A Blackboard-based Approach to Handwritten ZIP Code Recognition

Jonathan J. Hull, Sargur N. Srihari, Ed Cohen, Leonard Kuan,
Peter Cullen and Paul Palumbo

State University of New York at Buffalo
Department of Computer Science
Buffalo, New York 14260
hull@cs.buffalo.edu

ABSTRACT

A methodology for recognizing ZIP codes (postal codes) in handwritten addresses is presented. The method uses many diverse pattern recognition and image processing algorithms. Given a high-resolution image of a hand-written address block, the solution invokes routines capable of hypothesizing the location of the ZIP Code, segmenting and recognizing ZIP Code digits, locating and recognizing City and State names, and looking-up the results in a dictionary. The control structure is not strictly sequential, but in the form of a black-board architecture that opportunistically invokes routines as needed. An implementation of the methodology is described as well as results with a data-base of grey level images of handwritten addresses (taken from live mail in a U.S. Postal Service mail processing facility). Future extensions of the approach are also discussed.

1. INTRODUCTION

Determining the destination postal code from a handwritten address is a problem of great importance to automated mail processing. Presently, 15 percent of first class mail handled by the United States Postal Service (USPS) is handwritten (either hand-printed or cursive) [6]. While this percentage may seem small, it translates into nearly ten billion mail pieces annually [7]. Less than five percent of this is handled by present optical character recognition (OCR) systems, thus leaving a huge volume of mail that must be handled by costly semi-automatic or manual processes.

Handwritten addresses have been automatically sorted in some countries (particularly Japan) by requiring that the postal code be printed in boxes at known locations on a mailpiece. This is sometimes accompanied by examples of how digits should be printed. This makes the problem much simpler because the postal code is easily located, the digits are separated, and the digits are formed with fewer variations.

In the United States, the solution based on placing stringent requirements on the address has been ruled out for several reasons (e.g., public acceptability). Instead, USPS has required that approaches be developed to read ZIP Codes in completely unconstrained handwritten addresses. This makes the problem much more difficult.

2. SYSTEM DESIGN AND CONTROL STRUCTURE

A blackboard model is a problem solving system which consists of three major components: knowledge sources, blackboard data structure, and control procedures [2]. The blackboard data structure holds the current solution state of the blackboard. The control procedures invoke the knowledge sources in an opportunistic manner to update the solution state of the blackboard. Blackboard systems can coordinate many types of knowledge sources, where each knowledge source can contribute its information at appropriate times.

For the Handwritten ZIP Code Recognition System (HZRS), the blackboard model provides flexibility, a coordinated method of integrating different knowledge sources, and a means of hierarchical data organization. HZRS requires that many different levels of knowledge be accessed while processing the image. Even in thresholding the image (normally considered a low-level process), higher level information (such as the expected number of lines, the expected size of characters, etc.) can be used to determine the success of the thresholding step. Many other steps in the HZRS also require access to different levels of information. The organization of the blackboard in our system is represented in Figure 1.

The control flow is shown in Figure 2 and is described below. The input to HZRS is assumed to be a high-resolution (300 ppi) image of a correctly oriented address block. The first processing

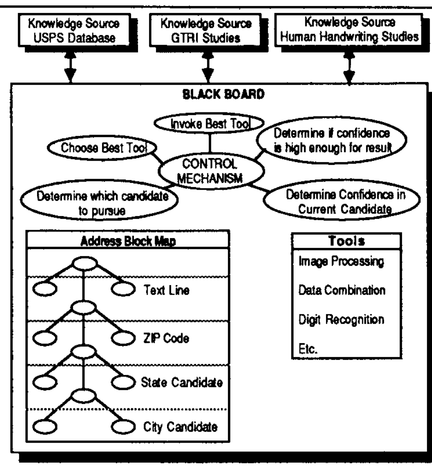

Figure 1. Organization of blackboard structure. The organization is divided into four separate areas: external information (top), control structure (middle), information relevant to the current address block (lower left), and available routines (lower right).

Int. Conf. on Pattern Recognition (ICPR), Rome, Italy, Nov. 14-17, 1988, 111-113.

step of HZRS is thresholding the image. An algorithm similar to that discussed in [3] is used in HZRS. Then, a specialized routine determines whether the image is machine-printed or handwritten. If an address is determined to be machine printed, the control structure will either process the image (with a simpler algorithm) or indicate that the image should be sent through a different process.

If the image is handwritten, the image will be segmented into text lines. If the text lines are not easily segmentable, rethresholding may be necessary. If the line segmentation is successful, the text lines are further segmented into words and characters. Again, rethresholding can occur if this segmentation step is not successful. Next, each character is identified as either a text character or a digit. This identification will not be completely reliable, so each character identification will have an associated confidence value.

Using the character identifications, text line segmentation, and word segmentation, a map of the address block is developed. The map approximates the contents of each line (i.e., the top line may contain a proper name, and the second line may contain a street address, etc.). The mapping process is not completely reliable, so each text line may have several content interpretations, and each interpretation has an associated confidence value.

Based on this mapping, a likely ZIP Code candidate is chosen from the text line with the highest confidence of containing a ZIP Code. The ZIP Code candidate is segmented into digits. If the segmentation does not produce the expected number of digits (for valid USPS ZIP Codes), one of several corrections is made: a new ZIP Code candidate is chosen, the current ZIP Code candidate is merged with another word, or the current ZIP Code candidate is split into 2 or more words. If a ZIP Code candidate is likely to be (or not to be)

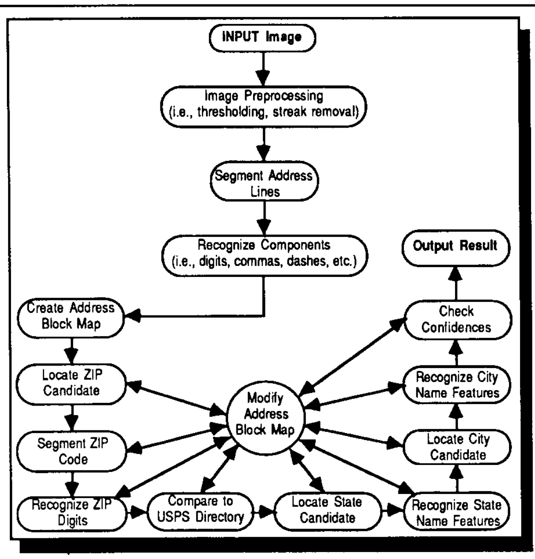

Figure 2. Control Flow of the HZRS. The graph shows the many ways control can flow depending on the requirements for a given address block image.

a ZIP Code, the address block map is modified to reflect a change in confidence values. Corrections and resegmentations are repeated on ZIP Code candidates until either a potential ZIP Code is found (and segmented) or it is determined that no likely ZIP Code candidate exists. Once a likely ZIP Code candidate is found and its digits are identified, other information from the image can be used, such as state and city names.

A likely state candidate is chosen from the group of segmented words. The state candidate is identified and assigned a confidence value. As with the ZIP Code, a poor confidence value may result in either choosing a new state candidate, merging state candidates, or splitting the state candidate to get a more likely state candidate. State names have the added difficulty that they may contain more than one word. A similar process is applied to recognize the City name. Note that as different candidates in the address block are determined to be (or not to be) State name, City names, etc., the address block map is modified to reflect these findings.

If a reasonable mapping of the address block is determined, and the City, State, and ZIP Code information is deemed sufficiently consistent, the ZIP Code search is complete and the destination address is reported. Otherwise, any of the steps in acquiring the City name, the State name, or the ZIP Code may be repeated.

3. ZIP CODE RECOGNITION

HZRS recognizes a ZIP Code by segmenting it into isolated digits and recognizing them. This approach is usually most successful for machine-printed text where the boundaries for characters are either well-determined or can be guessed with a reasonably high degree of accuracy. However, this approach is less successful in unconstrained handwritten addresses where the text can be cursive or printed and the digits can be touching.

Our approach to ZIP Code segmentation is similar to that discussed in [5]. In this method it is assumed that the number of digits present in the image is known. Simple connected component analysis is used to attempt to locate that number of digits. If this fails, the components are analyzed and those that contain more than one digit are segmented. Figure 3 shows an example of this process.

Our method for digit recognition uses a *parallel* approach in which three algorithms are applied to each image and the results are combined to reach a decision. A similar approach has been used before where algorithms are applied hierarchically [4]. Our method

Figure 3. Example of ZIP code segmentation

Int. Conf. on Pattern Recognition (ICPR), Rome, Italy, Nov. 14-17, 1988, 111-113.

uses three algorithms: (i) a template matching algorithm to make a decision based on the overall holistic characteristics of the image, (ii) a mixed statistical and structural classifier used on features extracted from the contour of the character, and (iii) a structural classifier used on information about the size and placement of strokes in the image. An example of stroke decomposition and the features extracted from a numeral is shown in Figure 4. These three algorithms were chosen because they utilize different types of information in the image and thus have a better chance of compensating for each others' weaknesses.

The results of the three digit classifiers are combined with a decision tree. The tree contains seven tests, an example of which is "If the decision of each classifier agrees, then output that decision." The performance of the classifier was tested with a training set of 1754 digits and a distinct testing set of 8129 digits. 91.0 percent of the testing set was correctly classified, there was an error rate of 1.7 percent and a reject rate of 7.3 percent.

4. EXPERIMENTAL RESULTS

Experiments have been conducted with a preliminary version of the system discussed in this paper. A complete system is under development. The current version of the system contains the following image processing routines:

1. Thresholding;
2. Testing for machine or handwritten text;
3. Horizontal line removal
4. Text line segmentation;
5. Text word segmentation;
6. ZIP Code location;
7. ZIP Code segmentation;
8. Parallel digit recognition algorithm;
9. ZIP Code dictionary lookup;

The system has been tested on 2000 grey level images of handwritten address blocks that were taken from live mail [1]. 31 percent of the ZIP Codes were read correctly, 9.4 percent of the images did not contain a ZIP Code and were rejected. Thus, 40.4 percent of the pieces were correctly processed. An erroneous ZIP Code was assigned to 2.3 percent of the pieces and 57.3 percent contained a valid ZIP Code and were rejected.

Figure 4. An example of the stroke decomposition and feature description of the numeral "3".

5. CONCLUSIONS

A complete system for handwritten address ZIP code recognition was described. A blackboard based control structure was established for the system, and several component image processing algorithms that can be invoked by the control structure were presented.

Future work on this approach will include continued development of the control structure including an improved inference technique. Several image processing algorithms still have to be incorporated, including better line and word segmentation. Also, a technique for handwritten word recognition is needed to improve the performance of City and State name recognition.

References

1. J. J. Hull, S. N. Srihari, E. Cohen, C. L. Kuan, P. Cullen and P. Palumbo, "A blackboard-based approach to handwritten ZIP Code recognition," *United States Postal Service Advanced Technology Conference*, Washington, D.C., May 3-5, 1988, 1018-1032.
2. H. P. Nii, "Blackboard Systems: The Blackboard Model of Problem Solving and the Evolution of Blackboard Architectures," *The AI Magazine*, Summer, 1986, 38-53.
3. N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man, and Cybernetics SMC-9*, 1 (January, 1979), 63-66.
4. A. Perez and C. Y. Suen, "Computer recognition of totally unconstrained handwritten ZIP codes," *International Journal of Pattern Recognition and Artificial Intelligence 1*, 1 (March, 1987), 1-15.
5. M. Shridhar and A. Badreldin, "Recognition of isolated and simply connected handwritten numerals," *Pattern Recognition 19* (1986), 1-12.
6. S. N. Srihari, J. J. Hull, P. W. Palumbo and C. Wang, "Automatic address block location: Analysis of images and statistical data," 86-11, Department of Computer Science, State University of New York at Buffalo, April, 1986.
7. USPS, Annual Report of the Postmaster General, 1983.

Acknowledgment

This work was supported by the Office of Advanced Technology of the United States Postal Service. We wish to thank Dr. Tim Barnum and Gary Herring of USPS and Dr. John Tan of Arthur D. Little, Inc. for their guidance and encouragement.

Int. Conf. on Pattern Recognition (ICPR), Rome, Italy, Nov. 14-17, 1988, 111-113.