A Keyword Spotting System of Korean Document Images

Il-Seok Oh*, Yoon-Sung Choi*, Jin-Ho Yang*, and Soo-Hyung Kim**
* Department of Computer Science, Chonbuk National University, Korea
** Department of Computer Science, Chonnam National University, Korea

Abstract

This paper presents a keyword spotting system for Korean document images. A two-stage retrieval scheme is proposed aiming at accelerating the retrieval process. In the first stage, a very coarse feature set is used to select the candidate words very fast, and the second stage chooses the relevant words among the candidate words using a fine feature set. The paper presents a prototype system including the document image analysis module and the keyword spotting module. Preliminary experimental data show that the system is promising both in speed and retrieval rates.

I. Introduction

Once the old paper documents have been scanned and stored in a database, the most important services to be provided to the users by the digital libraries is the retrieval capability. The retrieval service for the scanned document images should be at the same or at least conceptually similar level as the one for the electronically generated documents for the seamless contents service [Oh02]. For the Western document images, lots of researches have been done [Doermann98, Mitra00] and some systems are available [SCRIB02].

In Korea, the NDL (National Digital Library) project have been largely funded, and as a result enormous amount of old paper document are in the database in a digital image format and are serviced to Korean people via Internet [NDL02]. However, it appears that the advanced services like the full-text retrieval are not attempted for the document images.

This paper presents a keyword spotting system of the Korean document images. A two-stage coarse-to-fine retrieval scheme was devised and tested using a coarse profile feature set and a fine wavelet feature set. A prototype system including the document image analysis module and the keyword spotting module has been developed and evaluated. A reasonable retrieval rate has been obtained and about 6 times speedup was possible by the two-stage scheme.

II. Feature Extraction and Word Image Matching

A word image is segmented into character images by analyzing the vertical projection. After size-normalizing the character image into 32*32 array, two kinds of features are extracted. To obtain the first
feature set, four profiles are computed from top, bottom, left, and right sides. Each profile is represented as a one-dimensional array with 32 values. By averaging the 32 values, a coarse 4-dimensional feature set is obtained.

For the second feature set, the character image is wavelet-transformed using Harr operator. The effectiveness of Harr wavelet features for the task of Korean word image retrieval was demonstrated in [Kim01]. The wavelet coefficients are taken as features. Since the coefficients with large values are more significant in representing the original image, only a small set of the $K$ largest coefficients are used as the features. By doing so, a $K$-dimensional feature vector is obtained and it is used as a fine feature set. In our experiment, $K$ was 30.

Using the coarse and fine feature sets, a two-stage retrieval scheme can be implemented. Figure 1 illustrates the scenario where a digital library archive has a huge amount of word images. We assume the page images are already segmented into word images.

![Figure 1. Two-stage retrieval scenario.](image)

The matching between the query and document image is accomplished by sliding $n$ character-wide window one character forward at a time over the segmented text document where $n$ represents the number of characters in the query word. In the first stage, the coarse matching is performed using the 4-dimensional profile feature set. This stage aims at selecting candidate words in a very fast speed. The matching is accomplished using the Euclidean distance between the query (Q) and target (T) word images.

In the following formula, $q^i$ and $t^i$ represent 4-dimensional profile feature vectors.

First stage:

- Query word: $Q = (q^1, q^2, \ldots, q^n)$ and $q^i = (x_{1i}, x_{2i}, x_{3i}, x_{4i})$
- Target word: $T = (t^1, t^2, \ldots, t^n)$ and $t^i = (y_{1i}, y_{2i}, y_{3i}, y_{4i})$
- Distance between characters: $d^i = \left(\sum_{k=1}^{4} (x_{ki} - y_{ki})^2\right)^{1/2}$
- Distance between Q and T: $d^\text{av} = \sum_{i=1}^{n} d^i / 4$
- Matching rule: $d^i < p_1$, $1 \leq i \leq 4$, and $d^\text{av} < p_2$

The matching for the second stage is accomplished in a similar manner. The modifications are from the 4-dimensional profile features to 30-dimensional wavelet features. The threshold parameters for the matching condition are $p_3$ and $p_4$. The parameters in the matching conditions must be properly set to get a good retrieval rates. The $p_1$ and $p_2$ must satisfy a high recall rate while they may allow a low precision. High rates both in recall and precision must be produced by $p_3$ and $p_4$. They were set empirically in our
experiments.

III. System Implementation and Evaluation

A prototype system was implemented using C++ language on IBM-PC platform. The page image layout analysis module developed by Chonnam National University [Kwang01] was installed into the system. Figure 2 illustrates the system where Figure 2(a) shows the output from the page image analysis module and Figure 2(b) the spotted words.

An evaluation was performed using 238 page images scanned from technical journals in Korean language. The page images are in fairly good printing quality and the scanning was done in 300 dpi binary tone. The speed was 199,918 words per second. The retrieval rate was 99.6% recall and 97.5% precision. When one-stage processing with 30-dimensional wavelet feature set was applied, similar recall and precision were obtained and the speed was 33,063 words/second. Based on these experimental data, a conclusion can be drawn that the two-stage scheme accelerates the retrieval module by about 6 times without losing the retrieval rate.

Another evaluation was performed on the medium-quality document images scanned in 200 dpi from Journal of Korean Information Science Society. The recall and precision were 91.0% and 88.0%, respectively.

IV. Conclusions

A prototype system for keyword spotting in Korean document images was developed and a preliminary evaluation was presented. The system can be used either for the on-line keyword spotting of a reasonable volume of archive or for the pre-indexing of large-volume archive. Future works include the
retrieval rate improvement using a feature selection algorithm and robust algorithm development for low-quality document images.

References


