Letter-Level Shape Description by Skeletonization in Faded Documents

Rahul Singh Michael C. Wade Nikolaos P. Papanikolopoulos

Artificial Intelligence, Robotics, and Vision Laboratory
Department of Computer Science and Engineering
University of Minnesota, Minneapolis, MN 55455

Abstract

We present a method for determining the skeletal shape description for letters in texts faded due to ageing and/or poor ink quality. The proposed algorithm is interesting in that it neither involves assumptions about demarcation of object regions from the background, nor does it require pixel connectivity in the text regions. Consequently, it may be applied for obtaining the shape descriptions of "sparse" regions, which are characteristic of letters in faded documents. Given the pixel distribution for a letter or a word from a faded document, the method involves an iterative evolution of a piecewise-linear approximation of the principal curve of this pixel distribution. By constraining the principal curve to lie on the edges of the Delaunay triangulation of the shape distribution, the adjacency relationships between regions in the shape can be detected and used in evolving the skeleton. The approximation of the principal curve, on convergence, gives the final skeletal shape. The skeletonization is invariant to Euclidean transformations and is adaptive in terms of the topology of the underlying shape distribution as well as in the number of units needed for the piece-wise approximation of the principal curve.

1 Introduction

Analysis and processing of information from degraded printed texts is an important problem in contemporary document image analysis. In this paper, we consider a special case of this problem, where the image degradation is due to fading of the printed text. Such a document defect model can occur, among others, due to ageing of the document, exposure of printed material to conditions of high humidity, poor scanning, as well in photocopiers and faxes with low toner, or due to poor thresholding. The complexity of this problem can be gauged from the example in Figure 1,

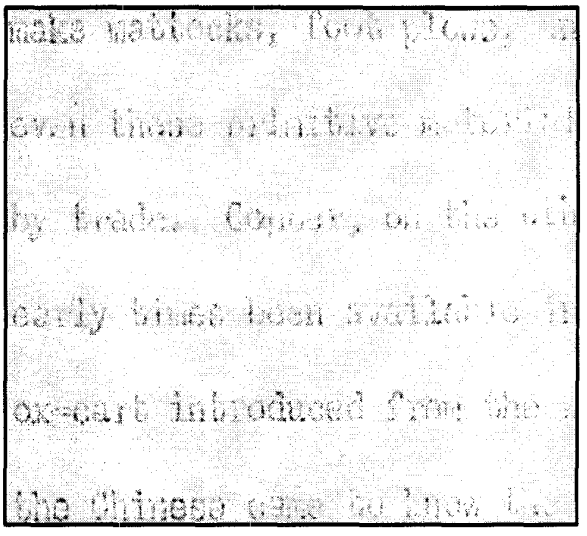
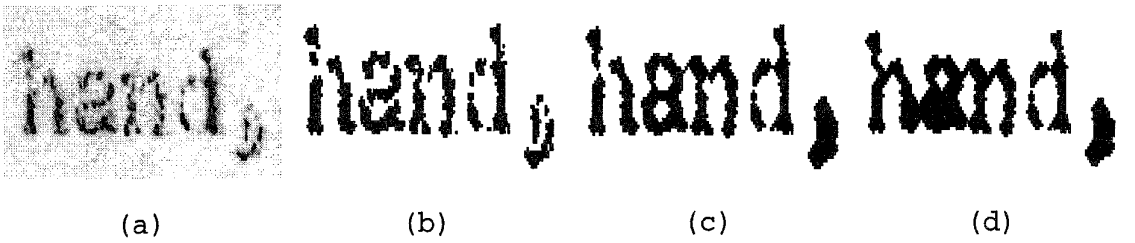


Figure 1: Section from a text faded due to ageing.

where a section from a scanned page of an old text is shown.

A fundamental issue, in document analysis in general and our problem in particular, is that of representing the letters or words by using their *essential characteristics*. At the present state-of-the-art, no ideal representation scheme is known and different representations have their own advantages and disadvantages. Representation using boundary feature extraction is precluded in the given domain due to the sparseness and the breaks in the character images. Representation based on topological features like angle of a crossbar or position of holes, have been found to increase in complexity as image degradation occurs [3]. In this paper, we represent the shape of the letters by using the shape skeleton, obtained by approximating the *principal curve* of the shape distribution. The concept of principal curves was introduced in Statistics [7], to

0-8186-8606-5/98 $10.00



(a) (b) (c) (d)

Figure 2: Dilation-Erosion on faded words: (a) gray-scale image, (b) after 1 dilation-erosion operation, (c) after 2 dilation-erosion operations, (d) after 3 dilation-erosion operations

summarize patterns exhibited by points in a scatterplot. The principal curve is a non-linear generalization of the first principal component and is defined for both sparse and non-sparse distributions. It may be noted here, that skeletonization of faded images is ill-posed in the sense of the conventional techniques, due to the anisotropy introduced inside the image regions by the lack of pixel connectivity. Furthermore, in the case of highly faded documents, conventional filtering techniques like median filtering or morphological operations such as dilation-erosion are ineffective. In Figure 2, we present some results of performing a sequence of dilation-erosion operations on a faded word. Note that not only there is no *a priori* way to determine the number of dilation-erosion operations which may be needed, but distortion of topological features as well as merging of adjacent letters may also occur. In contrast, the approach proposed by us does not involve changing the input data. Rather, we propose an alternate definition of the shape skeleton. This definition is valid not only for faded shapes, but may also be applied to conventional ones.

We begin this paper with an overview of the prior research in the processing of degraded texts (Section 2). The proposed method is formulated in Section 3. In Section 4 some issues related to the evolution of the shape skeleton are considered. Results of applying the method to various shapes are reported in Section 5. The conclusions of the present research and possible directions of future work are reported in Section 6.

2 Previous Work

A variety of approaches have been proposed for processing degraded documents. Explicit quantitative models describing imaging defects due to printing and errors in digitization were proposed in [1]. Based on the hypothesis that much of the noise introduced during scanning and transmission is random, bitmap averaging [8] was used to cancel out the noise and produce images with smooth outlines. Hidden Markov Model along with a level building dynamic programming algorithm was used for the recognition of connected and degraded text in [4]. In this work structural analysis of the letters was performed using a line adjacency graph after median filtering had been used to reduce the noise and ensure connectivity. Topological features like holes in the object, vertical bars, and corner opens were used in [14] to recognize facsimile documents. Image degradations due to salt-and-pepper noise and partial omission of figures was handled in [13]. The method was based on producing gray level images from binary images by computing the density of black pixels in an $N \times N$ grid placed on each pixel. A similar approach was also used in [3]. Our work differs significantly from the above mentioned approaches in that, given a faded image, the issue of its shape representation is directly addressed by computing its shape skeleton. Unlike [8], we do not assume that the sensor noise is uniformly random. Furthermore, our approach does not require selection of hand-crafted features of the type used in techniques based on topological features.

Recent efforts towards solving the problem of obtaining skeletons for images which may lack pixel-level connectivity include an entropy based method [5] and the use of neural models [6]. The entropy based technique in [5] is based on computing the circular range containing the maximal information for each pixel. The symmetry score of the pixel distribution in the circular range is then computed. This symmetry information is treated as a grey-scale image which is thinned to obtain the shape skeleton. In [6], a flow-through self-organizing map (SOM) is used to obtain the shape skeleton of a connected set of points. The SOM is initialized with a linear topology and evolves to arc patterns, forked patterns or circular patterns based on thresholds on the angle formed at each map unit by its neighbors (for changing from linear to forked patterns) and the distance between two map units (for evolving from linear to circular patterns).

3 The Proposed Method

3.1 Principal Curves

Principal curves are smooth, non-linear generalization of principal components [7] and provide non-linear dimensionality reduction. Unlike the formulation in regression where the dimensions of the data are treated in terms of a predictor-response relationship, principal curves summarize the data by treating each of the dimensions symmetrically. For a given data distribution, its principal curves are smooth, self-consistent curves that pass through the *middle* of the distribution.

Let $\mathcal{X}$ be a random set distributed according to the probability density $\mathbf{p}$ in $\Re^d$. Let $\mathbf{E}(\mathcal{X}) = 0$ and let $\mathcal{X}$ have finite second moments. Denote by Γ a compact, non self-intersecting, and unit speed curve in $\Re^d$, parameterized over $\Lambda \subseteq \Re$.

Let λ_Γ be the projection index, $\lambda_\Gamma : \Re^d \rightarrow \Re$, such that

$$\lambda_\Gamma(x) = \sup_\lambda \{\lambda : \| x - \Gamma(\lambda) \| = \inf_\xi \| x - \Gamma(\xi) \| \} \quad (1)$$

Note that the projection index $\lambda_\Gamma(x)$ assigns to each $x \in \Re^d$, a point $\Gamma(\lambda)$, which for the given value of λ is the closest to x. Since Γ is compact, there exists at least one such point on Γ. It may be shown (see [7]), that the Lebesgue measure of the set of points having more than one closest point on Γ is zero. Principal curves are defined as [7]:

Definition 1 *The curve Γ is called a principal curve of the density $\mathbf{p}$ if $\mathbf{E}(\mathcal{X} \mid \lambda_\Gamma(\mathcal{X}) = \lambda) = \Gamma(\lambda)$ for almost every λ.*

The self-consistency of the principal curve implies that for any point on the principal curve, the average of all data points projecting onto it coincides with this point on the curve. The algorithm for computing the principal curve consists of iterating (until convergence) the following steps with successively smaller scatterplot smoothing span:

1. *Projection step*: For each data point, find its projection on the curve. The projection is determined by finding the closest point on the curve (in the orthogonal sense).
2. *Conditional-Expectation step*: Scatterplot smooth the projected values along the curve length.

The relation of the principal curve to the medial axis of a shape follows from its property of self-consistency. Furthermore, the non-linear nature of the principal curve is important in obtaining smooth, unidimensional representations of (possibly) high dimensional, complex distributions. Since principal curves are well defined for both dense (4 or 8 connected) and sparse data, skeletal descriptions based on them can be obtained both for images of faded documents, as well as the conventional non-faded documents.

3.2 Shape Skeletons Based on Principal Curves

The principal curve algorithm provides a basis for obtaining the skeletal description of both sparse and non-sparse shapes. However, crossovers of regions can not be faithfully represented by principal curves, due to the fact that by definition, principal curves do not self-intersect. Our method involves two results from research in Neural Networks to solve this problem.

- For a given distribution we obtain a discrete approximation of its principal curve by using the batch formulation of the SOM algorithm [10, 15]. The approximation of the principal curve thus obtained is parametrized according to the topological coordinates of the SOM units. Invariance to Euclidean transformations follows, since in the batch method the final position of the map units is invariant to the order of presentation of the data [12].
- Instead of the traditional topologies used with SOM (*linear* and *grid*), a minimum spanning tree (MST) on the map units is used to define the neighborhood relationships. Such a topology can provide better representation of highly non-linear distributions [9]. In this topology, arcs are assigned between map units and the length of the arcs is defined as the distance between the nodes in the input space. By definition, the sum of the length of all the arcs in the tree is minimal. The sections of the spanning tree between the branching points thus provide a piece-wise approximation of the principal curve for the data points projecting onto these sections. Furthermore, the connectedness and thinness of the minimum spanning tree is ideal for obtaining a skeletal representation. The use of a MST topology also obviates the need to define threshold based criteria as in [6] for evolution of the skeletal topology.

Let $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_\mathcal{D}\}$ be the set of input vectors (data), and $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_\mathcal{U}\}$ be the units of the SOM respectively. Also let the location of unit $\mathbf{u}_j$

($j \in [1, \mathcal{U}]$), in the sample space be $\mathbf{W}_j$ and let its coordinate location in the topological space of the map be $\tilde{\mathbf{j}}$. Let $\mathcal{D}$ denote the size of the data set and $\mathcal{U}$, the number of units in the map respectively. The batch-mode SOM algorithm consists in iterating the following two steps:

1. *Voronoi tessellation of the input data*: The data is partioned into Voronoi regions of the units. The centroid of each Voronoi region is computed along with its size.

$$\arg\min_j (\| \mathbf{x}_k - \mathbf{W}_j \|) \quad k = 1, \cdots, \mathcal{D}$$

   Conceptually, this step may be considered as a single iteration of the Lloyd vector quantization algorithm (K-means algorithm).

2. *Kernel smoothing on the centroids in the topological space*: The units are updated by a weighted centroid of the data. The weights of each datum are determined by the Voronoi region it belongs to and correspond to the neighborhood function of the flowthrough SOM update formulation:

$$\mathbf{W}_j = \frac{\sum_{p=1}^{\mathcal{U}} C(\tilde{\mathbf{p}} - \tilde{\mathbf{j}}) \mathcal{M}_p \mathcal{S}_p}{\sum_{p=1}^{\mathcal{U}} C(\tilde{\mathbf{p}} - \tilde{\mathbf{j}}) \mathcal{S}_p} \tag{2}$$

   where $\mathcal{M}_p$ is the centroid of the Voronoi region defined by map unit $\mathbf{u}_p$ and $\mathcal{S}_p$ is the number of samples in this region. $C(.)$ is a monotonically decreasing neighborhood function defined in the topological space. In the present work, a Gaussian was used as the neighborhood function.

4 Issues in the Evolution of the Skeleton

The approximation of the skeletal representation for a given shape distribution depends not only upon the map topology, but on the number of map units as well. Since it is not possible to determine *a priori* the number of map units for an unknown shape distribution, an adaptive procedure to change the map size is desirable. A smooth evolution of the skeleton can occur if the perturbation to the skeletal shape during the addition and/or deletion of map units is minimal. To this end, we follow the approach suggested in [6] and add a new unit in the middle between two existing units $\mathbf{u}_i$ and $\mathbf{u}_j$ if $\| \mathbf{W}_i - \mathbf{W}_j \| > \delta_{\mathbf{max}}$. Similarly, two existing units $\mathbf{u}_k$ and $\mathbf{u}_l$ are merged into a single one if the distance between their weight vectors $\| \mathbf{W}_i - \mathbf{W}_j \| < \delta_{\mathbf{min}}$. For a given shape distribution skeletal descriptions of varying details can be obtained by suitably tuning the parameters $\delta_{\mathbf{min}}$ and $\delta_{\mathbf{max}}$.

While the spanning tree topology provides an intrinsic method to span complex shapes, it is unable to represent certain topological properties like the closure of circular regions. Circular regions can be represented by knowing the adjacency relations between the Voronoi regions of the shape as represented by their Delaunay triangulation. We define the *neighborhood* relationship between two Voronoi regions $\mathbf{V}_\mathcal{P}$ and $\mathbf{V}_\mathcal{Q}$ with centroids $\mathbf{u}_\mathcal{P}$ and $\mathbf{u}_\mathcal{Q}$ based on their second-order Voronoi polyhedra $\mathbf{V}_{\mathcal{PQ}}$, where $\mathbf{V}_{\mathcal{PQ}}$ is defined as [11]:

$$\begin{aligned} \mathbf{V}_{\mathcal{PQ}} = \{ \mathbf{x}_i \in \Re^d : & \| \mathbf{x}_i - \mathbf{W}_\mathcal{P} \| \leq \| \mathbf{x}_i - \mathbf{W}_\mathcal{Q} \| \\ & \leq \| \mathbf{x}_i - \mathbf{W}_\mathcal{K} \|, \forall \mathcal{K} \neq \mathcal{P}, \mathcal{Q} \} \end{aligned} \tag{3}$$

It may be noted (see [11] for a proof), that

$$\mathbf{V}_{\mathcal{PQ}} \neq \phi \Leftrightarrow \mathbf{V}_\mathcal{P} \cap \mathbf{V}_\mathcal{Q} \neq \phi \tag{4}$$

The computation of the neighborhood relationships can then be done by determining for each $x_i \in \mathcal{X}$, the centroids $\mathbf{W}_k$ and $\mathbf{W}_l$, such that

$$\| \mathbf{x}_i - \mathbf{W}_k \| \leq \| \mathbf{x}_i - \mathbf{W}_j \|, \forall j \tag{5}$$

and

$$\| \mathbf{x}_i - \mathbf{W}_l \| \leq \| \mathbf{x}_i - \mathbf{W}_j \|, \forall j, j \neq k \tag{6}$$

The relations in Eqs. (5) and (6) can be implemented by joining, for each datum, the two closest centroids during the Voronoi tessellation in the batch SOM algorithm. Since the shape distributions we are dealing with form sparse manifolds, the above relations lead to an induced Delaunay triangulation representing the neighborhood property expressed in Eq. (3). Skeletal representation of circular regions are obtained by joining two disjoint units in the SOM if there is an edge between them in the induced Delaunay triangulation. The occurrence of small loops in the skeleton can be prevented by allowing two map units to join only if the resultant cycle has more than K edges. We use the value of $K = 3$ in all our experiments.

5 Experimental Results

The algorithm has been applied on several words from different faded documents. The pertinent document pages were scanned and bi-level thresholding was performed using moment preserving thresholding. Skew estimation and correction was performed using the algorithm in [2]. After skew correction, horizontal

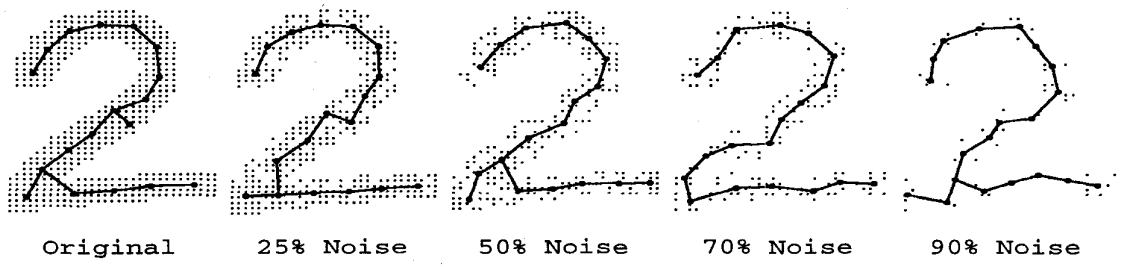


Figure 3: Skeletonization performance under decreasing SNR. The original image is a dilated handwritten numeral. Uniformly distributed random noise is added to the shape.

projection profiles were used to separate lines in the text. The separation of words and individual letters was done using vertical projection profiles. To evaluate if disconnected regions belong to a single letter, the difference between the total width of these regions and the running average of the width of letters in the document was computed. These regions were grouped as a single letter if the product of this difference and the separation between the regions was less than an empirically established threshold. The proposed algorithm was invoked on each letter, and the final skeletal shape was obtained on convergence.

The robustness of the algorithm was tested on images with additive random noise. Uniformly distributed random noise was generated in the bounding box containing the image. A desired *signal to noise ratio* (SNR) was obtained by counting the number of foreground pixels which had their values flipped. An example from this experiment is presented in Figure 3. It may be noted that this type of noise is not typical of the noise caused due to image fading.

The skeletal representation of words from the third, fourth, and fifth lines of the text shown in Figure 1 are presented in Figure 4. The reader may note the breaks in the skeletons caused due to excessively large gaps within many individual letters.

6 Conclusions and Future Work

This paper presents a new method for skeletonizing letters in faded texts. Due to the lack of contiguity in the image regions, conventional skeletonization techniques are either inapplicable for such shapes or perform poorly. Our approach is based on obtaining the shape skeleton by approximating the principal curve of the sparse images. The approximations are computed by an algorithm, which is based on the batch formulation of the SOM. In our implementation both the topology and the size of the map are adaptively determined, based on the input shape distribution. Experimental results indicate that the method performs robustly for words from documents with varying degrees of fading. In the present formulation, the algorithm depends on the $\delta_{\mathbf{min}}$ and $\delta_{\mathbf{max}}$ parameters. Their values are currently determined by using a global estimation technique. If the noise distribution has a local character, then such estimations are often poor. We also aim for developing a method for letter-level segmentation, based on the adjacency relations obtained using the proposed method.

Acknowledgements

The authors would like to thank Vladimir Cherkassky, who has participated in and co-authored [16], the skeletonization algorithm. Thanks are due to the James J. Hill Reference Library, St. Paul Minnesota for the faded text used as an example in this paper. This research was supported by the NSF through Grants #IRI-9410003 and #IRI-9502245.

References

[1] H. S. Baird. Document image defect models. In H. S. Baird et. al., editor, *Structured Document Image Analysis*, pages 546–556. Springer-Verlag, 1992.

[2] H. S. Baird. The skew angle of printed documents. In L. O'Gorman and R. Kasturi, editors, *Document Image Analysis*, pages 204–207. IEEE Computer Society Press, 1995.

[3] M. Bokser. "Omnidocument Technologies". *Transactions of the IEEE*, 80(7):1066–1078, 1992.

[4] C. B. Bose and S-S Kuo. "Connected and Degraded Text Recognition Using Hidden Markov Model". In *Proceedings of the 11th International Conference on Pattern Recognition*, pages 116–118, 1992.

[5] Y. S. Chen and Yu. T. Yu. "Thinning Approaches for Noisy Digital Patterns". *Pattern Recognition*, 29(11):1847–1862, 1996.

[6] A. Datta, S. K. Parui, and B. B. Chaudhuri. "Skeletal Shape Extraction from Dot Patterns by Self-Organization". In *Proceedings of the 13th International Conference on Pattern Recognition*, volume 4, pages 80–84, 1996.

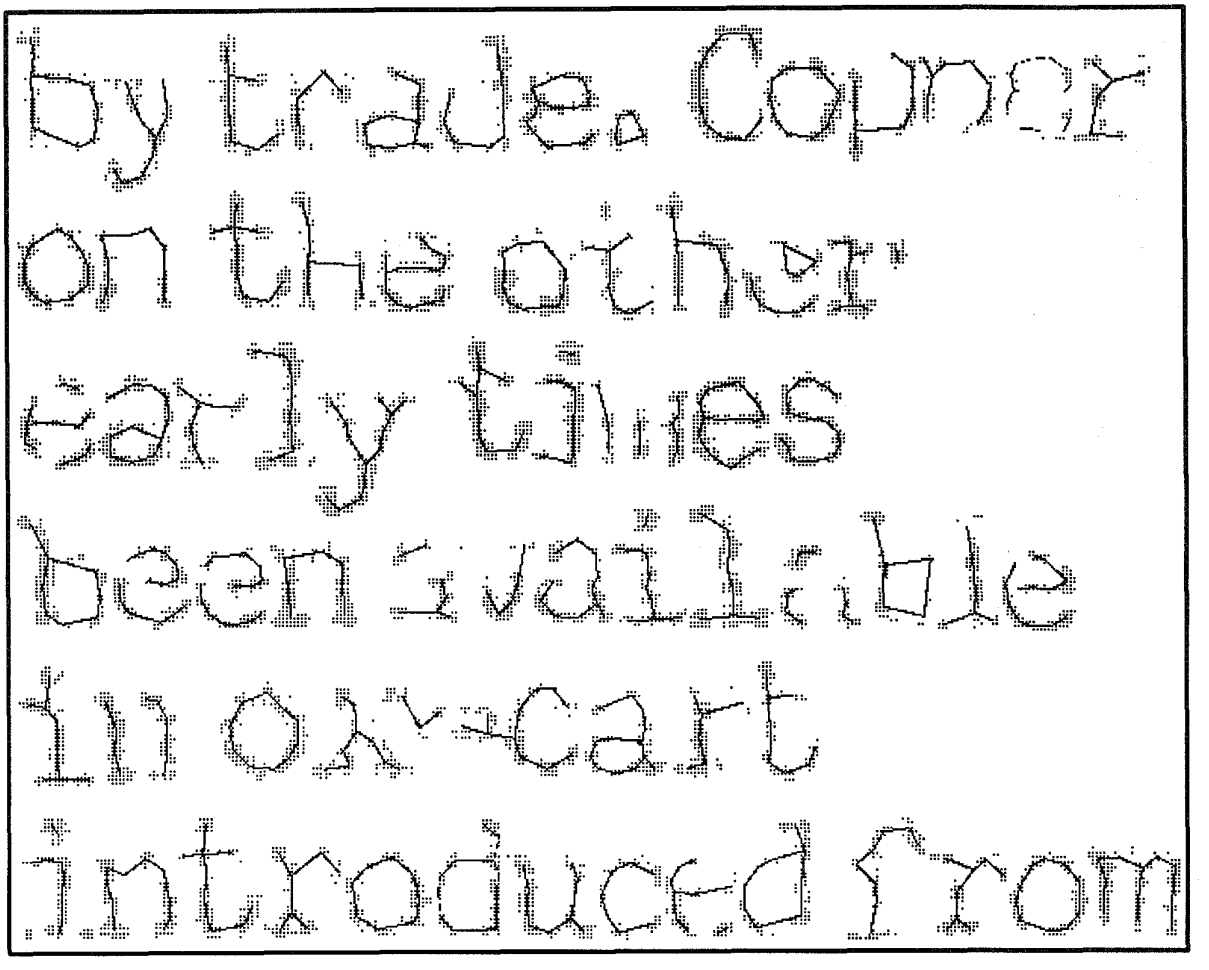


Figure 4: Skeletal shape extraction for select lines from the faded text in Figure 1.

[7] T. Hastie and W. Stuetzle. "Principal Curves". *Journal of the American Statistical Association*, 84(406):502–516, 1989.

[8] J. D. Hobby and T. K. Ho. "Enhancing Degraded Document Images via Bitmap Clustering and Averaging". In *Proceedings of the Fourth International Conference on Document Analysis and Recognition*, volume 1, pages 394–400, 1997.

[9] J. A. Kangas, T. K. Kohonen, and J. T. Laaksonen. "Variants of Self-Organizing Maps ". *IEEE Transactions on Neural Networks*, 1:93–99, 1990.

[10] S. P. Luttrell. "Derivation of a class of training algorithms". *IEEE Transactions on Neural Networks*, 1(2):229–232, 1990.

[11] T. Martinetz and K. Schulten. "Topology Representing Networks". *Neural Networks*, 7(3):507–522, 1994.

[12] F. Mulier and V. Cherkassky. "Self -Organization as an Iterative Kernel Smoothing Process". *Neural Computation*, 7(6):1165 –1177, 1995.

[13] M. Oguro, T. Akiyama, and K. Ogura. "Faxed Document Image Restoration Using Gray Level Representation". In *Proceedings of the Fourth International Conference on Document Analysis and Recognition*, volume 2, pages 679–683, 1997.

[14] G. Raza, A. Hennig, N. Sherkat, and R. J. Whitrow. "Recognition of Facsimile Documents Using a Database of Robust Features". In *Proceedings of the Fourth International Conference on Document Analysis and Recognition*, volume 1, pages 444–448, 1997.

[15] H. Ritter, T. Martinetz, and K. Schulten. *"Neural Computation and Self -Organizing Maps: An Introduction"*. Addison -Wesley, Reading, Massachusetts, 1992.

[16] R. Singh, V. Cherkassky, and N. P. Papanikolopoulos. "Determining The Skeletal Description Of Sparse Shapes". In *Proceedings of the IEEE International Symposium on Computational Intelligence in Robotics And Automation*, pages 368–373, 1997.