Classification and Authentication of One-dimensional Behavioral Biometrics

John V. Monaco
Pace University, Pleasantville, NY 10570, USA
vmonaco.com

Abstract

For some behavioral biometrics, only the timestamps of a recurring event may be available. This is the case for the recently proposed random time interval (RTI) biometric in which a user repeatedly presses a single button. A dynamical systems approach is taken to deal with biometrics which are inherently one-dimensional. The methodology uses the minimum description length principle to find the optimal time delay embedding for a time series and an optimization to the multivariate Wald-Wolfowitz test for efficiently comparing time series of different lengths. Promising classification and authentication results are achieved on several experimental datasets, utilizing event timestamps only. Classification accuracy ranged from 16.2% to 44.1% and authentication EER from 32.8% to 12.7%. The proposed methodology was also used to achieve first place in the 2014 EMVIC, with 39.6% classification accuracy. All code is made available for experiment reproducibility[1].

1. Introduction

Behavioral biometrics may be placed into two different categories: active, passive. An *active* behavioral biometric is one that requires a user to interact with a system, such as keystroke, mouse, and touchscreen. Thus, many active behavioral biometrics are commonly associated with computer interaction since a computer requires human operation. A *passive* behavioral biometric is one that can be observed without system interaction. Examples of passive behavioral biometrics are gait and voice. This paper is primarily concerned with active behavioral biometrics.

Various active behavioral biometrics induce different cognitive loads depending on the complexity of the task involved. In [14], Newell proposed the time scale of human action. Adapted by [12], a logarithmic time scale is broken up into four different categories, shown in Figure 1. The categories each roughly encapsulate actions that belong to a particular psychological theory.

[1] http://vmonaco.com/ijcb2014

Scale (sec)	Time Units	System	World (theory)
10^7	Months		SOCIAL BAND
10^6	Weeks		
10^5	Days		
10^4	Hours	Task	RATIONAL BAND
10^3	10 min	Task	
10^2	Minutes	Task	
10^1	10 sec	Unit task	COGNITIVE BAND
10^0	1 sec	Operations	
10^{-1}	100 ms	Deliberate act	
10^{-2}	10 ms	Neural circuit	BIOLOGICAL BAND
10^{-3}	1 ms	Neuron	
10^{-4}	100 µs	Organelle	

Figure 1: Newell's time scale of human action [14, 12]

Human actions may fall into one of four different categories based on the time scale of a recurring event. Generally, actions that occur higher up in the table are more difficult to quantity and effectively classify [13]. The difficulty comes as a result of greater behavioral variability. For example, the cognitive band enters consciousness where humans are "in control" of the actions they make, whereas there is little control over actions that take place in the biological band.

Traditional behavioral biometrics rely on the presence of a rich stream of information to authenticate or identify an individual. Normally, a behavioral biometric sample consists of a sequence of events where each event occurs at time t and may contain additional attributes. In the case of keystroke biometrics, an event sequence would be a series of keystrokes ordered by the key press time. Event attributes include the name of the key that was pressed and the duration it was held down for.

It is the additional information contained in the attributes of each event that allows researchers to define a rich set of features on the event sequence. Events are normally categorized according to their attributes. For example, in keystroke biometrics events are commonly identified by the name of the key that was pressed. Descriptive statistics can

then be taken on sets of event classes. Features such as the mean duration of the "E" key, or variance of vowel key durations, are typical in keystroke biometrics. Events with no additional information cannot be categorized in this way, thus the events themselves are indistinguishable. This is the case for a one-dimensional time series that contains only event timestamps. This work presents a methodology for classifying and authenticating one-dimensional behavioral biometrics.

The research contributions of this work are the following:

1. A time and space optimization to the multivariate Wald-Wolfowitz test, coined the *approximate multivariate Wald-Wolfowitz test* (AMWW), empirically shown to yield acceptable results
2. Proposed methodology for classification and authentication of one-dimensional behavioral biometrics
3. A simple extension of the methodology for classifying multivariate time series
4. Experimental classification and authentication results on several publicly available datasets, including other proposed sources of one-dimensional behavioral biometrics.

The rest of the paper is organized as follows: Section 2 defines the RTI biometric and reviews work that has been done. Section 3 uses Takens' Theorem and several heuristics to reconstruct system dynamics. The optimization to the Wald-Wolfowitz test is covered in Section 4. Section 5 explains the classification and authentication methodology, while experimental results and conclusions are presented in Section 6 and Section 7, respectively.

2. Random-time interval biometric

Let an event e_t be some action taken by a user at time t. An event could be a key press, mouse movement, or some high level cognitive action. It may also carry additional information with it, such as the action that occurred (e.g. press or release), the key name, or screen coordinates. However, events from a one-dimensional (1-D) biometric do not contain any of this information. At each event, only the time the event occurred is known. A 1-D behavioral biometric sample consists only of a time series of event timestamps, $[t_1, t_2, \ldots, t_n, t_{n+1}]$ from events $[e_{t_1}, e_{t_2}, \ldots, e_{t_n}, e_{t_{n+1}}]$. Commonly, the interval between timestamps is taken, to give a series of random time intervals (RTI) of length n:

$$y = [t_1 - t_0, t_2 - t_1, \ldots, t_{n+1} - t_n]$$

An RTI biometric may be the result of a user repeatedly performing a single action, such as pressing a key on a keyboard. It can also arise in situations where additional information, if any, cannot be obtained and only the event timestamps are known. Such a scenario is encountered during the observation of encrypted web traffic, anonymized phone logs, or anonymized email message history.

With the introduction of the RTI biometric in 2009 [11], there has not yet been extensive research on its applicability. The author used the Wald-Wolfowitz (WW) test statistic as a distance measure and reconstructed the samples in feature space with multidimensional scaling (MDS). The time series were uniformly embedded with parameter selection guided by the net class separability in feature space. With an RTI dataset consisting of 40 users repeatedly pressing a single key on a keyboard (hereafter this action referred to as *key-blow*), who each supplied 10 samples of 127 time intervals (or 128 key-blows), the EER of various classifiers ranged from 19.80% to 5.40%. The best performance was achieved by a minimum class support vector machine (MCVSVM) trained on the optimized vectorial representations of the RTI signals.

Human-generated random number (HGRN) sequences have also been studied. These could be thought of as an RTI operating at a higher cognitive level, where the RTI in [11] operates at the biological band of Newell's Time Scale. Though randomness and unexpected decisions play an important role in psychology, there has not been much research which considers HGRN as a biometric. In [17], HGRN sequences were shown to be predictable, and it is well known that many HGRN sequences lack true randomness. In [4], both the RTI and HGRN are considered as behavioral biometrics. Data was collected from 30 participants, who each provided 10 random number sequences. The subjects were given an explanation of randomness using the "balls in a hat" model and then told to generate random number sequences for a given duration. Wavelet decomposition and approximate entropy were used as features for a SVM classifier to achieve an equal error rate of 4.3% on HGRN sequences. A comparative study on key-blow RTI was performed with data collected from the same participants. Again, 10 samples were provided by each participant, with each sample consisting 150 key presses (slightly longer than [11]) to achieve an EER of 4.7%.

The multivariate Wald-Wolfowitz test [2] has been successfully applied to the eye movement biometric. In [15], the WW test is used to compare eye movement samples where each sample is an 8-dimensional time series composed of the velocity and acceleration of each eye in horizontal and vertical directions. A k-nearest neighbor (kNN) algorithm using the WW similarity as a distance measure achieved good results, with an ACC1 (percentage of correctly classified samples) of 91.5% and 82% on two different datasets.

A key-blow RTI dataset is used in the following sections

as an illustrative example of concepts defined. The dataset is from the labeled portion of data made available by the 2014 RTI competition and is fully described in Section 6.

3. Takens' Theorem

Given a series of observations, it is possible to reconstruct the dynamics of the original system through time delay embedding. This remarkable theorem is due to Florin Takens [19]. A time delay embedded series is a transformation of the original time series y, such that

$$\mathbf{x}_i = [y_i, y_{i-\tau}, y_{i-2\tau}, \ldots, y_{i-(d_e-1)\tau}]$$

The time lag is given by τ and the number of dimensions by d_e. Takens' embedding theorem states that the evolution of $\mathbf{x}_i$ will approach the evolution of the dynamical system state given that d_e is large enough. In other words, assume that only one variable of a multi-dimensional system can be measured. The dynamics of the original system can be reconstructed, up to a homomorphism, with observations made on the single variable.

The theorem requires a few assumptions. The observation function must be at least twice differentiable, there must be enough data, and the data sampled often enough [18]. In practice, these assumptions will never be met due to the discretized nature of any data on a computer. Despite this, results of the theorem can still be observed on experimental data [18].

3.1. Uniform time delay embedding

There are two key parameters in Takens' time-delay embedding procedure: the embedding dimension, d_e and the time-lag, τ. The goal is to find $\mathbf{x}_i$ such that $\mathbf{X} = [\mathbf{x}_{d_e \times \tau}, \mathbf{x}_{d_e \times \tau + 1}, \ldots, \mathbf{x}_n]$ is a faithful representation of the system dynamics. Here, $\mathbf{X}$ is the $d_e \times (n - d_e \times \tau)$ matrix formed by the series of time delay embedded vectors. This type of embedding is *uniform* since the time lags between each dimension are equal. An *irregular* (or non-uniform) embedding is one with a variable time lag. The estimation of time lag and embedding dimension are described in the following sections.

3.1.1 Time lag

The time lag is found first, using a heuristic which does not depend on the embedding dimension. The value of τ will determine the shape of the embedded time series. The time lag should be chosen so that the data are separated as much as possible in reconstructed phase space while still being "close". Ideally, the data should not be correlated too much [1].

The mutual information (MI) of a time series is given by (1) and can be used to estimate τ. Usually, the value of T would be chosen where the first local minimum of $I(T)$ occurs [1].

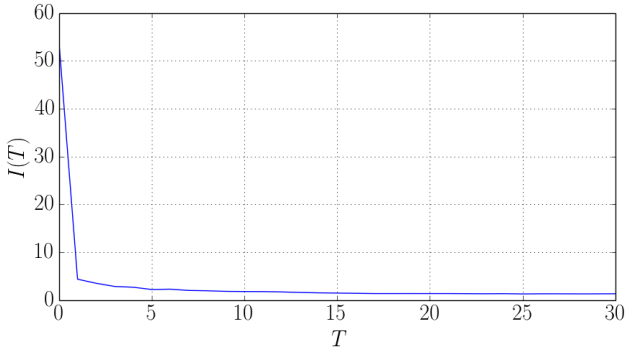


Figure 2: Mutual information of the key-blow RTI dataset as the time lag is increased

$$I(T) = \sum_{n=1}^{N} P(y_n, y_{n+T}) \log_2 \frac{P(y_n, y_{n+T})}{P(y_n)P(y_{n+T})} \quad (1)$$

The mean mutual information over all of the samples in the key-blow RTI dataset is given in Figure 2. It can be seen that there is no local minimum. Despite this, the MI can still be used by choosing the value of T where $I(T)/I(0) \approx \frac{1}{5}$ [1]. In Figure 2, we have $I(0) = 52$ and $I(1) = 5$. Thus, $\tau = 1$ is chosen as the time lag, coming as close as possible to the desired ratio. Note that this corresponds well to one of the time lags found in [11] on a similar dataset, where a functional measuring the net class separability is used as a filter for choosing an embedding strategy.

3.1.2 Embedding dimension

A geometrical approach is used to estimate d_e. The method of *false nearest neighbors* (FNN) [8] is a technique where the embedding dimension is continually increased, and changes in dynamics of the system in each higher dimension are observed. As the embedding dimension increases, points which are actually close to each other in the dynamical system should separate slowly compared to points which are not actually close and can only be distinguished in higher dimensions.

The method of FNN proceeds as follows. Consider the unordered vectors $\mathbf{x}_i \in \mathbf{X}$ as points in $\mathbb{R}^{d_e}$. Each vector $\mathbf{x}_i$ has a closest neighbor, $\mathbf{x}_i^{NN}$. Let the index of $\mathbf{x}_i^{NN}$ be i^{NN}. If $\mathbf{x}_i$ and $\mathbf{x}_i^{NN}$ are actual nearest neighbors in the system of dimension d_e, then the distance between them will increase slowly as the embedding dimension is increased from d_e to $d_e + 1$. The normalized increase in distance is given by (2).

$$R = \frac{|y_{i-(d_e+1)\tau} - y_{i^{NN}-(d_e+1)\tau}|}{\| \mathbf{x}_i - \mathbf{x}_i^{NN} \|} \quad (2)$$

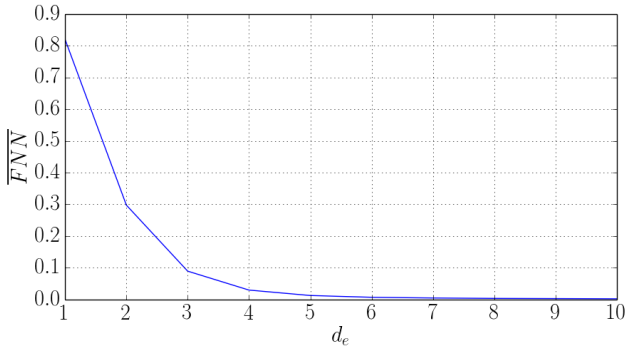


Figure 3: Ratio of false nearest neighbors in the key-blow RTI dataset as the embedding dimension is increased

A pair of points are FNN if $R \geq R_T$, where R_T is a predefined threshold. Generally, the embedding dimension can be found by increasing d_e until most points are not FNN. [18] suggest using $R_T = 15$, although this value depends on the data available and is ultimately chosen at the discretion of the practitioner.

Using a value of $R_T = 15$, the embedding dimension is found for the key-blow RTI dataset. With the time lag of $\tau = 1$ found in the previous section, the proportion of points in each sample for which $R \geq R_T$ is found. The mean proportion of points for which $R \geq R_T$ over the entire dataset is found by taking taking the average proportion from each sample, denoted by $\overline{FNN}$. The results are shown in Figure 3.

A reasonable choice for d_e that results in less than 5% of points being FNN in each sample is found to be $d_e = 4$. Not much is gained by increasing the embedding dimension beyond this, so $d_e = 4$ is chosen as the embedding dimension of the key-blow RTI dataset. Note that an embedding dimension of $d_e = 7$ was found on a similar dataset in [11]. Care must be taken in selecting an embedding dimension which is too high, though. The unnecessary degrees of freedom will often end up being filled by noise in the data and not the dynamics of the system itself.

3.2. Irregular embedding

The need for non-uniform embedding is apparent for time series with more than one dominant periodicity. A uniform embedding, as described above, will not be able to handle situations in which a time series contains both low and high frequency components. In [5], what's called an *irregular,* or *non-uniform,* embedding is introduced. This is one such that the time lag τ is not fixed. Instead, a lag vector l is used to create the time-delay embedded vectors,

$$\mathbf{x}_i = [x_{i-l_1}, x_{i-l_2}, x_{i-l_3}, \ldots, x_{i-l_{d_e}}]$$

where $0 \leq l_1 \leq l_i < l_{i+1} \leq l_{d_e}$. The variable time lags are able to reconstruct systems that operate on multiple time scales. This allows the dynamics of a system with multiple periodic components to be captured. Both [18, 5] go into more detail on irregular embedding.

It can be seen that there is now a combinatoric explosion with the possibilities of lag vectors. A search algorithm is needed, guided by a heuristic for how well an irregular embedding allows for reconstruction of system dynamics.

3.2.1 MDL principle

The minimum description length (MDL) principle states that a model should faithfully describe the data without being overly complex [3]. In [18], the MDL principle is used to guide the selection of time lags. For a sequence of n time delay embedded vectors $\mathbf{X}$, the description length is given by (3),

$$DL(\mathbf{X}) = \frac{d}{2} \ln \left[\frac{1}{d} \sum_{i=i}^{d} (y_i - \bar{y})^2 \right] + d + DL(d) + \frac{n-d}{2} \ln \left[\frac{1}{n-d} \sum_{i=d+1}^{n} e_i^2 \right] \quad (3)$$

where $\overline{y}$ is the mean of the first d terms in the original time series, d is the maximum embedding lag l_{d_e}, $DL(d)$ is the description length of the integer d [16], and e_i is the model prediction error at index i. For sufficiently long time series, the first three terms in (3) have little effect on the total description length and $DL(\mathbf{X})$ becomes asymptotically equivalent to the model prediction error. The model prediction error can be computed by a *drop-one-out* approach [18].

3.2.2 Optimal embedding: a greedy approach

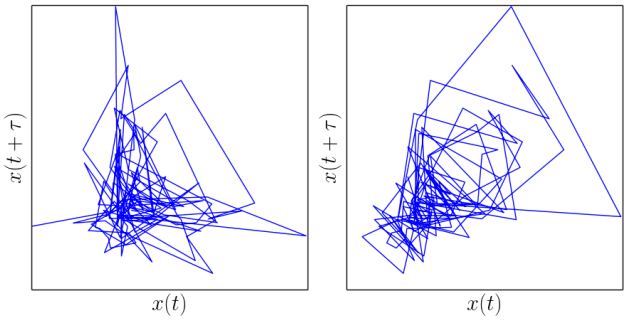
A simple greedy algorithm is used to choose an optimal embedding strategy. As will be seen in the experimental results, this generally leads to better classification accuracy than a uniform embedding found by MI and FNN alone. An embedding window, $d_w = d_e \times \tau$, is defined as the maximum lag to search within. The algorithm is given in Algorithm 1.

Starting with a single lag of $[1]$ in line 1, a local optimal choice is made on each iteration by concatenating a new lag to the lag vector in line 6. Note that in lines 6 and 9, $\cup$ denotes the concatenation of lag k with lag vector l. The loop from lines 4-13 steps through each possible lag up to d_w and only keeps the new lag if the description length is reduced. It is possible for the lag vector to be out of sorted order, since the inner loop goes from i to d_w on each iteration. Although the inner loop may terminate early when a

Algorithm 1 Greedy optimal embedding strategy

INPUT N one-dimensional time series, $\{y_j\}$
OUTPUT embedding strategy l

$l := [1]$
$mdl := \infty$
$dl := \frac{1}{N}\sum_{i=0}^{N} DL(y_i, l)$
for i in $range(2, d_w)$ **do**
  for k in $range(i, d_w)$, not in l **do**
    $dl := \frac{1}{N}\sum_{j=0}^{N} DL(y_j, l \cup [k])$
    if $dl < mdl$ **then**
      $mdl := dl$
      $l := l \cup [k]$
      break
    end if
  end for
end for



(a) Uniform embedding (b) Irregular embedding

Figure 4: Key-blow RTI sample with uniform embedding determined by MI and FNN and irregular embedding determined by MDL

new MDL is found, the algorithm terminates only after the outer loop has finished.

To see the effects of an irregular embedding chosen by MDL versus a uniform embedding determined by MI and FNN, consider an illustrative example. A single randomly chosen sample from the key-blow RTI dataset is uniformly embedded with $d_e = 4$ and $\tau = 1$, as determined in the previous sections. This is shown in Figure 4a. The same sample is embedded with the optimal lag vector for the key-blow RTI dataset, as determined by Algorithm 1, and is shown in Figure 4b. An embedding window of $d_w = d_e \times \tau = 4$ is used. In each figure, two dimensions are selected at random for visualization.

A quick visual analysis shows that an attractor appears to emerge in the irregularly-embedded sample, whereas the uniformly-embedded sample lacks a richer structure. The optimal lag vector used to embed the sample in Figure 4b is $l = [1, 2]$. In this case, the lag vector is equivalent to a uniform embedding with $d_e = 2$ and $\tau = 1$. In general, the embeddings found by the MDL principle exhibit better dynamics than uniform embedding, and the superiority of MDL as a heuristic for embedding parameters will be empirically shown in the experimental results.

4. Multivariate Wald-Wolfowitz Test

The Wald-Wolfowitz (WW) test, originally proposed by [21], is a nonparametric test to determine whether two samples come from the same distribution. The multivariate WW test [2] is a generalization of the WW test that is able to make this determination for samples coming from distributions of any dimension.

The test works as follows. Consider the set of vectors from two time delay embedded samples with embedding dimension d_e. Take the embedded vectors from both samples and construct the minimum spanning tree (MST) over all observations in $\mathbb{R}^{d_e}$. A *run* is a segment of the tree that traverses vectors from only one sample. Runs are separated by edges which connect nodes (i.e. embedded vectors) from different samples. In the case that both samples came from the same distribution, the branches of the tree will likely encounter vectors from each sample, resulting in a large number of runs. If the observations came from different distributions, then the branches will traverse the vectors of one sample and then the other sample, resulting in relatively few runs.

The number of runs is used in the WW statistic to determine whether the samples originated from the same distribution. More precisely, the WW statistic and the expected number of runs are given by (5) and (4), respectively, where m and n are the number of vectors in each sample and $N = m + n$.

$$E(R) = \frac{2mn}{N} + 1 \tag{4}$$

$$W = \frac{R - \frac{2mn}{N} - 1}{\left(\frac{2mn(2mn-N)}{N^2(N-1)}\right)^{\frac{1}{2}}} \tag{5}$$

It has been shown that $E(R)$ has a standard normal distribution [2].

4.1. Optimizing the Wald-Wolfowitz Test

Since the WW statistic relies on construction the MST over both samples, it can be show that this function is $O(N^3)$. A simple assumption can dramatically reduce the complexity of this function, making it practical to compare large samples.

Consider only the distances to the nearest k neighbors of each vector in the embedded samples. This can efficiently be computed and represented as a sparse matrix with

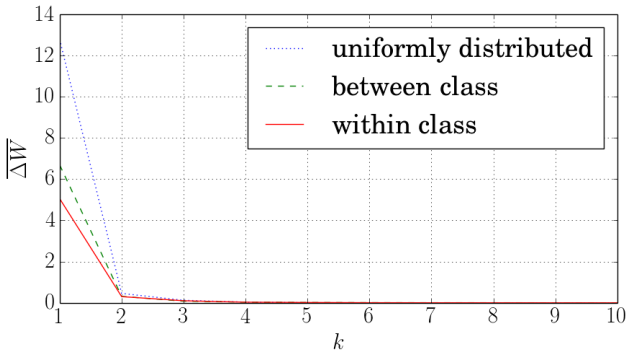


Figure 5: Difference between MST and AMST in computing W

non-zero elements containing the Euclidean distances between neighboring points, as opposed to the full distance matrix needed to compute the true MST. Construct the *approximate minimum spanning tree* (AMST) using the sparse matrix that contains distances to the k-nearest neighbors of each vector. The *approximate multivariate WW* (AMWW) statistic can be computed from the AMST. The cost is then only $O(Nk\ \log(Nk))$. Note that for $k = N - 1$, the true MST is given.

Because a sparse distance matrix is used to construct the AMST, distances of zero are ignored (i.e. zero distances indicate no edge between nodes in the distance graph). To avoid the occlusion of points lying directly on top of each other, a small Gaussian noise is added to each dimension of the embedded vectors before the $N \times k$ distances are found.

For most time series, the AMST is close to the MST and has a negligible effect on W. This is shown with synthetic and experimental data. Let $W^{(k)}$ denote the AMWW similarity obtained with the AMST for k neighbors. With a uniform embedding of $d_e = 4$ and $\tau = 1$ (derived in Section 3), the WW test is applied to 1000 random pairs of samples in the key-blow RTI dataset. The mean difference $\overline{\Delta W^{(k)}} = \frac{1}{1000} \sum_{i=0}^{1000} |W_i - W_i^{(k)}|$ is taken for $1 \leq k \leq k_{max}$ over all of the samples, where W_i is the true WW statistic for sample i. This is performed for the within-class and between-class samples separately. In addition to this, 1000 pairs of synthetic samples uniformly distributed in the unit hypercube of $\mathbb{R}^4$ are generated, with n and m ranging between 10^2 to 10^3. The results are shown in Figure 5.

The difference quickly drops off, showing that the experimental data is *well behaved* and the AMWW gives a good approximation to the true WW statistic. With large enough k, there is no significant difference between W and $W^{(k)}$. For the rest of the paper, the AMWW statistic with a value of $k = 10$ is used and W denotes $W^{(10)}$

5. Identification and verification methodology

The methodology for identification and classification are as follows. First all of the samples are embedded with either a uniform or irregular lag vector. Next, W is found between the unknown sample and every enrolled sample. The mean similarity $\overline{W}_c$ is taken between the unknown sample and each class c. For identification, the class which maximizes $\overline{W}_c$ is chosen as the label for the unknown sample. In this sense, the WW statistic acts as a distance measure in a kNN classifier with $k = 1$.

Authenticating an unknown sample proceeds similar to classifying an unknown sample, with authentication decisions made by a global threshold. After $\overline{W}_c$ has been found for each class in the training dataset, a normalized linear weight is assigned to each class based on its similarity to the unknown sample. The weight of the claimed identity c is then compared to a global threshold to make an authentication decision.

Consider a simple example. Suppose there are 10 classes in the training dataset. $\overline{W}_c$ is found between the query sample with claimed identity I and each class c in the labeled data. The similarity measures are sorted in decreasing order, i.e with the most similar class first. A normalized linear weight is assigned to each class based on its position in the sorted list. The first sample gets $\frac{10}{10} = 1$, the second sample $\frac{9}{10} = 0.9$, and so on, with the last sample having a weight of $\frac{1}{10} = 0.1$. An authentication is performed by checking whether the normalized linear weight of $\overline{W}_I$ is above some threshold w_{thresh}, where $0 \leq w_{thresh} \leq 1$. The false acceptance rate (FAR) and false rejection rate (FRR) are determined by adjusting the threshold from 0 to 1. The receiver operating characteristic (ROC) curve is the tradeoff between FAR and FRR, determined by the parameter w_{thresh}.

For both classification and authentication results, a leave-one-out cross validation (LOOCV) is used. Classification accuracy is reported as the percentage of correctly classified samples (ACC1). Authentication accuracy is reported as the equal error rate (EER) on the ROC curve. For n users with m samples, there $n^2 \times m$ authentications, with $n \times m$ positive authentications and $n(n-1) \times m$ negative authentications. The classification and authentication algorithms are intentionally kept simple since the effect of embeddings and viability of other RTI sources are the primary focus of the paper, not maximizing classification accuracy.

6. Experimental results

The key-blow RTI dataset used in this paper consists of the labeled samples provided in the 2014 RTI contest, one of the official 2014 IJCB competitions. Each time series in the dataset contains 129 time intervals from 130 key-blow events (i.e. repeatedly hitting a single key 130 times). There are a total of 60 users and 7 samples per user in the dataset.

Table 1: RTI dataset results

Embedding	ACC1(%)	EER(%)
[11] Uniform $d_e = 7, \tau = 1$	NA	19.80
[11] Uniform $d_e = 4, \tau = 3$	NA	14.51
Uniform $d_e = 7, \tau = 1$	38.3	15.4
Uniform $d_e = 4, \tau = 3$	31.9	16.2
Optimal uniform $d_e = 4, \tau = 1$	41.4	14.9
Optimal irregular $l = [1, 2]$	44.1	14.0

Each user provided samples on 5 separate days.

6.1. Uniform vs non-uniform embedding

Classification and authentication accuracies were obtained for both uniform and irregular embeddings of the key-blow RTI dataset. Uniform and irregular embedding parameters were determined in 3.1 and 3.2. Results for $d_e = 4$, $\tau = 3$ and $d_e = 7$, $\tau = 1$ are also shown for comparison with the results in [11], which utilized a similar key-blow RTI dataset of 128-blow samples and 40 users with 10 samples per user. The classification and authentication results are summarized in Table 1.

The EERs shown in Table 1 from [11] were found by a comparable classifier that used a global thresholding scheme similar to the one presented in Section 5. This was not the best performance achieved by [11], as more sophisticated classification algorithms were used to obtain lower error rates.

6.2. Higher derivatives

The RTI can be thought of as the first derivative of a timestamp sequence, while higher derivatives of the original time series may also yield useful information. The same methods described can be used on each of several derivatives of a single time series.

A summation rule is used as a simple way of combining similarity scores from multiple classifiers. Summing classifier outputs can be particularly effective if independence is assumed between classifier scores [10]. With the lag vector $l = [1, 2]$, the $\overline{W}$ scores from the first derivative (i.e. the RTI) and second derivative are added together before making a classification or authentication decision. In this way, a separate score is obtained from each derivative time series, and a single final score is achieved for the sample in question. This approach led to the best results on the key-blow RTI dataset, with an EER of 12.6% and ACC1 of 48.1%.

6.3. Additional results

Several publicly available datasets are considered as 1-D behavioral biometrics in addition to the key-blow RTI dataset. In the following experiments, only the event timestamps in each dataset are used. The size of each dataset is reduced to match the size of the key-blow RTI dataset. Each sample is limited to 130 events, with 7 samples per user and 60 users, unless otherwise noted.

The CMU keystroke password benchmark dataset [9] contains the keystroke timings of 51 users entering an 11-character password. Each sample consists of the key press and release timestamps, for a total of 22 events (this is the only dataset where each sample contained less than 130 events). Thus, each sample as a RTI contains 21 time intervals. With 400 samples per user in the original dataset, a subset of 7 randomly-selected samples per user were chosen to match the size of other datasets used.

A mouse motion dataset is used, which contains mouse motion events from 58 students taking online exams over a semester [13]. Samples were collected on mostly homogeneous hardware and each event is the motion delta generated by the movement of a standard desktop mouse. Although the motion deltas, or alternatively the screen coordinates, may contain valuable information, only the event timestamps are considered in the spirit of analyzing 1-D behavioral biometrics. Again, samples were limited to 130 events and 7 samples per user.

A free-text keystroke dataset contains responses from students answering open-ended questions [20]. Similar to the password dataset, an event could be either the press or release of a key. A subset of this dataset was used, which contained 60 randomly-selected users and 7 randomly selected samples per user. Samples were limited to 130 events, equal in size to the key-blow RTI dataset.

Finally, the web history database provided by [7] contains the web page visit times of 454 users. The timestamp, anonymized web page id, visit type, and previous visit location is given for each record. For a comparative study, a subset of the data was used. The reduced web history dataset contained 60 randomly selected users with 7 sessions, limiting each session to 130 web page visits. Only the event timestamps from each sample were used for classification and authentication.

The classification and authentication results of all 5 datasets are shown in Table 2. The mean event frequency of each dataset is also shown for a rough placement in Newell's Time Scale. Web history clearly falls into the rational or social band, while password, mouse, keystroke, and key-blow are in the cognitive band. The irregular embedding for each dataset was determined by the methods described in 3.2, with $d_w = d_e \times \tau$, where d_e and τ were found by the methods in 3.1.

7. Conclusions

For the key-blow RTI dataset, the MDL principle served as a better heuristic for choosing embedding parameters than MI and FNN. Lower error rates are seen when an irregular lag vector found by a greedy search algorithm is used

Table 2: Additional datasets results

Dataset	Freq. (Hz)	Embed.	ACC1(%)	EER(%)
Pass.	8.5	$[1, 2]$	33.3	18.3
Mouse	9.0	$[1, 2, 3]$	34.5	17.1
Keyst.	4.2	$[1, 2]$	43.1	12.7
Key-blow	3.4	$[1, 2]$	44.1	14.0
Web	6.8×10^{-5}	$[1, \ldots, 9]$	16.2	32.8

to embed the samples before applying the non-parametric Wald-Wolfowitz test. Using an approximation to the MST in the WW statistic also significantly decreased the cost of comparing samples with no apparent loss of accuracy.

Choosing a heuristic for estimating time-delay embedding parameters is an important consideration. MDL naturally leads to a model which is descriptive for the dataset in question. Whether or not this is superior to a discriminative heuristic is a problem left for future work. Additionally, it may be possible to increase classification and authentication accuracies by using transformations of the original time series, such as taking higher derivatives, although this area is relatively unexplored. The methods described can be utilized on multivariate time series by simply summing the W scores for each component (as described in 6.2). This technique was employed by the author in the Second Eye Movements Verification and Identification Competition (EMVIC) [6], to place 1^{st} with a classification accuracy of 39.6%. Derivatives of the horizontal and vertical positions of eye movement coordinates were be taken with W computed for each series.

Finally, with anonymity becoming a valid concern, this work has some interesting applications due to the very few and general assumptions that are made: all that is required is the timestamp of a recurring event. That event could be a keystroke, web page visit, or possibly timestamps from a phone log or encrypted network traffic, although these applications are left for further investigation. Over time, simply the observation of a recurring event at given intervals may be enough to identity or authenticate an individual, even if no other information is leaked.

References

[1] H. D. Abarbanel, R. Brown, J. J. Sidorowich, and L. S. Tsimring. The analysis of observed chaotic data in physical systems. *Reviews of modern physics*, 65(4):1331, 1993.

[2] J. H. Friedman and L. C. Rafsky. Multivariate generalizations of the wald-wolfowitz and smirnov two-sample tests. *The Annals of Statistics*, pages 697–717, 1979.

[3] P. D. Grünwald, I. J. Myung, and M. A. Pitt. *Advances in minimum description length: Theory and applications*. MIT press, 2005.

[4] E. Jokar and M. Mikaili. assessment of human random number generation for biometric verification. *Journal of medical signals and sensors*, 2(2):82, 2012.

[5] K. Judd and A. Mees. Embedding as a modeling problem. *Physica D: Nonlinear Phenomena*, 120(3):273–286, 1998.

[6] P. Kasprowski, O. V. Komogortsev, and A. Karpov. First eye movement verification and identification competition at btas 2012. In *Biometrics: Theory, Applications and Systems (BTAS), 2012 IEEE Fifth International Conference on*, pages 195–202. IEEE, 2012.

[7] R. Kawase, G. Papadakis, E. Herder, and W. Nejdl. Beyond the usual suspects: context-aware revisitation support. In *HT*, pages 27–36, 2011.

[8] M. B. Kennel and H. D. Abarbanel. False neighbors and false strands: A reliable minimum embedding dimension algorithm. *Physical review E*, 66(2):026209, 2002.

[9] K. S. Killourhy and R. A. Maxion. Comparing anomaly-detection algorithms for keystroke dynamics. In *Dependable Systems & Networks, 2009. DSN'09. IEEE/IFIP International Conference on*, pages 125–134. IEEE, 2009.

[10] J. Kittler, M. Hatef, R. P. Duin, and J. Matas. On combining classifiers. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 20(3):226–239, 1998.

[11] N. A. Laskaris, S. P. Zafeiriou, and L. Garefa. Use of random time-intervals (rtis) generation for biometric verification. *Pattern Recognition*, 42(11):2787–2796, 2009.

[12] I. MacKenzie. *Human-Computer Interaction: An Empirical Research Perspective*. Elsevier Science, 2012.

[13] J. V. Monaco, J. C. Stewart, S.-H. Cha, and C. C. Tappert. Behavioral biometric verification of student identity in online course assessment and authentication of authors in literary works. 2013.

[14] A. Newell. *Unified Theories of Cognition*. William James Lectures. Harvard University Press, 1994.

[15] I. Rigas, G. Economou, and S. Fotopoulos. Human eye movements as a trait for biometrical identification. In *Biometrics: Theory, Applications and Systems (BTAS), 2012 IEEE Fifth International Conference on*, pages 217–222. IEEE, 2012.

[16] J. Rissanen. *Stochastic complexity in statistical inquiry*, volume 511. World scientific Singapore, 1989.

[17] M.-A. Schulz, B. Schmalbach, P. Brugger, and K. Witt. Analysing humanly generated random number sequences: a pattern-based approach. *PloS one*, 7(7):e41531, 2012.

[18] M. Small. *Applied nonlinear time series analysis: applications in physics, physiology and finance*, volume 52. World Scientific, 2005.

[19] F. Takens. Detecting strange attractors in turbulence. In *Dynamical systems and turbulence, Warwick 1980*, pages 366–381. Springer, 1981.

[20] C. C. Tappert, S. Cha, M. Villani, and R. S. Zack. Keystroke biometric identification and authentication on long-text input. *Int. Journal Information Security and Privacy (IJISP)*, 2010.

[21] A. Wald and J. Wolfowitz. On a test whether two samples are from the same population. *The Annals of Mathematical Statistics*, 11(2):147–162, 1940.