Simultaneous Highlighting of Paper and Electronic Documents

Jonathan J. Hull and Dar-Shyang Lee
Ricoh California Research Center
Menlo Park, CA
{dsl, hull}@rsv.ricoh.com

Abstract

The ability to automatically record the marks applied to paper documents on their electronic originals would preserve the information represented by those annotations. Users could even lose the original paper document. The marked-up version could be re-generated by merely re-printing it.

We describe a solution that saves an electronic representation for the highlights users commonly apply over the top of machine-printed text. A unique combination of algorithms is presented that maps the image captured from a pen scanner affixed to a highlighting pen onto text strings in electronic documents. Documents are automatically located in a large database using characteristics of the highlighted text. We describe here the system components, including the image recognition algorithms, and discuss their performance in finding a unique mapping from an image of text onto a sequence of words in an electronic document within a large database.

1. Introduction

Electronic documents are becoming increasingly prevalent in our lives. URL's for World Wide Web documents are widely advertised and every day, more content, such as scientific papers, is made available on the Internet. However, users often still prefer to read paper documents. The advantages of paper include its high resolution, persistence, portability, lack of a power requirement, easy re-generation by printing, and easy modification by addition of written annotations.

Highlights are one kind of written annotation that are very popular [6]. These are colored markings that are applied over the top of machine-printed text. They represent information added to a document at some cost in cognitive effort. As such, they are worthy of preserving in an electronic form that could be retrieved later. Ideally, this would even allow users to lose the original paper document and still be confident they could locate the annotated electronic version. This would be a significant advantage for users who have difficulty finding old documents.

We propose a method for capturing highlights as they are applied to paper documents and recording them on the electronic originals from which they were produced. This is suitable for paper documents for which the corresponding electronic original *exists* and is *accessible*. This is often a practical problem because it implies users must take some explicit action to prepare the electronic document for later access. That is, they must anticipate that at some future time they will highlight the document.

We provide access to electronic originals for most paper documents in an office with a document management system, known as the Infinite Memory Multifunction Machine (IM^3) [5], that saves an electronic version of every printed, copied or faxed document as a side-effect of processing them. Users make no explicit decision about whether any particular document is captured. Instead, *every* document is captured without asking the users.

Highlights are captured from a pen scanner attached to a highlighting marker. We use a commercially available pen scanner. Ideally, the scanning electronics would be embedded in the highlighting pen or in cartridges that would hold highlighting markers. Highlights are mapped onto the correct electronic document by combining the recognition results from one or more highlighted sequences of words (phrases).

There are several existing approaches that provide an interface between paper and electronic documents. They typically record the path followed by a pen when users write. Examples include video cameras focused on desktops or sheets of paper [2]. Another solution embeds small gyroscopes in a pen [8]. In both cases, the motion of a pen is transformed to a raster image. However, no means are provided for identifying the page being written on. While some solutions are obvious, such as scanning pre-printed bar codes before writing, this requires that bar codes be printed on a page before the user writes on it. This makes it difficult to use either of these methods for writing on an existing document and having that ink trace be registered with the electronic original. One solution to

Proceedings of the International Conference on Pattern Recognition (ICPR'00)
1051-4651/00 $10.00 @ 2000 IEEE

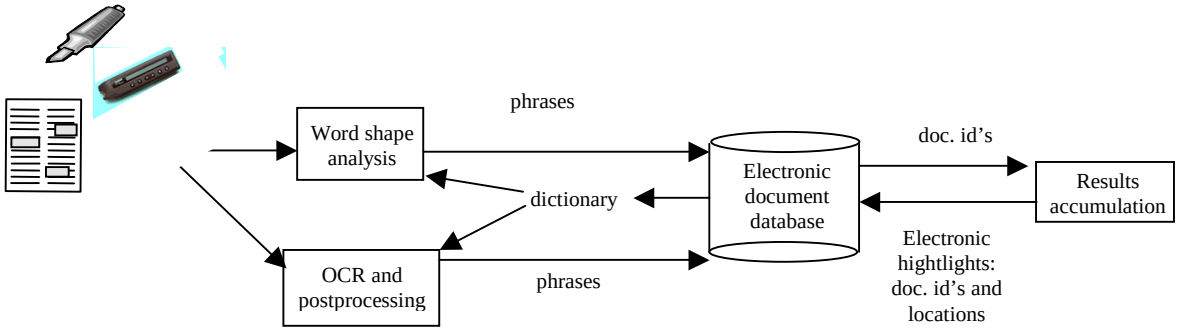

Figure 1. System design for simultaneous paper and electronic document highlighting.

this problem is to use specially prepared paper [3] that distributes an identification mark across its surface.

Another paper-electronic document interface uses a small video camera focused on the tip of a pen [1]. Highlights were captured by the camera. Users could manually indicate electronic actions that would be taken the next time the highlighted text was seen. However, no methodology was provided for automatically mapping highlighted text onto a collection of documents.

2. System Design

The design for the simultaneous annotation system is shown in Figure 1. A marker is combined with a pen scanner. Users can highlight documents as they normally do. Images of the highlighted words are transferred to a word shape analysis technique and an OCR process. Their results form queries to an electronic document database. Identifiers for documents that match the queries are accumulated across different annotations. When the document being processed is identified, electronic versions of the highlights are passed to the document database. They identify a document and the position of each annotation within the document.

The objective of the word shape analysis routine is to identify rough characteristics such as the number of characters in each word, whether they contain holes, and whether they are capitalized or lower case [4, 7]. Given a set of connected component features, a rule–based system calculates a set of candidate identifications for each character. These are composed into a regular expression that is used to query the dictionary. This returns a list of words that match the given constraints. A dynamic programming solution for phrase lookup is then applied independently to the lists of word candidates returned by word shape analysis and the OCR module. See Figure 2 for an example.

The electronic document database used in our experiments is the IM^3 system mentioned earlier. It has been in daily use at our laboratory for more than four years and during that time has captured over 70,000 documents containing more than 300,000 pages. About 20 users contributed data to this system, on average, over this period.

ASCII text is extracted from every document entered in the IM^3 and a full text index is constructed. This allows us to pose word-based conjunctive, disjunctive, and phrase queries. They return identifiers (id's) for the IM^3 documents that match the query.

The results accumulation module receives the list of IM^3 document id's found in the full text index. It identifies a unique document that contains a given set of images. Its underlying principle of operation is that while one image may map onto many documents, a number of images from the same document are highly likely to map onto that document and only that document. The major unknown factors are the number of images and the number of words needed in each image to obtain high accuracy.

3. Experimental Results

There are several factors that are important for technical success. One is the accuracy in recognition processing. It should produce sets of candidates for each word that contain the correct choice. The other key factor for success is the performance of results accumulation. It should need only a small number of short phrases to locate unique documents in a large database. This would provide an easy-to-use index for a document that is encoded directly in short sequences of words. I.e., it is a steganographic key that everyone can see but which does not modify the appearance of the document.

1051-4651/00 $10.00 @ 2000 IEEE

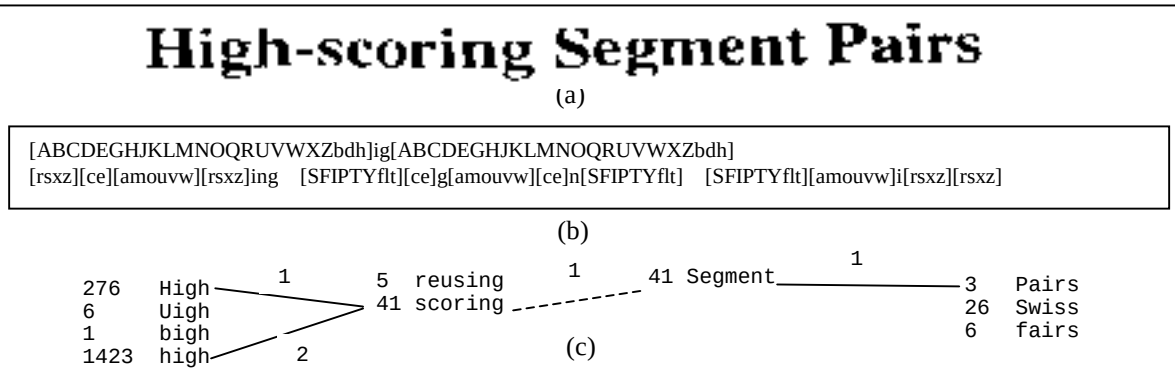

Figure 2. An example image from the pen scanner (a), regular expressions that locate words with the same shape (b), and word alternatives and their frequencies (c). Links between words show successful pair-wise dictionary lookups and their frequencies. The correct choice is determined at the "scoring Segment" lookup.

Recognition processing comprises word shape analysis and OCR plus post-processing. Others have investigated the performance of word shape in detail and reported better than 90% word accuracy [7]. We expect similar performance. The commercial OCR package we're using (from IRIS) is tuned for pen scanners. It typically produces better than 95% correct character recognition performance. The addition of custom dictionaries and post-processing should improve this.

The performance of the results accumulation module was investigated in two phases. First the number of documents that contain one or more phrases was determined. Phrases were randomly selected from each document in a subset of the IM^3 database. These were used as queries to the text index and the number of documents returned by each query, individually and in common, were calculated. This estimates the number of annotations needed to identify a unique document.

The document collection used in these experiments was derived from one user's IM^3 database. These were gathered from late 1995 through the end of 1999 and contain every document printed or photocopied by that user. Altogether, this comprises 9967 documents with 34,564 pages.

The second phase of the experiments estimated the effect of duplicates on the performance observed in the first phase. Duplicates were a significant consideration because the test collection contained many form letters and versions of documents that were printed several times during their lifetime.

3.1. Phrase matching

The ability of one or more phrases to identify a unique document was investigated by randomly selecting phrases of those lengths from each document in the test set. Numbers of phrases (N) from 1 to 4 were considered as well as a range of phrase lengths (PL) from 3 to 6. For each document in the database, N lines were randomly chosen that contained at least PL words. A starting position within each line was also randomly chosen and PL words beginning at this position were used as a query.

	Number of words per phrase			
Number of phrases	3	4	5	6
1	29% 65	35% 52	38% 28	39% 27
2	42% 21	45% 16	45% 10	46% 9
3	47% 9	48% 8	49% 6	49% 6
4	50% 6	50% 6	52% 5	52% 4

Table 1. Performance of phrase matching. The percentage of documents uniquely specified and the average number of documents are shown.

The results of this experiment are shown in Table 1. The percentage of documents uniquely identified by the chosen test phrase are indicated as well as the average number of documents found. This characterizes the number of alternatives the results accumulation module receives for any combination of phrase queries. The results show that performance improves significantly as more phrases with longer words are used. Performance peaks with about 50% of the queries returning only the correct choice. The average peaks at 4.

3.2. Duplicate Documents

The presence of duplicate documents is an obvious concern. The test database was taken from live data and was not preprocessed in any way. It contains many drafts of the same document as well as large numbers of form letters that differ only in the destination address. The prevalence of duplicates and their effect on performance was estimated by examining the results for 4 six-word queries. Each of the approximately 10,000 documents was compared to the files returned by multiple phrase matching. The percentage of unique vocabulary in the test document that occurred in each matching document was determined. This gives a rough approximation of whether they are the same document. If the percent unique vocabulary in common exceeds a threshold, we say they are the same (i.e., *duplicates*).

The results of the duplicate detection experiment are shown in Table 2. It is seen that when exact duplicates (100% vocabulary in common) are considered equivalent to the query document, the percent uniquely specified increases to 73%. This improves to 90% when 90% common vocabulary is the threshold and 98% when 75% common vocabulary is the threshold. The average number of returned documents decreases to 1. Examination of several cases shows that most of them are instances of the same two-paragraph form letter. This leads to the conclusion that when the percent unique peaks at around 50% in Table 1, an exact match is in fact being located. If more than one document is present, it is most likely a duplicate of the original.

Duplicate detection threshold					
75%	80%	85%	90%	95%	100%
98% 1	96% 2	94% 2	90% 3	85% 3	73% 3

Table 2. Effect of duplicates on matching highlights to electronic documents.

Duplicates are thus not a significant issue. They are an obvious concern, but we expect to compensate for them in the design of the user interface on the highlighting scanner or at retrieval time.

4. Conclusions

We presented a novel combination of components that allow users to highlight text in a paper document and have those highlights simultaneously recorded on the original electronic version. This preserves the highlights and guarantees that they can be re-generated if the paper document is lost. When used in combination with a document management system that retains electronic originals for every printed, copied, or faxed document, users could pick up almost any document, highlight it, and be confident those highlights would be recorded. No special preparation of the paper or electronic original would be required. Experimental results on a real-world collection of almost 10,000 documents captured over 4 years showed that a small number of short highlights can often uniquely specify a document. This demonstrates the feasibility of simultaneous paper-electronic document highlighting. Future work will consider issues that must be addressed before making this capability available to the public.

5. References

[1] T. Arai, D. Aust, and S.E. Hudson, "PaperLink: A technique for hyperlinking from real paper to electronic content," Proceedings of the ACM 1997 SIGCHI Conference, Atlanta, GA, March 22-27, 1997, 327-334.

[2] H. Bunke, T.V. Siebenthal, T. Yamasaki, and M. Schenkel, "Online handwriting data acquisition using a video camera," Fifth International Conference on Document Analysis and Recognition, Bangalore, India, September 20-22, 1999, 573-576.

[3] M. Dyteman and M. Cooperman, "Intelligent paper," Electronic Publishing, Artistic Imaging, and Digital Typography, R. Hersch, J. Andre, and H. Brown (eds), Springer Verlag, April, 1998, 392-406.

[4] J.J. Hull, "Hypothesis generation in a computational model for visual word recognition," IEEE Expert, v. 1, no. 3, Fall 1986, 63-70.

[5] J.J. Hull and P.E. Hart, "The infinite memory multifunction machine (IM^3)," DAS98: Pre-proceedings of the Third IAPR Workshop on Document Analysis Systems, Nagano, Japan, November 4-6, 1998, 49-58.

[6] C. Marshall, "Toward an ecology of hypertext annotation," Proceedings of ACM Hypertext'98, Pittsburgh, PA, June 20-24, 1998, 40-49.

[7] A.L. Spitz, "Shape-based word recognition," International Journal of Document Analysis and Recognition, v. 1, no. 4, May, 1999, 178-190.

[8] C. Verplaetse, "Inertial proprioceptive devices: self-motion-sensing toys and tools," IBM Systems Journal, v. 35, nos. 3 and 4, 1996, 639-650.

1051-4651/00 $10.00 @ 2000 IEEE