Adaptive Detection of Missed Text Areas in OCR Outputs: Application to the Automatic Assessment of OCR quality in Mass Digitization projects

Ahmed Ben Salah, Nicolas Ragot, Thierry Paquet

Adaptive Detection of Missed Text Areas in OCR Outputs: Application to the Automatic Assessment of OCR quality in Mass Digitization projects

Ahmed BEN SALAH [ab], Nicolas RAGOT [c] and Thierry PAQUET [b]

[a] Bibliothèque nationale de France, Quai François-Mauriac, Paris XIII, France;
[b] Université de Rouen-LITIS, Avenue de l'Université, Saint-Étienne-du-Rouvray, France;
[c] Université François Rabelais Tours-LI, 64 avenue Jean Portalis, Tours, France

ABSTRACT

The French National Library (BnF*) has launched many mass digitization projects in order to give access to its collection. The indexation of digital documents on *Gallica* (digital library of the BnF) is done through their textual content obtained thanks to service providers that use Optical Character Recognition softwares (OCR). OCR softwares have become increasingly complex systems composed of several subsystems dedicated to the analysis and the recognition of the elements in a page. However, the reliability of these systems is always an issue at stake. Indeed, in some cases, we can find errors in OCR outputs that occur because of an accumulation of several errors at different levels in the OCR process. One of the frequent errors in OCR outputs is the missed text components. The presence of such errors may lead to severe defects in digital libraries. In this paper, we investigate the detection of missed text components to control the OCR results from the collections of the French National Library. Our verification approach uses local information inside the pages based on Radon transform descriptors and Local Binary Patterns descriptors (LBP) coupled with OCR results to control their consistency. The experimental results show that our method detects *84.15%* of the missed textual components, by comparing the OCR ALTO files outputs (produced by the service providers) to the images of the document.

Keywords: OCR results assessment, Text detection, Texture characterization

1. INTRODUCTION

The collections of the French National Library (BnF) contain works of art and documents that have marked the history of the humanity. For preservation purpose most of these documents are not easily accessible. To present these documents to the public, the French National Library has launched its first mass digitization program in 1996. Gallica†, BnF's digital library, contains the images produced since then. Before 2006, document queries on Gallica were performed using bibliographic references of the documents. Since 2006, to allow full text search, the BnF has launched mass digitization projects including textual transcriptions. This is a big issue since the document collections from the National Library of France are very variable in terms of physical characteristics. Indeed, they contain documents that come from several periods of printing.

The transcriptions are carried out by OCR providers that deliver ALTO[1] files. These files include the word transcriptions with associated word confidence computed by the OCR and the segmentation results. In mass digitization projects, the BnF ask the OCR providers either Raw Quality or High Quality (HQ) transcriptions. Raw Quality is directly obtained by the OCR output when the estimated word recognition rate, computed using the word confidence scores, is between 60% and 99%. High Quality requires an estimated word recognition

Ahmed.Ben.Salah.: E-mail: ahmed.ben-salah@bnf.fr, Telephone: +33 (0) 1 53 79 41 64
Nicolas.Ragot.: E-mail:nicolas.ragot@univ-tours.fr, Telephone: +33 (0)2 47 36 14 31
Thierry.Paquet.: E-mail:Thierry.Paquet@univ-rouen.fr, Telephone: +33 (0)2 32 95 50 13
This work is done in the context of the ANR project *Digidoc*
*Bibliothèque nationale de France
†http://gallica.bnf.fr

rate of 99.99%. Of course, current technologies cannot provide High Quality outputs by themselves, and most of the time, manual corrections are required. It has been estimated that below an estimated word recognition rate of 85%, manual corrections would be both tedious and costly. Therefore manual corrections are asked only for documents that exhibit OCR outputs with an estimated recognition rate higher than 85%. Of course, the reliability of commercial OCR systems depends highly on the physical characteristics of the document. Moreover, the quality of the transcription can only be estimated by the system that provides a word confidence ratio for each recognized word. This is why an OCR verification stage should be introduced at the BnF before integrating the digitized documents into the Gallica digital library. Full human quality control would be too long when performed on the whole digitized corpora (around 30000 pages a day). Consequently, automating this control or part of it can be very useful.

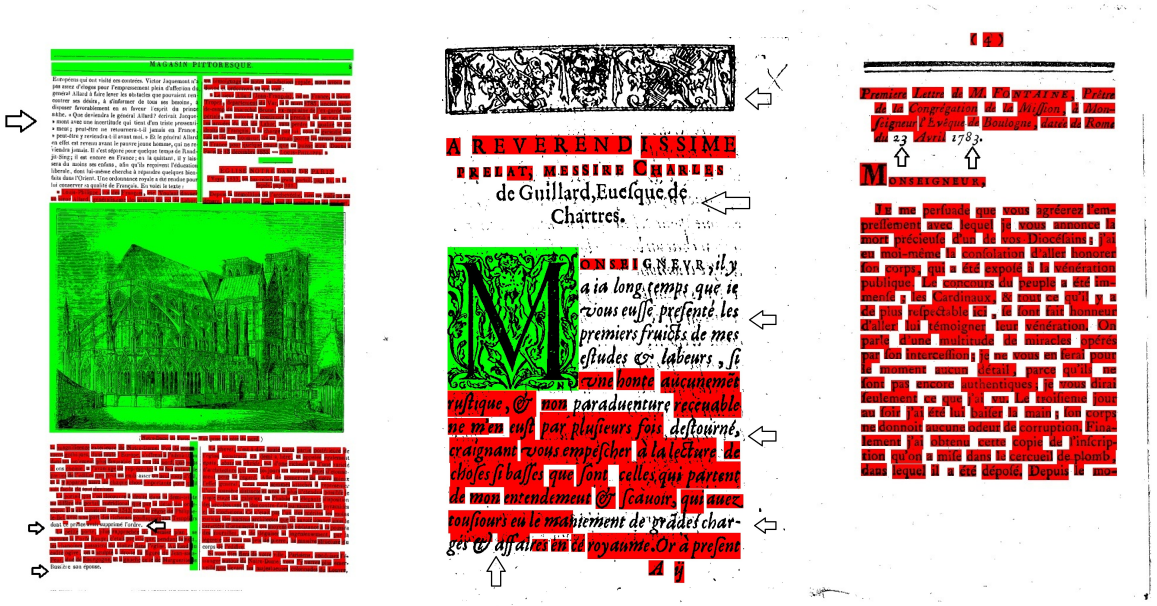
The BnF digitization department has identified that most of OCR errors are coming from a significant amount of missed text components in the OCR outputs. This is why this study focuses on the detection of these missed textual components. Because of the complexity of OCRs architecture,[2] it is rather difficult to identify the reasons of these defects. Sometimes they occur at the pre-processing stage (such as binarization). Sometimes, defects come from the segmentation process, while some other defects come from the character recognition stage. The approach we propose to detect the missing textual components does not concentrate on the defects of each of the processing stages. It provides a single methodology able to detect the missed textual components such as characters, words or text blocs. In fact, our approach exploits the local information inside a given page and the ALTO result produced by the OCR provider in order to detect the missed textual components. More precisely, textual areas detected by the OCR are used as training samples for a classifier whose task is to discriminate between text and other elements in the document image. Such page dependent procedure provides a generic methodology able to perform on any kind of document and particularly on old documents for which standard OCRs perform moderately well and require a lot of verification efforts at the BnF.

In section two of this paper, we analyze the various situations of missed textual elements. We try to explain the reasons of these defects by referring to the complexity of the OCR processing chain and we explain why these defects sometimes escape from quality control. In the third part, we describe the methodology of the proposed approach. Next, we report the experiments that have been carried out on a set of document images coming from the BnF collections. Then we conclude this article with some possible improvements of this work.

2. OVERVIEW OF OCR DEFECTS

An OCR system is composed of multiple processing stages organized sequentially. It includes preprocessing and binarization of the image, segmentation, and character and symbol recognition. This sequential organization tends to cumulate (even amplify) failures done at each stage. Nagy et al[3] show that preprocessing and segmentation failures can result from physical defects of document such as ink defects and paper defects. Furthermore, document image defects such as low contrast and background noise may lead to wrong segmentation results. Regarding the character recognition stage, errors may come from exotic character styles or may be due to the presence of noise. These errors cannot be detected easily by analyzing the OCR output only, since they are forgotten during the OCR process either during preprocessing and segmentation stage or during the recognition process when they induce a low word confidence.

According to the BnF digitization department, most of the errors in OCR outputs are missed textual components. The examples presented in figure 1 show the diversity of the cases that can be encountered. In this figure, the red and green areas refer respectively to text and illustration elements detected by the OCR. Figure 1 (I) shows a large missed text block. It is rather difficult to explain these failures, whereas some nearly similar components are correctly detected in the same document. Sometimes one single word is missed in a well segmented block (cf. figure 1 (III)). In this case, we may assume that text in italic has low word confidence scores and may be rejected at the end of the character recognition stage. Also in some cases, as in figure 2, one single character is missed probably due to its exotic shape.Figure 1 (II) also shows some missed textual components in the OCR output. These missed elements might occur because of different phenomena. For example, the presence of noise (bleed through effect) may disturb the character recognition stage. Some text areas are also missed in the title, although the character shapes do not exhibit any difficulty for the recognition stage. Finally, we may conjecture that the italic style has also disturbed the recognition stage. One can also notice that the illustration



(I) (II) (III)

Figure 1. The labeled mask areas : (I) case of a missing section in the OCR output, (II) case of a missing words, sentences and graphical elements in the OCR output, (III) case of missing words in the OCR output. The red bounding boxes represent the recognized words by the OCR, the green bounding boxes represent the detected graphical elements by the OCR

in the upper part of the image is missed although it should be easily detectable since it does not overlap with any other components. The reason for this error is not clear.

Most of the OCR systems try to assess the quality of the output delivered. Most of the time, it is done at the final step, estimating the performance of the character recognition stage by the computation of a word confidence score. This one is obtained by combining the confidence of the character recognizer with the word matching score to a dictionary. This word confidence score ranges from 0 to 1 but its definition depends on the OCR provider and on the OCR technology. Based on it, the page quality score is the average word confidence over a page after rejection of the poorly recognized textual elements (those with a low word confidence). The maximization of the page quality score can then be obtained either by maximizing the recognition rate or simply by increasing the number of rejected text components. As a consequence, documents with a high page quality score may have missed textual components. Thus, if one look only at the page quality score to assess the quality of the digitization process it may leads to inconsistencies. This is why, in order to control the quality assessment of the output of a mass digitization process, it is necessary to design a post processing stage able to quantify the OCR defects in number of missed textual components.

3. DETECTION OF MISSING TEXT COMPONENTS IN OCR RESULTS

3.1 Overview of the method and state of the art

Within this study, our aim is not to improve a segmentation process by providing feedbacks to an OCR or to the OCR provider but to give to the BnF new quality insigths about the documents by detecting textual components (blocks, lines, words, and characters) that the OCR was not able to find. Let us call them the missed components. Compared to the huge amount of works dedicated to document image segmentation, this study investigates new incites of page layout segmentation oriented towards document segmentation verification. This subject of detection of missed components in OCR outputs has not been much considered until now, but appears to be of primary importance for managing digitization workflows at the BnF.

Considering that the BnF's collection of documents is very heterogeneous - since it covers several centuries of printed documents - we must investigate a general method to detect the missed components, that does not

Figure 2. Case of a missing character in well segmented word

depend neither on the typography nor on the style of documents. We assume that the missed components can only occur in empty areas or background e.g. outside the word bounding boxes and outside the graphical regions already detected[‡]. The idea of our approach is to learn the properties of each element detected by an OCR and to try to find similar elements in the empty regions. In this way our method fits very well with typographic and structural characteristics of the current document to detect missed components inside it.

According to the state of the art presented by Chen et al,[4] low levels features such as color, texture and shape characteristics are usually used to extract textual elements. Zhong et al[5] assumes that spatial characteristics of textual elements can be used for a connected component analysis to detect some text inside an image. Other works, such as the approach proposed by Smith et al[6] use the vertical edges of text with a predefined template to group edges and find the text elements in an image. Other approaches use *a priori* knowledge to locate the text elements. For example, Sobottka et al's[7] approach is based on specific text features, such as baseline feature, in order to identify text in images. In fact, strings are characterized by a top and bottom baseline that help to determine the presence of text in an image area. While this approach is applicable with serif fonts, it may not give good results for old documents that have exotics fonts. Garcia et al.[8] use the fact that text strings contain edges in many orientations to detect text components in an image. The proposed method combines the variation of edge orientation computed in local area with edge features to describe the text components. The main advantage of such model-driven methods and *a priori* methods is that they are very fast. However, they adapt hardly with heterogeneous documents collections. In fact, model-driven approaches are generally used for specific fields of application, while the BnF collection includes several kinds of documents (press, scientific books, maps, letters, Fax....) and several printing technology, this kind of approach is not suitable in our context.

A texture approach was proposed by Wu et al[9] to detect and locate the text in an image. For each pixel, energy features are computed on derived images at different scales. The family of texture descriptors allows the characterization of document images with low-level information. This kind of features allows to locate text components without using any further information about the physical structure model or the typographical model of the documents. This characteristic corresponds to the diversity of the BnF's collection and this is why we used a texture approach to detect the missed components. According to the state of art presented by Journet et al,[10] there are four types of texture descriptor: statistical descriptors, probabilistic descriptors, frequency descriptors and geometric descriptors.

Statistical methods such as Grey Level Co-occurrence matrix proposed by Haralick et al[11] characterize the regularity, the repetition and the contrast of textures. Other statistical approaches, such as the work of Uttama et al[12] use the autocorrelation function to describe the texture pattern. The statistical approaches study the grayscale intensities of pixels to compute the texture feature. According to Mryka,[13] the statistical methods have a good performance. However, they are not very suitable for old document images. In fact, the grayscale

[‡]Graphical regions misclassified as textual regions and the reverse cases are assumed far less frequent, and do not fall within the scope of this study

images of old documents are similar to binary images due to the printing technologies. Therefore, the variation of pixel intensity in the image may be caused by digitization process and not by texture. For this reason we cannot apply the statistical methods in our context. Caron et al[14] use an approach that belongs to the family of probabilistic methods to identify areas of interest in a image of natural scene. This approach use the power law (Zipf's law) to perform the detection of the textual areas.

Raju et al[15] propose another method for text detection from complex color images that belongs to the family of frequency methods. This approach performs text localization using a Gabor function based on multi-channel filtering of the intensity component of the image. But according to Journet et al,[10] in some cases , eg. old documents, the frequency approaches for texture description may do some misclassification between the drawn illustrations and the text.

The geometric methods characterize the texture shapes and their spatial relationships. Tuceryan[16] uses geometric moments to characterize textual texture in document image. Another method proposed by Journet et al[10] for text/graphic separation is based on an autocorrelation function to study the main orientation of texture. In fact, in multi-resolution process, the texture of the text components has only one main direction at different scales. Whereas, the texture of graphic components has different main directions at different scales. The geometrical approaches are applicable both on binary images and grayscale images. Also, some geometric features as main direction of texture are suitable for the characterization of any kinds of fonts or graphics components (printed or drawn). This is why we have chosen to develop a new approach based on geometric techniques to detect the missed text components.

In our approach, the areas detected by the OCR provide a labeled mask of the image (cf. figure 1) which separates the background information from the foreground information (either graphics or textual elements). The foreground information is used for training a specific text and graphic detector assuming that the information provided by the OCR is reliable. Then the background information is analyzed using this detector in order to detect the missed textual components, if any. The proposed approach must deal with all kind of documents in the BnF's collection. This is why we need to train classifiers with the local characteristics of each page (both text and graphic elements). However, the characteristics of background and inter-word space are learned using a set of images [§] that includes binary images and grayscale images. Learning is performed by training four SVMs classifiers using the signature of the pixels of one given area as positive examples and others as counter-examples (one versus all classification scheme). The choice of the classifier was made after several experiments of supervised classifiers. The best results were obtained using SVM with linear kernel. After this learning step, pixels falling outside the initial labeled mask are classified as text, graphics, inter-word spaces and background. The inter-word spaces areas are then merged with empty areas in one single class corresponding to background areas. A final connected component analysis is conducted on text regions in order to provide an estimation of the number of missed textual components in the image. More details about the connected component analysis are given in the section 3.3.

3.2 Characterization of the areas

The proposed method was conceived so as to adapt to the various document contents that can be encountered with either textual or graphical content. Following this general goal, we choose to characterize regions with general purpose texture features that can be applied both on black and white and gray level images, since the BnF collections contain both kind of images. The multi-scale Radon features[18] have been chosen for their ability to characterize oriented textures such as textual images. LBP features[19] have also been chosen as general purpose texture descriptors. Finally, a paper to ink transition descriptor and the mean pixel intensities have been introduced so as to characterize textual areas. So for each pixel we have extracted 12 features to characterize the textures of the image. The following subsections summarize these feature descriptors.

[§]To characterize the background regions, we built a set of background images that includes binary images and grayscale images. This set of images contains some image defects (such as white noise) and some physical defects (such as small inkblot) .

3.2.1 Radon transforms descriptors

The Radon transform allows identifying the main direction of a texture. This information is very useful in our work. Indeed, according to Journet et al,[10] printed areas should exhibit a constant orientation at different scales. On the contrary, the main direction of the texture should change at each scale on graphical areas. The behavior of Radon transform on background areas is similar to its behavior on graphical areas: various orientations at each scale. However, the intensities of its responses are weak on background areas compared to those on informative areas. The Radon transform computes projections of the image along some specified directions by rotating the image around its central point. According to Deans,[18] the Radon transform of $f(x, y)$ is the line integral of f parallel to the y'-axis (along y-axis).

$$R_k(\theta) = \int_{-\frac{l}{2}}^{+\frac{l}{2}} \int_{-\frac{l}{2}}^{+\frac{l}{2}} (f(x' \cos(\theta) - y' \sin(\theta), x' \sin(\theta) + y' \cos(\theta)) d_{y'} d_{x'} \tag{1}$$

In this study, we applied the Radon transform on three windows of size 128×128 ($k = 1$), 64×64 ($k = 2$) and 32×32 ($k = 3$) and seven directions $\theta \in \{0, 30, 45, 60, 90, 120, 150\}$ so as to identify the main orientations of the texture at each scale. Thus our Radon index $\delta_k(i, j)$ is given by:

$$\delta_k(i, j) = \arg\max_{\theta}(\max(R_k(\theta))) \tag{2}$$

with $\theta \in \{0, 30, 45, 60, 90, 120, 150\}$
From the $\delta_k(i, j)$ value, 7 texture descriptors $(f_1, ..., f_7)$ are computed. The first one measures the consensus of the principal direction at each scale. This means that for each pixel, we compute the principal direction of the R function at each of the three scales. If the principal direction is the same, our descriptor value is 2. If the main direction is the same only in two scales its value is 1. Otherwise, its value is 0.

$$f_1(i, j) = (\delta_1(i, j) == \delta_2(i, j)) + (\delta_2(i, j) == \delta_3(i, j)) \tag{3}$$

The next three features are the median over the maximum projection of each orientation computed at each of the three scales. By definition, the maximum intensities of orientations are large on well inked areas(illustrations), less important on printed areas and very low on background regions which are very bright most of the time.

$$f_{k+1}(i, j) = median(\max(R_k(\theta))) \tag{4}$$

with $\theta \in \{0, 30, 45, 60, 90, 120, 150\}$ and $k = 1, 2, 3$.
Similarly, the variance of maximum projection over each direction is computed at each scale. The variance is large on regions that exhibit several directions. It is low on textual areas.

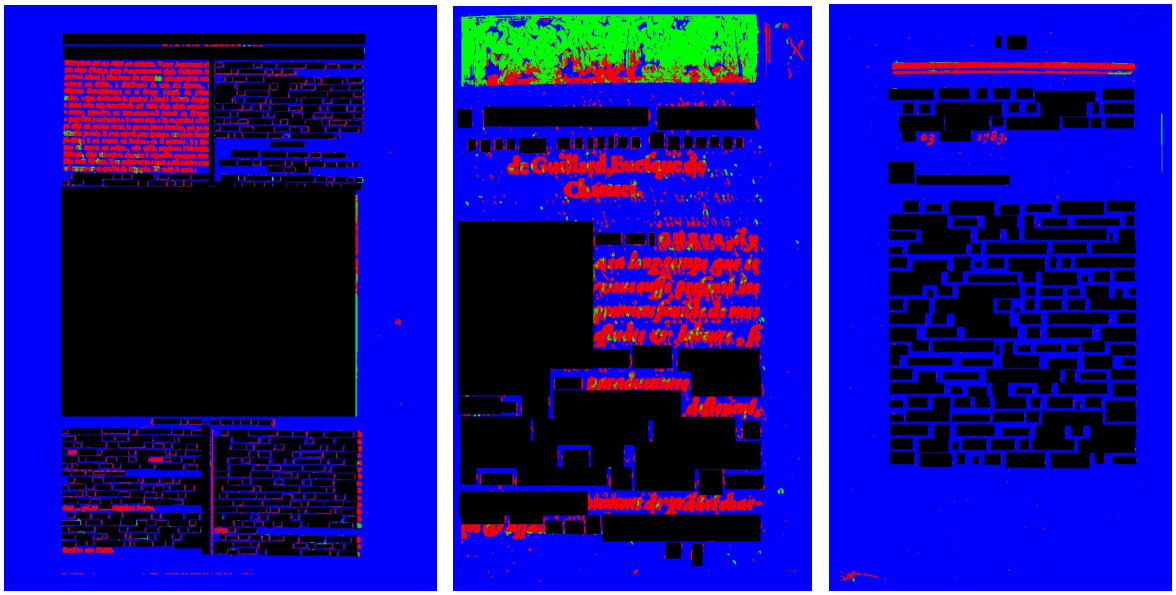
$$f_{k+4}(i, j) = std(\max(R_k(\theta))) \tag{5}$$

with $\theta \in \{0, 30, 45, 60, 90, 120, 150\}$ $\delta_k(i, j)$ and $k = 1, 2, 3$.

3.2.2 Local Binary Patterns descriptors

The original implementation of the LBP operator worked with the eight neighbors of a considered pixel. The sign of the difference between intensity of the central pixel and the intensity of neighboring pixels is used as a threshold to product a signature. This represents information on regular patterns in the image, in other words on the texture.
Let (x_c, y_c) a considered pixel with gray level g_c and $g_p(p = 0, ..., P - 1)$ be the gray values of the P neighboring pixels on the circle of radius $R(R > 0)$. The LBP descriptors used here are given by equation 6.

$$f_{k+7} = LBP_{P,R}(x_c, y_c) = \sum_{p=0}^{P-1} s(g_p - g_c) 2^p \tag{6}$$



(I) (II) (III)

Figure 3. Results of applying our approach of detecting missed components on the example of figure 1: red pixels indicate the missed text elements, green pixels indicate the missed graphical elements and blue pixels indicate the empty regions

with $s(x) = \begin{cases} 1 x \geq 0 \\ 0 x < 0 \end{cases}$, $k = 1, 2, 3, 2^p$ for the weight of difference and for $(k+7 = 8)$ $(P = 8, R = 1.0)$; for $(k+7 = 9)$ $(P = 12, R = 2.5)$; for $(k + 7 = 10)$ $(P = 16, R = 4.0)$. This analyzes the regularity of the texture. Moreover, it is invariant to gray scale shifts since the signs of the differences of $g_p - g_c$ are only considered.

3.2.3 Paper and ink transition descriptor

We also introduced the characterization of textual areas by measuring its average paper / ink transition. The descriptor we designed is inspired from the works of Journet et al.[10] Considering a sliding window of size 11×11 (the size was chosen empirically), we compute the average, over the window, of the line accumulations of horizontal differences between two adjacent pixels:

$$f_{11}(i,j) = Avg_{j \in J'}(\sum_{i' \in I'} |p_{i,j} - p_{i+1,j}|) \tag{7}$$

with I' and J' the size of the window and $p_{i,j}$ the gray level of the pixel of coordinate (i, j)

3.2.4 Mean intensity of pixels

Finally we present here the last feature that we have used in the signature of the texture. The pixel intensity is an important feature to separate background pixels from ink pixels. In fact, this characteristic is very reliable to separate the pixels that exist in the spaces between the words from ink pixels. The intensities of the pixels are very sensitive to image noise. To avoid this problem we used the average intensity of pixels instead of pixel intensity in the texture signature. The average intensities of pixels are measured on a window size of 5×5.

$$f_{12}(i,j) = Avg_{j \in J'}(\sum_{i' \in I'} p_{i,j}) \tag{8}$$

with I' and J' the size of the window and $p_{i,j}$ gray level of the pixel of coordinate (i, j)

3.3 Detection of missed text components

As mention in section 3.1, the classification of the pixels located in the background is performed by 4 SVMs classifiers to identify text, graphic, inter-word and background areas. After classification step, the inter-word

spaces areas are then merged with background areas in one single class corresponding to background areas. Figure 3 shows the results of this classification step at pixel level. The red pixels are the pixels classified as text, the green pixels are pixels classified as graphics and blue pixels are background pixels. Black pixels are the pixels that belong to the labeled mask provided by the OCR.

Our quality verification approach should provide an estimation of the number of missed text components at word level in order to be suitable with the BnF's process of OCR quality control. In most cases, the missed text components detected by the classifiers represent one or two characters. This is why it is needed to merge the neighbouring detected text components to form words or pieces of word. This is performed using a distance transform and a dilatation process. The distance transform[20] measures the distance that separates the background pixels from text components detected in the image (cf. figure 5 (b)). We then group all the pixels surrounding a text component with this component, if the distance is less than δ ($\delta = 3$) pixels to the component. The value of δ was chosen because it corresponds in most cases to the distance that separates two characters in a same word. This way, we obtain the envelopes of missed text components (called here Detected Missed Text Components $DMTC$). The proposed approach tends to make some detection errors on the examples when there are transparency defects or binarization defects (cf. figure 3 (II)). To eliminate noisy elements, we remove those that are smaller than the smaller text element detected by OCR in ALTO file. The number of detected missed components denoted by $number(DMTC)$ is the number of elements remaining after noise removal operation. Figure 5 (c) shows the results of this detection and localization step.

4. EVALUATION OF OUR ALGORITHM

4.1 Quantitative evaluation

The proposed method has been evaluated on a set of 256 images taken randomly from Gallica's dataset. This set of selected images cover five centuries of printing to maximize the variety of fonts and physical structures. The images are either binary or gray levels images. Their resolution is 300dpi. The ground truth of our dataset was produced manually during a verification stage of the automatic OCR outputs provided in ALTO files, using a BnF tool that allows to correct the segmentation and the word transcription. The ground truth that we have is also in ALTO format.

The visual results (cf. figure 3) show that our system is able to identify correctly the missed text components as well as missed graphic elements[¶]. In some cases, some graphical components can be identified as textual elements when they have similar characteristics, such as in the case of separators between the textual elements (cf. figure 3(III)). But according to the visual results, our approach tends to make few over detection mistakes.

To evaluate quantitatively our approach, we first have to compute the True Missed Text Components ($TMTC$). This is done thanks to a matching process at image level between the segmentation performed by the OCR (in the ALTO output file) and the ground truth (also in ALTO). The matching is done according to this rule for each text component in the ground truth: if more than 80% of the surface of a text component in the ground truth (GT_{text}) is covered by text components in the OCR output, we consider that the text component is well detected by the OCR. Otherwise, the difference of the surface between each (GT_{text}) and the OCR text components output becomes a $TMTC$ which is included in the image of the true missed text components (cf. figure 5 (a)). The number of textual components in this images is called: $number(TMTC)$.

The detection rate of our approach is obtained using a matching process between the true missed text components previously obtained and the image of envelopes of the missed text components detected by our approach ($TMTC \cap DMTC$). This process is rather different than classical segmentation algorithms since our aim is not the same: if more than 90%[‖] of a $TMTC$ is covered by a $DMTC$, we consider that the corresponding $TMTC$ is well detected. This way, our approach is able to detect 1886 missed text components (word or part of words) over the 2242 missed text components in the ground truth, which corresponds to a recall rate of 84.15% (cf. equation 9). We have also evaluated the false alarms (wrong detections) generated by our approach. In this

[¶]One should remember that the aim of our approach is to detect the missed text components in the OCR outputs and not to operate an accurate segmentation.

[‖]This threshold is rather strict but avoids to consider partial detection of missed elements as true positives.

Figure 4. Detected missed components (Defects images)

case, in compliance with the visual results of the proposed approach, we can estimate that our approach has a good precision rate: 94.73% (cf. equation 10).

$$Recall = \frac{number \quad (TMTC \quad \cap \quad DMTC)}{number \quad (TMTC)} \tag{9}$$

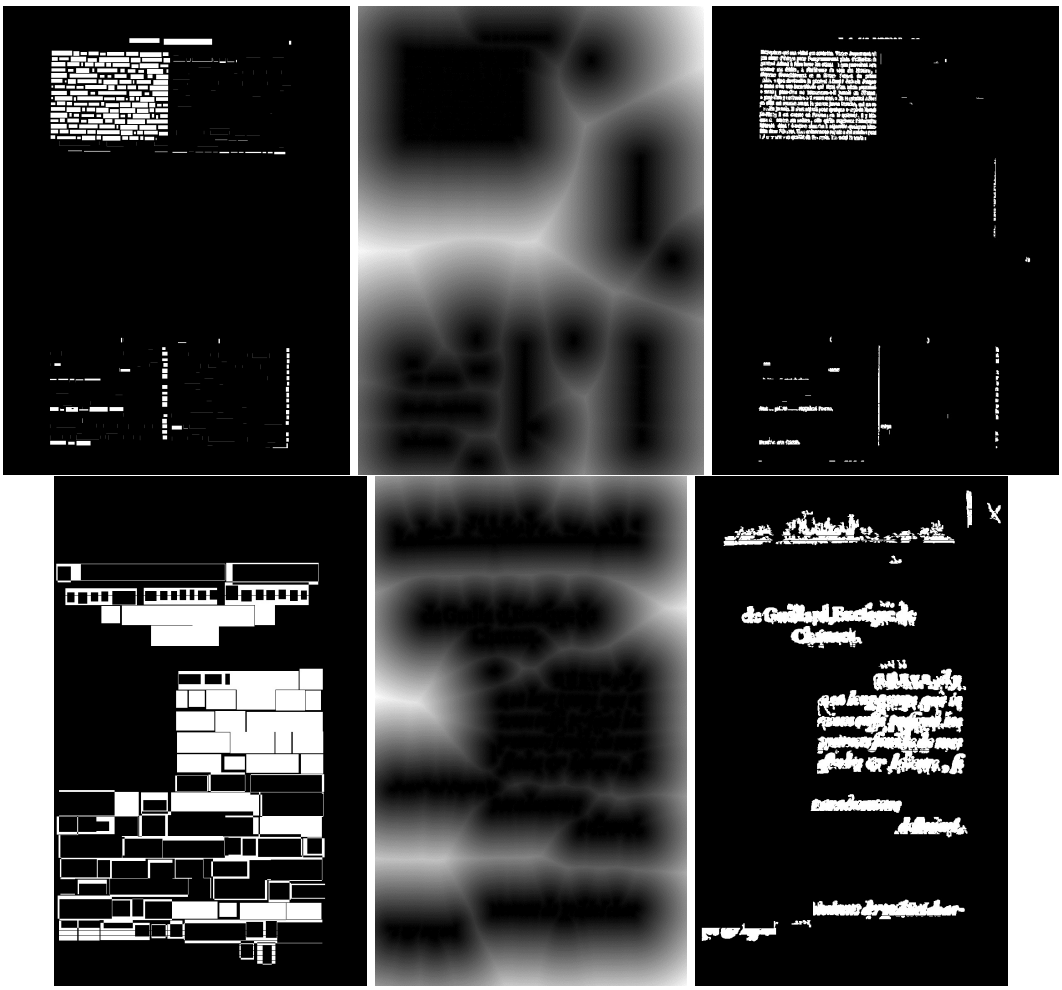
$$Precision = \frac{number \quad (TMTC \quad \cap \quad DMTC)}{number \quad (DMTC)} \tag{10}$$

4.2 Potential use of the approach

Considering the good results of the approach, the system will be integrated into the OCR quality workflow at the BnF as an assistant tool for the OCR controlers. At present, the system is able to provide a visual feedback (cf. figure 4) to the controlers as well as the number of missed textual components (DTMC). The visual result include the original page image, the OCR elements (black boxes) and the missed elements detected by our approach (The green areas refer to the missed graphic elements and the red areas refer to the missed textual elements). With this feedback, the user will be able to build a quality report and to justify a potential rejection of the OCR outputs. Furthermore, the integration of our approach into the OCR quality workflow will give us the possibility to estimate the performance of the approach on a large dataset derived from the BnF workflow.

5. CONCLUSION AND PERSPECTIVES

This paper presents an original approach to control OCR output and provide quality assessment of digitization process by detecting missed text components in OCR outputs without using ground truth. Usually, OCR systems provide outputs with a good confidence rate considering the detected elements (text or graphics). Consequently,



(a) (b) (c)

Figure 5. The needed information to conduct the evaluation of our approach:(a) The ground truth of the missed elements, (b) the images of distance transform of detected components. (c) The images of detected components.

these output files can be used to learn the specific properties of the components already detected (text and graphics). Moreover, they can be used to reduce the search space to detect missed text components, most of the time located outside the previously detected areas. This detection methodology based on page properties make our approach more robust to the diversity of styles and fonts that can be encountered in a wide range of documents, which has been verified considering the BnF collections.

The proposed approach uses 12 descriptors based on the Radon Transform, Local Binary Patterns and textual frequency descriptors to characterize pixels inside an image. Then, a SVM is used to classify the background pixels of the current image (those that were not previously classified by the OCR as text or graphic). Based on the classification results missed text components are finally localized.

Experimental results were performed on a heterogeneous set of images including old document images and new document images. They have shown the robustness of the proposed approach which has a good recall rate: 84.15% of missed text components are detected. Moreover, it does not cause many false alarms since it has a good precision rate of 94.73%.

Nevertheless, our method has also some drawbacks. The main one is that we assume the areas already detected by the OCR are without errors. But, sometimes, errors can occur and then the classifier can learn wrong characteristics. To limit these defects we think to work not only at page level but also at book level or at least considering groups of images. Another thing we would like to investigate is to validate our experimental results by an empirical study to evaluate how our approach can be useful for people in charge of the control at the BnF.

REFERENCES

[1] [*ALTO - Analyzed Layout and Text Object, as of version 2.0 maintained*], Library of Congress.

[2] Marosi, I., "Industrial OCR approaches: architecture, algorithms, and adaptation techniques," in [*Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*], *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series* **6500** (Jan. 2007).

[3] Rice, S., Nagy, G., and Nartker, T., [*Optical Character Recognition: An Illustrated Guide to the Frontier*], The Kluwer International Series in Engineering and Computer Science, Kluwer Academic Publishers (1999).

[4] Chen, D., "Text detection and recognition in images and video frames," *Pattern Recognition* **37**, 595–608 (Mar. 2004).

[5] Zhong, Y., "Locating text in complex color images," *Pattern Recognition* **28**, 1523–1535 (Oct. 1995).

[6] Smith, M. A. and Kanade, T., "Video skimming for quick browsing based on audio and image characterization," tech. rep., Computer Science Department, Pittsburgh, PA (July 1995).

[7] Sobottka, K., Bunke, H., and Kronenberg, H., "Identification of text on colored book and journal covers," in [*In Proceedings of the 5. Int. Conference on Document Analysis and Recognition*], 57–63 (1999).

[8] Garcia, C. and Apostolidis, X., "Text detection and segmentation in complex color images," in [*Proceedings of the Acoustics, Speech, and Signal Processing, 2000. on IEEE International Conference - Volume 04*], *ICASSP '00*, 2326–2329, IEEE Computer Society, Washington, DC, USA (2000).

[9] Wu, V., Manmatha, R., and Riseman, E. M., "Finding text in images," in [*ACM DL*], 3–12 (1997).

[10] Journet, N., Ramel, J.-Y., Mullot, R., and Eglin, V., "Document image characterization using a multiresolution analysis of the texture: application to old documents," *IJDAR* **11**, 9–18 (Sep 2008).

[11] Haralick, R. M., Shanmugam, K., and Dinstein, I., "Textural Features for Image Classification," *IEEE Transactions on Systems, Man, and Cybernetics* **3**, 610–621 (Nov. 1973).

[12] Uttama, S., Ogier, J.-M., and Loonis, P., "Top-down segmentation of ancient graphical drop caps: lettrines," in [*Proceedings of 6th IAPR International Workshop on Graphics Recognition*], 87–96 (Aug. 2005).

[13] BEYER, M. H., "Glcm texture : A tutorial. technical report," tech. rep. (2000).

[14] Caron, Y., Charpentier, H., Makris, P., and Vincent, N., "Power law dependencies to detect regions of interest," in [*DGCI*], 495–503 (2003).

[15] Raju, S. S., Pati, P. B., and Ramakrishnan, A. G., "Text localization and extraction from complex color images," in [*ISVC*], 486–493 (2005).

[16] Tucceryan, M., "Moment-based texture segmentation," *Pattern Recogn. Lett.* **15**, 659–668 (July 1994).

[17] Theodoridis, S. and Koutroumbas, K., [*Pattern Recognition*], Academic Press, Orlando, FL, USA (2008).
[18] Deans, S. R., [*The Radon transform and some of its applications*], A Wiley-Interscience Publication, New York, USA (1983).
[19] Ojala, T., Pietikäinen, M., and Harwoodl, D., "A comparative study of texture measures with classification based on featured distributions," *Pattern Recognition* **29**, 51–59 (Sep 1996).
[20] Bailey, D. G., "An efficient euclidean distance transform," *International Workshop on Combinatorial Image Analysis* **7**, 394–408 (Nov 2004).