Analysis of Annotations on Documents for Recycling Information in Collaborative Business Activities

Koichi Kise, Nobuyuki Kondo, Tomohiro Nakai, Keinosuke Matsumoto

Department of Computer Science and Intelligent Systems,
Graduate School of Engineering, Osaka Prefecture University
1-1 Gakuencho, Sakai, Osaka 599-8531, Japan
kise@cs.osakafu-u.ac.jp

Abstract

In order to make collaborative business activities fruitful, it is essential to know characteristics of organizations and persons in more details and to gather information relevant to the activities. In this paper, we propose a notion of "information recycle" that actualizes these requirements by analyzing documents. The key of recycling information is to utilize annotations on documents as clues for generating users' profiles and for weighting contents in the context of the activities. We also propose a method of extracting annotations on paper documents just by pressing one button with the help of techniques of camera-based document image analysis. Experimental results demonstrate that it is fundamentally capable of acquiring annotations on paper documents on condition that their electronic versions without annotations are available for the processing.

1 Introduction

Promotion and encouragement of collaborative business activities require knowing more about business partners starting from the whole organizations such as companies or divisions and ending up each person of the organizations. Electronic means that help acquiring knowledge about organizations and persons are often referred to as "modelling" or "profiling".

Since a huge amount of information have already been on networks such as the Internet, people may think that it is possible to capture most of the activities of organizations and persons by analyzing on-line transactions and flow of information such as document transmission. However it covers only limited aspects of the activities; in addition to on-line communications, people often use ordinary off-line media as means of formal and informal communications. A considerable amount of implicit knowledge that is important for modelling and profiling organizations and persons is scattered into such media. It is therefore necessary not to limit the media to pure electronic ones but to cover seamlessly a wide variety of media for making the profiles more fruitful (Dengel, Junker & Weisbecker, 2004).

Another and more important problem for supporting collaborative business activities is how to cope with the information overload caused by a huge amount of available information. Although the amount on both electronic and paper media is rapidly growing, our ability of dealing with information improves only at a considerably lower rate. Finding right information at the right time, or just-in-time handling of information is lifeblood in collaborative business activities. However, it is not so easy to put it into practice due to the above problem of the media.

In this paper, we propose a new notion called the *recycle of information* for boosting our ability of information handling. The key of the information recycle is both how to know users' and organizations' profiles and how to weight information from the context of collaborations without forcing users to do extra work. To this end, we focus on roles of annotations on documents, since annotating documents is a natural way of interacting with their contents and it often reflects the context of activities and the view of the user. In other words, the analysis of annotations enables us both to obtain user profiles and to weight contents of documents in a context sensitive manner.

From the viewpoint of not disturbing users, it is necessary to obtain information about the annotations based on a natural way with the minimum effort of users. As a possible way to meet this requirement, we propose the analysis and extraction of annotations from camera-captured images of documents. The results of preliminary experiments

show that it is fundamentally capable of acquiring annotations from images of documents on condition that their electronic originals are available for processing.

The organization of this paper is as follows. In Section 2, we define the recycle of information and the roles of annotations in it. Section 3 is devoted to describe our camera-based approach to the analysis of annotations. In Section 4, we show some results of preliminary experiments about extracting annotations from camera-captured images of documents. Section 5 is to conclude what we have learned from the discussion of recycling information as well as the results of experiments.

2 Recycling Information and Annotations

Let us begin with discussing the notion of recycling information for collaborative business activities and the roles of annotations for realizing the recycle.

2.1 Recycle of Information

Environmental issues are often discussed by referring to several terms whose first letter is R. The number of words depends on authors / speakers: from 3R's to 7R's. We focus here on the 4R's version: *Reduce*, *Refuse*, *Reuse* and *Recycle* whose meanings are summarized as follows:

- *Reduce* refers to decreasing the amount of items consumed.
- *Refuse* indicates another possibility of reducing the amount by not receiving unnecessary items.
- *Reuse* means the activity not to destroy items by using them again.
- *Recycle* is a process of transforming items into resources or materials for reproducing new items.

These activities have their own roles for protecting environments in our life.

In addition to the above issues, we also have different environmental issues in the IT world. For example, in our everyday life we are facing a huge amount of unwanted e-mails, or SPAMs. On the Internet, we have a huge number of web pages that are far beyond browsing. The Google currently indexes about 8 billion pages, which are about three times as many as the amount of pages a person can browse in his/her whole life (70 years) with the pace of one page browse per second. Such a situation of the Internet is often referred to as "information flood" or "information overload". In addition, the Internet is sometimes said to be a huge dump of information garbage, because most of the information on the Internet is useless to a lot of people. However it is also worth noting that such information is sometimes very useful to people with a special interest.

The same holds true for information environments in business activities. In the business scenes, we also receive a lot of SPAMs and other e-mails that are not really meaningful to business activities. We are now receiving a great deal and kind of information to which we have to spend time to determine whether or not it is worth reading. We also often produce a lot of documents which do not necessarily interest all people in the organization we belong to. This results in deteriorating further the information environment in business activities.

In order to solve the above problems, we would like here to consider applying 4R's to the information environment as follows:

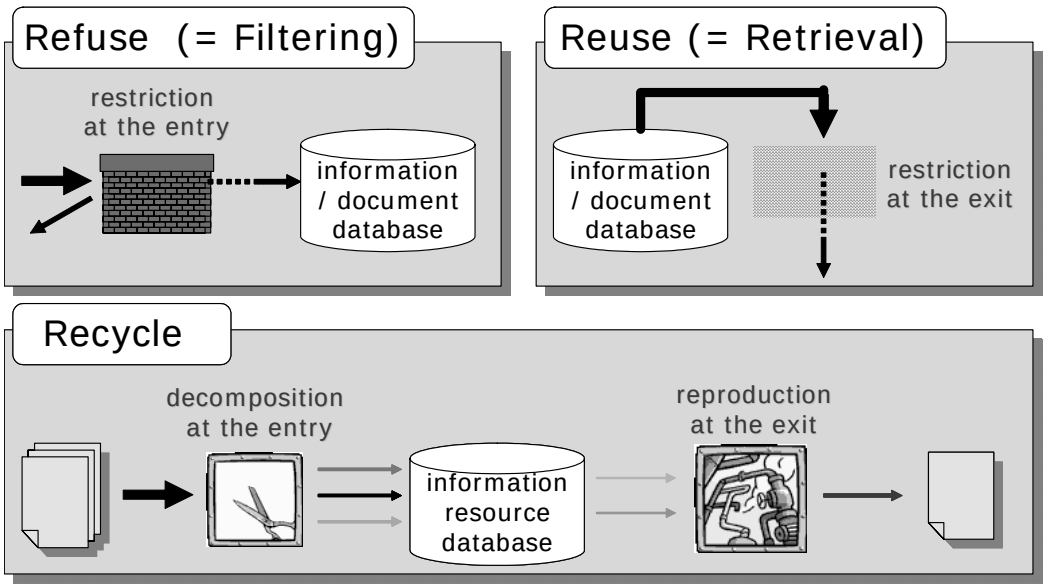


Figure 1: Refuse, reuse and recycle of information.

- Reduce: On the Internet it is almost impossible to control people who produce unwanted information. In the business activities, "reduce" can be considered to be the selection of right recipients of the information.
- Refuse corresponds to the research field called *filtering* or *routing* of information as shown in Fig.1. A good example is a SPAM filter for e-mails.
- Reuse means the repetitive use of information as it is. In this sense, *retrieval* of information or documents corresponds to "reuse" in the information environment.
- Recycle indicates the decomposition of information into "resources" and the reproduction of new information from the decomposed resources. There are no exactly corresponding techniques or research fields for the recycle of information.

The first three are to deal with information as a whole and as it is, such as to send or not, to select or not and to retrieve or not. In this sense, the information in these three is from the viewpoint of its producer(s). On the other hand, the last one "recycle" indicates the decomposition and the reproduction. Thus the information generated by "recycle" is from the viewpoint of its recipient(s). In order to ease the problem of information environments, not only the first three but also the last one should be pursued.

In the context of business activities that require collaborations, the recycle of information allows us to access information essential to our business activities in a user-adapted form. The process of user adaptation requires user profiles which also play an important role to support collaborations.

In order to realize the recycle of information, it is at least necessary to solve the following issues:

- How to decompose compound information into reusable resources.
- How to construct user profiles.
- How to reproduce meaningful information based on the decomposed information resources and user profiles.

For the issue of decomposition, a lot of efforts have so far been made. Major topics related to the decomposition include automated text summarization (Mani & Maybury, 1999), information extraction (Pazienza, 1999), mining of text and web (Chakrabarti, 2003), and structuring of web (Ohno, Maeda, Kise & Matsumoto, 1999), (Nanno, 2003). For the construction of user profiles, efforts have been made especially in the context of Web (Ohno, Maeda, Kise &

Matsumoto, 1999), (Cingil, Dogac & Azgin, 2000) and information retrieval (Light & Maybury, 2002). In contrast to these two topics, only few attempts have so far been made on the last issue. Another important point is that most of the above efforts are on manipulation of electronic media: user profiles are extracted through the interaction with the computer such as clicks, browsing and type-in. We consider that this covers only a limited aspect of collaborative business activities.

2.2 Roles of Annotations

In our daily business activities, we deal with a lot of documents; both reading and writing documents are indispensable parts of the activities. When we write new documents, it is relatively rare to make them from scratch; in most of the cases we have some references that serve as a foundation of the new documents. Thus writing documents can also be regarded as the process of recycling information. When we read documents, we sometimes write annotations on the margin of documents to represent our view of the contents. We consider that the annotations provide us fruitful clues for realizing recycle of information in collaborative business activities as follows:

- Annotations as a source of user profiles: annotations often reflect a user's view to the contents of a document. They represent what the user is or is not interested in. Thus clues to user profiles can be obtained by analyzing the annotations.
- Annotations as a source of weights on contents: if the profile of a user has already been described, his / her annotations can be considered to be weights to the contents of a document from the viewpoint of the profile.
- Annotations as a source of collaborations: documents are often read by multiple people. Since annotations reflect users' interests, users may share their interests if they annotate the common parts of a document.

3 Analysis of Annotations

In this section, we describe the proposed method of analyzing annotations on documents.

3.1 Possible Means of Capturing Annotations

Documents are nowadays provided as in two different forms: electronic and printed. We have already had several ways of capturing annotations that is often called "pen-based computing" (Subrahmonia & Zimmerman, 2000).

For electronic documents, tablets are widely used for capturing annotations. When they are utilized as a mean of annotating documents, a major problem is that a user is forced to write not on a document but on a tablet. In order to solve this problem, tablets with LCD displays as well as tablet PCs have been proposed. Along with the capability of InkML (W3C, 2004), these devices provide us a new way of interacting with documents (Shilman & Wei, 2004). However, unnaturalness of writing on a display seems to prevent them from disseminating.

We also have several ways for capturing annotations on paper documents. Ultrasonic pens give us a mean of keeping annotations in electronic formats while writing on normal paper. A disadvantage of this device is the discrepancy between electronic and written annotations caused by the slip of sensors. Another and more important device is the "Anoto" pen (Anoto) which captures its position by sensing fine dots on the special Anoto paper. Since it is guaranteed that patterns of fine dots are unique in all sheets of paper, no discrepancy is caused by the Anoto pen.

Although the above methods have an advantage that the annotations can be separated from the original documents from the beginning, they require special devices to realize this advantage. The use of such special devices may limit users and/or scenes in which annotations are made. In order to support annotations in more natural ways, it is required to use normal writing materials such as ordinary pens and sheets of paper. It is also desirable to use ordinary devices when capturing written annotations.

A possible way of meeting these requirements is to use digital cameras for capturing written annotations. A major disadvantage of this approach is that it is generally hard to separate annotations from the original documents.

3.2 Camera-Based Analysis of Annotations

In this subsection, we propose a new method of analyzing annotations written in normal writing materials with the help of analysis techniques of camera-captured documents (Doermann, Liang & Li, 2003), (Yamada, 2004).

3.2.1 Overview

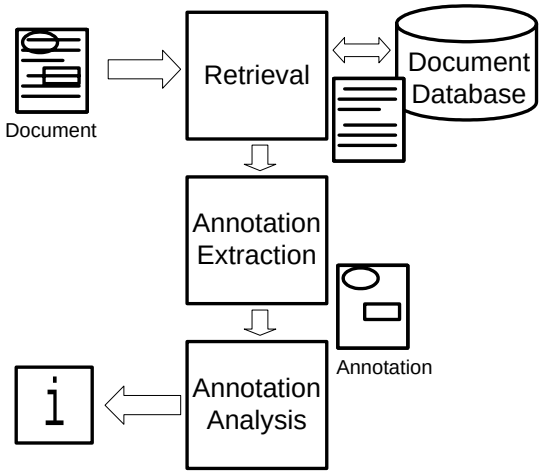


Figure 2: Overview of the system.

Figure 2 shows the overview of the proposed system. As stated above, a major disadvantage of camera-captured annotations is the difficulty of separation of annotations from imaged documents. In order to solve this problem, we employ a database of electronic documents. In recent years most of the printed documents are produced by printing their electronic versions. Thus it is not so unrealistic to assume that electronic equivalents are accessible when analyzing paper documents with annotations.

The processing of extracting annotations consists of three steps: retrieval, annotation extraction and annotation analysis. At the first step, a camera-captured document with annotations is employed to retrieve its electronic version without annotations (Nakai, Kise, Iwamura & Matsumoto, 2005). At the next step, annotations are extracted by subtracting the document image produced based on the electronic version from the camera-captured document image. At the final step, the structure of annotations is analyzed and each piece of annotations is attached to the corresponding object(s) in the electronic document. For example, an underline is attached to the textline under which the underline exists, and handwritten lines and arrows are utilized to attach annotations to objects.

In this paper, we describe the second step of annotation extraction in more details. The processing of this step consists of three smaller steps: preprocessing, dewarping, and extraction. Preprocessing is to obtain binary images from camera-captured color images for the succeeding steps of processing. Dewarping is to normalize perspectively skewed documents. Extraction is the process of subtracting the image of the electronic document from the normalized camera-captured image after adjusting the size and the location of images.

3.2.2 Preprocessing

The preprocessing step includes binarization and noise reduction. Binarization is first applied to the camera-captured color images. In our method, we employ the adaptive binarization algorithm provided by the OpenCV library (OpenCV). Next, noise reduction is applied to the binary image so as to remove too small and too large objects in the image. Most of the too small objects are caused by the salt-and-pepper noise. Too large objects often correspond

to annotations. Note that the removal of too large objects is just for the sake of the next step "dewarping"; after the dewarping, the large objects are recovered to extract annotations.

3.2.3 Dewarping

Documents that are not frontal-parallel to the camera's image plane undergo a perspective distortion. As shown in the lower left part of Figure 3, textlines laid out in parallel in the original document are not parallel in the image under the influence of the perspective distortion.

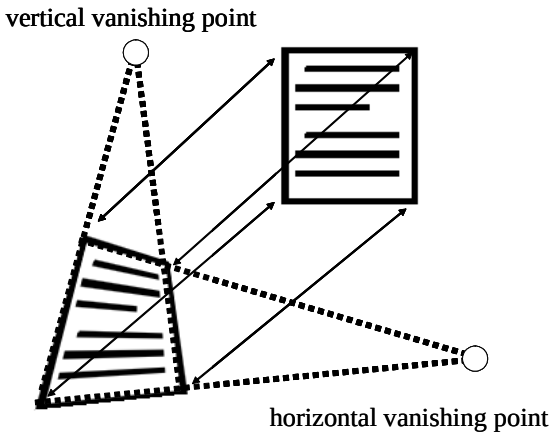
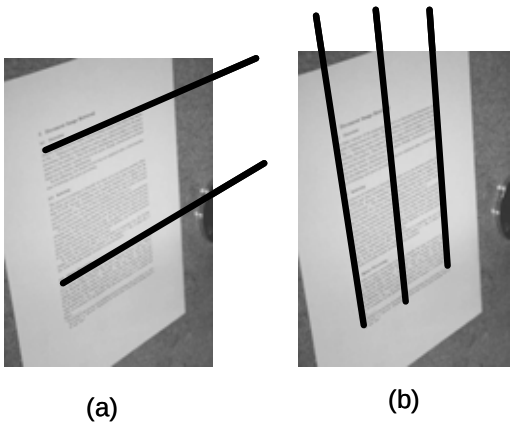


Figure 3: Perspective distortion and vanishing points.

The perspective distortion can be normalized by finding four pairs of corresponding points as shown in Figure 3. The horizontal and vertical vanishing points give us a clue to find such pairs. We employ the method proposed by Clark and Mirmehdi (Clark & Mirmehdi, 2002) for estimating the vanishing points. Their method is based on the projection profile of the image: for every hypothesized horizontal vanishing point, a projection profile from the point is calculated for measuring the confidence as the squared-sum of the projection profile. Figure 4 (a) illustrates an example of the estimated horizontal lines (the thick lines in Figure 4(a)) which define the horizontal vanishing point. The vertical vanishing point is likewise estimated. In order to deal not only with fully justified paragraphs but also with left and right justified as well as centered paragraphs, three lines shown in Figure 4 (b) are estimated from leftmost, center and rightmost points of each line. The pair of lines with the highest confidence is selected to find vertical vanishing point.

After finding the vanishing points, the image is normalized by applying the perspective transformation defined by the four points. We also utilize the module provided by the library OpenCV.

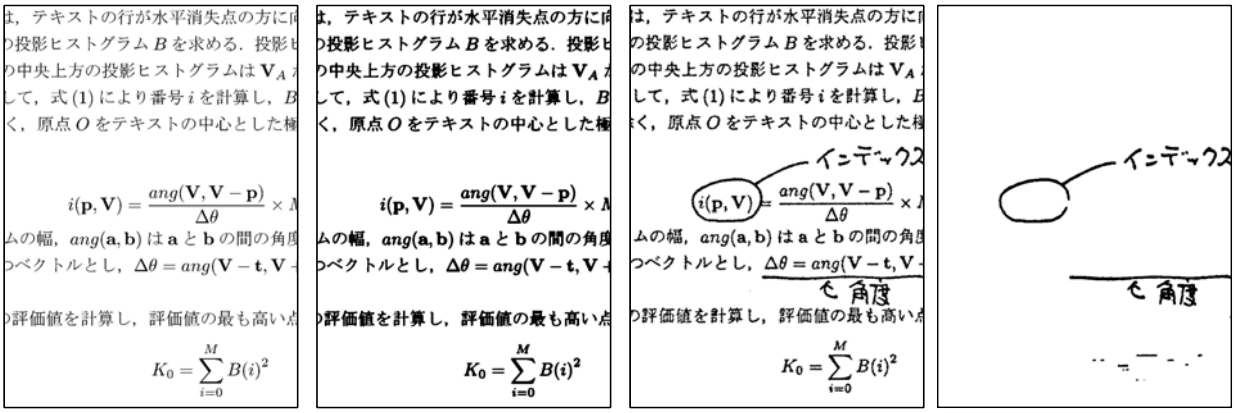


(a) (b)

Figure 4: Dewarping by estimating horizontal (a) and vertical (b) vanishing points.

3.2.4 Extraction

In order to extract annotations, the following three steps of processing are applied: global matching, local matching and subtraction. Global matching is to determine a rough location and size of the camera-captured image against the image from the electronic document. Projection profiles are again utilized for this purpose. Next the globally matched images are divided into $n \times n$ sub-regions for fine adjustment of the location. We currently use $n = 5$. In the current implementation, the location of each sub-region is independently adjusted; it is better to impose some restrictions to penalize inconsistent moves among the neighboring sub-regions.



(a) (b) (c) (d)

Figure 5: Subtraction of images. The original image (a) is blurred to obtain the robustness of the errors of matching. The blurred image (b) is subtracted from a camera-captured image (c) to extract annotations (d).

After determining the size and location of images, subtraction of images is applied. An example is shown in Figure 5. The original image generated from the electronic document is first blurred for improving the robustness to the errors of matching. The blurred image is then subtracted from the camera-captured image to extract annotations as the residual.

4 Experimental Results

In order to evaluate the proposed method, experiments were made on camera-captured documents with annotations.

4.1 Conditions and Evaluation

As the input device, we used a digital camera “Canon EOS Kiss” (6.3 million pixels) with the lens “EF-S 18-55mm USM”. The number of images was 36 all of which were with annotations. The results were evaluated using the criteria called recall (R), precision (P) and F-measure (F) defined as follows:

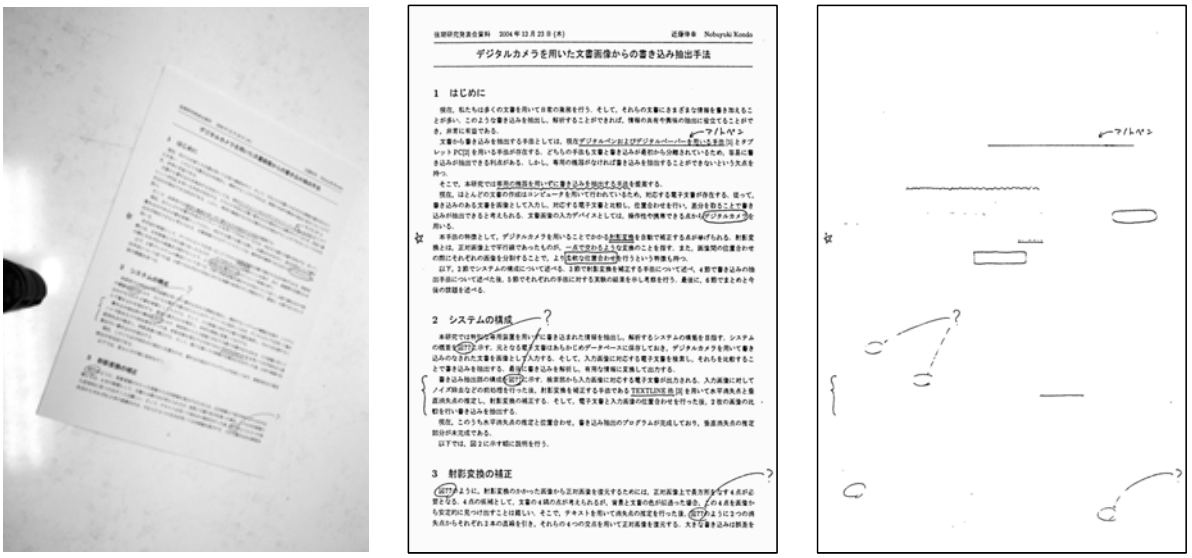
$$R = A/B, \quad P = A/C, \quad F = 2PR/(R+P),$$

where A, B and C indicate the number of pixels of correctly extracted annotations, the number of pixels of correct annotations, and the number of pixels of extracted annotations, respectively.

4.2 Results and Discussions

As the overall average, R=49.2%, P=34.0% and F=40.2% were obtained. Some examples of processing results are shown in Figures 6-8. Figure 6 illustrates a successful case of Japanese documents. Most of the annotations were successfully extracted without producing noise. An example of English documents is shown in Figure 7. Although some noise was on the upper part of the extracted annotations as shown by the dotted ovals in Figure 7(c), major parts of the annotations were successfully extracted. Figure 8 shows an erroneous case. For this image, as shown by the thick line in Figure 8, the estimation of the vertical vanishing point was failed due to the large annotation indicated by the dotted oval in Figure 9(c). This caused the error of matching of the images and resulted in extracting a large number of pixels as annotations. Most of the errors we encountered during the experiments were caused by the same reason.

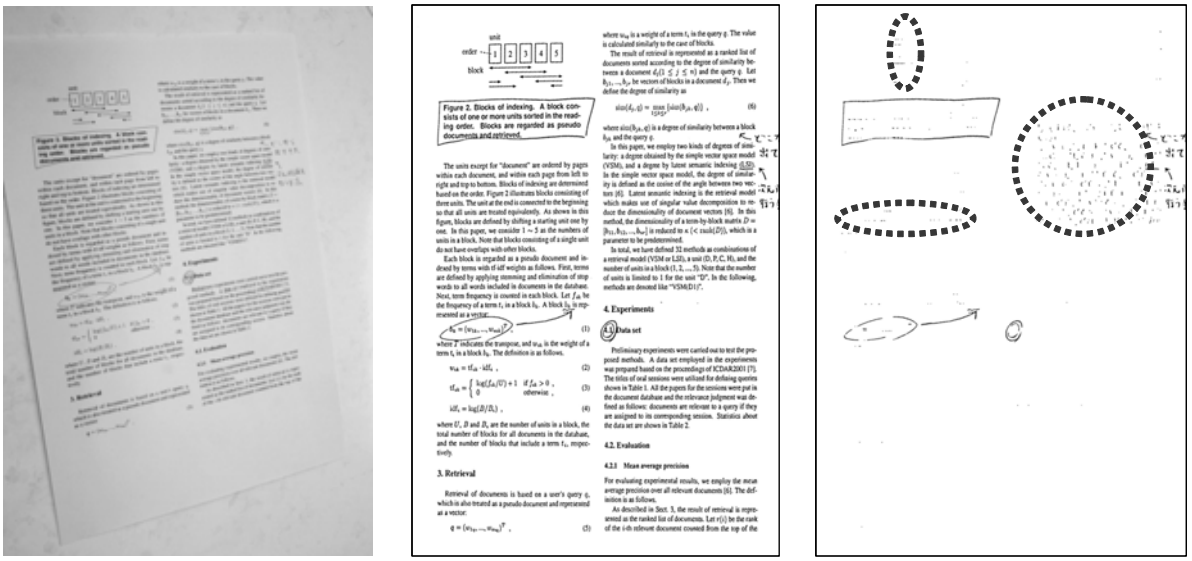
The above experimental results show that the proposed method is fundamentally capable of extracting annotations from camera-captured images, though several steps including estimation of vanishing points need to be improved.



(a) (b) (c)

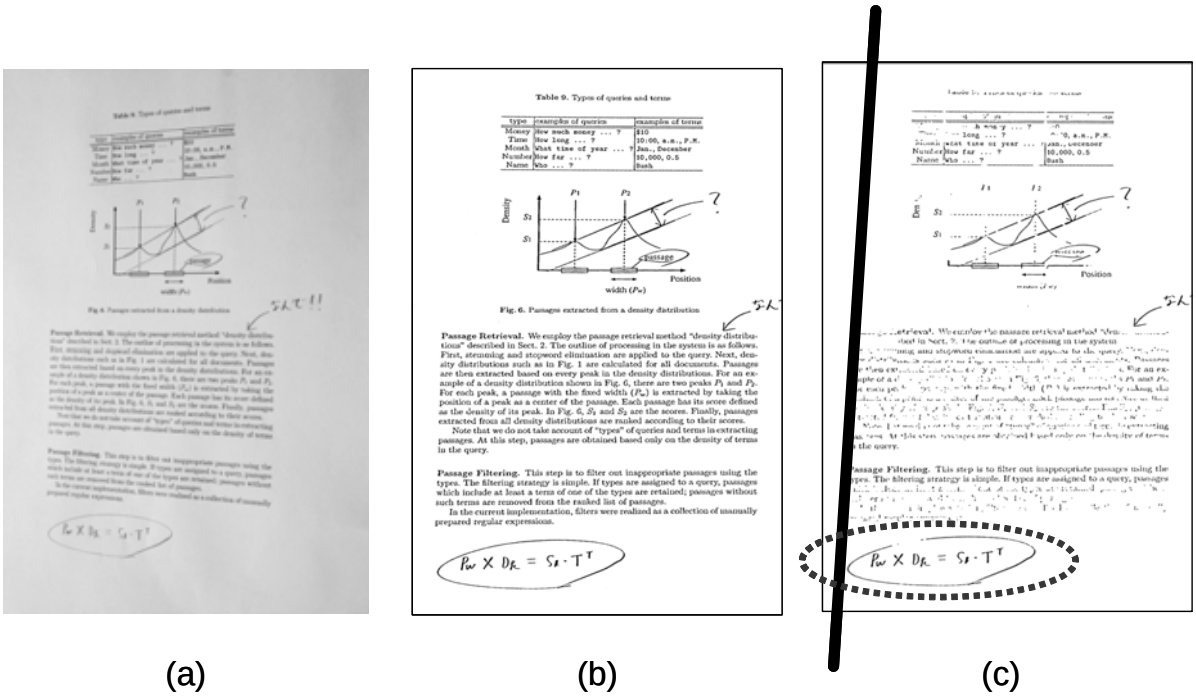
Figure 6: Successful case (Japanese document). R = 58.2%, P = 83.4%, F = 68.5%.

(a) input image, (b) normalized image, and (c) extracted annotations.



(a) (b) (c)

Figure 7: Successful case (English document). R = 69.1%, P = 69.6%, F = 69.3%.



(a) (b) (c)

Figure 8: Failure case (English document). R = 10.4%, P = 14.6%, F = 12.2%.

5 Conclusion

In this paper, we have discussed the need of "recycle of information" as a mean of supporting collaborative business activities. As compared to "reduce", "refuse" and "reuse" of information, "recycle" is characterized by the point that it decomposes information into resources and then reproduces new information using the decomposed resources. We have also pointed out that in order to realize the recycle of information, it is necessary to extract profiles of users and information contents. In the business activities, such profiles can be extracted from the annotations of documents, since documents are important and widely used information media in the activities. It has also been said that the extracted annotations can be utilized for finding clues of collaborations in the activities.

As a mean of extracting annotations from documents, we have proposed the system which takes as input camera-captured documents with annotations and extract annotations with the help of documents without annotations. The process of extraction of annotations was implemented as the first step towards the automated extraction and manipulation of annotations. From the experimental results we confirm that the proposed method is capable of extracting annotations if the dewarping is successful.

Future work includes the implementation and test of the remaining parts of the system. Utilization of extracted annotations for recycling information is another important work to be explored in the future.

References

Dengel, A., Junker, M., & Weisbecker, A. Eds. (2004): Reading and Learning: Adaptive Content Recognition, Springer.

Mani, I., & Maybury, M.T. Eds. (1999): Advances in Automatic Text Summarization, The MIT Press.

Pazienza, M.T., Ed. (1999): Information Extraction: Towards Scalable, Adaptable Systems, Springer.

Chakrabarti, S. (2003): Mining the Web: Discovering Knowledge from Hypertext Data, Morgan Kaufmann Publishers.

Ohno, S., Maeda, E., Kise, K., & Matsumoto, K. (1999): Estimation of User's Interests and WWW retrieval Based on Topic Extraction from HTML Files, Trans. IEE of Japan, 199-C (11), 1316-1322 [In Japanese].

Cingil, I., Dogac, A., & Azgin, A. (2000): A Broader Approach to Personalization, *Communications of the ACM*, 43 (8), 136-141.

Light, M., & Maybury, M.T. (2002): Personalized Multimedia Information Access, Communications of the ACM, 45 (5), 54-59.

Nanno, T., Saito, S., & Okumura, M. (2003): Structuring Web Pages Based on Repetition of Elements, In *Proceedings of the Second International Workshop on Web Document Analysis*, 7-10.

Subrahmonia, J., & Zimmerman, T. (2000): Pen Computing: Challenges and Applications, In *Proceedings of International Conference on Pattern Recognition*, (2), 2060-2066.

W3C (2004): InkML – The Ink Markup Language, from http://www.w3.org/2002/mmi/ink.html

Shilman, M., & Wei, Z. (2004): Recognizing Freeform Digital Ink Annotations, In *Proceedings of 6th International Workshop on Document Analysis Systems*, 322-331.

Anoto: from http://www.anoto.com/

Doermann, D., Liang, J., & Li, H.(2003): Progress in camera-based document image analysis, In *Proceedings of International Conference on Document Analysis and Recognition* '03, 606-616.

Yamada, K. (2004): Consideration on Character and Document Media Recognition and Understanding for Ubiquitous Information Interface, In *Technical Report of IEICE*, PRMU2003-229, 87-94 [In Japanese].

Nakai, T., Kise, K., Iwamura, M., & Matsumoto, K. (2005): Document Image Retrieval Based on Cross-Ratio and Hashing, In *Technical Report of IEICE*, PRMU (to appear) [In Japanese].

OpenCV: Open Source Computer Vision Library, from http://www.intel.com/research/mrl/research/opencv/

Clark, P., & Mirmehdi, M. (2002): Recognising text in real scenes, *International Journal of Document Analysis and Recognition*, 4, 243-257.