Effects of Time Normalization on the Accuracy of Dynamic Time Warping

Olaf Henniger　　Sascha Müller

***Abstract*— This paper revisits Dynamic Time Warping, a method for assessing the dissimilarity of time series. In particular, this paper provides theoretical and experimental evidence showing that uncritical normalizing the length of the time series to be compared has a detrimental effect on the recognition accuracy in application domains such as on-line signature recognition, where the length of compared time series matters for their classification as match or non-match.**

I. Introduction

Dynamic Time Warping (DTW) is an algorithm for aligning two time series (sequences of values at successive points in time) that are similar, but out of sync and generally not of exactly the same length, in such a way that the "distance" between the two time series is minimal. This minimal distance is used to characterize the dissimilarity of the two time series. If their distance is greater than a threshold value, the two time series are not regarded as similar; otherwise, they are recognized as similar. DTW has been used for many years in various application domains [1], including speech recognition (e.g. [2], [3]), "data mining" (e.g. [4]), and the recognition of on-line signatures, handwritten signatures captured by means of a graphic tablet and/or a special pen in the form of time series for the pen position (x and y coordinates) and possibly other values such as the pen-tip pressure (e.g. [5]). Section II of this paper briefly recapitulates the basic DTW algorithm and useful improvements reducing the runtime and memory requirements.

In on-line signature recognition, DTW may be applied to the entire signature or to signature segments [6]. DTW may be applied in combination with methods for the statistical, spatial, and spectral analysis of signatures [7] or as a rather self-sufficient method. The recently standardized signature data interchange formats [8] are suitable not only for encoding acquired on-line signature data that serve as a starting point for feature extraction, but also support encoding of time series to be compared directly by DTW algorithms. Using the improvements reviewed in Section II-D, the time and space complexity of DTW algorithms can be reduced significantly, making them efficient enough for application in devices with scarce resources, like mobile platforms [9] or even smart cards [10].

Even though in the different application domains essentially similar DTW algorithms are used, there are domain-specific differences: In [11], C.A. Ratanamahatana and E. Keogh argue for normalizing the length of the time series to the same length by interpolating and resampling before applying DTW in data mining domains such as handwritten-character recognition, video retrieval, or text mining. The reason why they recommend length normalization is that same-length time series in a database can be indexed very easily, to reduce the search space for fast one-to-many comparison, but time series of different lengths cannot be indexed so easily [4]. [11] states that

> "an extensive literature search through more than 500 papers dating back to the 1960's failed to produce any theoretical or empirical results to suggest that simply making the sequences to be of the same length has any detrimental effect."

The present paper shall fill this gap by providing experimental evidence showing that, in application domains such as on-line signature recognition, making the time series to be of the same length does have a detrimental effect on the recognition accuracy. Even without looking at the empirical results, a simple Gedanken-experiment in Section III may convince the reader of the detrimental effect on the recognition accuracy that normalizing the length of the time series may have in on-line signature recognition. Section IV presents empirical results obtained using a publicly available subset of the database [12], consisting of on-line signature samples of 100 persons. For signatures of different length, DTW has been carried out with both, the original lengths and normalized lengths, and the recognition accuracy in both cases has been compared.

II. Dynamic Time Warping

A. Optimization problem

Let $A = (a_i)_{i=1,\dots,I}$ and $B = (b_j)_{j=1,\dots,J}$ be two time series of length I and J that represent a biometric sample to be verified and a biometric reference, respectively. For on-line signatures, each a_i and b_j itself is a vector consisting of the x and y coordinates of the pen position at a sample point and possibly other values, like the pen-tip velocities in x and y direction or the pen-tip pressure. Let $\mathrm{distance}(a_i, b_j)$ be a point-to-point distance between a_i and b_j. The point-to-point distance between a_i and b_j can be defined in different ways, e.g. as the Euclidean distance or as the absolute value norm of the difference vector of a_i and b_j.

The goal of DTW is to determine a mapping between the time series A and B minimizing the sum of the point-to-point distances of corresponding sample points while observing certain side conditions. The side conditions are that the start

Olaf Henniger (henniger@sit.fraunhofer.de) is with the Fraunhofer Institute for Secure Information Technology, Darmstadt, Germany.

Sascha Müller (mueller@sec.informatik.tu-darmstadt.de) is with the Technische Universität Darmstadt, Germany.

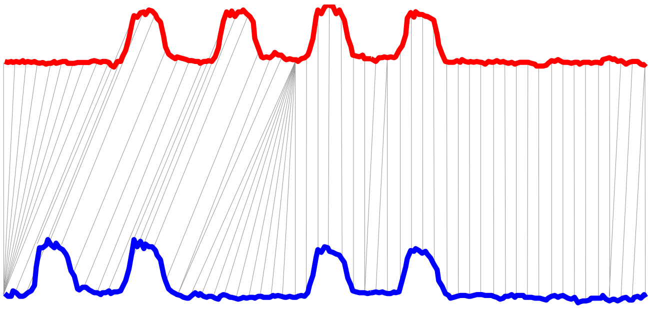


Fig. 1. “Warped” time series

and end points of the signatures to be compared must be mapped one on the other, that the temporal ordering of the sample points must be maintained, and that no sample point be omitted. Fig. 1 (from a presentation version of [4]) visualizes the idea. The time axes are in some way stretched in some places and squeezed in other places to map sample points in an optimal way such that the distance of the two time series is as small as possible.

More formally, let w be a warping function that assigns to each k, $1 \leq k \leq K$ with $\max(I, J) \leq K < I + J$, an index pair (i_k, j_k) such that $w(1) = (1, 1)$, $w(K) = (I, J)$, and $0 \leq i_k - i_{k-1} \leq 1$ and $0 \leq j_k - j_{k-1} \leq 1$ for $2 \leq k \leq K$. The set of index pairs $W = \{(1, 1), \ldots, (I, J)\}$ described by a warping function w can be considered as a distortion path through the plane spanned by the time axes of A and B (see Fig. 2). Let $\mathcal{W}$ be the set of all possible distortion paths for A and B. If the distortion path is the diagonal from $(1, 1)$ to (I, J), then the time distortion is linear. The farther off the diagonal it is, the more non-linear it is.

The accumulated distance $d_W(A, B)$ between A and B for a given distortion path W is the sum of the point-to-point distances $\mathrm{distance}(a_i, b_j)$ along this distortion path:

$$d_W(A, B) = \sum_{(i,j) \in W} \mathrm{distance}(a_i, b_j).$$

The goal is to determine a distortion path for which the accumulated distance between A and B is minimal:

$$d(A, B) = \min_{W \in \mathcal{W}} d_W(A, B).$$

This is the optimal distortion path. The minimal distance between A and B quantifies the dissimilarity of A and B. Usually, the optimal distortion path is acquired by dynamic programming (see Section II-C). Dynamic programming is a time-consuming procedure. The computing time is reduced by looking for the optimal distortion path only within a limited band (Sakoe/Chiba band [3], see Section II-D).

B. *Preprocessing of on-line signatures*

Care must be taken that values are in the same range when the point-to-point distances are determined. The goal of preprocessing is to suppress insignificant information that is expression of random variation without losing information about biometric characteristics of an individual. Preprocessing also helps to even out differences between data captured with different capture devices. It may include:

- Translation: The values of A and B should be linearly translated to be in the same domains.
- Rotation: If the direction of the signatures is different, one of them should be rotated to fit to the other one.
- Normalization and scaling: Not all signatures of a person have the same size. Their x and y coordinates should be scaled appropriately.[1]
- Removal of the linear component from the x-time regression line: The linear component of the x coordinates (due to writing from left to right) may be removed.
- Applying a low-pass filter: High frequency portions of the sample should be removed to suppress noise.

In the experiment described in Section IV, translation and scaling were applied. The rotation of the samples, which is a crucial preprocessing step in most instances, was omitted since the samples of the used test database are already aligned horizontally. As this step was omitted for both, DTW with the original lengths and with normalized lengths, this has no influence on the relative performance results.

C. *Dynamic programming*

The number of possible monotonically increasing paths from $(1, 1)$ to (I, J) is very large, so testing the length of all possible distortion paths to find the shortest one is too costly. Finding the shortest distortion path is achieved by dynamic programming algorithms.

Let $D = (d_{i,j})_{i=1,\ldots,I; j=1,\ldots,J}$ be a matrix whose entries $d_{i,j}$ are the accumulated point-to-point distances of A and B along optimal distortion paths to (i, j). The accumulated distance $d_{i,j}$ represents the length of the optimal distortion path from $(1, 1)$ to (i, j). The idea is to recurrently calculate $d_{i',j'}$ from already known values $d_{i,j}$ with $i \leq i'$ and $j \leq j'$ until $d_{I,J}$ is reached. $d_{I,J}$ represents the minimal accumulated distance $d(A, B)$ between A and B. Fig. 2 visualizes the idea: The shortest distortion path from $(1, 1)$ to a certain point (i, j) (e.g. $P = (5, 7)$) can be constructed by connecting (i, j) to that predecessor whose accumulated distance from $(1, 1)$ is shortest. Possible predecessors are $(i - 1, j)$, $(i - 1, j - 1)$, and $(i, j - 1)$.

The accumulated distance from $(1, 1)$ to (i, j) is obtained by adding the point-to-point $\mathrm{distance}(a_i, b_j)$ to the accumulated distance from $(1, 1)$ to that predecessor with the shortest accumulated distance from $(1, 1)$. It is sufficient to save an accumulated distance value for each matrix entry since for biometric verification the result of interest is the accumulated distance for the optimal distortion path and not the actual distortion path. The algorithm can be expressed in pseudo-code as follows:

```
FOR $i = 1$ TO $I$ $d_{i,0} = \infty$ NEXT $i$
FOR $j = 1$ TO $J$ $d_{0,j} = \infty$ NEXT $j$
$d_{0,0} = 0$
```

[1] Note that even though also not all signatures of a person take the same amount of time, the length in time should not be normalized because it helps to distinguish between originals and forgeries as will be seen later.

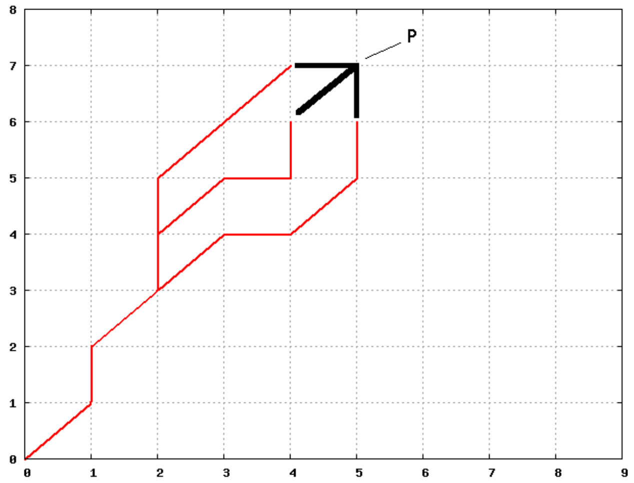


Fig. 2. Possible distortion paths

```
FOR $i = 1$ TO $I$
    FOR $j = 1$ TO $J$
        $d_{i,j} = \min(d_{i-1,j}, d_{i-1,j-1}, d_{i,j-1}) + \text{distance}(a_i, b_j)$
    NEXT $j$
NEXT $i$
RETURN $d_{I,J}$
```

The additional column and row with index 0 are introduced to deal with the special cases $i = 1$ and $j = 1$.

D. Improvements of the algorithm to reduce runtime and memory requirements

The time and space complexity of the DTW algorithm is one of its drawbacks. Since the accumulated distance matrix has $I \cdot J$ entries, the complexity is $O(n^2)$ where $I \approx J \approx n$. This is especially problematic for embedded systems that often have very limited resources.

This complexity can be reduced significantly. For matching biometric samples, the optimal distortion paths run near the diagonal. Therefore, it is sufficient to consider only paths lying completely inside a band around the diagonal from $(0, 0)$ to (I, J), as shown in Fig. 3a. This is achieved by adjusting the boundary conditions of the *for*-loops and setting $d_{i,j} = \infty$ for all (i, j) with $|i - j| > r$. Such a band [3] is called Sakoe/Chiba band. It contains all index pairs (i, j) with $|i - j| \leq r$. The width of the band is $2r$. If the matrix is quadratic, the slope of this band is 45°and it holds $O(n)$ index pairs.

The Itakura parallelogram [2], illustrated in Fig. 3b, may include even less index pairs. Here, the basic idea is to allow only small distortions in the beginning and end of the time series, while bigger ones are accepted in the middle. The implementation is a bit more elaborate than the Sakoe/Chiba band, and its use is not as widespread.

Using either of these approaches, the time and space complexity of DTW can be reduced to $O(n)$. The required memory space can even be constant (space complexity $O(1)$) when using Sakoe/Chiba bands or Itakura parallelograms of fixed size, independent from the length of the time series.

Both approaches are easier to implement with quadratic distance matrices, but also work with non-quadratic ones. The accumulated distance matrix is quadratic if $I = J$, which is generally not the case. There are two ways to solve this problem:

- Adapt the algorithm, so it can handle Sakoe/Chiba bands or Itakura parallelograms around diagonals with a slope angle other than 45°.
- Normalize A and B by interpolating one time series and resampling it with the same number of sample points as the other. Then apply the DTW algorithm in its simpler, quadratic form.

Normalizing the length is easy to implement, but may degrade the recognition accuracy. This will be investigated in the following sections.

In the experiment described in Section IV, we used a DTW implementation with a Sakoe/Chiba band both with and without length normalization. The optimal width of the band was determined empirically as 10% of $\max(I, J)$.

III. Gedanken-experiment

Assume that a skilled forger is able to perfectly replicate the shape of a signature, but doing so, takes him longer than the original signing does. This assumption appears plausible. If the forged signature is length-adjusted to the reference signature before determining the distance between them both, then the temporal dissimilarity between a perfectly shaped forgery and an original is evened out, and the forgery cannot be detected. Therefore, the elapsed time can be expected to matter in distinguishing between original and forged on-line signatures.

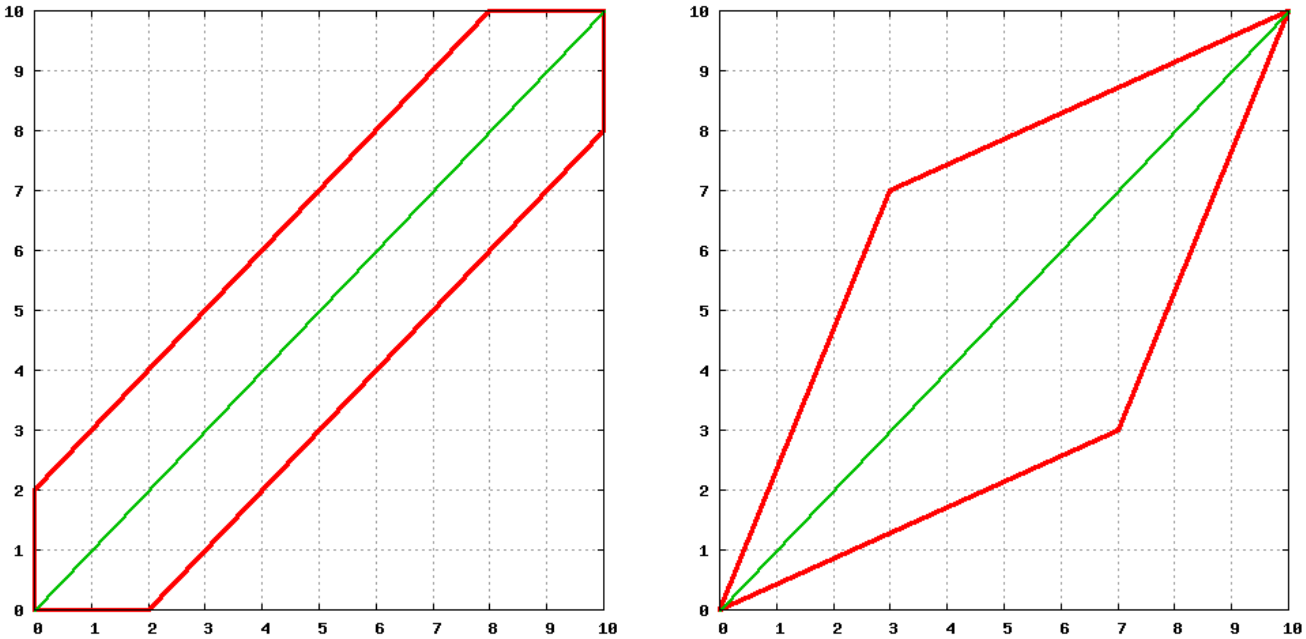
Because it is not clear how big the detrimental effect of length normalization is in practice, this is investigated in the following section.

IV. Empirical experiment

A. Design of the experiment

The hypothesis is that normalizing the on-line signatures to the same length in time will lead to more false accepts of forged signatures than if the length in time of the on-line signatures is not normalized. We have tested this hypothesis by trying both, normalizing the lengths of signatures to the same length before DTW comparison and doing the DTW comparison with the original lengths.

A publicly available subset of the database [12], consisting of signature samples of 100 persons has been used for the experiment. For each person's signature, there are 25 genuine samples and 25 skilled forgeries. In addition to the x and y coordinates, each sample point vector includes a numeric value representing the associated pen pressure. Each of the 250 genuine signature samples has been compared both to the other 24 genuine samples of the same person and to the 25 forgery attempts. All distance values resulting from these comparisons have been recorded. The recognition accuracies in both cases have been calculated from the recorded distance values and been compared with each other.



(a) Sakoe/Chiba band

(b) Itakura parallelogram

Fig. 3. Reducing the number of possible distortion paths

B. Metric used

From the recorded distance values, for each genuine sample, the sample equal error rate (sEER) that is achieved when comparing this sample with the corresponding genuine and forged samples has been calculated. This metric can also be applied in the a-posteriori quality assessment of handwritten signatures [13].

The sEER is determined as follows: The sample false match rate (sFMR) of a biometric sample is the proportion of attempted forgeries of this sample that are falsely declared to match that particular sample. The sample false non-match rate (sFNMR) of a biometric sample is the proportion of should-be matching samples that are falsely declared not to match that particular sample. Both sFMR and sFNMR depend on the threshold chosen for declaring a match or non-match. The sEER of a biometric sample is the value of sFMR and sFNMR at that threshold where both sFMR and sFNMR are equal.

C. Results

Fig. 4 gives an overview of the sEER's of all 100 probands, with time normalization (left column) and without time normalization (right column). Apparently, in many cases the difference is significant. For example, look at the sEER's of the signatures of person 9. With time normalization the mean sEER for person 9 is 12.3%, the lowest value being 5.7%. Without time normalization, i.e. when comparing the samples in their original unaltered form, all sEER's of person 9 are less than 3%.

Furthermore, the sEER's of persons 83 and 93 show that time normalization may cause outliers that are not there without time normalization: One genuine signature of person 93 has a sEER of more than 45%, while the mean value of the other sEER's for that person is less than 5%. This outlier occurs only when time normalization is applied.

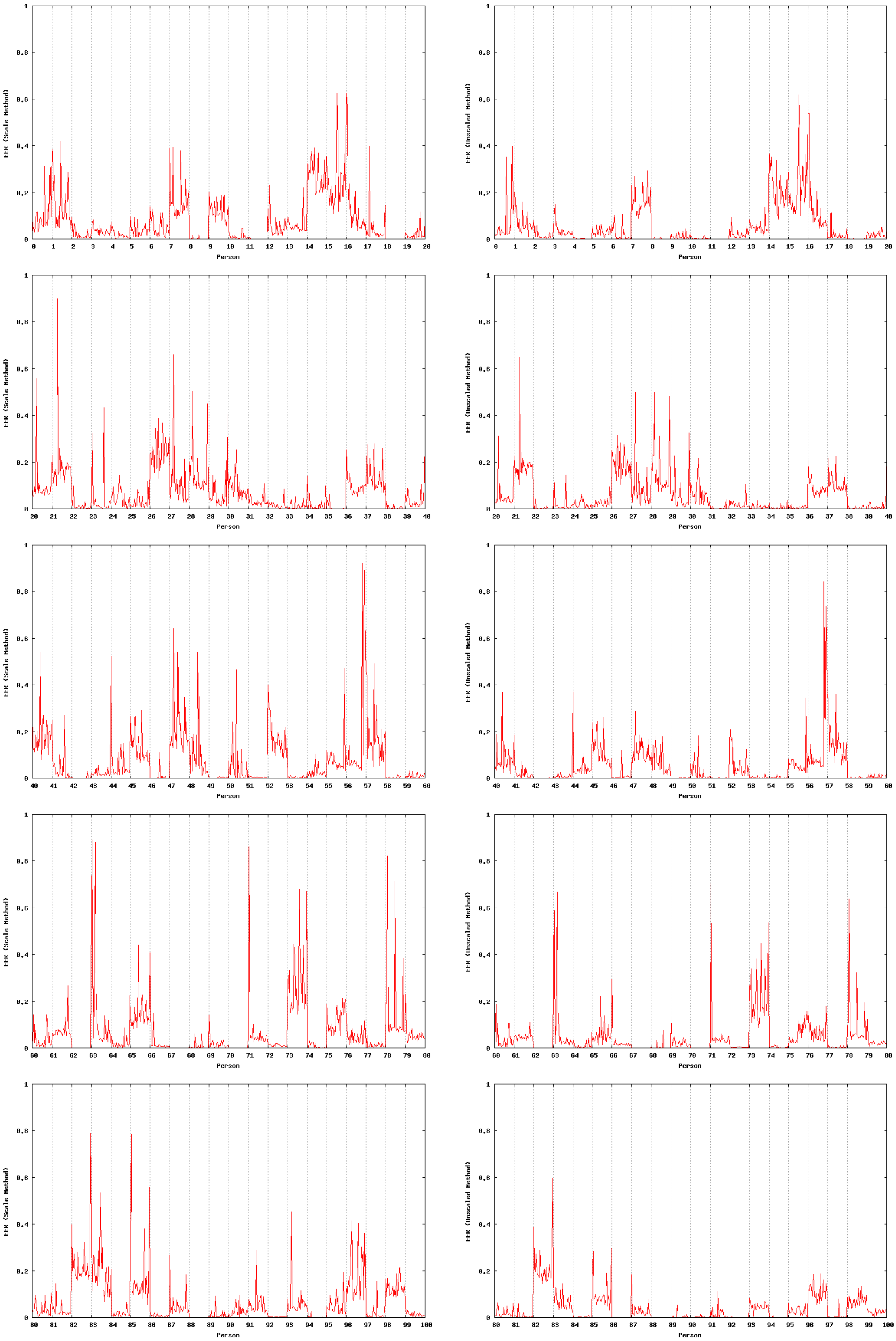
The mean sEER of all signatures with respect to the DTW algorithm using time normalization is 7.1%. The variance is 0.0114. Using no time normalization, the overall recognition accuracy improves significantly. The mean sEER decreases to 4.82%, the variance to 0.0065. This shows the detrimental effect of normalizing the length of on-line signatures on the recognition accuracy.

Furthermore, as could be expected, the length normalization has no significant effect on the distance of genuine signatures to the other genuine signatures of the same person. The mean value of these distances is 38.945 with time normalization and 39.852 without time normalization.

V. Summary

This paper has shown that length normalization reduces the recognition accuracy in application domains where the length of the compared time series matters for their classification as match or non-match. The results of [11] are limited to application domains where the length of the time series does not matter.

The lengths of the time series to be compared can be normalized to the same length if one tries, for instance, to recognize the shape of handwritten characters. However, the time series to be compared should not be normalized to the same length if one tries to verify handwritten signatures, except on the condition that the elapsed time is taken into account as a separate feature whose comparison score is to be fused in some way with the length-normalized DTW distance score. Often, forged signatures imitate the shape of



(a) DTW with time normalization

(b) DTW without time normalization

Fig. 4. Sample equal error rates

the original quite well, but it takes longer to write them. This difference is suppressed if the lengths of the two signatures are normalized to the same length and if the elapsed time is not otherwise taken into account.

Acknowledgments

The authors are grateful to E. Keogh for having inspired this investigation and for a fruitful e-mail discussion that helped to clarify some aspects of this paper. The authors are also grateful to J. Ortega-Garcia and J. Fierrez-Aguilar for making a subcorpus of the MCYT database available for research purposes.

References

[1] J.B. Kruskal. An overview of sequence comparison. In D. Sankoff and J.B. Kruskal, editors, *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*, pages 1–44. Addison-Wesley, 1983.

[2] F. Itakura. Minimum prediction residual principle applied to speech recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-23(1):67–72, 1975.

[3] H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-26(1):43–49, 1978.

[4] E. Keogh and C.A. Ratanamahatana. Exact indexing of dynamic time warping. *Knowledge and Information Systems*, 7(3):358–386, March 2004.

[5] A. Kholmatov and B. Yanikoglu. Identity authentication using improved online signature verification method. *Pattern Recognition Letters*, 26:2400–2408, 2005.

[6] B. Wirtz. Stroke-based time warping for signature verification. In *3rd International Conference on Document Analysis and Recognition*, pages 179–182, Montréal, Québec, Canada, 1995.

[7] C. Schmidt. *On-line Unterschriftenanalyse zur Benutzerverifikation*. PhD thesis, RWTH Aachen, 1999.

[8] Information technology – Biometric data interchange formats – Part 7: Signature/sign time series data. Final Draft International Standard ISO/IEC 19794-7, 2007.

[9] M. Skrbek. Signature dynamics on a mobile electronic signature platform. In R. Grimm, H.B. Keller, and K. Rannenberg, editors, *GI-Jahrestagung – Schwerpunkt "Sicherheit – Schutz und Zuverlässigkeit"*, Frankfurt am Main, Germany, 2003.

[10] O. Henniger and K. Franke. Biometric user authentication on smart cards by means of handwritten signatures. In D. Zhang and A.K. Jain, editors, *1st International Conference on Biometric Authentication*, number 3072 in Lecture Notes in Computer Science, Hong Kong, China, 2004. Springer.

[11] C.A. Ratanamahatana and E. Keogh. Everything you know about dynamic time warping is wrong. In *Workshop on Mining Temporal and Sequential Data, in conjunction with the International Conference on Knowledge Discovery and Data Mining*, Seattle, WA, USA, 2004.

[12] J. Ortega-Garcia, J. Fierrez-Aguilar, D. Simon, J. Gonzalez, M. Faundez-Zanuy, V. Espinosa, A. Satue, I. Hernaez, J.-J. Igarza, C. Vivaracho, D. Escudero, and Q.-I. Moro. MCYT baseline corpus: a bimodal biometric database. *IEE Proceedings Visual Image Processing*, 150(6):395–401, 2003.

[13] S. Müller and O. Henniger. Evaluating the biometric sample quality of handwritten signatures. In *2nd International Conference on Biometrics*, Seoul, South Korea, 2007.