A Multi-Hypothesis Approach for Off-Line Signature Verification with HMMs

Luana Batista, Eric Granger and Robert Sabourin
Laboratoire d'imagerie, de vision et d'intelligence artificielle (LIVIA)
École de technologie supérieure
1100, rue Notre-Dame Ouest, Montréal, QC, H3C 1K3, Canada
lbatista@livia.etsmtl.ca, {eric.granger, robert.sabourin}@etsmtl.ca

Abstract

In this paper, an approach based on the combination of discrete Hidden Markov Models (HMMs) in the ROC space is proposed to improve the performance of off-line signature verification (SV) systems designed from limited and unbalanced training data. This approach is inspired by the multiple-hypothesis principle, and allows the system to choose, from a set of different HMMs, the most suitable solution for a given input sample. By training an ensemble of user-specific HMMs with different number of states, and then combining these models in the ROC space, it is possible to construct a composite ROC curve that provides a more accurate estimation of system's performance during training and significantly reduces the error rates during operations. The experiments performed by using a real-world SV database with random, simple and skilled forgeries, indicated that the proposed approach can reduce the average error rates by more than 17%.

1 Introduction

Signature verification (SV) systems seek to authenticate the identity of an individual, based on an analysis of his/her signature, through a process that discriminates a genuine signature from a forgery [1]. In off-line SV, the signature is available on a sheet of paper, which is later scanned in order to obtain a digital representation. These systems are relevant in many situations where handwritten signatures are currently used, such as cashing checks, transactions with credit cards, and authenticating documents [8].

Though not fully explored in literature, it has recently been shown that the Receiver Operating Characteristic (ROC) curve - where the true positive rates (TPR) are plotted on the *y axis*, while the false positive rates (FPR) are plotted on the *x axis* - provides a powerful tool for evaluating, combining and comparing off-line SV systems [2] [10]. By taking into account several operating points (thresholds), the ROC curve allows to analyze these systems under different classification costs [5].

Among several well-known classification methods used in off-line SV, the discrete Hidden Markov Model (HMM) - a finite stochastic automata used to model sequences of observations - adapts easily to the dynamic characteristics of the occidental handwriting [9]. Despite the fact that the HMM is a generative classifier [3] which generally requires much training data to perform well, the HMM-based off-line SV systems are often designed from limited and unbalanced data. This occurs because the dataset used to model a writer generally contains a reduced number of genuine signatures against several random forgeries [10].

In this paper, an approach based on the combination of discrete HMMs in the ROC space is proposed to improve performance of off-line SV systems designed from limited and unbalanced data. By training an ensemble of user-specific HMMs with different number of states and then combining these models in the ROC space, it is possible to construct a composite ROC curve that provides a more accurate estimation of system's performance during training. Moreover, during operations, the corresponding operating points - which may be selected dynamically according to the risk associated with the input samples - can significantly reduce the error rates. This approach follows the multiple-hypothesis principle [7], which request the system to propagate several hypothesis throughout the recognition steps, generating a hierarchical tree of possible solutions.

The rest of this paper is organized as follows. Next section describes the issues that impact the performance of off-line SV systems. Then, Section 3 describes the proposed approach and its advantages. In Section 4, the experimental results are shown and discussed. Finally, Section 5 presents the conclusions of this work.

2 Problem Statement

In this paper, it is assumed that the off-line SV system is composed of several local classifiers, where each one is a

978-0-7695-3725-2/09 $25.00 © 2009 IEEE
DOI 10.1109/ICDAR.2009.5

discrete HMM trained with data corresponding to a specific writer, and that the performance of the whole system is measured by a single averaged ROC curve. Averaging methods have been used to group ROC curves from different classifiers [5] [11]. Ross [11], for example, proposed a method to generate averaged ROC curves which takes into account user-specific thresholds. At first, for each user i, the cumulative histogram of the impostor scores (regarding that user) 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 = negatives (forgeries) correctly classified / total of negatives) and that it may be associated with different thresholds. Figure 1 shows an example where the thresholds associated with γ = 0.3 are different for users 1 and 2, that is $t_{user1}(0.3) \cong 5.0$ and $t_{user2}(0.3) \cong 5.5$.

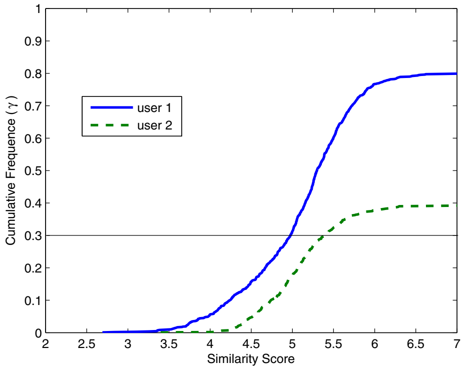
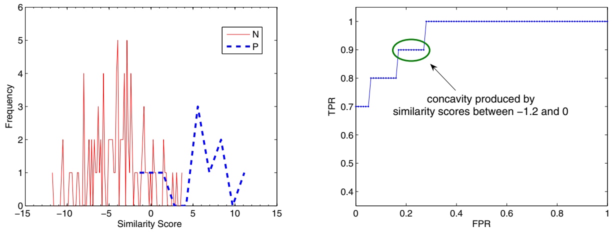


Figure 1. Cumulative histogram of the impostor scores regarding two different users.

In off-line SV, where the dataset used to model a writer signature generally contains a reduced number of genuine samples against several random forgeries, it is common to obtain ROC curves with concave areas. In general, a concave area indicates that the ranking provided by the classifier in this region is worse than random [6]. Figure 2 (a) shows an example of score distribution in an off-line SV system. The positive class, P, contains only 10 genuine samples of a given writer, while the negative class, N, contains 100 samples of forgeries. Due the limited amount of samples in the positive class, the resulting ROC curve (see Figure 2 (b)) presents three concave areas, which correspond to low-quality predictions [6]. For example, the similarity scores between -1.2 and 0 provide $TPRs$ of 90%. The result of averaging the ROC curves related to the models of 100 different writers, by using the Ross's method, is illustrated by Figure 3. Note that the imperfections of individual ROC curves are hidden within the average points, which can be observed with any averaging algorithm.



(a) (b)

Figure 2. Typical score distribution and respective ROC curve of a writer model.

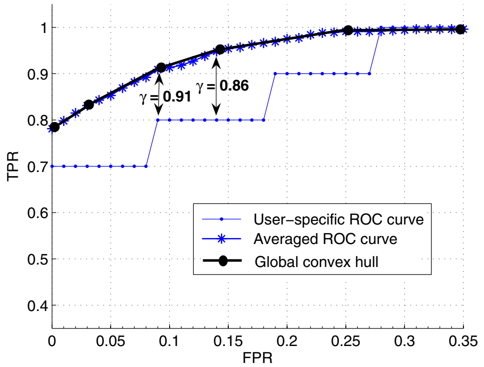


Figure 3. Averaged ROC curve obtained by applying the Ross's method to 100 different user-specific ROC curves.

The drawback of using an averaged ROC curve can be observed during the selection of optimal thresholds in the respective convex hull. Given two γ in the convex hull, γ_1 and γ_2, where each one minimizes a different set of costs [13], γ_2 should provide a TPR higher than γ_1 whenever $\gamma_1 > \gamma_2$. However, regarding an user-specific ROC curve, γ_1 and γ_2 may fall in a same concave area, providing identical $TPRs$. An example is illustrated by Figure 3, where $TPR(\gamma = 0.86)$ is higher than $TPR(\gamma = 0.91)$ in the global convex hull, but, in the user-specific ROC curve, $TPR(\gamma = 0.86)$ is equal to $TPR(\gamma = 0.91)$.

3 A Multi-Hypothesis Approach

Based on the combination of HMMs trained with different number of states, the approach proposed in this section provides a solution to repair concavities of user-specific ROC curves while generating a high quality averaged ROC curve. Three steps are involved in the proposed multi-hypothesis approach: model selection, combination and averaging.

3.1 Model Selection

During the model selection step, a ROC outer boundary is constructed in order to encapsulate the best operating points provided by HMMs trained with different number of states. The utilization of different HMMs is motivated by the fact that the superiority of a classifier over another may not occur on the whole ROC space [5]. In off-line SV systems where the optimal number of states for a HMM is found empirically by a cross-validation process [4] [9], it is often observed that the best HMM is not superior than the other intermediate/sub-optimal HMMs in all operating points of the ROC space.

Given a set of ROC curves generated from different classifiers associated with a same user, the process consists in splitting the *x axis* $\in [0, 1]$ into a number of bins, and within each bin finding the pair (FPR, TPR) having the largest value of TPR. While the ROC outer boundary is being generated, the best HMM_s, where s is the number of states, is automatically chosen for each operating point. Figure 4 shows an example of an user-specific ROC outer boundary constructed from ROC curves of two different classifiers, HMM_7 and HMM_9. The corresponding convex hull is composed of three vertices, p, q and r, where p and q are associated with HMM_9, and r is associated with HMM_7.

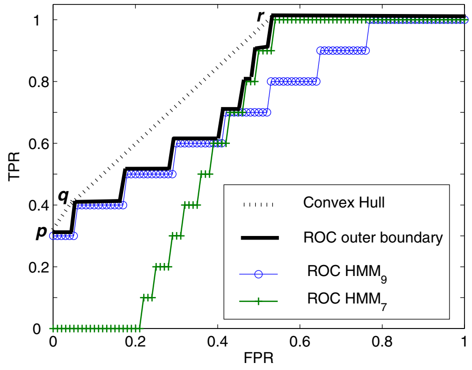


Figure 4. ROC outer boundary constructed from ROC curves of two different HMMs.

3.2 Combination

In the second step, the combination method proposed by Scott et. al [12] is applied to the ROC outer boundaries in order to repair concavities. Given two vertices A and B on the convex hull, it is possible to realize a point C, located between A and B, by randomly choosing between A and B. The probability of selecting one of the two operating points is determined by the distance of C regarding A and B. Equations 1 and 2 indicates how a given FPR_C can be obtained.

$$P(C = A) = \frac{FPR_C - FPR_B}{FPR_A - FPR_B} \qquad (1)$$

$$P(C = B) = 1 - P(C = A) \qquad (2)$$

In the convex hull illustrated by Figure 4, any point in the segments $\overline{pq}$ and $\overline{qr}$ is realizable by combination. The expected operating point (FPR, TPR) is given by equations 3 and 4 [12].

$$FPR = (P(C = A) \cdot FPR_A) + (P(C = B) \cdot FPR_B) \qquad (3)$$

$$TPR = (P(C = A) \cdot TPR_A) + (P(C = B) \cdot TPR_B) \qquad (4)$$

3.3 Averaging

Finally, a process based on the Ross's method [11] is used to group the operating points already computed during the combination step. Given a γ, the process searches the pairs $\{TPR_i(\gamma), FPR_i(\gamma)\}$ where $FPR_i(\gamma) = 1 - \gamma$ (recalling that γ can be viewed as the TNR, and that $FPR = 1 - TNR$). Then, the operating points corresponding to a same γ are averaged and used to generate the averaged ROC curve.

In the test phase, γ is used to retrieve the set of user-specific HMMs/thresholds which will be applied to unknown data. Figure 5 illustrates two possible situations linking the averaged ROC curve and an user-specific ROC curve. In the first case, γ falls directly in HMM_7. While in the second case, the requested γ is obtained combining classifiers HMM_7 and HMM_9. That is, each test sample must be randomly sent either to HMM_7 or to HMM_9, according to the probabilities given by equations (1) and (2).

4 Simulation Results

The brazilian signature database [9] was 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 300 ppi. 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 by 4320

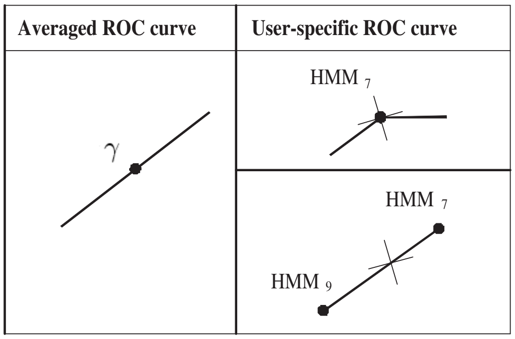


Figure 5. Retrieving user-specific HMMs from an averaged ROC curve.

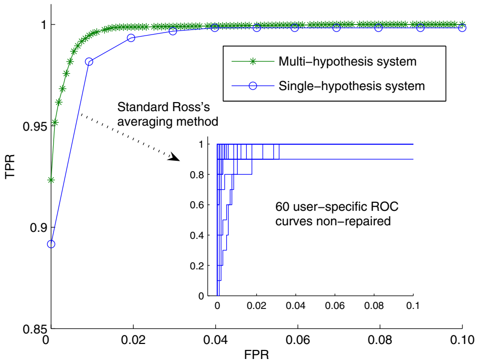


Figure 6. ROC curves of the single- and multi-hypothesis systems.

genuine samples supplied by 108 individuals and is mainly used for designing codebooks. DB_{exp} contains 60 writers, each one with 40 samples of genuine signatures, 10 samples of simple forgery and 10 samples of skilled forgery. For each writer, 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 model.

The signature images are represented by means of density of pixels, extracted through the grid segmentation scheme described in [9]. In experiments not presented here, a codebook with 35 symbols, CB_{35}, was constructed by applying the *k-means* algorithm to the first 30 signatures of each writer in DB_{dev}. Then, CB_{35} was applied to DB_{exp}, which is composed of 60 authors. The ROC curve of the multi-hypothesis system is indicated by the star-dashed line in Figure 6. Whereas the circle-dashed line represents the baseline or single-hypothesis system used for comparisons. In this system, only the HMM which performs the best in the cross-validation process[1] is considered for generating an user-specific ROC curve. This means that all operating points of an individual ROC curve are associated with a same HMM. Besides, the user-specific ROC curves are directly averaged, without repairing, by using the standard Ross's method [11] (see the inner graphic in Figure 6). As expected, the multi-hypothesis system provided a higher ROC curve, and, due the Scott's method [12], a superior number of operating points was obtained. Note that both ROC curves were generated by using a validation set which contains only genuine samples and random forgeries.

In the following phase, the operating points of the multi-hypothesis ROC curve (given by γ) were used to retrieve the user-specific HMMs/thresholds, and apply them to the test set. Table 1 presents the error rates on test for some γ values in both single and multi-hypothesis systems. Since there are three types of forgeries in the test set, the average error rate is calculated as $AER(\gamma) = (FNR(\gamma) + FPR(\gamma)_{random} + FPR(\gamma)_{simple} + FPR(\gamma)_{skilled})/4$.

Table 1. Error rates (%) on test.

Results obtained with the single-hypothesis system

γ	FNR	$FPR_{rand.}$	$FPR_{simp.}$	$FPR_{skil.}$	AER
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

Results obtained with the multi-hypothesis system

γ	FNR	$FPR_{rand.}$	$FPR_{simp.}$	$FPR_{skil.}$	AER
0.96	4.13	4.67	8.70	54.81	18.07
0.97	4.15	3.17	7.35	52.23	16.72
0.98	4.48	2.50	5.51	47.05	14.88
0.99	5.55	1.17	4.00	39.63	12.58
1	12.83	0	1.17	20.50	8.62

In general, the multi-hypothesis system provided smaller error rates. Moreover, the $FPR(\gamma)_{random}$ are closer to the expected error rates given by 1-γ. Other results obtained with the multi-hypothesis system are presented in Table 2.

Finally, in order to analyze the impact of repairing individual ROC curves, the proposed approach was applied only to the 20 writers having ROC curves with concavities. On average, 75.91% of the problematic writers had their $AERs$ on test enhanced with the multi-hypothesis system in the region between $\gamma = 0.9$ and $\gamma = 1$. This represents 25.30% of the whole population, which may indicate a considerable amount of users in a real world application. Only 18.64% of the problematic writers performed better with

[1] Given a set of HMMs trained with different number of states, the cross-validation process selects the HMM providing the highest training probability [9].

Table 2. Additional error rates (%) on test obtained with the multi-hypothesis system.

γ	FNR	$FPR_{rand.}$	$FPR_{simp.}$	$FPR_{skil.}$	AER
0.991	5.71	1.00	3.83	38.01	12.13
0.992	5.81	0.67	3.83	37.38	11.92
0.993	6.00	0.67	3.50	36.15	11.58
0.994	6.17	0.67	3.50	35.90	11.56
0.995	6.30	0.67	3.17	34.80	11.23
0.996	7.10	0.67	2.67	33.73	11.04
0.997	7.63	0.67	2.67	32.48	10.86
0.998	8.21	0.67	2.67	30.60	10.54
0.999	9.37	0.17	1.91	26.83	9.57

the single-hypothesis system. For the remaining 5.45%, both systems performed equally. Figure 7 presents the results for $\gamma = 0.96$, where the improvements obtained with the multi-hypothesis system are located in the positive side, that is, below the dotted line. With the single-hypothesis system, the author indicated by the arrow had an AER of 22.5%. When using the multi-hypothesis system, the respective AER was reduced to 4.9%, that is, 17.6% lower.

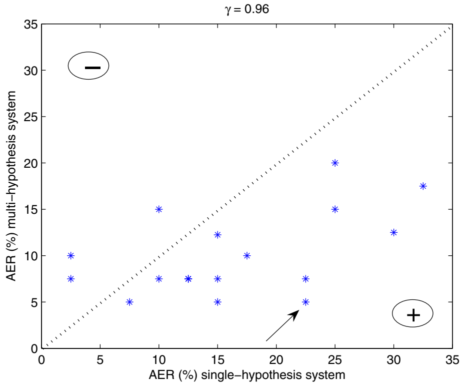


Figure 7. User-specific $AERs$ obtained on test with the single- and multi-hypothesis systems.

5 Conclusions

Based on the combination of HMMs trained with different number of states, this paper proposed a multi-hypothesis approach used to improve performance of off-line SV systems designed from limited and unbalanced data. The experiments carried out on the brazilian SV database showed that the proposed approach leads to smaller error rates on unseen data ragarding not only the baseline system, but also another single-hypothesis system developed with the same SV database [9][2]. The multi-hypothesis approach can be used for dynamic selection of the best classification model based on the risk associated with the input samples. In a banking application, for example, the decision of using a specific operating point may be associated with the amount of the check. In the simplest case, for an user that rarely signs high value checks, big amounts would require operating points related to low $FPRs$, such as provided by γ =0.999 or γ =1; while normal amounts would require operating points related to low $FNRs$, since the user would not feel comfortable with frequent rejections. Finally, this approach can be easly adapted to any type of neural or statistical classifiers designed to solve similar two-class problems.

References

[1] 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.

[2] H. Coetzer and R. Sabourin. A human-centric off-line signature verification system. In *Int. Conf. on Document Analysis and Recognition (ICDAR 2007)*, pages 153–157, 2007.

[3] C. Drummond. Discriminative vs. generative classifiers for cost sensitive learning. In *Canadian Conference on Artificial Intelligence. Lecture Notes in Artificial Intelligence.*, pages 479–490, 2006.

[4] A. El-Yacoubi, E. Justino, R. Sabourin, and F. Bortolozzi. Off-line signature verification using hmms and cross-validation. In *IEEE Workshop on Neural Networks for Signal Processing*, pages 859–868, 2000.

[5] T. Fawcett. Roc graphs: Notes and practical considerations for data mining researchers. *Machine Learning*, 2004.

[6] P. Flach and S. Wu. Repairing concavities in roc curves, 2003.

[7] H. Fujisawa. *Robustness Design of Industrial Strength Recognition Systems*. Springer, 2007.

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

[9] 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.

[10] L. Oliveira, E. Justino, and R. Sabourin. Off-line signature verification using writer-independent approach. In *Int. Joint Conference on Neural Networks (IJCNN)*, pages 2539–2544, 2007.

[11] A. Ross. *Information Fusion in Fingerprint Authentication*. PhD thesis, Michigan State University, 2003.

[12] M. Scott, M. Niranjan, and R. Prager. Realisable classifiers: improving operating performance on variable cost problems, 1998.

[13] F. Tortorella. A roc-based reject rule for dichotomizers. *Pattern Recognition Letters*, 26:167–180, 2005.

[2] Error rates reported in [9]: FNR = 2.17%, $FPR_{rand.}$ = 1.23%, $FPR_{simp.}$ = 3.17%, $FPR_{skil.}$ = 36.57%, AER = 10.78%.