Proceedings of the 18th International Conference on Pattern Recognition (ICPR2006), Hong Kong, August 20-24, 2006, IEEE-CS Press, pp. 1062-1065.

Flexible Text Recovery from Degraded Typewritten Historical Documents[†]

A. Antonacopoulos and C. Casado Castilla

Pattern Recognition and Image Analysis (PRImA) Research Lab
School of Computing, Science and Engineering, University of Salford
Greater Manchester, M5 4WT, United Kingdom
http://www.primaresearch.org

Abstract

The conversion of large collections of historical typewritten documents into digital libraries and archives is met with significant challenges that standard recognition techniques cannot address. The condition and individual nature of characters in these degraded documents necessitate a departure from existing thresholding approaches. This paper presents a flexible approach designed to overcome the difficulties presented by such documents by flexibly analysing each individual character and cautiously repairing it. The main sources of OCR errors are successfully addressed and reliable corrective actions are taken.

1. Background

There is a significant and pressing need to convert collections of decaying historical documents into digital archives and libraries. One application of considerable historical as well as administrative interest —and the focus of this paper— is the analysis of large collections of administrative documents of the 20th century.

The fact that the majority of office documents and official correspondence of the 20th century are *typewritten* introduces certain unique challenges. First, each character is produced *independently* of the rest as a result of a person applying force to the corresponding typewriter key. In contrast to printed documents, each individual character within a document (or word even) may appear stronger of more faint than its neighbours (in direct relation to the amount of force used when pressing the corresponding key). The difference can be considerable in some cases.

Second, a typewritten document may not be produced in its entirety in an ordinary sequential manner. Instead the paper may be removed at some point and reinserted to make corrections and further additions. This can result in a non-uniform skew angle throughout the document and in non-uniform spacing between text lines or characters.

Finally, it was typical in the case of official documents to produce a carbon copy at the same time. Usually, the carbon copy was produced on a very thin paper (a.k.a. Japanese paper) which has prominent texture. Due to the mechanical nature of the typing process (the force from the typewriter key has to be transferred through the original paper and through the carbon sheet before a character is produced on the carbon copy) the characters of the carbon copy are usually blurred.

Historical typewritten documents are also affected by problems of ageing and repeated use, manifesting themselves as discolouration, disintegration of document parts, stains, punch holes, tears, rust from paperclips etc.

This paper focuses on the relatively difficult case of degraded carbon copy documents—see Fig. 3(a).

As perhaps expected, OCR systems fail to recognise the vast majority of the characters in this document class. The main reasons are, the presence of background texture, faint characters that appear broken and blurred characters that are filled-in and/or touching with others. These are acknowledged challenges for any OCR system.

From the above it is evident that the recovery of text from such degraded historical documents is a crucial stage in a digital archive conversion application.

It must be noted, for completeness, that the pre-OCR enhancement of degraded typewritten characters has been the subject of earlier research [1]. That approach however, addresses artefacts that are of different nature than those found on degraded historical documents and deals only with bilevel images.

In the more general case of *thresholding* in historical documents, it is perhaps obvious that global methods are not appropriate due to the non-uniformity of the text and the background in the presence of artefacts as mentioned earlier. A widely cited locally adaptive method is that of Niblack [2] on which a number of variants have been proposed, mainly to deal with the noise resulting when the background is not uniform. Two representative Niblack-derived approaches are: the more established method of Sauvola *et al.* [3] and more recently the method proposed by Gatos *et al.* [4]. While these approaches offer significant advantages over Niblack for

[†] This work has been supported in part through the EU grant IST-2001-33441.

the documents considered, the resulting quality is still not sufficient for the most degraded documents.

The authors, with other colleagues in their laboratory, have implemented and experimented with various such thresholding techniques [5]. That work demonstrated that the indiscriminate application of *any* thresholding approach (global or local) does not yield as good results as when a method is applied only to the *segmented* text.

This paper presents a new flexible text recovery method for the degraded documents in question. The flexibility consists in two aspects. First, the method is segmentation-based. Given the individual nature and appearance of each typewritten character on a page, the proposed approach analyses and treats each character separately. This concept is a unique characteristic of this method. Second, a novel combination of thresholding methods is proposed to repair most of the structural problems that contribute to OCR errors.

The next section describes the proposed method in more detail, with each stage explained in a separate subsection. Section 3 presents representative experimental results and discusses the effectiveness of the approach.

2. The method

In response to the individuality of typewritten documents, to the artefacts of the carbon copies and to the degradation through use and storage, the proposed approach introduces flexibility at two levels. Each character is treated individually and the problems of varying character intensity and stroke strength are addressed by a novel combination thresholding approach.

First, the positions of individual characters are located in the image. This process actually solves the problem of touching characters (significant for OCR). The actual characters are then more precisely localised, excluding as much background as possible. Two different thresholding methods take place next, in parallel. One cautiously identifies three types of pixels: those that are definitely foreground or background as well as those that could be either. The other method is a more aggressive binarisation method [3]. The results from each method are combined to produce a ternary image of each character, which is then prepared and submitted to the OCR process. The whole approach is summarised in the diagram in Fig. 1.

2.1. Character position location

This first stage exploits the fixed-pitch characteristic of a typewriter font to locate the position of characters in the document page. While the expected dimensions of a typewritten character "box" may be (approximately) known, the location of the actual character positions is not straightforward (the typewriter grid is irregular). To compensate for such artefacts as uneven spacing and non-straight baselines (resulting often from removal and re-insertion of the paper in the typewriter) and the problems of historical documents, a flexible approach is followed.

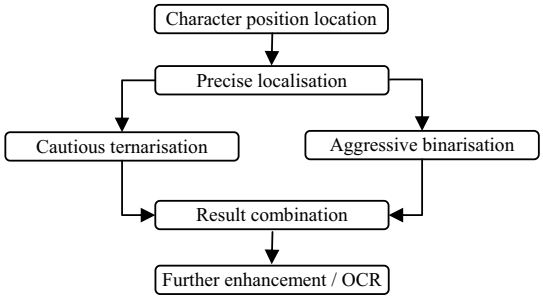

Fig 1. Outline diagram of the proposed approach.

Based on a vertical projection-profile analysis of the page (calculated on a suitably pre-processed greyscale version of the original – used only for this step), all possible grid positions where textlines could be expected to fall are examined and validated by assessing the minimum cost of each segmentation over the whole page (so that almost all valid cuts are made on profile valleys). The height of each textline is then adjusted accordingly.

To increase flexibility, character positions are located within each textline independently of the others (in practice, there is enough mis-alignment of the notional typewriter grid to necessitate this). The process is similar to the location of the textlines but the analysis takes place on the horizontal projection-profile of each textline.

It must be noted that only the positions actually containing characters are retained for further processing. An illustration of the located character positions can be seen on the left-hand side in Fig. 2.

2.2. Character localisation

The goal of this stage is to precisely localise the characters, excluding as much background as possible. This aids the correct estimation of the parameters for the combination thresholding approach.

The content area of each character position, as identified in the previous step, is first smoothed using a Gaussian filter (5x5 , $\sigma = 1.0$). The Gaussian filter has the advantage of smoothing the background and the strokes of the characters (attenuating noise) without impacting on the significant structural features of the characters.

The histogram of each box is then stretched using a sigmoid function and the resulting box contents become almost bilevel (the background is significantly lighter while the foreground is significantly darker). Such a hard-line treatment is necessary to ensure that no character parts are lost into the background.

The next step involves the tight fitting of a bounding box around each character by examining the horizontal

and vertical projection-profiles within each character position. The stretched image of each character position is then discarded and the localised character rectangles are noted within the original (smoothed) character boxes (see right-hand side of Fig. 2).

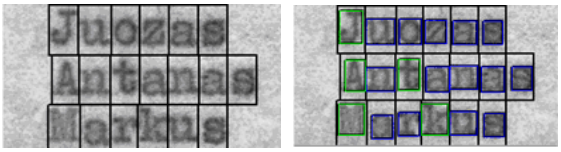

Fig 2. Position location (left) and precise localisation of characters (right).

2.3. Combination thresholding

As any thresholding approach is essentially an information reduction process it is crucial, especially for this type of documents, that neither any faint strokes are lost nor any character holes and cavities are filled in.

To maximise flexibility this stage applies two methods, in parallel, to each localised character bounding box and adaptively combines their results. The first objective is to repair broken characters by identifying those parts of the character strokes that can be confidently determined as foreground and to reliably indicate those parts of the strokes that could be merged to complete each character. The second objective is to avoid merging any background pixels with the character strokes as that will damage further the very dark and blurred characters.

2.3.1. Cautious ternarisation. This method was developed to classify pixels within the localised bounding box into three categories: pixels that can be confidently considered as part of the strokes, those that can be confidently considered as background and those that cannot be confidently classified outright.

After extensive experimentation it was observed that, due to the precise localisation, the pixels that can be confidently classified as foreground constitute the darkest 40% of the greylevel histogram of the localised character bounding box. Those pixels are turned black.

Similarly, it was determined that the pixels that can be confidently classified as background (even in the very noisy target documents) are those represented in the above histogram as being lighter than a specific threshold. That threshold is determined from the greylevel histogram of the pixels belonging only to the *outer* character position box (i.e., the pixels that lie in the complement of the localised bounding box, within the outer character position rectangle). The value of that threshold is calculated as the average of the most repeated value and the darkest one in that histogram. The pixels classified as background are turned to white.

The remaining pixels, corresponding to the mid-range values of the localised bounding box histogram, are those that cannot be reliably classified at this stage. Those pixels retain their greylevel value. An example of the result of the cautious ternarisation method can be seen in Fig. 4(b).

2.3.2. Aggresive binarisation. While existing methods do not perform sufficiently in noisy and degraded documents such as those considered here, they can still be useful in aggressively reconstructing broken character strokes (even though they may over-merge stroke parts – see next section). A good example of such a method is that of Sauvola and Pietikainen [3]. As mentioned earlier, that method is a "specialisation" of Niblack [2] and it is here further adapted by adding a preceding step of smoothing, much in the same spirit as the initial smoothing described in Section 2.2. The same Gaussian filter is used here but as the underlying Niblack method is very sensitive to noise, it is applied twice. Experimentation has shown that this is adequate to satisfy the objectives of this step.

The result is a binary image —see Fig. 3(c).

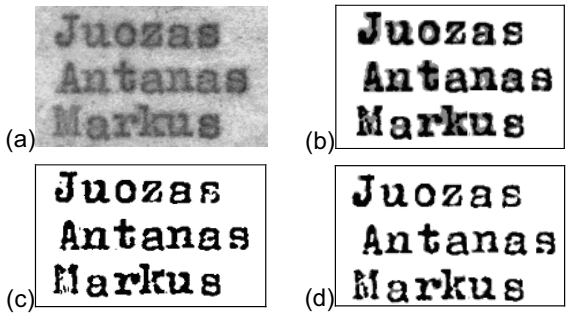

Fig 3. (a) Original image, (b) cautious ternarisation result, (c) aggressive binarisation result, and (d) combination result.

2.3.3. Result combination. Both the cautious ternarisation (CT) and the aggressive binarisation (AB) methods are designed and used for different purposes. The CT method provides reliable results but is cautious in not joining any strokes when there is doubt. On the other hand, the AB method aggressively joins broken strokes but also tends to join strokes that should not be connected (e.g. the free ends of "s" and "a" with the main body of the character). A number of deformations on the strokes are also observed in the results of the AB method.

The result-combination step aims to draw from the advantages of each method and to produce the final result by comparing corresponding pixels in each resulting localised character bounding box (all other pixels outside the localised boxes are turned white).

The rationale of the combination strategy is as follows. If the CT and AB pixels are both either black (foreground) or white (background), the final result is that value (effectively the value of the CT).

If the CT pixel is white and the AB is black, then the AB pixel is noise and the resulting pixel will be white (again, the confidence of the CT method is trusted).

If the CT pixel is grey (denoting uncertainty whether it is foreground or background) and the AB is black, then the result is a weighted mean of the CT and AB pixel values (to indicate a stronger possibility of being black). Being careful not to combine strokes that should not be joined together, the result is weighted towards the value of the CT pixel. The new pixel value is calculated as (2*CT + 100)/3. Note that the AB value is 'represented' by a value of 100 as it is, naturally, 0.

For completeness, in the cases that the AB pixel is white and the CT happens to be grey or black, the resulting pixel will be white. This decision is made as the AB method tends to err towards producing more black pixels (and therefore the fewer whites are trusted more).

An example result of the combination strategy is shown in Fig. 3(d).

2.4. Recognition

This stage further prepares the enhanced character images and sends the final result to be recognised. First, the already identified (Section 2.1) character positions are adjusted by suitably adding space (where necessary) between characters and between textlines.

Off-the-shelf OCR is used to recognise the recovered characters. In an attempt to obtain results with a standard widely available platform the authors have used the Microsoft Office Document Imaging (MODI) library that has been developed by ScanSoft and is free to download.

3. Results and conclusions

The approach presented in this paper has been designed to overcome the difficulties presented by the nature of degraded historical typewritten documents by flexibly analysing each individual character and cautiously repairing it (while pinpointing ambiguous areas).

A comparative illustration of results obtained by the proposed method and existing ones on a representative image part can be seen in Fig. 4. The improvement offered by this method is notable.

In quantitative terms, preliminary results demonstrate considerable improvement. On average 91.3% of the characters are correctly recognised in contrast to 77.2% achieved by OCR alone in the original image and 88.5% if the approach of Gatos *et al.* [4] is applied instead before OCR. It should be noted that the OCR system introduces its own additional errors (included in the above figures). and that the proposed method is also advantageous over previous ones as it has already segmented multi-column documents and therefore avoids reading order errors.

In qualitative terms, extensive experimentation shows that the most significant situations that cause OCR errors have been addressed, mainly the recovery of faint characters, the suppression of the textured/noisy background as well as making a reliable attempt at repairing broken character strokes (while indicating the potential areas of missing stroke segments). Furthermore, the proposed approach inherently solves the remaining problem of touching (merged) characters since it isolates each character position before treating its content. At the same time important features such as cavities and holes are preserved.

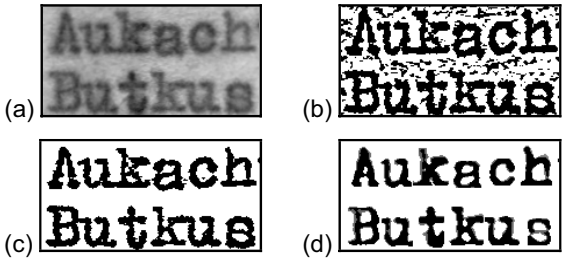

Fig 4. (a) Original image, (b) Sauvola and Pietikainen [3] result, (c) Gatos *et al.* [4] result, and (d) proposed method result.

References

[1] M. Cannon, J. Hochberg and P. Kelly. "QUARC: A Remarkably Effective Method for Increasing the OCR Accuracy of Degraded Typewritten Documents", *Proceedings of the 1999 Symposium on Document Image Understanding Technology (SDIUT'99)*, Annapolis, MD, May 1999, pp. 154-158.

[2] W. Niblack, *An Introduction to Digital Image Processing*. Prentice-Hall, London, 1986.

[3] J. Sauvola and M. Pietikainen, "Adaptive document image binarization", *Pattern Recognition*, Vol. 33, 2000, pp. 225-236.

[4] B. Gatos, I. Pratikakis, and S.J. Perantonis, "An Adaptive Binarization Technique for Low Quality Historical Documents", *Proceedings of the 6th IAPR Workshop on Document Analysis System (DAS 2004)*, Florence, Italy, September 8-10, 2004, Springer LNCS (3163), pp 102-113.

[5] A. Antonacopoulos and D. Karatzas, "Semantics-Based Content Extraction in Typewritten Historical Documents", *Proceedings of the 8th International Conference on Document Analysis and Recognition (ICDAR2005)*, Seoul, South Korea, August/September 2005, ISBN: 0-7695-2420-6, pp. 48–53.