Assessment of the Quality of Handwritten Signatures Based on Multiple Correlations

Richard Guest
University of Kent
School of Engineering, Canterbury, UK
r.m.guest@kent.ac.uk

Olaf Henniger
Fraunhofer IGD
Darmstadt, Germany
olaf.henniger@igd.fraunhofer.de

Abstract

Assuring the quality of individual biometric samples is important for maintaining the discriminatory power of biometric recognition systems as biometric data of low-quality are likely to be mismatched. This paper presents an investigation into the assessment of the quality of handwritten signatures, predicting the performance or 'utility' of individual signature samples in automated biometric recognition. The prediction of utility is based on multiple correlations with static and dynamic signature features. First, the utility of handwritten signature samples from publicly available databases is assessed by comparing them with each other using commercial automatic signature verification engines. The samples are classified into four quality bins (excellent, adequate, marginal, and unacceptable quality) with totally ordered bin boundaries. Then, the correlation of multiple static and dynamic signature features with utility is analysed to find features that can be used for predicting the utility of samples. Our results show that it is possible to predict the utility of handwritten signature samples using a multi-feature vector.

1. Introduction

Not all handwritten signature samples donated to an automatic signature verification system are equally well suited for the automated recognition of the persons from whom they are acquired (see e.g. [1]). The 'utility' of a biometric sample, i.e. the usefulness of the sample for telling genuine and forged samples apart, can be expressed by a *quality score* assigned to that sample. The quality score of biometric samples can be used, for instance, for deciding whether the re-acquisition of data is deemed necessary [2] or for weighting partial results in multi-biometric systems. The question of how to objectively predict the utility of handwritten signature samples is the topic of this paper.

The importance of a sample-based quality score across all biometric modalities is evidenced by the fact that fields for holding biometric sample quality scores have been introduced into several ISO/IEC biometric data structures [for example 3] thereby indicating the community's and industry's desire to store quality information within sample data formats. In these data formats if a biometric sample quality score is reported, valid values are integers between 1 and 100 though some quality assessment algorithms such as [6] provide less than 100 valid quality values. Quality scores in the range 1–25 are to indicate unacceptable quality, in the range 26–50 marginal quality, in the range 51–75 adequate quality, and quality scores in the range 76–100 represent excellent quality.

Related work on open-access public-domain algorithms for predicting the utility of biometric samples has concentrated on image-based biometric modalities such as finger images [6] and iris images [9]. For handwritten signatures, there are proprietary algorithms in use for assessing sample complexity during the enrolment process. Brault and Plamondon [10] have already introduced a coefficient of difficulty of imitation of handwritten signatures. Müller and Henniger [11] reported a correlation investigation between a number of signature features and utility scores estimated by comparing signature samples from a dataset with each other. Among the individual features that were assessed there were a number that allow a faint prediction of a signature's production stability, but none that allow a prediction of a signature's forgeability. Stronger correlations with utility may be found by looking at feature vectors instead of individual features, as it has been done in [6] for the quality of finger images. Other studies [12] have utilised methods for the assessment of signature quality based on a forensic 'attack' assessment model and explored the effect on quality within the synthetic generation of signatures. This paper extends [11] by assessing a wider range of character features and exploring a multiple-feature correlation with utility.

2. Measures of quality

The quality of an individual biometric sample can be expressed through three separate elements: the *character*, the *fidelity*, and the *utility* of a sample [14]. Below is a description of each of the quality measures, and how it relates to handwritten signatures.

Utility refers to the usefulness of an individual sample for telling genuine and forged samples apart. The utility of a genuine sample is high if the sample-specific false match rate and false non-match rate both are low. Conversely, utility is low if the sample-specific false match rate and false non-match rate both are high. The sample-specific false match rate is the proportion of forgeries for a specific genuine sample falsely declared to match this sample. The sample-specific false non-match rate is the proportion of genuine samples from the same source as the specific genuine sample falsely declared not to match this sample. Due to differences in feature extraction and comparison algorithms, the utility of a specific sample depends on the particular comparator used. The utility of an individual sample can be estimated *a-posteriori* by comparing it with the corresponding genuine and forged samples from a database. Therefore, utility also depends on the underlying dataset.

Character refers to measurable characteristics (or features) from a particular sample. A signature with good character would be one that is a good representation of the typical signature from a user (i.e. signature characteristics showing low intra-individual variability) and that enables a comparator to distinguish between different signers (signature characteristics showing high inter-individual variability and difficult to forge). A signature could be described as having poor character if the sample is too simple, resulting in low inter-individual variability or being easy to forge, or if an outside factor affected the signing process, producing an atypical signature for the signer in question. This, of course, is also influenced by the natural intra-signature variability within a subject's samples [15]. An example of this may be if the signer produced an erroneous movement within the signing process, for instance by jolting the hand and introducing an anomalous writing stroke.

Fidelity refers to the degree of similarity of the captured signature to the original signing process. The majority of factors influencing fidelity are related to issues of the data capture apparatus. In the case of a signature sample, high fidelity would be noted if the sample contains a high level of detail relating to the motion and position of the pen during the signing process. Fidelity may be affected by a number of factors in a dynamic signature capture environment: low spatial resolution within the capture device, inadequate capture device sensitivity to pressure, capture device sampling rate too low, electronic noise and jitter or a non-linear reporting characteristics of the capture device. With a static image capture environment (i.e. an image produced on paper and scanned) fidelity can also be affected in a number of ways: ink used not identified successfully by the optical reader, resolution of the scanner too low resulting in aliasing of an image or background noise is also recorded as part of the signature.

3. Datasets

Two publicly available signature datasets were used in the analysis:

i) Dataset Two of the SVC (Signature Verification Competition) 2004 [17]: consists of 40 sets of signature data each from a different test subject. For each test subject the set contains 20 genuine signatures and 20 forged signatures produced by five other test subjects. The dataset therefore totals 800 genuine and 800 forged signatures. Note that the 'genuine' signatures were devised by each test subject specifically for the data collection. As they were not the signatures that the test subjects were familiar with for everyday purposes, they have a higher intra-individual variability than normal signatures do.

ii) Signature subset of the MCYT dataset [18]: The data consists of 100 sets of signature data each from a different test subject. For each test subject the set contains 25 genuine signatures and 25 forged signatures produced by five other test subjects. The dataset therefore totals 2500 genuine and 2500 forged signatures. To create the forgeries, the potential impostors had the original signatures available on paper and were allowed to practice the signatures, to look at the original while forging, and even to retrace the original. Unlike the SVC database, the MCYT genuine signatures were real signatures used in day-to-day transactions. Samples were collected from test subjects over multiple time-separated sessions, with the aim of capturing the natural intra-individual variance present in signatures.

The data stored for each signature consists in both datasets of a time series of data vectors containing information on pen coordinates, pressure, pen orientation and pen tip status (pen on or off the writing surface). We divide each dataset randomly into equally-sized disjoint training and testing subsets, leaving the subsets of signatures for the same test subject undivided. The training dataset is used to form ordinal regression models using selected feature data to predict utility. The testing data is used to examine whether the utility prediction models are unbiased**.** As both datasets contain samples captured under homogeneous conditions, a correlation between fidelity and utility cannot be calculated and therefore will not be considered in this study. The influence of fidelity could be observed by using samples captured with different devices and different temporal and spatial sampling rates. An influence of different capture devices on the utility of signature samples has been observed before [19] and has led to the specification of best practices in data acquisition [20].

4. A-posteriori assessment of sample quality

The a-posteriori assessment estimates the utility of a genuine sample by comparing it with all other genuine and forged samples for the same signer. We utilise a four-compartment binning approach to sample-utility assignment. The obtained a-posteriori utility values form the basis for building a sample-quality prediction model later in Section 5.

Two signature verification SDKs were used to calculate the utility of signature samples. For the purposes of this study, the comparators are treated as ‘black boxes’ as we do not target at a comparative technology evaluation, but at a generalised sample quality assessment. Both systems are off-the-shelf commercial signature comparators that are deployed in real-life applications. In the following they are referred to as comparator 1 and comparator 2. As both systems are commercial engines not more is known about them than that they both make use of both dynamic and static features extracted from each signature. Both signature comparators return a continuous *similarity score* (ranging from 0 to 100) alongside a binary match/non-match decision about a template and the probe sample.

Each of the two comparators forms an enrolment template using three reference signatures, against which a probe signature is to be compared. In conducting the experiments we deviate slightly from general usage in that, alongside comparing the probe signature to a template formed by three genuine signatures, we assess forgery performance by comparing against a template formed by three skilled forgeries. In this way, we can compare an individual genuine signature (used as probe signature in both cases) against both genuine and forged signatures.

For each of the comparators, the following method was used for each genuine signature:

1. Template formation using three other genuine and, separately, three forged signatures of the same signer. Successful formation of a template was assessed using the default template formation score validation setting within the comparator.
2. Successfully formed templates were used to verify the genuine signature using the comparator’s default verification threshold. The similarity score was noted for use in the utility prediction.

Stages 1 and 2 described above were repeated for the formation of a total of six genuine and six forged templates for each genuine signature in the SVC 2004 dataset, using 3 x 6 = 18 of the 19 other genuine signatures and 18 of the 20 forged signatures available, and eight genuine and eight forged templates for each genuine signature in the public subset of the MCYT dataset, using 3 x 8 = 24 of the 24 other genuine signatures and 24 of the 25 forged signatures available. All templates were formed using disjoint triplets of reference signatures.

4.1. Binning utility assessment method

The similarity scores from the two comparators are used to assign a utility score to each genuine sample. The utility scores increase with increasing quality. We follow an approach similar to that outlined in [6, 14] adapted to the available datasets and comparators and, in contrast to [6, 14] enforcing totally ordered bin boundaries:

I. For each of the available comparators c_h, $1 \le h \le n_c$ ($n_c = 2$), for each of the datasets D_i, $1 \le i \le n_D$ ($n_D = 2$), and for each genuine signature $j \in D_i$:
   1. Compare j with the n_{gj} templates formed from genuine signatures of the same person to produce n_{gj} genuine similarity scores s_{gjk}, $1 \le k \le n_{gj}$.
   2. Compare j with the n_{fj} templates formed from forged signatures of the same person. The results are n_{fj} impostor similarity scores s_{fjk}, $1 \le k \le n_{fj}$.
   3. Insert j into set T if all its genuine similarity scores are larger than all its impostor similarity scores.
   4. Compute the arithmetic mean $\overline{s_{gj}}$ and standard deviation σ_{gj} of j’s n_{gj} genuine similarity scores and the arithmetic mean $\overline{s_{fj}}$ and standard deviation σ_{fj} of j’s n_{fj} impostor similarity scores.
   5. Compute the normalised difference between the mean of genuine similarity scores and the mean of impostor similarity scores:

$$z_j = \frac{\overline{s_{gj}} - \overline{s_{fj}}}{\sqrt{\frac{\sigma_{gj}^2}{n_{gj}} + \frac{\sigma_{fj}^2}{n_{fj}}}} \tag{1}$$

II. For each of the available comparators c_h, $1 \le h \le n_c$ ($n_c = 2$), for each of the datasets D_i, $1 \le i \le n_D$ ($n_D = 2$), and for each genuine signature $j \in D_i$, bin j into one of L bins based on membership to T and z_j. A strategy for $L = 4$ is shown in Table 1. Use the bin number assigned to a signature j as its utility score q_{jh} regarding comparator c_h.

III. The rounded-up arithmetic mean of j’s utility scores is j’s n_c overall utility score $\overline{q_j}$:

$$\overline{q_j} = \frac{\sum_{h=1}^{n_c} q_{jh}}{n_c} \tag{2}$$

4.2. A-posteriori results

The 3300 genuine signatures within the two datasets were binned as shown in Table 2. The majority of signatures are not binned in broad agreement by the two comparators. For almost two thirds of the signatures, the utility bins assigned by the two comparators differ. This may be due to differences in the accuracy of the comparators or due to the fact that the comparators use different feature sets for comparison and that some signatures are more readily verifiable based on this or that feature set. This leads to the conclusion that a specialised quality prediction model may be constructed for each comparator under consideration. However, it would be nice to construct a generic signature quality prediction model independent of particular comparators. A rounded-up average utility across the two comparators is computed as an overall utility score. Table 2 also shows the binning when the overall utility score is used.

While in the MCYT dataset the overall utility of less than 4% of the signatures is unacceptable or marginal, in the SVC 2004 dataset this is the case for more than 23% of the signatures. This shows that the quality of the devised signatures is lower than that of real signatures (see Section 3).

5. A-priori assessment of sample quality

For the prediction of a sample's quality, we establish correlations between multiple signature features and the overall utility scores. A vector of 69 commonly used global features was extracted from each signature. These features can be divided into two sub-groups: static features, relating to the drawn outcome of the signature, and dynamic features, relating to temporal measurements from the signature production such as pen velocity and execution time. For the extraction of the static features, signature images have been recreated from the time-series data. Further information about these features can be found in [15].

The objective of a first experiment was to investigate whether individual features can be used to predict the utility of a sample. To explore the relationships between the individual features and the overall utility scores described in Section 4.1, correlation coefficients were calculated (the overall utility score for each signature was used rather than the dataset specific scores to assess the generic nature of the quality binning calculation). A significant strong correlation between the overall utility score and a feature indicates a potential for the individual feature to reveal quality information about signature samples. Correlations were calculated across the two signature comparators and the two datasets, resulting in four correlation coefficients for each feature. Each correlation was tested for statistical significance.

Observing the results from the features values to utility score correlations, it was obvious that none of the individual features were strongly correlated (the highest with a correlation of 0.18239 (number of pixels enclosed within loops). It is interesting to note, however, that there is a proportional mixture of static and dynamic features indicating the importance of assessing constructional elements of signature formation alongside static shape and image-based measurement in the prediction of signature utility. Observing the mean correlations across all features for each of the various dataset and feature type groupings within the investigation reveals negligible differences, showing the generality of the technique.

5.1. Correlation and modelling between multiple signature features and utility

A number of features have shown individually significant classification performance albeit with a low correlation; however, jointly used, they provide a more accurate classification result. Whether the correlation between multiple signature features and utility is satisfactory is investigated in this section.

Let the features $x_1, \ldots, x_n$, which have been calculated for all available genuine signatures, be random potential "predictor variables". There are four ordinal quality classes Q_i in this study: the bin numbers or utility scores 1, 2, 3, 4. The classification task here is to determine which class is the most likely class of a sample based on the extracted features. A multiple regression equation for predicting the utility q can be expressed as follows:

$$f(q) = a + b_1 x_1 + \ldots + b_n x_n \tag{3}$$

The values for a and b_i, $1 \leq i \leq n$, are determined in such a way that the probability that the utility class is correctly predicted from the observed values of the extracted features is maximised over all training data. Such criterion is referred to as Maximum Likelihood Estimation in the literature [16]. The left side of equation (3), i.e. the "criterion variable", is the multinomial logit [16] function of the predicted utility q. The logit function of utility score Q_i is the natural logarithm of the odds of Q_i, i.e. the ratio of the probabilities for and against Q_i. The choice of modelling the logit function is preferable to modelling q itself, because q has an irregular distribution, while the logit of q follows a chi-square distribution.

Utility prediction models were formed in SPSS using ordinal regression on the training sets. This process was undertaken in two stages for each model. Firstly all extracted features were entered into a modelling process and the output model noted in terms of each feature's contribution to the output function. In the second stage only features that were flagged as making a significant contribution to the model ($p < 0.05$) from the first stage were included in an ordinal regression process. The output model from the second stage formed the final model for a particular training set. To assign the ground-truth utility class membership to each sample (in both training and testing datasets), the overall utility (i.e. the rounded-up average of the utility scores obtained with the two different comparators) was used.

5.2. A-priori modelling results

To assess the accuracy of each utility prediction model we divide each of the datasets into equally-sized, disjoint training and testing subsets and analyse the percentage of correctly identified utility class membership cases in the testing sets. We also assess the performance of optimally formed models from one dataset on samples from the other dataset, thereby assessing the dataset-independence of a model. For assessing the performance of a general dataset-independent model, we form disjoint training and testing sets using an equal number of samples from each of the two datasets within each set. In this way we can assess the performance of a general dataset-independent model.

Table 3 details the results of the utility prediction experiments arranged in accuracy order. The results show that the lowest prediction accuracy is achieved when the training and testing data come from different datasets. This confirms the expectation that the predicted value depends on the properties of the dataset used for training the predictor. There are large differences between signature styles in the SVC 2004 and MCYT datasets: While the SVC 2004 dataset contains a great portion of Chinese characters, the MCYT dataset contains predominantly Spanish signatures. The properties of the dataset used for training the predictor should as much as possible reflect the properties of the data whose quality is to be predicted.

We have also explored the ‘within-one binning accuracy’ – that the utility prediction model predicts a binning classification within +/- one bin accuracy. Table 6 shows a good ‘within-one binning accuracy’ as long as similar data are used for training and testing.

The following nine features were selected by the regression method with a mix between dynamic (four) and static (five) features contributing to the model, again highlighting the contribution of including both types of information in an assessment: travel distance of pen on paper, maximum *x* acceleration, vertical median midpoint, (vertical median midpoint – y_{min}) / (y_{max} – y_{min}), *x* maximum jerk, *y* pixel centroid (mean pen-down *y* pixel position), *x* direction minimum velocity / *x* direction average velocity and moment p=0, q=1 on 'negative' image. These features may be considered in forming a generic quality prediction model.

6. Conclusions and discussion

In this paper we have presented our findings from experiments to predict the utility of signature samples within automatic signature verification systems based on characteristic feature vectors of those signature samples. Our results show that it is possible to predict the utility of a sample using a multi-feature vector as a predictor. The accuracy of the utility prediction depends on the training data used for building the utility prediction model. Our experiments show that the best results are obtained when a utility prediction system is tuned to a specific dataset and comparator. However, we have also shown that a generic model can be constructed which performs similarly to single dataset models.

The more similar the training data is to real-life data, the more accurate the quality prediction will be. There may be a gap between skilled forgeries in signature datasets collected for research purposes and those produced in practice, where forgers may be well motivated and skilled.

The study shows that the proposed technique can have direct and transparent application within a practical automatic signature verification implementation using both common static and dynamic features to predict sample quality. Of course, the features that we used as predictor variables, although commonly used within the community, may not be optimal in terms of quality prediction accuracy. For example, local features applied to specific regions of interest within the signature may yield improved correlation. Nevertheless, this study has provided an insight into the features which may be practically used to provide foresight information on sample quality and performance within an automatic signature verification system. Further work will continue to expand the range of features explored, judicious selection of character features in an a-posteriori distance metric and further develop the multiple-regression modelling prototype study to produce a normalised quality metric.

Table 3: Binning accuracy

Training	**Testing**	**Test Samples**	**Binning Accuracy**	**Within-1 Binning Accuracy**
MCYT (train)	MCYT (test)	1250	68.72%	97.84%
Combined (train)	Combined (test)	800	55.75%	97.75%
SVC (train)	SVC (test)	400	55.00%	97.25%
SVC (all)	MCYT (all)	2500	30.28%	75.40%
MCYT (all)	SVC (all)	800	20.50%	63.37%

7. Acknowledgements

The authors gratefully acknowledge the support of the Leverhulme Trust in the production of this work. The authors are also grateful to J. Ortega-Garcia and J. Fiérrez-Aguilar for making a subcorpus of the MCYT database available for research purposes.

References

[1] O. Henniger, S. Müller. Handwritten signature on-card matching performance testing. In J. Fierrez, J. Ortega-Garcia, A. Esposito, A. Drygajlo, M. Faundez-Zanuy, eds., *Proc. of the Internat. Conf. on Biometric ID Management and Multimodal Communication*, Madrid, Spain, 2009. Springer (LNCS vol. 5707)

[2] J. Fàbregas, M. Faundez-Zanuy. On-line signature verification system with failure to enrol management, *Pattern Recognition*, 2009, vol. 42, pp. 2117–2126.

[3] Information technology – Biometric data interchange formats. Part 1: Framework. Internat. Standard ISO/IEC 19794-1, 2011.

[4] F. Alonso-Fernandez, J. Fierrez, J. Ortega-Garcia, J. Gonzalez-Rodriguez, H. Fronthaler, K. Kollreider, J. Bigun, A Comparative Study of Fingerprint Image-Quality Estimation Methods, *IEEE Transactions on Information Forensics and Security*, vol. 2, no. 4, pp. 734–743, Dec. 2007.

[5] E. Tabassi, P. Grother, and W. Salamon. Iris quality calibration and evaluation – Performance of iris image quality assessment algorithms. NIST Interagency Report 7820, NIST, Gaithersburg, MD, USA, 2011

[6] J.-J. Brault, R. Plamondon, A complexity measure of handwritten curves: Modeling of dynamic signature forgery, *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 23, no. 2, pp. 400–413, Mar./Apr. 1993

[7] S. Müller, O. Henniger. Evaluating the biometric sample quality of handwritten signatures. In S.-W. Lee and S.Z. Li, eds., *Advances in Biometrics*, LNCS vol. 4642, 2007. Springer

[8] L. Ballard, D. Lopresti, F. Monrose, Forgery quality and its implications for behavioral biometric security, *IEEE Trans. Systems, Man, and Cybernetics, Part B*, vol. 37, no. 5, Oct. 2007, pp. 1107–1118

[9] Information technology – Biometric sample quality – Part 1: Framework. Internat. Standard ISO/IEC 29794-1, 2009,

[10] N. Ragot, J. Fortune, N. Vincent, H. Cardot, Study of temporal variability in on-line signature verification. In *Proc. 11th Internat. Conf. on Frontiers in Handwriting Recognition (ICFHR)*, Montréal, Québec, Canada, 2008.

[11] D.-Y. Yeung, H. Chang, Y. Xiong, S. George, R. Kashi, T. Matsumoto, G. Rigoll. SVC2004: First international signature verification competition. In D. Zhang, A.K. Jain, eds., *Proc. Internat. Conf. on Biometric Authentication (ICBA)*, LNCS vol. 3072, 2004. Springer

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

[13] S.J. Elliot. A comparison of on-line dynamic signature trait variables across different computing devices. PhD thesis, Purdue University, 2001

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

[15] R. Guest, Age stability in handwritten dynamic signature verification systems, *Pattern Recognition Letters*, 27 (10), pp. 1098–1104, 15 July 2006.

[16] D.W. Hosmer, S. Lemeshow, *Applied Logistic Regression*. Wiley, New York, 1989.

Table 1: Binning Conditions

Bin no. (utility score)	**Quality**	**Binning condition**	**Description**
4	**Excellent**	$j \in T$	All genuine similarity scores are larger than any impostor similarity score.
3	**Adequate**	$j \notin T$ and $z_j \geq \min(\{z_j\colon j \in T\})$	Some impostor similarity scores are larger than some genuine similarity scores, but z_j is greater than or equal to that of some samples of excellent quality.
2	**Marginal**	$0 < z_j < \min(\{z_j\colon j \in T\})$	z_j is lower than for any sample of excellent quality, but greater than 0.
1	**Unacceptable**	$z_j \leq 0$	The mean of the impostor scores is greater than the mean of the genuine similarity scores.

Table 2: Frequency Distribution of Utility Scores

Bin no. (utility score)	**Quality**	**MCYT dataset** **Frequency of Utility score regarding comparator**			**SVC dataset** **Frequency of Utility score regarding comparator**		
		1	**2**	**Overall utility score**	**1**	**2**	**Overall utility score**
1	**Unacceptable**	51	2	1	49	11	4
2	**Marginal**	357	208	91	366	250	183
3	**Adequate**	1456	423	702	198	260	391
4	**Excellent**	636	1867	1706	187	279	222