International Conference on Advanced Computing Technologies and Applications (ICACTA-2015)

Automatic Removal of Handwritten Annotations from Between-Text-Lines and Inside-Text-Line Regions of a Printed Text Document

P. Nagabhushan, Rachida Hannane, Abdessamad Elboushaki, Mohammed Javed*

Department of Studies in Computer Science, University of Mysore, Mysore-570006, India

Abstract

Recovering the original printed text document from handwritten annotations, and making it machine readable is still one of the challenging problems in document image analysis, especially when the original document is unavailable. Therefore, our overall aim of this research is to detect and remove any handwritten annotations that may appear in any part of the document, without causing any loss of original printed information. In this paper, we propose two novel methods to remove handwritten annotations that are specifically located in between-text-lines and inside-text-line regions. To remove between-text-line annotations, a two stage algorithm is proposed, which detects the base line of the printed text lines using the analysis of connected components and removes the annotations with the help of statistically computed distance between the text line regions. On the other hand, to remove the inside-text-line annotations, a novel idea of distinguishing between handwritten annotations and machine printed text is proposed, which involves the extraction of three features for the connected components merged at word level from every detected printed text line. As a first distinguishing feature, we compute the density distribution using vertical projection profile; then in the subsequent step, we compute the number of large vertical edges and the major vertical edge as the second and third distinguishing features employing Prewitt edge detection technique. The proposed method is experimented with a dataset of 170 documents having complex handwritten annotations, which results in an overall accuracy of 93.49% in removing handwritten annotations and an accuracy of 96.22% in recovering the original printed text document.

* Corresponding author: Tel: +91-97-4116-1929
E-mail address: javedsolutions[at]gmail.com; mohammed.javed.2013[at]ieee.org

Keywords: Handwritten annotation removal; Marginal annotation removal; Between-text-line annotations; Inside-text-line annotation.
1. Introduction

Annotating a printed text document refers to the process or the act of adding annotations or writing critical commentary or explanatory notes into the machine printed text documents. Adding annotations at different regions of the printed text document is the habit of many people whenever they read any document. These annotations may be some important observations or corrections marked in their own style; this makes the problem of recovering the original text document very challenging, particularly when the original document is not available.

A reader can add different types of annotations in any part of the document; the type of annotations depend on the information being read by the reader. For example, the annotations can be lines underscoring or side scoring to highlight keywords, sentences or part of a paragraph and so on, or even enclosing such a part with a circular, elliptical or contour shapes holding a word or a sentence of interest, or it can be simply a question mark (when the reader does not understand the content), also it can be in the form of more frequent annotations that are handwritten comments (in which the related idea is extrapolated or the missed details are added), or it can be any other type of external remark that can be added anywhere in a document including: in marginal area of the document, in between-text-line regions, inside-text-line regions crossing over the printed text lines. Based on the regions where annotations appear, we classify them into the following categories: marginal annotations, between-text-line annotations, inside-text-line annotations, and overlapping annotations (see Fig. 1).

There are a few attempts reported in the literature related to handwritten annotations in printed text documents. However, we understood that the nature of the annotations considered in is predefined in structure and location. The method described in by Mori and Bunke, have achieved high extraction rates by limiting the colors or types of annotations. Although the correction is successful, their method has the limitation on colors of annotations as well as type of documents. Their method does not focus on the extraction of annotations but on the utilization of extracted annotations. Guo and Ma proposed a method that involves the separation of handwritten annotations from machine printed text within a document. Their algorithm is based on the theory of hidden Markov models to distinguish between machine-printed and handwritten materials, and the classification is performed at the word level. Handwritten annotations are not limited to marginal areas and also their approach can deal with document images having handwritten annotations overlaid on machine-printed text. However, their extracted annotations are limited to some predefined characters; their method cannot extract handwritten line drawings which is also a frequently used annotation. Y. Zheng et al. proposed a method to segment and identify handwritten text from machine printed text in a noisy document; their novelty is that they treat noise as a distinguished class and model noise based on selected features. Trained Fisher classifiers are used to identify machine printed text and handwritten text from noise. Lincoln Faria da Silva et al. proposed a method that involves the recognition of the handwritten text and machine printed text in a scanned document image. New features for classification of the handwritten text and machine printed text are proposed as well: Vertical Projection Variance, Major Horizontal Projection Difference, Pixels Distribution, Vertical Edges, and Major Vertical Edge. Unfortunately, their method can work only when the machine printed text and handwritten texts are separated. However, in our proposed method, we explore the feature of vertical Prewitt edge in order to remove the handwritten annotations inside the printed text line regions.

To our best knowledge, the proposed methods in the literature do not present a systematic approach for handling the different annotations that appear in printed text documents. Therefore, in order to systematically handle the different annotations coined in this paper, we propose a Handwritten Annotation Removal System (HARS) based on the different types of annotations that a document can have (see Fig. 2). However in this paper, we specifically focus on detecting and removing annotations located in between text lines and inside text line regions. Rest of the paper is organized as follows: In section 2, we detail the proposed methods to remove between-text-line and inside-text-line annotations. In section 3, the experimental results of our proposed methods are reported, and the last section 4 concludes the paper.

2. Proposed Methods

The overall proposed system (HARS) for detecting and removing handwritten annotations consists of five major stages (see Fig. 2), where the scanned annotated document is taken as an input and converted to a binary image; this
A binary image in the first stage is pre-processed to remove noise coming out of the document scanning. The approach used here is described in Peerawit and Kawtrakul. Later, a pre-processing to correct any skew present in the document image is applied. We use the method, which is provided by Cao and Li. Further, the second stage of removing marginal annotations (See Fig. 2) has been attempted by Abdessamad et al. They use the vertical and horizontal projection profiles in order to detect the marginal boundary and remove the annotations in one stretch. Because of annotations near the marginal area, the marginal boundary detected is approximate. Therefore, they use the concept of connected component analysis to bring back the printed text components cropped during the process of marginal annotation removal. In this research work, we propose two novel methods for removing between-text-line annotations and inside-text-line annotations. The proposed methods are presented in the following subsections.

2.1. Removal of Between-text-line Annotations

The handwritten annotations that are located between any two adjacent printed text line regions are called between-text-line annotations (see Fig. 3).

Adding a chord to $T$ creates a cycle in $T$, called a fundamental cycle of $G$ with respect to $T$. Each chord of $G$ creates a different cycle in $T$, and thus there exists a one-to-one correspondence between the chords of $G$ and the fundamental cycles of $G$ with respect to $T$. A cut-set of $G$ is a set of edges with the property that the removal of the edges in $C$ from $G$ disconnects $G$, but the removal of any proper subset of $C$ from $G$ does not disconnect $G$. Deleting a branch from a spanning tree $T$ of $G$.

In this section, the proposed model for removal of between-text-line annotations is developed (see Fig. 4) in which we present a two-stage algorithm for removing these annotations. They are described as follows:

2.1.1 Detection of the sure printed text lines

We use the connected component labeling algorithm to determine the average character size of the characters in the document, which is computed statistically. However, the descender-less characters in a printed text line have essentially same base regions, whereas the handwritten annotations located in between-text-lines region have different base regions. Each connected component is represented by the midpoint of its base region (see Fig. 5), and then the vertical projection of these midpoints is obtained (see Fig. 6).
Fig. 4: Proposed model for removal of between-text-line annotations

Each printed text line region contains high density of connected components producing a high peak at its base region. Therefore, the high peaks in the histogram (in Fig. 6) represent the sure printed text lines. To identify these peaks we compute the mean ($\mu_1$) of all peaks in the histogram (H1), then select all peaks above the mean ($\mu_1$), and label them as a sure printed text lines. If there are any two peaks which are close to each other by a distance less than character size, then the largest peak is considered as the baseline of the text line (see Fig. 7).

2.1.2 Detection of the probable or missed text lines

The reason that some small printed text lines are missed from the previous stage is that, their peaks are below the mean ($\mu_1$) obtained from histogram (H1) (see Fig. 6). To retrieve them, the procedure involves the removal of all higher peaks from the histogram (H1), followed by computation of mean ($\mu_2$) of all peaks from the remaining histogram (H2) (see Fig. 8), then for every peak $P$ above the mean ($\mu_2$) in the histogram (H2), Consider $P$ as base line of text line in the document if:

$$\text{distance}[P, R] > 2 \times \text{character size} \quad \forall R \in \text{PriBL}$$  \hspace{1cm} (1)
where PriBL is set of selected peaks in the first iteration.

**Abstract**

Tech savvy is becoming persuasive and pervasive.

Concept of OCR less auto-reading of the children’s handwritten text is a potential problem.

---

**Fig. 7:** Baseline detection of sure printed text lines

**Fig. 8:** Computation of the mean of peaks in histogram (H2)

---

Fig. 9 shows a sample document after the recovery of probable missed text lines.

After the detection of all the printed text lines in the annotated document, the next stage involves the removal of handwritten annotations between text lines; the method below demonstrates the process of removing of between-text-line annotations.

**Step-1:** Compute the average character size for every text line in the document.

**Step-2:** For every two consecutive text lines Ti and Ti+1 remove the annotations between region of Ti and region Ti+1 of with the following formula

\[
T_i + (0.5) * F_{Z_i} < An < T_{i+1} - (F_{Z_{i+1}} + 0.6) * F_{Z_{i+1}}
\]

where Ti: Text line of index i, FZ: Font size of the text line Ti, and An: the annotation.

**Abstract**

Tech savvy is becoming persuasive and pervasive.

Concept of OCR less auto-reading of the children’s handwritten text is a potential problem.

---

Fig. 9: Baseline detection for probable printed text lines

Fig. 10: Removal of Between-Text-Line Annotations

During this process, all annotations that appear in between-text-line regions are removed (see Fig. 10). The next stage in the proposed HARS (see Fig. 2) involves the removal of handwritten annotations from inside-text-line regions, which is discussed as follows.

### 2.2 Removal of Inside-text-line Annotations

The handwritten annotations that are located inside the rectangular area in which the entire printed text line fits into are called inside-text-line annotations. We distinguish here two different types of annotations. The first type is overlapping annotations: which overlap with the printed text lines. The other type of annotations is normal annotations located in some free area inside the text line region but not crossing the printed text (see Fig. 11).
In this section, a proposed model for removal of free area inside-text-line annotations is developed (see Fig. 12). The different stages involved in the proposed method are:

![Proposed model for removal of inside-text-line annotations](image)

**2.2.1 Merging of connected components at word level**

The merging algorithm of connected components is used to segment the printed text at word level (see Fig. 13) on which the next operations will be performed. The algorithm takes as input two neighboring connected components, and checks if they are in the same position and the difference between their heights is less than half character size, if so, then both of them are merged. This procedure will continue until no merging case remains and results in connected components of words from which the different features will be extracted.

**2.2.2 Computation of density distribution**

The vertical projection profile of the connected components at word level is obtained (see Fig. 14), and the profile is empirically divided in three regions (see Fig. 15) which are:

Region1: Height of $(0.6) \times \text{character size}$ in the top part of the bounding box, called Ascender region.
Region2: Height of $\text{character size}$ in the middle part of the bounding box, called Middle region.
Region3: Height of $(0.5) \times \text{character size}$ in the bottom part of the bounding box, called Descender region.

It is visible that the vertical projection profile is almost similar and consistent for all machine printed words (see Fig. 15). Moreover, in case of printed text, the highest density of black pixels occurs in the middle part of the projection profile; therefore the ascender region and descender region are known to have low pixel densities. However, the density distribution of the profile for handwritten annotations depends on the type of annotation, and it
changes from one annotation to another (see Fig. 16). Experimenting with 50 handwritten words, we observed that all three regions densities are all most similar. Based on this evidence, we conclude that the variance in densities of three regions for machine printed text is higher compared to the variance in case of handwritten annotation.

We compute the variance of three regions densities using the formula:

\[ \sigma_{cc}^2 = \frac{1}{a} \left( D_A^2 + D_M^2 + D_D^2 \right) - \mu_{cc}^2 \]

where \( D_A \): Ascender density, \( D_M \): Middle density, \( D_D \): Descender density, \( \mu_{cc} \): The mean of three regions densities

\[ \mu_{cc} = \frac{1}{a} \left( D_A + D_M + D_D \right) \]

and \( D \) = Number of black pixel in the region / region area

### 2.2.3 Vertical Edge Detection

In a printed text document where black pixels represent foreground (printed text) and white pixels represent the background of the document; the vertical edge is a set of connected pixels that are stretched in vertical direction. A spatial filter mask is based on the directional Prewitt filter (see Fig. 17). Fig. 18 shows the result of applying this mask on a handwritten and on a machine printed word. Vertical edge is also an important feature used here to distinguish the handwritten annotation from machine printed text. This feature also remains invariant to some overlapping annotations inside the connected components. We detect two important features related to vertical edges which are extracted from the connected components of words: number of large vertical edges and the major vertical edge.

\[
\begin{bmatrix}
1 & 0 & -1 \\
1 & 0 & -1 \\
1 & 0 & -1
\end{bmatrix}
\begin{bmatrix}
-1 & 0 & 1 \\
-1 & 0 & 1 \\
-1 & 0 & 1
\end{bmatrix}
\]

Fig. 17: Spatial filter masks
Extracting large vertical edges: Vertical edge which has the height more than $(2/3) \times \text{character size}$ is considered as large vertical edge and more than 65% of the English alphabets have more than one vertical edge. (see Fig. 19).

\[ \text{length}(LVE) \geq \frac{2}{3} \times \text{character size} \]  

where $LVE$ is the large vertical edge.

As observed in Fig. 20, number of large vertical edges in a printed word is greater than the number of large vertical edges in a handwritten word. The following steps show how to compute the large vertical edge parameter:

Step-1: Compute the number of the large vertical edge $LVE$.

Step-2: Compute the parameter $D_{LVE}$ and store it as a feature.

\[ D_{LVE} = \frac{\text{number of LVE}}{\text{length}(BB)} \]  

where BB: Bounding Box.

Extracting major vertical edge: Major vertical edge is the one with larger number of pixels in the component. It is also considered as a feature for distinguishing the handwritten annotations from printed words in the free area of the text line regions of a document. In this case we observe that, the difference between the longest edge and the height of its connected component in case of machine printed text is less compared to the case of handwritten annotation; this difference is called major vertical edge (see Fig. 21).

The following steps show how to compute the major vertical edge parameter:

Step-1: Compute the max or Height of the Major Vertical Edge.

Step-2: Compute the parameter $D_{MVE}$ and store it as feature:

\[ D_{MVE} = \frac{H_{MVE}}{\text{Height}(BB)} \]  

where, BB: Bounding Box.
2.2.4 Distinguishing between machine printed and handwritten word

The Table 1 below summarizes the features used to distinguish between machine printed text and handwritten annotation based on the features described above.

The classification rules are learnt in the training phase of 30 documents randomly selected from 170 dataset used here, the result obtained for classification are described in Fig. 22.

| Features            | Machine printed text | Handwritten annotation |
|---------------------|----------------------|------------------------|
| Variance of density distribution | High                  | Low                    |
| Number of large vertical edges        | High                  | Low                    |
| Major vertical edge             | High                  | Low                    |

Table 1: Feature wise comparison between machine printed text and handwritten annotation

Fig. 22: The procedure to remove handwritten annotations obtained experimentally from 30 training documents

3. Experimental Results

To our best knowledge, we could not find any standard dataset with complex handwritten annotations, which suits our research requirements. Therefore, we have collected 170 text documents from literary books and scientific articles, and given to heterogeneous group of students for marking random annotations of their choice. They generate the annotations in their own styles and using their own pens with different thickness. Dataset document images with annotations were obtained by using a scanner of 300 dpi. We compute the accuracy of our proposed methods in two ways.

3.1 Accuracy of Removing Handwritten Annotations

In order to compute the accuracy of removing between-text-line and inside-text-line annotations, we use the formula

$$\text{Accuracy}_{HA} = 1 - \frac{B}{A}$$

where $A$ is the amount of expected annotations to be removed, and $B$ is the amount of removed annotations. The experimental results show an average removal accuracy of 93.5%. Most of the documents in our dataset have accuracy above 93.49% (see Table 2).

Table 2: Result of removing the annotations

| Accuracy     | Below 93.49% | Above 93.5% |
|--------------|--------------|-------------|
| No. of Documents (%) | 38           | 62          |

Table 3: Result of recovering original printed documents

| Accuracy     | Below 96.22% | Above 96.23% |
|--------------|--------------|--------------|
| No. of Documents (%) | 20           | 80           |
3.2 Accuracy of Recovering the Original Document

To compute the accuracy of recovering the original document by removing between-text-line and inside-text-line annotations, we use the formula

\[ \text{Accuracy}_{OD} = 1 - \frac{\beta - \alpha}{\alpha} \]  

(8)

where \( \alpha \) is the expected cleaned document without handwritten annotations, and \( \beta \) is recovered document after the removal of annotations. Through our experiments, the average accuracy of getting the original document is 96.22% in which most of the documents of our dataset give accuracy above the average (see Table 3).

We also compute the correlation coefficient between the original document and the processed document by using the formula

\[ r = \frac{1}{n-1} \sum_{i=1}^{n} \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right) \]  

(9)

where the number of pixels is given by \( n \), the pixels of the original document are represented by \( x_i \) and the pixels of the processed document are represented by \( y_i \). The mean of the \( x_i, y_i \) is denoted by \( \bar{x} \) and \( \bar{y} \). The standard deviation of the \( x_i, y_i \) is denoted by \( s_x, s_y \) respectively. However, the overall correlation between the original document and the processed document for 170 images of our database gives an average correlation of 0.9625.

Therefore, it is clear that most of the documents that represent 83% from the dataset have a correlation above 0.9625. Further, the individual accuracies of three components shown in HARS (see Fig. 2) is tabulated in Table 4.

| Table 4: Cumulative accuracy of three components of HARS |
|---------------------------------|-----------------|-----------------|
| Accumulative Accuracy (%)       | Removing Handwritten Annotations | Recovering Original Document |
|--------------------------------|----------------------------------|------------------------------|
| Marginal                        | 91.11                           | 98.34                        |
| Between-Text-Line regions       | 94.14                           | 95.90                        |
| Inside-Text-Line regions        | 91.45                           | 95.60                        |

Finally with the experiments carried out, we observe an overall error of 6.51%, which is due to the presence of overlapping annotations within the printed text line regions. Overlapping annotations have not be tackled in this paper, which could be anticipated as an extension work to the proposed research.

4. Conclusion

In this research paper, we have described and discussed the methods to remove both of handwritten annotations located in between-text-lines and inside-text-line regions of a printed text document. The experimental results show that our methods produce consistent and adequate results under reasonable variation of type of annotations. We demonstrate the proposed idea on a dataset of 170 documents which gives an average of 0.9625 in correlation coefficient between the original document and the processed document, and an overall accuracy of 93.49% for removal of handwritten annotations, and 96.22% in case of recovering the original printed text document. However, as the handwritten annotations overlapping with printed text are not yet processed, we got only 6.51% as reduction in the accuracy of removing of the annotations. Therefore, our proposed algorithm removes any annotations inside text line regions except those annotations, which are over the printed text.

References
1. Cao Y., Li H.. Skew detection and correction in document images based on straight-line fitting. P R Letters, 24(12):1871–1879, 2003.
2. Da Silva L. F., Conci A., Sanchez A.. Automatic discrimination between printed and handwritten text in documents. Published in Computer Graphics and Image Processing (SIBGRAPI), XXII Brazilian Symposium, pages 261–267, 2009.
3. Elboushaki A., Hannane R., Nagabhushan P., Javed M. Automatic removal of handwritten annotations in marginal area of printed text document. International Conference on ERCICA, Published by Elsevier, 2014.
4. Guo J. K., Ma M. Y. Separating handwritten material from machine printed text using hidden markov models. ICDAR, pages 436–443, 2001.
5. Mori D., Bunke H. Automatic interpretation and execution of manual corrections on text documents. In P. S. P. W. H. Bunke, Editor, Handbook of Character Recognition and Document Image Analysis, World Scientific, Singapore, pages 679–702, 1997.
6. Peerawit W., Kawkwalek A. Marginal noise removal from document images using edge density. In Proc. Fourth Information and Computer Eng. Postgraduate Workshop, 2004.
7. Zheng Y., Li H., Doermann D. The segmentation and identification of handwriting in noisy document images. In Lecture Notes in Computer Science, 5th International Workshop DAS 2002 Princeton, NJ, USA., pages 95–105, 2002.