Interval versus Histogram of Symbolic Representation Based One-Class Classifier for Offline Handwritten Signature Verification

Mohamed Anis Djoudjai[0000-0001-5302-5023] and Youcef Chibani[0000-0002-7957-7456]
Laboratoire d'Ingénierie des Systèmes Intelligents et Communicants
Faculty of Electronics and Computer Science
University of Sciences and Technology Houari Boumédiene
32, El Alia, Bab Ezzouar, 16111, Algiers, Algeria
{ma.djoudjai, ychibani}@usthb.dz

Abstract. This paper proposes a comparison study of using Interval and Histogram of Symbolic Representation (ISR and HSR) based One-Class classifiers, namely OC-ISR and OC-HSR, respectively, applied to the offline signature verification. Usually, symbolic verification models are built straightforward from the feature space. The proposed work explores an alternative approach based on the use of feature-dissimilarities generated from Curvelet Transform (CT) for building the OC-ISR and the OC-HSR classifier. For the OC-ISR classifier, a new weighted membership function is proposed for computing the similarity values between a dissimilarity query vector and a targeted ISR model. The experimental evaluation performed on the well-known public datasets GPDS, CEDAR, and MCYT, reveals the proposed OC-ISR's superiority over the OC-HSR classifier. Moreover, the proposed verification model based on the OC-ISR classifier outperforms the last similar work reported in the literature on the GPDS-160 dataset by 0.99%, 0.8%, and 0.35% of Average Error Rate (AER) for 5, 8, and 12 reference signatures, respectively.

Keywords: SDA, histogram, interval, one-class classification, dissimilarity, signature verification.

1 Introduction

Automating biometric recognition systems using offline handwritten signatures can offer two distinct applications which are signature identification and signature verification. The former aims to attribute an identity to a query signature belonging to a writer enrolled in a database. While the latter aims to verify the authenticity of a query signature allegedly belonging to a writer, whether it is a genuine or a forgery. Nevertheless, signature verification is a more challenging problem for researchers according to the state-of-the-art performances achieved during the last two decades, and therefore represents the focus of the present paper. Generally, an Offline Handwritten Signature Verification System (OHSVS) is composed of three main modules which are prepro-

cessing, feature generation, and classification. Since the main contribution of the present paper concerns the classification module, the present paper focuses only on attempting to develop this module. Hence, the classification methods proposed in the literature for OHSVSs can be divided into two categories: Multi-Class Classifiers (MCCs) and One-Class Classifiers (OCCs). OCCs represents an alternative of MCCs when negative examples are not available during the training step. Hence, the OCC concept is desirable for signature verification cases, since only genuine signatures (positive class) contained in a bank database, for example, are available for training the OHSVS.

The classifiers proposed in the literature for OHSVSs are built following one of the two approaches: Writer Dependent (WD) and Writer Independent (WI). The WD consists of building a model for each writer using its genuine signatures. On the other hand, the WI approach is based on building one single model for all writers involved in the database. This later uses the dissimilarity concept where the pattern recognition problem becomes a bi-class problem namely target and reject class [1]. Several MCCs have been explored for building OHSVSs such as Hidden Markov Models (HMMs), Support Vector Machines (SVMs), Neural Networks, and Deep Learning or an ensemble of combined classifiers [1-2]. On the other hand, few authors have explored the use of OCC such as the OC-SVM [2].

Recently, a new OCC based on Symbolic Data Analysis ($OC-SDA$ classifier) method has been introduced for OHSV. Generally, the symbolic models are constructed either *via* intervals (ISR) or histograms (HSR) [3] using exclusively straightforward features such as Curvelet Features and Local Binary Patterns (LBP) features [4]. In this investigation work, a comparative evaluation is proposed using the $OC-SDA$ classifier through its two models namely the $OC-ISR$ and $OC-HSR$ model, for offline signature verification. The symbolic verification models proposed in this work are constructed on the feature-dissimilarity space. Dissimilarities are generated from the Curvelet Transform (CT) feature space [5]. Moreover, a new membership function is proposed for computing the similarity values between a dissimilarity of the feature vector and the model. The proposed system is based on WI parameters where the same configuration parameters are set for all writers evolved into the database.

The remainder of this paper is organized as follows. Section II presents a brief review of Symbolic Data Representation (SDR) and its extension for classification. Next, a detailed description of the proposed system is presented in section III. To evaluate the performance of the proposed system, various experiments performed on three offline signature datasets: GPDS-300, CEDAR, and MCYT datasets, are presented in section IV. Finally, a conclusion and perspective work is provided in the last section.

2 Brief Review of Symbolic Data Representation

Usually, the classes are represented by a simple sample mean (or median or the like). However, such a representation of classes doesn't provide a real description of *intra-class* variability. Thus, an alternative approach for representing the aggregation observation (i.e. samples) of the same class can be performed through the SDR concept either

via intervals (ISR) or histograms (HSR). In the beginning, this concept has been proposed for analyzing (SDA) and clustering complex data [3]. Later, the SDA applicability has been extended for reducing large datasets and for one-class classification problems [4] namely the $OC - SDA$ classifier.

The base idea of symbolic representation models consists of representing symbolically each feature component by either a set of intervals (ISR) or histograms (HSR). Hence, the symbolic model according to the ISR concept can be described as follows:

$$ISR = \{If_1, If_2, \dots, If_P\} \tag{1}$$

where If_k represents the feature interval associated to the k^{th} feature component such that $k = \{1,2, \dots, P\}$, and P represents the size of the feature vectors. For generating the inferior and superior bounds of the feature intervals (If_k), different statistical metrics can be used such as $mean$ and $standard\ deviation$ [4].
On the other hand, a writer can be described symbolically using the HSR concept as follows:

$$HSR = \{ (If_1^t, \pi_1^t) ; (If_2^t, \pi_2^t) ; \dots ; (If_P^t, \pi_P^t) \} \tag{2}$$

where If_k^t represents the t^{th} feature subinterval of the k^{th} feature component such that $t = \{1,2, \dots, N_{bins}\}$, and N_{bins} is the number of subintervals tuned experimentally. While π_k^t is the frequency probability attributed to the t^{th} bin of the histogram HSR_k, associated to the k^{th} feature interval (If_k), such that:

$$\pi_k^t = \frac{N_{f_k}}{N} \tag{3}$$

where N is the number of reference signatures, and N_{f_k} is the number of features found within the t^{th} subinterval belonging to If_k.

3 Proposed System

The proposed verification scheme is presented in Fig. 1, and the details of each step are described in next sections.

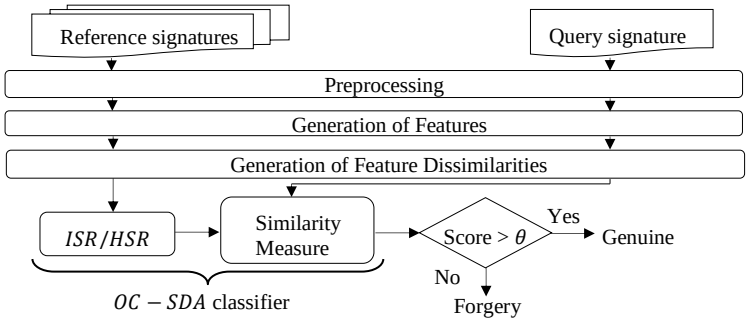

Fig. 1. Scheme of the proposed offline verification system.

3.1 Preprocessing

For this step, an efficient binarization method is specifically performed on the signature image using Local Iterative Method (LIM), followed by a simple signature extraction. LIM is performed through an iterative process for finding the binarization threshold in a sliding window using the mean and the standard deviation [6].

3.2 Generation of Features

For generating features, the Curvelet Transform (CT) is considered in this paper for its efficiency in extracting edges and other singularities along curves. Contrary to the wavelet transform, CT has a high degree of directional specificity elements contained into the curvelet pyramid [7]. Aiming to capture more effectively the local information, the signature image is subdivided into an equi-space grid images before applying the CT. Fig. 2 depicts an example of generating a grid of 3×3.

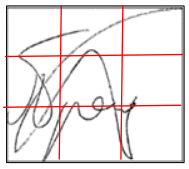

Fig. 2. Example of an equi-space grid image with 3×3.

Hence, a wrapping CT is performed on each grid at the scale j and orientation k, which allows generating curvelet coefficients namely $C_{j,k}$. Next, the energy E is calculated at each scale j and orientation k such as:

$$E(j,k) = \sum_{t_1} \sum_{t_1} \left| C_{j,k}(t_1, t_2) \right| \tag{4}$$

Finally, the feature vector is constructed by concatenating all energy components issued from all grid images.

3.3 Generation of Feature-Dissimilarities (GFD)

Usually, the building of symbolic verification models is based on using straightforward features. In this investigation work, an alternative approach is proposed using the feature-dissimilarities. It consists of performing an absolute difference between each pair of N reference feature vectors of size P namely F_i, and F_l, such that i (and l) $= \{1,2, \dots, N\}$ without redundancy i.e. $i \neq l$, such as $D_u = |F_i - F_l|$, where $u = \{1, \dots, U\}$ and U is the total number of *intra-class* feature-dissimilarity vectors issued from N reference signatures. Each vector D_u of size P can be described as $D_u = \{d_{1u}, d_{2u}, \dots, d_{Pu}\}$, where each feature-dissimilarity component d_{ku} is generated with the respect of its position such as: $d_{ku} = |f_{ki} - f_{kl}|$, where $k = \{1,2, \dots, P\}$. Finally, a matrix namely Matrix of Feature Dissimilarities (MFD) of size $P \times U$ is built

for each writer containing all feature-dissimilarity components taking the following form:

$$\Theta = \{d_{ku}; k = 1, \dots, P; u = 1, \dots, U\} \quad (5)$$

MFD is then handled for creating the writer's model as described in the next section.

3.4 Building the $OC - SDA$ classifier

Basically, two steps are required for building the $OC - SDA$ classifier: Creating the symbolic model and computing the similarity values which is described in the next section (verification process). In this work, two types of Symbolic Representation models (SRM) are considered namely the ISR and HSR models.

Creating the ISR model. The first step consists of creating an interval of feature dissimilarities (instead of features) namely ID_k for each k^{th} feature-dissimilarity component. More precisely, the inferior and superior bounds of ID_k are calculated for each k^{th} column of the matrix Θ using simply the *minimum* and the *maximum* metric, such as:

$$ID_k = [\, d_k^-, d_k^+] \quad (6)$$

Then, each ID_k is symbolically represented by an adaptive weighted distribution function namely ϑ_k, inspired from the real distribution of training feature-dissimilarities. Its mathematic formula is described as follows:

$$\vartheta_k = \begin{cases} 1 & if \quad d_k^- \leq d_{uk} \leq \mu_k \\ e^{-\lambda \cdot d_{uk}} & if \quad \mu_k < d_{uk} \leq d_k^+ \\ 0 & otherwise \end{cases} \quad (7)$$

where λ is a unique control parameter tuned experimentally during the design step, and μ_k is the mean value computed for each ID_k. Hence, the writing style of a writer according to the proposed ISR model is then defined as follows:

$$ISR = \{\vartheta_1, \vartheta_2, \dots, \vartheta_P\} \quad (8)$$

Creating the HSR model. Genuinely, the same interval of feature-dissimilarities ID_k provided in eq. (6) is considered for this symbolic model, and modulated by symbolic histograms, as described in section 2. Hence, the writing style of a writer is defined by a set P of symbolic feature-dissimilarity histograms namely HSR_k, such that $k = 1, \dots, P$.

3.5 Verification process

To verify the authenticity of a query signature represented by a query feature vector of size P namely F, the first step is to perform a straightforward absolute difference between F and the N reference feature vectors F_i such that $i = \{1,2, \dots, N\}$, which belongs

to the claimed writer. This step allows providing N query vectors of feature-dissimilarities having P components namely $D_{qi} = \{d_{q1i}, d_{q2i}, d_{q3i}, \dots, d_{qPi}\}$. Next, a specific similarity measure between all D_{qi} and the generated symbolic model of the claimed writer is performed such as:

$$Sim(D_{qi}, ISR) = \frac{1}{P}\left(\sum_{k=1}^{P} \vartheta_k\right) \tag{9}$$

or:

$$Sim(D_{qi}, HSR) = \frac{1}{P}\left(\sum_{k=1}^{P}\sum_{t=1}^{N_{bins}} \pi_k^t\right) \tag{10}$$

Consequently, N output scores ranged between 0 and 1 are then generated namely $S_q = \{s_{q1}, s_{q2}, s_{q3}, \dots, s_{qN}\}$. In the sequel, a selection rule based on the *maximum* metric is performed for selecting only one representative output score namely s_{qmax}. Finally, the selected score (s_{qs}) is compared to a threshold θ for accepting or rejecting (i.e. genuine or forgery) the query signature (Sig_q) according to the following rule:

$$Sig_q \in \begin{cases} accepted & if \quad s_{qmax} > \theta \\ rejected & otherwise \end{cases} \tag{11}$$

The threshold θ is tuned during the design step using the reference signatures for each writer.

4 Experimental Results

4.1 Dataset Description and Evaluation Criteria

Three offline handwritten signature datasets are used for evaluating the proposed system: GPDS, CEDAR, and MCYT. The GPDS signature dataset [8] contains 300 writers, each one has 24 Genuine Signatures and 30 Forgery Signatures designated as GS and FS, respectively. The CEDAR signature dataset [9], contains 55 writers where each one has 24 GS and 24 FS. While the MCYT dataset [10] which represents genuinely a part of a bimodal database is composed of 75 writers where each one has 15 GS and 15 FS. For the evaluation step, four well-known metrics are used which are "False Rejection Rate" (FRR), "False Acceptance Rate" (FAR), "Average Error Rate" (AER), and the Equal Error Rate (EER).

4.2 Experimental Setup

For designing the proposed system, a small set of M writers is selected randomly from the GPDS-300 dataset ($M = 30$). Next, among the 24 GS of each writer, only 5 GS ($N = 5$) are selected as reference signatures and used for building the symbolic model. The 19 remaining GS are used for finding the optimum configuration parameters by minimizing the AER. On the other hand, three signature datasets namely GPDS,

CEDAR, and MCYT datasets, are considered during the testing phase for evaluating the proposed system.

After the preprocessing step, two parameters should be found during the design step namely N_x and N_y, which represents the number of image partitions performed per line and per column, respectively. Hence, the best configuration found during the design step is $N_x = 3$ and $N_y = 3$. For the classifier parameters, the optimal number of N_{bins} is required when using the $OC - HSR$ classifier. Thus, Table 1 shows the evolution of the AER and EER versus N_{bins} values using five reference signatures ($N = 5$). For better convenience, the obtained AER using the Global Threshold (GT) and Local Threshold (LT) are designated as AER_{GT} and AER_{LT}, respectively.

Table 1. Training results achieved by the $OC - HSR$ classifier for various number of bins.

N_{bins}	**3**	4	5	6	7
AER_{GT} (%)	**13.51**	16.29	20.70	22.98	24.14
AER_{LT} (%)	**12.13**	15.8	19.05	21.6	23.05
EER (%)	**11.49**	14.79	16.84	18.8	21.19

As can be seen, performances decrease gradually as long as N_{bins} increase. Hence, the best performances are obtained for $N_{bins} = 3$. Actually, the use of feature-dissimilarities justifies this result, since the range of feature-dissimilarity values of each component is small. Consequently, there is no need to subdivide again its value greater than 3. In the other hand, the optimal value of λ parameter is required for building the proposed $OC - ISR$ classifier. For better observing the effect of λ values, λ is taken within the interval [0.0001, 5]. Hence, Table 2 shows the evolution of the AER and EER versus λ values using five reference signatures ($N = 5$). For better convenience, only representative results are reported in the table.

Table 2. Training results achieved by the $OC - ISR$ classifier for various value of λ.

λ	0.0001	0.001	0.005	0.01	0.05	0.1	**0.5**	1	2	3	5
AER_{GT} (%)	11.21	11.55	11.43	10.38	10.12	9.85	**8.69**	9.54	10.04	10.28	10.64
AER_{LT} (%)	10.02	10.61	10.52	9.85	9.41	8.97	**7.57**	8.43	9.12	9.61	9.90
EER (%)	8.54	8.82	8.71	7.87	6.61	6.02	**5.42**	5.90	6.51	7.05	8.01

As clearly seen, the optimal value of λ is 0.5 corresponding to the best training verification performance offering 8.69%, 7.57% and 5.42% for AER_{GT}, AER_{LT} and EER, respectively. For better understanding the effect of λ in the verification process, Fig. 3 illustrates the real distribution of training feature-dissimilarities superimposed with the proposed weighted distribution function (ϑ_k) for different value of λ.

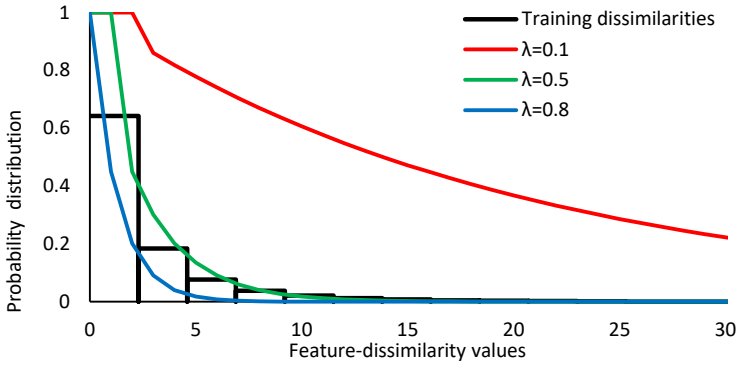

Fig. 3. The probability distribution of training feature dissimilarities superimposed with the proposed weighted membership function ϑ_k for three values of λ (low, optimal and high).

As can be clearly seen, when λ parameter takes small values, the width of ϑ_k shape is large (i.e. red color). While, when λ parameter takes high values, the width of ϑ_k shape is narrow (i.e. blue color). In contrast, almost the same shape of the real dissimilarity distribution is obtained for $\lambda = 0.5$ which corresponds exactly to the optimal λ value reported in Table 2.

4.3 Experimental Evaluation

In this section, the results achieved on the GPDS-300 and GPDS-160 dataset are presented. Adding to that, a blind test is performed on the two signature datasets namely CEDAR and MCYT datasets using five reference signatures ($N = 5$).

Table 3. Verification performances achieved for both $OC - HSR$ and the $OC - ISR$ classifiers.

Dataset	GPDS-300		GPDS-160		CEDAR		MCYT	
Method	*HSR*	***ISR***	*HSR*	***ISR***	*HSR*	***ISR***	*HSR*	***ISR***
AER_{GT} (%)	23.16	**19.14**	22.45	**18.09**	19.23	**15.25**	14.28	**9.44**
AER_{LT} (%)	21.04	**17.06**	19.57	**16.62**	17.53	**12.87**	12.41	**7.97**
EER (%)	19.15	**15.34**	17.69	**12.48**	14.71	**10.84**	9.38	**5.54**

It is clearly shown that the $OC - ISR$ classifier allows obtaining the best performances on all used datasets. Indeed, 15.34% 12.48% 10.84% and 5.54% of EER are obtained for GPDS-300, GPDS-160 and CEDAR and MCYT dataset, respectively. Besides, the signature verification scheme using the local decision threshold allows getting as expected a better performance. The obtained performance on blind datasets especially on MCYT dataset, demonstrates the robustness and the flexibility of the proposed system even when few reference signatures are available.

4.4 Comparative Analysis

For better evaluating the performance of the proposed verification system based on the $OC - ISR$ classifier against the last similar work [4], Table 4 depicts the comparison of cross-validation results achieved on the GPDS-160 dataset using 5, 8, and 12 reference signatures.

Table 4. Comparative analysis from the last similar work on the GPDS-160 dataset.

Work	Descriptor	Classifier	Weighted function	N	$AER_{LT}(\%)$	σ(%)
Alaei et al. [4]	LBP features	$OC - ISR$	Trapezium	5	17.61	0.9
	-	-	-	8	13.85	1.69
	-	-	-	12	11.47	1.99
Proposed	CT dissimilarities	$OC - ISR$	Exponential	5	**16.62**	**0.85**
	-	-	-	8	**13.05**	**0.93**
	-	-	-	12	**11.12**	**1.57**

As highlighted in Table 4, the best performances are achieved by the proposed system on the GPDS-160 dataset for different reference signatures. Indeed, an improvement of 0.99%, 0.8% and 0.35% in AER_{LT} is reported for 5, 8 and 12 reference signatures, respectively. Moreover, the stability of the proposed system is better, according to the standard deviation values. Furthermore, the proposed system requires adjusting the only OC-SDA classifier parameter which is set for all writers. In contrast, Alaei et al. [4] adjust the classifier parameter for each writer which requires more computations. Hence, these results show the effectiveness of the proposed suitable exponential weighted distribution function used for building the $OC - ISR$ classifier against the trapezium weighted function proposed in [4]. Adding to that, the use of dissimilarities seems more suitable for designing symbolic verification models than straightforward features. Indeed, it allows better defining the *intra-class* variability *via* only a few reference signatures.

5 Conclusion

This paper aimed to investigate the use of the $OC - SDA$ for handwritten signature verification. Usually, the symbolic verification models are built straightforward from features. For better capturing the *intra-class* variability, the dissimilarities generated from the curvelet transform are proposed for building the $OC - SDA$ classifier. Hence, two types of the $OC - SDA$ classifier are proposed in this work which are the $OC - ISR$ and the $OC - HSR$ classifiers. For the $OC - ISR$ classifier, a new weighted function based on decreasing exponential distribution is proposed, which is genuinely inspired from the real distribution of training dissimilarities. The experimental evaluation conducted on the three datasets namely GPDS, CEDAR and MCYT dataset, have shown an encouraging improvement offered by the proposed $OC - ISR$ over the $OC -$

HSR classifier. In addition, the proposed verification model based on the $OC - ISR$ classifier outperforms the symbolic verification model proposed in the last similar work. For future work, an interesting work is to use the deep learning for generating features to improve the verification process when using the $OC - ISR$ classifier.

Acknowledgement

This work was supported by the Direction Générale de la Recherche Scientifique et du Développement Technologique (DGRSDT) grant, attached to the Ministère de l'Enseignement Supérieur et de la Recherche Scientifique, Algeria.

References

1. Bertolini D, Oliveira LS, Justino E, Sabourin R (2010) Reducing forgeries in writer-independent off-line signature verification through ensemble of classifiers. Pattern Recognition 43:387-396. doi: 10.1016/j.patcog.2009.05.009.
2. Guerbai Y, Chibani Y, Hadjadji B (2015) The effective use of the one-class SVM classifier for handwritten signature verification based on writer-independent parameters. Pattern Recognition, 48(1):103–113.
3. Billard L, Diday E (2003) Symbolic Data Analysis: Definitions and Examples. Technical Report, available at http://www.stat.uga.edu/faculty/LYNNE/Lynne.html.
4. Alaei A, Pal S, Pal U, Blumenstein M (2017) An Efficient Signature Verification Method Based on an Interval Symbolic Representation and a Fuzzy Similarity Measure. IEEE Transactions on Information Forensics and Security 12(10):2360-2372. doi:10.1109/TIFS.2017.2707332.
5. Hadjadji B, Chibani Y, Nemmour H(2017) An efficient open system for offline handwritten signature identification based on curvelet transform and one-class principal component analysis. Neurocomputing, 265:66-77. doi: 10.1016/j.neucom.2017.01.108.
6. Djoudjai MA, Chibani Y, Abbas N (2017) Offline signature identification using the histogram of symbolic representation. The 5th International Conference on Electrical Engineering-Boumerdes (ICEE-B), Boumerdes, pp 1-6. doi:10.1109/ICEE-B.2017.8192092.
7. Candès E, Donoho D (1999) Curvelets - A surprisingly effective non-adaptive representation for objects with edges. Curves and Surface Fitting: Saint-Malo,Vanderbilt, University Press, Nashville, pp. 105-120.
8. Vargas J, Ferrer M, Travieso C, Alonso J (2007) Off-line Handwritten Signature GPDS-960 Corpus. In *Ninth International Conference on Document Analysis and Recognition (ICDAR)*, Curitiba, Brazil, pp 764-768. doi: 10.1109/ICDAR.2007.4377018.
9. Kalera MK, Srihari S, Xu A (2004) Offline signature verification and identification using distance statistics. International Journal of Pattern Recognition and Artificial Intelligence 18(07):1339 1360. doi: 10.1142/S0218001404003630.
10. Ortega-Garcia J et al. (2003) MCYT baseline corpus: a bimodal biometric database. In IEEE Proceedings Vision, Image and Signal Processing 150(6):395-401. doi:10.1049/ip-vis:20031078.