Multilingual Document Image Retrieval Based on a Large-Scale Database

Kazutaka TAKEDA[†], Koichi KISE[††], and Masakazu IWAMURA[††]

† School of Engineering, Osaka Prefecture University
†† Graduate School of Engineering, Osaka Prefecture University
1-1 Gakuen-cho, Naka, Sakai, Osaka, 599-8531 Japan
E-mail: †takeda@m.cs.osakafu-u.ac.jp, ††{kise,masa}@cs.osakafu-u.ac.jp

Abstract The document image retrieval method that can support various languages has already been proposed. However, the application of this method for a large-scale database consisting of documents in various languages has not been shown yet. In this paper, we experimentally evaluate the effectiveness of this method for a large-scale database of documents in Japanese, Chinese and Korean. From the experimental results, we have confirmed that this method realizes more than 93% accuracy on a 10,000 pages database for each language.

Key words Document image retrieval, LLAH, Large-scale database

1. Introduction

Document image retrieval is a task which finds document images corresponding to a given query from a database of a large number of document images. In the document image retrieval, various types of queries have been employed [1]. In a camera-based version of the document image retrieval, its queries are documents captured with digital cameras. If this document image retrieval method is realized, various services are linked to documents by capturing paper documents. Examples of the various services include a markerless Augmented Reality [2] and a camera-pen [3].

As one of document image retrieval methods, the retrieval method based on a hashing technique called Locally Likely Arrangement Hashing (LLAH) has already been proposed [4]. In this method, images are retrieved based on their features which consist of geometric invariants. It has been shown that more than 95% accuracy is accomplished with about 100 ms retrieval time on a 10,000 pages database in English [5].

In the above method, feature points are extracted from centroids of word regions. Therefore, if the language of documents is such as Japanese and Chinese in which words are not separated, stable feature points can not be extracted. For this reason, there is a problem that the above method can not retrieve document images of such languages.

As a solution of this problem, a document image retrieval method for various languages using LLAH has already been proposed [6]. This method is an expanded version of LLAH. In this method, document images written in such as the above languages can also be retrieved. However, application of this method for large-scale databases has not been verified yet.

In this paper, we experimentally evaluate the effectiveness of this method for large-scale databases of Japanese, Chinese and Korean documents. We constructed a 10,000 pages database for each language and experimented. From the experimental results, it has been shown that more than 99% accuracy is accomplished for the queries captured from the front for each language. We have confirmed that this method has effectiveness for large-scale databases of these languages.

2. English document image retrieval using LLAH

We first explain the LLAH [4] for languages in which words are separated such as English and the retrieval process with it.

2.1 Overview of processing

Figure 1 shows the overview of processing of document image retrieval by LLAH. First, at the step of feature point extraction, a document image is transformed into a set of feature points. Next, the feature points are inputted into the storage step or the retrieval step. These steps share the step of calculation of features. In the storage step, every feature point in the image is stored independently into the document image database using its feature. In other words, a document image is indexed by using each feature point. In the retrieval step, the document image database is accessed with features to retrieve images by voting. We explain each step in the following.

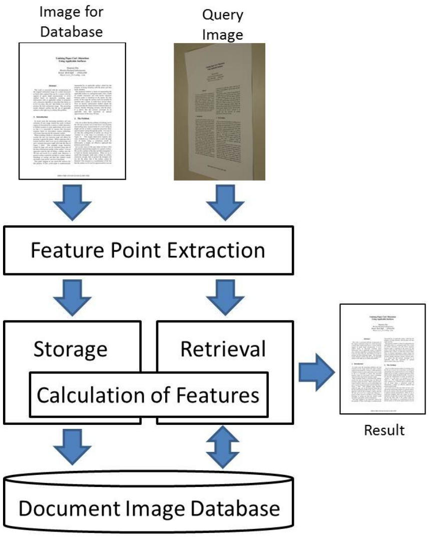


Figure 1 Overview of processing

2.2 Feature point extraction

In LLAH, the matching of a document image is based on the arrangement of feature points. Therefore, feature points should be extracted stably even under the influence of projective transformation and noise. For this reason, centroids of word regions as feature points are employed. Since centroids of word regions are extracted relatively stably, it is effective for stable calculation of features.

The processing is as follows. First, the input image is adaptively thresholded into the binary image. Next, the binary image is blurred using the Gaussian filter. The blurred image is adaptively thresholded again. Finally, centroids of word regions are extracted as feature points.

2.3 Calculation of features

The feature is the value which represents a local arrangement of feature points. The matching of feature points is based on the feature. In order to realize successful retrieval, a robust feature should be utilized. Hence, we employ geometric invariants which are invariant to geometric distortion. In concrete term, an affine invariant is utilized. The affine invariant is defined using four coplanar points ABCD as follows:

$$\frac{P(A,C,D)}{P(A,B,C)} \qquad (1)$$

where P(A,B,C) is the area of a triangle with apexes A, B and C.

In order to increase the discrimination power of the feature, multiple affine invariants calculated from $m(>4)$ feature points are employed as the feature. Furthermore, in order to deal with errors of feature point extraction, multiple features are calculated from nearest $n(>m)$ feature points. First, nearest n points from the feature point of interest are identified. Next, all combinations of m points taken from n points are examined. Therefore, we can obtain $\binom{n}{m}$ features from one feature point. Then, a sequence of discretized affine invariants $(r_{(0)},\cdots,r_{(\binom{m}{4}-1)})$ are obtained from all possible combinations of 4 feature points taken from m feature points. We utilize this sequence of affine invariants as the feature.

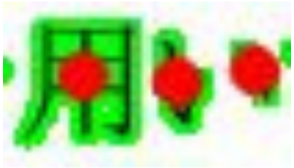


Figure 2 Feature points are extracted from connected components

2.4 Storage

Every feature point is stored to the database in accordance with its feature. The index H_{index} of the hash table is calculated by the following hash function:

$$H_{\text{index}} = \left(\sum_{i=0}^{\binom{m}{4}-1} r_{(i)}k^i\right) \bmod H_{\text{size}} \qquad (2)$$

where $r_{(i)}$ is a discrete value of the invariant, k is the level of quantization of the invariant, and H_{size} is the size of the hash table. The item (document ID, point ID, $r_{(0)},\cdots,r_{(\binom{m}{4}-1)}$) is stored into the hash table where chaining is employed for collision resolution.

2.5 retrieval

In LLAH, retrieval results are determined by voting on documents represented.

First, the hash index is calculated for the feature point of a query image in the same way as in the storage step. The list of items is obtained by looking up the hash table. For each item of the list, the corresponding document ID we cast a vote for if it has the same feature $(r_{(0)},\cdots,r_{(\binom{m}{4}-1)})$. Finally, the document which obtains the maximum votes is returned as the retrieval result.

3. Retrieval of document images in various languages using LLAH

Next, we explain the LLAH for various language which contain languages with no space between words [6].

3.1 Feature points extraction

In the previous method of LLAH, feature points are extracted as centroids of connected components of words obtained by the Gaussian filter. However, it is difficult to obtain word regions in Japanese and Chinese in which words are not separated. Therefore, another feature point extraction method is required. We employ centroids of connected components as feature points. As shown in Figure 2, connected components can be obtained from a character or a part of a character. Due to low resolution and defocusing, it is difficult to obtain a camera-captured document image where fine strokes are completely separated. For this reason, input images are

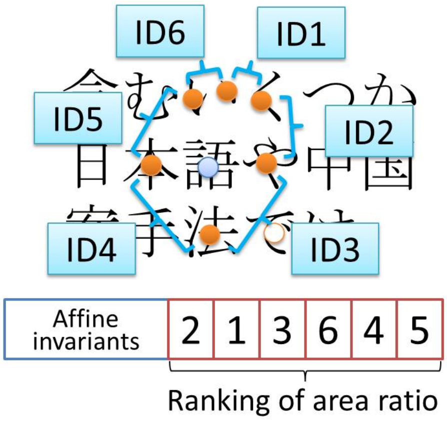


Figure 3 Additional features are the ranks of area ratios of connected components

blurred using the Gaussian filter to combine fine strokes and their adjacent connected components. This feature point extraction method is equivalent to the previous method except for a smaller mask size of the Gaussian filter.

3.2 Additional feature

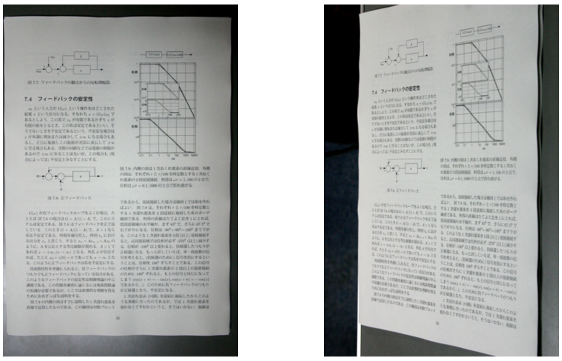
In Japanese and Chinese documents, most characters consist of one connected component obtained by Gaussian filter. Moreover, most characters are placed at equal spaces. Therefore, lattice-like arrangements of feature points are dominant. Discriminative features are hardly extracted from such arrangements. In order to solve this problem, additional features based on area ratios of connected components are introduced.

Figure 3 shows the example of additional features. In this figure, a feature is calculated from 6 feature points taken from nearest 7 feature points. First, ID is configured between adjacent feature points. Next, the area ratio of two adjacent connected components is calculated. Then, IDs are sorted in descending order according to the area ratios. Finally, we employ a sequence of the sorted IDs as additional feature. In this figure, the ID 2 has the largest area ratio among the 6 IDs.

4. Experiments

4.1 Experimental overview

In order to confirm the effectiveness of LLAH for large-scale databases of Japanese, Chinese, and Korean documents, we performed experiments to examine accuracy and retrieval time. We constructed a database for each language. Each database contained 10,000 pages of document images. For each language, 100 pages were selected from each database and printed. These printed documents were captured with a digital camera as queries. The size of captured images is 3000 × 2000. Shooting angles are 90°, 60° and 45° for each printed document. Figure 4 shows examples of captured images. We used a computer with 2.8GHz CPU and 128GB memory.



(a) 90°

(b) 45°

Figure 4 Examples of query image

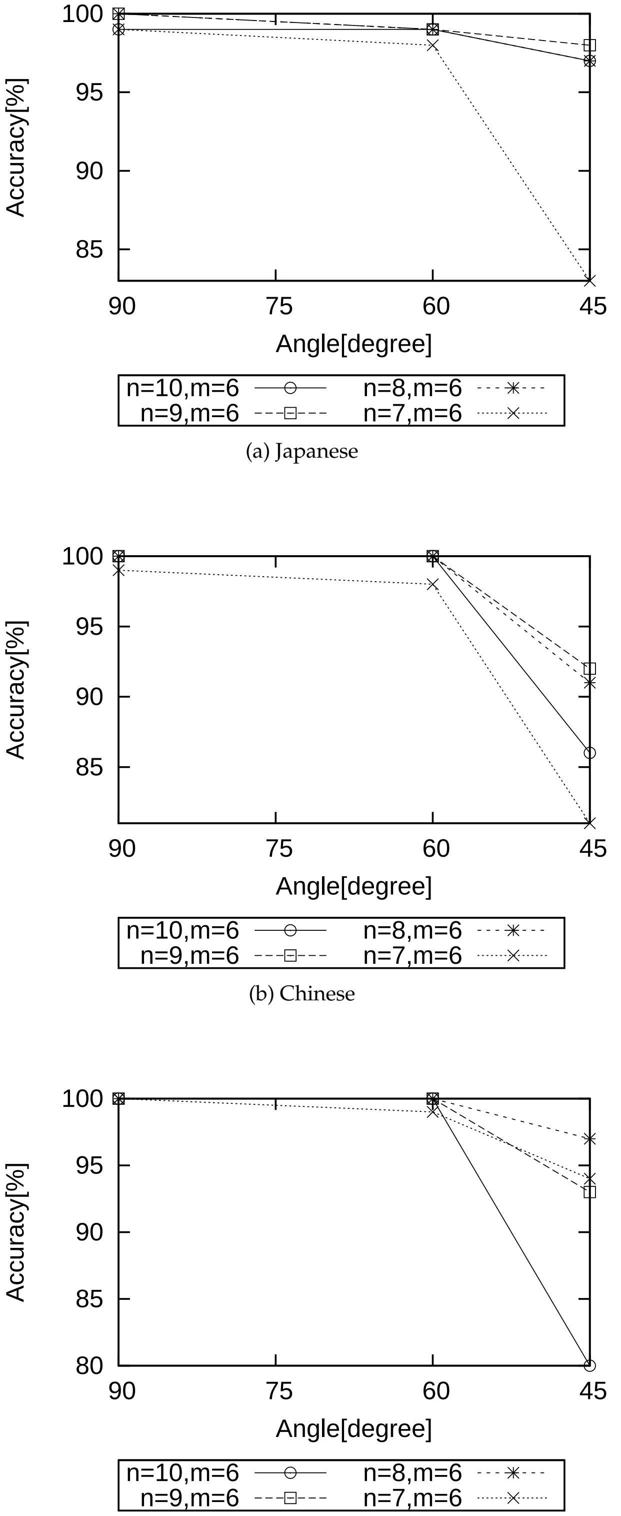
4.2 Ex.1 : relationship between accuracy and shooting angles

We first examined relationship between accuracy and shooting angles. The performance of LLAH changes depending on the parameters n and m to decide the number of features. In this experiment, we evaluated accuracy with $m = 6$, $n = 10, 9, 8, 7$. The numbers of levels of discretization was 10.

Figure 5 shows relationship between accuracy and the shooting angle about all the combinations of n and m in (a) Japanese, (b) Chinese and (c) Korean. As the shooting angle becomes small, the accuracy reduces. This is because the neighborhood structures of feature points have been changed by projective distortion. In other words, the requirement corresponding m feature points taken from nearest n feature points are not satisfied. However, we obtained positive results with $n = 8$ for all three languages. The reasons are as follows. The larger the value of n is, the more features are calculated from one feature point. Therefore, a large n is more likely to improve the robustness against variance of nearest points by projective distortion. In spite of this, with $n = 10$, the accuracy reduced. This is because the discrimination power of features dropped due to the enormous number of features stored in the database. As a result, the accuracy was reduced by increasing false votes. From these results, $n = 8$ is the best value for the accuracy.

Figure 6 shows a document image which is liable to fail in retrieval. The reasons are as follows. To obtain the effective features, many feature points are required. However, this document image has a small number of feature points due to few characters. Therefore, it is difficult to obtain effective features from this document image. For this reason, this document image is hard to be retrieved.

From the above-mentioned results, we have confirmed that LLAH has effectiveness for a large-scale database of documents in various language in which words are not separated.



(a) Japanese

(b) Chinese

(c) Korean

Figure 5 Accuracy of retrieval

4.3 Ex.2 : relationship between the parameter *n* and retrieval time

Next, we evaluate relationship between the parameter *n* and retrieval time. Retrieval time here is the time needed for retrieval processing of one query, excluding the time for the feature point extraction, which is around 0.7 sec. In this experiment, we evaluated retrieval time with $m = 6$, $n = 10,9,8,7$. The number of levels of discretization was 10. As a query, we employed images captured with 60° shooting angle.

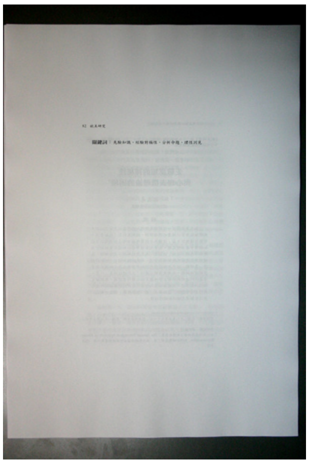


Figure 6 Example of failure

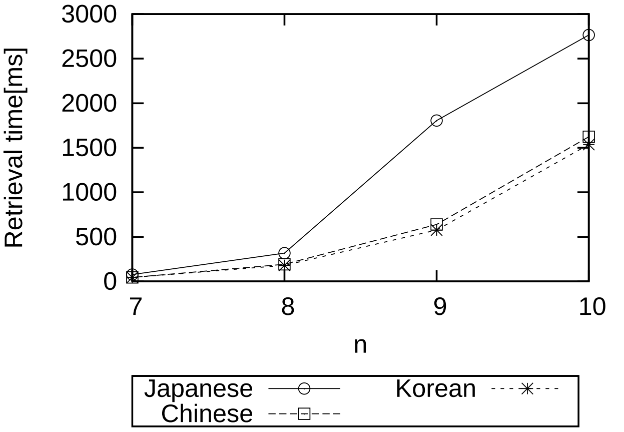


Figure 7 Additional features are the ranks of area ratios of connected components

Figure 7 shows relationship between the parameter *n* and retrieval time. As *n* increases, the retrieval time becomes long. In particular, the time largely increases with $n = 10$. This is because the number of computed invariants and access to the database have been increased by enlarging the value of *n*. The acceptable retrieval time is with $n = 7,8$.

5. Conclusion

In this paper, we have evaluated the effectiveness of LLAH for large-scale databases of Japanese, Chinese and Korean documents. We constructed a 10,000 pages database for each language and performed experiments. From the experimental results, we confirmed the effectiveness of LLAH for the large-scale database of documents in each language. Our future work includes inspecting effectiveness for the larger scale databases and dealing with the problem that this method can not retrieve documents of few characters.

Acknowledgment

This research was supported in part by the Grant-in-Aid for Scientific Research (B)(22300062) from Japan Society for the Promotion of Science (JSPS).

References

[1] D. Doermann, "the indexing and retrieval of document images: a survey", Computer Vision and Image Understanding, vol. 70, no. 3, pp. 287-298 (1998).

[2] A.I. Comport, E Marchand, M. Pressigout, and F. Chaumette, "Real-time markerless tracking for augmented reality: the virtual visual servoing framework", IEEE Transactions on Visualization and Computer Graphics, vol. 12, no. 4, pp. 615-628 (2006).

[3] K. Kise, M. Chikano, K. Iwata, M, Iwamura, S. Uchida and S. omachi, "Expansion of Queries and Databases for Improving the Retrieval Accuracy of Document Portions", Proceedings of the 9th IAPR International Workshop on Document Analysis Systems (DAS2010), pp. 309-316 (2007).

[4] T. Nakai, K. Kise and M. Iwamura: "Camera based document image retrieval with more time and memory efficient llah", Proceedings of Second International Workshop on Camera-based Document Analysis and Recognition (CBDAR2007), pp. 21-28 (2007).

[5] T. Nakai, K. Kise, and M. Iwamura, "Use of affine invariants in locally likely arrangement hashing for camera-based document image retrieval", Lecture Notes in Computer Science (7th International Workshop DAS2006), vol. 3872, pp. 541-552 (2006).

[6] T. Nakei, K. Kise, and M. Iwamura, "Real-Time Retrieval for Images of Documents in Various Languages using a Web Camera", Proceedings of the 10th international Conference on Document Analysis and Recognition (ICDAR2009), pp. 146-150 (2009).