Written Letter Analysis

by

Hriday Ravindranath

Introduction:

Written letter Analysis commonly referred to as Handwriting Analysis has been a popular subject of research due to its vast applications in fields like signature analysis, address reading, converting written text into hyper-text etc. There have been several complexities that arose during written letter analysis especially when there are no constrains on the text to be analysed.

Handwriting analysis can be broadly classified into 2 sub categories [1]:-

- **Online Recognition:**
  Here the user is directly connected to the system using an electronic pen or a touch-screen and recognition is carried out in real-time
- **Offline Recognition:**
  Here recognition is carried out on handwritten text that is captured using a scanner or a camera. Thus the text is treated as an image.

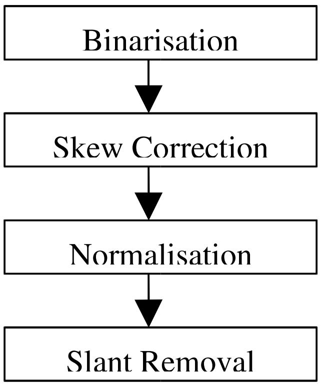
However both on-line and offline methods are similar in many respects except for the fact that offline methods does not have any temporal information. However using certain heuristic methods and prior knowledge (for example in the Roman script, writing is always performed from left to right) we can determine the direction of the strokes with respect to time.

Handwriting analysis over the years has been carried out in various ways depending on the script and language. For purpose of discussion a few methods for the Roman Script will be described below. But general classification methods that can be used for various scripts are discussed in [2, 3].

Method:

A. Preprocessing:

The first task is pre-processing which has 4 subtasks [1]:-



Step 1:

Binarisation:

After the image is capture by a sensor it is converted into its binary equivalent using various algorithms (see [4]). Usually before or after binarisation the image is filtered to remove noise.

Step 2:

Skew Correction:

Skew Correction methods (based on horizontal projection profiles, contours etc see [5,6,7] are used to align the paper document with the coordinate system of the scanner.

Step 3:

Normalisation:

Due to the variation of stroke, normalisation methods such as thinning or skeletonization are used. This normalises the width of each stroke to 1 pixel (see [8]).

Step 4:

Slant Removal:

Due to the fact that the slant of handwritten text varies from user to user, slant removal methods are used. Various methods are used depending on the type of task.

B. Recognition

The task of Handwriting analysis can be approached in 3 ways namely [1]

1. Recognising Isolated Characters
2. Isolated words
3. Unconstrained text of words

1. Recognition of isolated characters:

Recognition of isolated characters is the simplest of the 3 tasks. Here each character is assigned to one of the given classes. After preprocessing the image is subjected to feature extraction. There are several feature extraction methods that have been proposed such as moments, quantities derived from series expansion, structural features, component analysis etc.

After feature extraction methods, the feature vector is fed into a classifier. Classifier such as Bayes classifier can be used (see [9]). There are several classifiers and one such study that compares them are in [10].

2. Isolated Word Recognition:

One way of approaching word recognition is segmenting the word into isolated characters and then using the methods described above. However segmenting words can be extremely difficult. There are 3 approaches that can be used namely [1]

1. Holistic Method
2. Segmentation based approach
3. Segmentation-free approach

2.1. Holistic Method:

Here instead of segmenting the word into isolated characters, the word is classified as a whole entity using a dictionary. An order group of features (left to right) are loops, ascenders, descenders, face up and face down valleys and shape descriptors. The features of unknown input are matched against known prototypes (see [11]).

Limitations:

- Can not deal with a large number of classes
- Has to be used in conjunction with the other 2 methods.

2.2. Segmentation based approach:

This method segments the word into smaller units. Since there is no actual way of splitting the word into its individual characters with out prior knowledge of the word, over-segmentation is carried out. Here the image of the word is broken up into several entities called graphems [1]. After segmentation, all possible combinations of adjacent graphems are fed into a recogniser. The recogniser returns an ordered list of class names and a confidence for each class. Now depending on the confidence, the best sequence of characters matching the word is obtained. After the isolated characters are obtained, they are classified as described above. To improve efficiency a dictionary is used as a post-processing phase.

Limitations:

- Segmentation and Grapheme recombination are based on human intuition.

2.3. Segmentation-free Approach:

This approach is based on Hidden Markov Model (HMM) see [12]. The main advantage of using this model is that it can cope with the following problems:

- Noise and shape variations.
- Variable length of feature vectors.
- Segmentation problem.

This is due to the fact that HMMs are stochastic models. An individual HMM is constructed for each individual pattern class. For small vocabulary, an HMM is built for each word, but for larger vocabularies, HMMs are built for each character and character models are concatenated to word models.

3. Word Sequence Recognition:

In order to recognise a sequence of words, the input is segmented into words using various segmenting algorithms and then performing one of the recognition methods described above. The most common segmentation method is detecting spaces between connected components of the input. This task is relatively easier than segmenting words into characters. However in some cases segmentation can be difficult, hence HMM can be used for segmentation.

References:

[1] H. Bunke. Recognition of cursive roman handwriting – past, present and future. In Seventh International Conference on Document Analysis and Recognition, pages 448–461, 2003.

[2] H. Bunke and P. Wang, editors. Handbook of Character Recognition and Document Image Analysis. World Scientific, 1997.

[3] S. Mori, H. Nishida, and H. Yamada. Optical Character Recognition. John Wiley and Sons, Inc., 1999.

[4] L. O'Gorman and R. Kasturi, editors. Document Image Analysis. IEEE Computer Society Press, 1995.

[5] R. Bozinovic and S. Srihari. Off-line cursive script word recognition. IEEE Trans. on Pattern Analysis and Machine Intelligence, 11(1):68–83, Jan. 1989.

[6] S. Madhvanath and V. Govindaraju. Local reference lines for handwritten phrase recognition. Pattern Recognition, 32(12):2021–2028, 1999.

[7] M. Morita, J. Facon, F. Bortolozzi, S. Garnes, and R. Sabourin. Mathematical morphology and weighted least squares to correct handwriting baseline skew. In 5th Int. Conference on Document Analysis and Recognition, pages 430–433, 1999.

[8] C. Suen and P.Wang, editors. Thinning Methodologies for Pattern Recognition. World Scientific, 1994.

[9] U. Kressel and J. SchÂ¨urmann. Pattern classification techniques based on function approximation. In [13], pages 49–78.

[10] C.-L. Liu, H. Sako, and H. Fujisawa. Performance evaluation of pattern classifiers for handwritten character recognition. Int. Journal on Document Analysis and Recognition, 4:191–204, 2002.

[11] S. Madhvanath and V. Govindaraju. Local reference lines for handwritten phrase recognition. IEEE Trans. on Pattern Analysis and Machine Intelligence, (23):149–164, 2001.

[12] A. Kundu. Handwritten word recognition using hidden Markov model. In [13], pages 157–182.

[13] H. Bunke and P. Wang, editors. Handbook of Character Recognition and Document Image Analysis. World Scientific, 1997.