2009 10th International Conference on Document Analysis and Recognition

Real-Time Retrieval for Images of Documents in Various Languages using a Web Camera

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

Abstract

We propose a real-time retrieval method for document images in various languages. In this method, queries are images of documents captured by a web-camera. The document images corresponding to the queries are retrieved from the document image database in real time. Since we have already proposed a document image retrieval method for English documents, the proposed method is an extension for retrieval of documents in various languages. In the previous English document image retrieval method, only centroids of word regions are used as feature points. Therefore it cannot be applied to some languages including Japanese and Chinese due to no separation between words and periodic arrangements of characters. In the proposed method, additional features are introduced to realize real-time retrieval for document images in various languages.

1. Introduction

Recently, digital cameras are in widespread use. In the field of mobile phones, camera phones have become very popular devices. Nowadays the quality of a mobile phone camera is comparable to that of an ordinary digital camera. Therefore we are in the situation that ordinary people always have high resolution digital cameras.

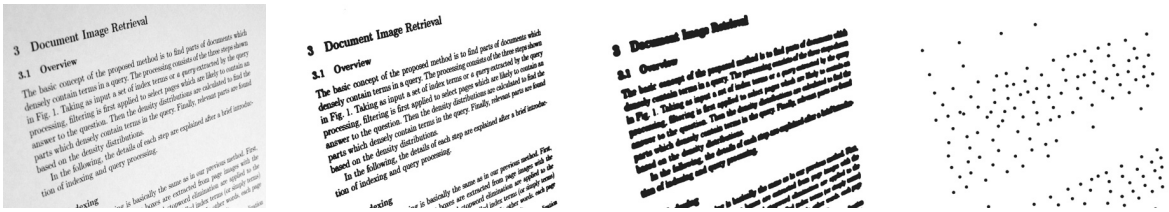
Services which utilize camera-captured images are attracting attention. Recognition 2-dimensional barcodes [1] is one of the typical services. This technique enables to extract information from special black-and-white patterns on objects by capturing them with a digital camera. The patterns have distinctive appearances in order for robust recognition. However, such an approach has a problem that the distinctive patterns sometimes spoil appearances of the recognized objects. It has also a limitation that adding patterns to existing objects is difficult. In order to retrieve information from documents using barcodes, they have to be printed with their barcodes. It is difficult to add or modify the barcodes afterward.

Therefore an approach to extract information using appearances of objects as keys is desired. In this approach, an appearance of an object is linked with its relevant information in the database. From a captured image of the object, the relevant information is retrieved and provided to users. This process requires object recognition to identify objects based on their appearances [3]. However, when a digital camera is used as an input device, appearances of objects vary due to change of camera angles toward objects. Therefore object recognition of camera captured images is a difficult problem.

Several researches to solve this problem for the case of documents have been published [4]. We also proposed an image retrieval method for documents named LLAH (Locally Likely Arrangement Hashing) [5]. It enables to find the document image corresponding to the camera-captured image from the database. In this method, images are retrieved based on their feature descriptors which consist of geometric invariants. Since geometric invariants are invariant to perspective distortion, LLAH is robust to change of camera angle. LLAH is also known for fast retrieval which enables a real-time retrieval [6].

In the previous methods, LLAH has been applied only for English documents. This is because LLAH requires stable feature points. Since English documents have a space between words, stable feature points can be extracted from centroids of word regions. On the other hand, scripts of other languages do not necessarily have a space between words. For this reason, usage of LLAH for documents in other languages has not been shown yet. If a technique depends on characteristics of a language, applying the technique to different languages sometimes causes difficult problems [2]. However, applying LLAH to other languages is beneficial for users of the languages. It is more significant in the countries such as Japan where digital cameras are widely used. Therefore image retrieval of documents in various languages using LLAH is a difficult but important

978-0-7695-3725-2/09 $25.00 © 2009 IEEE
DOI 10.1109/ICDAR.2009.125



(a) Input image. (b) Binarized image. (c) Connected components. (d) Feature points.

Figure 1. Feature point extraction.

problem.

In this paper, we propose an application of LLAH to documents in various languages. It is not easy to extract stable feature points from documents of languages such as Japanese and Chinese in which words are not separated. In the proposed method, we adopt centroids of connected components as feature points extracted from these documents. In Japanese and Chinese documents, characters are placed evenly spaced apart. Therefore arrangements of feature points have less discrimination power. It is crucial for LLAH because descriptors are calculated from arrangements of feature points. In the proposed method, additional descriptors are calculated from areas of connected components. We implemented a real-time retrieval system using the proposed method. From experimental results of the system, we confirmed effectiveness of the proposed method in documents in various languages.

2. English document image retrieval using LLAH

LLAH is an image retrieval method. First, feature points are extracted from an image. Then local descriptors are calculated from arrangements of the feature points. Using the local descriptors, the image corresponding to a query image is retrieved from an image database. Since the descriptors consist of geometric invariants, LLAH can deal with images which suffer from perspective distortion.

2.1. Feature Point Extraction

An important requirement of feature point extraction is that feature points should be obtained identically even under perspective distortion, noise, and low resolution. To satisfy this requirement, we employ centroids of word regions as feature points.

The processing is as follows. First, the input image (Fig. 1(a)) is adaptively thresholded into the binary image (Fig. 1(b)). Next, it is blurred using the Gaussian filter. Then, the blurred image is adaptively thresholded again (Fig. 1(c)). The resulting blobs are assumed to be word regions. Finally, centroids of the word regions (Fig. 1(d)) are extracted as feature points.

2.2. Calculation of Local Descriptors

The descriptor of LLAH has following features.

- A descriptor is defined for each feature point.

  In order to realize robustness and availability under occlusion, a descriptor has locality.

- A descriptor is calculated using geometric invariants.

  In order for invariance to perspective distortion which occurs in camera-captured images, geometric invariants are used. In concrete term, the affine invariant is used. An affine invariant is defined using four coplanar points ABCD as follows:

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

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

- A descriptor consists of more than one geometric invariants.

  In order to increase discrimination power of a descriptor, multiple affine invariants calculated from multiple feature points are used. Since an affine invariant is calculated from four points, more than one affine invariants can be calculated from more than four feature points. In concrete, a descriptor is $(r_{(0)}, \cdots, r_{({}_mC_4-1)})$ calculated from m neighboring points where $r_{(i)}$ is an affine invariant. All possible combinations of four points from m points are used.

- More than one descriptors are calculated for each feature point.

  In order to deal with errors of feature point extraction, multiple descriptors are calculated from nearest $n(> m)$ points. In concrete, ${}_nC_m$ descriptors are calculated. All possible combinations of m points from n points are used.

2.3. Storage and Retrieval

In LLAH, images are retrieved using a hash table. Descriptors of images in the database are preliminarily calculated and stored in the hash table. When a query image is given, descriptors with the same value are retrieved from the hash table. By voting images in the database, the document image corresponding to the query image can be retrieved.

2.4. Real-time document image retrieval

A real-time document image retrieval system with a web camera [6] has been proposed as an application of document image retrieval using LLAH. In this system, each frame image captured by a web camera is used as a query image. Since retrieval by LLAH is fast, simple repetition enables real-time retrieval.

3. Retrieval of document images in various languages using LLAH

3.1. Feature points extraction

In the previous method of LLAH, feature points are extracted as centroids of connected components of words obtained by Gaussian filter. This method works well in English documents since it have spaces between words. Therefore it is also effective in Latin (e.g. French and Spanish) documents. Even though they are not written in alphabets, documents in languages with spaces between words (e.g. Arabic and Hindi) are also applicable to the previous method.

On the other hand, another feature point extraction method is required for documents in languages with no space between words (e.g. Japanese and Chinese). We propose a feature point extraction method which takes centroids of connected components as feature points. As shown in Fig. 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. Therefore input images are blurred using a Gaussian filter to combine fine strokes and their adjacent connected components. This feature point extraction method is equivalent to the previous method except for smaller mask size of the Gaussian filter.

3.2. Additional descriptor

In Japanese and Chinese documents, most characters consist of one connected component. Moreover, most characters are placed evenly spaced apart. Therefore lattice-like

Figure 2. Feature points are extracted from connected components.

arrangements of feature points are dominant in most documents. Discriminative descriptors are hardly extracted from such arrangements. In the proposed method, additional descriptors based on areas of connected components are introduced

Stability is crucial for additional descriptors. Areas of connected components can vary with change of image capturing conditions. Therefore additional descriptors should be robust to change of connected components' areas. For example, under dim illumination, captured image become dark. In this case, connected components have larger areas.

We focus on magnitude relation among areas of connected components. The largest connected component tends to have the largest area under certain degree of change in condition. Since magnitude relation among areas of connected components hardly changes, we adopt ranks of connected components' areas as descriptors.

An example of a descriptor is shown in Fig. 3. In this picture, a descriptor is calculated from 6 points selected from 7 nearest points. Numbers are given to the 6 points in the order corresponding to their angles from the center point. In the same manner as the previous method, affine invariants are calculated from centroids of connected components. Since ${}_6C_4 = 15$ combinations of 4 points can be obtained from 6 points, 15 affine invariants are calculated. This descriptor is shown as (Invariant 1, $\cdots$, Invariant ${}_6C_4 = 15$) in Fig. 3. The additional descriptor in the proposed method is shown in its right side. In this example, the point 3 has the largest connected component among the 6 points. The second largest connected component belongs to the point 6. Like this, numbers of points are aligned in the order of areas. Finally, (3, 6, 1, 5, 2, 4) is obtained as an additional descriptor.

As shown in Fig. 3, the additional descriptor is added to the existing descriptor. It increases discrimination power of the descriptor. On the other hand, it also decreases stability. In order to avoid decrease of stability, smaller number of discretization is used in the proposed method.

3.3. Retrieval

As stated in **3.1**, languages with spaces between words and those without spaces have different effective feature

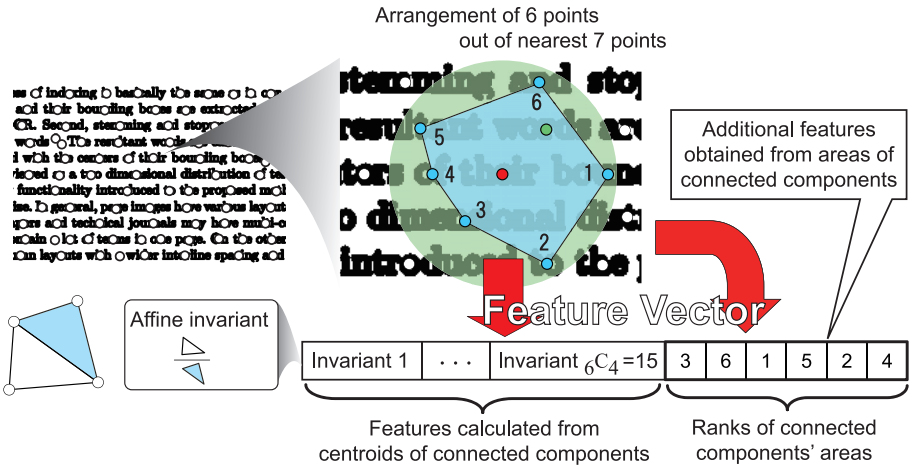


Figure 3. Additional descriptors are the ranks of areas of connected components.

points. Centroids of word regions are stable in the former ones and centroids of connected components are stable in the latter ones. If languages of documents can be recognized before retrieval, appropriate feature points can be extracted. However such an approach requires an extra process to recognize languages. It is also undesirable in terms of accuracy. When the language recognition fails, the remaining retrieval process also fails.

In the storage process of the proposed method, two sets of feature points are extracted as centroids of word regions and those of connected components from a document image. They are individually stored in two databases. In the retrieval process, two sets of feature points are extracted from a query image. For each of them, the retrieval process is performed to its corresponding database. As a result, two numbers of votes are obtained for each document in the database. The retrieval result is determined based on weighted sum of them.

4. Experimental results

In order to confirm effectiveness of the proposed method, we performed experiments of real-time document image retrieval using documents in various languages. We used documents of the following 10 languages: Japanese, English, Arabic, Chinese, French, Hindi, Korean, Russian, Spanish and Tamil. The database contained 1,000 pages of document images, 100 pages for each language. For each language, 10 pages are extracted from the database as queries. Queries are 100 pages of printed documents. In the experiment, a web camera was fixed in 6cm above surface of the captured document. Query documents were put and replaced continuously under the camera. Size of the captured image is 1600×1200. Figure 4 shows examples of captured

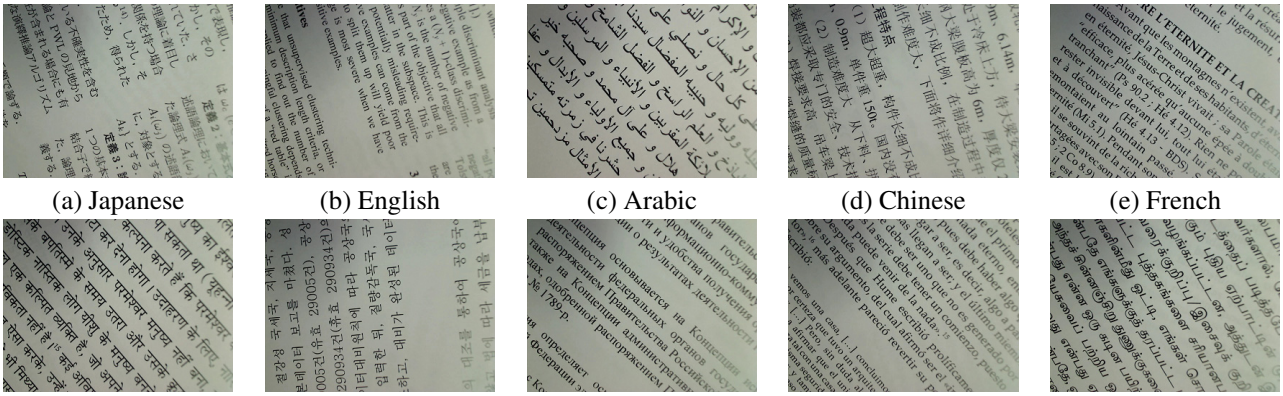
Table 1. Accuracy of retrieval.

Script	Accuracy
Japanese	9/10(90%)
English	9/10(90%)
Arabic	10/10(100%)
Chinese	7/10(70%)
French	10/10(100%)
Hindi	9/10(90%)
Korean	10/10(100%)
Russian	7/10(70%)
Spanish	10/10(100%)
Tamil	10/10(100%)
Total	91/100(91%)

images. The number of frame images is 985. Each page was captured in about 10 frames. Retrieval of a frame image is considered as successful if the corresponding document image obtained the largest number of votes. In this experiment, success or failure of retrieval was determined for each page. When more than half of frames which include the page were successful, retrieval of the page was considered as successful. We used a client PC with 2.2GHz CPU and 2GB memory, a server computer with 2.8GHz CPU and 32GB memory. As for parameters of LLAH, $n = 7$ and $m = 6$ were used. The number of discretization was 7.

Accuracy of retrieval is shown in Table. 1. Average processing time was 359ms and average frame rate was 2.78fps. As shown in Table. 1, high accuracy was obtained. Especially for Latin languages (English, French and Spanish), almost 100% accuracy was obtained.

On the other hand, relatively low accuracy was obtained



(a) Japanese (b) English (c) Arabic (d) Chinese (e) French

(f) Hindi (g) Korean (h) Russian (i) Spanish (j) Tamil

Figure 4. Examples of document images used in the experiment.

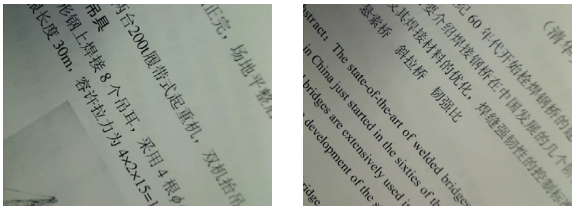


Figure 5. Failure cases. Left: the number of characters was insufficient. Right: different scripts (English and Chinese) are included in the captured image.

for Chinese and Russian documents. As for Chinese, insufficient number of characters and emergence of different scripts as shown in Fig. 5 seem to cause failures. English and Chinese scripts have different effective feature points: centroids of word regions for English and centroids of connected components for Chinese. Since different feature points are stored and retrieved individually, a page with different scripts is difficult to be retrieved successfully.

As for Russian, narrow captured regions caused failure. Russian documents have relatively fewer words due to shape of their Cyrillic alphabets. Therefore narrow captured regions result in insufficient feature points.

As stated above, relatively low accuracy was obtained for some languages. However, more than 70% accuracy for all 10 languages shows effectiveness of the proposed method.

5. Conclusion

In this paper we proposed an extension of our image retrieval method LLAH for retrieval of documents in various languages. The proposed method includes the feature point extraction method from documents of languages which have no space between words. It also includes additional descriptors based on areas of connected components. We implemented a real-time document image retrieval system with the proposed method. From the experimental results we confirmed effectiveness of the proposed method in documents in various languages. Our future work includes improving accuracy for some languages and confirming effectiveness of the proposed method in other languages.

References

[1] Qr code.com. from http://www.denso-wave.com/qrcode/index-e.html.

[2] F. Chang. Retrieving information from document images: Problems and solutions. *International Journal on Document Analysis and Recognition, Special Issues on Document Analysis for Office Systems*, 4:46–55, 2000.

[3] K. Kise, K. Noguchi, and M. Iwamura. Memory efficient recognition of specific objects with local features. In *Proc. of the 19th International Conference of Pattern Recognition (ICPR2008)*, WeAT3.1, Dec. 2008.

[4] X. Liu and D. Doermann. Mobile retriever - finding document with a snapshot. In *Proceedings of Second International Workshop on Camera-Based Document Analysis and Recognition (CBDAR2007)*, pages 29–34, 2007.

[5] T. Nakai, K. Kise, and M. Iwamura. Camera based document image retrieval with more time and memory efficient llah. In *Proceedings of Second International Workshop on Camera-Based Document Analysis and Recognition (CBDAR2007)*, pages 21–28, 2007.

[6] T. Nakai, K. Kise, and M. Iwamura. Real-time document image retrieval with more time and memory efficient llah. In *Proceedings of Second International Workshop on Camera-Based Document Analysis and Recognition (CBDAR2007)*, pages 168–169, Sept. 2007.