Available online at www.sciencedirect.com
SCIENCE DIRECT
Pattern Recognition 39 (2006) 1074–1087
PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY
www.elsevier.com/locate/patcog

Rule-based cleanup of on-line English ink notes

Zhouchen Lin[a,*], Rongrong Wang[b], Heung-Yeung Shum[a]

[a]*Microsoft Research, Asia, Zhichun Road #49, Haidian District, Beijing 100080, PR China*
[b]*Fudan University, Handan Road #220, Shanghai 200433, PR China*

Received 24 November 2004; accepted 9 August 2005

Abstract

Recently, many pen-based devices have enabled people to input digital ink naturally. Often, there is smear and correction when writing. This not only makes the document dirty and look unpleasant, but also affects the handwriting recognition when recognition is called for. As the first paper to address the ink cleanup problem, we present our ink cleanup system that removes the smear and correction, so that the document becomes cleaner and more legible and the handwriting recognition rate could also be improved. The algorithms are rule-based and are capable of dealing with the most common cases that may happen during writing, including self-overtracing of a single stroke, inter-overtracing between strokes, correction, touch-up, insertion and wrong writing order. Experimental results show that our system is effective in cleaning the ink note and is promising in increasing the recognition rate as well.

Keywords: Document analysis; Document and text editing; Handwriting analysis; Handwriting recognition

1. Introduction

Handwriting is the most important way to expand human memory and facilitate communication. With the increase of the computing power of computers, people are relying more on handwriting recognition technologies [1–6] to convert documents into texts and graphics so that the documents can better be stored, shared, retrieved, and so on. Recently, the flourish of mobile working, particularly the emergence of PDA, Tablet PC, etc., are enabling people to produce more and more handwritten words and even ink documents. When writing on such devices, it is common that people write something erroneous and then fix them. The resultant smear and correction not only make the words or document dirty and look unpleasant, but also decreases the recognition accuracy if such function is called for. Although both off-line and on-line handwriting recognition have been studied by many scholars during the past decades [3–6], to our best knowledge, we have not seen that the handwriting cleanup problem is addressed in the literature. Usually, the preprocessing of an on-line handwriting recognizer only includes data smoothing, signal filtering, dehooking and break corrections [4], etc. This may be partly because the cleanup problem is not very important in the past as there is still room for improving the recognition accuracy for those "clean" words. Another reason may be that in the past the smeared or corrected words were relatively few so that people did not take them seriously. The third reason may lie in the belief that the training process of the handwriting recognizer can automatically deal with the smearing and correction as long as these cases happen in the training samples. Unfortunately, such a belief is just a misconception because the deteriorated words account for relatively small portion of the training samples and thus will have little effect on the training. The fourth reason may be that cleanup is a much less severe problem in off-line handwriting recognition and people may convert on-line recognition to off-line recognition to bypass this problem. However, the accuracy of off-line recognition is usually lower than that of on-line recognition [3,4]. Nowadays, the increasing demand on higher recognition accuracy disallows such conversion. Therefore, we have

* Corresponding author. Tel.: +86 10 62617711x3143; fax: +86 10 88097306.
E-mail addresses: zhoulin@microsoft.com (Z. Lin), rrwang@fudan.edu.cn (R. Wang), hshum@microsoft.com (H.-Y. Shum).

0031-3203/$30.00
doi:10.1016/j.patcog.2005.08.013

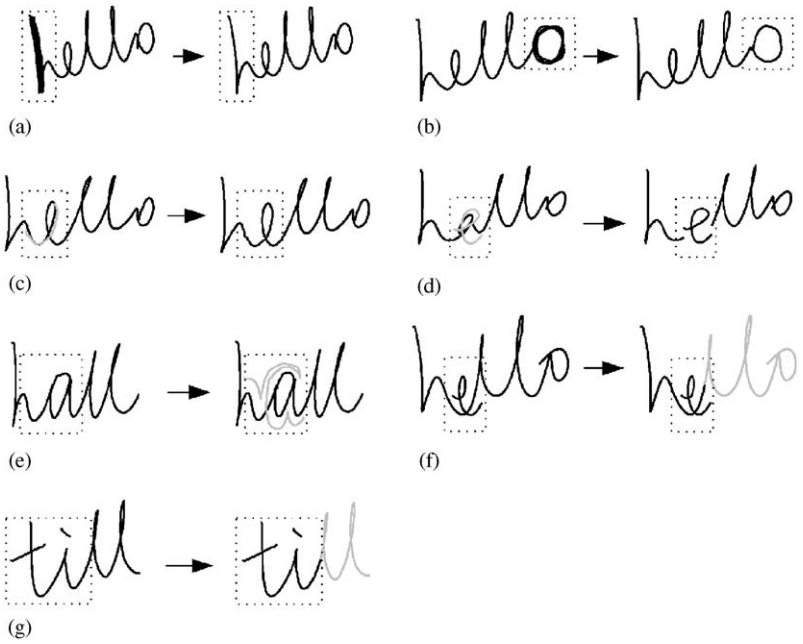

Fig. 1. The cases dealt with by our ink cleanup system and the corresponding desired results. (a) and (b) Self-overtracing by folding and looping, respectively. After cleanup, the overtracing parts are simplified. (c) Inter-overtracing. The overtracing parts are merged. (d) Correction. The background "a" is replaced by the foreground "e". (e) Touch-up. Two strokes become one stroke, with the writing order shown in grey curve on the right. (f) Insertion. "e" is inserted between "h" and "llo". The single stroke for "hllo" is broken into two strokes (black and grey stroke on the right). (g) Late stroke. The t-bar and i-dot are rearranged near to their stems. In the example, the t-bar is arranged before its stem because its stem is the first downward stroke piece of the stroke. And the i-dot is right after its stem. Note that the stroke containing the i-stem is broken into two strokes.

to consider the recognition with degraded words. And our another motivation is to make the on-line ink documents more readable. In this paper, as primary investigation, we seek to handle these two issues in a unified system.

Our system adopts rule-based approaches because the data collection is a hard problem. Although theoretically the deteriorated words can be collected from large amount of data set, detecting the words that we want is not trivial. For example, collecting words with incorrect writing order of strokes may require the visualization of the stroke order and heavy human examination. Moreover, data labeling is also hard, e.g., specifying what the cleaned strokes should be and the correct writing order for the cleaned words is not easy. Due to such difficulties, we have to apply complicated rules that are summarized from our observations to clean the "bad" words. As a result, our rule-based algorithms can only work on English handwriting. In addition, as we conceive our system as the preprocessing of general handwriting recognizers, we do not utilize any recognition results to assist our processing. Therefore, we can only analyze the geometric shapes of strokes and their relationship in detail. Finally, as the number of "bad" words are usually less than "good" words, we have to make the algorithms conservative so that those "good" words will not be processed.

Our system is designed to deal with the following kinds of word quality degradation (Fig. 1) that we believe are the most common when writing English words or documents:

- Self-overtracing, i.e., the folding and looping of a single stroke. The "duplicated" parts of a stroke will be simplified (Figs. 1(a) and (b)).
- Inter-overtracing, i.e., the overlapping of two strokes. The two strokes will be merged into one (Fig. 1(c)).
- Character correction, i.e., the replacement of characters. Some characters will be replaced by others that are written later at the same position (Fig. 1(d)).
- Touch-up, i.e., changing a character by adding a short stroke. The two strokes will be merged and the writing order within the merged stroke may be rearranged (Fig. 1(e)).
- Insertion, i.e., adding a missing character between two characters that are already written. The strokes will be inserted at their intended position so that the time order corresponds to their horizontal order (Fig. 1(f)). When the two characters around the missing character are written in one stroke, the stroke will be broken at a specific point.
- Late stroke, i.e., the dot and the bar of "i", "t", and so on, are not written right before or after their stems (Fig. 1(g)). The late strokes will be rearranged so that the temporal order complies with the spatial order. Their stems may be severed if they are connected to other characters.

The rest of this paper is organized as follows. Section 2 describes the pre-processing of the cleanup algorithms and