Segmentation-free Word Spotting for Handwritten Arabic Documents

G. Khaissidi[1], Y. Elfakir[1], M. Mrabti[1], Z. Lakhliai[1], D. Chenouni[1], M. El yacoubi[2]

[1] LIPI, ENS, FES, Maroc
[2] SAMOVAR, Télécom SudParis, CNRS, Université Paris-Saclay, France

Abstract **— In this paper we present an unsupervised segmentation-free method for spotting and searching query, especially, for images documents in handwritten Arabic, for this, Histograms of Oriented Gradients (HOGs) are used as the feature vectors to represent the query and documents image. Then, we compress the descriptors with the product quantization method. Finally, a better representation of the query is obtained by using the Support Vector Machines (SVM).**

Keywords **— Word Spotting, Arabic Handwritten Documents, Histograms Of Oriented Gradients (HOG), Image Recognition, Classification.**

I. Introduction

The search for information in Arabic manuscripts is not a simple process, for this, the conception of recognition system knows today a great expansion and seems as a necessity so as to exploit the wealth of information contained in ancient manuscripts. The manual manipulation repetitive of fragile documents could destroy them, for this, many digitization projects have been developed such as DEBORA (Digital access to Books Of Renaissance) [1], EAMMS (Electronic Access to Medieval Manuscripts) [2], Better Access to Manuscripts and Browsing of Images (BAMBI) [3] and Manuscript Access Through Standards for Electronic Records(MASTER)[4]…treat Latin scripts. For this reason above, and in order to develop a complete system for recognition Arabic handwriting, the first step for the creation of this system is presented in this article. In the survey of literature, it is found that many researchers of keyword spotting methods are inspired by one of two following categories:

- Learning-based methods use supervised machine learning techniques to train models of the words that the user wants to spot.
- Example-based methods, receive as input an instance of the word that the user wants to retrieve.

Most of the authors prefer a learning-based approaches for applications where the keywords to spot are a priori known and fixed. In same handwritten Arabic documents, the segmentation step is not usually easy “Fig.1”, any segmentation errors affect the subsequent word representations and matching steps, this dependence motivated the researchers to move towards complete segmentation-free methods.

From literature survey of handwritten word spotting techniques, we found that some researchers has done work on the handwritten Arabic documents where a million documents are written in various disciplines between the seventh and fourteenth centuries, whereas many works treat a Latin’s manuscripts documents. For a given query image, Y. Leydier and al. [5] extract and encode interest points by a simple descriptor based on gradient information. The word spotting is then performed by trying to locate zones of the document images with similar interest points. Only the ones sharing the same spatial configuration than the query model are returned.

Kamble and Hegadi [6] extract the features of handwritten Marathi

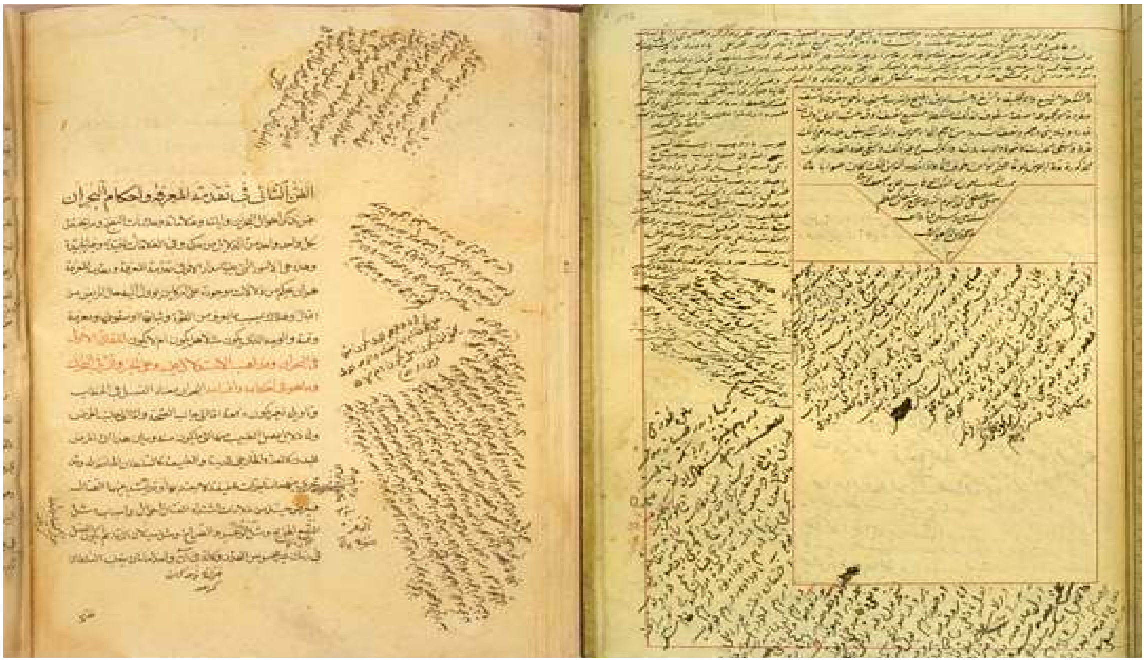


Fig. 1. Process of the proposed system

DOI: 10.9781/ijimai.2016.411

characters using Rectangle Histogram Oriented Gradient, in this work; Feed-Forward Artificial Neural Network is used for classification step. For a query image, Rath et al. [7] extract discrete feature vectors that describe query, which are then used to estimate similarity after training step of the probabilistic classifier. Jon et al. [8] represent the document with a grid of HOG descriptors and use a sliding-window approach for word spotting in document images. A similar approach is used in [9] which the retrieval step is performed by using an exemplar support vector machine framework.

The approach used in [10] combines three slanted windows, vertical, left and right. They proposed this method to remedy with the problem of writing inclination, overlapping and diacritical marks. HMM-based classifier is used at the decision step. Kessentini et al. propose to use Multi-stream hidden Markov Models for off-line handwritten Arabic word recognition [11]. The approach used combines density based features and contour based features of sliding windows. In [12], each character is a state in Variable Duration Hidden Markov Models and has a variable duration to model a character model of multiple segments.

The remainder of this paper is organized as follows: Section 2 present the indexation system process. Section 3 describes HOG feature extraction approach for word-spotting and product quantization method. Afterwards, in section 4, classification using SVM classifier is presented. Section 5 discuss experimental results. Finally, conclusions are summarized in section 6.

II. Present Work

In the present work, the document images have been preprocessed to enhance them after scanning the collected datasheets. For this purpose, a model for the restoration of the degradations [13], which uses a series of multi-level classifiers [14] applied to document images. Then, the window slid on the image document Histograms of Oriented Gradients (HOG) features are then extracted from each word image to represent and to compare the query with the region of the document. The application of Product Quantization [15] to encode the HOG descriptors reduce the size of the descriptors provides better performance in time of descriptor computation. Finally, Support Vector Machines is used to produce a better representation of the query and to classify feature vectors. Identical positive set is produced by slightly the window around the query and sample negative set is obtained by taking a sampler random regions. The regions with high similarity will be used in reranking step. The general process of the proposed system is shown in “Fig.2”.

The proposed indexation system is achieved in the following steps:

- Image preprocessing
- Feature Extraction
- Product Quantization
- SVM training set
- SVM classifying set
- Reranking

III. HOG Feature Extraction

Histogram of oriented gradients feature descriptors are used in computer vision and image processing for the object recognition purpose. The main idea behind the HOG descriptors is that local object appearance and shape within an image can be described by the distribution of intensity gradients or edge directions. The implementation of these descriptors can be achieved by dividing the image into small connected regions, called cells, and for each cell computing a histogram of gradient directions; histograms are also normalized based on their energy (regularized L2 norm). The combination of these histograms represents the descriptor.

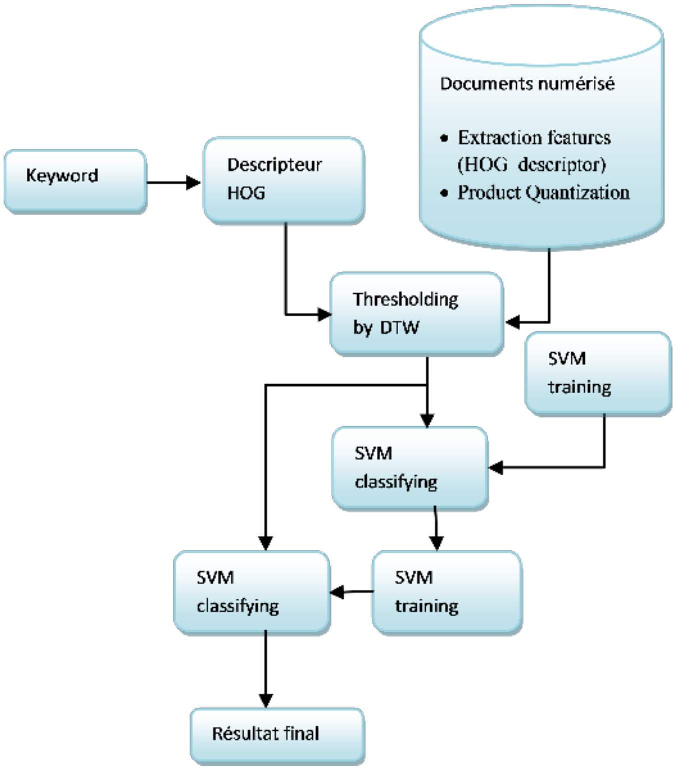


Fig. 2. Process of the proposed system.

A. Feature extraction

The application of HOG features descriptors and Product Quantization for word spotting in handwritten Arabic documents is the main contribution of this paper. The feature extraction is a most important part of word spotting system, that mean transforming the input query into the set of features. Histogram Oriented Gradient is used to detect and extract feature of Arabic handwritten documents “Fig. 3”. Initially, we remove noise for sliding-window and region of documents with Gaussian filter. After smoothing, the Sobel kernel is used to calculate the horizontal and vertical components of the gradients. Let, is the smoothed image and the horizontal and vertical

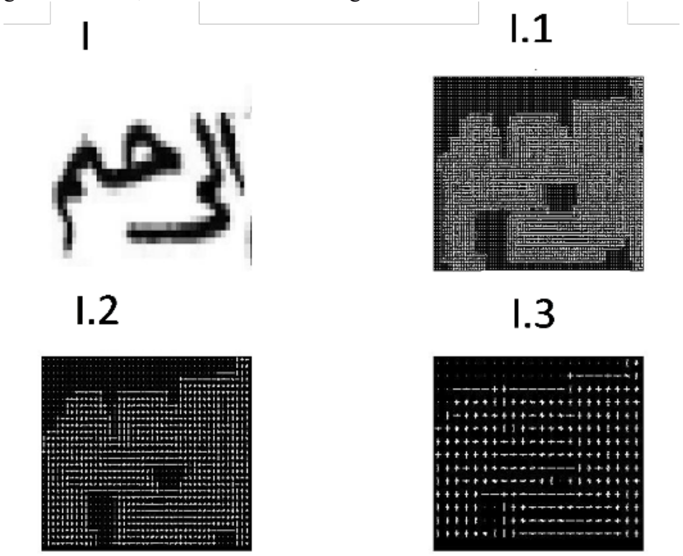


Fig. 3. Rectangle HOG of 2*2, 4*4 and 8*8 block size of Ibn Sina (I1, I2, I3) dataset

components of image gradient is Ix(x,y) and Iy(x,y) respectively.

$$I_x(x,y) = I_s * \begin{bmatrix} -1 & 0 & +1 \\ -2 & 0 & +2 \\ -1 & 0 & +1 \end{bmatrix} \text{ and } \quad I_y(x,y) = I_s * \begin{bmatrix} +1 & +2 & +1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix}$$

The magnitude M(x, y) and direction D(x, y) of the gradient at pixel (x, y) in the smoothed image are computed as follows:

$$M(x,y) = \sqrt{I_x^2(x,y) + I_y^2(x,y)}$$

$$D(x,y) = \tan^{-1}\left(\frac{I_x(x,y)}{I_y(x,y)}\right)$$

Then, histogram of all blocks can be computed using the block size of character; each pixel is assigned in certain category according to its gradient direction, the feature extraction process is achieved in the following steps:

1. Query
2. Gradients in X and Y direction
3. Dividing query into Blocks
4. Feature extraction
5. Normalize the values
6. Feature concatenation form all blocks
7. Descriptor

B. Product Quantization

The main drawback of sliding-window based methods is the cost of re-computing the descriptors of every image with every new query. However, it's possible to be pre-computing and storing the HOG descriptors, that necessary a large amount of memory to keep them. For this, we propose to encode the HOG descriptors by means of Product Quantization (PQ) proposed by Jégou et al. in [15]. This method both to reduce the amount of memory needed to store the descriptors and reduce the computational cost of searching of the descriptor. This technique has shown excellent results on approximate nearest neighbor tasks. The idea is to decompose the space into a Cartesian product of low-dimensional subspaces and to quantize each subspace separately.

IV. Classification

The main objective of query recognition system is to achieve robust performance to identification of query. In our case we have used SVM classifier with linear function for the recognition. The Width of the margin between the classes is the major optimization criterion, the empty area around the decision boundary, defined by the distance to the nearest training pattern. These patterns called support vectors, which finally define the function for classification.

The kernel linear function used in SVM classification is:

$$K(x_i, x_j) = x_i^T \cdot x_j$$

In a linear model, separating hyper-plane has equation

$$w^T \cdot x + b = 0$$

Considering a binary classification problem with training data:

$$\{(x_1, y_1)........(x_i, y_i)\} \text{ where } x_i \in (\mathrm{N}, \mathrm{P}) \text{ and } y_i \in \{+1, -1\}$$

The SVM attempts to find the hyper-plane < w, b > that maximizes the margin.

The positive and negative support vectors respectively is :

$$w^T \cdot x_+ + b = +1 \text{ and } w^T \cdot x_- + b = -1 \text{ so}$$

$$\frac{w}{\|w\|} \cdot (x_+ - x_-) = \frac{w^T(x_+ - x_-)}{\|w\|} = \frac{2}{\|w\|}$$

So we can deduce that maximize the margin amounts to minimizing. This can be casted as an optimization problem as:

$$\arg_w \min \frac{1}{2}\|w\|^2 + c_1 \sum_{(x_+, y_+) \in \mathrm{P}} L(y_+ w^T x_+) + c_2 \sum_{(x_-, y_-) \in \mathrm{N}} L(y_- w^T x_-) \qquad c_{1,2}$$

Is a regularization parameter, $y_+ = +1$ and $y_- = -1$

$x_+ \in \mathrm{P}$ is constructed by deforming the query, To produce the negative set , the sample random regions over all the documents $x_- \in \mathrm{N}$.

V. Experiments

In this section, we present the result of the approach proposed for searching query on Ibn Sina handwritten manuscripts datasets. MATLAB is used to measure all score and running times of the different sections (computing the HOG descriptors, calculating the scores with query and training the SVM). "Table 1" shows the mean average precision of the approach proposed and the time of descriptor computation per image document.

As we see in "Fig. 5" and "Fig. 6", the precision and recall mean recall results depending on the query's size.

$$\text{Precision} = \frac{|(\text{relevant document}) \cap (\text{retrieved document})|}{|(\text{retrieved document})|}$$

$$\text{Recall} = \frac{|(\text{relevant document}) \cap (\text{retrieved document})|}{|(\text{relevant document})|}$$

$$\text{mAP} = \frac{\sum_{q=1}^{Q} P(q)}{Q}$$

TABLE I
TIME OF DESCRIPTOR COMPUTATION AND MEAN AVERAGE PRECISION

Dataset	Time computation	Mean average precision
Ibn Sina	125s	68.4%

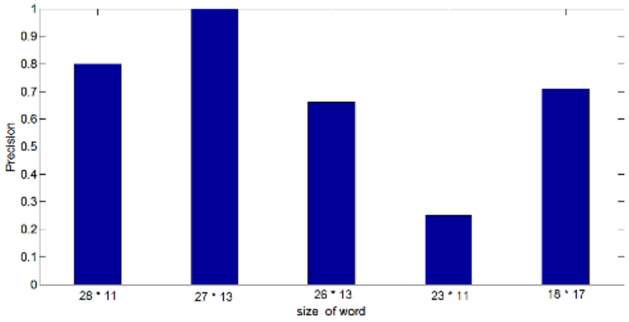


Fig. 4. Precision for different size of word .

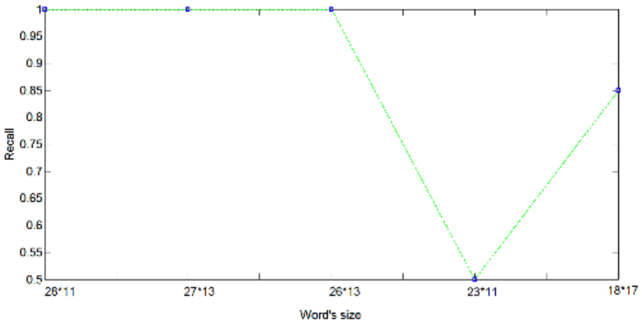


Fig. 5. Recall for different size of word.

VI. Conclusion

In this paper we have presented segmentation-free word spotting method for Handwritten Arabic documents using HOG descriptor. This method has been evaluated on a large amount of handwritten Ibn Sina. The experimental results "Table 1" and "Fig. 5" show that the used of HOG based feature extraction method and SVM classifier with linear function provides good results in mAP and the time of descriptor computation. In future, this work can extend to enhance the performance and reducing the time of descriptor by adding some more relevant features.

References

[1] DEBORA: projet européen n°. LB 5608 A. Coordinateur R. Bouché, juin 2000.179 pages

[2] http://www.hmml.org/eamms/index.html.

[3] CALABRETTO, Sylvie; BOZZI, Andrea; PINON, Jean-Marie, décembre 1999. "Numérisation des manuscrits médiévaux", le projet européen BAMBI, in: Actes du colloque Vers une nouvelle érudition: numérisation et recherche en histoire du livre, Rencontres Jacques Cartier, Lyon.

[4] BURNARD, Lou. ; ROBINSON, PETER. "Vers un standars européen de description des manuscrits", le projet Master. Document numérique. 1999, vol 3, no 1-2, p.151-169.

[5] Y. Leydier, A. Ouji, F. Lebourgeois, H. Emptoz, 2009. "Towards an omnilingual word retrieval system for ancient manuscripts", Pattern Recognit. 42 (2009) 2089–2105.

[6] M.Kamble, S.Hegadi, 2015. "Handwritten Marathi character recognition using R-HOG Feature", in: International Conference on Advanced Computing Technologies and Applications (ICACTA), Procedia Computer Science 45 (2015) 266 – 274015.

[7] T. Rath, V. Lavrenko, and R. Manmatha, 2003. "Retrieving historical manuscripts using shape", Technical Report, Center for Intelligent Information Retrieval Univ. of Massachusetts, Amherst.

[8] J. Almazán, A. Gordo, A. Fornés, E. Valveny, 2014. "Segmentation-free word spotting with exemplar SVMs", Pattern Recognition, 47 (12), pp. 3967–3978.

[9] Y. Elfakir, G. Khaissidi, M. Mrabti, D. Chenouni, "Handwritten Arabic Documents Indexation using HOG Feature," International Journal of Computer Applications", vol. 126, no. 9, pp. 14–18, Sep. 2015

[10] R.A. Mohamad, L. Likforman-Sulem, C. Mokbel, 2009. "Combining slanted-frame classifiers for improved HMM-based Arabic handwriting recognition", IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (7) (2009) 1165–1177.

[11] Y. Kessentini, T. Paquet, A. Ben Hamadou, "Off-line handwritten word recognition using multi-stream hidden Markov models", Pattern Recognition Letters 31 (1) (2010) 60–70.

[12] A. Kundu, T. Hines, J. Phillips, B. Huyck, L. Van Guilder, 2007. "Arabic handwriting recognition using variable duration HMM", in: 9th International Conference on Document Analysis and Recognition (ICDAR), pp. 644–648.

[13] M .Cheriet, R. Farrahi Moghaddam, R. Hedjam, 2013. "A learning framework for automation and optimization of document binarization methods", Computer Vision and Image Understanding 117(3): 269-280.

[14] Y.Elfakir, G. Khaissidi, M. Mrabti, "Traitement des documents anciens par les classificateurs multi-niveaux", Colloque International sur le Monitoring des Systèmes Industriels, ENSA Marrakech, CIMSI14.

[15] H. Jégou, M. Douze, C. Schmid, "Product quantization for nearest neighbor search", IEEE Trans. Pattern Anal. Mach. Intell. 33 (1) (2011) 117–128.

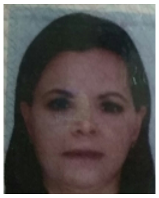
Ghizlane KHAISSIDI, National PhD holder in 2009 of the Sidi Mohamed Ben Abdellah University in Fez in Image Processing and Computer. Currently Professor at the National School of Applied Sciences (ENSA), University USMBA Fez (Morocco) and member of Lab of computering and interdisciplinary physics (L.I.P.I) at (E.N.S.F), Her research activities concern the image processing and its applications in medicine, heritage preservation (indexing of old manuscripts), societal dimension of applications (applications for the blind and visually impaired), handwriting and machine print recognition.

Youssef Elfakir received his Master degrees in 2013, both from the Department of Physics of the Faculty of Sciences Dhars Elmehraz, Fes University Marocco. He is currently a PhD student in Computer Science Engineering at the National School of Applied Sciences. His main research interests are image preprocessing, image analysis of visual manuscript and word spotting.

Mostafa MRABTI, state PhD holder in 1996 of the Sidi Mohamed Ben Abdellah University in Fez Automatic and Signal Processing. Currently Professor at the National School of Applied Sciences (ENSA), University USMBA Fez (Morocco) and member of Lab of computering and interdisciplinary physics (L.I.P.I) at (E.N.S.F),. His research activities concern the error correcting codes, implementation of signal processing algorithms on dedicated circuits, image processing, indexation of old manuscripts, handwriting and machine print recognition..

Mounîm A. El Yacoubi obtained his PhD in Signal Processing and Telecommunications from the University of Rennes I, France, in 1996. During his PhD, he was with the Service de Recherche Technique de la Poste (SRTP) at Nantes, France where he developed software for Handwritten Address Recognition that is still used for real-life Automatic French mail sorting. He was a visiting scientist for 18 months at the Centre for Pattern Recognition and Machine Intelligence (CENPARMI) in Montréal, Canada. He then became an associated professor (1998-2000) at the Catholic University of Parana (PUC-PR) in Curitiba, Brazil. From 2001 to 2008, he was a Senior Software Engineer at Parascript, Boulder (Colorado, USA), a world leader company in automatic processing of handwritten and printed documents (mail, checks, forms). Since June 2008, he is a Directeur d'Etudes / Professor at the University of Paris Saclay, Telecom SudParis within the Intermedia team. His main interests include Machine Learning, Statistical Pattern Recognition, Video, Image and Signal Processing, Human Gesture and Activity recognition, Human Robot (Computer) Interaction, e-Health, Video Surveillance, Biometrics, Smart Cities, Information Retrieval, and Handwriting and Machine Print Recognition.

Driss Chenouni received the Ph.D. degree in physics from the University of Montpellier II, France, in 1989, and the State Doctor's degree from the University of Fes, Morocco, in 1996. He is currently a Lab director of computering and interdisciplinary physics (L.I.P.I), at (E.N.S.F), and a Director of the Ecole Normale Supérieur at Sidi Mohammed Ben Abdellah University (USMBA), Fez, Morocco. His current research interests include Multi-Agent systems, Enterprise Architecture, Modeling, Web services, Autonomic computing, image processing and indexation of old manuscripts.



Zakia Lakhliai received the Ph.D. degree in physics from the University of Montpellier II, France, in 1987, and the State Doctor's degree from the University of Fes, Morocco, in 1996. She is currently a member of Lab of computering and interdisciplinary physics (L.I.P.I), at (E.N.S.F), and a Professor in the Superior School of technology (E.S.T.F.) at Sidi Mohammed Ben Abdellah University (USMBA), Fez, Morocco. Her current research interests signal and image processing, indexation of old manuscripts.