Biometric Hashing for Handwriting: Entropy based Feature Selection and Semantic Fusion

Tobias Scheidat, Claus Vielhauer
Dept. of Computer Science, Univ. of Magdeburg, Universitätsplatz 2, 39106 Magdeburg, Germany

ABSTRACT

Some biometric algorithms lack of the problem of using a great number of features, which were extracted from the raw data. This often results in feature vectors of high dimensionality and thus high computational complexity. However, in many cases subsets of features do not contribute or with only little impact to the correct classification of biometric algorithms. The process of choosing more discriminative features from a given set is commonly referred to as feature selection. In this paper we present a study on feature selection for an existing biometric hash generation algorithm for the handwriting modality, which is based on the strategy of entropy analysis of single components of biometric hash vectors, in order to identify and suppress elements carrying little information. To evaluate the impact of our feature selection scheme to the authentication performance of our biometric algorithm, we present an experimental study based on data of 86 users. Besides discussing common biometric error rates such as Equal Error Rates, we suggest a novel measurement to determine the reproduction rate probability for biometric hashes. Our experiments show that, while the feature set size may be significantly reduced by 45% using our scheme, there are marginal changes both in the results of a verification process as well as in the reproducibility of biometric hashes. Since multi-biometrics is a recent topic, we additionally carry out a first study on a pair wise multi-semantic fusion based on reduced hashes and analyze it by the introduced reproducibility measure.

Keywords: Biometrics, biometric hashing, collision, handwriting, measures, reproducibility, semantic fusion

1. INTRODUCTION

The authentication of information and persons is one important part of the IT security. For user authentication, today there are three main methods: secret knowledge, personal possession and biometrics. While secret knowledge is based on confidential information, which is only known by the authorized user (e.g. password), personal possession methods use private token in possession of the authorized person (e.g. smart card). Common problems of both, knowledge and possession are caused by the fact that the authentication object can be handed over, stolen or get lost, and so it can be possibly used by non-authorized persons. User authentication based on biometrics provides a solution for this problem, because the authentication object is directly linked to the body and/or the behavior of the person. Methods based on passive biometric traits use static information of a part of a person's body (e.g. fingerprint), while active methods are based on dynamic information obtained from an action performed by a person (e.g. handwriting).

The generation of hash values based on biometric input is a recent topic in current biometric research. One goal of the determination of a fix hash value for a biometric trait of one person from its fuzzy input data is to assure either authenticity and integrity or confidentiality and privacy of biometric information. Another aim can be the generation of unique strong keys for cryptographic purposes, since the biometric information of a person is available anytime and anywhere, without the need to remember secret information or to present a special token. A further field of application could be the embedding of biometric hashes into parts of ID documents as digital watermarks. For example, embedding of biometric information (e.g. based on the handwritten signature) into the face image of a travel document, in addition to storage on integrated chips, appears feasible. Here one problem is the limited capacity of the cover medium as described in [1], [2] and [3] and the capacity/transparency tradeoff problem. Consequently, storage of additional information such as biometric data by means of digital watermarking is possible only up to a certain degree without influencing the quality of the cover medium significantly. Thus embedding compact biometric hashes rather than complex raw data may enable alternative storage paradigms for biometric information in documents.

In the literature today, a number of approaches can be found describing methods for the generation and use of biometric based hashes. In the following we present a small selection of publications without neglecting others. In [4] the authors

present a method to calculate a cryptographic key based on a spoken password. Therefore an *n*-dimensional (*n=12*) vector of cepstral coefficients is used as well as an acoustics model, which is speaker dependent. Based on these components segmentation is carried out in order to create different types of features as basis of a so called feature descriptor which can be used as hash value. The biometric hashing method described by Vielhauer et al. in [5] is based on online handwriting biometrics and determines a feature vector of statistical parameters. These parameters are transformed into a hash value space using an interval mapping function, which results in a hash vector as feature vector representation based on an individually statistical model of each user or a given user group, which is determined during enrollment process. In section 2 this method is described in more detail since we used this algorithm for the evaluations in this paper. A method for the generation of reference data in the field of biometric face recognition is described in [6]. Based on the assumption that each element of the feature vector fluctuates by a certain value, a number of Gauss functions can be created for each registered person. The values determined in this way can be used as reference data. The parameters of the transformation are necessary for the authentication process and have to be stored on a capable device (e.g. smart card). In [7] the authors propose a biometric hash generation based on a human's fingerprint where the representation of fingerprint minutia points is based on complex numbers. Symmetric complex functions are used as hash functions, and a corresponding matching method is also proposed. In [8] a function based feature extraction for dynamic signatures is presented, which combines discrete wavelet transform and discrete Fourier transform in order create a compact representation. Using a pseudo random number generator and a special mixing method (so-called BioPhasor) based on user tokens, the biometric information is transformed in a one-way manner. The last steps are a discretization to avoid the reconstruction of the original biometric trait and a Gray encoding to ensure the discriminatory power of the hash.

One problem of biometric systems may be caused by a high number of features used as representation of the biometric trait. In some cases not all of them contribute to the authentication or hash generation process significantly. Here the reason may lie in correlation between individual features or the fact that a feature is not suitable to represent a biometric trait. Feature selection methods try to reduce such a feature set without or with marginal consequences for the applications result. Another case is the use of sensors with different technical characteristics for the same application. For example, in handwriting verification some sensors support features such pen tip pressure or the pen orientation angles azimuth and altitude (e.g. some graphical tablets), while other devices are not able to acquire those data from the writing process (e.g. personal digital assistants). Those features and their derivates should be disregarded to save computation power and runtime resources.

In this paper, we present a feature selection strategy based on the information provided by each element of a biometric hash vector used as feature representation of dynamic handwriting. To determine the information gain provided by an individual feature, the entropy analysis introduced by Shannon in [9] is used. Since one consequence of the feature selection is a reduction of the number of representative elements within the hash vector also a fusion of hashes is proposed to preserve or extend its dimensionality after applying feature selection.

This paper is structured as follows: The next section introduces the Biometric Hash algorithm, which is used in our work for the generation of hashes based on dynamic handwritten data as well as for the verification of biometric handwriting. In the third section, a feature selection strategy is described, which bases on an empirical entropy analysis. The fourth section explains the fusion strategy of combining biometric hashes based on different handwritten semantics. The evaluation database, methodology and the results with regard to biometric error rates and reproducibility of the hashes are described in the fifth section. The last section concludes this paper and gives an overview of future work in this field of biometric research.

2. BIOMETRIC HASHING

The method for the generation of hashes based on dynamic handwriting, the Biometric Hash algorithm, is initially introduced in [5] and enhanced in [10] and [11]. The functionality of the last version of this algorithm is based on *69* statistical features derived from measurements of horizontal and vertical pen position *x(t)* and *y(t)*, pen tip pressure *p(t)* and pen azimuth and altitude *Θ(t)* and *Φ(t)* respectively are taken from a digitizer device.

The Biometric Hash algorithm (see Figure 1) calculates the statistical feature vector containing *k=69* statistical parameters (online and offline features), which are transformed into a hash value space by a so-called interval mapping function, in Figure 1 denoted as key generation. This mapping results in a feature vector representation $\vec{b} = (b_0, ..., b_{k-1})$ supported by a person's or person group's specific statistical model and consisting of an interval matrix (*IM*), which is

obtained during the enrollment process. The left part of Figure 1 shows the five discrete signals, which are taken from the digitizer tablet during verification process. Five signals are used by the key generation module to determine the current hash vector $\vec{b}$, which is compared to a stored reference vector $\vec{b}_{\text{Ref}}$ against some decision threshold value T in the hash authentication module. The authentication is performed by calculation of the Canberra Distance between the vector $\vec{b}$ obtained from the current presented handwriting and the reference vector $\vec{b}_{\text{Ref}}$. Finally, a verification results in a binary True/False decision with respect to the current biometric data and the given threshold T.

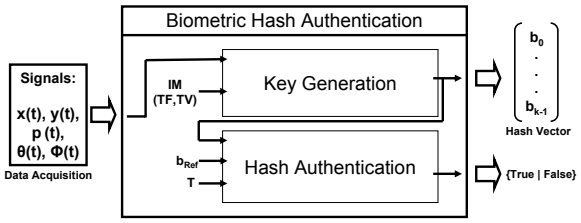


Figure 1. User authentication based on Biometric Hash

For the hash generation, the Biometric Hash algorithm can be used without the Hash Authentication module. In this case the system's output is the hash vector created based on helper data (*IM*) and current data instead of the authentication decision.

For the determination of the *IM*, there are two possibilities to affect the length of the mapping interval: local parameterization by the tolerance vector (*TV*) and global parameterization by the tolerance factor (*TF*). Both parameters are used to scale the individual interval length, which is necessary to map a statistical feature vector element into a hash vector element. Therefore, the *TV* consists of *k* individual values, where each is assigned to the corresponding feature to compute the interval length. The *TF* is one scalar value, which is used for the calculation of each single interval length. For the evaluation described in this paper, the tolerance vector is determined using the statistical features derived from handwriting data of a training set of users (see section 5.1). The tolerance factor is set to *3* for all tests, unless otherwise indicated.

In this paper we study the influence of the entropy based feature selection and multi-semantic fusion on both, verification mode (see section 5.3.2) and hash generation mode (see section 5.3.3).

3. EMPIRICAL ENTROPY BASED FEATURE SELECTION

The hash vector, generated by the Biometric Hash algorithm, has a dimensionality of *69*, based on the used *69* statistical features extracted from the handwriting process. Some of the features may have no or little contribution to the hash generation process. If it is possible to find and suppress those features, which have neither influence on the reproducibility of the hash vectors nor on the verification performance, the calculation complexity and time are likely to be decreased. In order to reduce the dimensionality of the hash vector, we carry out an entropy analysis of the single vector components. This means, we study the information content of each single component of the generated biometric hash vector, based on the information entropy method, which is presented by Shannon in [9].

In [10] an entropy analysis is suggested to estimate the possible value space that can be used by the biometric hash algorithm for dynamic handwriting. The examined algorithm is the same as we are using in our work described in this paper.

Determination of entropy

The entropy *H* introduced by Shannon in [9] describes the degree of information, which can be contained in a given set of data. It can be determined using the following equation:

$$H = -\sum_{i=1}^{n} p_i \log p_i \qquad (1)$$

In this formula, p_i is the occurrence probability of the *i-th* value while *n* is the number of different values observed for an index *z*. To translate this statement to the individual entropy of the biometric hash vector components, we build the following model: There is a set of *m* biometric hash vectors $BH=BH_1,\ldots,BH_m$ with $BH_j=(b_{j,0},\ldots,b_{j,k-1})$ for $j=1,\ldots,m$,

created on a given set of samples from different users on the same semantic class and hash generation parameterization, and containing *k* elements. Then the occurrence probability p_i describes how often a single value occurs in all biometric hash vectors on the same component index *z* (*z=0,...,k-1*). It can be calculated by the ratio of the number of occurrences of the *i-th* value and the sum of occurrences of all values on index *z*. The entropy *H(z)* of the *z-th* element of the biometric hashes is calculated based on the equation (1) and the corresponding probabilities of the values stored on index *z* in the hash vectors.

In order to carry out a feature selection based on the information contents of the single elements of the biometric hash, only those features are taken over into a new reduced feature vector, which have an entropy higher than *0*. To prove the concept of our feature selection method based on entropy analysis, we run verification tests and/or hash reproducibility tests before and after feature selection. We assume that if we take over only the features into the new feature set with entropy higher then *0*, the verification results as well as the hash reproducibility will change only within a minimal magnitude. The corresponding results and the underlying evaluation methodology are presented in section 5.

4. MULTI-SEMANTIC FUSION APPROACH

In this section we present a new biometric fusion strategy based on the pair wise combination of the reduced biometric hash vectors of two semantic classes. In the context of biometric handwriting, semantics are alternative written contents in addition to the signature. Semantics can be based on the additional factors of individuality, creativity and/or secret knowledge, e.g. by using passwords, personal identification numbers or arbitrary symbols. In [11], Vielhauer shows that the usage of such alternative contents may lead to similar results as the usage of the signature in context of online handwriting based authentication performance.

A multi-biometric system bases in general on one of three fusion levels ([12]) depending on the point of fusion within the single biometric components used: feature extraction level, matching score level or decision level. Biometric components can be for example modalities, algorithms or units. The data itself or the extracted features are fused at the *feature extraction level*. At the *matching score level,* the matching scores of all components involved are combined by the multimodal system. In order to parameterize the fusion, matching scores of the different components may be weighted with regard to their individual authentication performance for example.. For a fusion on *decision level* each component involved is processed completely and the individual decisions are fused to a final decision.

Based on the number of biometric components involved in the fusion process Ross and Jain differentiate in [12] between the following four scenarios for automatic biometric fusion: single biometric trait – multiple sensors, single biometric trait – multiple classifiers, single biometric trait – multiple units and multiple biometric traits.

Since the fusion proposed in this paper is executed on the feature extraction level in the hash domain based on hashes of different semantics, it is called *multi-semantic hash fusion*. It can be assigned to the single biometric, multiple units' stage.

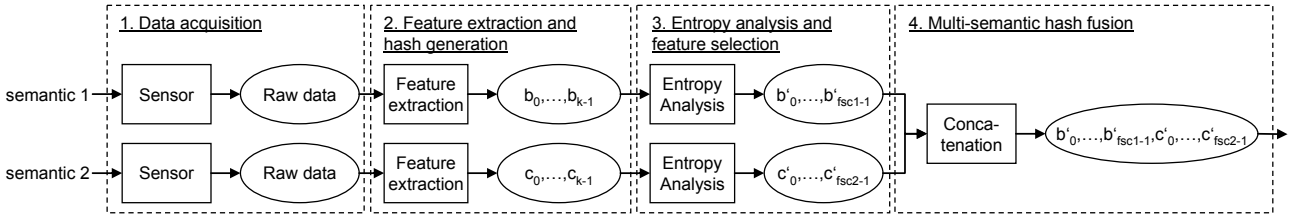


Figure 2. Multi-semantic hash fusion scheme

Figure 2 shows the process steps required for the new multi-semantic hash fusion approach from the presentation of biometric traits up to the output of the fused hash vector. The first step is the data acquisition of two semantics, which are the input for the next step, feature extraction and hash generation. In this process step, a statistical feature vector is calculated from each of the raw data. Then, biometric hash vectors *b* and *c* are derived from the statistical values, where each hash element is based on one corresponding statistical feature element. Thus, for statistical feature vectors and biometric hash vectors, the dimensionality (*k*) is equal. In the next (optional) step the entropy based feature selection is carried out to reduce the given hash vectors *b* and *c* by the information density of its elements. The *fsc* (feature select count) as dimensionality of the fused reduced hash vectors *b'* and *c'* is the sum of their individual dimensionalities *fsc1* and *fsc2*, respectively. The dimensionalities of the reduced hash vectors *b'* and *c'* generated from both semantics (*fsc1* for

semantic 1 with $fsc1 \leq k$, $fsc2$ for semantic 2 with $fsc2 \leq k$) are not necessarily being equal at this stage. The fusion of the two hashes is the last process step, which is carried out as concatenation of both reduced hashes.

5. EVALUATION

This section firstly describes the test data used for evaluation. Following, our methodologies are presented, which are used to study the influence of the reduced feature sets to the biometric handwriting verification as well as to the biometric hash generation. For the latter a new method is introduced to measure the reproducibility and collision of the hashes. Finally, the results for both, verification and hashing are presented and discussed.

5.1 Evaluation database

The entire test set is based on *86* users, which have donated *10* handwriting samples for four different semantics. Semantics are alternative written contents to the conventionally used signature, which is generally associated with biometric handwriting authentication. In our test setup we use the four semantics *PIN*, *Place*, *Pseudonym* and *Symbol*. The *PIN* is given as a sequence of the five digits "77993". Using this semantic the individual kind of writing plays a more important role than the content to recognize a person as its self or distinguish him/her from other users. The semantic *Place* is the individual answer to the question "Where are you from?". This answer includes personal knowledge in a certain degree which, however, is not absolutely secret. Since most of the test subjects refrained from donating their real signature, we use the semantic *Pseudonym* as anonymous substitution of the individual signature. The Pseudonym is a name freely chosen by the writer, which had be trained several times before the acquisition. The freely chosen *Symbol* holds individual creative characteristics and additionally provides a knowledge based component in form of the sketched object (e.g. order of single strokes to create the symbol).

In order to determine the individual entropy of each element of the hash vectors as described in section 3, and to evaluate if there any influence of the suppressed features, a training set (hereafter set *T*) of *17* users and an evaluation set (hereafter set *E*) of *69* users are extracted from the entire set of *86* persons. Both sets are disjoint and structured as follows: From the *10* handwriting samples $S=S_1,\ldots,S_{10}$ of each person and each semantic the first *5* samples $S_1,\ldots,S_5$ are taken to create *5* reference sets, using a leave-one-out strategy. This means a combination of *5* choose 4, i.e. *5* different references ($R=R_1,\ldots,R_5$) are created, containing *4* handwriting samples each. From these references, each is used to create an interval matrix (*IM*) as basis for the biometric hash generation. Based on these interval matrices and the remaining samples $S_6,\ldots,S_{10}$, *5* biometric hashes are created for each user of set *T* and set *E* respectively.

The entropy analysis is executed based on set *T*, which is also used to determine the tolerance vector *TV* based on all users of set *T*. The evaluations in form of a biometric error rate analysis and/or a Hamming Distance based histogram analysis are carried out on set *E*.

5.2 Evaluation methodologies

In this subsection the methodologies are described to study the influence of the suppressed features on the authentication performance based on biometric hashes as well as on the reproducibility of the hashes. As a novel measurement to determine the reproduction ability of a biometric hash generation algorithm, the Hamming Distance based histogram analysis is presented.

5.2.1 Biometric error rate analysis

Due to the fact, that the authentication performance of a biometric system cannot be measured directly, it has to be determined empirically. Basis for this determination are typically biometric error rates: While the FNMR (false non match rate) calculates the ratio between the number of rejected authorized persons and the entire number of authentication attempts, the FMR (false match rate) describes the ratio between accepted non-authorized users and the entire number of authentication attempts. A common measurement in biometric research provides the EER (equal error rate), where FNMR and FMR are identical. It can be used as normalized reference point for comparison in terms of one scalar value of biometric systems.

In this paper, we use the biometric error rates to ensure, that the reduction of the feature set using the entropy analysis has no or only a marginal influence to the verification performance of the biometric handwriting system. However, our

main goal is to improve the reproduction of biometric hashes based on dynamic handwriting. For this reason we analyze this reproduction performance by using the Hamming Distance to compare the reference and current hashes.

5.2.2 Hamming Distance based histogram analysis

One goal of the generation of hashes from biometric input data is to produce a unique reproducible value for a person, which can be used for user authentication or as cryptographic key, for example. The problem here lies in the fact that it is impossible to acquire identical biometric data from a person at different times. Based on the data obtained in this way the hash generation method looks to generate identical hashes from data of the same person and/or different hashes from data of different users. In order to provide a measure for the degree of the reproducibility and/or false generation of such hashes, we suggest the Hamming Distance ([13]) as already shown in [11]. In this context, the Hamming Distance measure determines the number of positions, where two biometric hashes are different and returns a value between *0* and the number of elements. In the optimal case a biometric hash method generates the same value each time a person presents the corresponding biometric trait, and the Hamming Distance will be *0*. In equation (2), x and y are the biometric hash vectors of dimension k we want to compare, and x_i and y_i are the corresponding elements of x and y at index i. The direct comparison of x_i and y_i is *0* if the two elements are equal and *1* else. The Hamming Distance between the hashes x and y is the sum of the results of all single comparisons.

$$hd(x,y)=\sum_{i=0}^{k-1} dist(x_i,y_i)\ with\ dist(x_i,y_i)=\begin{cases}0, if\ x_i=y_i\\1, else\end{cases} \tag{2}$$

Derived from the properties of cryptographic hashes, error rates to estimate the performance of biometric hash algorithms should be considered in the reproduction and the collision in addition to FNMR, FMR and EER, known from biometric verification and identification based on thresholds. In our Hamming Distance based histogram analysis, we compare all generated biometric hashes of each person to each other hash of the same person to calculate the intra-class reproduction rate (hereafter *reproducibility rate*, *RR*). The inter-class generation rate (hereafter *collision rate*, *CR*) is determined by the comparison of a person's biometric hashes with the hashes of all other users. Both rates are logged in two histograms, one for the reproducibility rate and one for the collision rate. In the best case, each comparison between hashes of the same person and semantic should be result in a *0*, while the comparison between hashes of two persons should be greater than *1*.

In order to have an indicator of the trade-off relation between RR and CR an additional measure is introduced here: the *collision reproducibility ratio* (*CRR*). It is the result of the division of CR by RR. Since one aim of biometric hashing is to reproduce hashes of each person with a high degree, while hashes of different persons should be different, the CRR should be very small.

5.3 Results

This subsection describes the outcomes of the entropy based feature selection and the influence of the used semantic on the suppressed features. Also the results are presented for the verification and the hash reproducibility, each determined on the complete feature set and/or the reduced feature set after feature selection. The corresponding tests are carried out on the single semantics as well as on their pair wise fusion.

5.3.1 Feature Selection

As described in section 3 an empirical feature selection is carried out based on an entropy analysis. For each semantic the entropy of the individual elements of the biometric hash vector is calculated to determine its contribution to the hash generation. All elements, which result in an entropy of *0* are suppressed, while all others are taken over in the new reduced feature set.

In Table 1 the results of the entropy analysis are shown for all four semantic classes. It can be seen that the number, the selected features as well as the entropy of selected features differ between the semantics. Table 1 shows in the first column (*Feature Number*) the number of each feature f_j with $j=0,...,k-1$ and $k=69$. The cells of the other four columns, which are related to the four semantics, containing a value higher than *0*, if the corresponding feature is selected, in the other case the cells hold a *0*. As shown in the last row of Table 1, the *fsc* (feature select count) as number of features taken over into the reduced feature set, amounts *33* for semantic PIN, *40* for Place and Pseudonym, and *39* for Symbol.

The rows in the table, that are printed bold and italic, containing those features, which were selected based on the entropy analysis for all four semantic classes in our test scenario. Contrarily to these *19* features, there are *12* features, which were not selected for any of the four semantics (see rows containing *0* and marked with gray background). In our future work for both groups it should be analyzed, whether these are features, which should be selected or suppressed, respectively, in any case. The remaining features should be investigated regarding the influence of the semantic used.

Table 1. Entropy values of each individual feature per semantic class

Feature Number	PIN (77993)	Place	Pseudonym	Symbol
f_0	***0,1247***	***0,3652***	***0,1008***	***0,1149***
f_1	***0,2738***	***0,2203***	***0,2527***	***0,2800***
f_2	0	0	0,0239	0,1008
f_3	0,2203	0,0923	0,0923	0
f_4	0,0923	0	0	0
f_5	***2,7045***	***0,7493***	***1,8195***	***2,3521***
f_6	***1,2569***	***1,4625***	***1,6610***	***0,9826***
f_7	0	0,0239	0,1200	0
f_8	***1,0306***	***1,0729***	***1,9549***	***0,9465***
f_9	0	0,0432	0,0239	0
f_{10}	0	0,1608	0,1008	0
f_{11}	0	0,0923	0,1008	0
f_{12}	0	0	0	0
f_{13}	***0,9887***	***1,3283***	***0,7533***	***0,9505***
f_{14}	***1,1157***	***0,3204***	***0,7488***	***0,3737***
f_{15}	***0,4946***	***0,4140***	***0,2203***	***0,3119***
f_{16}	***1,1072***	***0,6770***	***0,9150***	***0,5226***
f_{17}	0,0239	0,1008	0	0
f_{18}	0	0	0	0
f_{19}	0	0	0	0
f_{20}	0	0	0	0
f_{21}	0	0	0	0
f_{22}	***1,3233***	***1,0697***	***0,9538***	***1,3464***
f_{23}	0	0	0	0
f_{24}	0	0	0	0
f_{25}	0	0,0923	0	0,1690
f_{26}	0,1008	0,0923	0	0,0239
f_{27}	0,0239	0	0	0,2375
f_{28}	0	0,1037	0	0,0239
f_{29}	***0,3228***	***0,2738***	***0,2527***	***0,3035***
f_{30}	0,0923	0,1844	0	0
f_{31}	0	0	0,0718	0,0923
f_{32}	0	0,0239	0,0239	0,0923
f_{33}	***0,0479***	***0,0923***	***0,5743***	***0,0606***
f_{34}	0	0	0,7721	0,1844

Feature Number	PIN (77993)	Place	Pseudonym	Symbol
f_{35}	0,1162	0,0239	0	0
f_{36}	0	0	0	0
f_{37}	0,0239	0	0,2447	0,5068
f_{38}	0	0	0	0,0239
f_{39}	0	0,0239	0,0671	0,3679
f_{40}	***0,0923***	***0,4867***	***0,0239***	***0,0239***
f_{41}	0	0,4239	0,2082	0
f_{42}	0	0,3191	0,1401	0,0769
f_{43}	0,4119	0	0,4825	0
f_{44}	0,3652	0,0923	0,4140	0
f_{45}	0	0,1008	0	0
f_{46}	0	0	0,1102	0,0239
f_{47}	0,0923	0	0	0,2273
f_{48}	0	0	0	0
f_{49}	0,1608	0	0,0432	0
f_{50}	0,0923	0	0	0
f_{51}	0	0,0923	0	0,0432
f_{52}	0	0	0	0,2082
f_{53}	0	0	0	0
f_{54}	0,0239	0	0	0,0923
f_{55}	***1,3867***	***1,1882***	***1,0123***	***1,5966***
f_{56}	0	0	0	0
f_{57}	***0,1348***	***0,0769***	***0,1733***	***0,1211***
f_{58}	0	0,0769	0,0432	0,0923
f_{59}	0	0,0923	0,1149	0
f_{60}	0	0	0	0,0923
f_{61}	***0,4225***	***0,3228***	***0,1037***	***0,4140***
f_{62}	0	0	0,3228	1,0000
f_{63}	0	0	0	0
f_{64}	***2,4016***	***1,9914***	***1,8672***	***1,8992***
f_{65}	***0,0923***	***0,3889***	***0,5723***	***1,0000***
f_{66}	***0,1008***	***0,2328***	***0,4501***	***0,5690***
f_{67}	0	0,0923	0	0
f_{68}	0	0	0,0769	0
fsc	**33**	**40**	**40**	**39**

For a detailed description of the *69* statistical features used to generate the biometric hash, the interested reader is referred to the literature ([5], [10], [11]).

5.3.2 Biometric error rate analysis

In order to prove the idea to use the entropy analysis as basis for a feature selection strategy in a biometric verification system, we compare the verification results of the single semantics and their multi-semantic fusion before and after feature selection to each other. In the case that there is no loss of information by suppression of features with an entropy equal to *0*, the EER should not change significantly after feature selection.

As shown in Table 2, there is no change in the EER values after the feature selection for all four semantics as well as for their fusion. The fusion is based on the matching score level and uses a mean rule, which weights both scores with a value of *0.5* and summates the results to a final fused score. This observation shows that there seems to be no contribution of the suppressed features to the verification process. Note, for this evaluation, we assume that there is no temporal dependence between semantic 1 and semantic 2 (e.g. EER of fusion of PIN and Symbol is equal to EER of

fusion of Symbol and PIN). Thus, the outcome of the fusion is symmetric with respect to the sequence the semantics taken into account, leading to the triangular characteristics of the results presented in Table 2.

As shown in Table 2 the best single verification result with respect to the EER is reached using the semantic Symbol with *EER=0.047*. The second best result is based on semantic Place (*EER=0.058*). Another observation from Table 2 is that all pair wise fusion combinations improve the results determined by the corresponding semantics. Here the lowest EER of *0.022* is calculated based on the combination of Place and Symbol.

Table 2. Equal error rates per semantic class and their pair wise fusion before (EER_{before}) and after (EER_{after}) feature selection

	single		Multi-semantic fusion					
			Symbol		Pseudonym		Place	
Semantic	**EER_{before}**	**EER_{after}**	**EER_{before}**	**EER_{after}**	**EER_{before}**	**EER_{after}**	**EER_{before}**	**EER_{after}**
PIN	0.077	0.077	0.028	0.028	0.051	0.051	0.033	0.033
Place	0.058	0.058	0.022	0.022	0.038	0.038		
Pseudonym	0.094	0.094	0.036	0.036				
Symbol	0.047	0.047	-	-				

5.3.3 Hamming Distance based histogram analysis

Reproducibility of hashes using single semantics

Figure 3, Figure 4 and Table 3 show the results of the Hamming Distance based histogram analysis. In the second and third row of Table 3 the reproducibility rate of genuine hashes by the corresponding genuine users is shown for the complete (row *all*) and the reduced feature set (row *reduced*) in dependency of the semantic class. The fourth and fifth rows are showing the collision rate for the complete and reduced feature sets, while the sixth and seventh rows present the collision reproducibility ratio. The reproducibility rate of genuine hashes increases in all four cases by a minimal magnitude after the feature selection. For example, the highest change is reached by the semantic PIN with a relative improvement of approx. *5%*, while the smallest change is caused by semantic Symbol with approx. *3.2%* (see Table 3). On the other side, after feature selection the collision rates for all four semantics degrade compared to those before feature selection. Here the highest decline with regard to the generation of genuine hashes by random imposters amounts approx. *49%* based on semantic Symbol. The smallest degradation of *24%* is calculated for the semantic Place (see Table 3).

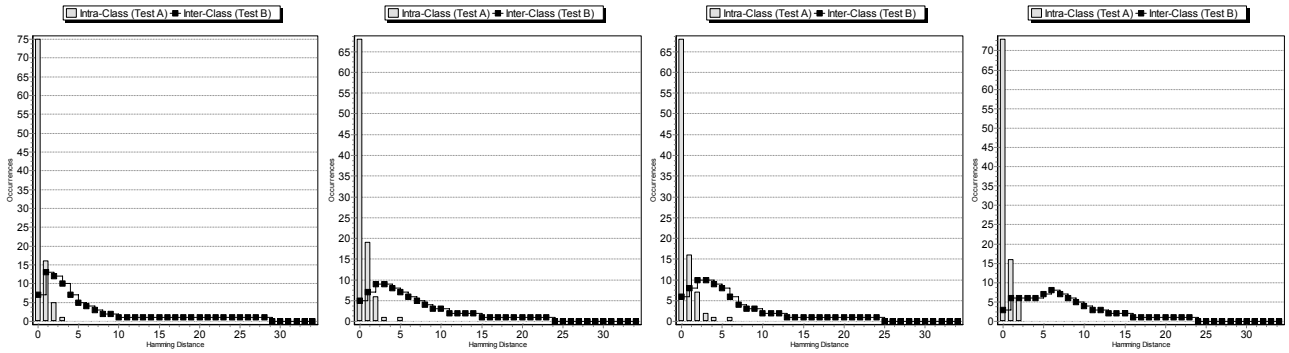
Table 3. Reproducibility and collision rate for complete and reduced feature sets per semantic class

Measurement	**Features**	**PIN**	**Place**	**Pseudonym**	**Symbol**
RR	**all**	75.94	68.46	68.35	73.91
	reduced	79.77	71.42	70.67	76.29
CR	**all**	7.81	5.74	6.13	3.87
	reduced	10.71	7.12	8.96	5.75
CRR	**all**	0.103	0.084	0.090	0.052
	reduced	0.134	0.100	0.127	0.075

The results show that there is a dependency between the reproducibility rate and/or collision rate and the features used, even if some of them have zero intra-class entropy. Thus, the entropy analysis based feature selection leads to an improvement of the reproducibility of genuine hashes, which must be paid with a higher false generation of hashes produced by other users.

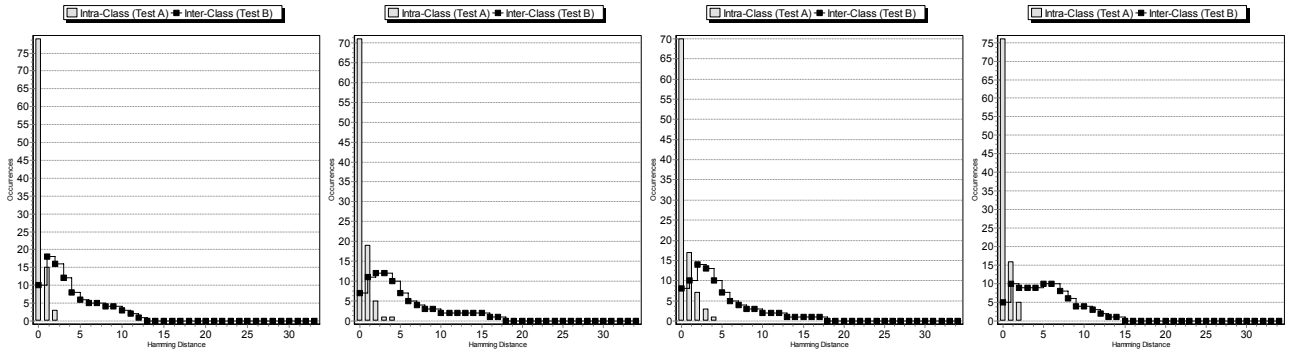
In order to try to improve the reproducibility rate further more, we have performed additional test with an increasing tolerance factor (*TF*) as parameter of the Biometric Hash algorithm. The *TF* is a scalar value, which is used to scale the width of the mapping interval during the hash generation. The results presented in Table 3 are determined using a *TF* of *3*. Figure 5 shows the reproducibility rate (upper lines) and collision rates (lower lines) in dependency on the *TF* before and after feature selection. The incrementing of the *TF* leads in the first two steps (*TF=7*, *TF=10*) to an improvement of the reproducibility rate. Above a *TF* of *10* the values are changing only in small magnitude and then it seems that they are converging to a certain limit by reaching saturation. This observation can be made for all four semantics, but with varying limits. This effect of saturation cannot be observed for the analysis of the biometric error rates (FNMR/FMR/EER) in our experiments. Further, the graphs for the collision rate are increasing faster than the

corresponding lines of the reproducibility rate for increasing values for TF. The slope declines with increasing TF, however, saturation is not reached at *TF=40*.



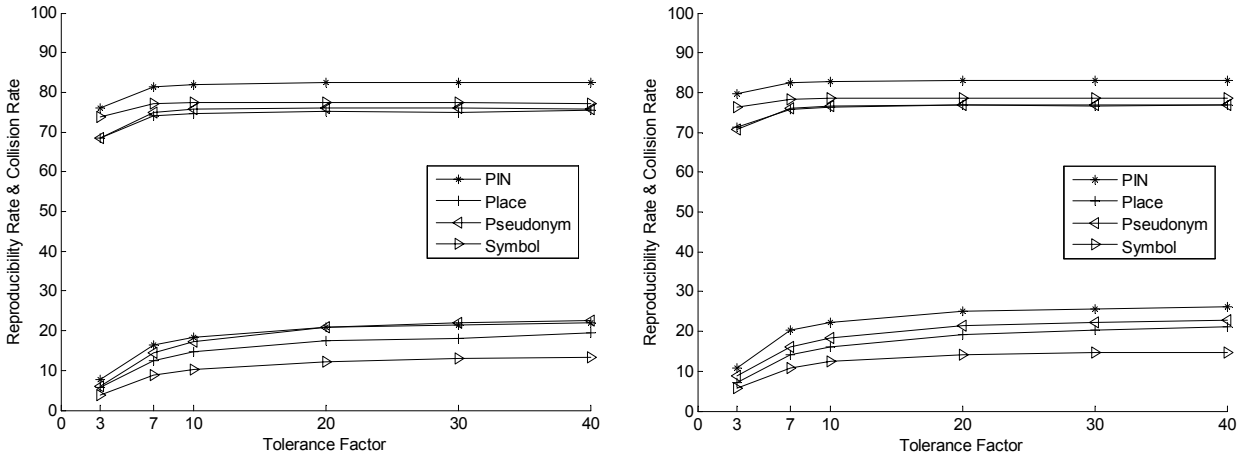
(a) PIN (b) Place (c) Pseudonym (d) Symbol

Figure 3. Reproducibility rate (bars) and collision rate (line) before feature selection



(a) PIN (b) Place (c) Pseudonym (d) Symbol

Figure 4. Reproducibility rate (bars) and collision rate (line) after feature selection



(a) (b)

Figure 5. Reproducibility rates (upper graphs) and collision rates (lower graphs) in dependency on the tolerance factor before (a) and after (b) feature selection

Reproducibility of multi-semantic fused hashes

The concatenation of two different hashes generates a new hash with a higher dimensionality based on two existing hashes. Thus, the reproducibility of the new hash depends on the reproducibility of the both hashes involved. Based on this assumption, we can state that the reproducibility of fused hashes cannot be better than the worst reproducibility rate of each of the hashes used for the fusion. For the evaluation in this paper, the order of concatenation is not taken in consideration. This means, the results are identically for RR, CR and CRR, respectively, independently whether semantic 2 is concatenated to semantic 1 or vice versa.

Please note that the length of the fused biometric hash vector depends on the length of the two single hash vectors of the semantics involved. Thus, in some cases it is not possible to generate hashes with the same length from each pair wise concatenation of the semantics. One solution could be to fill the hashes with values generated by other mechanisms, e.g. based on the written content or the id of the user.

Table 4. Reproducibility and collision rate for complete and reduced feature sets for pair wise semantic hash fusion

Semantic 1	Measurement	Features	Semantic 2		
			Symbol	Pseudonym	Place
PIN	RR	all	57.04	55.13	54.32
		reduced	63.59	61.22	59.59
	CR	all	1.00	0.88	1.53
		reduced	1.59	2.29	2.70
	CRR	all	0.017	0.016	0.028
		reduced	0.025	0.037	0.045
Place	RR	all	52.64	48.06	
		reduced	56.00	50.84	
	CR	all	0.75	0.98	
		reduced	1.27	1.70	
	CRR	all	0.014	0.020	
		reduced	0.023	0.033	
Pseudonym	RR	all	52.99		
		reduced	56.23		
	CR	all	0.50		
		reduced	0.83		
	CRR	all	0.009		
		reduced	0.015		

Table 4 shows the results of the pair wise multi-semantic hash fusion. The intersections of rows and columns of the different semantics are showing the corresponding fusion results for the reproducibility rate (*RR*), collision rate (*CR*) and collision reproducibility ratio (*CRR*) and the two feature sets (*Features – reduced/all*) each. As assumed, the first observation is, that the fusion results for the reproducibility rate are worse than the results obtained based on the single semantics. For example, for the fusion of PIN and Symbol RRs of *57.04%* and/or *63.59%* for the complete and reduced, respectively, feature set were calculated. This corresponds to a relative degradation of approx. *28%/25%* in comparison to the best single result determined for the PIN (*RR=75.94%* before, *79.77%* after feature selection). Further, the collision rates are significantly lower than those of the single semantics involved. Here the relative decline lies between *75%* and *92%*. The best CR of *0.50%* was determined for the fusion of the semantics Pseudonym and Symbol based on the complete feature set, while the corresponding RR amounts *92.55%*. The change of the CRs before and after the feature selection is smaller for the fused semantics than for the single semantics, which results in a smaller CRR compared to the single semantic based hashes.

6. CONCLUSIONS

In this paper, we present a strategy to identify features without any or little influence to the generation of biometric hashes, which is based on entropy analysis introduced by Shannon in [9]. Based on this strategy, a feature selection is carried out. Therefore, those features are taken over into a new reduced feature vector, which have an entropy higher than *0*. As a new measure we introduce the analysis of the biometric hash reproducibility rate based on the Hamming Distance. The reproducibility rate (RR) shows, how is the performance of a hash generation algorithm with respect to generate the same hash for the same person and the same written content. The collision rate (CR) is a measure for the generation of genuine biometric hashes by non-authorized users. The collision reproducibility ratio (CRR), as third introduced measure, indicates the relation between CR and RR. In order to find a suitable working point for a biometric

hash generation algorithm, one solution can be to minimize the CRR. Further, we have suggested a novel concept in the domain of multi-biometrics: Multi-semantic fusion of biometric hashes generated using different writing contents. Using an analysis based on the biometric error rates FNMR, FMR and EER, we have shown that there is no degradation in the recognition accuracy before and after feature selection for single semantics as well as for the pair wise multi-semantic fusion.

The experimental study of the influence of a reduced feature set has shown that the RR increases, if the number of features becomes smaller. This means, that the reproduction of genuine hashes becomes better by suppressing such features, which have an entropy equal to *0*. However, also the CR rises after feature selection. Thus, the probability of falsely generated genuine hashes by non-authorized users gets higher. This leads also to a higher CRR. If the parameterization of the hash generation is changed by using another tolerance factor (*TF*), an increasing can be observed for both, RR and CR. But, while a saturation appears for the RR at *TF=20*, the CR rises up to the maximum tolerance factor (*TF=40*) studied. Based on this observation it seems that there is no possibility to increase the RR further more by incrementing the *TF*. Thus, in order to improve the RR even more, other methods have to be studied, e.g. error correction mechanisms. In this case, one has also to keep track of the expansion of the CR as counterpart of the RR.

For potential applications in ID documents, due to limitations in the embedding capacity, the size of the hash as watermarking payload should be minimized. For handwriting based biometric hashes in this context the following important observations can be summarized: The entropy based feature selection reduced the size of the biometric hashes determined for the four semantics significantly. The number of hash elements deflates from *69* to *33* (*52.17%*) for PIN, to *40* (*42.03%*) for Place and Pseudonym and to *39* (*43.48%*) for Symbol. However, as shown in our biometric error rate analysis in section 5.2.1 the verification performance remains identical before and after the feature selection. For example, the best single result is determined for the individual Symbol with an EER of *0.047* before as well as after the feature selection. Thus, the size of the biometric hash as a potential watermarking payload was reduced by approx. 50% without any influence to the verification performance.

One aim of our future work is to enlarge the evaluation database to achieve more representative test results. To improve the reproducibility even more, also the parameterization can be adjusted to any user registered in the database by determining user specific tolerance vectors, which are used to calculate the mapping interval of the Biometric Hash algorithm.

Regarding the entropy based feature selection we are currently working on an enhancement of the selection strategy: In order to take the problems caused by intra-class variability and inter-class similarity into consideration, three entropy analysis steps are carried out sequently. The first calculates the entropy of the hash elements for each user separately, to find those features with a marginal variance, which will be taken over in the reduced feature set. In the second step, those features are sorted out which has no entropy over the biometric hashes of all genuine users, since they have no contribution to the discrimination between them. The last step will suppress such features, which have zero entropy in the comparison of genuine biometric hashes and imposter biometric hashes (random or skilled).

ACKNOWLEDGEMENTS

The work on biometric hashes with regard to verification and reproducibility is partly supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation, project *WritingPrint*). The work in this paper has been supported in part by the European Commission through the IST Programme under Contract IST-2002-507932 ECRYPT for aspects of application of biometric hashes as watermarking payload. The content of this publication is the sole responsibility of the University Magdeburg and their co-authors and can in no way be taken to reflect the views of the European Union.

We would particularly like to thank Prof. Jana Dittmann for her support and fruitful discussions in the context of our work.

REFERENCES

[1] Vielhauer, C., Steinmetz, R., "Approaches to biometric watermarks for owner authentification", Proc. SPIE - Security and Watermarking of Multimedia Contents III, Vol. 4314, 209-219 (2001).

[2] Picard, J., Vielhauer, C., Torwirth, N., "Towards Fraud-Proof ID Documents using Multiple Data Hiding Technologies and Biometrics", Proc. SPIE - Electronic Imaging, Security and Watermarking of Multimedia Contents VI, Vol. 5306, 416 – 427 (2004).

[3] Vielhauer, C., Kalker, T., "Security for Biometric Data", Proc. SPIE - Security and Watermarking of Multimedia Contents VI, Vol. 5306, 642 – 652 (2004).

[4] Monrose, F., Reiter, M. K., Li, Q., Wetzel, S., "Using Voice to Generate Cryptographic Keys", A Speaker Odyssey, The Speech Recognition Workshop, Crete, Greece (2001).

[5] Vielhauer, C., Steinmetz, R., Mayerhöfer, A., "Biometric Hash based on Statistical Features of Online Signature", Proc. of the Intern. Conf. on Pattern Recognition (ICPR), Quebec City, Canada, Vol. 1 (2002).

[6] Sutcu, Y., Sencar, H. T., Memon, N., "A Secure Biometric Authentication Scheme Based on Robust Hashing", Proc. of the 7th workshop on Multimedia and security, ACM Press, New York, U.S.A. (2005).

[7] Tulyakov, S., Farooq, F., Govindaraju, V., "Symmetric Hash Functions for Fingerprint Minutiae", International Workshop on Pattern Recognition for Crime Prevention, Security and Surveillance, Bath, UK (2005).

[8] Kuan, Y.W., Teoh, A.B.J., Ngo, D.C.L., "Secure Hashing of Dynamic Hand Signatures Using Wavelet-Fourier Compression with BioPhasor Mixing and 2N Discretization", EURASIP Journal on Advances in Signal Processing, vol. 2007, Article ID 59125 (2007).

[9] Shannon, C. E., "A mathematical theory of communication", Bell System Tech. J. 27, 379-423, 623-656 (1948).

[10] Vielhauer, C., Steinmetz, R., "Handwriting: Feature Correlation Analysis for Biometric Hashes", In: Bourlard, H., Pitas, I., Lam, K., Wang, Y. (Eds.), EURASIP Journal on Applied Signal Processing, Special Issue on Biometric Signal Processing, Hindawi Publishing Corporation, Sylvania, OH, U.S.A. (2004).

[11] Vielhauer, C., [Biometric User Authentication for IT Security: From Fundamentals to Handwriting], Springer, New York (2006).

[12] Ross, A., Jain, A.K., "Multimodal biometrics: an overview", Proc. 12th European Signal Processing Conference (EUSIPCO), Vienna, Austria, 1221–1224 (2004).

[13] Hamming, R.W., "Error-detecting and error-correcting codes", Bell System Technical Journal XXVI (2): 147-160 (1950).