A Hough Based Algorithm for Extracting Text Lines in Handwritten Documents

Laurence Likforman-Sulem, Anahid Hanimyan, Claudie Faure

Ecole Nationale Supérieure des Télécommunications, CNRS-URA 820
46 rue Barrault, 75013 Paris, France

Abstract

The method herein proposed detects text lines on handwritten pages which may include either lines oriented in several directions, erasures, or annotations between main lines. The method has a hypothesis-validation strategy which is iteratively activated until the end of the segmentation is reached. At each stage of the process, the best text-line hypothesis is generated in the Hough domain. taking into account the fluctuations of the text-line components. Afterwards, the validity of the line is checked in the image domain using a proximity criteria which analyses the context in which is perceived the alignment hypothesed. Ambiguous components belonging to several text lines are also marked.

1. Introduction

Optical reading necessitates segmenting the text image into physical components before performing symbol recognition. A text has a linear structure which can be described, at the symbol level, by a string of characters or words. The physical components corresponding to this linear stucture are the lines of text.

Methods such as projections, smearing and Hough Transform have been successfully applied to the detection of text lines in printed documents [9]. They could also be applied to handwritten documents in the case of regular or simplified texts. One characteristic of handwritten documents is layout variability. For constrained documents, such as bank checks and forms, the text appears within *a priori* given text zones. Knowledge about position and orientation is used to drive the extraction of text lines. For unconstrained handwritten documents, methods adapted to the layout variability of such documents must be defined. A method for grouping components into text lines from images of unconstrained handwritten documents has been recently proposed [4].

In the present paper, a different approach is proposed, based on Hough transform. Hough transform has been applied for skew detection in printed documents [2], for string detection in engineering documents [1], for line detection in printed texts [7] or simplified handwritten texts [6]. Hough transform can also be used to detect stroke orientation in handwritten words [3] [8].

In our approach, text lines are extracted in handwritten documents using an iterative hypothesis-validation strategy. Information gathered from both the Hough domain and the image are combined. At each stage of the process, a text-line hypothesis is obtained by searching the best alignment of connected components in the Hough domain. Afterwards, the validation occurs in the image domain using contextual information which enables the rejection of alignments of components which have no perceptual relevance. No assumption is made about orientation or position of the text lines.

2. Line selection in the Hough domain

The Hough transform is applied to the gravity centers of the connected components in the image. In the Hough domain, collinear alignments are searched in any direction. The process takes into account possible fluctuations of text lines, slight variations of the main direction, the irregularity of interlines and does not assume any privileged direction.

The Hough transform is a line to point transformation from the cartesian space to the polar coordinate space. A line in the cartesian coordinate space can be described by :

$\rho = x*\cos\theta + y*\sin\theta$

where ρ is the normal distance of the line from the origin and θ the angle between the x-axis and the normal line. A line corresponds to a point (ρ, θ) in the Hough domain which is quantized into cells. For each component gravity center in the image, the set of lines passing through that point for different discrete values of ρ and θ corresponds to a set of cells in the Hough domain. The cells are initialised to zero, and incremented by one, each time a point in the image (a gravity center) belongs to that line. Strong alignments correspond to cells with large values.

The starting point of our method is the work of Fletcher and Kasturi who proposed a Hough based method for detecting text strings in printed mixed text/graphics documents. Unlike printed documents, horizontal and

0-8186-7128-9/95 $4.00 © 1995 IEEE

vertical directions are not favored in unconstrained handwritten documents. Therefore in our case, horizontal and vertical text lines are not searched before other directions. Following these authors, fluctuations of the gravity centers around the main direction are taken into account when choosing the resolution along the ρ direction in the Hough domain and by defining a cluster including near ρ values around the primary cell. The primary cell corresponds to the cell in the Hough domain, which constitutes the best text line hypothesis.

For constructing the Hough domain, the resolution along the θ direction was set to 1 degree and the resolution along the ρ direction was set to $R = 0.2\ H_w$,with H_w being the average height of all connected components of the document image.

The first hypothesis generated in the Hough domain is the cell having the greatest count, say cell (ρ_0, θ_0). The clustering factor enables taking into account the possible fluctuations in the position of centroïds. These fluctuations depend on the height of connected components in the alignment. As the alignment is not yet defined, the average height of components is estimated by H_{moy} which is the average height of connected components which are 5-cells away from cell (ρ_0, θ_0). The same estimation is done in the work of Fletcher and Kasturi. The clustering factor is defined by :

$f_{clus} = H_{moy}/R$ if $85° \leq \theta_0 \leq 95°$

$f_{clus} = 0.5{*}H_{moy}/R$ if $0° \leq \theta_0 < 85°$ or if $95° \leq \theta_0 < 180°$

We distinguish between horizontal and inclined lines since the bounding rectangle of connected components which lay in inclined directions is higher than the rectangle of the same component in a horizontal direction.

In order to tolerate a variation of the line direction θ, a rectangular sub-domain x0, x1, z0, z1 in the Hough domain is defined in which the second hypothesis is searched (figure 1). The sub-domain is centered around the cell having the greatest count (ρ_0, θ_0) :

$x0 = \rho_0 - f_{clus}$ $\qquad$ $x1 = \rho_0 + f_{clus}$

$z0 = \theta_0 + \Delta\theta$ $\qquad$ $z1 = \theta_0 + \Delta\theta$

with $\Delta\theta=3$ cells corresponding to the rotation of the line by 1, 2 or 3 degrees around θ.

Cell (ρ_1, θ_1) is the cell having the second highest value in the rectangular sub-domain. The first hypothesis is (ρ_0, θ_0), the second one (ρ_1, θ_1). The primary cell is the best hypothesis, i.e. the one which best fits the data.

Let us call the *structure* of a cell (ρ, θ), the set of cells including the cell itself at a central position and the cells in a cluster $(\rho - f_{clus}, \rho + f_{clus})$ around the position (ρ, θ). Let us call n0 the value of the cell (ρ_0, θ_0) and n1 the value of the cell (ρ_1, θ_1). The second hypothesis is chosen instead of the first one if n1 is sufficiently large relative to n0 ($n1/n0 > 0.65$) and if the second structure contains more components than the structure associated to (ρ_0, θ_0). Consequently the line (ρ_1, θ_1) which is derived from line (ρ_0, θ_0) by a maximum rotation of 3 degrees, is chosen because it contains more elements. This principle can be also derived from perceptual organisation since the more elements an alignment includes, the stronger it is. In all other cases, the first hypothesis (ρ_0, θ_0) is chosen as the primary cell.

Those components which are grouped to form the line, are in the cells belonging to the structure associated with the primary cell. However some extra components may be added to the line if they immediately stand at the border of the cluster.

The alignment hypothesis found in the Hough domain does not necessarily correspond to text lines, hence a validation procedure is performed in the image domain.

3. Image domain

Alignments read by human beings correspond to visually perceived alignments. To be perceived, alignments occur in a context which enables the perception of actual text lines and inhibits the perception of other alignments. Without contextual information, the best alignments of components (the longest ones) may cross text lines. This happens when the height of the text is larger than its width.

The perceptual relevance of alignment (ρ, θ) found in the Hough domain is now checked. Components are first ordered along the line (ρ, θ), grouped into a string where inter-component distances are computed. Then for all components of the line, the number of nearest neighbors within the Hough line is compared with the number of nearest neighbors in the image which do not belong to the line. At this stage, ambiguous components that could belong to the line are searched.

3.1 Forming a string of components

Components are first ordered left to right by projecting their centroids c on the line (ρ, θ). Let us call c' the projection of c. The order of components is the order given by the x coordinates of the projected points c'. Computing distances between pairs of connected components has been studied in [5]. Here the inter-components distance d_l depends on the position of bounding boxes and their centroids relatively to the line. Distance d_l between two adjacent components is defined by the edge to edge distance of bounding boxes along line (ρ, θ) if they are crossed by line (ρ, θ). When one component c is not crossed by line (ρ, θ), the projection c' of its centroid on the line is used to compute an edge to point distance. If both do not intersect the line, the distance is a point to point distance between the projections of the centroids on line (ρ, θ). If components overlap and are intersected by the line, the distance is considered to be zero.

All components do not contribute to the computation of inter-component distances: those enclosed in another component, and those which are not crossed by line (ρ, θ) and whose centroid projection falls into the rectangle of one component belonging to the line.

Ambiguous components are those which are not included in the alignment (because their centroids are two far) but which are intersected by the line (ρ,θ). Ambiguous components result from the overlapping of two writing lines, when ascenders or descenders appear in the space of another line, or when the component is at the intersection of two lines running in different directions. The marking of ambiguous components provides valuable information used for further processing such as symbol recognition, or the separation of components into several parts, or the association of the component to a unique alignment. The string forming the line is re-ordered taking into account ambiguous components.

3.2 Perceptual relevance of an alignment

The evaluation of the relevance of the alignment is performed by checking its context. If most of the nearest neighbors of the components of the alignment string belong to the group of components forming the alignment, the alignment has both properties of direction continuity and proximity, and it will be accepted as a text line.

The internal nearest neighbors of one component c_i are the preceding and the following components of c_i on the line: c_{i-1} and c_{i+1}. The external nearest neigbors of c_i are defined as the components which do not belong to the line but which are within a distance d_e^i from c_i :

$$d_e^i = \frac{d_l\,(c_{i-1},\, c_i) + d_l\,(c_i,\, c_{i+1})}{2}$$

d_l is the distance between two adjacent components of the line. The distance between one component c_i on the line and a component c outside the line is an edge to edge distance along a line joining their respective centroids.

If the number of external neighbors is greater than the number of internal ones, the alignment is validated and components of the alignment are removed from the whole Hough domain. In the opposite case, the alignment is invalidated and components belonging to the alignment are only removed from the cluster associated with the primary cell in order that they will be able to participate in other alignments at a next iteration.

4. Results and conclusion

We have tested our method on several kinds of documents including rough drafts, address blocks, letters. Although the main direction is often roughly horizontal, this information has not been used *a priori*. The lack of data bases for handwritten documents and the fact that handwritten documents do not constitute a homogeneous class, both hinder the quantitative evaluation of segmentation methods. Figure 2 shows the segmentation result on a manuscript. Alignments found as text lines are crossed by a line, components belonging to the same line share the same identification number inscribed above their enclosing rectangles. In another example (figure 3), alignments which were found in the Hough domain, but invalidated in a second stage are crossed by a dashed line. These alignments were generally found at the beginning of the iterations, crossing the height of the text. On such images fluctuations can be observed, as well as insertions and annotations. Ambiguous components are inscribed in dashed rectangles. However, when fluctuation is combined with proximity of text lines, merging lines may appear.

The Hough domain and the image domain are used in combination for detecting text lines in unconstrained handwritten texts. The Hough transform becomes a powerful technique when coupled with a validation procedure. The validation enables the rejection of alignments of components occuring in a context which inhibits their perception.

Experiments performed on several kinds of handwritten documents detect fluctuating text lines, sloped annotations or annotations added between main lines, and mark ambiguous components. The process can be refined towards two directions. The first one is the post-processing of ambiguous components in order to split them or include them into a line. The second one consists in validating parts of alignments when they are composed of components belonging to an actual text line and additional spurious elements.

References

[1] Fletcher L.A., Kasturi R. Text string segmentation from mixed text/graphics images, *IEEE PAMI*, Vol 10, No 3, pp 910-918, 1988.

[2] Hinds S. C., Fisher J., D'Amato D. A document skew detection method using run-length encoding and the Hough transform, *Proceedings of the 10th IAPR*, Atlantic City, pp 464-468, 1990.

[3] Lecolinet E. (1994) Cursive script recognition by backward matching, in *Advances in handwriting and drawing : a multidisciplinary approach*, C. Faure, P. Keuss, G. Lorette, A. Winter (Eds), Europia, Paris, pp 117-135.

[4] Likforman-sulem L., Faure C. Extracting text lines in handwritten documents by perceptual grouping, in *Advances in handwriting and drawing : a multidisciplinary approach*, C. Faure, P. Keuss, G. Lorette, A. Winter (Eds), Europia, Paris, pp 21-38, 1994.

[5] Seni G., Cohen E. External word segmentation of off-line handwritten documents, *Pattern Recognition,* Vol 27, No 1, pp 41-52, 1994.

[6] Shapiro V., Gluhchev G., Sgurev V. Handwritten document image segmentation and analysis, *Pattern Recognition Letters* , No 14, pp 71-78, 1993.

[7] Srihari S. , Govindaraju V. Analysis of textual images using the Hough transform, in *Machine Vision and Applications*, 2, pp 141-153, 1989.

[8] Vincent N., Dargenton P., Emptoz H. Utilisation de la transformée de Hough dans la reconnaissance de l'écriture manuscrite, *Actes de CNED'92*, Nancy, Juillet 92, pp 294-301.

[9] Tang Y., Suen C., C. Yan, Cheriet M, Document analysis and understanding : a brief survey, ICDAR 91, Saint Malo, pp 17-31.

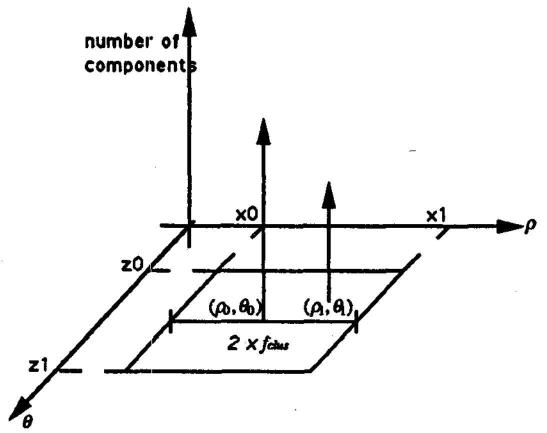

Figure 1 : rectangular zone in the Hough domain where hypotheses are searched.

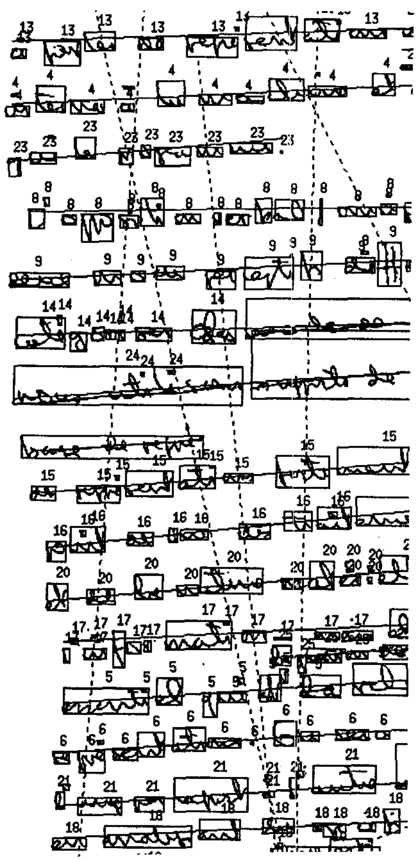

Figure 3 : segmentation result on a part of manuscript. Invalidated alignments are crossed by a dashed line

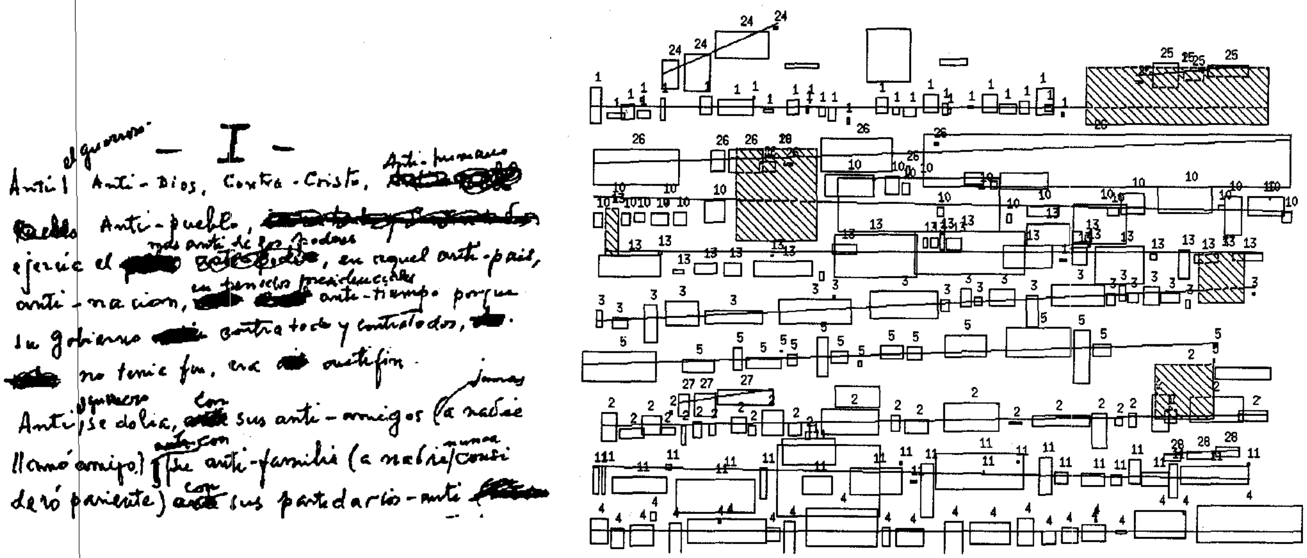

Figure 2 : original image (on the left) and the corresponding segmentation into lines (on the right)