Content-Oriented Categorization of Document Images

Takehiro Nakayama
FX Palo Alto Laboratory, Inc.
3400 Hillview Avenue
Palo Alto, CA 94304 USA
nakayama@pal.xerox.com

Abstract

We have developed a technique that categorizes document images based on their content. Unlike conventional methods that use optical character recognition (OCR), we convert document images into *word shape tokens*, a shape-based representation of words. Because we have only to recognize simple graphical features from image, this process is much faster than OCR. Although the mapping between word shape tokens and words is one-to-many, they are a rich source of information for content characterization. Using a vector space classifier with a scanned document image database, we show that the word shape token-based approach is quite adequate for content-oriented categorization in terms of accuracy compared with conventional OCR-based approaches.

1 Introduction

The number of documents available on the network is increasing with the development of the computational infrastructure. Accordingly, information retrieval has become one of the most important research topics in natural language processing (NLP). In the digital network world, documents are usually distributed in either text file or image format, where the former is a sequence of character codes (e.g., ASCII) and the latter is a bitmap. Although only text files are machine-readable and convenient from the viewpoint of information retrieval, many documents are available as images alone. They are easily generated by scanning hard-copy documents which the real world is massively using.

While most information retrieval systems have been designed for text files, there are some systems proposed for images. They convert images into text files using optical character recognition (OCR) to utilize existing NLP techniques. Even though state-of-the-art OCR creates noisy output with recognition errors (Rice, *et al.*, 1995), prior work has shown that OCR output is satisfactory for retrieval purposes (Ittner, *et al.*, 1995; Mittendorf, *et al.*, 1995; Myers and Mulgaonkar, 1995; Wenzel and Hoch, 1995). The inaccuracy of OCR can be largely mitigated. However, little attention has been paid to reducing the computational expense of OCR. OCR is a major bottleneck for information retrieval systems in terms of speed. For example, Myers and Mulgaonkar reported in their OCR-based information extraction system that the total processing time was dominated by character and word recognition processes (Myers and Mulgaonkar, 1995). This suggests an important question: *"how much NLP can be done without character recognition* (Church, *et al.*, 1994)*?"*

As an alternative technique to OCR, there is *word shape token processing* which converts images into a shape-based representation. It recognizes coarse character shape classes (*character shape codes*) rather than character codes. Because the number of character shape codes is small and they are defined by simple graphical features, their recognition from images is inexpensive. Word shape token processing has been proven to be of use for European language identification (Nakayama and Spitz, 1993; Sibun and Spitz, 1994). Also, its feasibility for content characterization has been discussed with the use of controlled (noise-free) on-line data set (Nakayama, 1994; Nakayama 1995; Sibun and Farrar, 1994). However, no analysis has been done with real document images, which are usually degraded in quality. In addition, a comparative evaluation between the word shape token-based and the OCR-based approach is needed.

We have developed a technique which automatically categorizes document images into pre-defined classes based on their content. It employs a vector space classifier drawn from many robust statistical techniques in information retrieval (see Salton, 1991). We show in this paper that our technique can categorize as accurately as the conventional OCR-based approach, while it can process much faster.

In the next section, we describe the definition of character shape codes and word shape tokens, and their generation from document images. In section 3, we outline the automated categorization system which we developed. In section 4, with the use of a topic-tagged document image database, we show the word shape token-based approach is quite adequate for content-oriented categorization in comparison with a conventional OCR-based system. In section 5, we discuss the experimental results and future work.

2 Character Shape Code and Word Shape Token

A character shape code is a machine-readable code which represents a set of graphically similar characters. A word shape token is a sequence of one or more character shape codes which represents a word. Character shape codes are defined differently by the selection of graphical features. In this paper, we consider the number of connected components, vertical location, and deep concavity as graphical features to classify characters. First, we identify the positions of the text lines as shown in figure 1. Second, we identify the character cells, and count the number of connected components in each character cell. Third, we note their position with respect to the text lines. Finally, we identify the presence of a deep eastward/southward concavity. In figure 1, vertical location classifies characters into three groups—{"l"} {"g"} {"a", "n", "u", "e"}; characters that occupy the space between the top and the baseline, characters that occupy the space between the x-height line and the bottom, and characters that occupy the space between the x-height line and the baseline, respectively. The last one is further classified by presence or absence of a deep eastward/southward concavity. Resultant groups are {"a", "u"} {"e"} {"n"}.

The defined character classes and the members for the ASCII character set are shown in Table 1. Once classification has been performed, the resulting character shape codes are grouped by word boundary and used as word shape tokens for the downstream processing. Figure 2 gives an example of generated word shape token representation with its original document image.

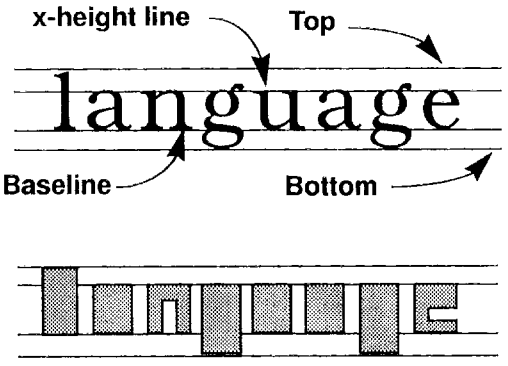

Figure 1: text line parameter positions (above) and connected components (below)

Table 1: character shape code membership

character shape code	members
A	A-Zbdfhklt0-9#$&@
x	amorsuvwxz
e	ce
n	n
i	i
g	gpqy
j	j
'-.:= !|	'"-,.:;=!? () /<> [] { } |

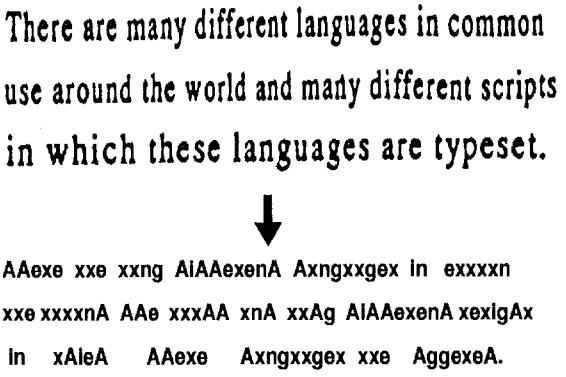

Figure 2: document image (above) and generated word shape tokens (below) note: there is an error (many - xxAg) in the second line due to a small ink drop

Our character shape code recognition doesn't require a complicated image analysis. For example, distinguishing "c" from "e" is a difficult task for OCR that requires a considerable computational expense (Ho and Baird, 1994), whereas they are in the same class in our representation (Table 1). Also, our process is free from font identification which is mandatory for OCR (for font identification complexity, see Zramdini and Ingold, 1993). As a result, the process of word shape token generation from images is much faster than current OCR technology.

While we save a computational expense, we lose some information which original document images have. Table 1 shows that the mapping between character shape codes and original characters is one-to-many—we use only seven character shape codes {**A x e n i g j**}[1] to represent all alphabetical characters.

1. We use **boldface** to represent the character shape codes.

This would seem to be very ambiguous. However, when used for mapping between word shape tokens and original words, the ambiguity is much reduced. We show this using a lexicon of 122,545 distinct word (surface-form) entries. When we transformed the lexicon into word shape token representation, the number of distinct entries was reduced to 89,065. This means one word shape token mapped to 1.38 words on average. Next, we extracted nouns, which are important content-representing words for information retrieval, from the lexicon. We were then left with 75,043 distinct word entries. Similarly, we obtained 57,049 distinct word shape tokens from them. This time, one word shape token mapped to 1.32 words. More importantly, most of them—49,953 of 57,049 word shape tokens (87.6%)—mapped to a single word.

3 Categorization System

We implemented a content-oriented categorization system to evaluate the word shape token-based approach in comparison with the OCR-based approach. The system, which uses the vector space classifier, consists of three main processes as shown in figure 3.

First, the system transforms the test document image into a sequence of word shape tokens as described in the previous section, where conventional systems perform OCR to generate a sequence of ASCII encoded words.

Next, it generates a document profile through the following stages:

Stage 1. The system removes punctuation marks. Note that they are distinguishable from alphabetical characters in the character shape code representation (Table 1).

Stage 2. The system removes word shape tokens corresponding to stop-words. In this process, it may also remove some non stop-words because of the one-to-many mapping between word shape tokens and words. In the OCR-based approach, it removes stop-words.

Stage 3. The system computes frequencies of word shape tokens to generate a document profile. The document profile $\boldsymbol{D}_i$ is represented as a vector of numeric weights, $D_i = \left(w_{i1}, w_{i2}, \ldots, w_{ik}, \ldots, w_{it}\right)$, where w_{ik} is the weight given the kth word shape token in the ith document, and t is the number of distinct word shape tokens of the ith document. We use the relative frequency between 0 and 1 as the weight. As for the OCR-based approach, read *word shape token* as *word*.

Finally, the system measures the degree of similarity between the document profile and a category profile. The category profile $\boldsymbol{D}_j$ is also represented as a vector derived in the same manner from a collection of topic-tagged document images. The system uses the cosine measure to compute the similarity:

$$sim(D_i, D_j) = \frac{\sum_{k=1}^{t} \left(w_{ik} \cdot w_{jk}\right)}{\sqrt{\sum_{k=1}^{t} \left(w_{ik}\right)^2 \cdot \sum_{k=1}^{t} \left(w_{jk}\right)^2}}$$

The greater the value of $sim(\boldsymbol{D}_i, \boldsymbol{D}_j)$, the more the similarity between $\boldsymbol{D}_i$ and $\boldsymbol{D}_j$. For each prepared category profile, the system computes the similarity to assign the test document to the most similar category[1].

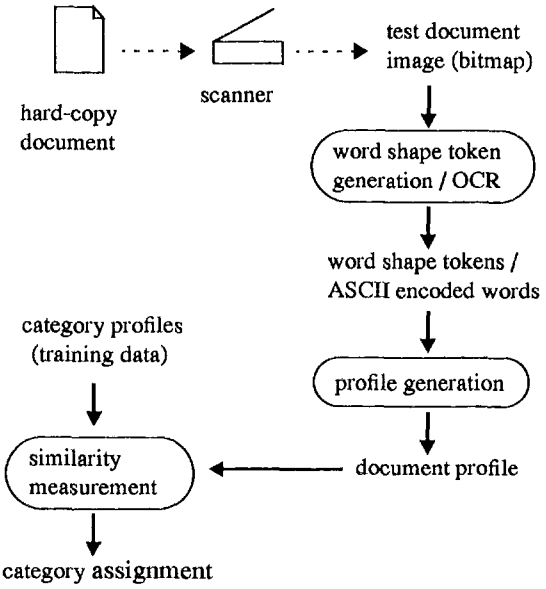

Figure 3: categorization process

4 Performance Assessment

We have constructed a document image database to compare our categorization approach with the conventional OCR-based approach. First, we carefully chose ten topic categories with strong boundaries. In general, the accuracy of an automated categorization system is evaluated by contrast with the expert judgements. However, experts don't always agree on the judgements. For an unbiased comparative experiments between the two approaches, we chose relatively specific topics. Resultant topic categories are affirmative action, Internet, stock market, local traffic,

1. In this paper, documents are always assigned to a single category.

Presidential race, Athletics (MLB), Giants (MLB), PGA golf, Tokyo subway attack, and food recipe. Second, we manually collected the body potion of 50 newswire articles for each category; 500 documents in total. They were clearly relevant to a single category and much less relevant to the other categories. Third, we printed them using a 300-dpi laser printer, and made *n*th generation photo-copies from them to degrade images by quality. In the photo-copy process, documents were degraded due to spreading toner, low print contrast, paper placement angle, paper flaws, and so on. Finally, we scanned the hard-copy documents of the first, the third, and the fifth generation with a 300-dpi scanner. As a result, we obtained 500 topic-tagged document images for each *n*th generation photo-copies (n = 1, 3, 5). Figure 4 shows scanned image samples. The average size of the original documents was 647, and ranged from 63 to 2,860 words. The standard deviation was 377.

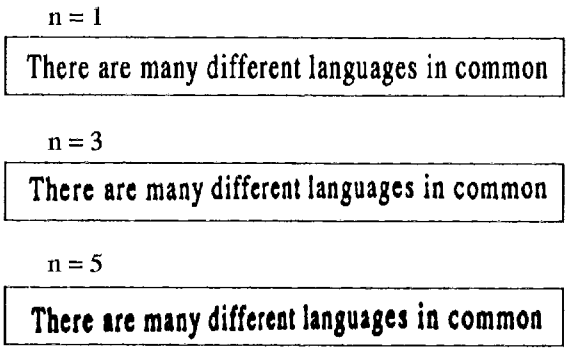

Figure 4: scanned image samples from *n*th generation photo-copy

We transformed the document images into word shape tokens and ASCII encoded words, where we randomly took 30 images for each category (300 in total) as training data to generate category profiles, and tested the remaining 20 images (200 in total). We used ScanWorX OCR (Xerox Imaging Systems)[1] for the ASCII encoding.

Table 2 shows the processing time for the transformation of all images on a SPARCstation 10 (Sun Microsystems). Although it had not been optimized, word shape token generation was 8 to 52 times faster than OCR. The difference increased with progression of n (n = 1, 3, 5). The OCR speed was highly dependent on image quality. Also, its word recognition accuracy was affected by image quality—96.3%, 92.8%, and 80.7% for the first, the third, and the fifth generation copies, respectively. It is well understood that OCR is slower and generates numerous errors for lower quality images (Taghva, et al., 1994). On the other hand, word shape token generation was a little faster for lower quality images. This unfavorable result was mainly caused by the lack of character segmentation function. Some characters touched each other in lower quality images, and were treated as a single character in the process of word shape token generation. Consequently, the number of characters to process became small.

Table 2: processing time (second) for word shape token (WST) generation and OCR

	image quality (*n*th generation photo-copies)		
	n = 1	n = 3	n = 5
WST	1860	1814	1702
OCR	15408	32322	87986

Our system categorized the test documents in word shape token and ASCII format as described in the previous section. As shown in Table 3, the accuracy of the word shape token-based approach for higher quality images (n = 1, 3) was nearly equal to that of the OCR-based approach. For lower quality images (n = 5), the former was significantly lower than the latter. Table 4 and 5 show the accuracy of the two approaches as a function of the size of test documents. When images were in higher quality (n = 1, 3), there was little correlation between the accuracy and the size. When they were in lower quality (n = 5), the OCR-based approach had stronger correlation between the accuracy and the size than the word shape token-based approach. This can be explained as follows: In the statistical categorization, it is generally difficult to get good accuracy when the size of the test document is small. In the OCR-based approach with the first and the third generation copies (n = 1, 3), the test documents were large enough for this categorization task. When the OCR encountered the fifth generation copies (n = 5), it garbled many words. Most of them were transformed into ill-formed (unknown) words[2] rather than mistaken for other words. These ill-formed words were ignored in our similarity measurement. Thus, they didn't act as a negative factor, but virtually made the size of the test document smaller. On the other hand, in the word shape token-based approach with the first and the third generation copies (n = 1, 3), the test documents were similarly large enough. When it encountered the fifth generation copies (n = 5), it also garbled many words. But, this time, they were mistaken for other word shape tokens (e.g., many - xxAg in Fig. 2), and acted as a negative factor to reduce the accuracy.

1. This is one of the state-of-the-art OCRs in terms of speed and accuracy, see Rice, *et al.*, 1995.

2. ScanWorX outputs a word with a reject mark when it is unable to recognize or is unsure in recognition (e.g., meterii~g).

Table 3: categorization accuracy for the word shape token-based and the OCR-based approach (number of correctly assigned documents / number of test documents)

	image quality (*n*th generation photo-copies)		
	n = 1	n = 3	n = 5
WST	193/200 (97%)	192/200 (96%)	154/200 (77%)
OCR	196/200 (98%)	196/200 (98%)	189/200 (95%)

Table 4: accuracy of the word shape token-based categorization as a function of the size of test documents

	size of test documents (number of words)		
	0 - 400	400 - 800	800 -
n = 1	50/51 (98%)	84/86 (98%)	59/63 (94%)
n = 3	51/51(100%)	81/86 (94%)	60/63 (95%)
n = 5	39/51 (76%)	62/86 (72%)	53/63 (84%)

Table 5: accuracy of the OCR-based categorization as a function of the size of test documents

	0 - 400	400 - 800	800 -
n = 1	50/51 (98%)	85/86 (99%)	61/63 (97%)
n = 3	50/51 (98%)	85/86 (99%)	61/63 (97%)
n = 5	44/51 (86%)	84/86 (98%)	61/63 (97%)

5 Discussion

From the experimental results in the previous section, our hypothesis that word shape token-based approach is quite adequate for content-oriented categorization was strongly supported at least for the document images from first and third generation photo-copies. This means that the mapping ambiguity between word shape tokens and original words was acceptable for the categorization purpose. The accuracy drop observed with the fifth generation photo-copies was not due to the mapping ambiguity but was caused by recognition errors. Unlike OCR which attempts to correctly recognize each word using lexical information, word shape token generation is only faithful to the original image. Thus, it makes many errors with low quality images, whereas OCR indicates illegible characters. Indicating diffidence is better than incorrect recognition for categorization. It would be possible to utilize lexical information in word shape token representation for reducing errors. However, we must pay attention to its computational expense.

Although it is arguable whether word stemming algorithms contribute to improving the categorization accuracy (Riloff, 1995), we desire to develop an algorithm for word shape token representation. It would be of use for other information retrieval applications such as word-spotting. We feel the word shape token representation is sufficient for locating some suffixes with accuracy. For example, 1,651 words were with suffix "*-tion*" in the lexicon of 122,545 distinct word entries. We obtained a set of word shape tokens from them. The set mapped to only 25 words without the suffix[1]. Similarly, word shape tokens from all 8,077 words with suffix "*-ing*" mapped to only 20 words without the suffix[2].

Because all capital letters map to **A** (Table 1), it is difficult to identify words with only capital letters, which are sometimes important content-representing words (e.g., acronyms). We need to find a graphical feature to distinguish some capital letters from others, considering the complexity of image analysis.

When we extend the word shape token processing to other applications, it is important to note that the word shape token representation is only meaningful for the computer and hardly human-friendly. Thus, it should be used in unsupervised systems with no human interaction required. Our technique would be useful for an automated incoming fax sorting by the content. Also, it would be used as an automated dictionary selector for the OCR which uses domain-specific dictionaries.

6 Conclusion

Several studies have suggested that OCR output is satisfactory for information retrieval in terms of accuracy. However, OCR is a major bottleneck for information retrieval systems in terms of speed.

We have described a technique to generate word shape tokens from document images, and have shown that this shape-based representation can be generated much faster than current OCR technology. Further, we have shown how word shape token processing can be applied to content-oriented categorization. In spite of the mapping ambiguity between word shape tokens and words, we have shown that the word shape token-based approach can categorize document images in good quality with nearly the same accuracy as the conventional OCR-based approach. When images are

1. e.g., **AxxAixn**-*fashion*, **exxeAixn**-*comedian*
2. e.g., **AexAing**-*destiny*, **Aing**-*tiny*

in poor quality, the accuracy drops significantly due to misrecognition of word shape tokens as opposed to OCR which indicates illegible characters rather than making errors.

Acknowledgments

We would like to thank Dan Kuokka for his comments, Ron Mann for his programming assistance, and Arlene Holloway for her constructing our document image database.

References

Kenneth W. Church, William A. Gale, Jonathan I. Helfman, and David D. Lewis. 1994. Fax: an alternative to SGML. In *Proceedings of the 15th International Conference on Computational Linguistics*. pages 525-529, Kyoto, Japan.

Tin Kam Ho and Henry Baird. 1994. Asymptotic accuracy of two-class discrimination. In *Proceedings of the Third Annual Symposium on Document Analysis and Information Retrieval*, pages 275-288, Las Vegas, Nevada.

David J. Ittner, David D. Lewis, and David D. Ahn. 1995 Text categorization of low quality images. In *Proceedings of the Fourth Annual Symposium on Document Analysis and Information Retrieval*, pages 301-315, Las Vegas, Nevada.

Elke Mittendorf, Peter Schauble, and Paraic Sheridan. 1995. Applying probabilistic term weighting to OCR text in the case of a large alphabetic library catalogue. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 328-335, Seattle, Washington.

Gregory K. Myers and Prasanna G. Mulgaonkar. 1995. Automatic extraction of information from printed documents. In *Proceedings of the Fourth Annual Symposium on Document Analysis and Information Retrieval*, pages 81-88, Las Vegas, Nevada.

Takehiro Nakayama and A. Lawrence Spitz. 1993. European language determination from image In *Proceedings of the Second International Conference on Document Analysis and Recognition*, pages 159-162, Tsukuba Science City, Japan.

Takehiro Nakayama. 1994. Modeling content identification from document images. In *Proceedings of the Fourth ACL Conference on Applied Natural Language Processing*, pages 22-27, Stuttgart, Germany.

Takehiro Nakayama. 1995. Text categorization using word shape tokens, In *Proceedings of the Second Conference of the Pacific Association for Computational Linguistics*, pages 207-217, Brisbane, Australia,.

Stephen V. Rice, Frank R. Jenkins, and Thomas A. Nartker. 1995. The fourth annual test of OCR accuracy. *Information Science Research Institute 1993 Annual Research Report, University of Nevada, Las Vegas*, pages 11-50.

Ellen Riloff. 1995. Little words can make a big difference for text classification, In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 130-136, Seattle, Washington.

Gerard Salton. 1991. Developments in automatic text retrieval, *Science, Vol. 253*, pages 974-980.

Penelope Sibun and David. S. Farrar. 1994. Content characterization using word shape tokens, In *Proceedings of the 15th International Conference on Computational Linguistics*. pages 686-690, Kyoto, Japan.

Penelope Sibun and A. Lawrence Spitz. 1994. Language determination: natural language processing from scanned document images. In *Proceedings of the Fourth ACL Conference on Applied Natural Language Processing*, pages 15-21, Stuttgart, Germany.

Kazem Taghva, Julie Borsack, Allen Condit, and Srinivas Erva. 1994. The effects of noisy data on text retrieval. *Journal of the American Society for Information Science. 45*, pages 50-58.

Claudia Wenzel and Rainer Hoch. 1995. Text categorization of scanned documents applying a rule-based approach. In *Proceedings of the Fourth Annual Symposium on Document Analysis and Information Retrieval*, pages 333-346, Las Vegas, Nevada.

Abdelwahab Zramdini and Rolf Ingold. 1993. Optical font recognition from projection profiles. *Electronic Publishing, Vol. 6(3)*, pages 249-260.