Multi-script Writer Identification Optimized With Retrieval Mechanism

Chawki Djeddi
LAMIS Laboratory
University of Tebessa
Tebessa, Algeria
c.djeddi@mail.univ-tebessa.dz

Imran Siddiqi
Department of GS & AS
Bahria University
Islamabad, Pakistan
imran.siddiqi@gmail.fr

Labiba Souici-Meslati
LRI Laboratory
Badji Mokhtar-Annaba University
Annaba, Algeria
souici_labiba@yahoo.fr

Abdellatif Ennaji
LITIS Laboratory
Rouen University
Rouen, France
abdel.ennaji@univ-rouen.fr

***Abstract*— identifying the writer of a handwritten document has been an active research area over the last few years with applications in biometrics, forensics, smart meeting rooms and historical document analysis. In this paper, we present a new writer identification system based on a retrieval mechanism. Texture based edge-hinge and run-length features are used to characterize the writing style of an individual. The effectiveness of the proposed system is evaluated on a total of 1583 writing samples in Arabic, German, English, French, and Greek from two different databases. The experimental evaluations reveal that reducing the search space using a writer retrieval mechanism prior to identification improves the identification rates.**

Keywords- edge-hinge features; multi-script handwritten documents; run-length features; writer identification; writer retrieval.

I. INTRODUCTION

Over the last few years, writer identification has been studied and applied in a wide variety of applications, such as, biometric recognition [1], personalized handwriting recognition systems [2], automatic forensic document examination [3], classification of ancient manuscripts [4], and smart meeting rooms [5]. It is defined as a behavioral handwriting-based recognition modality which proceeds by matching unknown handwritings against a database of documents with known writers and is considered as a promising research area today.

The potential applications of writer identification mentioned above have resulted in a renewed research interest of the document analysis and recognition community to improve writer identification methods. Recent advancements in writer identification have resulted in new research directions. These include introducing new features [6], examining the sensitivity of character size on accuracy of writer recognition [7], addressing the problem of writer identification in multi-script environments [8], detecting and removing ruling lines from handwritten documents [9], studying the effectiveness of model perturbed handwriting for writer identification [10], exploring the potential utility to differentiate persons by means of their on-line and off-line writings [11], extending the writer identification task to writer retrieval which is based on the selection of all documents authored by a writer [12], identifying the author of handwritten music scores [13], and, using immunological models in the writer identification task [14].

Especially, Siddiqi et al. [15] have proposed a two-step sequential combination of global and local features to improve writer identification performance. A texture based global analysis first performs a broad classification of writings followed by the use of local features to identify the writer of the query document. Our research is inspired by the same idea of reducing the search space prior to identification task. We, however, do not perform a pre-classification of writings. Instead, we integrate a retrieval mechanism as preprocessing stage to the writer identification system for improving the overall system performance.

The accuracy of writer identification systems is known to decrease as the size of the database increases [16]. This deterioration can be very significant for large-scale identification systems. In such cases, a writer retrieval mechanism that aims at retrieving all the documents written by a specific writer from the database may be used to reduce the search space for a writer identification system.

Based on the same idea, we propose a writer identification system optimized with a retrieval mechanism. The system comprises two main steps. A query document is first presented to the writer retrieval system which compares it with all the documents in the database and retrieves its Top-N nearest neighbors. In the next step, the query document is compared only with the Top-N documents returned by the retrieval system, the comparison being based on features other than those used in the first step.

In this work, we use the probability distributions of run-lengths [1, 17] and edge-hinges [16] as features to characterize the writing style of a writer. Two different scenarios are evaluated where one of these features is used for writer retrieval and the other for identification. The experiments carried out in a multi-script environment and read promising results.

978-0-7695-4774-9/12 $26.00 © 2012 IEEE
DOI 10.1109/ICFHR.2012.239

The rest of this paper is organized as follows: Section 2 gives a brief description of the databases used in our study. Section 3 describes the extraction of the proposed features. We then present the proposed system architecture followed by experimental results and analysis. Finally, we conclude with a discussion on some possible future enhancements to the existing system.

II. DATABASES

We have used two different databases for the experimental evaluation of the system, the IFN/ENIT database [18] and the GRDS database [19].

The IFN/ENIT database [18] is one of the most well known and widely used databases in problems such as handwriting recognition and writer recognition. It consists of forms with handwritten Tunisian town and village names collected from 411 writers, most of which filled 5 forms. The database was mainly designed for training and testing recognition systems for handwritten words and was also used in the ICDAR 2005 Arabic OCR competition [22]. Since the forms also contain the identity of the writer, the same database can also be used for the evaluation of writer identification and verification systems [1, 14, 20]. For our experiments, 1375 samples from 275 different writers have been used.

GRDS [19], a relatively new database, has been created by a research group from the Computational Intelligence Laboratory of the Institute of Informatics and Telecommunications in The National Center for Scientific Research "Demokritos", Greece. This database has been developed with the help of 26 different Greek writers; each copying eight different texts in four different languages (German, English, French, and Greek - two different texts per language). Among all documents, only the Greek documents were written in the native language of the writers. A part of this database was also used in the ICDAR 2009 Handwriting Segmentation Contest [21] while the totality of the database was used in the ICDAR 2011 Writer Identification Contest [19].

All the samples from the two databases were combined to build a multi-script database. This allows studying the effectiveness of the proposed system in a multi-script environment as well. The combined database comprises a total of 1583 documents by 301 different writers (275 Arabic and 26 Greek writers). The next section presents the features used in our system.

III. FEATURE EXTRACTION

The proposed system employs two texture based features to characterize the writer of the given handwritten documents. These features are the probability distributions of run-lengths [1, 17] and edge-hinges [16]. Each of these has been discussed in detail in the following.

A. Run-lengths

Run lengths are computed on a binary image of handwriting where the black pixels correspond to the ink trace and the white pixels correspond to the background, and, the probability distribution of these run-lengths is used to characterize the writing style of a writer. We can define a 'run' as a sequence of connected pixels which have the same color along a given direction. Note that, if A_iA_j is a run comprising pixels A_i, A_{i+1},…, A_{j-1},A_j with an identical color, pixel A_{i-1} must differ in color from pixel A_i, while pixel A_j must differ from pixel A_{j+1}.

During the calculation of run-lengths, the image is scanned in the four principle directions: horizontal, vertical, left-diagonal and right-diagonal. The normalized histogram of these run-lengths is interpreted as a probability distribution function characterizing its writer. The method considers horizontal, vertical, left-diagonal and right-diagonal white run-lengths extracted from the original image as well as black run-lengths (in the same four directions) extracted from the image after applying Sobel edge detection to generate an image in which only the edge pixels are "on".

We then define a run-length matrix P as follows. Each element P(i, j) of the matrix represents the number of runs with pixels of color equal to i and length of run equal to j along a specific direction. The size of the matrix P is N by K, where N is the number of colors in the image and K is equal to the maximum possible run length in the corresponding image.

A direction is defined using a displacement vector d(x, y), where x and y are the displacements for the *x-axis* and *y-axis*, respectively. As discussed earlier, the four principal directions that we can consider include right-diagonal (45°), vertical (90°), left-diagonal (135°) and horizontal (180°). Calculating the run-length encoding for each direction produces a total of four run-length matrices.

The four run-length matrices are converted into (normalized) vectors which are then concatenated to obtain a single vector characterizing the writer of a document. This naturally leads to the problem of the large dimensionality of the feature vector. The maximum possible length of a run is linked to the image size and may not be very meaningful. Moreover, most of the information is present in the initial columns of each feature vector. We therefore truncate each run-length to keep the first 100 columns only giving a feature vector of dimension 800 as summarized in Table 1.

TABLE I. SUMMARY OF RUN-LENGTH FEATURES

Feature		Dimension
Length of black runs	Horizontal	100
	Vertical	100
	Left diagonal	100
	Right diagonal	100
Length of white runs	Horizontal	100
	Vertical	100
	Left diagonal	100
	Right diagonal	100
Total		800

These run-length features provide information on the average width of letters, the density of writing, the structure of letters, the average size of letters, the ink width, the placement of characters, the regions enclosed inside the letters, the blank spaces between letters and words, the regularity and irregularity of handwriting and finally the slope in handwriting. The set of features discussed in the above section is similar to the one that was employed in the ICDAR 2011 Writer Identification Contest [19]. A part of these features was also used in the ICDAR2011 Arabic Writer Identification Contest [23] and in the ICDAR 2011 Music Scores Competition: Staff Removal and Writer Identification [24].

B. *Edge Hinge*

Edge-hinge distribution is a feature that characterizes the changes in direction of a writing stroke in handwritten text and is known to effective in characterizing the writing style [16]. The edge-hinge distribution is extracted by means of a window that is slided over an edge-detected binary handwriting image. Whenever the central pixel of the window is on, the two edge fragments (i.e. connected sequences of pixels) emerging from this central pixel are considered. Their directions are measured and stored as pairs. A joint probability distribution P(φ1, φ2) is obtained from a large sample of such pairs. An example of an angle pair is shown in figure 1.

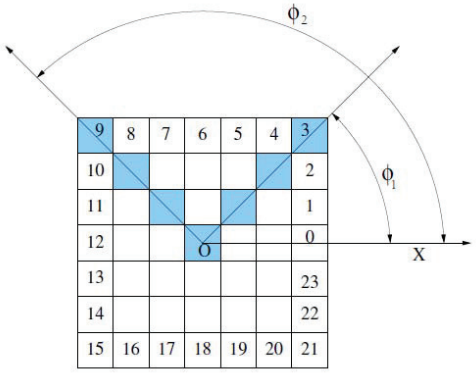

Figure 1. Example of an edge-hinge distribution (image reproduced from [16]).

IV. Writer Retrieval And Identification

As discussed earlier, the overall system comprises two main steps: retrieval and identification. The first step of writer retrieval is based on run-length features (edge-hinge features successively). The features are compared using the Manhattan (Cityblock) distance as metric. In this step, a small subset of *N* documents (*N* chosen to be from 2 to 10 in our case) most similar to the query are selected and the rest are discarded. This step therefore acts as filter excluding more than 99% (from 1581 to 1573 documents) of handwritten documents in the database and keeping less than 1% of the documents (from 2 to 10 documents) to be used in the next step.

In the second step, each of the documents returned by the retrieval step is compared with the query using edge-hinge features (run-lengths features successively). The documents are sorted with increasing distance to the query and those with minimum distances are assumed to be written by the same writer as that of the query image. Figure 2 shows the overall architecture of the proposed system.

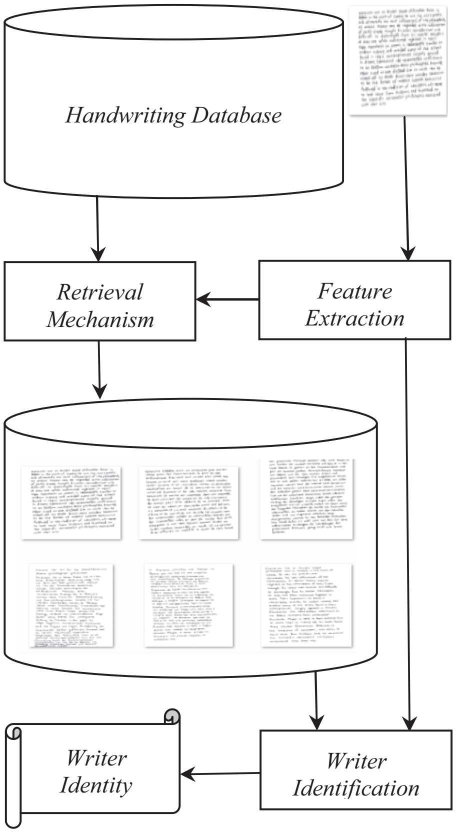

Figure 2. Overview of the proposed system.

V. Experimental Evaluations

This section presents the experiments performed to validate the effectiveness of the proposed features for writer identification and retrieval. We first discuss the metrics used for retrieval and identification and then present the performance of the system without and with writer retrieval. Section *B* presents the writer identification without retrieval; Section *C* presents and discusses the results of writer retrieval while the last section presents the results on writer identification with retrieval.

A. *Evaluation Metrics*

To evaluate the performance of our writer retrieval mechanism, precision and recall are employed; these metrics are widely used in information retrieval, and are defined as follows:

$$Recall = \frac{Number\ of\ relevant\ documents\ retrieved}{Number\ of\ relevant\ documents}$$

$$Precision = \frac{Number\ of\ relevant\ documents\ retrived}{Number\ of\ documents\ retrieved}$$

Precision concerns the proportion of correct-identity documents in the hit list. Recall concerns the proportion of correct-identity documents relative to the total number of documents in the database from the sought writer.

The performance of writer identification is quantified using the well known nearest neighbor rule. More specifically, we calculate the number of documents that were assigned to the correct writer in terms of TOP-1 choice. For every document in the database, the distance among all documents in this database will be calculated using the Manhattan distance metric. A ranked list will be produced and the evaluation will record the accuracy in terms of TOP-1.

B. *Writer Identification without Retrieval Mechanism*

In this section, we present the performance of run-length and edge-hinge features as well as that of their combination on writer identification without retrieval mechanism. The experiments are first conducted on the two databases individually and then on the combined database. The Top1 identification rates are summarized in Table 2. From Table 2, it can be noticed that identification rates vary: from 83.5% to 91.5% for the IFN/ENIT database, from 97.6% to 99.5% for the GRDS database and from 85.3% to 92.4% for the database comprising samples from both the databases.

TABLE II. IDENTIFICATION PERFORMANCE ON IFN/ENIT, GRDS AND IFN/ENIT MIXED WITH GRDS.

Features **Database**	**Run Lengths**	**Edge Hinge**	**Combination**
IFN/ENIT	83.5%	89.2%	91.5%
GRDS	97.6%	99.5%	98.6%
IFN/ENIT + GRDS	85.3%	90.5%	92.4%

C. *Writer Retrieval*

This section presents the performance of run-length and edge hinge features on writer retrieval. We compute the precision and recall for each query document (as discussed in Section *A*) and then determine the precision and recall ratios for the database. These results have been summarized in Tables 3-5 for the IFN/ENIT, GRDS and the mixed database respectively. In comparison with other two scenarios, very good precision and recall rates are observed on the GRDS database. Naturally, this can be attributed to the small number of writers and more samples per writers in this database.

TABLE III. RETRIEVAL PERFORMANCE ON IFN/ENIT DATABASE.

Feature **Top**	**Run-lengths**		**Edge Hinge**	
	Recall	***Precision***	***Recall***	***Precision***
1	20.9%	83.5%	22.3%	89.2%
2	38.0%	76.1%	42.3%	84.6%
3	51.8%	69.1%	58.9%	78.5%
4	62.3%	62.3%	70.6%	70.6%
5	67.0%	53.6%	74.3%	59.5%
6	70.3%	46.8%	76.7%	51.1%
7	72.7%	41.5%	78.6%	44.9%
8	74.5%	37.3%	80.2%	40.1%
9	75.9%	33.8%	81.4%	36.2%
10	77.3%	30.9%	82.6%	33.1%

TABLE IV. RETRIEVAL PERFORMANCE ON GRDS DATABASE.

Feature **Top**	**Run-lengths**		**Edge Hinge**	
	Recall	**Precision**	**Recall**	**Precision**
1	13.9%	97.6%	14.2%	99.5%
2	27.8%	97.4%	28.2%	98.6%
3	41.3%	96.5%	41.5%	96.9%
4	54.5%	95.3%	54.7%	95.8%
5	66.7%	93.5%	66.6%	93.3%
6	78.5%	91.6%	75.6%	88.2%
7	86.7%	86.7%	83.3%	83.3%
8	88.0%	77.0%	86.0%	75.2%
9	88.9%	69.1%	87.1%	67.7%
10	89.9%	62.9%	88.9%	62.2%

TABLE V. RETRIEVAL PERFORMANCE ON IFN/ENIT MIXED WITH GRDS DATABASE.

Feature **Top**	**Run-lengths**		**Edge Hinge**	
	Recall	**Precision**	**Recall**	**Precision**
1	20.0%	85.3%	21.2%	90.5%
2	36.7%	78.9%	40.4%	86.4%
3	50.5%	72.7%	56.5%	80.8%
4	61.3%	66.7%	68.4%	73.8%
5	67.0%	58.8%	73.3%	63.9%
6	71.3%	52.7%	76.4%	55.9%
7	74.5%	47.5%	79.2%	49.9%
8	76.3%	42.5%	80.8%	44.6%
9	77.7%	38.4%	82.0%	40.3%
10	79.0%	35.1%	83.3%	36.8%

D. *Writer Identification with Retrieval Mechanism*

The final series of experiments is aimed at studying the impact of integrating a retrieval mechanism in a writer identification system. Two different scenarios are considered. In the first one, the edge hinge distribution is used in writer retrieval while the run-lengths are used for identification while the inverse is employed in the second. The results of these experiments are reported in Tables 6 and 7.

Comparing with the performance of individual features (Table 2), it can be seen that identification rates improve by integrating the retrieval mechanism in the writer identification system. For the first scenario, the identification rates rise from 83.5% to 92.4% for the IFN/ENIT database, from 97.6% to 99.5% for the GRDS database and from 85.3% to 93.3% for the mixed database. Similar trend can also be observed in Table 7 for scenario II of evaluations.

From Tables 2, 6 and 7, it can be noticed that the achieved identification rates when we combine the run-length and the edge-hinge features are: 91.5% for the IFN/ENIT database, 98.6% for the GRDS database and 92.4% for the mixed database. By incorporating a retrieval mechanism, the identification rates are increased to 92.5%

for the IFN/ENIT database, 99.5% for the GRDS database and 93.3% for the mixed database.

The experimental results reported in this section clearly support the idea put forward in this paper i.e., a writer identification system can be optimized when the query document is compared with a top few retrieved documents (returned by a writer retrieval system) rather than the entire database.

TABLE VI. IDENTIFICATION RESULTS (SCENARIO I).

Scenario I			
Number of Retrieved Documents	*IFN/ENIT*	*GRDS*	*IFN/ENIT + GRDS*
2	88.4%	99.5%	89.8%
3	89.9%	99.5%	91.2%
4	90.5%	99.5%	91.7%
5	91.0%	99.5%	92.2%
6	91.2%	99.5%	92.3%
7	91.8%	99.5%	92.9%
8	92.1%	99.5%	93.1%
9	92.4%	99.5%	93.3%
10	92.2%	99.0%	93.1%

TABLE VII. IDENTIFICATION RESULTS (SCENARIO II).

Scenario II			
Number of Retrieved Documents	*IFN/ENIT*	*GRDS*	*IFN/ENIT + GRDS*
2	91.7%	99.5%	92.7%
3	92.5%	99.5%	93.3%
4	91.9%	99.5%	93.0%
5	91.8%	98.6%	92.7%
6	91.7%	98.1%	92.5%
7	91.3%	97.6%	92.0%
8	90.8%	97.1%	91.6%
9	90.8%	97.1%	91.6%
10	90.4%	97.1%	91.3%

VI. CONCLUSION AND FUTURE WORK

In this paper, we have proposed a writer identification system based on a retrieval mechanism which reduces the search space of the identification process. We used the probability distributions of run-length and edge-hinge features to characterize the handwritten documents. Two databases containing Arabic, German, English, French, and Greek samples are used to evaluate the effectiveness of the proposed approach and the experimental results reveal the usefulness of having a retrieval mechanism prior to identification.

Currently, our work is based on the extraction of global features, but further work will focus on the use of local features. An integrated system will be considered to combine both local and global features to produce more reliable classification accuracy. We are now conducting some experiments with larger databases containing samples from different scripts. Also, the proposed system can be extended to include a rejection threshold to reject any writers that are not a part of our databases.

REFERENCES

[1] Djeddi, C., Souici-Meslati, L.: “A texture based approach for Arabic Writer Identification and Verification”. In Proc. of the 1st International Conference on Machine and Web Intelligence, pp. 88-93, Algeria, 2010.

[2] A. Nosary, L. Heutte, and T. Paquet, “Unsupervised writer adaption applied to handwritten text recognition”, In Pattern Recognition vol. 37, 2004, pp. 385–388.

[3] M. Van Erp, L. Vuurpijl, K. Franke, and L. Schomaker, “The WANDA measurement tool for forensic document examination”, Journal of Forensic Document Examination, 16:103–118, 2005.

[4] I. Siddiqi, F. Cloppet and N. Vincent, “Contour Based Features for the Classification of Ancient Manuscripts”, In Proc of The 14th Conference of the International Graphonomics Society, France, 2009.

[5] M. Liwicki, A. Schlapbach, H. Bunke, S. Bengio, J. Mariéthoz, J. Richiardi.: “Writer Identification for Smart Meeting Room Systems“, IDIAP research report IDIAP-RR 05-70, 2005.

[6] I. Siddiqi, N. Vincent, “Text independent writer recognition using redundant writing patterns with contour-based orientation and curvature features”, in Pattern Recognition 43 (2010), pp : 3853 – 3865.

[7] M. Ozaki, Y. Adachi and N. Ishii, “Examination of Effects of Character Size on Accuracy of Writer Recognition by New Local Arc Method”, In Proc of The International Conference on Knowledge-Based Intelligent Information and Engineering Systems, Bournemouth, United Kingdom, LNCS, 2006, Volume 4252, 2006, pp : 1170-1175.

[8] U. Garain, T. Paquet, “Off-line Multi-script Writer Identification using AR Coefficients“,In Proc of the International Conference on Document Analysis and Recognition, Spain, pp. 991-995, 2009.

[9] J. Chen, D. Lopresti, E. Kavallieratou, “The Impact of Ruling Lines on Writer Identification”, In Proc of the 12th International Conference on Frontiers in Handwriting Recognition, India, pp :439-444, 2010.

[10] J. Chen, W. Cheng and D. Lopresti, “Using Perturbed Handwriting to Support Writer Identification in the Presence of Severe Data Constraints”, In Proc of the DRR XVIII, 2011, USA, pp. 78740G-1 - 78740G-8.

[11] A. Chaabouni, H. Boubaker, M. Kherallah, A.M. Alimi and H.E. Abed, “Combining of Off-line and On-line Feature Extraction Approaches for Writer Identification”, In Proc of the 11th International Conference on Document Analysis and Recognition, pp. 1299-1303, China, 2011.

[12] V. Atanasiu, L. Likforman-Sulem and N. Vincent, “Writer Retrieval—Exploration of a Novel Biometric Scenario Using Perceptual Features Derived from Script Orientation”, In Proc of the 11th International Conference on Document Analysis and Recognition, pp. 628-632, 2011.

[13] A. Fornés, J. Llados, G. Sanchez, “Writer Identification in Old Handwritten Music Scores”, In Proc of the 8th International Workshop on Document Analysis Systems, 2008, pp. 347-353.

[14] C. Djeddi and L. Souici-Meslati, “Artificial Immune Recognition System for Arabic Writer Identification”, In Proc of the 4th International Symposium on Innovation in Information & Communication Technology, pp. 159-165, Jordan, 2011.

[15] I. Siddiqi, N. Vincent, “Combining Global and Local Features for Writer identification”, in In Proc of the 11th International Conference on Frontiers in Handwriting Recognition, 2008, pp : 48 – 53.

[16] M. Bulacu, “Statistical Pattern Recognition for Automatic Writer Identification and Verification”. PhD thesis, University of Groningen, 2007.

[17] X. Tang, “Texture Information in Run-Length Matrices”. In IEEE Transactions on Image Processing, Vol. 7, No. 11, pp. 1602-1609, 1998.

[18] M. Pechwitz, S. Maddouri, V. M¨argner, N. Ellouze, H. Amiri, “IFN/ENIT-database of handwritten arabic words“. In Proc of Colloque International Francophone sur l’Ecrit et le Document, 2002, pp : 129 - 136.

[19] G. Louloudis, N. Stamatopoulos and B. Gatos, “ICDAR 2011 - Writer Identification Contest”, In Proc of the 11th International Conference on Document Analysis and Recognition, pp. 1475-1479, China, 2011.

[20] M. Bulacu, L. Schomaker, A. Brink, “Text-Independent Writer Identification and Verification on Off-Line Arabic Handwriting”. In Proc of the 9th International Conference on Document Analysis and Recognition, pp. 769–773, 2007.

[21] B. Gatos, N. Stamatopoulos and G. Louloudis, "ICDAR2009 Handwriting Segmentation Contest", In Proc of the 10th International Conference on Document Analysis and Recognition, pp. 1393-1397, Spain, 2009.

[22] Margner, V., Pechwitz, M., El Abed, H.: “ICDAR 2005 arabic handwriting recognition competition”. In Proc of the 8th International Conference on Document Analysis and Recognition, pp. 70- 74, 2005.

[23] A.Hassaine, S. Al-Maadeed, J.M. Alja’am, A. Jaoua and A. Bouridane, “The ICDAR2011 Arabic Writer Identification Contest”, In Proc of the 11th International Conference on Document Analysis and Recognition, pp. 1470-1474, China, 2011.

[24] A. Fornés, A. Dutta, A. Gordo and J. Llados, “The ICDAR 2011 Music Scores Competition: Staff Removal and Writer Identification”, In Proc. of the 11th International Conference on Document Analysis and Recognition, pp. 1511-1415, China, 2011.