ICFHR2012 Competition on Writer Identification - Challenge 2: Arabic Scripts

Abdelâali Hassaïne, Somaya Al Maadeed
Computer Science and Engineering Department
College of Engineering, Qatar University
Doha, Qatar
$\{hassaine, s\_alali\}$@*qu.edu.qa*

Abstract

Arabic writer identification is a very active research field. However, no standard benchmark is available for researchers in this field. The aim of this competition is to gather researchers and compare recent advances in Arabic writer identification. This competition has been hosted on Kaggle, it has attracted forty-three teams from both academia and industry. This paper gives details on this competition, including the dataset used, the evaluation procedure and description of participating methods and their performances.

1. Introduction

Writer identification helps forensic experts in taking their decisions regarding the authenticity of a certain document. It also makes it possible to improve the performance of handwriting recognition by the mean of personalized recognizers.

Writer identification is a very active research field; in the ICFHR 2010, not less than 6 papers addressed this research area [2, 5, 9, 12, 14, 22] and in ICDAR 2011, around 5 papers dealt also with this field [3, 6, 8, 7, 17].

Moreover, three contests have been organized within ICDAR 2011 on writer identification. One dealing with English, French, German and Greek [20], another one dealing with music scores [13], and we have organized a contest dealing with Arabic scripts [16].

This competition is a follow-up to the last year's edition. The aim of this competition is to allow researchers and industries working in writer identification or related fields to compare the performances of their systems on a new unpublished data.

This competition has been organized through *Kaggle* which is a platform for data prediction competitions. It allows companies, governments and researchers to post their data in order to have scientists from all over the world compete on it and produce optimum solutions [15].

This competition has attracted 43 participants, among which sixteen participants agreed to share their identities and short descriptions of their methods. In the next section, we describe the dataset used in this competition. Short descriptions of the participating methods are given in section 3. Evaluation procedure is described in section 4. Discussion and conclusions are presented in the two final sections.

2. Dataset

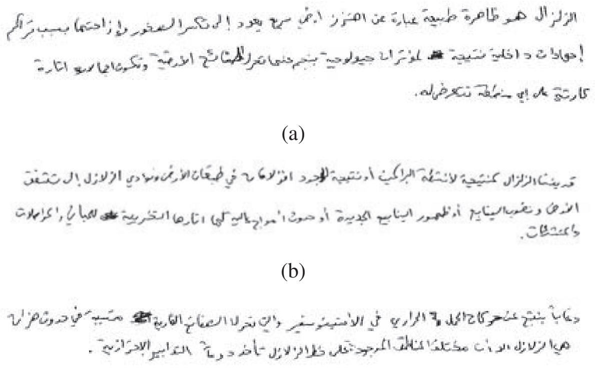
In this competition, 206 writers were asked to write three different paragraphs in Arabic. The first two paragraphs are used for training and the third one is used for testing. For some writers, the first two paragraphs have been removed from the training set in order to test the ability of systems to detect unknown writers. Also, in order to prevent participants from assigning each image to and only to one writer, some writers have reproduced the third paragraph more than once. Figure 1 shows an example of these paragraphs.

Images were acquired using an "EPSON GT-S80" scanner, with a 600 DPP resolution. Images were provided in PNG binary format. The binarization has been performed using Otsu's method.

Note that this dataset is a subset of a large dataset that will be made available progressively through different evaluation campaigns [1].

Motivated by the fact that most Kaggle users are data scientists without necessarily an image-processing background, we have provided all the participants a set of more than 30 features extracted from all the images. These features are based on the lengths of branches

978-0-7695-4774-9/12 $26.00 © 2012 IEEE
DOI 10.1109/ICFHR.2012.218



(a)

(b)

(c)

Figure 1. Example of three paragraphs written by the same writer.

in the skeleton, handwriting thickness, tortuosities, directions, curvatures, chain codes and codebook distributions. Those provided features correspond to histograms of several values, bringing the total length of feature vectors to 6272 values.
The whole dataset of images as well as the corresponding features can be downloaded from:
`http://www.kaggle.com/c/awic2012/data`

Participants were free to use the provided features or other extracted features or even a combination of both.

3. Participants

In this competition, 43 teams submitted a total of 578 entries. The final leaderboard can be found here:
`www.kaggle.com/c/awic2012/leaderboard`

3.1. Benchmarks

Some standard writer identification methods have been made available to all participants to serve as benchmarks. This methods were based on the Edge-Based Directional Features (EDBF) [4], with varying number of directions (4, 8, 12 and 16).

3.2. Participating methods

The following teams accepted to share their identities and/or a description of their methods.

Marcos Sainz This method uses Principal Component Analysis and normalization to filter out noise in the provided features and dramatically reduce the number of dimensions, then a 1 nearest neighbor classifier is used to assign each document to the closest class.

Sashi Dareddy The proposed method uses extensive feature selection using Boruta algorithm which utilizes a random forest classifier at its core [18]. This brought down the number of the provided features to about 500. Classification has been performed using Sparse Partial Least Squares with extensive parameter selection on 10-fold cross validation [19].

Wayne Zhang from the Department of Information Engineering, The Chinese University of Hong Kong. This method combines the provided features with the edge-hinge and grapheme features introduced in [25]. Kernel principal component analysis [23] is applied after applying a random sampling linear discriminant analysis [26] in order to reduce dimensionality. The distance between documents is obtained by averaging distances with regards to several feature sets. Finally, support vector machines have been used to detect unknown writers.

AWReS Submitted by Chawki Djeddi, from LAMIS, Tebessa University, Algeria, Labiba Souici-Meslati from LRI, Annaba University, Algeria and Abdellatif Ennaji from LITIS, Rouen University, France. This method is based on two types of features: edge-hinge features and run-lengths features. In addition, the provided features have also been used. Edge-hinge features estimate the joint distribution of edge angles in a writer's handwriting. They are constructed by performing an edge detection using a Sobel kernel on the input images, and subsequently, measuring the angles of both edge segments that emanate from each edge pixel. Run-lengths features [10] are determined on the binary image taking into consideration both the black pixels corresponding to the ink trace and the white pixels corresponding to the background. The probability distribution of horizontal, vertical, left-diagonal and right-diagonal black and white run-lengths has been used. Classification is performed using one against all support vector machines classifier. This method does not handle unknown writers.

steinke Submitted by the University of Applied Sciences and Arts. This participants did not provide us with any details of their method.

Luciferase This method classifies the provided features using a multilayer perceptron (MLP) which is a feedforward artificial neural network model that maps the sets of input features onto the corresponding most probable writer.

cess_northumbria Submitted by Muhammad Atif Tahir and Ahmed Bouridane from the Computer and Electronic Security Systems Research Group, University of Northumbria, UK. This method combined the proposed features with kernel collaborative representation and multiscale local binary patterns. This method have been successfully applied in face recognition [24].

ihata Submitted by Talha Karadeniz from the Middle East Technical University, Turkey. This method used SIFT and Brief descriptors which are extracted from 'relatively' dense keypoints (i.e. instead of keypoint detection, binarized pixel values are used). Estimated covariance matrices of the descriptors are subsequently merged in order to form the final feature vector of each image. Classification is then performed using 1 nearest neighbor.

YT Submitted by Yanir Seroussi from Monash University, Melbourne, Australia. The core of this approach was to use SVMs with a diffusion kernel, which has proven to be suitable for comparing instances based on histograms [11]. The basic diffusion kernel has been extended by weighting histograms for different features based on the cross-validated accuracy obtained when using each feature alone. To detect unknown writers, SVMs are used in a one-versus-all setup after setting a threshold on the distance from the hyperplane.

William Cukierski from the Center for Biomedical Imaging and Informatics, Rutgers University, USA. This method runs a principal component analysis on the provided features, tuning for the optimal dimension, which has been found to be around 80. Then, it runs a linear discriminant analysis on the principal components. In order to identify unknown writers, this method thresholds the logarithms of the unconditional predictive probability density of the sample observations.

Foxtrot This method uses a feature selection algorithm in order to reduce dimension of the provided features and then applies a 1 nearest neighbor classifier.

Ben Hamner from Kaggle. This method applied 1 nearest neighbor classifier on the provided features.

Han & Kilian Submitted by Kilian Mie and Han Wang, from the University of California, Berkeley. This method simply applies a distance learning metric after reducing the features to about 400.

D33B Submitted by Ahmed El Deeb from Microsoft Egypt. This method combines an identification approach with a verification approach. The identification approach uses an improved nearest neighbor algorithm by considering the neighborhoods as the hypercubes defined by the two samples for each class. Furthermore, the logarithm of the difference in each feature is used (instead of the absolute distance) in order to achieve more stability. The verification approach classifies the difference between feature vectors using a multilayer perceptron.

bfs Submitted by Wei LI from The Chinese University of Hong Kong. This method uses dual space linear discriminant analysis with simple regularization model for training with all the training samples, and transforms the whole data set using the resulting transformation matrix. Then, using a weighted k-nearest neighbor model to perform prediction on the testing set while mining confident samples from the testing set. Confident samples from the testing set are subsequently merged with the training set to improve classification.

Newell and Griffin Submitted by Andrew Newell and Lewis Griffin from University College London. At the core of this method is a system called oriented Basic Image Feature columns (oBIF columns) [21]. The description of oBIFs begins with Basic Image Features (BIFs). In this system every location in an image is assigned to one of seven classes according to local symmetry type, which can be dark line on light, light line on dark, dark rotational, light rotational, slop, saddle-like or flat. The class is calculated from the output of six Derivative-of-Gaussian filters. An extension to the BIF system is to include local orientation, depending on local symmetry type, to produce oriented Basic Image Features (oBIFs). As for the matching step, this method used a simple nearest neighbour classifier. The unknown writers were identified after assigning each test image to its nearest training image in oBIF column space.

4. Evaluation

Participants were asked to produce, for each image of the test set, the ID of the most probable writer (among the writers of the training set), or zero when it is most likely that the writer is unknown. The methods are ranked according to their identification rate (IR):

$$IR = \frac{number\ of\ images\ correctly\ identified}{total\ number\ of\ images}.$$

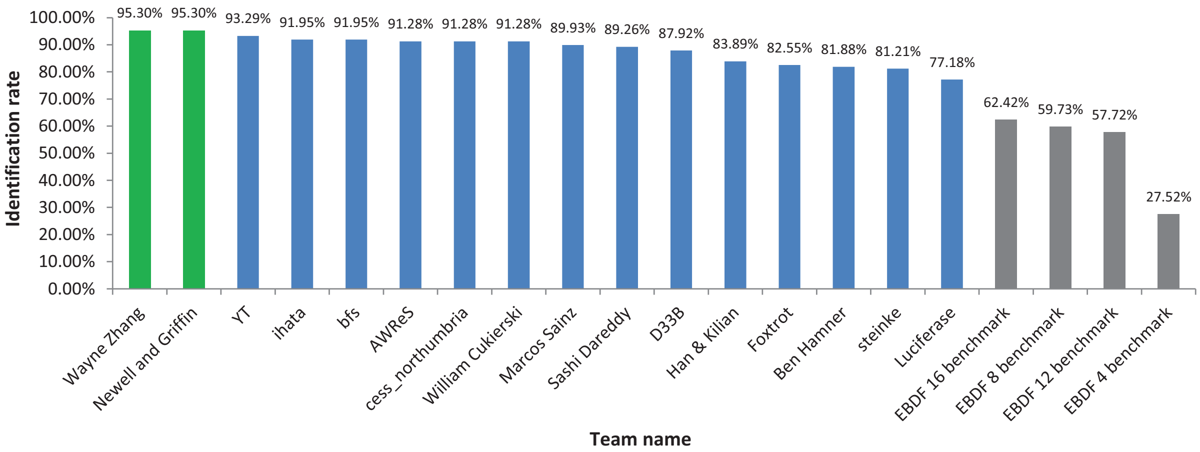


Figure 2. Identification rates of the participating methods.

It has to be noted that Kaggle displays a public leaderboard which allows participants to see how well they perform comparing to other methods. This public leaderboard is computed on a part of the test set which does not count toward the final standing (35% of the test set in this competition). The results are computed on the remaining part of the test set and are not shown to participants before the end of the competition.
Figure 2 shows the results of the above mentioned teams. The best performance is jointly achieved by "Wayne Zhang" and "Newell and Griffin".

5. Discussion

After analyzing the submitted methods, we noticed the following:

- Several participants used a feature reduction approach, using generally linear discriminant analysis has been generally preferred. This suggests that not all the provided features were discriminative. In order to further investigate the importance of the provided features with regards to this dataset, we have computed the identification rate of logistic regression classifiers based on each category of features separetly, the results are summed-up in figure 3. This clearly indicates that, curvatures, directions and chain code features are far more discriminative than the other features.

- It is interesting to note that several top-ranking participants made use of test data in unsupervised learning to improve their classification.

- Not all the participants handled the problem of unknown writers (most participants only assigned them to the closest writer). For those who handled unkwown writers, SVMs have generally been preferred. Table 1 sums up the performance of these methods in detecting unknown writers.
  Note that F-measure=2 $\cdot \frac{Precision \cdot Recall}{Precision + Recall}$.

Table 1. Performances of the participating teams in detecting unknown writers

Team name	Precision	Recall	F-measure
Wayne Zhang	0.6667	1	0.8
Newell and Griffin	0.4286	0.75	0.5455
YT	0.25	0.25	0.25
bfs	0.5	0.25	0.3333
William Cukierski	0.6	0.75	0.6667
Han & Kilian	0.6667	0.5	0.5714

6. Conclusion

This Writer Identification Contest for Arabic scripts has been organized in order to allow researchers and industries in writer identification or related fields to compare the performances of their systems on a new unpublished data. This contest has been organized through Kaggle and has also been made available to data scientists by providing a large set of features extracted from all the images.

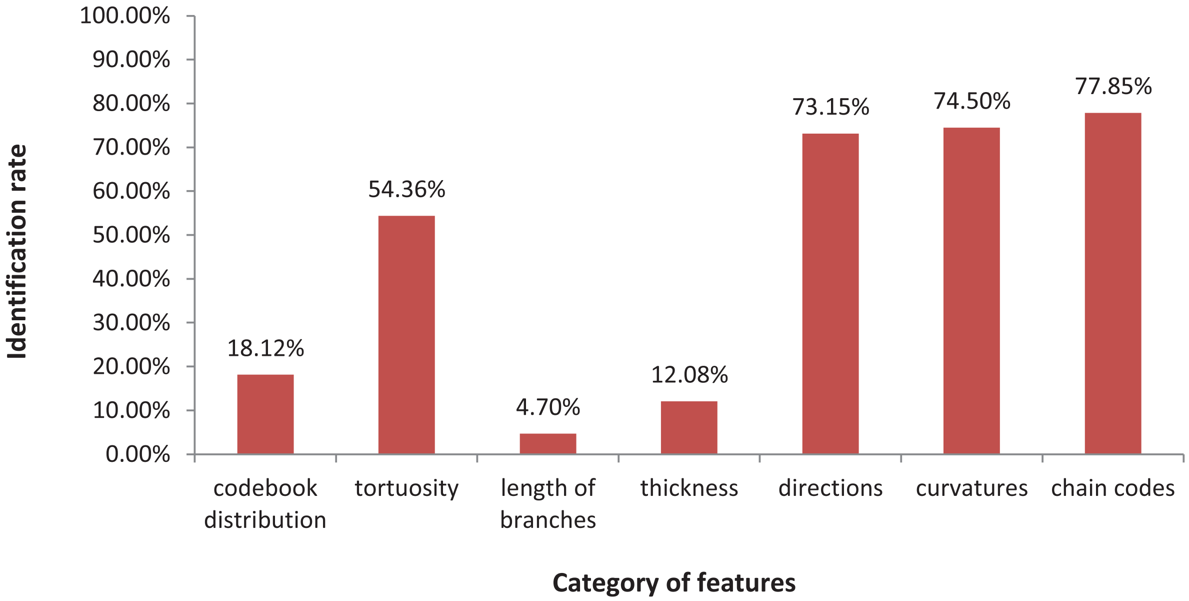


Figure 3. Identification rates of each category of features.

The objective of this contest is fulfilled by providing a comparison between all the participating methods and by making the benchmarking dataset publicly available. The best performance is jointly achieved by "Wayne Zhang" from The Chinese University of Hong Kong and "Newell and Griffin" from University College London. For future editions of this contest, it is planned to provide handwritings of a larger set of writers with different backgrounds in both Arabic and English languages in order to obtain a more detailed comparison between the systems.

Acknowledgment

This work is supported by Qatar National Research Fund (QNRF) grant through National Priority Research Program (NPRP) No. 09 - 864 - 1 – 128.
We would like to thank Kaggle for hosting this competition as well as all the participants. We would also like to thank all the volunteers who dedicated their valuable time and effort in the data collection process.

References

[1] S. Al-Maadeed, W. Ayoubi, A. Hassaïne, and J. Alja'am. QUWI: An Arabic and English Handwriting Dataset for Offline Writer Identification. In *Frontiers in Handwriting Recognition (ICFHR), 2012 International Conference on*, sep. 2012.

[2] G. Ball, S. Srihari, and R. Stritmatter. In *Frontiers in Handwriting Recognition (ICFHR), 2010 International Conference on, title=Writer Verification of Historical Documents among Cohort Writers*, pages 314 –319, nov. 2010.

[3] Q. A. Bui, M. Visani, S. Prum, and J. Ogier. Writer Identification Using TF-IDF for Cursive Handwritten Word Recognition. In *Document Analysis and Recognition (ICDAR), 2011 International Conference on*, pages 844 –848, sept. 2011.

[4] M. Bulacu, L. Schomaker, and L. Vuurpijl. Writer Identification Using Edge-Based Directional Features. In *Proceedings of the Seventh International Conference on Document Analysis and Recognition - Volume 2*, ICDAR '03, pages 937–, Washington, DC, USA, 2003. IEEE Computer Society.

[5] H. Cao, R. Prasad, and P. Natarajan. Improvements in HMM Adaptation for Handwriting Recognition Using Writer Identification and Duration Adaptation. In *Frontiers in Handwriting Recognition (ICFHR), 2010 International Conference on*, pages 154 –159, nov. 2010.

[6] H. Cao, R. Prasad, and P. Natarajan. OCR-Driven Writer Identification and Adaptation in an HMM Handwriting Recognition System. In *Document Analysis and Recognition (ICDAR), 2011 International Conference on*, pages 739 –743, sept. 2011.

[7] A. Chaabouni, H. Boubaker, M. Kherallah, A. Alimi, and H. Abed. Combining of Off-line and On-line Feature Extraction Approaches for Writer Identification. In *Document Analysis and Recognition (ICDAR), 2011 International Conference on*, pages 1299 –1303, sept. 2011.

[8] A. Chaabouni, H. Boubaker, M. Kherallah, A. Alimi, and H. Abed. Multi-fractal Modeling for On-line Text-Independent Writer Identification. In *Document Analysis and Recognition (ICDAR), 2011 International Conference on*, pages 623 –627, sept. 2011.

[9] J. Chen, D. Lopresti, and E. Kavallieratou. The Impact of Ruling Lines on Writer Identification. In *Frontiers in Handwriting Recognition (ICFHR), 2010 International Conference on*, pages 439 –444, nov. 2010.

[10] C. Djeddi and L. Souici-Meslati. A texture based approach for Arabic writer identification and verification. In *Machine and Web Intelligence (ICMWI), 2010 International Conference on*, pages 115 –120, oct. 2010.

[11] H. Escalante, T. Solorio, and M. Montes-y Gómez. Local histograms of character N-grams for authorship attribution. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 288–298, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[12] R. Fernandez-de Sevilla, F. Alonso-Fernandez, J. Fierrez, and J. Ortega-Garcia. Forensic Writer Identification Using Allographic Features. In *Frontiers in Handwriting Recognition (ICFHR), 2010 International Conference on*, pages 308 –313, nov. 2010.

[13] A. Fornes, A. Dutta, A. Gordo, and J. Llados. The ICDAR 2011 Music Scores Competition: Staff Removal and Writer Identification. In *Document Analysis and Recognition (ICDAR), 2011 International Conference on*, pages 1511 –1515, sept. 2011.

[14] A. Fornes, A. and and J. Llado ands. A Symbol-Dependent Writer Identification Approach in Old Handwritten Music Scores. In *Frontiers in Handwriting Recognition (ICFHR), 2010 International Conference on*, pages 634 –639, nov. 2010.

[15] A. Goldbloom. Data Prediction Competitions – Far More than Just a Bit of Fun. In *Data Mining Workshops (ICDMW), 2010 IEEE International Conference on*, pages 1385 –1386, dec. 2010.

[16] A. Hassaïne, S. Al-Maadeed, J. Alja'am, A. Jaoua, and A. Bouridane. The ICDAR2011 Arabic Writer Identification Contest. In *Document Analysis and Recognition (ICDAR), 2011 International Conference on*, pages 1470 –1474, sept. 2011.

[17] R. Jain and D. Doermann. Offline Writer Identification Using K-Adjacent Segments. In *Document Analysis and Recognition (ICDAR), 2011 International Conference on*, pages 769 –773, sept. 2011.

[18] M. Kursa, A. Jankowski, and W. Rudnicki. Boruta A System for Feature Selection. *Fundamenta Informaticae*, 101(4):271–285, 2010.

[19] K. Le Cao, P. Martin, C. Robert-Granie, and P. Besse. Sparse canonical methods for biological data integration: application to a cross-platform study. *BMC Bioinformatics*, 10(1):34, 2009.

[20] G. Louloudis, N. Stamatopoulos, and B. Gatos. ICDAR 2011 Writer Identification Contest. In *Document Analysis and Recognition (ICDAR), 2011 International Conference on*, pages 1475 –1479, sept. 2011.

[21] A. Newell and G. L.D. Natural image character recognition using oriented basic image features. In *Digital Image Computing: Techniques and Applications (DICTA), Proceedings of the International Conference on*, 2011.

[22] P. Purkait, R. Kumar, and B. Chanda. Writer Identification for Handwritten Telugu Documents Using Directional Morphological Features. In *Frontiers in Handwriting Recognition (ICFHR), 2010 International Conference on*, pages 658 –663, nov. 2010.

[23] B. Schölkopf, A. Smola, and K. Müller. Kernel principal component analysis. In W. Gerstner, A. Germond, M. Hasler, and J.-D. Nicoud, editors, *Artificial Neural Networks ICANN'97*, volume 1327 of *Lecture Notes in Computer Science*, pages 583–588. Springer Berlin / Heidelberg, 1997. 10.1007/BFb0020217.

[24] M. Tahir and A. Bouridane. Face Recognition using Kernel Collaborative Representation and Multiscale Local Binary Patterns. In *IET Image Processing Conference*, July 2012.

[25] L. van der Maaten and E. Postma. Improving automatic writer identification. In *Proc. of 17th Belgium-Netherlands Conference on Artificial Intelligence*, pages 260–266, 2005.

[26] X. Wang and X. Tang. Random sampling LDA for face recognition. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 2, pages II–259 – II–265 Vol.2, june-2 july 2004.