Local and Global Feature Selection for On-line Signature Verification

Jonas Richiardi, Hamed Ketabdar*, Andrzej Drygajlo
Signal Processing Institute
Swiss Federal Institute of Technology Lausanne (EPFL)

Abstract

In this paper we propose a methodology for selecting the most discriminative features in a set for on-line signature verification. We expose the difference in the definition of class between signature verification and other pattern recognition tasks, and extend the classical Fisher ratio to make it more robust to the small sample sizes typically found when dealing with global features and client enrollment time constraints for signature verification systems. We apply our methodology to global and local features extracted from a 50-users database, and find that our criterion agrees better with classifier error rates for local features than for global features. We discuss the possibility of performing feature selection without having forgery data available.

1 Introduction

Different types of features have been proposed to represent signatures in verification tasks: *local* features, where one feature is extracted for each sample point in the input domain, *global* features, where one feature is extracted for a whole signature, based on all sample points in the input domain, and *segmental* features, where the signature is subdivided into segments (typically based on velocity) and one feature is extracted for each segment. Within each feature type, many features are available to the signature verification system designer. Signature verification can be considered as a two-class pattern recognition problem, where the authentic user is a class and all her forgers are the second class. Feature selection refers to the process by which descriptors (features) extracted from the input-domain data are selected to provide maximal discrimination capability between classes. In previous work on feature selection for signature verification, statistical methods such as Linear Discriminant Analysis (LDA) have been applied for segmental features to obtain the discriminative power of each individual feature [2]. In [8], a statistical distance measure (feature-by-feature difference of means between two users scaled by standard deviation) is used to select the best feature subset out of a 42-features and a 49-features candidate list. In [1], a backwards search procedure starting from 44 global features is used with an equal error rate (EER) cost function to select a subset of features. Selection of local features based on classifier score (match distance called dissimilarity measure) is performed in [5]. Recently, a mix of 22 local and global features extracted from the SVC 2004 database were extracted and ranked individually by a "consistency" measure, essentially a difference of distance measure-specific means scaled by the standard deviations [9]. We reviewed a large number of global features (more than 150 extracted from 60 papers dating from 1983 to the present) before settling on an initial subset of the 46 that seemed to be the most commonly used in literature. To perform feature selection, we use a near-optimal feature space search algorithm (floating search) along with a modified version of the Fisher ratio as a cost function. In order to take into account effects of correlation between feature vector components, the cost is computed on whole feature vectors instead of individual features. The results of applying this method to global features on a 25-users database are reported in [6]. Our approach is as follows: a series of search steps (described in Section 3) starting from an initial set of features is used. At each search step a cost function (described in Section 2) measuring the discriminative ability of the whole feature subset is applied. We then correlate the cost function value with the mean equal error rates (EER) of a classifier for signature verification using global (Section 4) and local features (Section 5).

*now with IDIAP Research Institute, Martigny

2 Measuring the class separation ability of feature subsets

In order to evaluate a candidate feature vector subset at each search step, it is necessary to define an objective measure or cost function. The measure should

Proceedings of the 2005 Eight International Conference on Document Analysis and Recognition (ICDAR'05)
1520-5263/05 $20.00 © 2005 IEEE

be high when classes are more easily separable in feature space. Many types of cost functions can be used for feature selection. While it can be argued that ultimately the classifier error-rate (wrapper method) is the optimal criterion because the best feature vector will not be the same for different classifiers [4], running times can be prohibitive if many different feature subsets have to be tested. In the present work, we use a modified version of the Fisher ratio as a cost function, which we explain below.

The Fisher ratio [4] provides an intuitive mathematical framework for expressing the idea that within-class variability should be small, while between-class variability should be large. The within-class scatter matrix is defined as follows:

$$\mathbf{S}_w = \sum_{m=1}^{M} P_m \Sigma_m, \tag{1}$$

where M is the number of classes, P_m is the prior probability of class m, and Σ_m is an estimate of the class covariance matrix

The between-class scatter matrix is defined as:

$$\mathbf{S}_b = \sum_{m=1}^{M} P_m(\mu_m - \mu_0)(\mu_m - \mu_0)^T, \tag{2}$$

where μ_m is the mean for class m and μ_0 is the global mean vector, computed over all classes as follows:

$$\mu_0 = \sum_{m=1}^{M} P_m \mu_m, \tag{3}$$

From Eqs. 1 and 2, one computation that can be used for the "classical" Fisher ratio is:

$$J = \frac{trace(\mathbf{S}_b)}{trace(\mathbf{S}_w)} \tag{4}$$

Thus, J will be large when class samples (signature presentations) are narrowly clustered around their class means and class clusters are well separated.

2.1 Modifications to the Fisher Ratio

The Fisher Ratio relies mainly on estimates of class covariances and global covariances. Unfortunately, for global features the amount of data points available for training these covariances is very small, as only one multidimensional datapoint will be extracted from each signature. In our application scenario, users provide only 5 training signatures to keep enrollment times with reasonable limits. Therefore, it is expected that covariance estimates will be severely biased and inaccurate. While not ideal, a criterion based solely on estimations of the class means will prove more robust in these situations.

Another shortcoming with using the definition of Fisher ratio in Equation 4 as a class separability measure is that the between-class separation is only measured with respect to the global mean. Therefore, if the sum of class distances to the global mean stays the same, J will not become larger for classes that are pairwise further apart.

In an attempt to correct these issues, we add an Euclidean distance term representing the averaged two-by-two distance between class means E to the J criterion:

$$E = \frac{1}{\frac{1}{2}M(M-1)} \sum_{m}^{M} \sum_{n>m} dist(\mu_m, \mu_n) \tag{5}$$

$$J' = \sqrt{J} + E \tag{6}$$

Depending on the amount of data available, this criterion could be biased in favour of either the J or the E component by normalising each component with respect to its dynamic range over a development set, and experimenting with weights. In the current form, the E term tends to dominate for many feature subsets.

2.2 Definitions of class in feature selection for signature verification

The first possibility is to consider each user as a class, and to compute the modified Fisher ratio J' of Equation 6 using these classes (as many classes as users). This will measure the separability between authentic users, which is equivalent to measuring the separation ability of a feature subset for random impostors. We denote this definition of our measure J_ω.

The second possibility is to define two classes for each user: one authentic class and one forgery class. In this case the modified Fisher ratio J' is computed, for each user, with these two classes. Then, the mean over all the users is computed to produce what we call the "authentic-forgery" cost function $J_{\bar{\omega}}$.

This distinction is of practical importance for system designers: skilled forgeries may not always be available when developing an application, thus if the J_ω correlates well with the equal error rates of the classifier, we may perform feature selection while alleviating the need for forgery data. This is not necessarily an issue for research databases such as SVC2004 [12] or MCYT [10], which provide forgeries.

2.3 Data normalisation

Data normalisation is an important issue when using Fisher ratio-type measures of class separability for feature selection. Features used for on-line signature verification typically have very different dynamic ranges. As an example, the first moment feature (sum of squares of zero-meaned x and y components normalised by signature length) has small values in the 10^{-20} range, whereas the number of x values with positive velocity has large values in the 10^2 range.

Since the Fisher ratio is based on the trace of the class covariance matrices, it is easy to see that feature with larger dynamic ranges, which have larger variances than other features, will dominate the trace. Thus, they will carry more weight in the ratio and potentially more discriminative features will not be selected if their dynamic range is smaller.

The simple solution we have adopted is to normalise all features with respect to their max, min and mean values in the subset used to train the models. This was done both for global and local data prior to feature selection.

3 Searching the feature subset space

With a large number of initial features, exhaustive search of the feature subset space becomes computationally intractable, as an initial set of F features would result in $2^F - 1$ possible combinations. Many algorithms exist for reducing this time down to reasonable limits, amongst which genetic algorithms and floating search are popular choices and can offer comparable performance for a "medium" amount of initial features (20 to 50) [7].

Therefore, we use sequential forward floating search [11]. Floating search includes features in the current set based on recomputing the cost function for the "current set plus the candidate feature" to choose which non-included feature would bring the most increase in the cost function; it excludes features from the current set by selecting the feature whose removal would be the least damaging to the cost function. Once the floating search has reduced the initial feature set to a more tractable dimension, optimal (exhaustive) search can be performed on the reduced space of potential feature subsets.

In our experiments, we use floating search to reduce an initial set of 46 global features down to a vector of size 12, and likewise for local features we use floating search to reduce an initial set of 39 features to a vector of size 12. The choice of a final vector size of 12 was dictated by the fact that we intend to perform exhaustive search on the 4095 resulting subsets, and it may not be the optimal size.

4 Experiments: global feature selection

We used a 50-users subset of the MCYT database [10], which provides 25 authentic signatures per user and 25 "skilled" forgeries from 5 different forgers, which are allowed to practice their imitations being shown a static realisation of their target's signature. Our initial set of 46 global features is summarised in Table 2. The forward floating search algorithm was run on this initial set, using only authentic signatures, to produce a secondary subset of 12 features which maximise the cost function of Equation 6. These are shown in Table 1.

1	6	8	11	14	15
20	33	21	22	38	37 (y)

Table 1. Secondary subset of 12 global features selected by floating search (numbers reference Table 2)

It can be seen that four features carrying pressure information in the initial subset are present in the secondary subset. This is in line with many reported results in the litterature (a notable exception being [12]) that pressure is an important distinguishing feature between signers.

Secondly, to assess the ability of this criterion to predict classifier error rates, we performed exhaustive search on the 4095 possible subsets resulting from this secondary subset. The classifier we used is a Gaussian Mixture Model (GMM) classifier with 2 diagonal-covariance matrix Gaussian mixture components per user and 6 diagonal-covariance matrix Gaussian mixture components for the world model. The classifier is trained from 5 randomly chosen authentic signatures per user (since we are using global features, this means 5 d-dimensional data points, where d is the dimensionality of the feature subset under test), and tested on 20 held-out authentic signatures and 25 forged signatures, resulting in 1000 authentic test and 1250 skilled forgeries tests per feature subset. Figure 1 shows the resulting scatter plot between the J_ω criterion value and the mean EER computed over 5 random cross-validation runs. The linear correlation coefficient was computed to be -0.64. This result shows that the J_ω criterion can give an indication of the error rate to be expected when using a given global feature subset, and can be used as a preliminary step when exploring new databases or application scenarios. However, being a statistical criterion it may be too sensitive to the (low) amount of training data provided and as such may not be the most appropriate choice when dealing with global features.

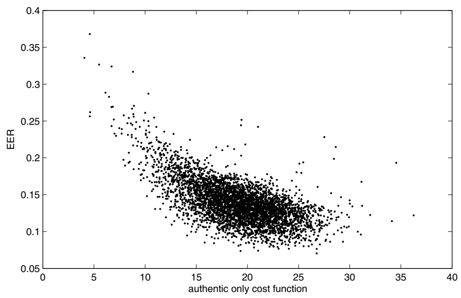

Figure 1. Scatter plot of GMM classifier EER and J_ω value for size 12 subsets ($\rho = -0.64$)

Another interesting result is to compare the cost functions $J_{\bar{\omega}}$, which defines two classes for each user: themselves and their forgers, and J_ω, which defines each user as a class (see Section2.2). Computing a correlation coefficient gives 0.78, meaning the two cost functions are highly correlated. This suggests that using a criterion for user-to-user separation (J_ω) can be used to select effective features (in the signature verification sense) based only on authentic samples, and that to this end forgeries may not be necessary. While this could be a database artefact due to the non-expert status of forgers, it is tempting to postulate that maximising the distance between legitimate users in feature space will also contribute to separate forgeries and authentic signatures.

5 Experiments: local feature selection

For local feature selection we used an initial set consisting of 13 base local features, their first order derivative (approximated using regression), and their second order derivative. Table 3 shows only the details of the base features, where the t subscript indicates that this feature is sampled at time instant t.

Again, floating search was used to find the 12-features subset that would maximise our cost function. The result of the floating search is displayed in Table 4, with features sorted by their order of inclusion (thus, from most significant contribution to least significant contribution). All signals provided directly by the tablet (pressure, y, azimuth, elevation, x) are included in the resulting feature subset. Pen orientation information, which was selected by the search algorithm, is thought to carry writer-specific information [3] and has been in use for a long time in signature verification. It is to be noted that, as is the case in speech recognition applications, derivative features can help in discrimination by compensating for improper independence assumptions between feature vectors, but carry less class-specific information than the base features, and as such only two of them (Δx and Δa) are included in the secondary subset.

3	2	12	13	1	4
5	8	11	14	6	21

Table 4. Secondary subset of 12 local features selected by floating search (numbers reference Table 3)

We trained and tested GMM-based classifiers using 32 diagonal-covariance matrix Gaussian mixture components for the user models and the world model. The local feature data was standardised to unit variance and zero-mean before modelling. Due to time constraints, we report results for 198 subsets only on the scatter plot of Figure 2; the latter suggests that J_ω is a better criterion for selecting local features than for global features, as the criterion is more correlated to the EER than for the global feature case. This can partly be explained by the fact that the covariance estimates which form part of the classic Fisher ratio (4) can now be estimated with about 3 orders of magnitude more data (one single, 2-seconds signature generates 200 local feature vectors if sampled at 100 Hz), and thus are likely to be more indicative of class scatter. The same argument goes for the pairwise distance measure term E (which is mostly dominating the J_ω criterion) and will have lower bias than in the global feature case. Lastly, the data standardisation we perform before modelling causes the distribution of local features to be closer to unimodal Gaussian.

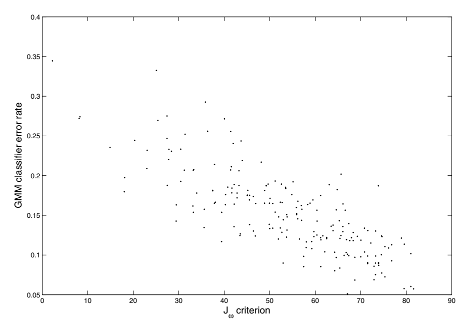

Figure 2. Scatter plot of GMM classifier EER and J_ω value for size 12 subsets ($\rho = -0.79$)

6 Discussion

We have shown a complete methodology for feature selection in signature verification and emphasised prac-

1: number of samples (T)	2: signature height (H)	3: signature width (W)
4: W to H ratio	5: T to W ratio	6: avg. velocity
7: max velocity	8: avg. velocity $\div$ max velocity	9: avg. x velocity
10: var. of x velocity	11: num. pts. with positive x velocity	12: RMS velocity
13: var. of velocity	14: pen down samples (T_d)	15: time of max velocity $\div T_d$
16: time of max x velocity $\div T_d$	17: RMS acceleration	18: avg. acceleration
19: var. of acceleration	20: avg. pressure	21: max pressure
22: point of max pressure	23: avg. azimuth	24: avg. elevation
25: avg. y velocity	26: x y velocity correlation	27: first moment
28: max pressure-min pressure	29: max x velocity	30: avg. x acceleration
31: max y velocity	32: avg. y acceleration	33: var. of pressure
34: point max. velocity $\div T_d$	35: num. points with negative x or y velocity $\div T_d$	
36: max. acceleration	37: num. points with positive x or y velocity $\div T_d$	
38-46: tangent histogram in 8 quadrants: $S_q = card\left\{\theta_t : (q-1)\frac{\pi}{8} < \theta_t < q\frac{\pi}{8}\right\} \div (T-1)$ where $k = 2, \ldots, T$ and $q = 1, \ldots, 8$		

Table 2. Initial set of 46 global features

1: horizontal position x_t	2: vertical position y_t	3: normal pressure p_t
4: path tangent angle θ_t	5: total velocity v_t	6: x velocity v_x
7: y velocity v_y	8: total acceleration a	9: x acceleration a_x
10: y acceleration a_y	11: log radius of curvature	12: pen azimuth
13: pen elevation	14-26: Δ(features 1-13)	27-39: Δ(features 14-26)

Table 3. Initial set of 39 local features

tical points such as data normalisation and theoretical points such as class definitions. One drawback of the Fisher-ratio based approach which we have not mentionned is that it assumes that the feature probability densities are unimodal Gaussians. Clearly, this will be an incorrect assumption for many features. Thus, other criteria such as mutual information or conditional mutual information might afford more flexibility in this regard. Non-parametric cost criteria might also be useful, especially when dealing with global features. This is an avenue of research we are currently pursuing.

7 Acknowledgements

We thank J. Ortega-Garcia and J. Fierrez-Aguilar for the provision of the MCYT signature sub-corpus.

References

[1] H. Crane and J. Ostrem. Automatic signature verification using a three-axis force-sensitive pen. *IEEE Transactions on Systems, Man, and Cybernetics*, 13(3):329–337, June 1983.

[2] J. Dolfing, E. Aarts, and J. van Oosterhout. On-line signature verification with Hidden Markov Models. In *Proc. International Conference on Pattern Recognition 1998*, volume 2, pages 1309–1312, Aug. 1998.

[3] K. Frank and L. Schomaker. Pen orientation characteristics of online handwritten signatures. In *Proceedings of the 11th Conference of the International Graphonomics Society (IGS2003)*, pages 224–227, November 2003.

[4] K. Fukunaga. *Introduction to statistical pattern recognition*. Academic Press, second edition, 1990.

[5] A. Jain, F. Griess, and S. Connell. On-line signature verification. *Pattern Recognition*, 35:2963–2972, 2002.

[6] H. Ketabdar, J. Richiardi, and A. Drygajlo. Global feature selection for on-line signature verification. In *Proc. 12th Conference of the International Graphonomics Society*, Salerno, Italy, June 2005.

[7] M. Kudo and J. Sklansky. Comparison of algorithms that select features for pattern classifiers. *Pattern Recognition*, 33:25–41, 2000.

[8] L. Lee, T. Berger, and E. Aviczer. Reliable on-line human signature verification systems. *IEEE Transactions on pattern analysis and machine intelligence*, 18(6):643–647, June 1996.

[9] H. Lei and V. Govindaraju. A study on the consistency of features for on-line signature verification. In *Proc. Structural, Syntactic, and Statistical Pattern Recognition: Joint IAPR International workshops, SSPR 2004 and SPR 2004*, pages 444–451, Lisbon, Portugal, 2004.

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

[11] P. Pudil, J. Novovicova, and J. Kittler. Floating search methods in feature selection. *Pattern Recognition Letters*, 15:1119–1125, November 1994.

[12] D.-Y. Yeung, H. Chang, Y. Xiong, S. George, R. Kashi, T. Matsumoto, and G. Rigoll. SVC2004: First international signature verification competition. In *Proceedings of the International Conference on Biometric Authentication (ICBA)*, July 2004.