A Multi-Classifier System for Off-Line Signature Verification Based on Dissimilarity Representation*

Luana Batista, Eric Granger and Robert Sabourin

Laboratoire d'imagerie, de vision et d'intelligence artificielle
École de technologie supérieure
1100, rue Notre-Dame Ouest, Montreál, QC, H3C 1K3, Canada
lbatista@livia.etsmtl.ca, {eric.granger, robert.sabourin}@etsmtl.ca

Abstract. Although widely used to reduce error rates of difficult pattern recognition problems, multiple classifier systems are not in widespread use in off-line signature verification. In this paper, a two-stage off-line signature verification system based on dissimilarity representation is proposed. In the first stage, a set of discrete HMMs trained with different number of states and/or different codebook sizes is used to calculate similarity measures that populate new feature vectors. In the second stage, these vectors are employed to train a SVM (or an ensemble of SVMs) that provides the final classification. Experiments performed by using a real-world signature verification database (with random, simple and skilled forgeries) indicate that the proposed system can significantly reduce the overall error rates, when compared to a traditional feature-based system using HMMs. Moreover, the use of ensemble of SVMs in the second stage can reduce individual error rates in up to 10%.

1 Introduction

Signature verification (SV) systems seek to authenticate the identity of an individual, based on the analysis of his or her signature, through a process that discriminates a genuine signature from a forgery [13]. SV systems are relevant in many situations where handwritten signatures are currently used, such as cashing checks, transactions with credit cards, and authenticating documents. In off-line SV, signatures are available on sheets of paper, which are later scanned in order to obtain a digital representation. Given a digitized signature, an off-line SV system will perform preprocessing, feature extraction and classification (also called verification) [3].

The utilization of multiple classifier systems (MCS) has been shown to reduce error rates of many challenging pattern recognition problems. However, MCS have received relatively little attention in the off-line SV community [1,4,6,15].

* This research has been supported by the *Fonds Québécois de la Recherche sur la Nature et les Technologies* (FQRNT).

By using just the best classifier, it is possible to loose valuable information contained in the other suboptimal classifiers. Moreover, it has been shown that, when a set of R classifiers is averaged, the variance contribution in the bias-variance decomposition decreases by 1/R, resulting in a smaller expected classification error [16]. Classifiers may be combined in parallel by changing (i) the training set, (ii) the input features and (iii) the parameters/architecture of the classifier. Multi-stage approaches, where each classification level receives the results of the previous one, is another way to use multiple classifiers to reduce the complexity of the problem.

Among several well-known classification methods used in off-line SV, the discrete Hidden Markov Model (HMM) [14] – a finite stochastic automata used to model sequences of observations – is known to adapt easily to the dynamic characteristics of the western handwriting [10]. The traditional approach consists in training a HMM only with genuine signatures. Therefore, the decision boundary between the impostor and genuine classes is defined later, by using a validation set that contains samples from both classes. Hence, an input pattern is assigned to the genuine class if its likelihood is greater than the decision threshold.

In contrast to this traditional system, Bicego [5] proposed a classification strategy (applied to the problem of 2D shape recognition) where both the genuine subspace, w_1, and the impostor's subspace, w_2, are modeled. Based on the dissimilarity representation (DR) approach – in which an input pattern is described by its distances with respect to a predetermined set of prototypes [12] –, the strategy consists in using a set of HMMs not as classifiers, but as a way to calculate similarity measures that define a new input feature space. The fact that two sequences O_i and O_j present similar degrees of similarity with respect to several HMMs enforces the hypothesis that O_i and O_j belong to the same class [5].

In this paper, a two-stage off-line SV system inspired by Bicego's DR concept [5] is proposed. Given a set of discrete HMMs trained with different number of states and/or different codebook[1] sizes, a greedy algorithm is employed to select the most representative ones that will be part of the first stage of the system. These HMMs can be viewed as feature extractors used to obtain the vectors of similarities. In the second stage, the vectors of similarities are used to train a SVM (or an ensemble of SVMs) whose objective is to provide the final decision. To analyze the system's performance, an overall ROC curve that takes into account user-specific thresholds is constructed. This curve also allows the system to dynamically select the most suitable solution for a given input pattern. This property can be useful in banking applications, for example, where the decision to use a specific operating point (threshold) may be associated with the value of a check.

Experiments performed with the Brazilian SV database [4] (with random, simple and skilled forgeries), indicate that the proposed system can significantly

[1] A codebook contains a set of symbols, each one associated with a cluster of feature vectors, used to generate sequences of discrete observations in discrete HMM-based systems.

reduce the overall error rates, when compared to a traditional feature-based system that uses a single HMM per writer. The paper is organized as follows. The next section presents the proposed approach. Then, Section 3 describes the experimental methodology and Section 4 presents and discusses the experiments.

2 Proposed System

In this section, a two-stage off-line SV system inspired by Bicego's DR concept [5] is proposed. In the first stage, a set of representative HMMs are used as feature extractors in order to obtain similarity measures (likelihoods) that populate new feature vectors. This idea is formally defined as follows.

Let $w_1 = \{\lambda_1^{(C_1)}, ..., \lambda_R^{(C_1)}\}$ be the set of R representative models of the genuine class C_1; $w_2 = \{\lambda_1^{(C_2)}, ..., \lambda_S^{(C_2)}\}$ be the set of S representative models of the impostor's class C_2; and $\mathcal{M}$ be the vector containing the representative models of both classes, that is, $\mathcal{M} = [w_1 \cup w_2]$. Given a training sequence $O_{trn} \in \{C_1|C_2\}$, its feature vector $\mathcal{D}(O_{trn}, \mathcal{M})$ is composed of the likelihoods computed between O_{trn} and every model in $\mathcal{M}$, that is,

$$\mathcal{D}(O_{trn}, \mathcal{M}) = \begin{bmatrix} P(O_{trn}/\lambda_1^{(C_1)}) \\ ... \\ P(O_{trn}/\lambda_R^{(C_1)}) \\ \\ P(O_{trn}/\lambda_1^{(C_2)}) \\ ... \\ P(O_{trn}/\lambda_S^{(C_2)}) \end{bmatrix}$$

After applying the same process to all training signatures from C_1 and C_2, the obtained feature vectors are used to train an ensemble of user-specific classifiers[2] (SVMs) in the second stage. During the test phase, the feature vector $\mathcal{D}(O_{tst}, \mathcal{M})$ is calculated for a given input sequence O_{tst}, and then sent to the ensemble of SVMs, which takes the final decision by majority vote. Figure 1 illustrates the proposed system, where three HMMs per subspace are used.

Observe that, if O_{tst} belongs to class C_1, the feature vector $\mathcal{D}(O_{tst}, \mathcal{M})$ should contain bigger values in the first R positions and smaller values in the remaining S positions (the inverse if O_{tst} belongs to class C_2), which allows to discriminate between the classes C_1 and C_2. In a feature-based approach, O_{tst} would be assigned to the class of the most similar model. However, this approach does not use all the information contained in a space of dissimilarities [5].

In order to obtain the most representative models to compose the subspaces w_1 and w_2, a greedy algorithm is used. Starting with an empty subspace, the models are incrementally added until a convergence criterion is reached. Basically, a model λ is chosen if its addition to the subspace minimizes the average

[2] In this paper, the term *user-specific classifier* is used to differentiate from systems where a same *global classifier* is shared by all users.

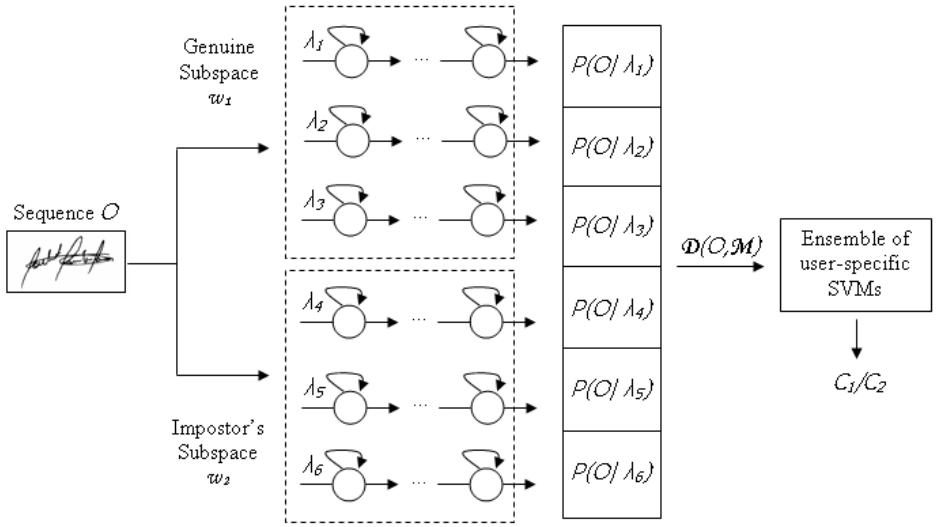


Fig. 1. Diagram of the proposed system.

error rate (AER) provided by a 1-NN classifier (with Euclidean distance) on the validation set. Algorithm 1 presents more details of this strategy.

3 Experimental Methodology

The Brazilian SV database [4] is used for proof-of-concept computer simulations. It contains 7920 samples of signatures that were digitized as 8-bit greyscale images over 400X1000 pixels, at resolution of 600 dpi. The signatures were provided by 168 writers and are organized in two sets: the development database (DB_{dev}) and the exploitation database (DB_{exp}). DB_{dev} is composed of 4320 genuine samples supplied by 108 individuals, and it is used for designing codebooks and to train the HMMs that will compose the impostor's subspace, w_2.

DB_{exp} contains 60 writers, each one with 40 samples of genuine signatures, 10 samples of simple forgery and 10 samples of skilled forgery. 20 genuine samples are used for training, 10 genuine samples for validation, and 30 samples for test (10 genuine samples, 10 simple forgeries and 10 skilled forgeries). Moreover, 10 genuine samples are randomly selected from the other 59 writers and used as random forgeries to test the current user-specific classifier. Each writer in DB_{exp} will, therefore, be associated to a genuine subspace, w_1.

The signature images are represented by means of density of pixels, extracted through a grid composed of cells of 16x40 pixels [2]. In order to generate the sequences of observations, a codebook with 35 symbols, denoted as CB_{35}, is employed (for more details regarding CB_{35} construction, see ref. [2]). For each

Algorithm 1 Selection of representative models.

Inputs:
(i) the validation set $\mathcal{V}$ composed of genuine signatures (C_1) and random forgeries (C_2)
(ii) the sets of available models Φ_1 and Φ_2, representing, respectively, C_1 and C_2
Outputs: the vector of representative models, $\mathcal{M}$

for each class C_i, $i = 1, 2$ **do**
  set $w_i \leftarrow [\,]$;
  set $j \leftarrow 0$; //j represents the number of positions of vector w_i
  repeat
    $j \leftarrow j + 1$;
    find the model λ_k in Φ_i that, when added to $w_i(j)$, provides the smallest AER of a 1-NN classifier using the validation set $\mathcal{V}$;
    remove λ_k from the list of available models Φ_i;
  until AER reaches a minimum value
end for
append, vertically, each $w_i(1..j)$; that is, $\mathcal{M} = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$

writer from both DB_{dev} and DB_{exp}, a set of HMMs is trained by using the *left-to-right* topology [14] with 20 genuine samples and different number of states; where the maximum number of states is given by the smallest sequence of observations used for training. Therefore, to compose w_1, there are a variable number of available HMMs that depends on the writer's signature size. On the other hand, to compose w_2, there are always 3296 available HMMs, taken from the 108 writers in DB_{dev}.

It is assumed that the overall system's performance is measured by an averaged ROC curve obtained from a set of user-specific ROC curves. The chosen averaging method [9] generates an overall ROC curve taking into account user-specific thresholds. At first, the cumulative histogram of random forgery scores of each user i is computed. Then, the similarity scores (thresholds) providing a same value of cumulative frequency, γ, are used to compute the operating points $\{TPR_i(\gamma), FPR_i(\gamma)\}$. Finally, the operating points associated with a same γ are averaged. Note that γ can be viewed as the true negative rate (TNR) and that it may be associated with different thresholds. Figure 2 shows an example where the thresholds associated with $\gamma = 0.3$ are different for users 1 and 2, that is $t_{user1}(0.3) \cong -5.6$ and $t_{user2}(0.3) \cong -6.4$.

To measure the system's performance during test, false negative rates (FNR) and false positive rates (FPR) are calculated by using the user-specific thresholds associated to different operating points γ of the averaged ROC curve. The average error rate (AER), also computed for different γ, indicates the total error of the system, where FNR and FPR are averaged taking into account the *a priori* probabilities.

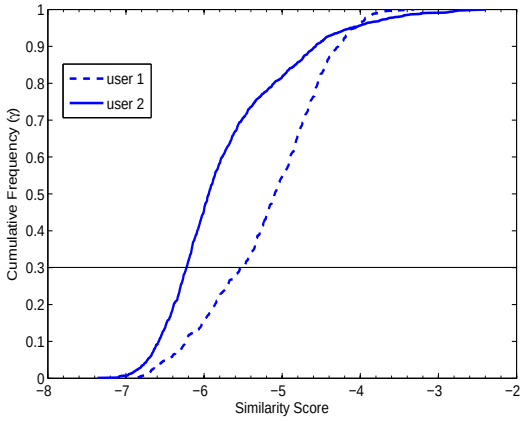


Fig. 2. Cumulative histogram of random forgery scores regarding two different users.

4 Simulation Results

This section presents two sets of experiments performed with the Brazilian SV database [4]. In the first one, the proposed DR-based system is compared to a traditional feature-based system that uses a single HMM per writer. In this case, only one SVM per writer is employed in the second stage of the proposed system. In the second set of experiments, the impact of using ensembles of SVMs in the second stage is investigated.

4.1 DR-based approach *vs.* feature-based approach

Given the HMMs trained such as explained in Section 3 and the validation set – which contains 10 genuine signatures taken from DB_{exp} *versus* 1080 (108x10) random forgeries taken from DB_{dev}–, Algorithm 1 was applied to select the most representative models. This process was performed individually for each writer in DB_{exp}, and, on average, 2 models were selected for representing w_1, and 3 models, for representing w_2. Then, still using the validation set, the gridsearch technique with 10-fold cross-validation was employed in order to find the best parameters $\{c, \gamma\}$ of the two-class SVMs ($CSVC$) with RBF kernel [8]. Finally, the selected parameters were used to train a single SVM per writer, with 20 genuine signatures taken from DB_{exp} *versus* 2160 (108x20) random forgeries taken from DB_{dev}.

The averaged ROC curve representing the proposed system (obtained from the validation set) is indicated by the square-dashed line in Figure 3. The circle-dashed curve corresponds to a traditional (feature-based) HMM-based system designed such as described in Introduction. This baseline system uses only density of pixels as features and a single HMM per writer as classifier, where the

number of states is selected through the cross-validation procedure described in [10]. Table 1 (a) and (b) present the error rates on test for both systems regarding different operating points (γ). Note that the proposed system provided a reduction in AER from 2.5%, for $\gamma = 1$, up to 9.87%, for $\gamma = 0.95$.

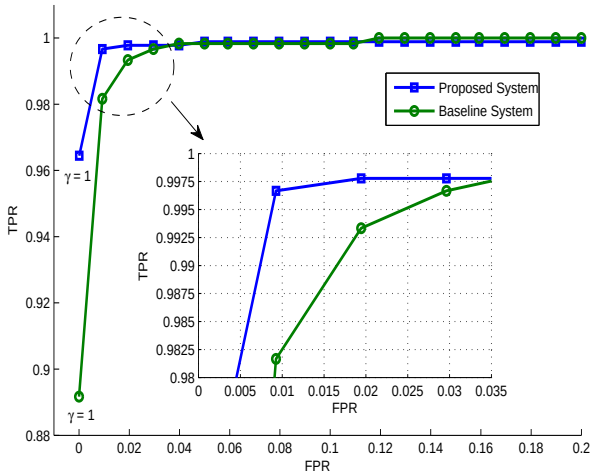


Fig. 3. Averaged ROC curves of baseline and proposed systems.

Table 2 shows the results provided by other systems that use the Brazilian SV database [4] and pixel density features. Except in [4], where DR is employed to design a global classifier, the refered articles propose feature-based approaches. Note that our system provides the smallest AER (see Table 1 (b), $\gamma = 1$).

4.2 DR-based approach using an ensemble of SVMs

The experiment presented in this section consisted in analyzing the impact of employing and combining different SVMs per writer in the proposed DR-based system. In order to generate the candidate classifiers, a different user-specific SVM is trained each time that a new model is selected by Algorithm 1. For example, given that w_1 is represented by models (a, b, c) and w_2, by (d, e), six candidate SVMs can be produced by using $w_1 \cup w_2$, that is, (a, d); (a, d, e); (a, b, d); (a, b, d, e); (a, b, c, d) and (a, b, c, d, e).

Once the set of candidate classifiers is obtained, the algorithm ICON [17] is applied in order to incrementally construct the ensemble of SVMs. Like as Algorithm 1, ICON consists in a greedy process that, at each iteration, chooses the classifier that best improves system's performance (on validation data) when added to the current ensemble.

Table 1. Overall error rates (%) of baseline and proposed systems on the test data.

(a) Baseline system

γ	FNR	FPR_{random}	FPR_{simple}	$FPR_{skilled}$	AER
0.95	0.50	6.83	12.83	68.00	22.04
0.96	0.50	6.00	10.83	64.83	20.54
0.97	0.83	5.67	9.00	60.17	18.92
0.98	1.17	4.00	5.67	52.50	15.83
0.99	2.33	2.67	4.00	42.67	12.92
1	12.67	0.33	1.17	19.83	8.50

(b) Proposed DR-based system with single SVMs

γ	FNR	FPR_{random}	FPR_{simple}	$FPR_{skilled}$	AER
0.95	2.17	3.83	5.17	37.50	12.17
0.96	2.17	3.17	4.83	36.67	11.71
0.97	2.17	2.50	4.50	36.50	11.42
0.98	2.33	2.00	4.00	36.33	11.17
0.99	2.50	1.33	3.33	34.67	10.46
1	16.17	0.00	0.17	7.67	6.00

(c) Proposed DR-based system with ensembles of SVMs

γ	FNR	FPR_{random}	FPR_{simple}	$FPR_{skilled}$	AER
0.95	2.83	2.33	4.50	33.33	10.75
0.96	3.00	1.67	3.67	33.67	10.50
0.97	2.83	1.33	3.67	33.67	10.37
0.98	2.83	1.00	3.50	34.17	10.37
0.99	3.00	1.00	3.50	32.67	10.04
1	13.50	0.00	0.17	8.33	5.50

A measure called CI (from Chebishev's inequality) [7,11] is employed to evaluate the ensemble. It is computed as $CI = \sigma(\tau)/\mu(\tau)^2$, where σ and μ denote the variance and the average of the set of margins τ provided by the samples in the validation set, respectively. Given a sample x_i from class C_1, its margin τ_i is given by the difference between the number of votes assigned to the true class C_1, minus the number of votes assigned to class C_2. The ensemble providing the smallest CI value contains the strongest and less correlated classifiers [11].

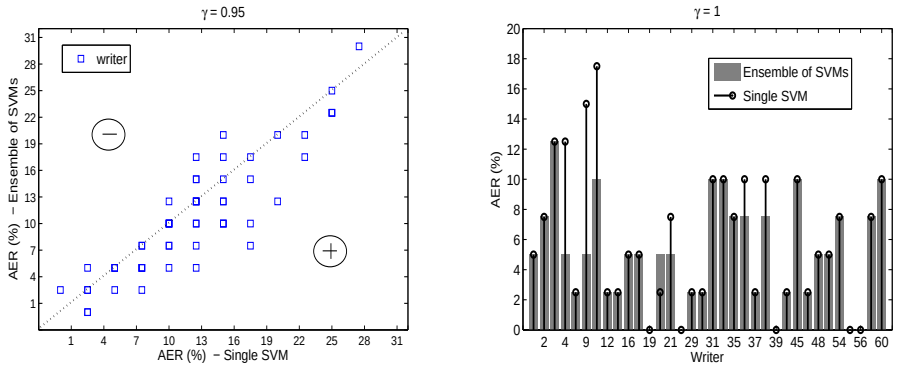
The overall error rates obtained on test by using majority vote are shown by Table 1 (c). Note that, except for $\gamma = 1$, the improvements were mostly related to $FPRs$. Figure 4 (a) presents the 60 individual $AERs$ for $\gamma = 0.95$. According

Table 2. Error rates (%) provided by other off-line SV systems.

Reference	FNR	FPR_{random}	FPR_{simple}	$FPR_{skilled}$	AER
Batista et al. [2]	9.83	0	1.00	20.33	7.79
Bertolini et al.[4]	25.32	3.8	4.48	7.8	10.35
Justino et al. [10]	2.17	1.23	3.17	36.57	7.87

to this graph, 48.33% of the writers had their *AERs* on test reduced (in up to 10%) with the use of ensembles – which may indicate a considerable amount of users in a real world application –, while 15% performed better with single SVMs. For the remaining 36.67%, both versions of the system performed equally.

Regarding $\gamma = 1$, only 34 writers were associated to ensembles by algorithm ICON. The remaining 26 writers kept using a single SVM with all models selected by Algorithm 1. In Figure 4 (b), observe that the *AER* was reduced in 7.5% for writers 4 and 10, in 10% for writer 9, and in 2.5% for writers 21, 26 and 38. Whereas for writer 20, the use of ensembles increased the *AER* in 2.5%.



(a) (b)

Fig. 4. Individual *AERs* obtained on test for $\gamma = 0.95$ (a) and $\gamma = 1$ (b), before and after using ensemble of SVMs. In (a), note that the writers that had their *AERs* reduced by using ensemble of SVMs are located below the dotted line.

5 Conclusions

In this paper, a two-stage off-line SV system based on DR [5] was proposed. In the first stage, a set of representative HMMs – trained with different number of states – was used to produce similarity measures to form new feature vectors. In the second stage, these vectors were input to one or more SVMs in order to provide the final classification.

When compared to a baseline system that uses a single HMM per writer, the proposed system provides a reduction in *AER* from 2.5%, for $\gamma = 1$, up to 9.87%, for $\gamma = 0.95$. One of the reasons for this improvement is the fact that both genuine and forger subspaces are modeled, which does not occur with traditional (feature-based) systems using HMMs. Moreover, feature-based approaches do not use all the information contained in a similarity space [5]. Finally, the use of ensemble of SVMs can reduce individual error rates in up to 10%.

The proposed approach may require greater computational complexity (training time and memory consumption) than a traditional approach due to the generation of the candidate HMMs and to the selection of the most representative ones. However, once the most representative HMMs are obtained (about 5 per writer in the experiments), all sub-optimal solutions can be discarded. As future work, we intend to employ different codebook sizes and different number of states to generate the set of candidate HMMs.

References

1. R. Bajaj and S. Chaudhury. Signature verification using multiple neural classifiers. *Pattern Recognition*, 30:1–7, 1997.
2. L. Batista, E. Granger, and R. Sabourin. Improving performance of hmm-based off-line signature verification systems through a multi-hypothesis approach. *International Journal on Document Analysis and Recognition, IJDAR*, 2009.
3. L. Batista, D. Rivard, R. Sabourin, E. Granger, and P. Maupin. State of the art in off-line signature verification. In B. Verma and M. Blumenstein, editors, *Pattern Recognition Technologies and Applications: Recent Advances*. IGI Global, 1st edition, 2007.
4. D. Bertolini, L. Oliveira, E. Justino, and R. Sabourin. Reducing forgeries in writer-independent off-line signature verification through ensemble of classifiers. *Pattern Recognition*, 43:387–396, 2010.
5. M. Bicego, V. Murino, and M. Figueiredo. Similarity-based clustering of sequences using hidden markov models. *Pattern Recognition*, 37(12):2281–2291, 2004.
6. H. Blatzakis and N. Papamarkos. A new signature verification technique based on a two-stage neural network classifier. *Engineering Applications of Artificial Intelligence*, 14:95–103, 2001.
7. L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.
8. C. Chang and C. Lin. Libsvm: a library for support vector machines. In *http://www.csie.ntu.edu.tw/˜ cjlin/libsvm*, 2001.
9. A. Jain and A. Ross. Learning user-specific parameters in a multibiometric system. In *International Conference on Image Processing (ICIP)*, pages 57–60, 2002.
10. E. Justino, F. Bortolozzi, and R. Sabourin. Off-line signature verification using hmm for random, simple and skilled forgeries. In *International Conference on Document Analysis and Recognition*, pages 105–110, 2001.
11. M. Kapp, R. Sabourin P., and Maupin. An empirical study on diversity measures and margin theory for ensembles of classifiers. In *10th International Conference on Information Fusion (Fusion 2007)*, pages 1–8, 2007.
12. E. Pekalska, P. Paclik, and R. Duin. A generalized kernel approach to dissimilarity based classification. *Journal of Machine Learning Research*, 2:2001, 2002.
13. R. Plamondon. *Progress in Automatic Signature Verification*. Word Scientific, Singapore, 1994.
14. L. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *IEEE*, 77(2):257–286, 1989.
15. C. Sansone and M. Vento. Signature verification: Increasing performance by a multi-stage system. *Pattern Analysis and Applications*, 3:169–181, 2000.
16. D. Tax. *One-class Classification*. PhD thesis, TU Delft, 2001.
17. A. Ulas, M. Semerci, O. Yildiz, and E. Alpaydin. Incremental construction of classifier and discriminant ensembles. *Information Sciences*, 179(9):1298–1318, 2009.