Document recto-verso registration using a dynamic time warping algorithm.

Vincent Rabeux, Nicholas Journet, Jean-Philippe Domenger

Document recto-verso registration using a dynamic time warping algorithm.

Rabeux Vincent
University of Bordeaux
LaBRi
Bordeaux
rabeux@labri.fr
shema

Journet Nicholas
University of Bordeaux
LaBRi
Bordeaux
journet@labri.fr

Domenger Jean Philippe
University of Bordeaux
LaBRi
Bordeaux
domenger@labri.fr

***Abstract*—Recto verso registration is an important step allowing detection of missing digitized pages, or location of the bleed-through defect over a page. An efficient way to restore or evaluate the bleed-through of a digitized document consists in analyzing at the same time both the recto side and the verso side. This method requires the two images to be aligned, registered.**

Without particular knowledge about document, recto verso registration is complex. Indeed, the only information that we can use to register the two is the bleed-through. Recto verso registration is complex because the recto's bleed-through is a highly degraded version of verso's ink pixels. Therefore, in this particular context, usual image comparison methods [1] are not very relevant.

Nevertheless, document recto verso registration algorithms has been proposed [2], [3] [4], but these methods have important time computation costs, are noise sensitive and even fail in some cases where bleed-through is too light. The previous techniques are based on a pixel to pixel approach where the bleed-through is considered to be just a set of grey pixels.

In this article, we consider the structure of the ink pixels on the verso page. The recto verso registration method presented here is based on the fact that bleed-through has the same structure that the ink on the verso side. The method registers the recto's bleed-through layout and the verso's ink layout, in two main steps, first a de-skewing algorithm is applied to both pages then, horizontal and vertical profiles are extracted and aligned with a dynamic time warping. The time complexity of our method is linear according to the image size. Moreover, experiments detailed at the end show the accuracy of our method.

I. Introduction

Document recto verso registration is an important step allowing detection of missing digitized pages, or location of the bleed-through defect over a page. The bleed-through defect is a well known type of degradation. It appears on mainly very old documents and is due to an important quantity of verso's ink that can be seen from the recto side [5]. This defect is of great interest. Indeed, documents suffering from bleed-through have their readability greatly decreased. Moreover, it results in a high OCR error rate [6]. For these reasons, research has been done to measure this defect in order to predict OCR error rates [7], and other works are able to restore degraded documents [8], [2], [9]. The main issue with bleed-through is that it cannot be identified by a global thresholding method because gray levels are often too similar to the ink. Nevertheless, there exists two kinds of restoration methods, those that use both side of the document's page (the recto and the verso) [8], [2] and the ones who do not [9]. Latters restore not only bleed-through but also any present noises. Bleed-through identification techniques that use the verso side of the page, select pixels on the recto corresponding to ink pixels on the verso. But, since both sides of the page are scanned separately, the verso side may be shifted or rotated of a few millimeters from the recto side. Since the optical parameters and configuration of the scanner is fixed, scale transformations are not applied. The registration process aims to find a transformation that will align the recto and the verso.

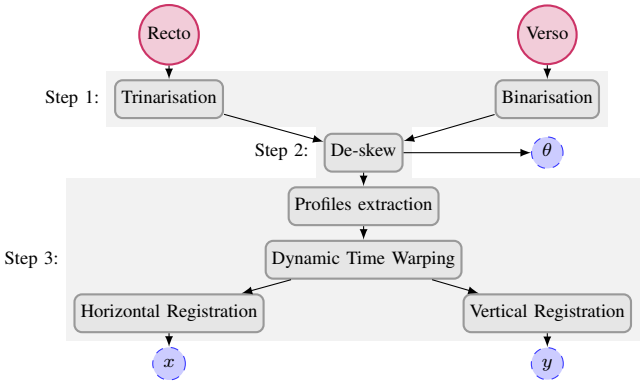

Figure 1. The overall registration method : dashed circles are the transformation parameters, θ for the rotation, x for the horizontal shift and y for the vertical shift .

Image registration methods aim at bringing two or more dataset in the same coordinate system [1]. Most of these methods are not suitable for the registration of a recto-verso pair since recto and verso have different image intensity and topography. The only relevant information that can be used in a recto verso registration is the bleed-through, but this information is a highly degraded version of verso's ink. Nevertheless, several recto-verso registration are proposed. In [2], [3] a parameter optimization method aims to find the appropriate transformation matrix that minimizes the

difference between the recto and the horizontal flipped verso. The difference between the two images is calculated by a cost function based of the grayscale value of each pixels. This registration method has a very high computation time. The second method [10] aims to lower this computation time by using a Fourrier-Mellin transformation. In [8], both methods are compared, and the results shows a slight advantage to the first one. Both methods do not provide a confidence measure that could be used to know if the registration has failed or not.

The registration method proposed in this article is based on the fact that bleed-through has the same structure than the ink on the verso side. Both layouts have the same lines, figures, etc. Our method is divided in several steps (see figure 1). We first roughly *trinarize* the recto, leading to three classes of pixels : the ink, the grey (containing bleed-through and noises), and the background. This first step extracts the bleed-through in order to register both images in the following steps. The second step estimates the skew angle by de-skewing the bleed-through on the recto side. In the third step, vertical profiles on both the bleed-through (recto side) and the ink (verso side) are extracted in order to align bleed-through and ink lines. At last, the same technic is used to extract and align horizontal profiles corresponding to text columns. Both horizontal and vertical profiles are registered using a **D**ynamic **T**ime **W**arping. Our method aims at registering the recto and verso image of a page, with an equivalent accuracy, and shorter computation times comparing to the state of the art. Moreover, the bleed-through extraction step allows us to propose a measure that can be used to know if the registration can be performed with a sufficient accuracy.

This article is organized according to the process steps of our method (see figure 1). First we will roughly identify the bleed-through pixels, then estimate the skew angle (θ) and the horizontal (x) and vertical (y) shifts. In the last section, experiments are made to statistically analyze the performance of our method.

II. BLEED-THROUGH AND NOISE PIXELS IDENTIFICATION

Our method is based on a layout analysis of both the bleed-through on the recto side and the ink on the verso side (which is horizontally flipped). The ink pixels on the verso side are identified by a global binarization method. We used Otsu's binarization method, but other binarization methods can be used. To extract the bleed-through, we need to *trinarize* the recto side of the document. The trinarization method is based on a *k-means* clustering which allows to separate the pixels into three clusters in which each pixel belongs to the cluster with the nearest mean. Other trinarization method can also be applied at this step. The global grayscale histogram has indeed three classes of pixels (I and B are two thresholds):

- (Set 1) Ink pixels : from 0 to I
- (Set 2) Grey Pixels : from I to B
- (Set 3) Background pixels : from B to 255

An example of this *trinarisation* method can be seen on figure 2 where bleed-through pixels are in the second class.

As shown on figure 2, the *trinarisation* method does not need to extract the bleed-through perfectly : spots and light grey ink pixels are also included in the bleed-through set. What matters at this step is that we identify enough bleed-through to extract informations about lines and columns of bleed-through.

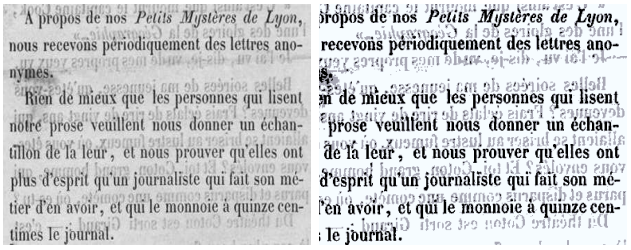

Figure 2. Left side : original document images. Right side: result of the trinarization method : pixels in set 1 are colored in black, pixels of set 2 in grey and pixels of set 3 in white.

III. RECTO'S AND VERSO'S SKEW ESTIMATION

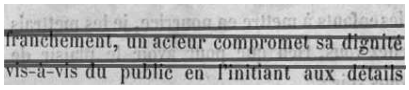

Figure 3. Two lines of ink and two lines of bleed-through. A grey line was traced on the bottom of bleed-through lines and on the top of the ink lines in order to visualize the different orientations (skew angle) between the latters.

The second step of our methods estimates de-skewing angles to find the registration rotation parameter. Several methods for de-skewing document images are proposed [11]. These methods estimate the skew angle of the ink and are known to have good results and be very accurate. This is well-suited for the estimation of the verso de-skew angle (θ_{verso}), but de-skewing the document recto side with the ink pixels is not sufficient. Indeed, bleed-through lines and ink lines orientations can be different. The figure 3 shows two lines of text and two lines of bleed-through, their orientations (skew angle) differs. To circumvent this problem, we use the de-skew technique on the extracted bleed-through pixels only. On the right side of figure 4 an image with only grey pixels (set 2) is shown.

The recto side and the verso side are de-skewed separately meaning that we have two estimated angles. One for the recto side named θ_{recto} and one for the verso side θ_{verso}. The rotation parameter θ of the registration method is $\theta_{recto} - \theta_{verso}$. For our experiments we used [11] to de-skew our images. An example of this step is shown in figure

4. Since the next steps of our method are computed on de-skewed images, the parameters θ_{recto} and θ_{verso} are kept.

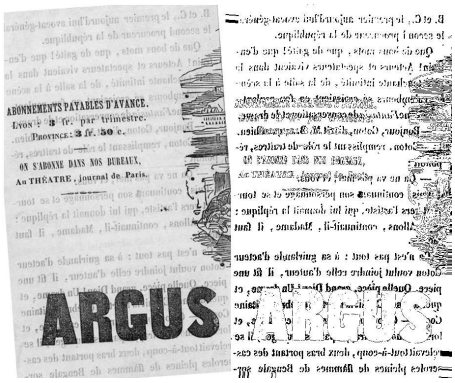

Figure 4. Results of the de-skew algorithm based on the extracted bleed-through and noise pixels : on the left the original document, on the right the de-skewed document (only the extracted bleed-through pixels are used).

IV. Lines and columns extraction for vertical and horizontal alignment.

This step of the method aims at finding the average vertical and horizontal shift between the recto side and the verso side. First the lines of bleed-through (recto side) and ink (verso side) are registered. Then the same procedure is applied to columns. Several lines extractions algorithms are proposed [12]. Our registration method is based on profiles method such as [13]. The vertical profile is obtained by summing pixels values along the horizontal axis for each y value. From the vertical profile, the gaps between the text lines in the vertical direction can be observed on figure 5 a. The vertical profile is not sensitive to writing characters fragmentation and therefore corresponds to our need since the bleed-through components are very fragmented. Vertical

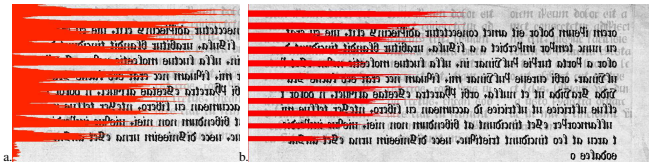

Figure 5. bleed-through lines extraction with vertical profile: a. the raw profile, b. the cleaned up profile: all un-relevant bins are removed.

profiles of bleed-through and ink are very similar. Even though the registration can be done on the two raw extracted profiles, the accuracy of the registration depends on the similarity of these profiles. This similarity depends on the quantity of noise and spots pixels since the bleed-through profile is computed from the extracted grey pixels. In order to have a more accurate registration of the two profiles, we need to clean up the recto's profile. To do so, every bins under a threshold T_0 is set to 0. The threshold T_0 aims to sort bins into two kinds : bins computed from lines or big spots ($profile[i] > T_0$) and bins computed form only noises ($profile[i] < T_0$). This thresholding process removes all bins resulting from only noises pixels. The profiles alignment process relies on white spaces which makes it more robust to noise.

Our experiments shows that the best value for T_0 is the trimmed mean. The trimmed mean is calculated by discarding 5% of the lowest and the highest scores and then computing the mean of the remaining scores. This trimmed means allows the threshold T_0 to be more accurate since bins that are outliers are not integrated in the mean computation. The result of this cleanup process is shown on figure 5 b.

Now that both profiles are similar enough, we need to find an optimal match between the two given histograms. This can be done using the well know **D**ynamic **T**ime **W**arping [14] algorithm. The DTW measures similarity between two sequences which may vary in time or speed, and can be back tracked in order to find the best alignment between the two sequences (ie: between the two profiles). The distance function given to the DTW has to evaluate the similarity of two bins in a profile. The distance between two bins is equal to the arithmetic difference of the two bins values :

$$d(i,j) = |R[i] - V[j]|$$

where R is the bleed-through and noise profile, V is the verso's ink profile, $i \in \{0..width\}$ and $j \in \{0..height\}$.

The DTW matrix backtracking results in a non-linear alignment in the vertical dimension of the two profiles. Bins can be $inserted$, $deleted$ or $matched$, (see explanation in figure 6) causing the shift value to vary along the profiles. This does not correspond to our needs since images lines cannot be $inserted$ or $deleted$ like profiles bins. In order to circumvent this problem we decided to compute the global mean shift value of $matched$ bins only (the subset $\{(b,a'),(c,b'),(d,c'),(e,d')\}$ in figure 6). This mean shift value is the vertical parameter of the registration process ((y) in figure 1).

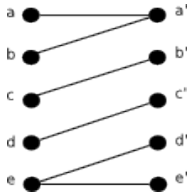

Figure 6. The DTW matching results in bins that are either $matched$: $\{(b,a'),(c,b'),(d,c'),(e,d')\}$, $inserted$: $\{(a,a')\}$ or $deleted$: $\{(e,e')\}$. Circles on the left and on the right correspond respectively to bins on the first profile and the second profile.

To find the last parameter of the registration (x), we apply the same technique on the vertical profiles of the images. Indeed, the same line alignment process can be applied

to columns represented by the vertical profiles. However, the accuracy of this last step depends on the number of columns in the document. The more there is the more the *DTW* will be accurate. But even if the vertical registration is more accurate (mean error of 1.81 pixels instead of 2.04 pixels for the horizontal registration) the mean difference is not relevant enough and in most cases one main column is enough to have an accurate registration.

V. Experiments

In order to compare the presented registration method to the state of the art, we made two kinds of tests. Experiments were made on several real old documents suffering from bleed-through, and to ensure the accuracy and computation times of our method tests were also made on a larger collection of auto generated documents.

There is, to our knowledge, no free corpus of document with their registration ground truths. In order to test our method, we manually registered a subset of documents suffering from bleed-through from the Collection BML[1].

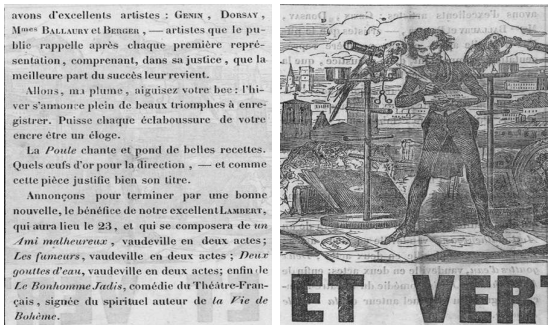

Figure 7. Example of our method and [2] on an extract of an old document suffering from bleed-through (recto on the right, verso on the left) : Manual registration : $x = 4$, $y = 1$, $\theta = 1$; Our method : $x = 5$, $y = 2$, $\theta = 1,3$, Method from [2] failed : $x = -34$, $y = -40$, $\theta = 1$

Both method presented similar accuracy results. But, on some documents, the parameter optimization technique [2] failed with errors reaching 50 pixels (lines height is 13 pixels) while our method was closer to the transformation parameters obtained by the manual registration (errors of 5 pixels). An example of such a document is shown in figure 7. The error obtained by [2] can be explained by the fact that since it tries to find the minimal difference between bleed-through and ink grey levels, the method fails when the bleed-through is not dark enough by aligning the recto's ink pixels with the verso's ink pixels. Our method is not affected by this problem since it is based on a bleed-through localization and not on its grayscale value.

In order to have a more reliable experiment, we also tested our method and [2] on generated documents. This allows us to create documents of different bleed-through levels, different quantity of texts, lines and columns (using the software presented in [15]). As a result we generated 48 rectos. To these rectos, we applied [9] to have 9 different images with different bleed-through levels (the bleed-through grayscale mean value varies from 255 to 159). This makes a total of 432 documents. This corpus of generated documents was also used in [7]. To each of these images, we applied a linear transformation with three random parameters for θ, x and y. x and y were chosen with maximum values corresponds to a realistic shift of 15% of our images.

As said the registration method proposed in [2] can only be applied to pages containing bleed-through. To have a fair comparison, we removed images that did not have bleed-through at all from the results tables. The results presented in table V confirm the previous tests on real documents. As a matter of fact, our method shows a better mean accuracy and its computation time is about 50 times faster than [2]. Also, [2] has a maximum error, on its horizontal error, of 39 pixels while our method has a maximum error of 11 pixels. These tests also shows that our method is very accurate on the vertical transformation parameter. Indeed the maximum error value is only 1 pixel, the mean error is 0.5 and the standard deviation is low (0.5).

Table I
Registration methods accuracy comparison. The registrations methods are implemented in C++ and tested on a computer with 8Go 1067 MHz DDR3 and an Intel Core i7 @ 2.8 Ghz

Registration Method	Skew Angle Error			
	Max	Min	Mean	Standard Deviation
Our Method	0.25	-0.03	0.15	0.06
Dubois's Method	18	0	7.19	4.45
	Horizontal Shift Error			
	Max	Min	Mean	Standard Deviation
Our Method	11	0	1.17	2.10
Dubois's Method	39	0	2.04	6.77
	Vertical Shift Error			
	Max	Min	Mean	Standard Deviation
Our Method	1	0	0.51	0.53
Dubois's Method	38	0	1.81	5.04
	Mean computation time			
Our Method	12s			
Dubois's Method	598s			

In oder to take into consideration documents on which the method failed, we normed the resulting DTW error to provide a confidence measure and re-named it m_{rv}. m_{rv} allows the detection of images with not enough bleed-through to be registered or any other documents that can not be registered with a sufficient accuracy by our measure.

We studied the measure m_{rv} and its relation to the accuracy of our method. Figure 8 a. shows this measure in relation to the vertical registration error. As shown, the measure can be used as a threshold to guaranty the accuracy of the registration. Registration errors happens when m_{rv} is higher than 0.19. All other registrations ($m_{rv} < 0.19$) are close to be pixel accurate.

These tests also shows some drawbacks of our methods.

[1] http://collections.bm-lyon.fr/

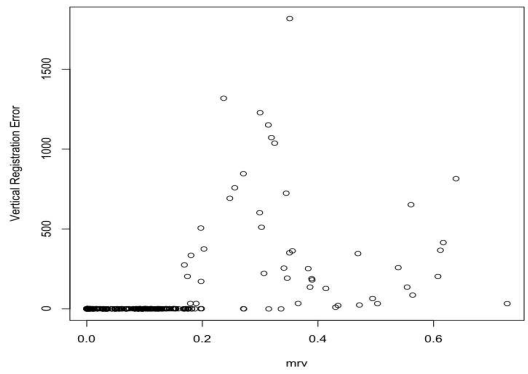

Figure 8. Relation between m_{rv} (x axis) and the vertical registration error (y axis). Accurate registrations (vertical error close to 0) have a m_{rv} value lower than 0.19. When m_{rv} is higher, we can not guaranty a pixel accurate registration.

First it does not work for white pages since latter can not be *trinarize*. Indeed, since this kind of images contains only 2 classes of pixels : the grey ones (containing the bleed-through) and the background pixel, the 3-means $trinarization$ method can not be used. To circumvent this problem we can detect if the image is a white page and then apply a 2-mean clustering method to extract the bleed-through. Second, some documents are sometimes sightly folded causing non-affine transformations. Our method and [2] do not take into considerations non-affine registrations. As a work around, we can use our method on images patches. This minimizes the remaining lag.

VI. Conclusion and research perspectives

In this article we present a new way to register a recto with its corresponding verso using a dynamic time warping algorithm to match horizontal and vertical profiles of both the bleed-through on the recto and the ink on the verso. This method shows that it is, in most cases, more accurate and faster than existing techniques. The computation times are divided by 50 compared to [2]. As a work perspective, we are working on applying this method to non-affine transformation (folded documents) by dividing the recto and the verso into patches. This technique has been experimented and seems promising, but we are having issues on finding the right patches (sizes and locations) to be registered.

Acknowledgment

We would like to thanks the on line document collection of Lyon (France), *Collection BML*[2], for letting us test our method on their documents. This work is done in the *Polinum*[3] project context, founded by Aquitaine region, the european community and Feder.

[2]http://collections.bm-lyon.fr/

[3]http://www.polinum.net/

References

[1] L. Brown, "A survey of image registration techniques," *ACM computing surveys (CSUR)*, Jan 1992. [Online]. Available: http://portal.acm.org/citation.cfm?id=146374

[2] E. Dubois and A. Pathak, "Reduction of bleed-through in scanned manuscript documents," in *IS AND TS PICS CONFERENCE*. SOCIETY FOR IMAGING SCIENCE & TECHNOLOGY, 2001, pp. 177–180.

[3] P. Dano, "Joint restoration and compression of document images with bleed-through distortion," pp. 1–72, May 2010.

[4] L. Hutchison..., "Fourier–mellin registration of line-delineated tabular document images," *International Journal on Document Analysis ...*, Jan 2006.

[5] G. Sharma, "Show-through cancellation in scans of duplex printed documents," *IEEE Transactions on Image Processing*, vol. 10, no. 5, pp. 736–754, 2001.

[6] L. Likforman-Sulem and J. Darbon..., "Enhancement of historical printed document images by combining total variation regularization and non-local means filtering," *Image and vision computing*, Jan 2011.

[7] V. Rabeux and N. Journet..., "Ancient documents bleed-through evaluation and its application for predicting ocr error rates (proceedings paper)," *spie.org*, Jan 2011.

[8] A. Tonazzini, G. Bianco, and E. Salerno, "Registration and enhancement of double-sided degraded manuscripts acquired in multispectral modality," in *Document Analysis and Recognition, 2009. ICDAR'09. 10th International Conference on*. IEEE, 2009, pp. 546–550.

[9] R. Moghaddam and M. Cheriet, "Low quality document image modeling and enhancement," *International Journal on Document Analysis and Recognition*, vol. 11, no. 4, pp. 183–201, 2009.

[10] S. Marchand and Desbarats..., "IBISA: Image-Based Identification / Search for Archaeology," vol. VAST-STAR, Short and Project Proceedings, St. Julians Malte, 09 2009, pp. 57–60.

[11] T. Breuel, "High performance document layout analysis," *Proceedings of the Symposium on Document ...*, Jan 2003.

[12] Likforman, "Text line segmentation of historical documents: a survey," pp. 1–25, Apr 2007.

[13] R. Manmatha and N. Srimal, "Scale space technique for word segmentation in handwritten manuscripts," *Proc. 2nd Int'l Conf. on Scale-Space Theories in Computer Vision*, pp. 22–33, 1999.

[14] R. Niels, "Dynamic time warping," *Artificial Intelligence*, Jan 2004.

[15] N. Journet, A. Vialard, and J. Domenger, "Analyse de fontes anciennes: de la génération de données synthétiques à la reconnaissance," *hal.archives-ouvertes.fr*, Jan 2010.