

Independent Component Analysis Segmentation Algorithm

Yan Chen and Graham Leedham
School of Computer Engineering
Nanyang Technological University
Singapore 639798
asgleedham@ntu.edu.sg

Abstract

In this paper we propose and investigate a new segmentation algorithm called the ICA (Independent Component Analysis) Segmentation Algorithm and compare it against other existing overlapping strokes segmentation algorithms. The ICA Segmentation algorithm converts the original touching or overlapping word components into a blind source matrix and then calculates the weighted value matrix before the values are re-evaluated using a fast ICA model. The readjusted weighted value matrix is applied to the blind source matrix to separate the word components.

The algorithm has been evaluated on 30 'overlapped' document images from the CEDAR letter database and another 30 degraded historical document images, which containing many different kinds of overlapping and touching words in adjacent lines. Quantitative analysis of the results by measuring text recall, and qualitative assessment of processed document image quality is reported. The ICA Segmentation Algorithm is demonstrated to be effective at resolving the problem in varying forms of overlapping or touching text lines.

1. Introduction

Many document images contain florid handwriting, which frequently exhibits extravagant loops in ascenders, descenders and upper case letters. These often result in touching or overlapping of words on adjacent lines. Separating the lines and words is a non-trivial task and the segmentation of touching or overlapping words on adjacent lines is an important stage in the processing of handwritten documents.

Numerous segmentation techniques [1] ~ [6] have previously been proposed for document images. Sabourin & Plamondon 1988 [1] proposed a technique

based on the extraction of textured regions characterized for handwritten signature image. Congedo et al 1995 [2] proposed a multiple segmentation algorithms for handwritten numeric strings. Feldbach & Tonnie 2003 [3] investigate a technique for historical document. Cheung et al [4] proposed a model-based segmentation technique for handwritten city name images from CEDAR address database. Sadri et al [5] investigate a foreground and background features based method for unconstrained handwritten numeral string. Daekeun et al [6] proposed a neural network based method for handwritten numeral strings.

Whilst these separation techniques have proven effective at segmenting words correctly if the handwritten text lines are not overlapping or touching, none has been shown able to produce consistently good results on the wide range of document images containing touching or overlapping handwritten strokes.

In this paper, we propose an ICA segmentation algorithm, which can be used effectively on handwritten document images containing many different kinds of overlapping and touching words in adjacent lines.

2. Independent Component Analysis

ICA of a random vector x consists of estimating the following generative model for the data:

$$x=As$$

where the latent variable (components) s_i in the vectors $s=(s_1, \dots, s_n)^T$ are assumed independent. The matrix A is a constant $m \times n$ 'mixing' matrix.

This is the simplest and widest used definition in most research on ICA. There are also other ICA definitions, which can be found in the literature [7, 8].

The ICA was chosen in this algorithm based on the three identification points of the ICA model, which are described in [9]:

- A. All the independent components s_i , with the possible exception of one component, must be non-Gaussian.
- B. The number of observed linear mixtures N must be at least as large as the number of independent components M , i.e., $N \geq M$.
- C. The matrix A must be of full column rank.

3. ICA Segmentation Algorithm Approach

The ICA-based Segmentation Algorithm focuses on handwritten documents, which contain overlap or touching words on adjacent lines. The flowchart of the whole algorithm is shown in Figure 1:

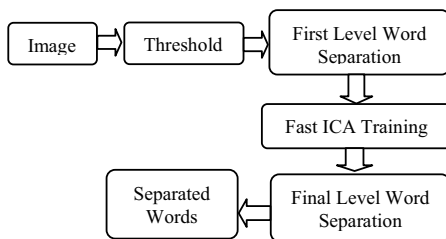


Figure 1. Flowchart of ICA-based segmentation algorithm

3.1. Vector Matrix (VM) Building for ICA Model Training

Vector Matrix ($VM=[x_1; x_2; x_3]$), is the source matrix for training the ICA model. It is a 3-by- ($r \times c$) matrix, where r and c are the row and column number of the whole touching word component respectively.

The outline of building the Vector Matrix for the ICA-based segmentation algorithm is:

1. Preprocessing and Thresholding
2. Overlapping Word Components Detection
3. Overlapping Word Components Area Classification
4. Fuzzy Area Loops Classification
5. Classify Word Components and Restore Grey-Scale Value
6. Achieving Vector Matrix

3.1.1. Preprocessing and Thresholding. Mean-gradient thresholding method [10], which works well in keeping strokes' internal loop and faint tips of words from handwritten images, is used here to obtaining a binary image.

3.1.2 Touching/Overlapping Word Components Detection. In most cases, the touching and overlapping of strokes happen in the characters which include loop descender, such as 'f', 'g', 'j', 'y' and so on as shown in Figure 2(a)-(d).

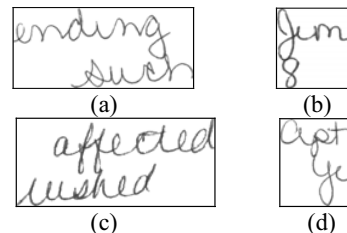


Figure 2. Illustration of loops overlapping with characters on adjacent Lines

3.1.3 Touching/Overlapping Words Region Classification. The detected overlap words component is first separated into Top Area, Bottom Area and Fuzzy Area. This classification is dependent on a left-to-right mapping histogram.

1. **Top Area:** the pixels in the first peak range of the histogram.
2. **Bottom Area:** if the last part of the histogram is flat, then the area from the last peak of the histogram to the end of the histogram is classified as Bottom Area; if the last part of the histogram is a peak, then the whole range of the peak is classified as Bottom Area.
3. **Fuzzy Area:** the unassigned area between the Top Area and Bottom Area is defined as Fuzzy Area.

3.1.4 Fuzzy Area Loops Classification. Firstly, the Laplacian of Gaussian filter is used to find edges of the image. This operator works well on handwriting binary image. Secondly, the closed edges in the Fuzzy Area are retained. In most of case, closed loops are contained in the strokes overlapped area. These closed edges are the edges of the loops inside the Fuzzy Area's strokes. Firm the experimental results, the maximum loop is the major part of overlapped stroke area. Finally, the Center Point Distance, the distance between the center point of two closed loops, is measured for each pair of closed loops in the Fuzzy Area to determine the closest loops. According to the position of two loops, they can be classified into upper loop or lower loop. These two loops are the major roles in overlapped component elementary classification.

3.1.5 Word Components Grey-Scale Value Restoration. Firstly, the upper loop is dilated N times, where N is equal to the width of the word strokes in pixels. The grey scale value of the connected words

where located in the dilated area is restored. The area higher and near the upper loop is simply separated from the overlapping words. Secondly, the lower loop can be restored using the same process to separate the lower word from the overlapping words.

In order to increase the accuracy of the separation in the overlapped region, the separated words and overlap words components are used as the source signals to train the ICA model.

3.1.6 Building the Vector Matrix. Vector Matrix (VM) is a source signal matrix for training the ICA model to separate the overlap words. VM is a 3-by- $(c \times r)$ matrix, where r and c are the row and column numbers of the whole overlap words component respectively. As shown in Figure 3(a), **A** is the overlap words component which has $c \times r$ images, **B** is the first level separated upper word, and **C** is the first level separated lower word. The first row of the components in VM are the pixel intensity values of the overlap words component image **A**, which have been converted from $r \times c$ (r is row numbers, c is column number) format to $1 \times (r \times c)$ (1 row, $r \times c$ columns) as shown in Figure 3(b). Then the upper class component **B** is converted from $r \times c$ to $1 \times (r \times c)$, as the second row of the VM shown in Figure 3(b), as well as the same conversion of the lower word image component **C** as the third row of the VM shown in Figure 3(b).

3.2. ICA Model Training

As shown in Figure 3(a) and Figure 3(b), the elements x_1 , x_2 and x_3 are the overlapping word components, the first level separated upper class word and the lower class word respectively. The vector matrix $x = [x_1 \ x_2 \ x_3]$ is the source signal of the ICA model.

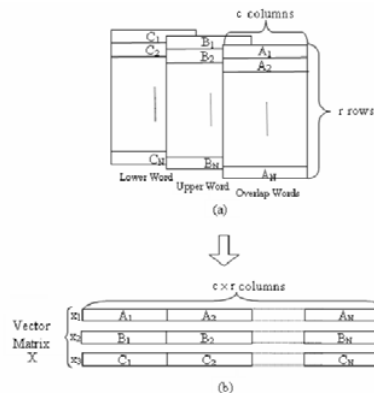


Figure 3. Conversion from Image to Vector Matrix

The ICA model used in the proposed ICA-based Segmentation Algorithm is a Fast ICA. **The Fast ICA algorithm** is a computationally efficient method for performing the estimation of ICA. The ICA model training as shown in Figure 4, $x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$ is the input

source signal of the Fast ICA model, where x_1 is the overlap words, x_2 is the initial separated upper class word, x_3 is the initial separated lower class word. Running the Fast ICA ($x = A \cdot s$) to find $W = A^{-1}$, where **A** is a constant $m \times n$ ‘mixing’ matrix. The vector **A** represents an underlying ‘cause’ of the image and is determined by training the linear image synthesis ICA model.

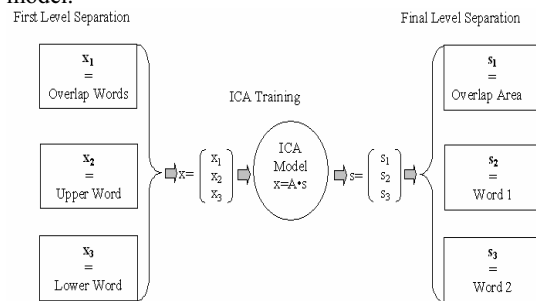


Figure 4. ICA-based segmentation approach

By definition, $s = W \cdot x$, where $W = A^{-1}$ [9]. The independent component $s = \begin{bmatrix} s_1 \\ s_2 \\ s_3 \end{bmatrix}$, where s_1 is the separated overlapped word 1; s_2 is the separated word 2; s_3 is the overlap area. s_1 and s_2 can be realigned to images which include the separated words respectively as shown in Figure 5.

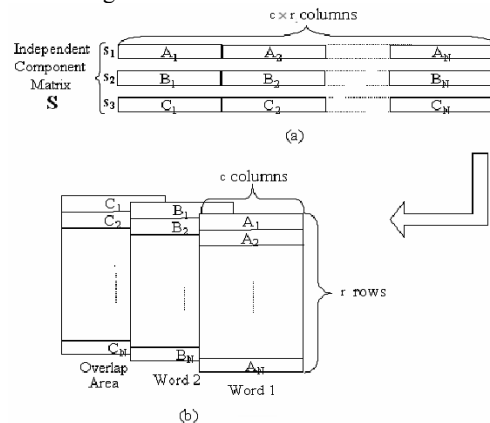


Figure 5. Conversion from Independent Components to Image

5. Experimental Results and Evaluation

5.1. First Level Separation

5.1.1 Overlap Words Component Detection. There are some touching and overlap words on adjacent lines, which are difficult to separate as shown in Figure 6. All the connected character components on adjacent lines in Figure 6 are detected as shown in Figure 7.

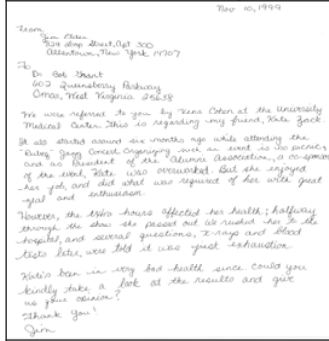


Figure 6. Typical CEDAR database image

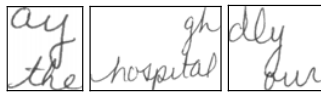


Figure 7. Connected components on adjacent lines

5.1.2 Region Classification of Touching / Overlapping Words Component. Figure 8(a) shows one of the detected overlapping components in Figure 7. Figure 8(b) is the histogram of the number of pixels in each row of Figure 8(a). The overlapping component is classified into three areas: Top Area, Bottom Area and Fuzzy Area as shown in Figure 9.

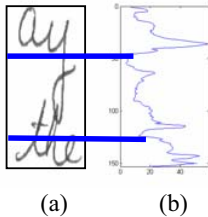


Figure 8. (a) Overlapping Component; (b) Histogram of the number of pixels in each rows of (a)

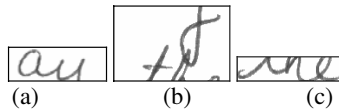


Figure 9. (a) Top Area; (b) Fuzzy Area; (c) Bottom Area

5.1.3 Fuzzy Area Loops Classification. As shown in Figure 10(a), the two closed edges are the edges of the loops inside the Fuzzy Area's strokes.

According to the position of two loops, they can be classified to upper loop or lower loop as shown in Figure 10(b) and Figure 10(c) respectively.

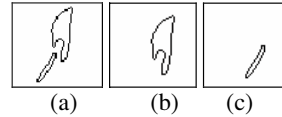


Figure 10. (a) Fuzzy Area's loops; (b) Fuzzy Area's Upper loop; (c) Fuzzy Area's Lower loop

5.1.4 Word Components Grey-Scale Value Restore.

The dilated upper loop area is shown in Figure 11(a). The grey scale value of the connected words located in the upper loop area is restored as shown in Figure 11(b). The area higher and near the upper loop is restored as shown in Figure 11(c). Hence, the first word is separated from the overlapping words.

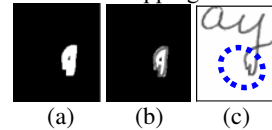


Figure 11. (a) Dilated Upper Loop Area; (b) Restored Upper Loop Area; (c) Restored Upper Words

The lower loop can be restored using the same process as shown in Figure 12.

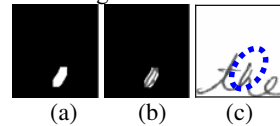


Figure 12. (a) Dilated Lower Loop Area; (b) Restored Lower Loop Area; (c) Restored Lower Words

It can be observed that the boundary of the result of the first level separated stroke region shown in the round dotted circle is very coarse for recognition. The result of the first level separation will be sent to ICA training for the final separation result.

5.1.5 Vector Matrix. The VM is built by the first level separated upper and lower words together with the

overlap words component. $VM = x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$, where x_1 is

the overlap words, x_2 is the first level separated upper class word, x_3 is the first level separated lower class word.

5.2. Final Level Separation based on Fast ICA Training

The weighted value matrix is obtained by training the Fast ICA: $W = \begin{bmatrix} -2.17 & 0.07 & 1.93 \\ 2.55 & -2.57 & 0.08 \\ 3.05 & -3.05 & -3.07 \end{bmatrix}$

The weighted value matrix W is multiplied by the source signal matrix x to adjust the separation results, the blown-up version results are shown in Figure 13.

Figure 13a is part of the lower word of overlap words, Figure 13b is shows the part of the upper separated word, and Figure 13c shows the overlapped area.

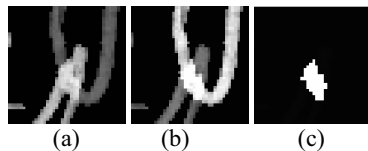


Figure 13. (a) Lower Word; (b) Upper Word; (c) Overlapping Area

Figure 14 shows the binary results of the final separated words. Within the dotted circle area, the much smoother boundary is shown on the overlapped area compared to the first level separated results in Figure 11(c) and Figure 12(c).

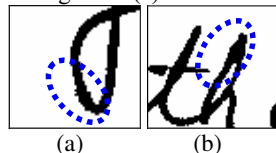


Figure 14. (a) Part of Separated Upper Word; (b) Part of Separated Lower Word

5.4. Evaluation & Conclusion

30 historical images, which exhibit extravagant loops in ascenders, descenders and upper case letters, were selected from the Library of Congress and another 30 handwritten images from the CEDAR letter database to train the algorithms. The images were characterized by high resolution of the scanned images with varying contrast of the handwriting.

The standard measure, recall [11], was used to quantitatively show the relative performance of the proposed method at separating the overlapping words on adjacent lines. Recall is defined as:

$$\text{Recall} = \frac{\text{Correctly Detected Groups}}{\text{Total Overlap Groups}}$$

The average recall value of above 60 images is 95.3%, which shows that ICA-based segmentation algorithm is a very effective method for segmentation of overlap words on adjacent lines.

None of existing segmentation techniques has been shown able to produce consistently good results on the

wide range of document images containing touching or overlapping handwritten strokes.

The proposed ICA segmentation algorithm has excellent performance on separating the overlapping words, which include loops in ascenders, descenders and upper case letters on adjacent lines. The method can be extended to separate other overlap patterns on adjacent lines.

6. References

- [1] R. Sabourin, R. Plamondon, "Segmentation of handwritten signature images using the statistics of directional data", 9th International Conference on Pattern Recognition, 1, Rome, Italy, 1988, pp: 282 - 285
- [2] G. Congedo, G. Dimauro, S. Impedovo, G. Pirlo, "Segmentation of numeric strings", 3rd International Conference on Document Analysis and Recognition, 2(2), 1995, pp: 1038 - 1041
- [3] M. Feldbach, K.D. Tonnie, "Word segmentation of handwritten dates in historical documents by combining semantic a-priori-knowledge with local features", 7th International Conference on Document Analysis and Recognition, 1, Montreal, Canada, 2003, pp: 333 - 337
- [4] K.W. Cheung, D.Y. Yeung, R.T. Chin, "Bidirectional deformable matching with application to handwritten character extraction", IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(8), 2002, pp: 1133 - 1139
- [5] J. Sadri, C.Y. Suen, T.D. Bui, "Automatic Segmentation of Unconstrained Handwritten Numeral Strings", 9th International Workshop on Frontiers in Handwriting Recognition, Tokyo, Japan, 2004, pp: 317 - 322
- [6] Y. Daekeun, K. Gyeonghwan, "An approach for locating segmentation points of handwritten digit strings using a neural network", 7th International Conference on Document Analysis and Recognition, Edinburgh, Scotland, 2003, pp: 142 - 146 vol.1
- [7] P. Comon, "Independent component analysis - a new concept", Signal Processing, 36, 1994, pp: 287-314
- [8] C. Jutten, J. Herault, "Blind separation of sources", Signal Processing, Part I: An adaptive algorithm based on neuromimetic architecture. 24, 1991, pp: 1-10
- [9] A. Hyvarinen, "Survey on Independent Component Analysis", Helsinki University of Technology, Finland, 1999.
- [10] C.G. Leedham, Y. Chen, K. Takru, J. Tan, and M. Li, "Comparison of some thresholding algorithms for text/background segmentation in difficult document images", 7th International Conference on Document Analysis and Recognition, 2, Edinburgh, Scotland, 2003, pp. 859 -865.
- [11] M. Junker, R. Hoch, "On the evaluation of document analysis components by recall, precision, and accuracy", 5th International Conference on Document Analysis and Recognition, Bangalore, India, 1999, pp: 713-716.