# Camera-based document image mosaicing using LLAH

Tomohiro Nakai, Koichi Kise and Masakazu Iwamura

Graduate School of Engineering, Osaka Prefecture University
1-1 Gakuen-cho, Naka, Sakai, Osaka, 599-8531 Japan

## ABSTRACT

In this paper we propose a mosaicing method of camera-captured document images. Since document images captured using digital cameras suffer from perspective distortion, their alignment is a difficult task for previous methods. In the proposed method, correspondences of feature points are calculated using an image retrieval method LLAH. Document images are aligned using a perspective transformation parameter estimated from the correspondences. Since LLAH is invariant to perspective distortion, feature points can be matched without compensation of perspective distortion. Experimental results show that document images captured by a digital camera can be stitched using the proposed method.

**Keywords:** Document image mosaicing, Camera-based document image processing, LLAH

## 1. INTRODUCTION

In the recent automated office environment, it is common for us to digitize printed documents into electronic documents. Scanners are typically used for this purpose. However, scanners have defects on usability that they lack portability and require connection to computers during scanning. Besides, scanners cannot be used to scan precious documents such as historical documents since scanning can damage them.

Digital cameras have been attracting attention as means of digitizing printed documents. By using digital cameras, documents can be digitized as document images with portable equipments and without direct contact to them. However, at present, digital cameras have relatively lower resolution than scanners. Therefore, in order to obtain document image using a digital camera with as well or higher resolution than scanners, a document have to be captured in several parts separately and the separated document images have to be assembled into a single large document image. Such stitching process is called “image mosaicing”.

In order to obtain a successful mosaic image, two or more document images with overlapped regions have to be aligned appropriately. Moreover, perspective distortion on document images has to be removed. Since a digital camera is used in arbitrary poses, resulting captured images suffer from perspective distortion. It is also required to obtain a mosaic image within reasonable processing time. Although there have been various methods of document image mosaicing, we still do not have one which solves all of the above problems. Most of the previous methods[3,4] are intended to stitch document images captured using a flatbed scanner. Therefore they cannot deal with camera-captured document images due to their perspective distortion. There is also a camera-based document image mosaicing method.[5] However, it requires difficult and burdensome processing to remove perspective distortion on document images. It might limit efficiency of the method. Therefore fast mosaicing of camera-captured document images is still a challenging problem.

In this paper, we propose a fast mosaicing method for camera-captured document images. The proposed method consists of three steps. In the first step, feature points are extracted from two document images and correspondences of them are computed using an image retrieval method called LLAH.[1] In the second step, the perspective transformation parameter to compensate perspective distortion between the two images is estimated. One of the two images is transformed using the parameter to fit the other one. In the third step, the aligned images are merged into a mosaic image. LLAH is known for its efficiency. Retrieval of a document image from the database containing 10,000 pages of documents takes about 30ms.[1] An application of real-time document image retrieval using LLAH is also proposed.[2] Since LLAH is a fast and perspective invariant method, feature points of images with perspective distortion are matched quickly without any preprocessing. In this way, fast mosaicing of camera-captured document images is realized.

## 2. RELATED WORK

There have been several image mosaicing methods. Whichello et al.[3] have proposed a method based on the correlation technique. In this method, the translation of two images $f(i,j)$ and $g(i,j)$ is computed from the cross correlation $c(x,y)$ defined as

$$c(x,y) = \sum_i \sum_j f(i,j)g(i-x,j-y). \tag{1}$$

When (x,y) is equal to the translation of $f(i,j)$ and $g(i,j)$, $c(x,y)$ has the maximum value since $f(i,j)$ and $g(i-x,j-y)$ matches. Then $f(i,j)$ and $g(i,j)$ are stitched using $(x,y)$ to make a mosaic image. The method also includes some ideas to speed up its processing. However, it is essentially unable to deal with distortion of an image other than translation. It is not robust to rotation, scaling and perspective distortion.

Isgró et al.[4] have proposed a feature based image mosaicing method. In this method, feature points are firstly extracted from one of two images to be stitched, and then the corresponding points in the other image are calculated. From the corresponding points, a Euclidean transformation parameter between the two images is estimated. The images are stitched using the Euclidean transformation parameter. In the method, an efficient processing is realized by step-by-step matching of feature points. However, the method is also unable to deal with scaling and perspective distortion because the Euclidean transformation includes only translation and rotation. Moreover, it cannot deal with significant rotation since the correlation technique is used for finding the corresponding points. Therefore the method cannot likewise realize mosaicing of camera captured document images.

Lian et al.[5] have proposed a mosaicing method for camera captured document images. Unlike the above methods, it realizes stitching of document images captured from arbitrary angles using a digital camera. In this method, perspective distortion of document images are firstly removed based on vanishing points estimated from text line direction and vertical character stroke direction. Then feature points of fronto-parallel document images are extracted and matched using PCA-SIFT.[6] Due to periodicity of document images, correspondences of points include many outliers. Hence inconsistent correspondences are filtered out. Then precise alignment is performed using cross-correlation block matching. A mosaic image is created from the aligned images. Since the method has no restriction on pose of the digital camera, it has significant flexibility. However, removing perspective distortion of camera captured document image performed in advance to matching of feature points is not a trivial task. It is difficult to extract textlines and character strokes stably under various kinds of disturbances. Besides, such processing requires much computation. Therefore the method lacks reliability and efficiency due to requirement of perspective distortion compensation.

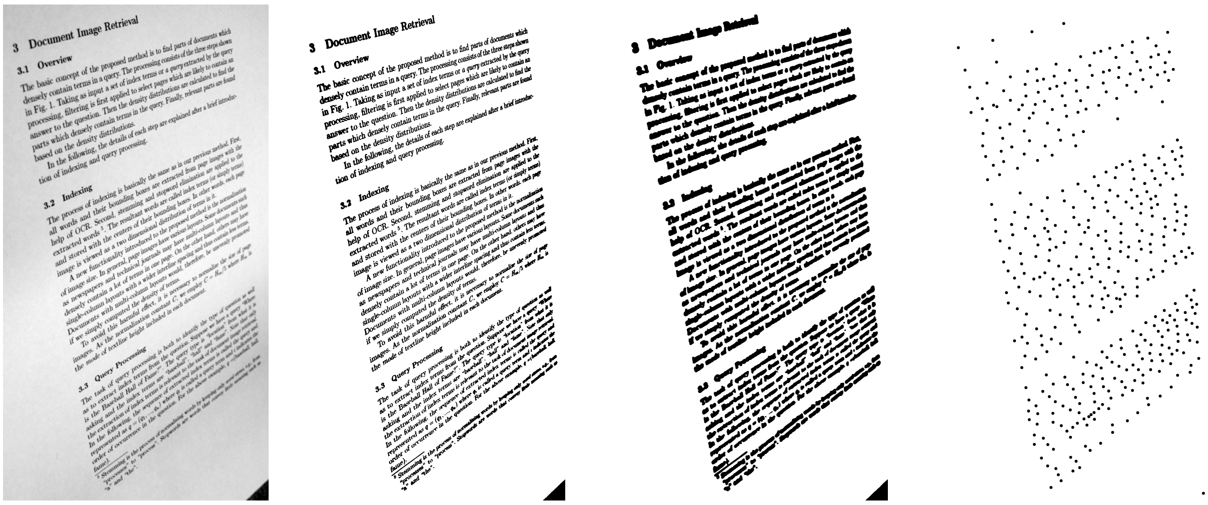
## 3. LOCALLY LIKELY ARRANGEMENT HASHING

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. Correspondences of feature points are subsidiarily obtained in the process of retrieval. Therefore matching of feature points of camera-captured document images can be realized using LLAH.

### 3.1 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.



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

Figure 1. Feature point extraction.

### 3.2 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:

  $$\frac{P(A,C,D)}{P(A,B,C)} \tag{2}$$

  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.

### 3.3 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. As shown in Fig. 2, identification numbers of document images (Document ID) are stored with descriptors. By voting images in the database using the Document ID, the document image corresponding to the query image can be retrieved. Since identification numbers of points (Point ID) are also stored, correspondences of feature points as shown in Fig. 3 can be obtained.

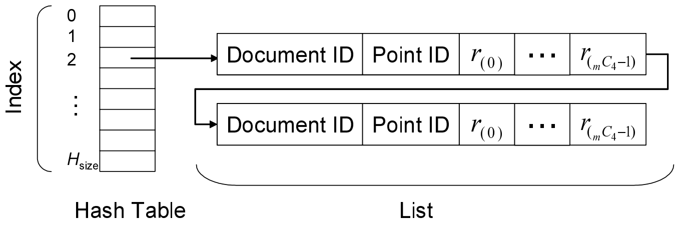


Figure 2. Configuration of the hash table.

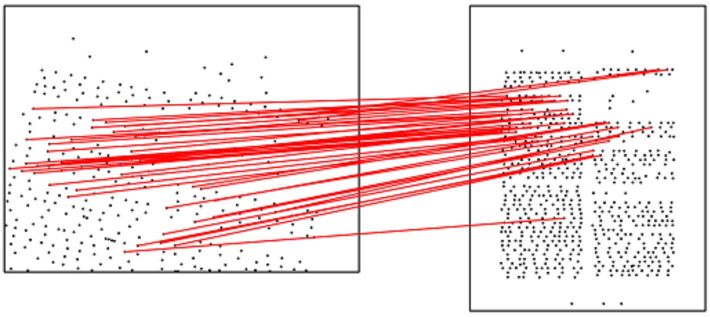


Figure 3. Retrieval result and correspondences of feature points.

## 4. DOCUMENT IMAGE MOSAICING USING LLAH

The proposed method consists of three following steps. Firstly, correspondences of feature points extracted from two camera-captured document images are calculated using LLAH. Then the perspective transformation parameter between planes of the document images is estimated from the correspondences. Finally, the document images are stitched to make a mosaic image.

In the process of retrieval using LLAH, corresponding points of images in the database are obtained for each feature point of a query image. Therefore corresponding feature points of two document images can be obtained by storing one of the two images in the database and retrieving the database using the other as a query image. In this case, only one document image is stored in the database.

The two document images are aligned using the correspondences of feature points. We adopt RANSAC[7] to estimate a perspective transformation parameter from the correspondences. The correspondences can include erroneous ones. RANSAC enables to filter out outliers and get an appropriate parameter which fits only in the majority of correspondences. Using the parameter, one of the two document images is transformed so that two document images on the same plane are obtained.

A mosaic image is created by a simple composition method. A mosaic image $h(x, y)$ is defined from two images $f(x, y)$ and $g(x, y)$ as follows,

$$h(x,y) = \begin{cases} f(x,y) & (x,y) \in F \text{ and } (x,y) \notin G \\ g(x,y) & (x,y) \notin F \text{ and } (x,y) \in G \\ \frac{f(x,y)+g(x,y)}{2} & \text{otherwise} \end{cases} \quad (3)$$

where $F$ and $G$ means regions of $f(x, y)$ and $g(x, y)$, respectively.

## 5. EXPERIMENTAL RESULTS

The proposed method was tested on four pairs of camera-captured document images. Document images were captured using Canon EOS 5D. Size of document images is 4368×2912. Image mosaicing was performed on a PC

Table 1. Average processing time.

| | Processing time [ms] |
|---|---|
| Extracting feature points | 8,464 |
| Matching feature points | 571 |
| Estimating transformation parameter | 2 |
| Creating mosaic image | 3,139 |

Table 2. The numbers of feature points.

| | Image 1 | Image 2 | Matched |
|---|---|---|---|
| Case 1 | 336 | 519 | 156 |
| Case 2 | 313 | 515 | 94 |
| Case 3 | 327 | 645 | 66 |
| Case 4 | 322 | 457 | 7 |

with 2.2GHz CPU and 2GB memory. Parameters mentioned in **3.2** were set to $n = 7$ and $m = 6$. Figures 4 – 7 show results.

Average processing time of each step is shown in Table 1. Since images used in this experiment have high resolution, extracting feature points which includes image processing took much time. However, it can be speeded up easily by scaling down images before feature point extraction.

Table 2 shows the numbers of feature points in the experiments of Figs. 4 – 7. In Table 2, Image 1 and Image 2 indicate the document images shown in top left and top right of Figs. 4 – 7, respectively. Table 2 also shows the numbers of matched feature points in middle rows of Figs. 4 – 7.

Figures 4 and 5 show successful cases. Many correct corresponding points are obtained by LLAH. Document images are successfully aligned using the perspective transformation parameter estimated from the correspondences. Note that the input image shown in Fig. 5 suffers from perspective distortion. It shows robustness of the proposed method to certain degree of perspective distortion.

In Figs. 6 and 7, slight misalignment occurs in local areas. We consider the slight misalignment is caused by change in position of feature points and non-perspective distortion. In the proposed method, feature points are centroids of word regions. However, the position of a centroid is affected by perspective distortion. In order to extract word regions, pixels of a word are connected by image processing including Gaussian filter. This process is applied to an image under perspective distortion. Calculation of centroids is also performed under perspective distortion. Therefore document images under different perspective distortion have slightly different centroids of word regions. In addition, alignment of images is performed by perspective transformation. However camera-captured document images suffer from non-perspective distortion such as lens distortion, which cannot be normalized by perspective transformation.

Errors as shown above are more likely to be caused under more significant perspective distortion. Therefore additional measures to compensate slight misalignment are necessary. For example, the correlation technique used in the previous methods could be effective.

Even though additional measures are required, the proposed method is still meaningful. When images are aligned roughly, search region of correlation can be limited. Therefore only a minor additional process may be required for precise alignment. Note also that mosaicing of excessively distorted document images as shown in Fig. 7 is almost inconceivable. When a user captures a printed document for mosaicing, he/she will hold a camera roughly in front of the document. As shown in Fig. 7, most part of an excessively distorted document image can be out of focus, which spoils the resultant image. Hence we can say that robustness to certain degree of perspective distortion is enough for a document image mosaicing method.

## 6. CONCLUSION

In this paper we proposed a mosaicing method of camera-captured document images. Since document images captured using digital cameras suffer from perspective distortion, alignment of them is a difficult task for previous methods. In the proposed method, correspondences of feature points are calculated using an image retrieval method LLAH. Document images are aligned using a perspective transformation parameter estimated from the correspondences. Since LLAH is invariant to perspective distortion, feature points can be matched without compensation of perspective distortion.

Experimental results show that document images captured by a digital camera can be stitched using the proposed method. Slight misalignment is confirmed under significant perspective distortion. However, it would be solved by some additional measures. Additional measures for precise alignment are included in our future work.

Although the experimental results show effectiveness of the proposed method, they do not include detailed evaluation. Quantitative evaluation of experimental results should also be done in our future work.

## REFERENCES

[1] Nakai, T., Kise, K., and Iwamura, M., "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)*, 21–28 (2007).

[2] Nakai, T., Kise, K., and Iwamura, M., "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)*, 168–169 (Sept. 2007).

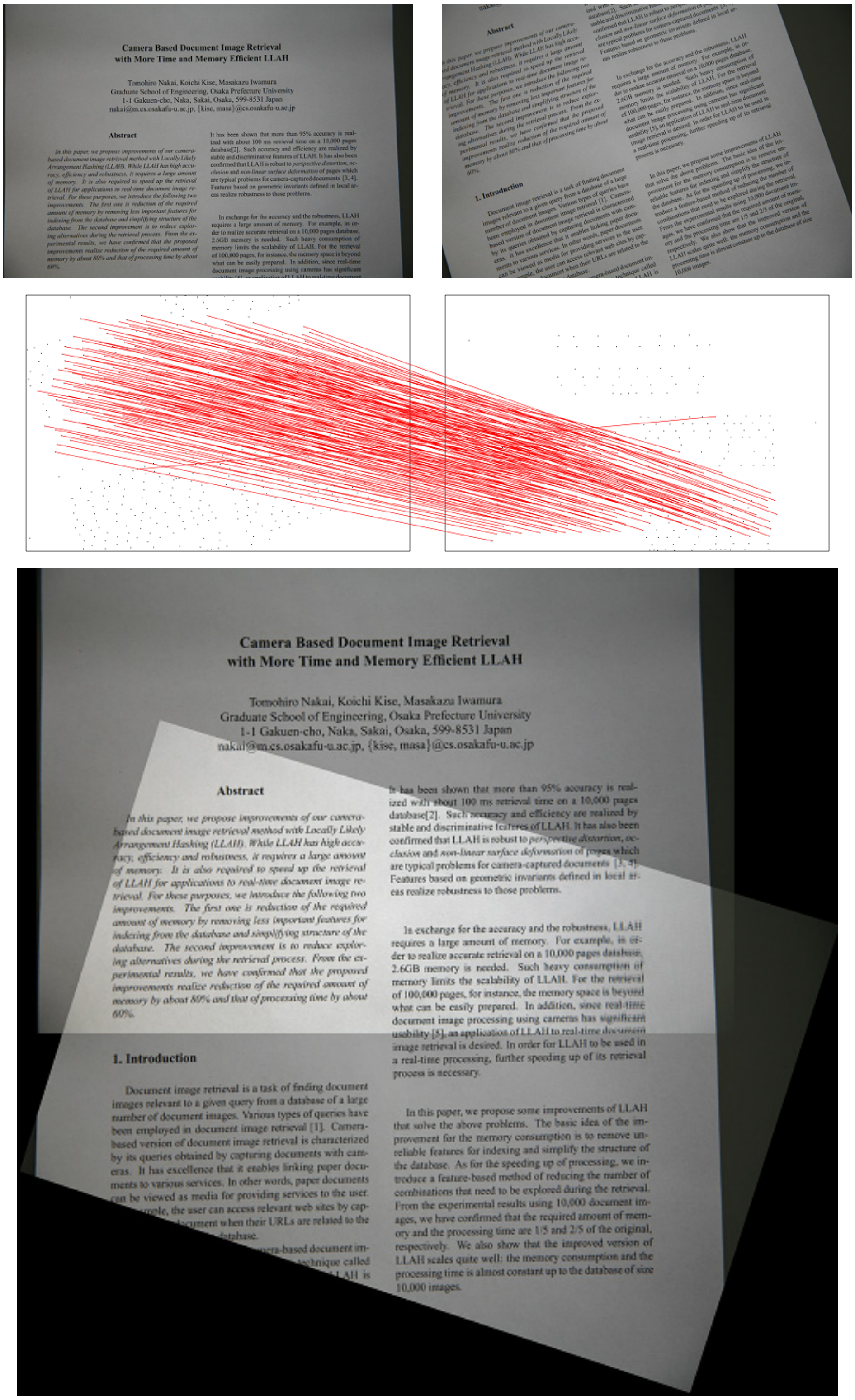
[3] Whichello, A. P. and Yan, H., "Document image mosaicing," in *Proceedings of ICPR 1998*, **2**, 1081–1083 (1998).

[4] Isgró, F. and Pilu, M., "A fast and robust image registration method based on an early consensus paradigm," *Pattern Recognition Letters* **25**(8), 943–954 (2004).

[5] Lian, J., DeMenthon, D., and Doermann, D., "Camera-based document image mosaicing," in *International Conference on Pattern Recognition*, (2006).

[6] Ke, Y. and Sukthankar, R., "Pca-sift: A more distinctive representation for local image descriptors," in *Proceedings of CVPR 2004*, **2**, 506–513 (2004).

[7] Fischler, M. A. and Bolles, C., "Random sample consensus: A paradigm for model fitting with application to image analysis and automated cartography," *Communications of ACM* **24**(6), 381–395 (1981).



Camera Based Document Image Retrieval
with More Time and Memory Efficient LLAH

Tomohiro Nakai, Koichi Kise, 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

*In this paper, we propose improvements of our camera-based document image retrieval method with Locally Likely Arrangement Hashing (LLAH). While LLAH has high accuracy, efficiency and robustness, it requires a large amount of memory. It is also required to speed up the retrieval of LLAH for applications to real-time document image retrieval. For these purposes, we introduce the following two improvements. The first one is reduction of the required amount of memory by removing less important features for indexing from the database and simplifying structure of the database. The second improvement is to reduce exploring alternatives during the retrieval process. From the experimental results, we have confirmed that the proposed improvements realize reduction of the required amount of memory by about 80% and that of processing time by about 60%.*

1. Introduction

Document image retrieval is a task of finding document images relevant to a given query from a database of a large number of document images. Various types of queries have been employed in document image retrieval [1]. Camera-based version of document image retrieval is characterized by its queries obtained by capturing documents with cameras. It has excellence that it enables linking paper documents to various services. In other words, paper documents can be viewed as media for providing services to the user.
...mple, the user can access relevant web sites by cap-
...cument when their URLs are related to the
...database.
...era-based document im-
...technique called
...LAH is

It has been shown that more than 95% accuracy is realized with about 100 ms retrieval time on a 10,000 pages database[2]. Such accuracy and efficiency are realized by stable and discriminative features of LLAH. It has also been confirmed that LLAH is robust to *perspective distortion, occlusion* and *non-linear surface deformation* of pages which are typical problems for camera-captured documents [3, 4]. Features based on geometric invariants defined in local areas realize robustness to those problems.

In exchange for the accuracy and the robustness, LLAH requires a large amount of memory. For example, in order to realize accurate retrieval on a 10,000 pages database, 2.6GB memory is needed. Such heavy consumption of memory limits the scalability of LLAH. For the retrieval of 100,000 pages, for instance, the memory space is beyond what can be easily prepared. In addition, since real-time document image processing using cameras has significant usability [5], an application of LLAH to real-time document image retrieval is desired. In order for LLAH to be used in a real-time processing, further speeding up of its retrieval process is necessary.

In this paper, we propose some improvements of LLAH that solve the above problems. The basic idea of the improvement for the memory consumption is to remove unreliable features for indexing and simplify the structure of the database. As for the speeding up of processing, we introduce a feature-based method of reducing the number of combinations that need to be explored during the retrieval. From the experimental results using 10,000 document images, we have confirmed that the required amount of memory and the processing time are 1/5 and 2/5 of the original, respectively. We also show that the improved version of LLAH scales quite well: the memory consumption and the processing time is almost constant up to the database of size 10,000 images.

Figure 4. Experimental result (1). Top row shows input images. Middle row shows correspondences of feature points obtained by LLAH. Bottom row shows the resulting mosaic image.

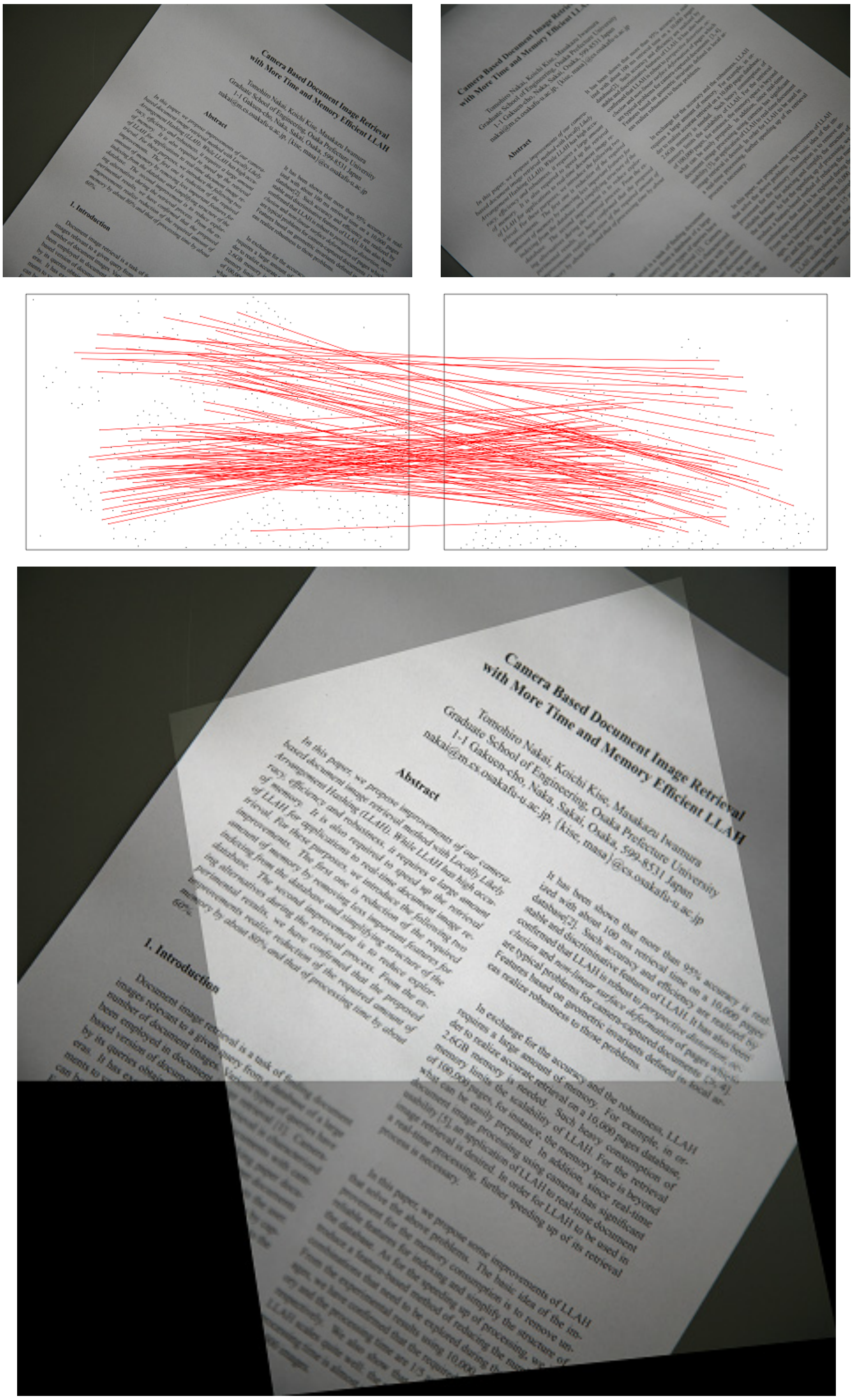


Figure 5. Experimental result (2). Top row shows input images. Middle row shows correspondences of feature points obtained by LLAH. Bottom row shows the resulting mosaic image.

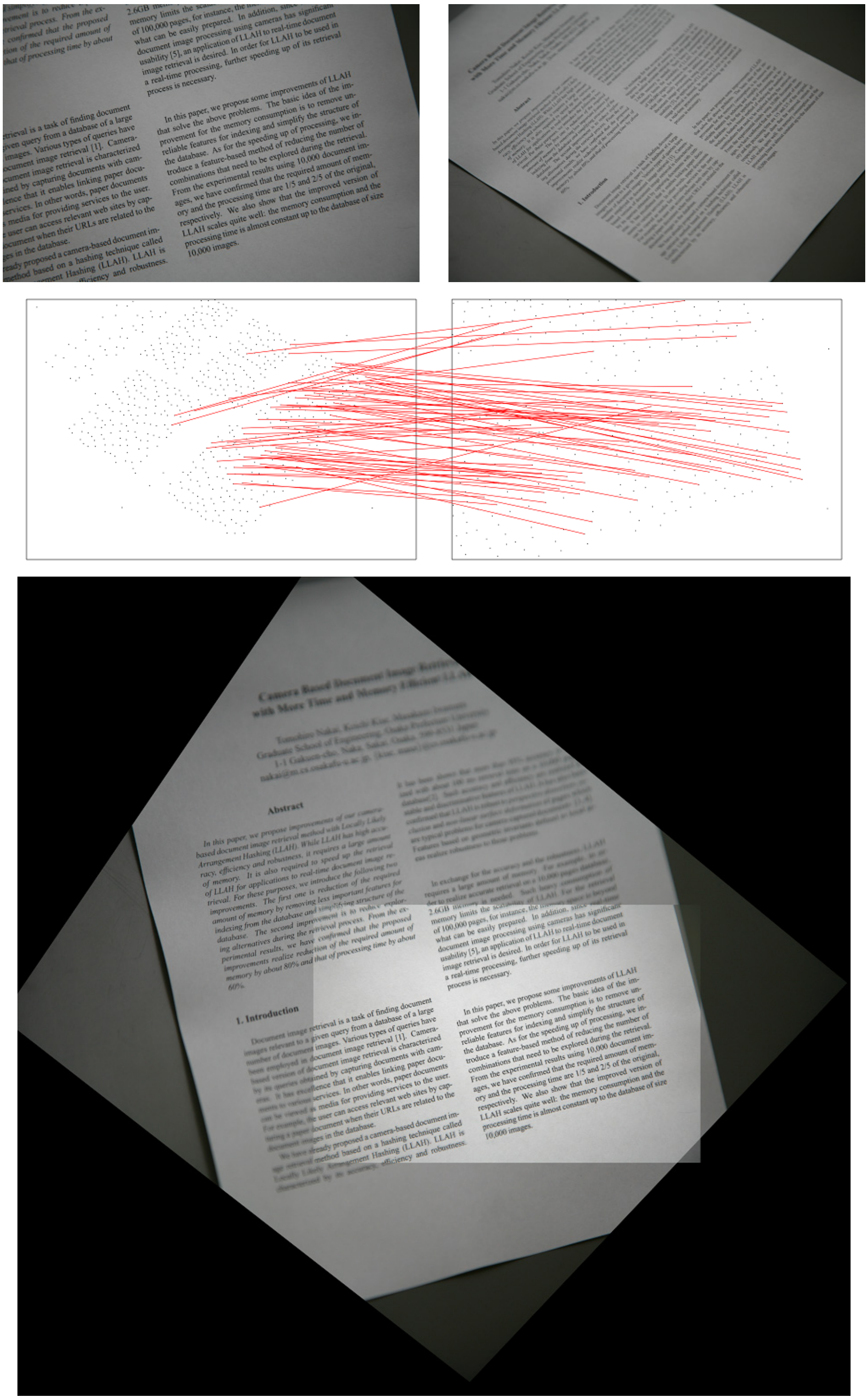


Figure 6. Experimental result (3). Top row shows input images. Middle row shows correspondences of feature points obtained by LLAH. Bottom row shows the resulting mosaic image.

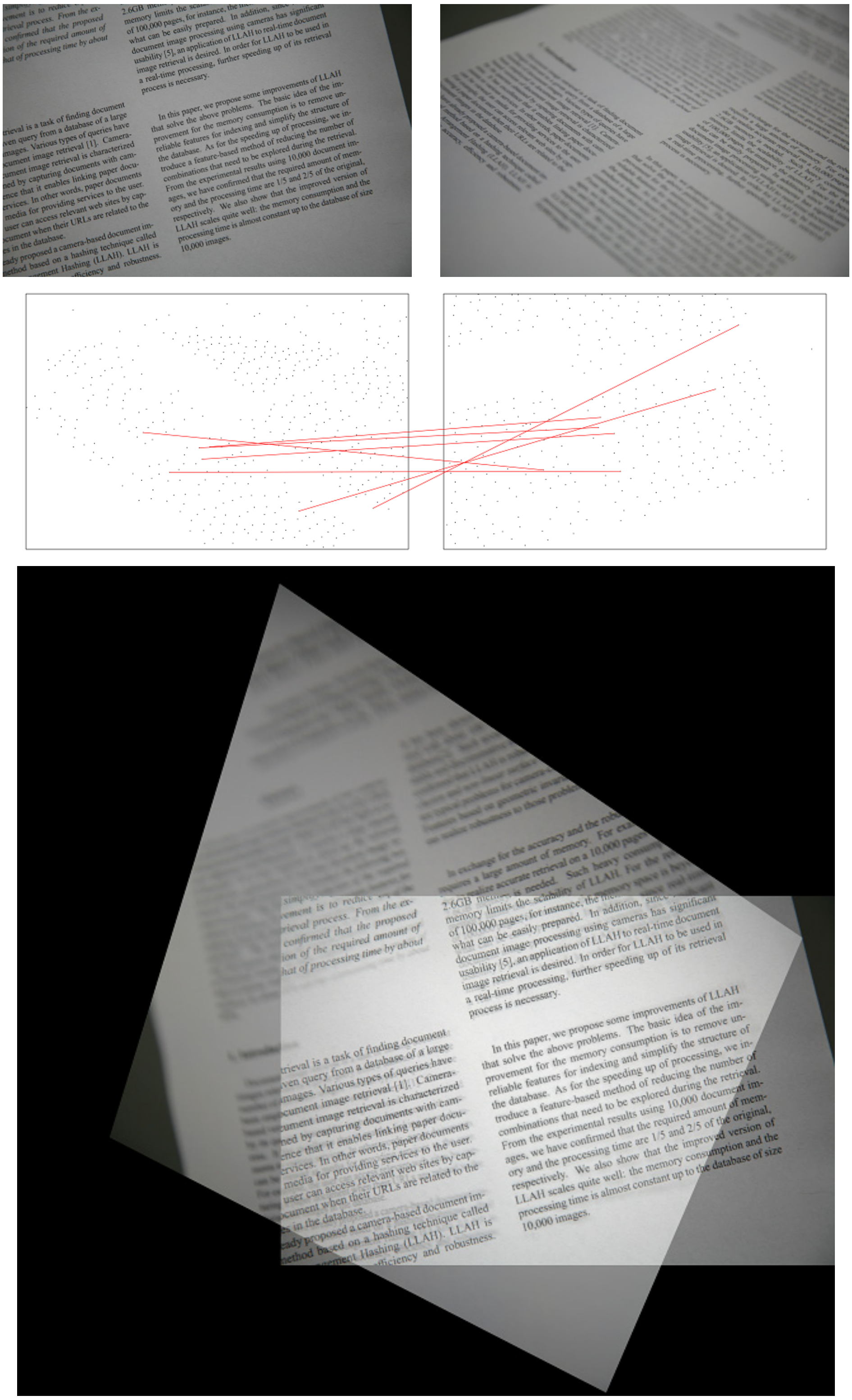


Figure 7. Experimental result (4). Top row shows input images. Middle row shows correspondences of feature points obtained by LLAH. Bottom row shows the resulting mosaic image.