OFF-LINE CURSIVE HANDWRITING SEGMENTATION

Ke Han and Ishwar K. Sethi

Vision and Neural Networks Laboratory, Department of Computer Science
Wayne State University, Detroit, MI 48202

Abstract

The segmentation of off-line cursive handwriting is an important step in cursive handwriting recognition. In this paper, a new approach is described for this task. The suggested approach uses a set of heuristic rules to determine possible letter boundaries in the image of a cursive word. The heuristic rules are based on associations that exist between certain geometric and topologic features and the English language characters. A segmentation system incorporating the proposed approach has been built to perform segmentation on postal address images. The system includes several preprocessing steps to extract cursive handwritten words from a postal envelope and normalization steps to allow variations in pen thickness and tilt in writing. The experimental results obtained thus far show that the proposed approach is capable of accurately locating the letter boundaries in cursive words.

1. Introduction

Despite of successes achieved in character recognition over the last three decades, the recognition of off-line cursive handwriting, including numerals, characters and signatures, still remains a challenging problem. It is mainly due to the difficulties involved in the segmentation of cursive handwriting to isolate individual characters.

Since cursive handwriting segmentation is a difficult task in an off-line environment, the use of word-level recognition approach in place of common character-level approach has been suggested to avoid the segmentation problem altogether. While word-level recognition strategy does avoid the difficult segmentation issue, the approach is limited in its discrimination capability and is suitable only for limited vocabulary applications.

In the work reported in this paper, we describe a new cursive handwriting segmentation scheme. The suggested scheme uses a set of heuristic rules to determine possible letter boundaries in the image of a cursive word. The heuristic rules are based on associations that exist between certain geometric and topologic features and the English language characters. A segmentation system incorporating the proposed approach has been built to perform segmentation on postal address images. The system includes several preprocessing steps to extract cursive handwritten words from a postal envelope and normalization steps to allow variations in pen thickness and tilt in writing. The experimental results obtained thus far show that the proposed approach is capable of accurately locating the letter boundaries in cursive words.

The organization of the paper is as follows. Section 2 outlines the proposed handwriting segmentation approach. Section 3 presents an off-line cursive handwriting segmentation system built upon the proposed approach. A segmentation system using the proposed scheme for processing postal address images is described in Section 4 along with its performance. The concluding remarks are presented in Section 5.

2. Cursive Handwriting Segmentation Approach

At the outset, it should be noted that cursive handwriting segmentation is a task which provides best results if recognition is integrated with it and the whole process is carried out in an iterative fashion as indicated in Figure 1. To initialize the iterative segmentation process of Figure 1, a basic cursive word segmentor is required. The proposed segmentation scheme is aimed at this basic segmentor.

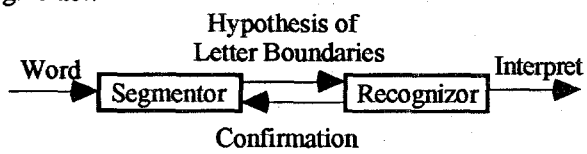


Figure 1. General Diagram of Character-level Handwriting Recognition System

The basic segmentor is based upon a set of associations that exist between a set of features and different groups of English language letters. From the graphological point of view, these features can be classified into two groups: regular and singular. The former includes geometric shapes such as horizontal bars, vertical bars and loops. The latter contains singularities and quasi-topologic features such as end points, branch points, crossing points, convex points,

0-8186-7128-9/95 $4.00

and concave points. The choice for these features is based on several previous works [1,2] where their use for describing handwriting has been clearly established.

In our segmentation scheme, the alphabet is divided into six basic character classes based on the presence of above features. These classes are:

Class 1: f, t, x.
Class 2: a, e, o, s.
Class 3: b, d, h, k, l.
Class 4: g, j, p, q, y, z.
Class 5: c, i, m, n, r, u, v, w.
Class 6: Initial Capital Character.

A character in Class 1 can be identified by a crossing point near the upper body line together with an ascender/descender and horizontal/vertical bars in the upper zone or lower zone. A center loop is the reliable indicator of a character in Class 2. In case the loop is not closed, a pair of convex point and concave point in the main body zone are reliable markers of the loop. A letter in Class 3 is indicated by an ascender in the upper zone and a couple of singular features in the main body zone. In addition, b or d should have a loop in the main body zone and in case the loop is not closed, the similar rule as in Class 2 will deal with it; h, k and l are captured by singular features in the upper zone and main body zone, and in most cases there is a small loop in the upper zone that will be helpful to identify them. The identification of characters in Class 4 can be carried out using the similar features as in Class 3 except that in Class 4 a descender is used instead of an ascender. All the characters in Class 5 can be segmented by analyzing curvature extrema. Based on the characteristics (upward, downward, leftward, or rightward), relationships and number of concavities, several extrema are combined together as a character of this class. If curvature extrema can not be detected because of writing styles, a sequence of pairs of end points and branch points are good indicators of characters in Class 5 and some of these points can be used to identify a instance in this class. The segmentation of initial capital characters is straightforward. They can be cut off at the middle of the first valid ligature or a certain distance from the beginning of a word based on what features are present.

In order to determine the letter boundaries based on this alphabet classification, several global coefficients, reference lines and zones of a word are calculated. The global coefficients include the estimated width of characters (EWC) and the estimated number of characters (ENC) in the word. The reference lines are upper line, upper body line, lower body line, and lower line. The reference zones are upper zone, main body zone and lower zone. When global and local information is available, the letter boundaries of the word are determined based on the distributions of all the features in the different zones and a set of heuristic rules. Finally, a sequence of physical segmentation points are located on ligatures between letters. These ligatures are provided by the off-line cursive handwriting tracing algorithm suggested by Lee and Pan [3].

3. Cursive Handwriting Segmentation System

The diagram of an off-line cursive handwriting segmentation system using the proposed approach is shown in Figure 2. The input of the system is a binary image of a cursive handwriting. Figure 3(a) shows an example input word. The output is a sequence of physical segmentation points.

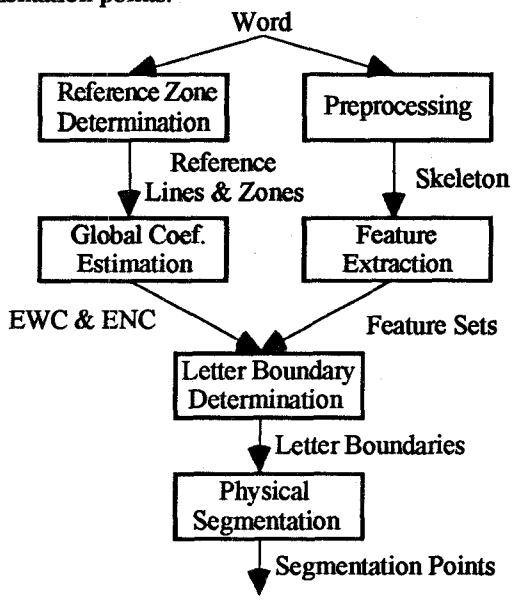
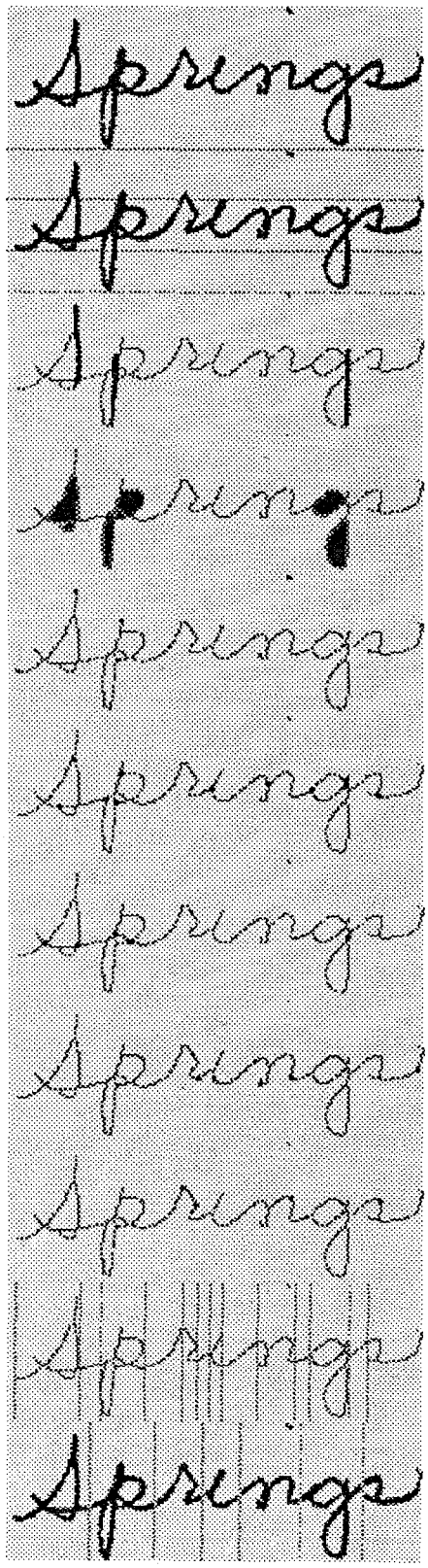


Figure 2. Diagram of Off-line Cursive Handwriting Segmentation System

The input word is processed through two paths. The first path consists of two stages. The first stage includes routines for finding reference lines and zones (Figure 3(b)). In the second stage, ENC and EWC are calculated based on the size of the input word and reference zones.

In the second path, the input word is preprocessed to generate the binary and skeleton images and then the regular and singular features are extracted from these two images. The regular features include horizontal bars, vertical bars and loops. The loops are extracted from the skeleton image using 4-connected component labeling algorithm. The horizontal and vertical bars are extracted from the binary image using the morphological hit-or-miss operation. The singular features include end points, branch points, crossing points, convex points, and concave points. To extract these features, the skeleton image is first decomposed into end points, branch points, crossing points, loops, and simple curves. The simple curves are the open curves without junction points. To extract convex points and concave points on the simple curves, the salient curve-point selection algorithm proposed by Fischler and Wolf [4] is used. Figure 3(c), (d), (e), (f), (g), (h), and (i) indicate the bars, loops, end points, branch points, crossing points, convex points, and concave points of the example input word, respectively.



(a) (b) (c) (d) (e) (f) (g) (h) (i) (j) (k)

Figure 3. A Series of Processing Results

In the letter boundary determination stage, global and local information are synthesized to generate a sequence of letter boundaries of the input word using the alphabet classification and a set of heuristic rules. The letter boundaries of the example input word are given in Figure 3(j). The principal heuristic rules are described below.

Loop determination. If there is a loop detected in the main body zone, then a character in Class 2 is identified and before and after the loop are two letter boundaries. If there is a loop detected in the upper zone, then after the loop is a letter boundary. If there is a loop detected in the lower zone, then before the loop is a letter boundary.

Crossing domination. A crossing point found in the upper zone is an indicator of a character in Class 1. A crossing point found in the main body zone indicates either a character in Class 1 or an end point of a ligature between two consecutive letters. A crossing point found in the lower zone is an end point of a ligature between two consecutive letters. In the case of a character in Class 1, two letter boundaries are determined according to EWC and features nearby. In the case of an end point, a letter boundary is produced near the crossing point.

Double t's splitting. If there are a loop and two crossing points found in the main body zone and/or the upper zone that are close together, double t's are located. The centroid of the loop is chosen as a cutoff point of these two t's that forms a letter boundary between two t's. The other two boundaries of two t's are on the opposite sides of the crossing points and are determined using Rule 2.

Dream team. A feature cluster in the upper zone and a feature cluster in the lower zone that are in a similar vertical position are markers of longer characters. According to the properties of relative features and EWC, proper letter boundaries may be determined.

Central masses. It is used to deal with characters in Class 5. In this case, a sequence of convex/concave feature points appear in the main body zone. Based on the characteristics and positions of these features and EWC, proper letter boundaries are determined.

Upward masses. If there are features in the upper zone and main body zone, they belong to characters in Class 3. Based upon characteristics and positions of these features and EWC, proper letter boundaries are determined.

Downward masses. If there are features in the lower zone and main body zone, they belong to characters in Class 4. Based upon characteristics and positions of these features and EWC, proper letter boundaries are determined.

Alignment. A couple of letter boundaries close to each other are aligned to the leftmost or rightmost one based on the features that produced these boundaries.

In the final stage, physical segmentation points are located at ligatures between letters of the input word. Here we consider a convex/concave point on a ligature as the best cutoff point and if there is no such a convex/concave point, the middle of a ligature is treated as the best cutoff point. The final physical segmentation points of the example input word is shown in Figure 3(k).

4. Application to Postal Address Images

A segmentation system incorporating the proposed cursive handwriting segmentation scheme has been applied to perform segmentation on postal address images. Since postal address images are gray level and contain several rows of words, the segmentation system first applies several preprocessing operations to extract these words and subsequently feed each of them into the proposed cursive handwritten segmentation system for segmentation. These preprocessing operations include thresholding, word extraction, word slant correction, and noise filtering.

Experiments have been performed on a database of 50 envelopes extracted from real mailpieces. These envelopes contain a total of 1119 cursive handwritten characters. For this database, the proposed segmentation system segmented 959 characters correctly, which corresponds to 85.7% accuracy rate. Some of the successful segmentation results are given in Figure 4.

Analyzing the failure cases produced by the proposed segmentation system, it is noticed that most errors are due to the inherent segmentation ambiguity of characters in Class 5. Other failures are generated by the writing defaults involved in the original input word. Some failure segmentation cases are shown in Figure 5.

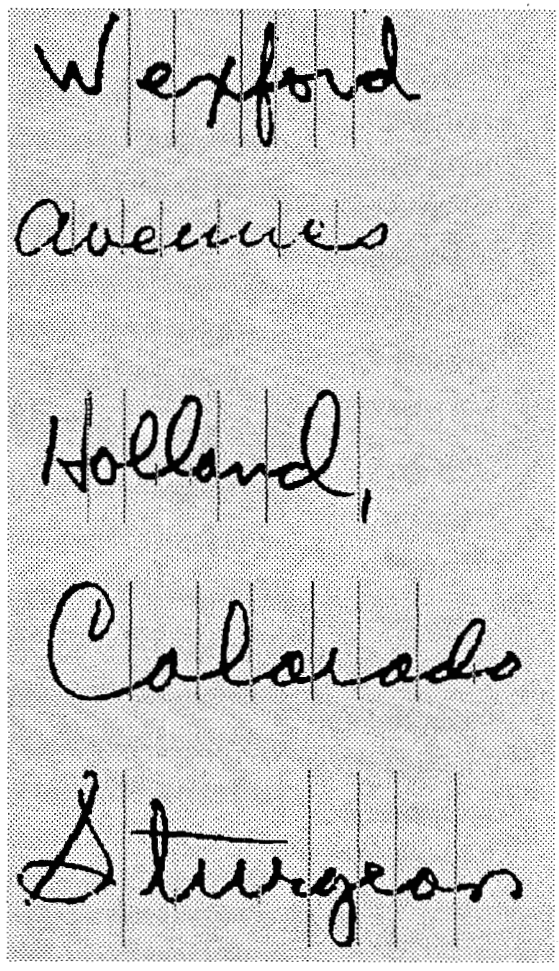


Figure 4. Some of Correctly Segmented Words

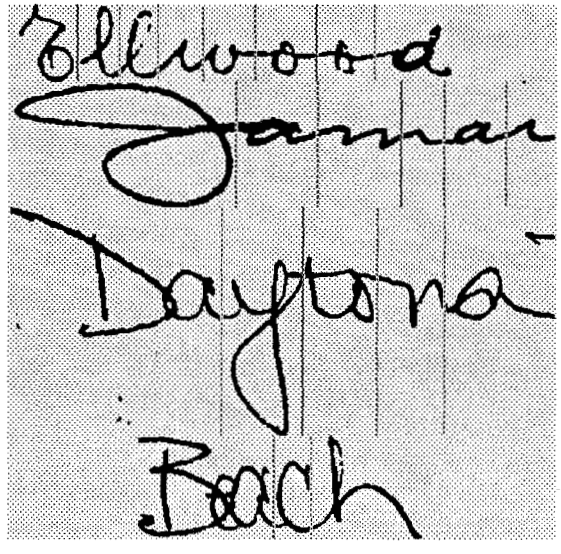


Figure 5. Some of Failure Cases Produced by System

5. Conclusion and Future Research Plan

In this paper, a new approach for off-line cursive handwriting segmentation has been described and a segmentation system incorporating the proposed approach has been built to perform segmentation on postal address images. The experimental results obtained by the segmentation system on postal address images show that the proposed approach is capable of accurately locating the letter boundaries in cursive words.

Obviously, there are a number of errors occurring in the segmentation, most of which have to do with the ambiguity of segmenting a sequence of specific letters. In order to break the difficulty, we think that combining word segmentation and character recognition is a better way to achieve a higher accuracy rate in a real situation and under a large vocabulary. This is currently being investigated.

References

1. H. Nishida and S. Mori, Algebraic description of curve structure, IEEE **PAMI-14**(5), 516-533 (1992).
2. J.J. Brault and R. Plamondon, A complexity measure of handwriting curves - modeling of dynamic signature forgery, IEEE **SMC-23**(2), 400-413 (1993).
3. S. Lee and J.C. Pan, Off-line tracing and representation of signatures, IEEE **SMC-22**(4), 755-771 (1992).
4. M.A. Fischler and H.C. Wolf, Locating perceptually salient points on planar curves, IEEE **PAMI-16**(2), 113-129 (1994).