A Duplicate Chinese Document Image Retrieval System

Yung-Kuan Chan
National Chung Hsing University, Taiwan, R.O.C.

Yu-An Ho
National Chung Hsing University, Taiwan, R.O.C.

Hsien-Chu Wu
National Taichung Institute of Technology, Taiwan, R.O.C.

Yen-Ping Chu
National Chung Hsing University, Taiwan, R.O.C.

INTRODUCTION

An optical character recognition (OCR) system enables a user to feed an article directly into an electronic computer file and translate the optically scanned bitmaps of text characters into machine-readable codes; that is, ASCII, Chinese GB, as well as Big5 codes, and then edits it by using a word processor. OCR is hence being employed by libraries to digitize and preserve their holdings. Billions of letters are sorted every day by OCR machines, which can considerably speed up mail delivery.

The techniques of OCR can be divided into two approaches: template matching and structure analysis (Mori, Suen & Yamamoto, 1992). The template matching approach is to reduce the complexity of matching by projecting from two-dimensional information onto one; the structure analysis approach is to analyze the variation of shapes of characters. The template matching approach is only suitable for recognizing printed characters; however, the structure analysis approach can be applied to recognize handwritten characters.

Several OCR techniques have been proposed, based on statistical, matching, transform and shape features (Abdelazim & Hashish, 1989; Papamarkos, Spilioties & Zoumadakis, 1994). Recently, integrated OCR systems have been proposed, and they take advantage of specific character-driven hardware implementations (Pereira & Bourbakis, 1995). OCR generally involves four discrete processes (Khoubyari & Hull, 1996; Liu, Tang & Suen, 1997; Wang, Fan & Wu, 1997):

1. separate the text and the image blocks; then finds columns, paragraphs, text lines, words, and characters;
2. extract the features of characters, and compare their features with a set of rules that can distinguish each character/font from others;
3. correct the incorrect words by using spell checking tools; and
4. translate each symbol into a machine-readable code.

The duplicate document image retrieval (DDIR) system transforms document formatted data into document images, then stores these images and their corresponding features in a database for the purpose of data backup. The document images are called duplicate document images. When retrieving a duplicate document image from the database, users input the first several text lines of the original document into the system to create a query document image. Then the system figures out the features of the image, and transmits to the users the duplicate document image whose image features are similar to those of the query document image (Nagy & Xu, 1997).

Some approaches have been proposed for the DDIR system. Doermann, Li, and Kia (1997) classified and encoded character types according to the condition that four base lines cross each text line, and uses the codes as the feature of the document image. Caprari (2000) extracted a small region from one document, assigned this region to the template (signature generation), and then scanned this template over a search area in another document. If the template also appears in the second document (signature matching), the two documents are classified as duplicates. Angelina, Yasser, and Essam (2000) transformed a scanned form into a frameset composed of a number of cells. The maximal grid encompassing all of the horizontal and vertical lines in the form is generated; meanwhile, the number of cells in the frameset, where each cell was created by the maximal grid, was cal-

culated. Additionally, an algorithm for similarity matching of document framesets based on their grid representations is proposed too. Peng, Long, Chi, and Siu (2001) used the size of each component block containing a paragraph text image in a duplicate document image and its relative location as the features of the duplicate document image.

The approaches mentioned previously are only suitable for stating the characteristics of an English document image. The characteristics of Chinese characters are quite different from those of English ones, and the strokes and shapes of Chinese characters are much more complicated than those of English characters. Chan, Chen, and Ho (2003) provided a line segment feature to represent a character image block and presented a duplicate Chinese document image retrieval (DCDIR) system based on this feature. The purpose of this short article is to give a brief overview of the duplicate Chinese DDIR systems.

BACKGROUND

Traditional information retrieval methods use keywords for textual databases. However, it is difficult to describe an image using exact information, and defining manually keywords is tedious or even impossible for a large image database. Moreover, some non-text components cannot be represented in a converted form without sufficient accuracy. One solution is to convert a document into digital images; meanwhile, some methods are applied to extract the features of the images. Based on the feature, some document images with database satisfying query requirements are returned.

A duplicate document image retrieval (DDIR) system has to own the following properties (Doermann, Li, & Kia, 1997):

- **Robust:** The features should be reliably extracted even when the document becomes degraded.
- **Unique:** The extracted features can distinguish each document image from others.
- **Compact:** The storage capacity required to hold the features should be as small as possible.
- **Fast:** The system needs a quick response with an answer to the query.
- **Scalable:** As more documents are processed, the size of the database could grow to tens of millions.
- **Accurate:** The system should accurately response with an answer, which satisfies the query requirement.

Unfortunately, many DDIR systems are vulnerable to poor qualities of document images, such as the scale, translation, rotation, and noise variants. Because of different resolution setup of a scanner, the same image may be scanned to become two images with different sizes. We call this phenomenon the scale variant. When an image is added with a great amount of noises, it may be regarded as a different image from the original one. It is named a noise variant image of the original one. In a particular document, images with rotation and translation variants may be generated owing to placing the document on different orientation angles or on different positions on a scanner. The variants mentioned previously will cause many troubles in feature extracting and image matching stages. They should be removed in advance.

A CHINESE DDIR SYSTEM

Many techniques about the DDIR system have been proposed (Caprari, 2000; Doermann, Li, & Kia, 1997; Peng, Chi, Siu, & Long, 2000; Peng, Long, Chi, & Siu, 2001). Since an English document mostly consists of approximately **70** commonly-used characters which contain **52** uppercase as well as lowercase English letters and punctuation marks, the classification and encoding procedure based on the feature of these characters' font types are possible. However, these techniques are only suitable for duplicate English document images, but not for duplicate Chinese document image retrieval (DCDIR) because the number of different Chinese characters is about **45,000**. What is more, the shapes of Chinese characters are complex, and many different characters have similar shapes to each other. Hence, there are several major problems with Chinese character recognition, that is, Chinese characters are distinct and ideographic, the size of a character is large, and there exist many structurally similar characters (Amin & Singh, 1996; Chan, Chen, & Ho, 2003).

It is necessary to develop a feature offering an excellent identification capability to classify Chinese characters by only using a little extra memory space. To reduce the extra memory space, it is feasible to segment a duplicate document image into blocks, each of which contains a set of adjacent characters, and then to extract the features from the blocks. Since the number of the blocks in a duplicate document image is much smaller than that of the characters in an identical duplicate document image, the feature dimensions are reduced greatly; however, its identification capability is lessened.

I. DCDIR System

The proposed duplicate document image retrieval system approximately includes three parts — image preprocessing, database creation, and document retrieval. This section will introduce these three parts in details.

A. Image Preprocessing

When scanning a document to generate a duplicate document binary image, the position of the document on the