A Deep Learning based System for Writer Identification in Handwritten Arabic Historical Manuscripts

Michel Chammas[a,b], Abdallah Makhoul[b], Jacques Demerjian[c], Elie Dannaoui[a]

[a] *Digital Humanities Center, University of Balamand, El-Koura, Lebanon*
[b] *Femto-ST Institute, UMR CNRS 6174, Université de Bourgogne Franche-Comté, Montbéliard, France*
[c] *LaRRIS, Faculty of Sciences, Lebanese University, Fanar, Lebanon*

Abstract

Determining the writer or transcriber of historical Arabic manuscripts has always been a major challenge for researchers in the field of humanities. With the development of advanced techniques in pattern recognition and machine learning, these technologies have been applied to automate the extraction of paleographical features in order to solve this issue. This paper presents a baseline system for writer identification, tested on a Historical Arabic dataset of 11610 single and double folio images. These texts were extracted from a unique collection of 567 Historical Arabic Manuscripts available at the Balamand Digital Humanities Center. A survey has been conducted on the available Arabic datasets and previously proposed techniques and algorithms. The Balamand dataset presents an important challenge due to the geo-historical identity of manuscripts and their physical conditions. An advanced Deep Learning system was developed and tested on three different Latin and Arabic datasets: ICDAR19, ICFHR20 and KHATT, before testing it on the Balamand dataset. The system was compared with many other systems and it has yielded a state-of-the-art performance on the new challenging images with 95.2% mean Average Precision (mAP) and 98.1% accuracy.

Keywords: Writer identification, historical documents, artificial intelligence, document

Email addresses: michel.chammas@balamand.edu.lb (Michel Chammas), abdallah.makhoul@univ-fcomte.fr (Abdallah Makhoul), jacques.demerjian@ul.edu.lb (Jacques Demerjian), elie.dannaoui@balamand.edu.lb (Elie Dannaoui)

analysis, Arabic manuscripts

1. Introduction

This paper introduces an adaptive deep learning based system that works on identifying the authorship of unidentified historical Arabic documents. This issue has always been a limitation for the study of historical texts, where a lot of documents have a lack of information about their origin, date, authorship, biometric [11] and paleographical features [9]. While many researchers worked to solve this issue, a lot of ambiguities and challenges remain in this area. Furthermore, the lack of Arabic datasets has limited the progress of testing algorithms on Arabic Handwritten documents. As a researcher at the Digital Humanities Center at the University of Balamand, I noticed the need to have an automated system that works on detecting those unidentified historical documents using machine learning [10].

The Digital Humanities Center has a unique database that contains a large number of digitized and transcribed manuscripts. The dataset consists of a large set of historical Arabic documents, more than 500 manuscripts owned by the center and hundreds imported from different areas in the middle east. This huge corpus is defined by the following important features: a large volume of textual heritage, a wide variety of text format, a broad period of time (from 13th to the 19th century), a vast geographic scope (from middle east and north Africa), and a wide variety of Vorlagen (translations). The average number of pages is around 150 per manuscripts, which means many thousands of pages and hundreds of authors. Around 60% of the documents have identified authors and dates while 40% are still unrecognized. This ratio is very ideal to train and test a deep neural network algorithm.

Different artificial neural networks (ANN) algorithms are being used recently for

writer identification like Recurrent Neural Network (RNN), Convolutional Neural Networks (CNN), Multilayer perceptron (MLP)... A survey was done in over different identification and verification systems, where CNN has been identified as the best model used in comparing to other methods like probability distribution function along with features extraction and encoding processes. CNN proved good results in predicting the genre and handedness of authors using documents from English and Arabic handwriting datasets. The developed system was very robust and accurate for identifying the authorship of historical documents, it has yielded the best performance when comparing it to state-of-the-art algorithms in this field. The system was compared with different systems on different datasets.

The novelty of this work stands on two perspectives: the uniqueness of the Arabic Historical dataset and the high accuracy of the system on both Latin and Arabic datasets. There is a lack of experiments on such dataset, where most of the previous studies were tested on Latin Historical and Non-Historical datasets. While the Arabic datasets were very limited to non-historical manuscripts and good physical condition. The challenges were to achieve high results on different Arabic and Latin datasets with the same system and to tackle the challenging physical conditions of the manuscripts without any pre-processing techniques. Mentioning that no previous study has been tested on such manuscripts. Most of the systems use pre-processing techniques to avoid the bad physical conditions of the manuscripts, while our system has proved to surpass any conditions. It was tested on four different Latin and Arabic datasets and it was compared with different systems. No previous studies have such wide comparison. Adding to this, the system results has surpassed all the other techniques.

The rest of this paper is organized as follows: Section 2 presents the related works in this field and available Arabic datasets. Section 3 describes the corpus used in this

study. The proposed system and its functionality are presented in Section 4. Section 5 evaluates the system results, and Section 6 concludes the paper.

2. Related works

This section reviews related works to writer identification for Arabic documents and the available datasets. According to Djeddi et al. [19] the first study in this field was in 2005, when Al Zoubaidy et al. proposed a global approach based on multi-channel Gabor filter and Grey Level co-occurrence Matrices (GLCM). They applied it to a small database of 500 samples for 20 writers. Many other researchers worked with the same methods for feature extraction and used traditional classifiers, where most experiments were using the IFN/ENIT database [29] [7]. This dataset consists of 2200 documents for 411 writers and each document contains 12 names of Arabic towns/villages. Djeji [19] used part of this dataset (650 documents for 130 writers) to compare basic classifiers like Support Vector Machine (SVM), k-Nearest Neighbors (kNN), and Naïve Bayes with the Artificial Immune Recognition System (AIRS) classifier. He extracted 640 features using GLCM and achieved the best result (85.77% accuracy) with AIRS when using all features together. Also, Hannad et al. tested the same dataset (all 411 writers) by applying the histogram of oriented gradients (HOG) for feature extraction and they achieved a better result (86.62% accuracy). Both datasets had a limited number of samples and texts.

Abdleazeem et al. [4] introduced another dataset in 2008, the Arabic Digit database (ADBase) which consists of 70000 Arabic digits for 700 writers. This dataset contains only digits and it was used by several researchers for writer identification [6].

Gazzah et al. proposed a new approach by using Multi-Layer Perceptron (MLP) for features extraction. He compared MLP with SVM and achieved better results on a small dataset of 180 documents for 60 writers. Al-Aziz el al. used spatial gray level dependence (SGLD) to classify eight different texture features: Correlation, Inverse

difference moment, Contrast, Angular second moment, Entropy, Mean, Sum of squares Covariance and Covariance. They worked on a dataset of old Arabic Manuscripts from the Mamluk and Ottoman age (90 documents from 10 books).

In the International Conference on Frontiers in Handwriting Recognition (ICFHR) in 2014 [33], two datasets have been used in the first edition of the Arabic Writer Identification competition: The Arabic Handwritten Text Images Database (AHTID/MW) and KHATT. AHTID/MW dataset was introduced for the first time in the competition. It contains 3710 text lines and 22,896 words written by 53 native writers of Arabic. KHATT was first introduced in ICFHR 2012 [26][25], it was the largest dataset in terms of pages (2000 paragraphs) for 1000 writers. According to Slimane et al. [33], four different systems competed on those two datasets: writer Identification Line Level/ Steerable Pyramids (WILL/SP), Edge-hinge and Multi-scale Run-length (EMR), Gaussian mixture model supervectors (GMM) and oriented basic image / scale-invariant feature transform (OBI/SIFT). The GMM model achieved the best results with 73.4% accuracy. This model was submitted by Christlein et al. [13], where they implemented RootSIFT as a feature extraction method. The SIFT descriptors were used to train the GMM, which serves as a Universal Background Model (UBM).

In 2014, Fecker et al. [21][20] presented a small Arabic historical dataset and tested it using basic features extraction methods: Contour Based Features (CON), OBI, SIFT and HOG. The dataset consisted of 4595 images from 29 writers where only 174 images of them were unknown. Then in 2017, they compared the same methods on two different Arabic datasets KHATT (non-historical) and WAHD (historical) [5]. In all experiments SIFT achieved the best results by far, so they compared different key point detector techniques by using SIFT as a key descriptor. They used pre-processing pipeline on the images before applying the feature extraction methods. They cropped the background to eliminate the noise and applied text segmentation. The WAHD dataset was previously introduced in 2017 by Abdelhaleem et al. [3] and it was tested using two classifiers:

GMM and OBI/SIFT. The dataset contains 43976 images for 322 writers, where all writers are known. According to their experiments GMM proved better results than the combination of OBI/SIFT. In 2019, Hannad et al. [23] improved the results on the KHATT dataset by using a combination of HOG and Gray Level Run Length (GLRL) Matrices.

Latest studies on Writer identification were made on Latin Historical datasets. Fiel et al. [22] and later Christlein et al. [17] showed the importance of using Convolutional Neural Network (CNN) in writer identification. Christlein et al. introduced the combination of using SIFT as a feature extraction method along with CNN as a classifier and proved a state-of-the-art results. After that, Lai et al. [24] introduced a new method by encoding Pathlet and SIFT for feature extraction and they clinched the best result in the ICDAR 19 competition. The best performing methods were using Vector of Locally Aggregated Descriptors (VLAD) encoding to create a global descriptor from the extracted features.

3. The corpus

The manuscripts in Arabic script have been produced constantly starting the seventh century. Since the eighteenth century, they have begun to coexist with printed books (incunabula). Arabic manuscripts were produced in Islamic and non-Islamic areas as well, either as stand-alone codices or as bi-lingual compilations. They cover a vast geographic scope extending from the Atlantic Ocean to the China Sea and from Zanzibar to the banks of the Volga. The surviving Arabic manuscripts cover a wide variety of scientific, cultural and religious topics. Although the study of Arabic codicology and paleography has been approached from philological and literary perspectives, this field is still insufficiently explored. Previous studies [8] have had almost exclusively focused on

studying the manuscripts individually rather than comprehensive quantitative analysis[1].

In this study, a special collection of Arabic manuscripts is used, it is described as follows:

Figure 1: Sample image from the Balamand Arabic Historical Manuscripts.

3.1. Volume

The corpus consists of 567 manuscripts collected from various collections conserved at Antiochian Orthodox monasteries and bishoprics in Lebanon and Syria [1]. These manuscripts are digitized at the University of Balamand[2]. The dataset contains 11610 digital photos[3]. For medium and large manuscripts, most of the photos are produced in portrait format; each photo corresponds to one page (recto or verso of one folio). In the case of small manuscripts, two pages (the verso and recto of two consecutive folios) are

[1]Mention the Arabic corpora in the domain of OCR and handwritten recognition.

[2]These manuscripts were digitized by the Saint Joseph of Damascus Manuscript Conservation Center (http://www.balamandmonastery.org.lb/index.php/about-the-center) and the Digital Humanities Centre (http://iohanes.uob-dh.org/?q=en/tags/digital-humanities).

[3]The total number of digitized pages exceeds the number of photos.

Figure 2: Sample image from the Balamand Arabic Historical Manuscripts.

combined into one landscape photo.

3.2. Date

The date of the manuscript is usually mentioned in the colophon[4]. Being located at the end of the codex, the colophon is exposed to various risks: lost folio, damage due to poor conservation conditions, illegal traffic and circulation, etc. For this reason, the date is susceptible to loss. In the used dataset, 409 manuscripts have their date recognized, whereas 158 manuscripts remain without any explicit information about their date.

3.3. Copyists

The information collected from the dataset led to identifying the names of 256 copyists who produced 329 manuscripts; a dataset may contain more than one manuscript copied by the same hand.

[4]"A statement providing information regarding the date, place, agency, or reason for production of the manuscript or other object" [2]

Figure 3: Sample image from the Balamand Arabic Historical Manuscripts.

Figure 4: Sample image from the Balamand Arabic Historical Manuscripts.

Table 1: Collection by date

Century	Number of manuscripts
19	163
18	122
17	96
16	20
15	5
14	2
13	1
Unkown	158

Table 2: Collection by copyist

Copyist status	Number of manuscripts	Percentage %	Number of Copyists
Identified	329	58	256
Unidentified	238	42	n/a
Total	567	100	n/a

3.4. Topology of the MSS (Geography)

The manuscripts are collected from 14 repositories in Lebanon and Syria:

3.5. Paleographical and codicological features

Since the paper used in these manuscripts is not lined, Miṣṭara[5] was used to define the writing area, the lines, the margins and the layout of the page. The dimensions of the written area and the number of lines reflect some features related to the geographical and chronological distribution [31]. Decorators and painters of these Christian manuscripts used to develop iconic and aniconic ornamentation (tahīb in Arabic). Aniconic ornamentation is based on three components: geometry (taṣṭīr), stylized vegetal motifs with arabesque (tawrīq), and epigraphy. Decorative elements are found at the beginning and

[5] A frame made of cardboard or occasionally of wood on which cords of various thickness could be stretched, corresponding to the text frame lines and guidelines [18].

Table 3: Collection by geographical location

Country	Repository	Number of manuscripts
Lebanon		422
	Balamand	197
	Bechmezin	15
	Bekhaaz	1
	Btaaboura	1
	Dimitrios	14
	Douma	16
	Haref	16
	Hamatoura	49
	Brumana	15
	Shwaya	62
Syria		145
	Hmayra	47
	Homs	56
	Lian	42

the end of text units. The text is usually written in black while the red is used for titles, drop caps, rubrics and punctuation. Western watermarked paper was used in Arabic oriental manuscripts by the end of the thirteenth century. Since the majority of the manuscripts in the corpus of this study are copied after that date, most of them are manufactured using western watermarked paper.

4. System description

The writer identification system was built on three main functions: feature extraction, classification and encoding. Normalization was used twice after the feature extraction and classification in order to improve recognition performance. Figure 7 shows the descriptive diagram of the system workflow, and the possible algorithms in each process based on the previous studies. The algorithms used in this system are in red.

4.1. Features extraction

First, image samples were localized for features extraction using an unsupervised Scale Invariant Feature Transform (SIFT) [15] method. The advantage of using SIFT is

Figure 5: Sample image from the Balamand Arabic Historical Manuscripts.

to preserve the sample properties and keeps them invariant from transformations such as scaling, rotation and translation. Image patches (32x32) are extracted from the document image at the center of each key point. Then, the image patches are mapped with their corresponding SIFT descriptors to create a 128-dimension vector.

Afterwards, a Principal Component Analysis (PCA) method (with whitening) [12] was applied on the SIFT descriptors to improve recognition performance and reduce vectors dimension, in order to obtain a compact representation.

This process reduces descriptors dimension from 128 to 32, which decreases the processing cost of the clustering.

4.2. Clustering

A random batch of descriptors (75 million) were selected from hundred millions extracted through SIFT. The descriptors were grouped into 5000 clusters (with an initial size of 15000) using mini batch k-means clustering algorithm [14]. The clusters were used as labels for image patches, which are labeled according to the closest cluster centre.

Figure 6: Sample image from the Balamand Arabic Historical Manuscripts.

Next, patches were extracted randomly (990k) with their corresponding labels (cluster centroids) from the training data (275 patches per document). The extracted patches were used to train a deep Convolutional Neural Network (CNN), which operates as a feature extractor [28]. The trained model is given a 32x32 patches with their associated clusters. The trained network was a Residual Neural Network with 20 layers (ResNet20), where 900k patches were used for training and 90k for testing. ResNet20 is composed of residual building blocks with multiple branches. Some branches contain two or more convolutional layers, while others simply forward the result of the preceding layers, consequently avoiding the other branch. These building blocks assist in the preservation of the model identity and allow deeper training [15].

A Universal Background Model (UBM) was created by extracting Local descriptors from the training documents using CNN (64-dimension vector). The descriptor vectors

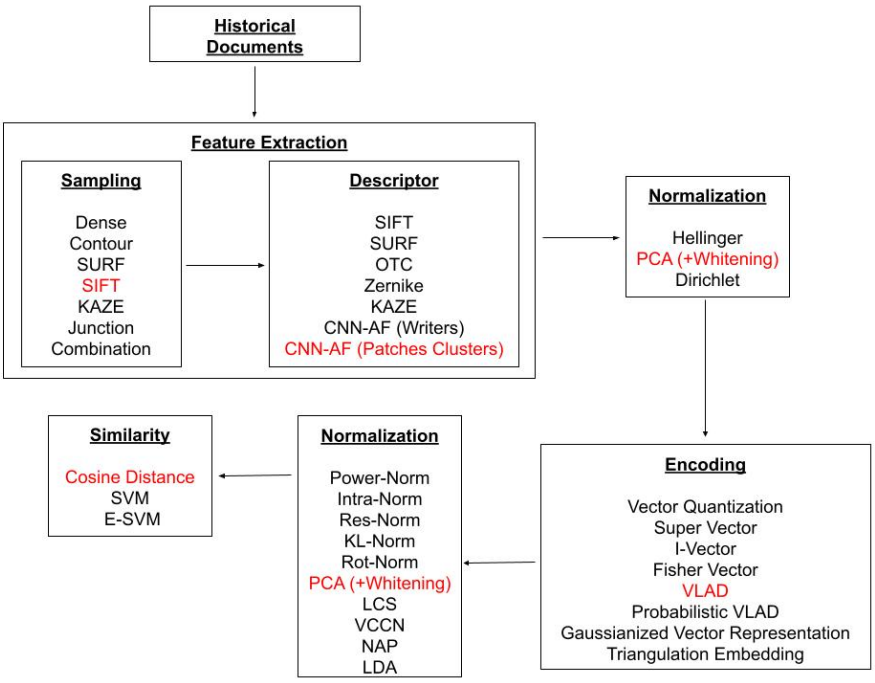

Figure 7: System workflow diagram for all algorithms in each process (used algorithms in red).

were clustered in 100 centroids using k-means.

4.3. CNN training

The image patches were mapped into their respective labels using CNN, which was used for extracting features [28]. The data were split for training (900k) and testing (90k). A deep residual network with 20 layers (ResNet20) was implemented to extract the local descriptors for each document [22]. Figure 7 shows ResNet20 accuracy. Following the feature extraction using CNN, An UBM model was created using trained descriptors, which will be clustered using k-means (100 clusters). At the end, l2-norm was applied for data normalization [15].

$$l2(v) = ||v||2 \tag{1}$$

$$||v||2 = \sqrt{a1^2 + a2^2 + a3^2} \tag{2}$$

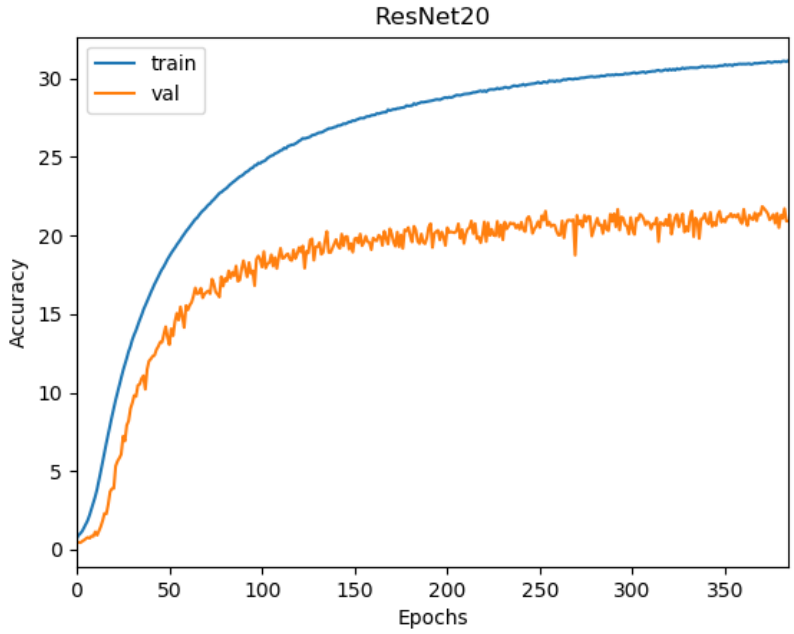

Figure 8: ResNet20 training and validation accuracy.

4.4. Encoding

A VLAD codebook [30] was created via k-means clustering of all the local descriptors and was used to encode each document [16]. The VLAD performs as UBM, where the local descriptors of each document were aggregated into a global descriptor vector with a dimension of 6400. Next, it was normalized via l2 norm [13].

A multi-VLAD approach was used to improve the encoding performance, where 5 different codebooks were created and concatenated into global descriptor vector with a dimension of 32000. At the end, the global descriptor was reduced to the dimension of 3200 via regularized PCA [16].

$$\Sigma = \frac{1}{n-1}\left((\mathbf{X} - \bar{\mathbf{x}})^T\ (\mathbf{X} - \bar{\mathbf{x}})\right) \tag{3}$$

$$\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{4}$$

4.5. Similarity

Finally, The documents were compared by computing the cosine distance between their VLAD vectors. An Exemplar-SVM (ESVM) with a linear kernel was implemented [27][15], which results an improved accuracy and precision.

5. Results

This system has been tested first on the ICDAR 2019 dataset, which contains a large collection of Latin manuscripts and it has proven important results comparing to other systems [9]. The dataset contains 3600 documents from 720 writers for the training set and around 20000 documents for 10000 writers for the testing set. Table 4 shows ICDAR 2019 submissions.

Table 4: The results with single feature extraction method [17]

	Method	**Accuracy [%]**	**mAP [%]**
Baseline	SRS-LBP (a) Classification	92.2	84.0
	SRS-LBP (b) Retrieval	93.1	86.8
SCUTT	SIFT	96.6	90.6
	Pathlet	96.0	89.8
Groningen	Hinge	88.4	75.6
	Co-Hinge	92.9	84.5
	QuadHinge	91.3	80.2
	Quill	88.3	76.0
	TCC	89.7	79.0
Tebessa	oBIFs	92.7	84.6
The proposed system	**SIFT**	**97.0**	**91.2**

Also, the system yielded the best results on another Latin dataset at the ICFHR 2020 competition [32] and clinched the first position. It was the only system to reach above 60% mAP on normal shape manuscripts (rectangular). The dataset contains more than 100k fragments for the training set and 20k for the testing set. Table 5 shows the overall results of ICFHR 2020 competition on rectangular and random shape (fragments) of manuscripts.

Table 5: ICFHR 2020 competition results [32]

	Method	**Accuracy [%]**	**mAP [%]**
Baseline	SRS-LBP	60.0	33.4
Groningen	FragNet	32.5	6.4
Belfort	TwoPathwriter	77.1	33.5
	TwoPathpage	61.1	25.2
Tebessa	oBIF	55.4	24.1
The proposed system	**ResNet20ssl**	**68.9**	**33.7**

The system has been also tested on the Arabic KHATT dataset (Non-historical documents) [26][25] and achieved great results. Table 6 shows the results for training, testing and validation sets. The KHATT dataset consists of 4000 small paragraphs were divided as follow: 2800 for training, 600 for testing and 600 for validation.

Table 6: KHATT dataset results

Partition	**Precision**	**Recall**	**F-Score**	**mAP**	**Accuracy**
Train	18.6 %	92.9 %	31.0 %	89.2 %	86.1 %
Test	19.5 %	97.3 %	32.4 %	95.5 %	94.0 %
Validate	19.3 %	96.3 %	32.1 %	92.8 %	90.3 %

And finally, the system has been tested on the first historical Arabic Manuscripts dataset and the results were as shown in Table 7. The dataset has a total 11610 images, they were divided 3110 for training and 8500 for testing. The training set contains random images from part of manuscripts. While, the testing set contains random images

from all the manuscripts where a large part of them not included in the training set. The images were used without any pre-processing techniques. The final result was a 95.2% mAP with 99.1% accuracy. The dataset will be published in order to test it with other systems and compare them with the results of this baseline system.

Table 7: Balamand Arabic Historical dataset results

Partition	Precision	Recall	F-Score	mAP	Accuracy
Train	95.8 %	61.9 %	75.2 %	95.4 %	99.0 %
Test	98.1 %	30.1 %	46.0 %	95.2 %	99.1 %

6. Conclusions

In this paper, the Balamand Arabic Historical Dataset is presented, which contains 11610 images from 567 unique manuscripts. The documents were collected from 14 different repositories between Lebanon between the 13th and 19th centuries. The state-of-the-art baseline system was tested on two Latin historical datasets ICDAR19 and ICFHR2020, and one Arabic non-historical KHATT dataset. The system showed high performance on all of them comparing to other systems. Then, the system was checked on the Balamand Arabic Historical dataset and yielded high results. Therefore, we can conclude that the ResNet20 has surpassed all other methods by far and proved high performance on Latin and Arabic datasets with different image shapes. For future work, we look forward to publish the dataset for researchers in order to compare other systems with our system results. Also, new end-end techniques will be applied in order to improve the obtained results.

7. Acknowledgements

This research is funded by the EIPHI Graduate School (contract "ANR-17-EURE-0002"). We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Quadro RTX 6000 GPU used for this research.

References

[1] (). *The Arabic Manuscripts in the Antiochian Orthodox Monasteries in Lebanon* volume 1–2. University of Balamand.

[2] (). P5: Guidelines for electronic text encoding and interchange. URL: `https://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-colophon.html`.

[3] Abdelhaleem, A., Droby, A., Asi, A., Kassis, M., Al Asam, R., & El-sanaa, J. (2017). Wahd: a database for writer identification of arabic historical documents. In *2017 1st International Workshop on Arabic Script Analysis and Recognition (ASAR)* (pp. 64–68). IEEE.

[4] Abdleazeem, S., & El-Sherif, E. (2008). Arabic handwritten digit recognition. *International Journal of Document Analysis and Recognition (IJDAR)*, *11*, 127–141.

[5] Asi, A., Abdalhaleem, A., Fecker, D., Märgner, V., & El-Sana, J. (2017). On writer identification for arabic historical manuscripts. *International Journal on Document Analysis and Recognition (IJDAR)*, *20*, 173–187.

[6] Awaida, S., & Mahmoud, S. (2011). Writer identification of arabic handwritten digits. In *First International Workshop on Frontiers in Arabic Handwritng Recognition, 2010*.

[7] Awaida, S. M., & Mahmoud, S. A. (2012). State of the art in off-line writer identification of handwritten text and survey of writer identification of arabic text. *Educational Research and Reviews*, *7*, 445.

[8] Bausi, A., Borbone, P. G., Briquel-Chatonnet, F., Buzi, P., Gippert, J., Macé, C., Melissakēs, Z., Parodi, L. E., Witakowski, W., & Sokolinski, E. (2015). *Comparative Oriental manuscript studies: An introduction*. COMSt.

[9] Chammas, M., Makhoul, A., & Demerjian, J. (2020). Writer identification for historical handwritten documents using a single feature extraction method. In *19th IEEE International Conference on Machine Learning and Applications (ICMLA 2020)*.

[10] Chandra, K., Kapoor, G., Kohli, R., & Gupta, A. (2016). Improving software quality using machine learning. In *2016 international conference on innovation and challenges in cyber security (ICICCS-INBUSH)* (pp. 115–118). IEEE.

[11] Chaurasia, P., Kohli, R., & Garg, A. (2014). *Biometrics minutiae detection and feature extraction*. LAP LAMBERT Academic Publishing.

[12] Chen, S., Wang, Y., Lin, C.-T., Ding, W., & Cao, Z. (2019). Semi-supervised feature learning for improving writer identification. *Information Sciences*, *482*, 156–170.

[13] Christlein, V., Bernecker, D., Honig, F., & Angelopoulou, E. (2014). Writer identification and verification using GMM supervectors. *IEEE Winter Conference on Applications of Computer Vision*, .

[14] Christlein, V., Bernecker, D., Hönig, F., Maier, A., & Angelopoulou, E. (2017). Writer identification using GMM supervectors and Exemplar-SVMs. *Pattern Recognition*, *63*, 258–267.

[15] Christlein, V., Gropp, M., Fiel, S., & Maier, A. (2017). Unsupervised feature learning for writer identification and writer retrieval. *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, .

[16] Christlein, V., & Maier, A. (2018). Encoding CNN activations for writer recognition. *2018 13th IAPR International Workshop on Document Analysis Systems (DAS)*, .

[17] Christlein, V., Nicolaou, A., Seuret, M., Stutzmann, D., & Maier, A. (2019).

ICDAR 2019 competition on image retrieval for historical handwritten documents. *arXiv [cs.CV]*, .

[18] Déroche, F. et al. (2005). Islamic codicology. *An Introduction to the Study of Manuscripts in Arabic Script*, .

[19] Djeddi, C., & Souici-Meslati, L. (2011). Artificial immune recognition system for arabic writer identification. In *International Symposium on Innovations in Information and Communications Technology* (pp. 159–165). IEEE.

[20] Fecker, D., Asi, A., Pantke, W., Märgner, V., El-Sana, J., & Fingscheidt, T. (2014). Document writer analysis with rejection for historical arabic manuscripts. In *2014 14th International Conference on Frontiers in Handwriting Recognition* (pp. 743–748). IEEE.

[21] Fecker, D., Asit, A., Märgner, V., El-Sana, J., & Fingscheidt, T. (2014). Writer identification for historical arabic documents. In *2014 22nd International Conference on Pattern Recognition* (pp. 3050–3055). IEEE.

[22] Fiel, S., & Sablatnig, R. (2015). Writer identification and retrieval using a convolutional neural network. *Computer Analysis of Images and Patterns*, (pp. 26–37).

[23] Hannad, Y., Siddiqi, I., Djeddi, C., & El-Kettani, M. E.-Y. (2019). Improving arabic writer identification using score-level fusion of textural descriptors. *IET Biometrics*, *8*, 221–229.

[24] Lai, S., Zhu, Y., & Jin, L. (2020). Encoding pathlet and sift features with bagged vlad for historical writer identification. *IEEE Transactions on Information Forensics and Security*, *15*, 3553–3566.

[25] Mahmoud, S. A., Ahmad, I., Al-Khatib, W. G., Alshayeb, M., Parvez, M. T., Märgner, V., & Fink, G. A. (2014). Khatt: An open arabic offline handwritten text database. *Pattern Recognition*, *47*, 1096–1112.

[26] Mahmoud, S. A., Ahmad, I., Alshayeb, M., Al-Khatib, W. G., Parvez, M. T., Fink, G. A., Märgner, V., & El Abed, H. (2012). Khatt: Arabic offline handwritten text database. In *2012 International Conference on Frontiers in Handwriting Recognition* (pp. 449–454). IEEE.

[27] Malisiewicz, T., Gupta, A., & Efros, A. A. (2011). Ensemble of exemplar-SVMs for object detection and beyond. *2011 International Conference on Computer Vision*, .

[28] Nguyen, H. T., Nguyen, C. T., Ino, T., Indurkhya, B., & Nakagawa, M. (2019). Text-independent writer identification using convolutional neural network. *Pattern Recognition Letters*, *121*, 104–112.

[29] Pechwitz, M., Maddouri, S., Märgner, V., Ellouze, N., & Amiri, H. (2002). Ifn/enit: database of handwritten arabic words.

[30] Rehman, A., Naz, S., & Razzak, M. I. (2019). Writer identification using machine learning approaches: a comprehensive review. *Multimedia Tools and Applications*, *78*, 10889–10931.

[31] D 'e roche, F. c c. o., & Rossi, V. S. (2012). *The manuscripts in Arabic characters*. Viella.

[32] Seuret, M., Nicolaou, A., Maier, A., Christlein, V., & Stutzmann, D. (2020). Icfhr 2020 competition on image retrieval for historical handwritten fragments. In *2020 17th International Conference on Frontiers in Handwriting Recognition (ICFHR)* (pp. 216–221). IEEE.

[33] Slimane, F., Awaida, S., Mezghani, A., Parvez, M. T., Kanoun, S., Mahmoud, S. A., & Märgner, V. (2014). Icfhr2014 competition on arabic writer identification using ahtid/mw and khatt databases. In *2014 14th International Conference on Frontiers in Handwriting Recognition* (pp. 797–802). IEEE.

Michel Chammas received his M.S. degree in Computer Science - Networking and Communication from the University of Balamand, Lebanon, in 2009. Since 2010, he has been a Researcher at the Digital Humanities Center at the University of Balamand. Michel is a Ph.D. student at the University of Franche-Comté, France. His research interests include Data Mining, Machine Learning, Internet of Things, Digital Humanities and CyberSecurity.

Abdallah Makhoul is a full professor is a full professor in Computer Science at University of Bourgogne - Franche-Comté (UBFC), France. He received the PhD degree in computer science from the University of Franche-Comté (UFC), France, in 2008. From 2009 to 2019, he has been an Associate Professor with the University of Franche-Comté. He is a member of the DISC department (Department of Computer Science and Complex Systems) of FEMTO-ST institute, France. He is the head the research team OMNI (Optimization, Mobility and Networking). His research focuses upon the following areas: distributed algorithms, Internet of things, programmable matter, e-health monitoring and real-time issues in wireless sensor networks. He has been a TPC chair and member of several networking conferences and workshops and guest editor and Reviewer for several international journals. He participated in several national and international research projects.

Jacques Demerjian received his PhD degree in Network and Computer Science from TELECOM ParisTech (Ecole Nationale Supérieure des Télécommunications - Paris) in 2004. Dr. Demerjian is a Professor at the Faculty of Sciences at Lebanese University. His main research interests include Body Sensor Network, Mobile Cloud Computing and Recommender Systems. He is an IEEE Senior Member.

Elie Dannaoui received the M.A. degree in Media Engineering for Education from the University of Poitiers, France, in 2006, and the Ph.D. degree in History from PIO, Rome, Italy, in 2012. Since 2018, he has been an Associate Professor with the University of Balamand. His research interests include Textual Encoding, Digital Cultural Heritage and Digital Philology.