Neighborhood Label Extension for Handwritten/Printed Text Separation in Arabic Documents

Ahmad-Montaser Awal, Belaïd Abdel

Neighborhood Label Extension for Handwritten/Printed Text Separation in Arabic Documents

Ahmad Montaser Awal
AriadNext - Pôle R&D Document
35000 Rennes, France
Email: montaser.awal@ariadnext.com

Abdel Belaïd
LORIA - Campus scientifique
54506 Vandoeuvre Ls Nancy, France
Email: abdel.belaid@loria.fr

***Abstract*—This paper addresses the problem of handwritten and printed text separation in Arabic document images. The objective is to extract handwritten text from other parts of the document. This allows the application, in a second time, of a specialized processing on the extracted handwritten part or even on the printed one. Documents are first preprocessed in order to remove eventual noise and correct document orientation. Then, the document is segmented into pseudo-lines that are segmented in turn into pseudo-words. A local classification step, using a Gaussian kernel SVM, associates each pseudo-word into handwritten or printed classes. This label is then propagated in the pseudo-word's neighborhood in order to recover from classification errors. The proposed methodology has been tested on a set of public real Arabic documents achieving a separation rate of around 90%.**

I. Introduction

The objective of the methodology proposed in this paper is the processing of scanned documents in order to extract handwritten information from the printed one. This separation is a very important step in an automatic reading system. In fact, such separation allows handling each text type with a specialized recognition system: Optical Character Recognition (OCR) for the printed text and Intelligent Character Recognition (ICR) for the handwritten text. Documents containing a mixture of handwritten and printed text usually result from annotating printed documents or filling forms, an example of such documents is given in Figure 1. Basically, the separation problem can be regarded as a segmentation task. The document is firstly segmented into basic units approximating lines, words or characters. Then, a classification method is applied to associate each basic unit into one of the two text categories: handwritten or printed.

The paper is organized as follows: A brief state of the art is presented in section II. The proposed approach is detailed in section III. Section IV gives some experiments performed on public data-set and we conclude the paper in section V.

II. State of the art

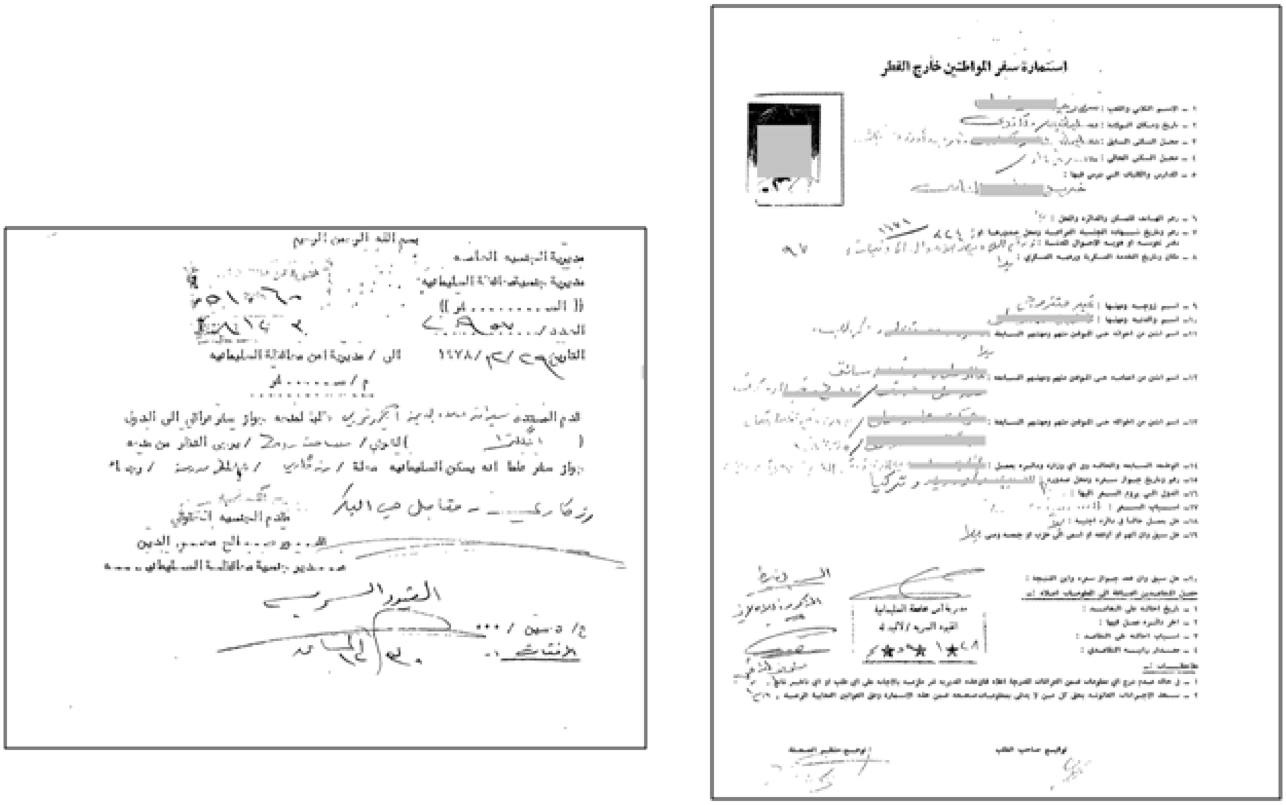
Handwritten and printed text separation problem is an active research area [4]. Classically, a separation system is composed of three main sub-parts: segmentation into basic units, classification of basic units, and label extension.

A. Segmentation

The segmentation method into basic units might differ depending on the script language. For example, handwritten and printed scripts are naturally separated into characters such as in Chinese [5], Lampung [25] and Hangul [27][28] scripts. Thus, segmentation into characters is sufficient to isolate mono-script text units. The authors in [5] use a two level X-Y cut algorithm to segment the document into text lines, and then into characters. Other language scripts, such as Latin and Arabic, are naturally separated into words (in Latin scripts) or piece of words (in Arabic scripts). Connected components might be grouped together in order to form the document basic units called pseudo-words [2], patches [10][11], or word blocks [9]. Geometrical proximity (height regularity) is used in [12][19] to merge connected components. Similarly, nearby connected components with overlapping pixels are merged together into one pseudo-word in [18]. The authors in [10][20] propose to avoid static distance thresholds by defining optimal word size dynamically for each document. They then apply a pixel growing algorithm to group pixels into patches with respect to the approximated word size. An Adaptive Run Length Smoothing Algorithm (ARLSA) [24] is used in [9] to obtain text lines. Then, a proximity criterion is estimated from the line's projection profile in order to merge close connected components into the same pseudo-word. A morphological closing by a 5×5 structuring element was used in [11] to find the pseudo-word patches. The segmentation at the word level is not necessary when text-lines are guaranteed to be mono-script. Projection profiles [6], or a smearing algorithm [7] are usually used in this context to estimate the inter-line separation distance and thus segment the documents into text lines.

B. Classification

Once the basic units are detected in the document, a classification method is required to associate each basic unit into the handwritten or the printed script. In most script languages, machine printed text is characterized by its regularity. Whereas, the handwritten text is of a more stochastic nature. In [6], three level structural descriptors are extracted



بسم الله الرحمن الرحيم

استمارة سفر المواطنين خارج القطر

Fig. 1. Example of Arabic documents containing handwritten and printed scripts [1]

from the Bangala and Devnagari scripts. These scripts are principally characterized by a head-line upper connection. This connection is less regular in handwritten text. A rule-based tree classifier was employed to achieve the classification task. A linear discriminant analysis was used in [7] to classify the extracted text lines based on features extracted from their upper and lower profiles. At the pseudo-words level, printed pseudo-words are more regular in terms of pixel distribution and connected component heights and alignment. The projection profile from observation windows was concatenated in [19] to feed an independent Hidden Markov Model for each script type. Da silva et al.[18] proposed a set of 11 features for each pseudo-word (structural features, projection profiles, pixel distribution). The classification task was then achieved by a rule based classifier. Another set of features was extracted in [12]: Gabor filter, crossing count histogram and bi-level co-occurrence. A Fisher classifier is then employed to associate the pseudo-word to the handwritten or the printed script. Fourier descriptors, Gabor filters, and Hu moments was extracted in [21]. A separate k-NN classifier was trained for each descriptor features, k-NN outputs were then combined using a simple majority vote. In [11], the authors extracted features at the pseudo-word and CCs levels. An alternative K-means classifier (called G-means) was used. The classification task was modeled directly by conditional random fields in [10], using a set of 23 features.

In this work, we use the pseudo-words as basic units since they are more stable than characters and text lines. Text lines are too global and may contain the two types of scripts. In addition, characters might be ambiguous in some languages between their handwritten and printed forms.

C. Label extension

The bottom up nature of the previous steps handicaps the separation system in difficult situations. In fact, the separation process is too easy for a human because he treats the incoming information in a global way integrating the neighborhood context. Neighborhood definition is not always an easy task. In image segmentation problems [13], labeling is performed at the pixel level. Though, an entity (pixel) neighborhood is naturally defined by pixel connectivity (4, 8, or others). This connectivity can be extended in the cases where the document is segmented into adjacent regions in a layout extraction problem [14]. The problem becomes more complex when the document is segmented into logical entities (lines, words ...) as it is the case in handwritten/printed text separation systems. The use of contextual information is introduced to recover from local classification errors. In this case, the contextual neighborhood of a basic unit is extended in order to have a more global decision. Label extension is only applicable on

[1] DATASET: Layer Separation - Arabic: http://lampsrv02.umiacs.umd.edu/projdb/project.php?id=61

characters or pseudo-words. In the case of text lines, the same label is already given to all the elements of the line.

Zheng et al. [12] proposed to only consider two horizontal neighbors to model a printed pseudo-word, whereas handwritten pseudo-words are not modeled to favor printed label propagation. A pseudo-word is considered as a neighbor of another if it satisfies proximity and similarity constraints. A convex hull distance is used in [11] to define the nearest 4 neighbors. The advantage of this distance is that it measures spatial similarity of the pseudo-words in addition to their spatial proximity. Neighbors number is increased to six in [10]. Euclidean distance between gravity centers of the pseudo words defines two vertical neighbors and four horizontal ones. We can notice that most of these previous works mainly focus on horizontal *local* neighbors of a pseudo-word.

The labels of pseudo-words belonging to the same neighborhood are then modified to be homogeneous. In [15][19], the decision was achieved by a majority vote of the neighbors. A similarity condition was added in order to guarantee the homogeneity of the ensemble. In [11][12], the optimal label configuration of the pseudo-words was modeled by a Markov Random Field (MRF). The MRF model measures the classification confidence by means of a dependence function. The similarity between the pseudo-word and its neighbors was measured by a similarity function. An inference method was required to propagate these two functions. This can be achieved by Gibbs quantification as in [12] where the performance was improved from 96.1% to 98.1%, or by a belief propagation method as in [11] where the performance is improved from 94.2% to 95.5%. The main drawback of MRF models is their complexity. Indeed, an enumeration of all labeling configurations is needed to find the optimal one. It has been shown in [10] that discriminative models were more efficient as the optimal labeling configuration can be directly modeled by a Conditional Random Field (CRF). An approximate comparison showed that using CRF model improves system's performance from 98.0% to 99.1%.

III. Proposed approach

We propose to segment the document into pseudo-lines and pseudo-words based on the proximity and the similarity of connected components. A multiclass SVM classifier is then used to categorize the segmented pseudo-words. A variety of methods is then explored for the label extension showing that the local neighborhood is not sufficient and a more global definition improves the global performance.

Firstly, we try to improve the document image quality by applying some necessary preprocessing. Salt and Pepper noise is eliminated by an adaptive k-fill algorithm [22], we propose to estimate the k value of the algorithm by quantifying the noise present in the image. We proceed then to correct the document skew by applying the RAST algorithm [23]. Finally, inappropriate connected components are discarded based on geometrical criteria.

A. Segmentation into pseudo-words

We propose to extract document pseudo-words by applying a two-level segmentation algorithm [2]. A proximity graph of the connected components is first constructed. Each isolated sub-graph is considered as a text pseudo-line. This step is based on the assumption that the document skew is already corrected and cannot handle multi-skew documents in which case a more adapted algorithm is required [1][16][26]. Then, an inter-word distance is dynamically estimated for each detected pseudo-line. This distance is used to construct in-line connected components graph. Similarly, isolated sub-graphs are considered as the pseudo-words of the currently processed pseudo-line.

B. Pseudo-word classification

A set of 11 groups of descriptors inspired from the state of the art [12][15][18] is extracted for each pseudo-word:

1) Connected components structural descriptors (8).
2) Hu moments (7).
3) Vertical profile (1).
4) Horizontal profile (4).
5) Pixel distribution (1).
6) Mean line (8).
7) Run length (20).
8) Crossing count (10).
9) Vertical segments (2).
10) Bi-level Co-occurrence (1D texture) (16).
11) Bi-level 2×2-gram (2D texture) (60).

Thus, a total of 137 features characterizes the pseudo-word. Afterwards, the classification task is achieved by an SVM (1-vs.-all) [17]. The classifier associates each pseudo-word to one of the classes handwritten (H), printed (P) with a confidence f.

C. Label extension

Firstly, we investigated the use of local neighborhood of pseudo-words.

a) k-NN: This first technique is based on grouping the k nearest neighbors of a given pseudo-word as its neighborhood. The label of the majority (over 50%) of the pseudo-words is extended to the central unit. Furthermore, small pseudo-words are not considered in order not to interfere with the label extension. Indeed, the accumulated count of the neighborhood pixels must be significant compared to the central unit.

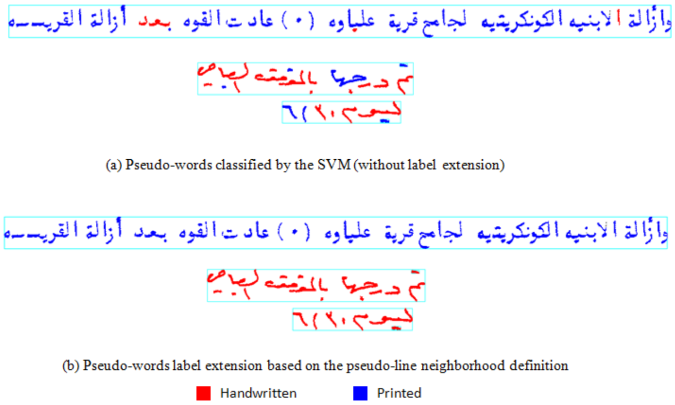
b) Confidence propagation: Based on the horizontal nature of text-lines, we propose the use of the nearest horizontal neighbor. However, the label of the neighbor is extended to the pseudo-word only if its classification confidence is stronger than that of the pseudo-word. The neighbor confidence is weighted by its distance in order to reduce the effect of far neighbors.

The main drawback of these two algorithms is the need to a predefined distance threshold which is image resolution dependent. Furthermore, their complexity is of $O(n^2)$ (where n is the number of pseudo-words).

c) Pseudo-lines: A novel neighborhood definition based on text pseudo-lines is proposed. In fact, a pseudo-line presents a logical relationship between its pseudo-words. Even if pseudo-lines are mixed and contains both scripts, a statistical study on the training data-set showed that around 93% of pseudo-lines contain a unique label. The label extension is thus applied as follows. The dominant class is first defined by the class the more present in the pseudo-line. In case of equality, the dominant class is that with highest average classification confidence of its pseudo-words. Let f_i be the classification confidence of the current pseudo-word, and h_i its height. The dominant class label is extended to a pseudo-word if it verifies the following condition $(f_i < cf) or (|h_i - H_D| < d)$ where $0 \leq cf \leq 1$ is the confidence threshold and d determines the regularity degree. Thus, a label will be changed into the dominant class label in one of the two following cases: 1) the classification confidence is low (first term of the condition) or 2) the pseudo-word has a similar height as the dominant class in which case the classifier decision is ignored (the second term). This later case is inspired from printed text lines where most of the words have a height similar to the height of the pseudo-line reflecting the regularity of the text line. This hypothesis is less present for handwritten text. We can notice from the algorithm 1 that the new method is less complex than the previously proposed one for two reasons. First, the complexity is of $O(n)$. Second, differently from the max distance threshold, the used thresholds cf and d are image resolution independent. In fact, they reflect the label extension freedom degree. In the case of the certainty factor $cf = 0$, a total confidence is given to the classifier and its decision is considered always true. On the other hand, $cf = 1$ indicates that the classifier decision is not considered. For the experiments held in this paper, we have set $cf = 0.9$. Similarly, a very small value of the regularity factor d indicates that a high regularity is required. Thus, only pseudo-words with the exact same height as the dominant class are changed. In contrast, a higher value of d gives more freedom even with heights highly different from the height of the dominant class. An example of pseudo-word label extension based on the pseudo-line neighborhood definition is given in Figure 2

Algorithm 1 Pseudo-line based relabeling

```
Input: List of pseudo-words $L$, Certainty factor $cf$, Regularity factor $d$
for each pseudo-line $l \in L$ do
    $C_D \leftarrow dominant\ class\ in\ l$
    $H_D \leftarrow median\ height\ of\ C_D$
    for each pseudo-word class $w$ in $l$ do
        if $(f_i < cf)$ or $(|h_i - H_D| < d$ then
            Assign the label $C_D$ to the current word $w$
        end if
    end for
end for
```


(a) Pseudo-words classified by the SVM (without label extension)

(b) Pseudo-words label extension based on the pseudo-line neighborhood definition

Fig. 2. Label extension based on pseudo-lines

IV. Experiments

The whole system was evaluated using a public Arabic documents data-set [1]. All documents are labeled at the pixel level. A total of 108 documents were used to train the SVM classifier and tune the system parameters, whereas another 25 documents were used as a test data-set. The separation system is evaluated using two metrics. The rate at the pixel level which is the ratio of pixels correctly labeled compared to the ground truth images [8]:

$$pix\_rate = \frac{\#pixels\ correctly\ classified}{\#total\ of\ the\ class\ pixels} \quad (1)$$

Another evaluation measurement is used in the literature based on the rate of pseudo-words labeled correctly:
Pseudo-word_rate=

$$\frac{\#Pseudo - words\ correctly\ classified}{\#total\ of\ pseudo - words} \quad (2)$$

Our system has reached very competitive results compared to those of the state of the art when applied to Latin script separation [3]. Experiments held for this paper on the Arabic documents data-set has shown very promising results. We can notice from the Table I that all label extension methods allow improving the pseudo-words separation rate. More specifically, the pseudo-line based extension outperforms the other two extension methods and allows reaching a very excellent rate on the printed pseudo-words level (97.3%). It also improves the handwritten pseudo-words separation rate from 63.9% to 81.6%.

V. Conclusion

A handwritten printed text separation method for Arabic documents has been proposed. A two level segmentation process is employed to segment the document into basic units (pseudo-words). It has been shown that local neighborhood

[1] DATASET: Layer Separation - Arabic: http://lampsrv02.umiacs.umd.edu/projdb/project.php?id=61

TABLE I
OUR SEPARATION SYSTEM EVALUATED ON A PUBLIC ARABIC DOCUMENTS DATA-SET

	Pseudo-word_rate		Pixels_rate	
	P%	H%	P%	H%
SVM	91.6	63.9	89.3	83.7
SVM + k-NN	93.3	71.1	90.1	83.1
SVM + Confidence propagation	93.4	64.8	90.5	83.2
SVM + Pseudo-line based extension	97.3	81.6	90.4	83.6

definition is not always sufficient to efficiently extend pseudo-word classification decision. In consequence, a new neighborhood definition has been proposed in order to obtain more global contextual information of the pseudo-words. Experiments have shown a significant improvement of the separation process when extending the pseudo-word classification label. An improved version of our system has been employed to the task of handwritten printed text separation in Latin documents. It achieves very results in a very competitive performances compared to the state of the art. More specifically, a third information layer (noise) is extracted improving the rappel of the two other layers (handwritten and printed). Furthermore, the segmentation method is adapted with the Latin script nature. We think that applying similar adaptations based on the Arabic script nature and introducing the noise information layer will significantly improve the system performance on Arabic documents.

REFERENCES

[1] Nazih Ouwayed and Abdel Belaïd, A general approach for multi-oriented text line extraction of handwritten documents, International Journal on Document Analysis and Recognition, Springer Verlag, vol. 14, n. 4, pp. 914-923 (2011)

[2] A. Belaïd, K. Santoch and V. P. d'Andecy, Handwritten and Printed Text Separation in Real Document, in Machine Vision Applications, vol. 2, pp. 9-21 (2013)

[3] A.-M. Awal, A. Belaïd and V. P. d'Andecy, Handwritten/printed text separation Using pseudo-lines for contextual re-labeling, in International Conference on Frontiers in Handwriting Recognition, pp. 29-34 (2014)

[4] R. Srivastva, A. Raj, T. Patnaik and B. Kumar, A survey on tenchiques of separation of machine printed text and handwritten text, International Journal of Engineering and Advanced Technology, vol. 2, no. 3, pp. 552-555 (2013)

[5] K.-C. Fan, L.-S. Wang and Y.-T. Tu, Classification of machine-printed and handwritten texts using character block layout variance, Pattern Recognition, vol. 31, n. 9, pp. 1275-1284 (1998)

[6] U. Pal and B. B. Chaudhuri, Machine-printed and hand-written text lines identification, Pattern Recognition Letters, vol. 22, n. 3, pp. 431-441 (2001)

[7] E. Kavallieratou and S. Stamatatos, Discrimination of Machine-Printed from Handwritten Text Using Simple Structural Characteristics, in International Conference on Pattern Recognition, vol. 1, pp. 437-440 (2004)

[8] F. Shafait, D. Keysers and T. M. Breuel, Performance Evaluation and Benchmarking of Six-Page Segmentation Algorithms, Pattern Analysis and Machine Intelligence, vol. 30, n. 6, pp. 941-954 (2008)

[9] K. Zagoris, L. Pratikakis, A. Antonacopoulos, B. Gatos and N. Papamarkos, Distinction between handwritten and machine-printed text based on the bag of visual words model, Pattern Recognition, vol. 47, n. 3, pp. 1051-1062 (2014)

[10] S. Shetty, H. Srinivasan and S. Srihari, Segmentation and Labeling of Documents using Conditional Random Fields, in Document Recognition and Retrieval IV, Proc. SPIE 6500, 65000U, (2007)

[11] X. Peng, S. Setlur, V. Govindaraju and R. Sitaram, Handwritten text separation from annotated machine printed documents using markov random fields, International Journal on Document Analysis and Recognition, vol. 16, n. 1, pp. 1-16 (2011)

[12] Y. Zheng, H. Li and D. Doermann, Machine Printed Text and Handwriting Identification in Noisy Document Images, Pattern Analysis Machine Intelligence, vol. 26, n. 3, pp. 337-353 (2004)

[13] X. He, R. S. Zemel et M. A. Carreira-Perpinn, Multiscaleconditional random fields for image labeling, chez CVPR (2004)

[14] S. Nicolas, J. Dardenne, T. Paquet et L. Heutte, Document Image Segmentation Using a 2D Conditional Random Field Model, chez Ninth International Conference on Document Analysis and Recognition (2007)

[15] R. Kandan, K. Arvind and A. Ramakrishnan, Localization of Handwritten text in Documents using moment invariants and Delaunay Triangulation, in International Conference on Computational Intelligence and Multimedia Applications, pp. 96-105 (2007)

[16] S. S. Bukhari, F. Shafait, T. M. Breuel, Coupled snakelets for curled text-line segmentation from warped document images. IJDAR, vol. 16, n. 1, pp. 33-53 (2013)

[17] C.-W. Hsu and C.-J. Lin, A comparison of methods for multi-class support vector machines, IEEE Transactions on Neural Networks, vol. 13, n. 2, pp. 415-425 (2002)

[18] L. F. da Silva, A. Conci and A. Sanchez, Automatic Discrimination between Printed and Handwritten Text in Documents, in Brazilian Symposium on Computer Graphics and Image Processing, pp. 261-267 (2009)

[19] J. K. Guo and M. Y. Ma, Separating Hadwritten Material from Machine Printed Text Using Hidden Markov Models, in International Conference on Document Analysis and Recognition, vol. 1, pp. 439-443 (2001)

[20] M. Shirdhonkar and M. B. Kokare, Discrimination between Printed and Handwritten Text in Documents, IJCA Special Issue on Recent Trends in Image Processing and Pattern Recognition, vol. 3, pp. 131-134 (2010)

[21] M. Benjlaiel, R. Mullot and A. M. Alimi, Multi-oriented handwritten annotations extraction from scanned documents, in International Workshop on Document Analysis Systems, pp. 126-130 (2014)

[22] K. Chinnasarn, Y. Rangsanseri and P. Thitimajshima, Removing Salt-and-Pepper Noise in Text/Graphics Images, in The Asia-Pacific Conference on Circuits and Systems, pp. 459-462 (1998)

[23] J. van Beusekom, F. Shafait and T. M. Breuel, Combined orientation and skew detection using geometric text-line modeling, International Journal on Document Analysis and Recognition, vol. 13, n. 2, pp. 79-92 (2010)

[24] N. Nikolaou, M. Makridis, B. Gatos, N. Stamatopoulos and N. Papamarkos, Segmentation of historical machine-printed documents using Adaptive Run Length Smoothing and skeleton segmentation paths, Image and Vision Computing, vol. 28, n. 4, pp. 590604 (2010)

[25] Junaidi, A., Grzeszick, R., Fink, G.A. and Vajda, S., 2013, August. Statistical Modeling of the Relation Between Characters and Diacritics in Lampung Script, 12th International Conference on Document Analysis and Recognition (ICDAR), pp. 663-667, (2013)

[26] V. Szilrd, T. Pltz, and G. A. Fink, Camera-Based Whiteboard Reading for Understanding Mind Maps, International Journal of Pattern Recognition and Artificial Intelligence, vol. 29, n. 03 (2015

[27] Ghosh, D., Dube, T. and Shivaprasad, A.P., Script recognitiona review, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32, n. 12, pp. 2142-2161, (2010)

[28] Kim, H.Y. and Kim, J.H., Hierarchical random graph representation of handwritten characters and its application to Hangul recognition, Pattern Recognition, 34(2), pp.187-201, (2001)