# EVALUATION OF FEATURES AND NORMALIZATION TECHNIQUES FOR SIGNATURE VERIFICATION USING DYNAMIC TIME WARPING

*David Fenton, Martin Bouchard and Tet H. Yeap*

University of Ottawa
School of Information Technology and Engineering
800 King Edward Avenue, Ottawa, Ontario, Canada K1N 6N5
{dfenton, bouchard, tet}@site.uottawa.ca

## ABSTRACT

This paper examines the use of different feature sets and normalization techniques for a signature and password verifier. The verifier made use of the Dynamic Time Warping (DTW) algorithm. Features that incorporated the pen velocity were found to be the strongest performers in tests against informed forgeries. Overall, password verification did not perform as well as signature verification. On average, the equal error rate for passwords was 2.7% higher than for signatures. Most signature and password features achieved their best performance when they were power-normalized, although normalization in both time and power was sometimes beneficial as well.

## 1. INTRODUCTION

Variations of the Dynamic Time Warping (DTW) algorithm have been considered for the automatic verification of handwritten signatures for over 20 years. Despite the popularity of the algorithm, little work appears to have been published on the relative strengths and weaknesses of different features and normalization techniques. This is most likely because the DTW algorithm can be computationally intense, particularly if the signatures are many samples in length, or if warping has to be performed to several reference signatures.

This paper reports a series of tests that were performed on 59 different combinations of feature sets and normalization techniques. The best configurations were identified for signatures and handwritten passwords using both global and per-class thresholds.

## 2. DATA COLLECTION AND PREPROCESSING

Signatures, passwords and forgeries were collected from volunteers using an "ePad-ink" tablet, manufactured by Interlink Electronics, Inc. The tablet captured the pen's $x$ and $y$ position, as well as the pressure, at a 100 Hz sampling frequency. As shown in Table 1, 37 signature classes and 38 password classes were used. For each class, a volunteer attended three signing sessions, contributing 10 signatures and/or 10 passwords each time. Passwords were generated from 9- to 11-letter English words with two letters substituted (e.g. 'abundances' might become 'atubdances'). Long passwords were chosen because a previous study had used only 5- to 6-letter passwords, and it was suspected that the short password length had influenced the reported error rates [1]. For both signatures and passwords, the strokes were concatenated and then adjusted to compensate for translation and rotation variance.

**Table 1**. Overview of the Data Set

| | Signatures | Passwords |
|---|---|---|
| # of classes | 37 | 38 |
| # of genuine specimens | 1110 | 1140 |
| # of informed forgeries | 1075 | 1065 |
| % of forgeries rejected by time verifier | 42.3% | 38.8% |

At each signing session, a volunteer also attempted five informed forgeries of another contributor's signature, and five forgeries of a password. The forger was shown an MPEG movie of the original signer's pen-tip motion. Both real-time and slow-motion movies were available, and forgers were allowed to watch them as many times as they wished. They were also allowed to practice as long as desired, and to discard poor attempts. At least 20 forgeries were collected for each class.

Many of the forgeries were of very poor quality, and could be rejected by a simple time verifier. If $\mu_t$ was the mean signing time of the signatures that comprised the genuine signature template, and $t_c$ was the signing time of a candidate signature, the candidate was rejected as a poor forgery unless:

$$-T_{neg} < t_c - \mu_t < T_{pos} \tag{1}$$

where $T_{neg} = 2.3$ seconds and $T_{pos} = 2.8$ seconds. These values were chosen based on a series of trials in which five-sample templates were chosen for each class. Using $T_{neg} = 2.3$ and $T_{pos} = 2.8$, the False Rejection Rate (FRR) measured across all classes was zero for 99% of trials, and always under 0.5% for the remaining 1% of trials. Using the simple time verifier, 42.3% of forged signatures and 38.8% of forged passwords were rejected. Only the remaining forgeries were considered in the subsequent experiments.

## 3. DTW VERIFIER

Dynamic time warping (DTW) is a simple iterative technique that provides optimal time alignment of two signals through the minimization of a distance measure. Sato & Kogure were the first to suggest the technique for signature verification applications, in 1982 [2].

This work was funded in part by Bell Canada and an Ontario Graduate Scholarship.

**Table 2**. Equal error rates for signatures. False Acceptance Rate was calculated only for the 57.7% of informed forgeries that passed the time verifier.

| Rank | Description | Normalization | Equal Error Rates: | |
|---|---|---|---|---|
| | | | Global threshold | Individual thresholds |
| 1 | Position & velocity | Power | 4.54% | 4.00% |
| 2 | Detrended velocity & tangential acceleration | Time & Power | 4.55% | 3.55% |
| 3 | Velocity | Power | 4.63% | 4.12% |
| 4 | Detrended velocity | Power | 4.68% | 4.10% |
| 5 | Velocity | Time & Power | 4.96% | 4.16% |
| 6 | Position & velocity | Time & Power | 4.99% | 3.92% |
| 7 | Detrended velocity | Time & Power | 5.01% | 4.15% |
| 8 | Detrended position & detrended velocity | Power | 5.02% | 4.55% |
| 9 | Detrended position & detrended velocity | Time & Power | 5.75% | 4.85% |
| 10 | Detrended velocity & tangential acceleration | Power | 5.81% | 4.89% |
| 11 | Detrended $y$ position & detrended $y$ velocity | Power | 6.19% | 5.26% |
| 12 | Position & pressure | Power | 6.38% | 5.42% |
| 13 | Detrended velocity | None | 6.41% | 5.44% |
| 14 | Detrended $y$ position, $y$ velocity & $y$ tangential acceleration | Power | 6.44% | 5.48% |
| 15 | Detrended $y$ position, $y$ velocity & $y$ tangential acceleration | Time & Power | 6.51% | 5.14% |

Because it offers solid performance with few user-selectable parameters, DTW has remained popular since then (see, for example, [3–7]). The DTW-based verifier described in [6] and [7] won the SVC2004 signature verification competition, which tested competing algorithms against a common signature database [8]. Against informed (skilled) forgeries, the winner achieved an average equal error rate of 2.89%.

Let a discretely sampled template signature with total time duration $N_t$ be described by a two-dimensional feature set, $x_t[n_t]$ and $y_t[n_t]$, $1 \leq n_t \leq N_t$. Similarly, a candidate signature of length $N_c$ may be described by $x_c[n_c]$ and $y_c[n_c]$, $1 \leq n_c \leq N_c$. At each point on an $N_t \times N_c$ grid, the distance measure:

$$d\left(n_t, n_c\right) = \sqrt{\left(x_t[n_t] - x_c[n_c]\right)^2 + \left(y_t[n_t] - y_c[n_c]\right)^2} \qquad (2)$$

is evaluated. Note that this function can be easily expanded by adding more feature dimensions under the square root, so long as the features are normalized in power beforehand. The best time alignment of the two signatures is described by a sequence $i[k]$ of template indices, and a sequence $j[k]$ of candidate indices. Both $i[k]$ and $j[k]$ are $K$ samples long, and are subject to the constraints $i[1] = 1$, $i[K] = N_t$, $j[1] = 1$, and $j[K] = N_c$. The dynamic time warping problem is to minimize the overall cost function [9]:

$$D[N_t, N_c] = \min_{i,j} \sum_{k=1}^{K} d(i[k], j[k]) \qquad (3)$$

In practice, this may be performed using dynamic programming. A forward pass is made through the $N_t \times N_c$ grid to determine the lowest cumulative cost at each point, then a backward pass is made to trace the best path. Various representations of the cumulative cost are discussed in References [3, 6, 9]; the one used in this study was:

$$D[n_t, n_c] = \min \begin{cases} D[n_t - 1, n_c] + d(n_t, n_c) + \rho \\ D[n_t, n_c - 1] + d(n_t, n_c) + \rho \\ D[n_t - 1, n_c - 1] + d(n_t, n_c) \end{cases} \qquad (4)$$

where the $\rho$ term is a parameter that punishes lateral (non-diagonal) movement through the cost matrix (set to 0 in this study). The only metric retained in this study was the lowest cumulative cost after the forward pass through the cost matrix, $D[N_t, N_c]$. Note that no normalization factor based on the template or candidate lengths was applied; factors such as $N_t$ or $\sqrt{N_t^2 + N_c^2}$ were found to be slightly detrimental to the verifier's overall performance.

The verifier used in this study was similar to that described in [6] and [7]. The template consisted of five genuine signatures. Two DTW normalization factors were calculated during enrolment: the average DTW distance between a template signature and the nearest of the remaining templates, and the average pairwise distance among template signatures. During verification, the minimum distance between the candidate and the five templates was calculated, as was the average distance between the candidate and all the templates. These distances were divided through by the template normalization factors. They were subsequently reduced to a single number by applying principal component analysis and retaining only the principal component. This number was then compared to a decision threshold to determine if the candidate was judged to be valid or forged.

## 4. EXPERIMENTAL PROCEDURE

The experiments compared four different types of normalization: (1) no normalization; (2) time normalization (cubic spline interpolation to a uniform length of 500 samples); (3) amplitude scaling to a uniform power of 1.0; and (4) time normalization followed by power normalization. Normalization was applied to features before DTW was performed, and was not related to the DTW template normalization previously discussed in Section 3.

The following features were evaluated with all four normalization techniques: position, detrended position, velocity, detrended velocity, pressure, tangential acceleration, normal acceleration, curvature, $y$-dimension velocity (by itself), and detrended $y$-dimension velocity. Jerk, the derivative of pressure, and the cumulative sum of

**Table 3**. Equal error rates for passwords. False Acceptance Rate was calculated only for the 61.2% of informed forgeries that passed the time verifier.

| Rank | Description | Normalization | Equal Error Rates: | |
|---|---|---|---|---|
| | | | Global threshold | Individual thresholds |
| 1 | Position & velocity | Time & Power | 6.84% | 5.76% |
| 2 | Velocity | Time & Power | 6.98% | 5.38% |
| 3 | Detrended velocity | Time & Power | 7.25% | 5.58% |
| 4 | Position & velocity | Power | 7.45% | 6.31% |
| 5 | Detrended velocity & tangential acceleration | Time & Power | 7.75% | 5.64% |
| 6 | Velocity | Power | 7.78% | 6.37% |
| 7 | Detrended velocity | Power | 7.88% | 6.46% |
| 8 | Detrended velocity | None | 8.50% | 7.23% |
| 9 | Velocity | None | 8.56% | 7.26% |
| 10 | Detrended $y$ position, $y$ velocity & $y$ tangential acceleration | Time & Power | 8.83% | 7.34% |
| 11 | Detrended position & detrended velocity | Power | 9.77% | 8.34% |
| 12 | Detrended $y$ position & detrended $y$ velocity | Power | 9.80% | 8.79% |
| 13 | Detrended $y$ position & detrended $y$ velocity | Time & Power | 10.27% | 8.93% |
| 14 | Detrended position & detrended velocity | Time & Power | 10.32% | 8.75% |
| 15 | Position | Power | 10.32% | 9.07% |

the position signal were tested without normalization. Unwrapped path tangent angle was evaluated with no normalization and with time normalization. Last, the following combinations of features were tested with power normalization, and with time & power normalization: position & pressure; position & velocity; detrended position & detrended velocity; detrended $y$ position & detrended $y$ velocity; detrended $y$ position, detrended $y$ velocity & $y$ tangential acceleration; detrended velocity & tangential acceleration; and tangential & normal acceleration. Altogether, 59 combinations of feature set and normalization were evaluated.

800 simulations were run for each signature class, each with a randomly selected template of five signatures and a verification set of 25 genuine signatures. The number 800 was chosen because it was a good compromise between execution speed and performance (smooth error curves and replicable results). The False Rejection Rate and False Acceptance Rate (FAR) were calculated at 51 decision threshold levels between 0.5 and 3.0 for each configuration under test (recall that because of the DTW template normalization factors, the average DTW cost for the template signatures was 1.0). These FRR and FAR calculations were performed for both a global decision threshold applied to all classes, and for individual thresholds applied on a per-class basis. The curves were subsequently upsampled using cubic spline interpolation to allow more accurate determination of the equal error rate (EER). The EER numbers presented in this paper for individual thresholds represent the average EER across classes, with all classes equally weighted.

## 5. EXPERIMENTAL RESULTS

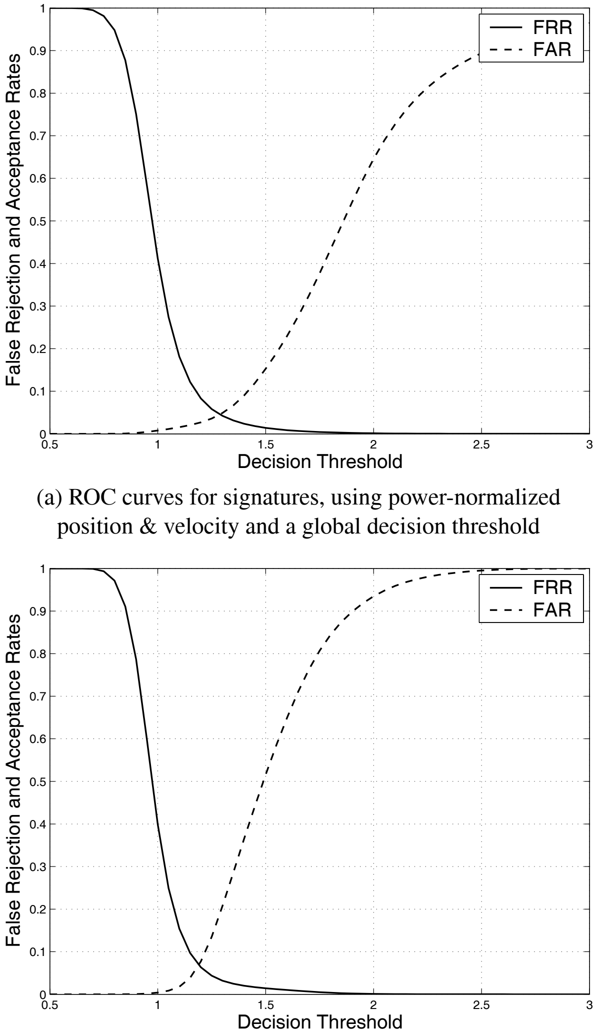
The 15 feature and normalization techniques that performed best for signatures and passwords are listed in Tables 2 and 3, respectively. The FRR and FAR curves of the top-performing features are depicted in Figure 1.

Password verification was found to be generally less effective than signature verification. The best EER achieved using a global threshold was 4.54% for signatures, and 6.84% for passwords. On average, signature features with global decision thresholds achieved EERs 2.7% lower than those achieved with the same password features. Using individual thresholds, the margin was 2.4%. This result was similar to that reported in [1], so it is clear that the use of longer passwords was not sufficient to close the gap with signatures. The error rates reflect the fact that the passwords were not as well practiced as the signatures. Many of the volunteers also chose to print their passwords, and that may have rendered them more susceptible to forgery.

While the reported error rates may seem high when compared to other publications, it is important to remember that testing was performed against informed forgeries rather than random forgeries (which are simply other signers' genuine signatures). In addition, the many forgeries that were rejected by the simple time verifier were not included in the error rate calculations. These forgeries were removed from consideration to reduce the number of computationally intensive warpings that had to be performed. If they had been included in the warping experiments, they would have produced large DTW metrics that were easily rejected, resulting in lower FAR and EER numbers.

The features that achieved the best performance were position & velocity; detrended velocity & tangential acceleration; velocity; and detrended velocity. Of the twelve normalization configurations that were tested for these features, eight ranked in the top ten performers for signatures, and nine were in the top ten for passwords. Other strong-performing features included position; position & pressure; detrended position & detrended velocity; detrended $y$-position & $y$-velocity; detrended $y$-position, $y$-velocity & $y$-tangential acceleration; and unwrapped path tangent angle. All of these occurred at least once in the top 20 for both signatures and passwords (EERs lower than 7.27% for signatures, and 11.52% for passwords).

The worst-performing features, regardless of the normalization technique applied, were jerk; the cumulative sum of the position signal; the derivative of pressure; curvature; and pressure. None of



(a) ROC curves for signatures, using power-normalized position & velocity and a global decision threshold

(b) ROC curves for passwords, using time- and power-normalized position & velocity and a global decision threshold

**Fig. 1**. Receiver Operating Characteristic curves.

these features ranked higher than 42 out of 59 (the best EER was 14.40% for signatures, and 19.21% for passwords).

Power normalization was by far the most important aspect of DTW preprocessing, even for individual features (whose two dimensions were normalized to an equal power). Power-normalized signals were consistently among the better performers, as is easily seen in Tables 2 and 3. On the other hand, time normalization performed more poorly than no normalization, probably because it helped to compensate for forgers' slower movements. Normalizing in both time & power typically gave results similar to power normalization by itself. In some cases, including the top-performing features of Table 3, there was a slight improvement, but most of the time, there was a slight degradation. As an example of these trends, consider detrended velocity. With no normalization, this feature achieved a 6.41% EER for signatures, and an 8.50% EER for passwords. Power normalization brought these numbers down to 4.68% and 7.88% respectively, while time normalization raised them to 10.21% and 12.00%. Combined time & power normalization was beneficial for passwords (7.25% EER), but slightly detrimental for signatures (5.01% EER).

Last, the use of a separate decision threshold for each signature class was beneficial. For signatures, individual thresholds performed an average of 1.1% better than global thresholds, and the margin was 1.3% for passwords. However, given that it would be extremely difficult to determine individual thresholds in an operational system unless signers contributed many template signatures, the global threshold EERs are more realistic numbers.

Altogether, these results should help researchers to select feature and normalization combinations that will be most effective against informed forgeries. Velocity-based features are most successful; in this study, combining the position and velocity signals in the DTW distance metric achieved the best results. In most cases, power normalization of the features was most beneficial, although there were a few cases in which normalization in both time and power yielded slightly better results.

## 6. REFERENCES

[1] M. Parizeau and R. Plamondon, "Relative performances of signature, handwritten password and initials for personal identification," *Proceedings of the 3rd International Symposium on Handwriting and Computer Applications*, Montreal, Canada, July 1987, pp. 164 – 166.

[2] Y. Sato and K. Kogure, "Online signature verification based on shape, motion, and writing pressure," *Proceedings of the 6th International Conference on Pattern Recognition*, Munich, Germany, 1982, vol. 2, pp. 823 – 826.

[3] M. Parizeau and R. Plamondon, "A comparative analysis of regional correlation, dynamic time warping, and skeletal tree matching for signature verification," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 12, no. 7, pp. 710 – 717, July 1990.

[4] B. Wirtz, "Stroke-based time warping for signature verification," *Proceedings of the 3rd International Conference on Document Analysis and Recognition (ICDAR 95)*, Montreal, Canada, Aug. 1995, vol. 1, pp. 179 – 182.

[5] G. Dimauro, S. Impedovo, R. Modugno, G. Pirlo, and L. Sarcinella, "Analysis of stability in hand-written dynamic signatures," *Proceedings of the 8th International Workshop on Frontiers in Handwriting Recognition*, Niagara Falls, Canada, Aug. 2002, pp. 259 – 263.

[6] A. Kholmatov, "Biometric identity verification using on-line and off-line signature verification," M.Sc. thesis, Sabanci University, 2003.

[7] B. Yanikoglu and A. Kholmatov, "An improved decision criterion for genuine/forgery classification in on-line signature verification," *Proceedings of the 13th International Conference on Artificial Neural Networks (ICANN 2003)*, Istanbul, Turkey, June 2003.

[8] "SVC 2004: First international signature verification competition (http://www.cs.ust.hk/svc2004/)," Feb. 2004.

[9] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, Inc., Englewood Cliffs, New Jersey, 1st edition, 1993.