Camera-Based Document Image Retrieval as Voting for Partial Signatures of Projective Invariants

Tomohiro Nakai, Koichi Kise, Masakazu Iwamura
Graduate School of Engineering, Osaka Prefecture University
Gakuen-cho 1-1, Sakai, Osaka, 599-8531 Japan
nakai@m.cs.osakafu-u.ac.jp, {kise, masa}@cs.osakafu-u.ac.jp

Abstract

We propose a method of document image retrieval using digital cameras. The proposed method takes as input a part or the whole of a document acquired as a query by a digital camera, and retrieves a document image that includes the query. For this purpose, it is required to solve the problem of "perspective distortion" of images, as well as to establish a way of matching parts of document images flexibly. These are achieved based on the following characteristics of the proposed method: (1) Indexing of document images using the projective invariants called the "cross-ratios", (2) Retrieval as voting for partial signatures of document images defined by the cross-ratios. From experimental results using digital cameras with high and low resolutions, we demonstrate the effectiveness of the proposed method.

1. Introduction

Global dissemination and rapid improvement of digital cameras and those attached to mobile phones heighten the need of camera-based document image analysis [1]. Researchers have tackled problems of this field including dewarping, character extraction and recognition [2] as well as applications such as translation with camera phones [3]. We are concerned here with another important topic "document image retrieval" using digital cameras.

Retrieval of electronic documents is necessary for efficient management of a large scale document database. Electronic documents can be retrieved by keywords when their textual contens are available. On the other hand, in order to retrieve documents whose textual contens are not available, another technique such as document image retrieval is needed. Document image retrieval is the task of retrieving documents represented as images by the query of scanned or captured document images. Digital cameras are more desirable as the input device of document image retrieval since they are less bothersome.

In the field of document image retrieval, a wide variety of methods have been proposed [4]. In addition to ordinary indexing and retrieval based on recognized text, special descriptors or *signatures* have been proposed for boosting speed and robustness of retrieval [5, 6]. However, it is unfortunately difficult to directly apply such existing methods to camera-captured documents, since most of them are for document images obtained by flat-bed scanners.

In order to realize camera-based document image retrieval, it is required to solve some problems we have not encountered with scanner-based retrieval. For example, geometric distortions are within similarity transformation for scanner-captured documents, projective transformation should be considered for camera-captured documents. In addition, it is generally required to query document images by their parts due to lower resolution and limited visual ranges of digital cameras. It is also difficult to achieve such querying by parts because of the higher degrees of freedom of projective transformation. Matching of document images are also more difficult by the same reason.

In this paper, we propose a method of camera-based document image retrieval aiming to solve the above problems. The characteristic points are as follows: (1) In order to cope with the projective distortions, indices (signatures) of document images are calculated based on the projective invariants called the "cross-ratios". (2) Indices constructed from local evidences (partial signatures of document images) allow us to query by parts. (3) Voting with a hash table enables us to make the retrieval flexible with a reasonable computational cost.

2. Proposed method

2.1. Fundamental ideas

There are several problems to be solved for achieving camera-based document image retrieval: images captured

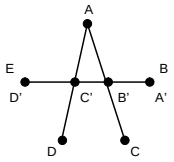


Figure 1. Cross-ratio.

by digital cameras can be projectively transformed, images may not include whole text regions, and resolution and illumination of images may be different from those in the database. Fundamental ideas for solving the above problems while keeping the computational cost feasible are as follows:

- Invariant-based indices $\cdots$ In order to make indices of document images projectively invariant, we calculate them using the cross-ratio which is known as an invariant of projective transformation. For five points ABCDE shown in Fig. 1, the cross-ratio is calculated as $(A'C' \cdot B'D')/(A'D' \cdot B'C')$. As feature points from which the cross-ratios are calculated, centroids of word regions are utilized, since they are robust to projective transformation and noises.
- Indices from local evidence $\cdots$ In order to achieve retrieval with partial images, each feature point is indexed based on the cross-ratios defined with local points.
- Retrieval as voting $\cdots$ In order to make retrieval computationally feasible, it is required to avoid the explicit matching of points between a query image and each database image. For this purpose, we employ voting for indices. In the proposed method, a document image which have the largest votes is regarded as correct.

2.2. Overview of processing

Figure 2 shows the overview of processing. At the step of feature point extraction, document images are transformed into a set of feature points. Then feature points are inputted into the registration step or the retrieval step. These steps share the step of calculation of indices.

2.3. Feature point extraction

Feature points should be obtained identically even under the perspective distortion, noise and low resolution. We employ centroids of word regions as feature points because they nearly satisfy this condition.

First, input images (Fig. 3(a)) are adaptively thresholded into binary images. Next, binary images are blurred using

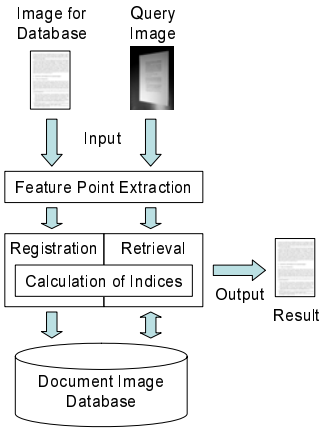
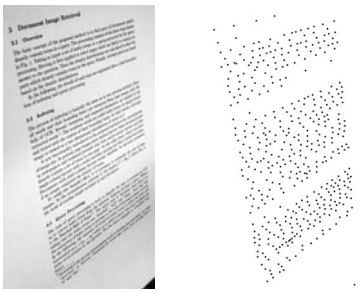


Figure 2. Overview of processing.



(a) input image (b) feature points

Figure 3. Feature point extraction.

the Gaussian filter whose parameters are determined based on an estimated character size (the square root of a mode of areas of connected components). Then, blurred images are adaptively thresholded again. Finally, centroids of word regions (Fig. 3(b)) are extracted as feature points.

2.4. Calculation of indices

In the proposed method, each feature point is characterized by the cross-ratios. Although it seems reasonable to calculate a cross-ratio for each feature point based on its five nearest feature points, it is not appropriate since in general the nearest points vary due to the projective distortion.

Another important problem is the discriminability of cross-ratios as illustrated in Fig. 4. In order to retrieve document images by voting, it is rquired to represent cross-ratios with k discrete values. Figure 4 shows the case with four discrete values. Suppose we have $\mathrm{cr}_o \sim \mathrm{cr}_3$ for documents A and B. Although the real values are different, their discrete versions are identical. Moreover, it is impossible to distinguish the documents A and B by counting the votes for each discrete value. Although the discriminability could

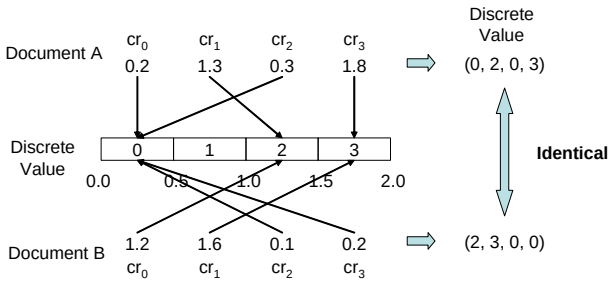


Figure 4. Discriminability of cross-ratios.

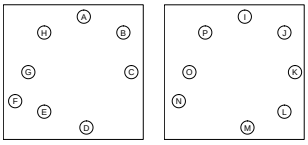


Figure 5. n points.

be improved by increasing the number k, it sacrifices the robustness to noise and image defects.

In the proposed method, we attempt to solve the first problem using local combinations of feature points. The index of a feature point is calculated not just from the five nearest point but from the n nearest points. It is often the case that $m(< n)$ points in the n points are kept unchanged under a normal projective distortion.

Let us explain in more details with Fig. 5 which represents the $n(= 8)$ nearest feature points for a feature point in a document image and those for the corresponding feature point in a query image. In this figure, $m = 7$ points ABCDFGH and IJKMNOP are common. Thus the common combination of feature points can be obtained by attempting all possible ${}_nC_m$ combinations. From the same combination of points, the common cross-ratios are obtained by combining all possible ${}_mC_5$ points for calculating cross-ratios such as ABCDF and IJKMN, ABCDG and IJKMO.

The second problem of discriminability is solved by taking into account the order of cross-ratios. In the case of Fig. 4, the cross-ratios are different if we consider them as the sequences (0,2,0,3) and (2,3,0,0). Note that if a feature point in a database image corresponds to that in a query image, the sequence should be identical. Consider again the case in Fig. 5. A sequence of cross-ratios are calculated for every m points. Suppose a series of letters ABCDF represent the cross-ratio defined by these points. If the points correspond with each other, the sequence from m points (ABCDF, ABCDG, ABCDH, BCDFG, BCDFH, ...) and its corresponding sequence (IJKMN, IJKMO, IJKMP, JKMNO, JKMNP, ...) become identical. Although it is rare to obtain all identical values of cross-ratios from real images, it is true that a certain amount of cross-ratios are common if a feature point in a database image corresponds to that in a query image.

1: **for all** $p \in$ {All feature points in a image for database} **do**
2: $P_n \leftarrow$ The nearest n points of p (clockwise)
3: **for all** $P_m \in$ { All m points combinations from P_n} **do**
4: **for all** $P_5 \in$ { All 5 points combinations from P_m} **do**
5: **for** $i = 0$ to 4 **do**
6: $cr_i \leftarrow$ The cross-ratio calculated with the ith point of P_5 as the starting point
7: **end for**
8: $H_{\rm index} \leftarrow$ The hash index calculated by Eq. (1).
9: Register (document ID, point ID, ${}_nC_m$pattern ID) using H_{index}
10: **end for**
11: **end for**
12: **end for**

Figure 6. Registration algorithm.

The following is the summary of calculation of indices. For each feature point, its n nearest points are obtained. Then all possible ${}_nC_m$ combinations of m points are generated. Indices are defined as ordered cross-ratios by taking ${}_mC_5$ combinations from m points in the fixed order.

2.5. Registration

Let us turn to the registration step. Figure 6 shows the algorithm of registration of document images to the database. In this algorithm, the document ID is the identification number of a document, and the point ID is that of a point. The ${}_nC_m$ pattern ID is the identification number given to every combination of m points from n points whose range is $0 \sim {}_nC_m - 1$. Similarly, the ${}_mC_5$ pattern ID is the identification number given to every combination of 5 points from m points whose range is $0 \sim {}_mC_5 - 1$.

For each combination of five points defined at the line 4, five cross-ratios are calculated at the lines 5∼7 as all cyclic permutations of the points such as ABCDE, BCDEA, CDEAB, DEABC and EABCD for the five points ABCDE [1].

Next, the index of the hash table is calculated by the following hash function:

$$H_{\rm index} = \sum_{i=0}^{4} cr_i (V_{\max} + 1)^i + pat(V_{\max} + 1)^5 \qquad (1)$$

where cr_i are the discrete values of five cross-ratios, $V_{\max}$ is the maximum value of the possible discrete cross-ratios and pat is the ${}_mC_5$ pattern ID.

A list (document ID, point ID, ${}_nC_m$ pattern ID) is registered into the hash table shown in Fig. 7. Chaining is used

[1] Another possible way is to utilize the feature called p^2-invariant which is invariant to the order of points.

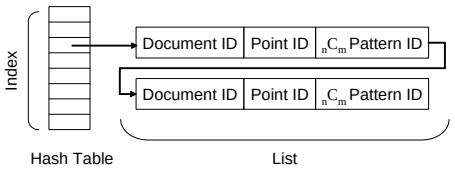


Figure 7. Configuration of the hash table.

1: **for all** $p \in$ {All feature points in a query image} **do**
2: $P_n \leftarrow$ The nearest n points of p (clockwise)
3: **for all** $P_m \in$ { All m points combinations from P_n} **do**
4: **for all** $P'_m \in$ { Cyclic permutations of P_m} **do**
5: **for all** $P_5 \in$ { All 5 points combinations from P'_m} **do**
6: **for** $i = 0$ to 4 **do**
7: $cr_i \leftarrow$ The cross-ratio calculated with the ith point of P_5 as the starting point
8: **end for**
9: $H_{\mathrm{index}} \leftarrow$ The hash index calculated by Eq. (1).
10: Look up the hash table using H_{index} and increment the corresponding cell of the first voting table.
11: **end for**
12: Increment the cell of the second voting table if the document ID in the first voting table has votes larger than the threshold l.
13: Clear the first voting table.
14: **end for**
15: **end for**
16: **end for**
17: Return a document image which has the largest votes in the second voting table.

Figure 8. Retrieval algorithm.

when a collision occurs. Not only document IDs but also point IDs and ${}_nC_m$ pattern IDs are registered in order for voting the number of matched ordered cross-ratios in the retrieval step.

2.6. Retrieval

The retrieval algorithm is shown in Fig. 8. In the proposed method, the number of matched cross-ratios is counted using the first voting table. If some cells have enough votes, the second voting table is incremented to determine retrieval results.

First, the hash index is calculated at the lines 6 $\sim$ 9 in the same way as in the registration step. At the line 10, the list shown in Fig. 7 is obtained by looking up the hash table. For each element of the list, the correspond cell of the first voting table is incremented.

These steps are repeated for every combination of five points from m points. Then, every cell of the first voting table is checked and the second voting table is incremented for all cells whose votes are more than l. Finally, a document with the largest number of votes is determined as a retrieval result.

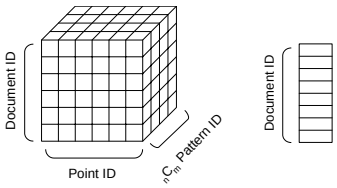


Figure 9. The first & second voting tables.

3. Experimental results

3.1. Overview

In the experimets, we employed a document image database consisting of different 50 pages of document images converted from PDF files of single- and double-column English papers. Examples of images in the database are shown in Fig. 10. Query images were captured using the following two cameras: (1) a high resolution normal digital camera: CANON EOS Kiss Digital (6.3 million pixels) with EF-S 18-55mm USM, (2) a low resolution digital camera attached to a mobile phone KYOCERA TK31(0.18 million pixels). Experiments were performed on a PC with Pentium 4 (2.4GHz) CPU and 768MB memory.

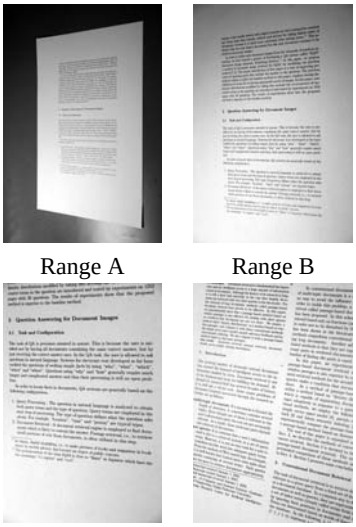
3.2. Ex. 1: with the high resolution camera

We first show the results of the experiment with the normal digital camera. Parameters described in Sect. 2 were set to $n = 8, m = 7, k = 9, l = 10$. As query images, 10 different pages were captured with four ranges (Fig. 11) and thus in total 40 images were employed. Range A covers the whole page, range B is for the whole text region, range C covers about half of the text region and range D is for a quarter of the text region. As shown in Fig. 11, images were captured to be projectively distorted to some extent.

Table 1 shows the results. From all query images, the correct document images obtained the largest numbers of votes. Hence accuracy of retrieval was 100% regardless of ranges. Moreover, correct document images obtained 5.11 times larger votes on the average than the largest votes of incorrect images. Therefore it would be easy to distinguish whether query images are included in the database by checking the numbers of votes.

As for the processing time, it takes much time since proposed method has a heavily nested algorithm. When parameters are set to $n = 8$ and $m = 7$, the number of times of hash access is ${}_nC_m \times m \times {}_mC_5 = 1176$ for each feature point. A narrower range enables us to access document images in shorter time because of the smaller number of feature points.

Figure 10. Images in database.



Range A Range B

Range C Range D

Figure 11. Four ranges of query images.

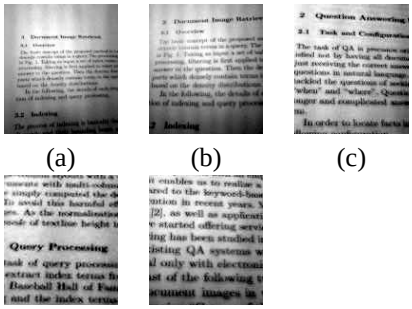
3.3. Ex. 2: with the low resolution camera

Next, we employed the query images captured by the low resolution camera shown in Fig. 12. As a result, retrieval succeeded on Figs. 12(b)∼(d) while failed on Figs. 12(a) and (e). In the case of Fig. 12(a), the resolution of the query image was too low to separate words and extract feature points. In the case of Fig. 12(e), the captured text region was too narrow to extract enough feature points.

This means that the range should be appropriately controlled to obtain correct images with such a low resolution camera. However, we consider that it is not a critical problem since the resolution of 0.18 million pixels is a low end

Table 1. Experimental results.

Range	A	B	C	D
Accuracy[%]	100	100	100	100
Processing time[sec]	231.6	173.1	157.6	118.1



(a) (b) (c)

(d) (e)

Figure 12. Query images captured by a mobile phone camera.

in available camera phones; those with the resolution of one million pixels or more would be enough to retrieve document images successfully.

4. Conclusion

We have proposed a method of camera-based document image retrieval which is characterized by (1) indexing with geometric invariants and (2) voting with hash tables. High accuracy of the proposed method with a normal digital camera was shown by the experimental results. The results also show the capability of retrieval with a low resolution devices such as cameras attached to mobile phones.

Future work includes the reduction of processing time and an extension of the proposed method to object retrieval in scene images.

References

[1] D. Doermann, J. Liang and H. Li: "Progress in camera-based document image analysis", Proc. ICDAR'03, pp. 606–616 (2003).

[2] P. Clark and M. Mirmehdi: "Recognising text in real scenes", IJDAR, **4**, pp. 243–257 (2002).

[3] Y. Watanabe, Y. Okada, Y-B. Kim, T. Takeda: "Translation camera", Proc. ICPR'98, pp.613–617 (1998).

[4] D. Doermann: "The Indexing and Retrieval of Document Images: A Survey", Computer Vision and Image Understanding, **70**, 3, pp.287–298 (1998).

[5] J. J. Hull : "Document Image Matching and Retrieval with Multiple Distortion-Invariant Descriptors", Document Analysis Systems, pp.379–396 (1995).

[6] A. F. Smeaton and A. L. Spitz: "Using Character Shape Coding for Information Retrieval", Proc. ICDAR'97, pp.974–978 (1997).