Writer Identification for Historical Arabic Documents

Daniel Fecker*, Abedelkadir Asi†, Volker Märgner*, Jihad El-Sana† and Tim Fingscheidt*

*Institute for Communications Technology
Technische Universität Braunschweig
Braunschweig, Germany
Email: {*fecker, maergner, fingscheidt*}@*ifn.ing.tu-bs.de*

†Ben-Gurion University of the Negev
Be'er-Sheva, Israel
Email: {*abedas, el-sana*}@*cs.bgu.ac.il*

***Abstract*—Identification of writers of handwritten historical documents is an important and challenging task. In this paper we present several feature extraction and classification approaches for the identification of writers in historical Arabic manuscripts. The approaches are able to successfully identify writers of multi-page documents. The feature extraction methods rely on different principles, such as contour-, textural- and key point-based and the classification schemes are based on averaging and voting. For all experiments a dedicated data set based on a publicly available database is used. The experiments show promising results and the best performance was achieved using a novel feature extraction based on key point descriptors.**

I. Introduction

Writer identification is the task of assigning a writer with known reference documents to a query document written by an unknown writer. This is usually done by comparing features of the handwriting with a database of documents with already known writers. Another related task is the writer retrieval which finds from a query document with known writer all related documents in a database by ranking them according to their similarity.

In this work we are focusing on writer identification for historical document analysis. The aim is to support historians and librarians working with such documents, especially in context with large scale databases.

The literature already includes a lot of approaches for the extraction of discriminative features for writer identification. Often, a technique is dedicated to a certain language. In [1] the terms micro and macro features were introduced. Micro features represent fine characteristics, e. g., on character level like grapheme-based approaches [2]. In our work, we focus on macro features which capture writer characteristics globally like contour-based [2] and textural features [3]. Recently, local feature based approaches which rely on key point descriptors were proposed which showed promising results [4], [5]. A detailed overview about this topic is given in [6].

This paper presents different feature extraction algorithms including novel methods and refinements of existing techniques and comparing their performances. Several classification schemes are proposed to compare multi-page documents, either on manuscript level or on page level. In addition, we present a data set for the writer identification task in Arabic historical documents. Manuscripts were collected from the Islamic Heritage project (IHP) [7]. To the best of our knowledge, it is possibly the first data set which is designated for the writer identification task in Arabic historical documents. Figure 1 shows example pages from the data set.

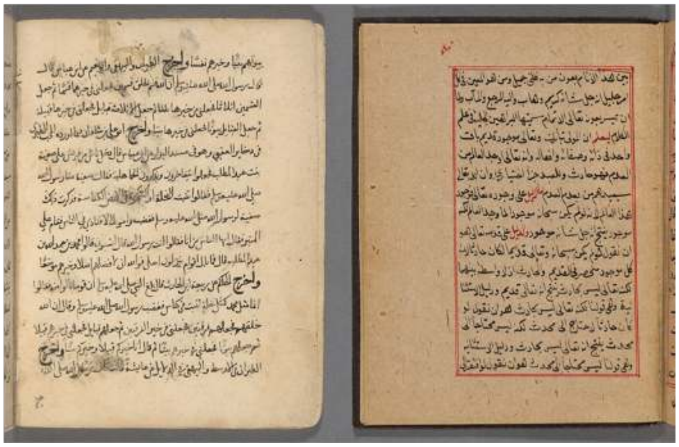


Fig. 1. Example pages of historical Arabic manuscripts.

The paper is organized as follows. In the next section the preprocessing of the historical manuscript pages used in our experiments is presented. In Section III the feature extraction algorithms are described, followed by explanation of the proposed classification schemes in Section IV. Our data set, the experimental setup, and the results of our experiments are presented in Section V. At the end, conclusions are given.

II. Preprocessing

This stage aims to minimize the inherent noise and obtain a clean binarized version of the manuscript pages. Toward this goal we define a pipeline of preprocessing steps. Separating

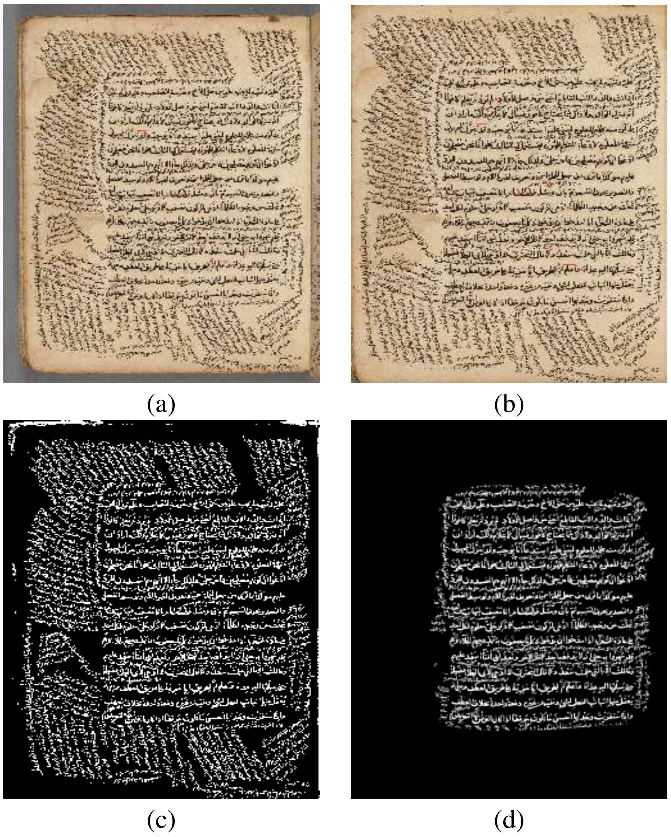


Fig. 2. Phases of the preprocessing pipeline. (a) Original image, (b) cropping scanning board, (c) binarization and (d) main text frame detection.

manuscript pages from the scanning board, binarization and detecting text in the main frame of the page are the main components of this pipeline (see Figure 2).

The difference between the colors of the scanning board and the manuscripts pages enables dividing the pixels into two clusters representing different colors. Since page color might slightly differ due to the inherent noise on historical images, we represent color in the CIELAB color space. This color space is widely considered as perceptually uniform for small color distances [8].

Detecting the main frame of the manuscripts pages relies on the fact that it contains horizontally oriented text. Horizontal text defines a unique texture as it usually has a uniform orientation and text line spacing. Due to this observation, we apply a Gabor filter as it had been found to be particularly appropriate to distinguish between texture representations [9]. The Gabor filter kernel is adapted per manuscript to adequately capture the properties of its writing style. The adaptation is done by determining the wavelength of the cosine factor of the Gabor kernel which we found to be strongly correlated with text line spacing. Hysteresis thresholding is used to binarize the response of the filter. At this stage of the pipeline we have a binary mask of the main text frame.

Masking the binarized images by the output of the previous step results on determining the main text frame of each manuscript page.

III. Feature Extraction

In this section several feature extraction methods are presented. While some are state-of-the-art techniques, other schemes are based on new modifications and improvements. The feature extraction methods rely on different basic principles, i. e., contour-, texture- and key point-based approaches.

A. Contour-Based Features

A first feature we use is the distribution of the directions of the writing contours (CON) [2]. The contour is obtained by tracing the exterior boundary of the writing in a binary image. This is also including the boundary of holes on the inside of the characters. The direction of the contour is specified by the angle

$$\phi = \arctan\left(\frac{u_{k+\delta} - u_k}{v_{k+\delta} - v_k}\right) \quad (1)$$

between two contour pixels with distance δ and the x-axis, with (u_k, v_k) denoting a position on the contour. The distance δ between two pixel positions (u_k, v_k) and $(u_{k+\delta}, v_{k+\delta})$ is defined as the step size on the contour.

A statistic of the contour directions is created by sorting all computed angles in an angle histogram with n bins spanning the interval $[0°, 180°]$. This histogram is normalized to a probability distribution $P(\phi)$ which is used as the feature vector.

In [2] a fixed δ was used. We propose a modified method which enables variable values for δ, allowing a more accurate description of the contour angles similar to [10]. In a first step, as initialization δ_0 is used and a reference angle ϕ_{δ_0} is computed. Afterwards, the distance is increased incrementally $\delta_{i+1} = \delta_i + 1$ and each time the angle $\phi_{\delta_{i+1}}$ is recomputed and compared to the reference angle. If

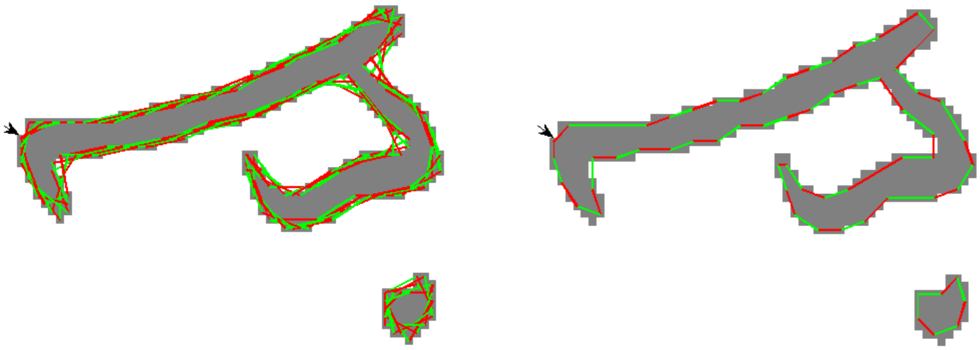
$$\phi_{\delta_{i+1}} \notin [\phi_{\delta_0} - \epsilon, \phi_{\delta_0} + \epsilon] \quad (2)$$

with ϵ being a predefined constant, we use ϕ_{δ_i} for building the histogram underlying $P(\phi)$.

Unlike pixel-wise computing the angle for each contour pixel like in [2], our algorithm uses the end point of the last computed angle as the new starting point. Fig. 3 shows an example using the approach of Bulacu et al. in Fig. 3(a) and our approach in Fig. 3(b). The starting point of the angle computation is in both cases the upper left corner, marked by the arrows, and the contour is traced clockwise. The pixel-wise computation of angles between a fixed amount of pixel in Fig. 3(a) introduces a lot of noise due to the wrong angles computed especially in the corners of the writing.

B. Histogram of Oriented Gradients

The histogram of oriented gradients (HOGs) are a popular descriptor that has been widely used for human detection [11]. The gradient histograms, which are computed over rectangular cells (R-HOG), are expected to capture the distribution of local edge directions. In this work we exploit the power and simplicity of HOGs as dense image descriptors to acquire a global



(a) (b)

Fig. 3. Example of the contour-based features. The pixel used for angle measurement on the contour are connected with lines. The alternating color (red and green) is just for demonstration purpose. (a) Original approach of Bulacu et al. [2], (b) our modified approach.

representation of an individual writing style. Concatenating the histograms over all the rectangular cells guarantees to capture a global signature of a given writer. The HOG parameters were adopted after a set of experiments that identified six cells and a nine bin histogram per cell as the appropriate values in terms of identification accuracy.

C. Oriented Basic Image Features

The oriented basic image (OBI) features are based on a multi-scale feature extraction which uses local symmetry and orientation [12]. This method originates from texture recognition [13] but was already used for natural image character recognition [12] and also won the Arabic writer identification challenge at the ICFHR 2012 [14]. Gaussian derivative filters with different orders and directions are used to form a filter-response space. From the filter responses seven features are constructed which approximate the local symmetry. Each image pixel is assigned to one of the seven features according to the largest expression. While orientations can be assigned to four of the features, the other three are rotationally invariant. Based on the symmetry features, orientations and scales a histogram is created which, after a normalization, results in the final feature vector.

D. Scale-Invariant Feature Transform

Scale-invariant feature transform (SIFT) descriptors are local features which are based on finding key points in an image [15]. In handwriting the key points lie on crossings, loops and peaks of the characters and describe the slant and the curvature [5]. To apply these descriptors to the task of writer identification, two approaches were already proposed. The first uses k-means clustering to apply a bag of words scheme [4] and the second uses supervised learning to form visual vocabularies [5]. Here, we propose a much simpler method which neither needs clustering nor supervised learning to generate a discriminative feature vector for the writer identification problem.

The SIFT algorithm extracts key points by using a scale-space of filter responses of differences of Gaussians. They are chosen by selecting local extrema through analyzing the neighborhood in a scale level and the surrounding scale levels. For each key point an orientation is assigned by analyzing the gradients in its local region. A key point descriptor vector $\mathbf{v}$ of $E = 128$ elements is generated by computing eight bins orientation histograms from a 4×4 region around the key point.

Let O be the amount of descriptor vectors which are extracted for an entity containing handwritten text. For each descriptor vector element $e \in \{1, ..., E\}$ a transformed vector $\tilde{v}_e$ is computed which contains all e-th elements of all O descriptor vectors

$$\tilde{\mathbf{v}}_e = \{\mathbf{v}_1(e), ..., \mathbf{v}_O(e)\}\,. \tag{3}$$

In a further step the cosine distance between all the vectors $\tilde{v}_e$ is computed and used finally as feature vector with $\sum_{e=1}^{E}(E - e) = 8128$ elements.

Those distances describe the relation of the local orientation and magnitude of the gradients in the regions around all the key points and thus capturing global information about the writing style of a writer.

IV. Classification

The classification task is to identify the writer of an unknown query manuscript Q by comparing it to an existing (training) data set S with N manuscripts of M already known writers and finding the manuscript which handwriting reveals the highest similarity to the query manuscript. The data set $\mathcal{S}$ is composed of all the feature vectors extracted for all pages of the data set

$$\mathcal{S} = \left\{ \mathbf{x}_1^{(1)}, ..., \mathbf{x}_{L_1}^{(1)}, ..., \mathbf{x}_1^{(N)}, ..., \mathbf{x}_{L_N}^{(N)} \right\} \tag{4}$$

with $L_i, i = 1, ..., N$ denoting the number of pages of the manuscript i. The absolute number of pages in $\mathcal{S}$ is

$$L = L_1 + L_2 + ... + L_N. \tag{5}$$

We present three different classification schemes: averaging, voting and weighted voting. These are working on different levels, i. e., averaging is working on manuscript level and the voting approaches are working on page level. Independent of the level, we employ the nearest-neighbor classifier which can be used with an arbitrary distance metric $D(\mathbf{a}, \mathbf{b})$, measuring the distance between the vectors $\mathbf{a}$ and $\mathbf{b}$. The idea behind the proposed classification schemes can also work with other classifiers, e. g., utilizing probabilities rather than distances.

A. Averaging

For each manuscript i in $\mathcal{S}$ an average feature vector

$$\bar{\mathbf{x}}^{(i)} = \frac{1}{L_i} \sum_{j=1}^{L_i} \mathbf{x}_j^{(i)} \tag{6}$$

is computed, resulting in an inferred data set

$$\hat{\mathcal{S}} = \left\{ \bar{\mathbf{x}}^{(1)}, ..., \bar{\mathbf{x}}^{(N)} \right\} \tag{7}$$

containing all the averaged feature vectors.

Similar to (6), the average vector of the query manuscript $\bar{\mathbf{x}}^{(Q)}$ is computed. The manuscript i^* with the smallest distance is identified by employing the nearest-neighbor classifier with an appropriate distance metric:

$$i^* = \arg \min_{\text{i}} D(\bar{\mathbf{x}}^{(Q)}, \bar{\mathbf{x}}^{(i)}) \tag{8}$$

where $\bar{\mathbf{x}}^{(i)} \in \hat{\mathcal{S}}$ and $i = 1, ..., N$. As a result, the writer of the manuscript $i = i^*$ is assigned to the query manuscript.

B. Voting

For the voting scheme the distances are computed on page level. For all the pages $k = 1, ..., L_Q$, with L_Q denoting the number of pages in the query manuscript, the page in $\mathcal{S}$ with the minimal distance of $\mathbf{x}_k^{(Q)}$ to all $\mathbf{x}_\kappa \in \mathcal{S}$, $\kappa = 1, ..., L$ is determined by

$$\kappa_k^* = \arg \min_{\kappa} D(\mathbf{x}_k^{(Q)}, \mathbf{x}_\kappa). \tag{9}$$

The corresponding manuscript i^* of the page κ_k^* is further used to find the writer of the query manuscript by analyzing the distribution of the assigned manuscripts in

$$\mathcal{I} = \{i_k^*\}, k = 1, ..., L_Q. \tag{10}$$

For this purpose, we generate a histogram $h_b(\mathcal{I})$, $b = 1, ..., B$ with the number of bins B equals either the number of manuscripts or the number of writers. For the latter, the information which manuscript belongs to which writer is additionally used to generate the histogram. We select the element which has the highest amount of occurrences in the histogram

$$i^* = \arg \max_{\text{b}} h_b(\mathcal{I}). \tag{11}$$

Note that i^* can either correspond to a writer or a manuscript. For the latter, the writer of the manuscript i^* is assigned to the query manuscript similar to subsection IV-A.

C. Weighted Voting

This approach is related to a soft decision classification approach. Here, we also use (9) to find the corresponding manuscript number to a query manuscript page, but additionally, we are computing the values of the minimal distances

$$d_k^* = \min_{\kappa} \left(D(\mathbf{x}_k^{(Q)}, \mathbf{x}_\kappa) \right) \tag{12}$$

for all $k = 1, ..., L_Q$ to determine weights. We then obtain weights by calculating

$$w_k = 1 - \frac{d_k^* - \min_{\kappa}(d_\kappa^*)}{\max_{\kappa}(d_\kappa^*)}. \tag{13}$$

The page with the smallest distance has the highest weight which equals one, all others have weights below one.

The further steps are similar to the voting scheme in section IV-B, except that the weights are used for generating the histogram $h_b(\mathcal{I})$.

V. Experiments

In this section the experimental setup and the results are described. The results are based on the data set described in section V-A by employing different feature extraction methods and classification schemes.

A. Data Set

Challenging data sets provide an empirical basis for research. To the best of our knowledge, this is the first data set for writer identification and verification in the context of Arabic historical documents. The data set consists of 60 different manuscripts that were written between the 13^{th} century and the early 20^{th} century. The manuscripts are available on-line[1] as part of the Islamic Heritage project (IHP) [7]. They are written mainly in Arabic script, except two, which were written in Ottoman script (still Arabic alphabet). While the data set has 24 known writers, one can find 6 generally unknown writers, too. From 11 of the writers multiple manuscripts are included in the data set ranging from 2 to 10 manuscripts per writer (in total 42 manuscripts). For the other 18 writers only one manuscript per writer is included in the data set.

The considered data set has in total $L = 4595$ manuscript pages. These pages contain different levels of noise which are common on historical documents. They contain ornamentations such as a red rectangle surrounding the main body text (see Figure 1). In addition, text had been added to the margins of some pages by unknown writers throughout the years. Another problem is that in some manuscripts the writer changes throughout the manuscript. Usually, in the end of each manuscript there are some full pages that contain text added by anonymous writers, e. g., students or owners.

[1] http://www.cs.bgu.ac.il/~abedas

TABLE I. RESULTS OF THE WRITER IDENTIFICATION IN ACCURACY

Features	Accuracy of	Accuracy of Manuscripts		
	Pages	Averaging	Voting	W-Voting
CON 1	0.482	0.619	0.691	0.714
CON 2	0.686	0.810	0.881	0.929
HOG	0.469	0.643	0.738	0.738
OBI	0.876	**0.929**	0.929	0.929
K-SIFT	**0.925**	0.595	**0.976**	**1.000**

TABLE II. PERFORMANCE ANALYSIS OF FEATURE EXTRACTION FOR ONE REFERENCE PAGE

Features	Vector size	Processing time
CON 1	12	64.1 s
CON 2	12	66.0 s
HOG	54	0.9 s
OBI	3969	10.5 s
K-SIFT	8128	16.2 s

B. Experimental Setup

Due to our limited data set size, we are using leave-one-out cross validation in our experiments. We employ this approach on manuscript level, i. e., one of the manuscripts is regarded as a query manuscript which writer needs to be determined, while for all other manuscripts the writer is known. We focus here on the ability of the presented features and classification schemes to correctly classify an unknown writer to a known writer without a reject decision. So we are using the 43 manuscripts from the 11 writers with more than one manuscript per writer as test data. For the training, however, we also use the other 17 writers with only one manuscript to have a broader variability.

As a distance metric for the nearest-neighbor classification we use the $\mathcal{X}^2$-distance which is commonly used in this field and defined as

$$D_{\mathcal{X}^2}(\mathbf{a},\mathbf{b}) = \sum_{\Delta=1}^{d} \frac{(a_\Delta - b_\Delta)^2}{a_\Delta + b_\Delta} \tag{14}$$

between two feature vectors $\mathbf{a}$ and $\mathbf{b}$ both of length d.

As feature extraction methods we use in our experiments the original contour-based approach from [2] (dubbed CON 1), our modified version from section III-A (CON 2), histogram of gradients-based approach (HOG), oriented basic image (OBI) features, and our key point-based approach (K-SIFT) from Section III-D.

We have observed that the already proposed key point-based approaches from [4], [5] are not applicable for our data set because large amounts of memory are needed to store all the descriptors ($>$ 50 GB, for around 10000 descriptor vectors per page). Furthermore, computing the clustering or modeling algorithms for such a huge amount of data would also be too computationally expensive.

C. Results

All the results we obtained are listed in Table I. Besides the accuracy for the correct classification of single pages to a writer, the accuracy for whole manuscripts based on the three classification schemes averaging, voting and weighted voting (denoted as w-voting), which are presented in Section IV, are depicted. The observations are that our key point-based approach (K-SIFT) achieved the best results for nearly all classification schemes. For the weighted voting approach a perfect classification of the writers from the manuscripts was achieved. In contrast the averaging classification yielded the worst results, which, we think, is due to the large size of the feature vector. Here, the OBI-based feature extraction achieved the best results. Furthermore, our modifications of the contour-based approaches (CON 2) achieved better results than the original approach (CON 1).

In general, however, we observed that the processing time for the feature extraction of the contour-based approaches was the largest, which can be problematic for large scale databases. Table II shows the performance of the different feature extraction methods for a reference page in terms of processing time of the feature extraction, running on an Intel Core i7 920@2.67 GHz with 18 GB of RAM. The HOG features had by far the shortest processing time but also achieved the worst results for now.

VI. CONCLUSIONS

In this paper we presented experimental results for the identification of writers in historical Arabic manuscripts. We have tested several feature extraction approaches which rely on contour-, textural- and key point-based principles and implemented several classification schemes for the identification of writers in multi-page documents. For our experiments, we assembled a dedicated data set based on the publicly available IHP collection. The experiments are showing promising results and the best performance was achieved using a novel feature extraction method based on key point descriptors yielding a perfect classification in combination with our weighted voting scheme.

Given that the presented features are text independent, they can also be applied to other languages. First tests with Latin historical manuscripts are looking promising. Future work will also include the introduction of a reject possibility in our classification framework and also experiments for writer retrieval. Furthermore, we are planning to enlarge the data set to obtain a more demanding problem which could also imply the use of additional techniques such as feature combination.

ACKNOWLEDGMENTS

The German Research Foundation (project: FI 1494-3-2), the Ministry of Science and Technology of Israel, and the Council of Higher Education of Israel is gratefully acknowledged for financial support.

Furthermore, we want to thank Waldemar Kuznik and Julia Kaehm who implemented parts of the feature extraction algorithms in their bachelor thesis.

REFERENCES

[1] S. Srihari, M. J. Beal, K. Bandi, V. Shah, and P. Krishnamurthy, "A Statistical Model for Writer Verification," in *Proc. of Int. Conf. of Doc. Anal. and Rec. (ICDAR)*, Seoul, Korea, August 2005, pp. 1105–1109.

[2] M. Bulacu, L. Schomaker, and A. Brink, "Text-Independent Writer Identification and Verification on Offline Arabic Handwriting." in *Proc. of Int. Conf. of Doc. Anal. and Rec. (ICDAR)*, Curitiba, Brazil, 2007, pp. 769–773.

[3] A. Al-Dmour and R. A. Zitar, “Arabic Writer Identification Based on Hybrid Spectral-Statistical Measures.” *J. of Exp. Theor. and Artif. Intell.*, vol. 19, no. 4, pp. 307–332, 2007.

[4] S. Fiel and R. Sablatnig, “Writer Retrieval and Writer Identification Using Local Features.” in *Proc. of Doc. Anal. Sys. (DAS)*, Queensland, Australia, March 2012, pp. 145–149.

[5] S. Fiehl and R. Sablatnig, “Writer Identification and Writer Retrieval Using the Fisher Vector on Visual Vocabularies,” in *Proc. of Int. Conf. of Doc. Anal. and Rec. (ICDAR)*, Washington, DC, USA, August 2013, pp. 545–549.

[6] Sreeraj and S. M. Idicula, “A Survey on Writer Identification Schemes,” *International Journal of Computer Applications*, vol. 26, no. 2, pp. 23–33, July 2011.

[7] Harvard University. Islamic Heritage Project (IHP). [Online]. Available: http://ocp.hul.harvard.edu/ihp/

[8] R. Cohen, A. Asi, K. Kedem, J. El-Sana, and I. Dinstein, “Robust Text and Drawing Segmentation Algorithm for Historical Documents,” in *Proc. of the Int. Workshop on Hist. Doc. Imag. and Process.* Washington, DC, USA: ACM, August 2013, pp. 110–117.

[9] I. Fogel and D. Sagi, “Gabor Filters as Texture Discriminator,” *Biological Cybernetics*, vol. 61, no. 2, pp. 103–113, June 1989.

[10] J. Sklansky and V. Gonzalez, “Fast Polygonal Approximation of Digitized Curves,” *Pattern Recognition*, vol. 12, no. 5, pp. 327 – 331, 1980. [Online]. Available: http://www.sciencedirect.com/science/article/pii/003132038090031X

[11] N. Dalal and B. Triggs, “Histograms of Oriented Gradients for Human Detection,” in *Proc. of Int. Conf. on Com. Vis. and Pat. Rec. (CVPR)*. San Diego, CA, USA: IEEE Computer Society, June 2005, pp. 886–893.

[12] A. J. Newell and L. D. Griffin, “Natural Image Character Recognition Using Oriented Basic Image Features.” in *Proc. of Int. Conf. Dig. Im. Comp.: Tech. and App. (DICTA)*, Queensland, Australia, December 2011, pp. 191–196.

[13] M. Crosier and L. D. Griffin, “Texture Classification with a Dictionary of Basic Image Features.” in *Proc. of Int. Conf. on Com. Vis. and Pat. Rec. (CVPR)*, Anchorage, AK, USA, June 2008, pp. 1–7.

[14] A. Hassane and S. Al-Madeed, “ICFHR 2012 Competition on Writer Identification Challenge 2: Arabic Scripts.” in *Proc. of. Int. Conf. in Front. of Handwr. Rec. (ICFHR)*, Bari, Italy, September 2012, pp. 835–840.

[15] D. G. Lowe, “Distinctive Image Features from Scale-Invariant Keypoints,” *Int. Journal of Computer Vision*, vol. 60, pp. 91–110, 2004.