A Method of Annotation Extraction from Paper Documents Using Alignment Based on Local Arrangements of Feature Points

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

Abstract

Annotations on paper documents include important information. We can exploit the information by extracting and analyzing annotations. In this paper, we propose a method of annotation extraction from paper documents. Unlike previous methods which limit colors or types of annotations to be extracted, the proposed method attempts to extract annotations by comparing a document image of an annotated document with its original document image for removing the limitations. The proposed method is characterized by fast matching and flexible subtraction of images both of which are essential to the annotation extraction by comparison. Experimental results have shown that color annotations can be extracted from color documents.

1. Introduction

Since paper documents have high legibility and portability, they are widely used as media for communication. Furthermore, it is common for us to mark up or annotate paper documents. Therefore annotations on paper documents have important information such as users' interests or knowledge. Such valuable information can be obtained by extracting and analyzing annotations on paper documents.

Several methods of annotations extraction have been proposed [1, 2, 3, 4]. These methods have achieved high extraction rates by limiting either colors or types of annotations. Since there are no limitation in annotating documents in the actual environment, annotation extraction methods with few limitations on annotations are needed.

Since most of printed documents are originated from their electronic version such as PDFs, a possible way is to assume the availability of the electronic originals without annotations and employ them to extract annotations. This type of annotation extraction is applicable when owners of electronic documents try to extract annotations from printed version of the documents. An example is as follows. In our daily business activities, we often distribute printed electronic documents for correction and collect annotated paper documents. In this case, annotations can be automatically extracted by the annotation extraction method since the original electronic documents are available.

For this strategy, problems to be solved are threefold: (1) how to make alignment of a document image with annotations (annotated document image) and its corresponding document image without annotations (original document image), (2) how to subtract two images with less influence of noise and errors of alignment, and (3) how to deal with color documents which may be annotated using color pens.

In this paper, we propose a method which attempts to solve the above problems. For the first problem, we employ a method of fast point matching based on local arrangements of feature points [5]. This method is characterized by the fast processing based on an unique indexing method using geometric invariants and the access to the indices using a hash table. For example matching 400 points in a page to the database including 10,000 pages needs less than 100 ms. [5]. For the second problem, we employ a simple search for finding the best match pixel within a limited local area. For the third problem, we utilize color clustering to deal with color documents as a collection of several color planes. This is also required for applying the point matching method designed for black-and-white pages. From the experimental results, it has been shown that most annotations can be extracted regardless of colors and types of annotations.

2. Related work

In order to extract annotations from printed documents, several methods have been proposed [1, 2, 3, 4]. They can be classified into the following two types. One is a group of methods that focus not on the extraction of annotations but on the utilization of extracted annotations, namely the auto-

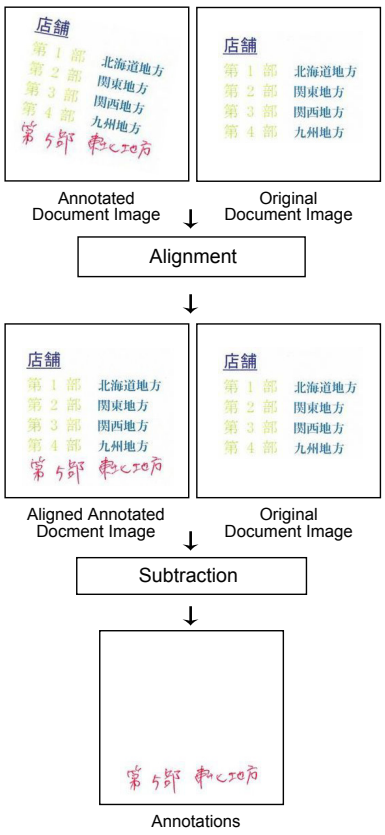


Figure 1. Overview of processing.

mated corrections of documents using annotations [1, 2]. In these methods, the color of annotations are predetermined to make the extraction easy and accurate. Although the correction is successful, such methods have the limitation on colors of annotations as well as documents. The other is a group of methods that focus on classification of connected components of annotated document images into handwritten and machine printed text[3, 4]. Although these methods have the advantage that they can extract handwritten annotations only from the annotated document images, extracted annotations are limited to characters; they cannot extract handwritten line drawings which are also frequently utilized as annotations.

3. Proposed method

3.1. Overview of processing

Our goal is to realize annotation extraction from annotated documents under the assumption that electronic version of original documents are available. It is easily conceived that annotations can be extracted by subtracting images of original documents from images of annotated documents. However this requires precise alignment of two images which is not a trivial task due to the influence of annotations. Different imaging conditions caused by the skew and translation of physical pages make the problem more difficult.

To solve the above problem, the proposed method employ the following two steps: “alignment” and “subtraction”, which are shown in Fig. 1 and described in the following sections.

3.2. Alignment process

In this process, an annotated document image is aligned with its original document image. In the process of alignment, it is required to correct the similarity transformation: translation, rotation and scaling of one image to the other. Handwritten annotations on the annotated document image generally disturbs the estimation of parameters of transformation.

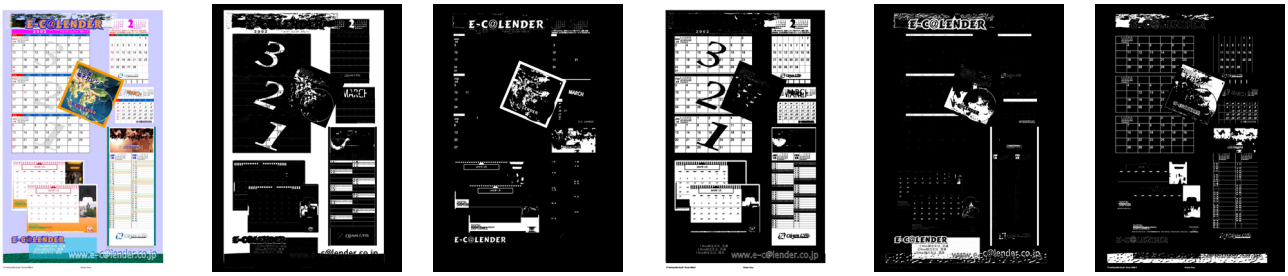
In order to solve this problem, we employ a matching method based on local arrangement of feature points [5]. The outline of processing is as follows. First the feature points are extracted from the annotated document image. Next points are matched to points extracted in advance from the original document image. Then resultant matches of points are utilized to estimate the parameters of similarity transformation. In this processing, the most important part is to find the point matching. It is prohibitive to try all possible matches due to the combinatorial explosion. The method of matching [5] enables us a fast processing whose order is almost $O(N)$ where N is the number of feature points.

Because the matching method [5] is designed for black-and-white images and feature points are extracted as centroids of word regions, additional processing is required for applying this method to color images that may also contain graphics and line drawings. Details are described below.

3.2.1 Feature point extraction

Points which characterize images are extracted as feature points both from the annotated document image and the original document image. Feature points should be robust to degradation. In the proposed method, centroids of connected components of each color obtained by color clustering in the RGB color space are used as feature points. Locations of plain colored regions such as characters are intended to be used as feature points. Arrangements of characters are distinctive and they can be easily extracted stably since their colors contrast with colors of their background.

Feature points are extracted as follows. A color clustering technique (k-means algorithm) is applied to the image.



(a) input image (b) images of color clusters

Figure 2. An image is decomposed into k images of color clusters.($k = 5$)

Figure 3. An example of point matching.

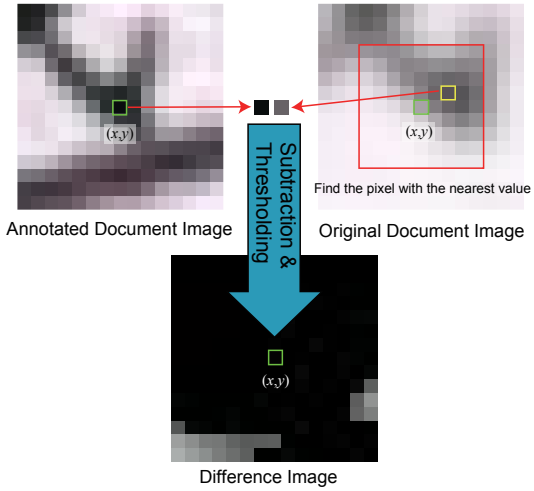


Figure 4. Subtraction and thresholding.

Based on the result of color clustering, the image is decomposed into k monochrome images as shown in Fig. 2. Connected components are extracted from these images. Centroids of connected components are then used as feature points for each color cluster.

3.2.2 Matching of feature points

Feature points of the annotated document image are matched with those of the corresponding original document image. By this stage, both the annotated document image and the original document image have k sets of feature points corresponding to k color clusters. Each color cluster of the annotated document image is matched with that of the original document image in advance of point-to-point matching. In the proposed method, a pair of color clusters which have smaller distance in the RGB color space are matched preferentially using the greedy algorithm.

Then point-to-point matching is performed between paired color clusters with the help of the matching method of [5]. An example of matching of feature points is shown in Fig. 3. As a result many pairs of points are obtained.

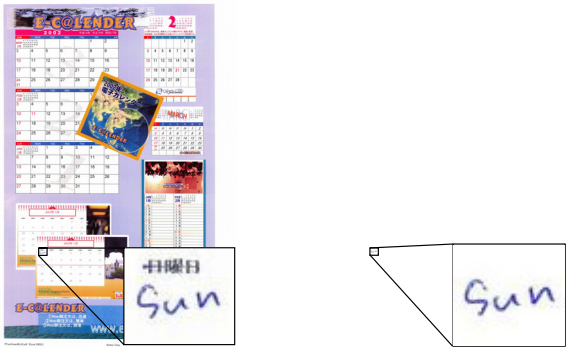
The next step is to estimate the parameters of similarity transformation. Note that, as shown in Fig. 3, the matching may contain outliers. In order to suppress their influence, the RANSAC algorithm [6] is applied to the match of feature points to estimate the parameters. Then the annotated document image is transformed using the parameters to make an aligned annotated document image.

3.3. Subtraction process

In this process, annotations are extracted by subtracting the original document image from the aligned annotated document image.

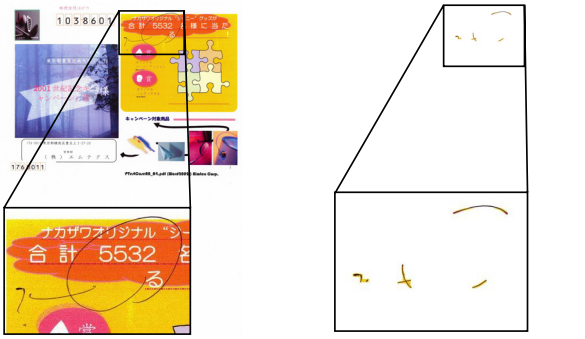
The aligned annotated document image and the original document image are compared for each pixel to apply subtraction and thresholding. In order to deal with misalignment caused by distortion during scanning and errors of feature point extraction, we have devised a subtraction method as follows.

Firstly, each pixel of the annotated document image is compared with its corresponding pixel of the original document image. As shown in Fig. 4, the corresponding pixel is



(a) Annotated document image (b) Extracted annotations

Figure 5. An example of successful cases.



(a) Annotated document image (b) Extracted annotations

Figure 6. An example of partially lost or noisy cases.

the pixel which has the nearest pixel value in the square area centered at the same coordinate. Small misalignment can be absorbed by finding the pixel of a similar value. Based of this correspondence, values of pixels of the original document image is subtracted from those of corresponding pixels of the annotated document image, so that a difference image can be obtained.

Next, a mask image is created from the difference image. Annotations are extracted by "AND" operation of the mask image and the aligned annotated document image.

4. Experimental results

Experiments were performed using 109 mostly color pages. In the experiments, original document images are derived from PDF files of the pages. To be precise, we applied the following procedure for obtaining the original document images. Although it is ideal to employ original document images directly produced from the PDF files, the resultant document images are far from the annotated document images. In order to make the images close enough it is required to employ image degradation models about scanning. In the experiments here this process was just skipped by also scanning original document images. Note that scanning of annotated and original document images was performed using a sheet feeder. Thus the resultant document images often suffer from similarity transformation.

Annotated document images are created by writing characters and line drawings using black, red and blue ball pens on several parts of the printed documents and scanning them in 600 dpi. For each page we apply three different patterns of annotations. Thus, in total, the number of annotated document images was 327.

The number of color clusters was set to $k = 5$. Parameters of the method of point matching [5] were set to $n = 5$, $m = 5$.

(a) Annotated document image (b) Extracted annotations

Figure 7. An example of failure cases.

Results of extraction were evaluated for each page. Extracted annotations were classified into three types: success for a whole page, partially lost or noisy, and failure as follows.

Success Most annotations are extracted without noises. (Fig. 5)

Partially lost or noisy Annotations are partially lost or some noises are included. (Fig. 6)

Failure Most annotations are lost or noises appear in the whole area. (Fig. 7)

Table 1 shows the results of processing, where original document images were classified into two types: with many "plain colored regions" and few of them. Because the method of alignment relies on the feature points extracted from plain colored regions, the latter is much harder to obtain correct answer.

For the former type of documents, 78% were successful and only 10% were classified into failure cases. In the suc-

Table 1. Experimental results.

	Success	Partially lost or noisy	Failure	Total
Many plain colored regions	103 (78%)	16 (12%)	13 (10%)	132 (100%)
Few plain colored regions	108 (55%)	34 (17%)	53 (27%)	195 (100%)
Total	211 (65%)	50 (15%)	66 (20%)	327 (100%)

Figure 8. Examples of images with gradation or pictures.

cessful cases, complicated pages with a lot of decorations such as shown in Fig. 5 are included. For some applications such as browsing of annotations by human, it would be enough to obtain partially lost or noisy results as shown in 6. In such a case the proposed method extract annotations from 90% of the pages. Although there still remains room for improvement, we consider that the results are promising as the first step toward the annotation extraction from color documents.

For the documents with few plain colored regions, on the other hand, the failure rate increased to 27%. This was mainly due to the figures with gradation or pictures as shown in Fig. 8. For such regions, results of color clustering varied due to slight change of scanning conditions, and thus the feature points were unstable. In order to solve this problem, it is required to employ a different method of feature point extraction.

In total, annotations in 65% of the total pages were correctly extracted. This means that the task of extracting color annotations from color document images are not easy task, but for documents with plain colored regions including the one shown in Fig. 5, the method is capable of extracting annotations from complicated colored documents.

5. Conclusion

In this paper, we proposed an annotation extraction method from paper documents. In this method, annotations are extracted by subtracting original document images from aligned annotated document images under the assumption that original electronic documents are available. The proposed method is characterized by the fast alignment process based on the point matching method and the flexible subtraction process.

The experimental results have shown that the proposed method was successful for extracting color annotations from documents printed in color with complicated foreground. The accuracy of 78% was obtained for images with many plain colored regions such as characters.

Future work includes improvement of feature point extraction in order to extract stable feature points from images which have few characters. It might be effective to utilize fruits of work in the field of computer vision such as Harris operators[7].

Acknowledgement

This research is supported in part by Grant-in-Aid for Scientific Research from Japan Society for the Promotion of Science (19300062, 19·7621).

References

[1] D. Möri and H. Bunke. Automatic interpretation and execution of manual corrections on text documents. In H. Bunke and P. S. P. Wang, editors, *Handbook of Character Recognition and Document Image Analysis*, pages 679–702. World Scientific, Singapore, 1997.

[2] J. Stevens, A. Gee, and C. Dance. Automatic processing of document annotations. In *Proc. 1998 British Machine Vision Conf.*, volume 2, pages 438–448, 1998.

[3] J. K. Guo and M. Y. Ma. Separating handwritten material from machine printed text using hidden markov models. In *Proc. 6th International Conf. on Document Analysis and Recognition*, pages 436–443, 2001.

[4] Y. Zheng, H. Li, and D. Doermann. The segmentation and identification of handwriting in noisy document images. In *Lecture Notes in Computer Science (5th International Workshop DAS2002)*, volume 2423, pages 95–105, 2002.

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

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

[7] C. Harris and M. Stephens. A combined corner and edge detector. In *Proc. 4th Alvey Vision Conf.*, pages 147–151, 1988.