The ICDAR2011 Arabic Writer Identification Contest

Abdelâali Hassaïne, Somaya Al-Maadeed, Jihad Mohamad Alja'am, Ali Jaoua
Computer Science and Engineering Department
College of Engineering, Qatar University
Doha, Qatar
$\{hassaine, s\_alali, jaam, jaoua\}$@*qu.edu.qa*

Ahmed Bouridane
CEIS, Northumbria University
Newcastle upon Tyne, UK
ahmed.bouridane@northumbria.ac.uk

***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 was hosted by Kaggle, it has attracted thirty participants from both academia and industry. This paper gives details on this competition, including the evaluation procedure, description of participating methods and their performances.**

***Keywords*-Arabic writer identification; Kaggle; performance evaluation;**

I. 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 ICDAR 2009, around 8 papers dealt with writer identification [6, 8, 11–13, 16, 17, 19] and in the ICFHR 2010, not less than 6 papers addressed also this research area [1, 2, 5, 7, 14, 15]. However, to the extent of our knowledge, never a competition has been organized in this field. 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and industries working in writer identification or related fields to compare the performances of their systems on a new unpublished data.

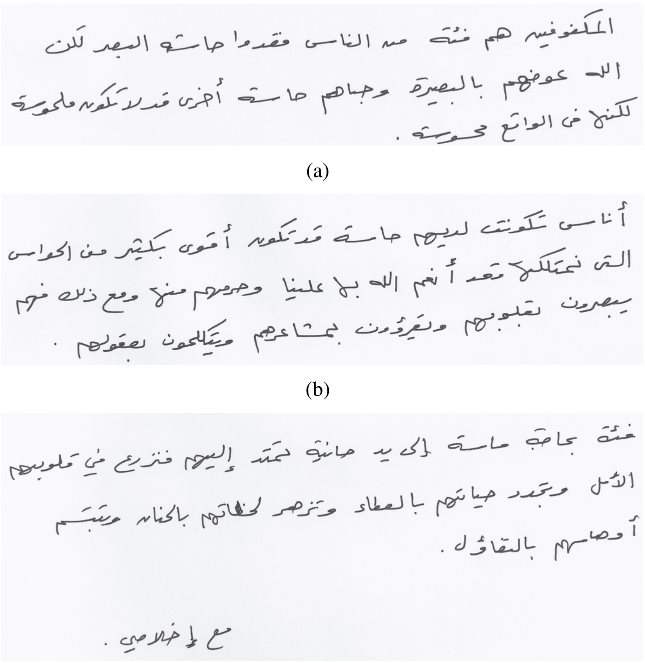
This competition has been organized through Kaggle which is a platform for data prediction competitions. It allows companies, governments and researchers to post their data in order to have scientists from all over the world compete on it and produce optimum solutions. Kaggle has achieved extraordinary results that have outperformed betting markets and advanced the state of the art in HIV research and chess ratings [10].

This competition has attracted thirty participants, among those seven 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 4. Evaluation procedure is described in section 5. Conclusion and future work are drawn in the final section.

II. Dataset

In this competition, 54 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. Images were provided in PNG color, gray and binary format. The binarization has been performed using the Otsu's method. Figure 1 shows an example of these paragraphs.



(a)

(b)

(c)

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

Motivated by the fact that most Kaggle users are machine learning scientists without necessarily an image-processing background, and also inspired by the German traffic sign recognition competition [18], we have provided all the participants a set of more than 70 features extracted from all the images. These features are based on the number of connected components, number of holes, moments, projections, distributions, position of barycenter, number of branches in the skeleton, Fourier descriptors, tortuosities, directions, curvatures and chain codes. A detailed description of these features will soon be available in an extended publication. Participants were free to use the provided features or other extracted features or even a combination of both.

III. Participating methods

In this competition, 30 teams submitted a total of 139 entries. The final leaderboard can be found here:
http://www.kaggle.com/c/WIC2011/Leaderboard
Among those, the following teams accepted to share their methodology:

Eu Jin Lok: Submitted by Eu Jin Lok from Deloitte Inc., Australia. This team used only the provided features. Because of the fact that there were too few instances to train the algorithm (2 paragraphs for each author), but there were way too many predictors (6,480), this method tried to tackle this problem by combining both train and test, factor analysing the predictors and reducing them down to about 300. The algorithms have then been trained to predict the target as multiple categories (53 authors). This method ended up working very well as it reduced training time and improved accuracy in prediction despite the target being multiple categories. Several classification algorithms have been tried including Logistic, BayesNet, SVM, RandomForest and Boosting. This method does not handle unknown writers.

Intelligentia: Submitted by Tri Kurniawan Wijaya from Knowledge Based Systems Group, Vienna University of Technology, Austria and Philips Kokoh Prasetyo from CrimsonLogic Inc., Singapore. This method used only the provided features, it applies several data preprocessing techniques and simple statistical analysis. First, training and testing data are put together. Then, the predictors are normalized to [0,1]. Then, the predictors which have high variance or which have 50% or more of the values equal to 0 or 1 are removed. Instead of having only 2 data for each author (from data training), 3 additional data are added which are max, min and average value of each attribute for the corresponding author. This leads to 5 data for each author.
As for the classification, the 1-nearest neighbor algorithm using Manhattan distance has been applied, this simple method outperforms other advanced distance measures such as cosine distance or euclidean distance.

Robin: Submitted by Enrico Glaab from University of Nottingham, UK. The author adapted his ensemble and consensus analysis of biological datasets method [9]. This method obtained significant performance improvements using a semi-supervised strategy for feature selection in combination with ensemble learning. Using the simplest possible classifier (kNN) and different distance measures (normalized Euclidean distance and Mahalanobis distance). The author applied different search algorithm for feature subset selection (greedy best first search and different evolutionary search methods) and scored the subsets on the leave-one-out cross-validation accuracy on the training data (in the 1. iteration) and repeated this analysis iteratively on the combined training and test set data (using the estimated test set labels from the previous iteration in the scoring function). This strategy provided significant performance improvements and might also be applicable in combination with other more successful learning algorithms used in this competition.

Team Shasta: Submitted by Greg Werner from George Washington University, USA. This team used simulated annealing to perform classification using the provided features. Given n pairs where n is the number of authors in the training set, we perform simulated annealing such that we ended up with n triplets. In our analysis, we performed a Monte Carlo simulation involving 100 runs of the simulated annealing process. Using basic probability, we accepted confidence levels for our triplets of 95% and over as matches. Whereas very obvious matches are brought out with this method, it can be more difficult to decide when there is a false match suggested. The simulated annealing process worked much better when we started from a known configuration rather than starting it from a random configuration. For this competition, we performed two separate "greedy" evaluations on each possible triplet. First, using all of the features provided by the author team, we calculated which test script gave the lowest score triplet for a given training pair. Second, we performed a rank heuristic between each training pair and each test script. Whichever test script was the best match on the most features was declared to be the winner. In practice, we applied the first heuristic which gave quite good results by itself and then refined gray areas with the second heuristic. We were able to also successfully pick out a "No Author" with the first heuristic.

UCL: This method has been submitted by Andrew Newell and Lewis Griffin from the Department of Computer Science, University College London, UK. At the core of this method is a system called oriented Basic Image Feature columns (oBIF columns) [3]. 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.

Wifahd: Submitted by Chawki Djeddi from Mathematics and Computer Science Department, Cheikh Larbi Tebessi University, Tebessa, Algeria and Labiba Souici-Meslati from LRI Laboratory, Computer Science Department, Badji Mokhtar University, Annaba, Algeria. In the feature extraction step, this method used run lengths features proposed in [4]. Run lengths are determined on the binary image taking into consideration either the black pixels corresponding to the ink trace and the white pixels corresponding to the background. The probability distribution of black and white run-lengths has been used in our writer identification experiments. There are four scanning methods: horizontal, vertical, left-diagonal and right-diagonal. We calculate the runs lengths features using the Grey Level Run Length Matrices and the histogram of run lengths is normalized and interpreted as a probability distribution. Our particular implementation considers horizontal and left-diagonal for white run-lengths and horizontal, vertical, left-diagonal and right-diagonal for black run-lengths. For the classification step, we combined four different classifiers: a multilayer perceptron (MLP), two Support Vector Machine classifiers (SVM One against all, SVM one against one) and a K-nearest neighbor classifier (K-NN) with Manhattan Distance Metric.

Wride: This method has been submitted by Laurens van der Maaten from Delft University of Technology, The Netherlands. It is based on two types of features: multi-scale edge-hinge features and grapheme features [20]. In addition, the provided chain code 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.

Grapheme features estimate the distribution by which a writer generates so-called graphemes. Graphemes are constructed by following the handwriting, and making a "cut" at locations where the sign of the y-direction of the handwriting changes. From the thus obtained graphemes, a codebook of prototypical graphemes is constructed using k-means clustering. Each writer may be considered a probabilistic generator of graphemes in the codebook.

Classification is performed by combining a 1-nearest neighbor classifier using Euclidean distance and a boosted logistic regressor. A classification is only accepted if the average of the two posteriors is higher than a certain threshold. If none of the writer labels satisfies the criterion, a label is not assigned (unknown labels are not handled by this method).

IV. Evaluation

Participants were asked to produce, for each image x of the test set, and for each writer i, a probability score $p(x,i)$ indicating how probable it is that image x is written by writer i ($i=0$ for unknown writer).

The systems are ranked according to their mean absolute error (MAE):

$$MAE = \sum_{N}^{x=1} \sum_{M}^{i=0} |p(x,i) - GT(x,i)|.$$

Where N is the number of images in the test set, M the number of writers and $GT(x,i)=1$ if the image x is written by the author i and $GT(x,i)=0$ otherwise.

We have also computed the identification rate (IR) which is the percentage of correctly identified writers.

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 (30% 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.

Table I shows the results of the above mentioned teams. The best performance is achieved by UCL team who managed to obtain a 100% correct classification.

Table I
PERFORMANCES OF THE PARTICIPATING TEAMS

Team name	**Rank**	**MAE**	**Identification rate**
UCL	1	0.00000	100%
Team Shasta	2	0.00400	89.19%
Wride	2	0.00400	81.08%
Eu Jin Lok	4	0.00551	78.38%
Intelligentia	5	0.00801	78.38%
Wifahd	6	0.00901	75.68%
Robin	7	0.03504	5.45%

These results are illustrated in figures 2 and 3.

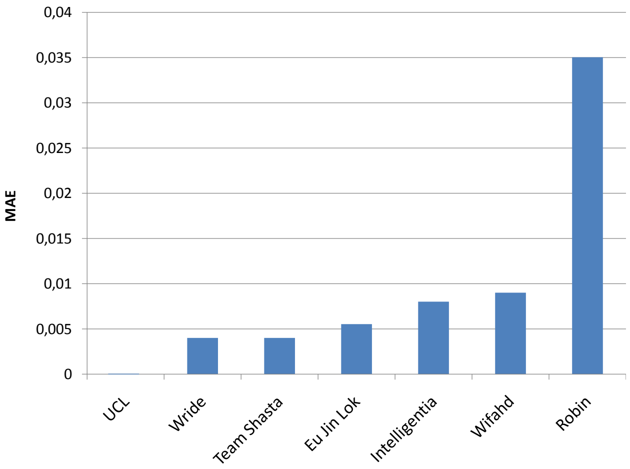


Figure 2. Mean absolute errors of the participating methods.

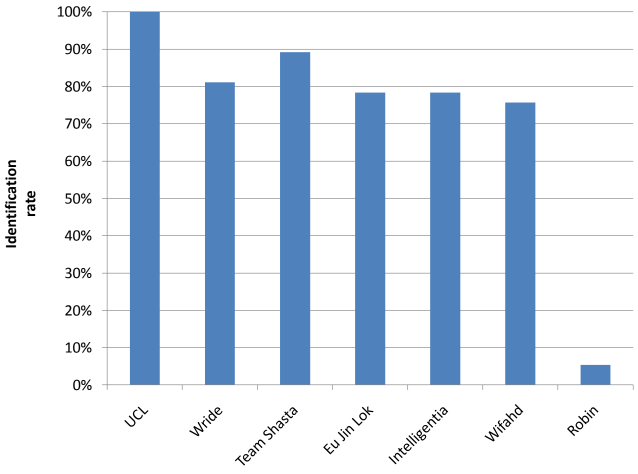


Figure 3. Identification rates of the participating methods.

V. Conclusion

This first Arabic Writer Identification Contest 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.

The objective of this contest is fulfilled by providing a comparison between all the participating methods and by making the benchmarking dataset freely available.

The winning method was the one submitted by Andrew Newell and Lewis Griffin from UCL.

For future editions of this contest, a large acquisition campaign is currently being organized. It is planned to collect handwritings of more than 1000 writers with different backgrounds in both Arabic and English languages in order to obtain more detailed comparisons 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] H. Cao, R. Prasad, and P. Natarajan. Improvements in HMM Adaptation for Handwriting Recognition Using Writer Identification and Duration Adaptation. In *2010 12th International Conference on Frontiers in Handwriting Recognition*, 2010.

[2] J. Chen, D. Lopresti, and E. Kavallieratou. The Impact of Ruling Lines on Writer Identification. In *2010 12th International Conference on Frontiers in Handwriting Recognition*, 2010.

[3] M. Crosier and L. Griffin. Using basic image features for texture classification. *International Journal of Computer Vision*, 88:447–460, 2010.

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

[5] R. Fernandez-de Sevilla, F. Alonso-Fernandez, J. Fierrez, and J. Ortega-Garcia. Forensic Writer Identification Using Allographic Features. In *2010 12th International Conference on Frontiers in Handwriting Recognition*, 2010.

[6] A. Fornes, J. Llados, G. Sanchez, and H. Bunke. On the use of textural features for writer identification in old handwritten music scores. In *Document Analysis and Recognition, 2009. ICDAR'09. 10th International Conference on*, pages 996–1000. IEEE, 2009.

[7] A. Forns and J. Llados. A Symbol-dependent Writer Identification Approach in Old Handwritten Music Scores. In *2010 12th International Conference on Frontiers in Handwriting Recognition*, 2010.

[8] U. Garain and T. Paquet. Off-Line Multi-Script Writer Identification Using AR Coefficients. In *Document Analysis and Recognition, 2009. ICDAR'09. 10th International Conference on*, pages 991–995. IEEE, 2009.

[9] E. Glaab, J. Garibaldi, and N. Krasnogor. Arraymining: a modular web-application for microarray analysis combining ensemble and consensus methods with cross-study normalization. *BMC Bioinformatics*, 10(1):358, 2009.

[10] 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.

[11] B. Li, Z. Sun, and T. Tan. Hierarchical shape primitive features for online text-independent writer identification. In *Document Analysis and Recognition, 2009. ICDAR'09. 10th International Conference on*, pages 986–990. IEEE, 2009.

[12] B. Li and T. Tan. Online Text-independent Writer Identification Based on Temporal Sequence and Shape Codes. In *Document Analysis and Recognition, 2009. ICDAR'09. 10th International Conference on*, pages 931–935. IEEE, 2009.

[13] D. Pavelec, LS. Oliveira, E. Justino, F.D.N. Neto, and LV Batista. Author Identification Using Compression Models. In *2009 10th International Conference on Document Analysis and Recognition*, pages 936–940. IEEE, 2009.

[14] P. Purkait, R. Kumar, and B. Chanda. Writer Identification for Handwritten Telugu Documents using Directional Morphological Features. In *2010 12th International Conference on Frontiers in Handwriting Recognition*, 2010.

[15] G. R. Ball, S. N. Srihari, and R. Stritmatter. Writer Identification of Historical Documents Among Cohort Writers. In *2010 12th International Conference on Frontiers in Handwriting Recognition*, 2010.

[16] F. Shahabi and M. Rahmati. A New Method for Writer Identification of Handwritten Farsi Documents. In *2009 10th International Conference on Document Analysis and Recognition*, pages 426–430. IEEE, 2009.

[17] I. Siddiqi and N. Vincent. A set of chain code based features for writer recognition. In *2009 10th International Conference on Document Analysis and Recognition*, pages 981–985. IEEE, 2009.

[18] J. Stallkamp, M. Schlipsing, J. Salmen, C. Igel, and G. Schner. The german traffic sign recognition benchmark. In *International Joint Conference on Neural Networks*, San Jose, California, July 31 - August 5 2011.

[19] G.X. Tan, C. Viard-Gaudin, and A.C. Kot. Impact of alphabet knowledge on online writer identification. In *Document Analysis and Recognition, 2009. ICDAR'09. 10th International Conference on*, pages 56–60. IEEE, 2009.

[20] 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.