theorem 1, the value of $d_{N-GLD}(X,Y)$ is obviously in [0, 1] and $d_{N-GLD}(X,Y)$ is zero if $X = Y$ and is a positive real number if $X \ne Y$, such that $d_{N-GLD}(X,Y) = d_{N-GLD}(Y,X)$. It has to be further shown that the triangle inequality holds for d_{N-GLD}, namely, $\forall X, Y, Z \in \Sigma^*$,

$$d_{N-GLD}(X,Y) + d_{N-GLD}(Y,Z) \ge d_{N-GLD}(X,Z). \tag{3}$$

Let $S(X,Y) = 1 - d_{N-GLD}(X,Y) = \frac{\mathrm{GLS}(X,Y)}{\mathrm{GLS}(X,X) + \mathrm{GLS}(Y,Y) - \mathrm{GLS}(X,Y)}$ for any $X, Y \in \Sigma^*$. Then, (3) can be rewritten as

$$1 + S(X,Z) \ge S(X,Y) + S(Y,Z). \tag{4}$$

If $S(X,Z) \ge S(X,Y)$ or $S(X,Z) \ge S(Y,Z)$, it can be directly obtained that

$$d_{N-GLD}(X,Z) \le d_{N-GLD}(X,Y) \text{ or } d_{N-GLD}(X,Z) \le d_{N-GLD}(Y,Z).$$

Accordingly, (3) is immediately satisfied. Thus, it is necessary to prove (3) only for the case in which $S(X,Z) < S(X,Y)$ and $S(X,Z) < S(Y,Z)$.

Using the relation between $S(X,Y)$ and $\mathrm{GLS}(X,Y)$, it is easy to get the following equations:

$$\mathrm{GLS}(X,Z) = \frac{S(X,Z)}{1 + S(X,Z)}(\mathrm{GLS}(X,X) + \mathrm{GLS}(Z,Z)), \tag{5}$$

$$\mathrm{GLS}(X,Y) = \frac{S(X,Y)}{1 + S(X,Y)}(\mathrm{GLS}(X,X) + \mathrm{GLS}(Y,Y)), \tag{6}$$

$$\mathrm{GLS}(Y,Z) = \frac{S(Y,Z)}{1 + S(Y,Z)}(\mathrm{GLS}(Y,Y) + \mathrm{GLS}(Z,Z)). \tag{7}$$

Applying (5), (6), and (7) to Step 4 in Theorem 1, (8) and (9) can be derived:

$$\begin{aligned}\mathrm{GLS}(Y,Y) &+ \frac{S(X,Z)}{1 + S(X,Z)}(\mathrm{GLS}(X,X) + \mathrm{GLS}(Z,Z)) \\ &\ge \frac{S(X,Y)}{1 + S(X,Y)}(\mathrm{GLS}(X,X) + \mathrm{GLS}(Y,Y)) \\ &+ \frac{S(Y,Z)}{1 + S(Y,Z)}(\mathrm{GLS}(Y,Y) + \mathrm{GLS}(Z,Z)),\end{aligned} \tag{8}$$

$$\begin{aligned}&\left[1 - \frac{S(X,Y)}{1 + S(X,Y)} - \frac{S(Y,Z)}{1 + S(Y,Z)}\right] \cdot \mathrm{GLS}(Y,Y) \\ &\ge \left[\frac{S(X,Y)}{1 + S(X,Y)} - \frac{S(X,Z)}{1 + S(X,Z)}\right] \cdot \mathrm{GLS}(X,X) \\ &+ \left[\frac{S(Y,Z)}{1 + S(Y,Z)} - \frac{S(X,Z)}{1 + S(X,Z)}\right] \cdot \mathrm{GLS}(Z,Z).\end{aligned} \tag{9}$$

From Theorem 1, $\mathrm{GLS}(X,X) \geq \mathrm{GLS}(X,Y)$ and $\mathrm{GLS}(Z,Z) \geq \mathrm{GLS}(Y,Z)$, so it is true that

$$\mathrm{GLS}(X,X) \geq S(X,Y)\cdot \mathrm{GLS}(Y,Y) \text{ and}$$
$$\mathrm{GLS}(Z,Z) \geq S(Y,Z)\cdot \mathrm{GLS}(Y,Y).$$

Since $S(X,Z) < S(X,Y)$ and $S(X,Z) < S(Y,Z)$ imply that

$$\frac{S(X,Y)}{1+S(X,Y)} - \frac{S(X,Z)}{1+S(X,Z)} > 0 \quad \text{and}$$
$$\frac{S(Y,Z)}{1+S(Y,Z)} - \frac{S(X,Z)}{1+S(X,Z)} > 0,$$

it is valid to write

$$\left[\frac{S(X,Y)}{1+S(X,Y)} - \frac{S(X,Z)}{1+S(X,Z)}\right]\cdot \mathrm{GLS}(X,X) \geq \left[\frac{S(X,Y)}{1+S(X,Y)} - \frac{S(X,Z)}{1+S(X,Z)}\right]\cdot S(X,Y)\cdot \mathrm{GLS}(Y,Y), \tag{10}$$

$$\left[\frac{S(Y,Z)}{1+S(Y,Z)} - \frac{S(X,Z)}{1+S(X,Z)}\right]\cdot \mathrm{GLS}(Z,Z) \geq \left[\frac{S(Y,Z)}{1+S(Y,Z)} - \frac{S(X,Z)}{1+S(X,Z)}\right]\cdot S(Y,Z)\cdot \mathrm{GLS}(Y,Y). \tag{11}$$

Using (9), (10), and (11), it can be further obtained that

$$\begin{aligned}&\left[1 - \frac{S(X,Y)}{1+S(X,Y)} - \frac{S(Y,Z)}{1+S(Y,Z)}\right]\cdot \mathrm{GLS}(Y,Y)\\ &\geq \left[\frac{S(X,Y)}{1+S(X,Y)} - \frac{S(X,Z)}{1+S(X,Z)}\right]\cdot S(X,Y)\cdot \mathrm{GLS}(Y,Y)\\ &+ \left[\frac{S(Y,Z)}{1+S(Y,Z)} - \frac{S(X,Z)}{1+S(X,Z)}\right]\cdot S(Y,Z)\cdot \mathrm{GLS}(Y,Y).\end{aligned} \tag{12}$$

Supposing that $\mathrm{GLS}(Y,Y) > 0$ (otherwise, (3) is clearly satisfied), (12) can be rewritten as (13) or (14):

$$1 \geq \frac{S(X,Y)[1+S(X,Y)]}{1+S(X,Y)} + \frac{S(Y,Z)[1+S(Y,Z)]}{1+S(Y,Z)} - S(X,Z)\frac{S(X,Y)+S(Y,Z)}{1+S(X,Z)}, \tag{13}$$

$$1 + S(X,Z)\frac{S(X,Y)+S(Y,Z)}{1+S(X,Z)} \geq S(X,Y) + S(Y,Z) \tag{14}$$

Because (14) yields (4), this means that (3) holds. □

4 Experimental Results

In order to demonstrate the benefits of the normalized GLD (NGLD) in a practical application, it has been used to solve the problem of handwritten digit recognition by the Approximating and Eliminating Search Algorithm (AESA) [20], [21], which is a fast algorithm for Brute-force Nearest Neighbor Search (BNNS). Because AESA is dependent on fulfillment of the triangle inequality, which NGLD satisfies if the operation weights are selected appropriately, NGLD-based AESA (NGLD-AESA) will always produce optimal results, but NED_1 or NED_2-based AESA (NED_1 or NED_2-AESA) will sometimes yield suboptimal results. Here, NGLD-AESA will be compared with NED_1-AESA, NED_2-AESA, and GLD-AESA (GLD-based AESA) in two handwritten digit recognition experiments.

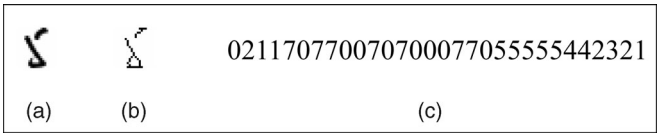

Fig. 1. The final string generated from a thinned binary image of the digit 5. (a) Gray-scale image. (b) Thinned binary image. (c) Final string or chain code.

The first experiment uses 4,000 strings (P1, 400 per digit) as the training set and 500 strings (T1, 50 per digit) as the test set. These strings represent the contours of the isolated sample digits, which are randomly selected from the MNIST database (containing 60,000 training digits from approximately 250 writers) constructed from NIST's Special Database 3 and Special Database 7 [27]. Each string is actually a chain code which was generated from the thinned binary version of its corresponding gray-scale digit image (using the Rosenfeld thinning algorithm [22]). The algorithm to compute a chain code begins by searching the first black pixel in raster order. Starting from this pixel, it outputs a string of numbers from 0 to 7 representing the eight possible directions from each pixel to its successive pixel along the contour of a thinned binary digit image until reaching an end point (a black pixel followed by no successive ones) or the first pixel again. In the computation, each branch or isolated part will be removed from the thinned binary image after its chain code is obtained, then the chain code of next branch or isolated part is repeatedly generated until no black pixels remain. If there are several branches or isolated parts in a digit image, all of their chain codes will be concatenated to produce the final single string. One example of final string generated from a thinned binary image of the digit 5 is illustrated in Fig. 1.

In each recognition experiment, the weight function γ is set with $\gamma(a,a) = 0, \gamma(a,b) = 1$, and $\gamma(a,\lambda) = \gamma(\lambda,b) = 1$ for any $a, b \in \Sigma$. The NGLD defined by such γ obviously satisfies the triangle inequality according to Theorem 2 and so does the corresponding GLD. However, the NED_1 and the NED_2 may violate the triangle inequality in a particular data set. In fact, for the NED_1 on the training set P1, there are only 44 strings and 82 triplets (X,Y,Z) possessing a negative looseness, which is defined as

$$h_1(X,Y,Z) = \mathrm{NED}_1(X,Y) + \mathrm{NED}_1(Y,Z) - \mathrm{NED}_1(X,Z).$$

But, for the NED_2, there are in total 1,600 strings and 18,056 triplets (X,Y,Z) possessing a similar negative looseness, namely,

$$h_2(X,Y,Z) = \mathrm{NED}_2(X,Y) + \mathrm{NED}_2(Y,Z) - \mathrm{NED}_2(X,Z).$$

For simplicity and clarity, the set of the 1,600 strings is also called the nontriangular set and the histogram of the 18,056 triplets (X,Y,Z) is illustrated in Fig. 2, where the horizontal axis is the normalized looseness, meaning the absolute value of $h_2(X,Y,Z)$ divided by the average NED_2 distance between the strings in the training set P1 and the vertical axis is the relative frequency, meaning the number of triplets (X,Y,Z) with a certain looseness divided by the total number of all negative-looseness triplets. It is shown in Fig. 2 that only a very small fraction of negative-looseness triplets violate the triangle inequality severely, while most of them do so only slightly.

The second experiment uses 1,202 strings (P2) as the training set and 398 strings (T2) as the test set, where P2 and T2 are obtained by directly partitioning the nontriangular set into two parts according to a proportion of about 3:1.

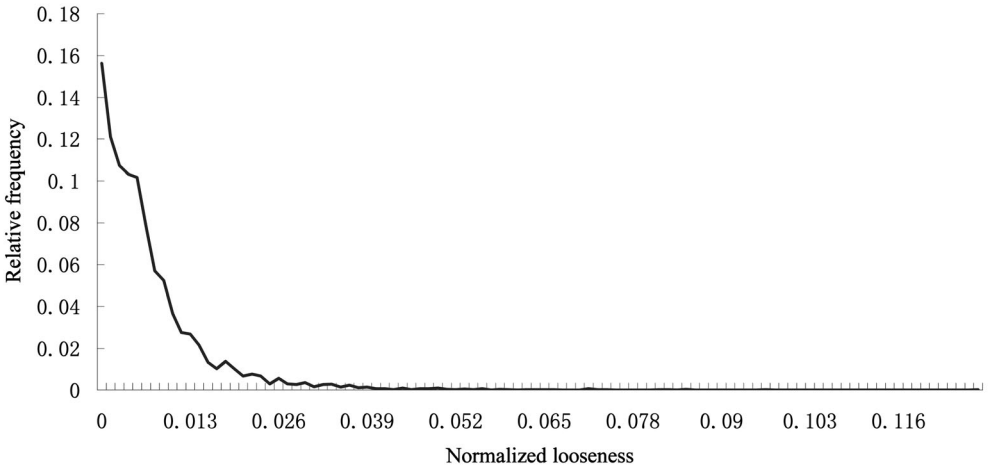

Fig. 2. The histogram of negative $h_2(X, Y, Z)$ on the training set P1.

TABLE 1
Results in Two Experiments Using AESA (BNNS) with Four Different Edit Distances

Accuracy	NGLD	NED_1	NED_2	GLD
P1, T1	80.44% (80.64%)	81.04% (81.04%)	80.44% (80.64%)	79.04% (77.84%)
P2, T2	68.34% (68.84%)	67.84% (69.35%)	68.09% (68.84%)	67.09% (65.83%)

The results in the two experiments are summarized in Table 1. In the first experiment, it is worth noting that the NGLD-AESA accuracy (80.44 percent) does not quite equal the NGLD-BNNS (NGLD-based BNNS) accuracy (80.64 percent) and similar relations often hold for NED_1, NED_2, and GLD. This is because the AESA algorithm uses "Approximation" as a heuristic strategy to find a nearest neighbor which may not be the first one determined sequentially by the BNNS (when there are two or more nearest neighbors), even in cases satisfying the triangle inequality. It can also be seen that the accuracy of NGLD-AESA is similar to that of NED_1 and NED_2-AESA, but higher than that of GLD-AESA. Furthermore, it is easy to notice that the NGLD-AESA accuracy (80.44 percent) is slightly lower than the NED_1-AESA accuracy (81.04 percent) in the P1-T1 experiment, as may be partially explained by the fact that some NGLD nearest neighbors lead to incorrect recognition but are eliminated by NED_1-AESA. However, the runtime of NGLD-AESA (644s on a DELL GX280 computer) is much less than that of NED_1-AESA (8,159s or 1,231s, respectively, when using DPNED [11] or FPNED [15] on the same computer) in this experiment.

Additionally, the NGLD-AESA accuracy (68.34 percent) is slightly higher than the NED_1-AESA accuracy (67.84 percent) and the NED_2-AESA accuracy (68.09 percent) in the second experiment, which implies that the NGLD may perform slightly better than other normalized edit distances if the triangle inequality is violated to a certain degree in a particular data set. The NGLD-AESA performs only slightly better because there are just a few correct nearest neighbors in those samples which are wrongly eliminated by the NED_1 or NED_2-AESA for violating the triangle inequality.

Finally, it must be noted that the accuracies we report in these two experiments are much lower than those reported elsewhere in the literature. which are close to 95 percent [21], but correspond to smaller training and testing sets (50 or so writers). Because data sets built from about 250 writers are used to generate chain codes here, lower accuracies are to be expected. Since it is our intention to show the relative advantage of NGLD over other normalized distances, absolute accuracies are not so important for our purposes. The P2-T2 accuracies are much lower than the P1-T1 accuracies mainly because the training set P2 is much smaller than P1 and the test set T2 contains relatively more strings having incorrect nearest neighbors in P2.

5 Conclusions

In this paper, a new normalized edit distance has been presented as a simple function of the string lengths and the Generalized Levenshtein Distance. The main contribution of the paper is to prove that the new distance is a metric valued in [0, 1] under common conditions and demonstrate, by using AESA in hand-written digit recognition, that it can generally achieve similar accuracies to two other normalized edit distances, yielding slightly better results if the triangle inequality is violated to a certain degree. Since no other normalized edit distance has been shown to be a metric, this work is significant in that regard. As future work, we plan to identify situations where the new distance is appropriate and study its performance in applications such as phylogenetic tree construction, where all three basic properties of a distance metric between two sequences are usually required at the same time [14]. Moreover, we will also consider the problems of how to use the presented techniques to normalize a distance between histograms [23], [24] or a local alignment score between strings [25], [26] in a provably more rigorous way.

ACKNOWLEDGMENTS

This work was supported by the Natural Science Foundation of Beijing (4052005), PHR(IHLB) and the Beijing Municipal Education Commission (Km200310005013). The authors would like to thank Associate Editor Dr. Daniel Lopresti for carefully checking a large number of grammatical and typographical errors and all of the anonymous reviewers, whose comments were of great value. Also, the authors are indebted to Professors Guo Jun and Zhang Honggang who works at the Pattern Recognition and Intelligent System Lab, Beijing University of Posts and Telecommunications.

REFERENCES

[1] K. Kukich, "Techniques for Automatically Correcting Words in Text," *ACM Computing Surveys,* vol. 24, pp. 377-439, 1992.
[2] G. Navarro, "A Guided Tour to Approximate String Matching," *ACM Computing Surveys,* vol. 33, no. 1, pp. 32-88, Mar. 2001.
[3] D. Sankoff and J.B. Kruskal, *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison.* Addison-Wesley, 1983.
[4] A. Levenshtein, "Binary Codes Capable of Correcting Deletions, Insertions and Reversals," *Soviet Physics Doklady,* vol. 10, no. 8, pp. 707-710, 1966.
[5] P.H. Sellers, "On the Theory and Computation of Evolutionary Distances," *SIAM J. Applied Math.,* vol. 26, no. 4, pp. 787-793, 1974.
[6] R.A. Wagner and M.J. Fischer, "The String-to-String Correction Problem," *J. ACM,* vol. 21, no. 1, pp. 168-173, Jan. 1974.
[7] W.J. Masek and M.S. Patterson, "A Faster Algorithm Computing String Edit Distances," *J. Computer Systems Science,* vol. 20, pp. 18-31, Feb. 1980.
[8] J.L. Peterson, "Computer Programs for Detecting and Correcting Spelling Errors," *Comm. ACM,* vol. 23, pp. 676-687, 1980.
[9] R.L. Kashyap and B.J. Oommen, "The Noisy Substring Matching Problem," *IEEE Trans. Software Eng.,* vol. 9, pp. 365-370, 1983.
[10] B.J. Oommen, "Recognition of Noisy Subsequences Using Constrained Edit Distances," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 9, pp. 676-685, 1987 (corrections in vol. 10, pp. 983-984, 1988).
[11] A. Marzal and E. Vidal, "Computation of Normalized Edit Distance and Applications," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 15, no. 9, pp. 926-932, Sept. 1993.
[12] A. Weigel and F. Fein, "Normalizing the Weighted Edit Distance," *Proc. 12th IAPR Int'l Conf. Pattern Recognition,* vol. 2, Conf. B: Computer Vision and Image Processing, pp. 399-402, Oct. 1994.
[13] P.N. Yianilos, "Normalized Forms for Two Common Metrics," Report 91-082-9027-1, revision 7 July 2002, NEC Research Inst., 1991, http://www.pnylab.com/pny/.
[14] H.H. Otu1 and K. Sayood, "A New Sequence Distance Measure for Phylogenetic Tree Construction," *Bioinformatics,* vol. 19, no. 16, pp. 2122-2130, 2003.
[15] E. Vidal, A. Marzal, and P. Aibar, "Fast Computation of Normalized Edit Distances," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 17, no. 9, pp. 899-902, Sept. 1995.
[16] A.N. Arslan and O. Egecioglu, "An Efficient Uniform-Cost Normalized Edit Distance Algorithm," *Proc. 1999 and Int'l Workshop Groupware String Processing and Information Retrieval Symp.,* pp. 8-15, Sept. 1999.
[17] A.N. Arslan and O. Egecioglu, "Efficient Algorithms for Normalized Edit Distance," *J. Discrete Algorithms,* (special issue on matching patterns) vol. 1, no. 1, pp. 3-20, 2000.
[18] E. Vidal, F. Casacuberta, J.M. Benedi, M.J. Lloret, and H. Rulot, "On the Verification of Triangle Inequality by Dynamic Time-Warping Dissimilarity Measures," *Speech Comm.,* vol. 7, pp. 67-69, 1988.
[19] A.H. Lipkus, "A Proof of the Triangle Inequality for the Tanimoto Distance," *J. Math. Chemistry,* vol. 26, no. 1-3, pp. 263-265, 1999.
[20] E. Vidal, "New Formulation and Improvements of the Nearest-Neighbour Approximating and Eliminating Search Algorithm (AESA)," *Patter Recognition Letters,* vol. 15, no. 1, pp. 1-7, 1994.
[21] J.R. Rico-Juan and L. Mico, "Comparison of AESA and LAESA Search Algorithms Using String and Tree-Edit Distances," *Pattern Recognition Letters,* vol. 24, pp. 1417-1426, 2003.
[22] A. Rosenfeld, "A Characterization of Parallel Image Thinning Algorithms," *Information and Control,* vol. 29, no. 3, pp. 286-291, 1975.
[23] S.-H. Cha and S.N. Srihari, "On Measuring the Distance between Histograms," *Pattern Recognition,* vol. 35, no. 6, pp. 1355-1370, 2002.
[24] F. Serratosa and A. Sanfeliu, "Signatures versus Histograms: Definitions, Distances and Algorithms," *Pattern Recognition,* vol. 39, no. 5, pp. 921-934, 2006.
[25] A.N. Arslan, O. Egecioglu, and P.A. Pevzner, "A New Approach to Sequence Comparison: Normalized Sequence Alignment," *Bioinformatics,* vol. 17, no. 4, pp. 327-337, 2001.
[26] A.N. Arslan and O. Egecioglu, "Dynamic Programming Based Approximation Algorithms for Sequence Alignment with Constraints," *INFORMS J. Computing,* vol. 16, no. 4, pp. 441-458, 2004.
[27] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-Based Learning Applied to Document Recognition," *Proc. IEEE,* vol. 86, no. 11, pp. 2278-2324, 1998.