文書画像検索

Document Image Retrieval

徐　盈輝[*]　伊東　秀夫[*]　大黒　慶久[**]
Yinghui XU　Hideo ITOH　Yoshihisa OHGURO

要　旨

オフィス文書の電子化が進む中，画像データの検索ニーズが高まってきている．本論文では，内容ベースの画像検索のための新しい手法を提案する．我々は画像データの特徴を記号化し，かつ１次元化することで，テキスト検索と同様の高速な画像検索を可能にした．この手法では，対象とする画像のピクセルを，二値化した後，縦方向と横方向に射影を取り，１次元の特徴量を生成する．この特徴量は，スムージングを施した後，ランドマークと呼ぶ特徴点の列として正規化される．そして，各ランドマークに記号を割り当てることで，元の画像全体を１つの記号列に変換する．最後に，従来のテキスト検索技術を，これら記号列に適用することで，画像検索の機能を実現する．いくつかの性能評価の実験を通し，その効果を検証した．

ABSTRACT

Digital documents are a mixture in modern office working environments. With the digitalizing era coming, the drop in the cost of digital data storage, easy-reading device, PDA. mobile device, etc. are become popular nowadays. It is critical to be able to manage and retrieve office images to satisfy user information needs. In this paper, we propose a novel framework for content-based image retrieval using symbolic time series. Given an image, the foreground pixels (usually the black pixels) projections along the horizontal direction and vertical direction are generated to be the primitive representation. projection profile of an image is inherently a sequence of data which is the same as the one dimensional time series data. In our framework, data cleaning and data reduction are introduced in the data normalization module to enhance the data quality, and then turning points, called landmarks, along the normalized data sequence are extracted. Next, symbolic representation for the extracted landmarks are produced. Finally, the converted symbol sequence enable us to use the conventional text retrieval indexing strategy and thus Ngram model is adopted to gauge the similarity of the symbolic time series. Experiment results confirmed the effectiveness of the proposed approach.

* グループ技術開発本部　オフィスソリューション技術開発センター
Office Solution Technology Development Center, Corporate Technology Development Group
** コントローラ開発本部　GW開発センター
GW Development Center, Controller Development Division

1. Introduction

With the coming of the digitalization era, the huge amount of document images has left a tremendous need for robust ways to access the information these images contain. Printed documents are often scanned for archiving or in an attempt to move toward to a paperless office and stored as images but without adequate index information. To enable us to make good use of digitalized resources, especially document images, we have developed a content-based image retrieval system which is a marriage of our efficient full text index engine and the novel image feature symbolization technique.

Several approaches for document image retrieval are summarized in a survey [1]. The existent approaches can be divided into two categories, optical character recognition (OCR) based approach and image feature based algorithms. While the state-of-art OCR technology can produce very accuracy results, the sensitiveness to image resolution, heavy computation cost and language dependency limit its applications. Different from the OCR based algorithms, image feature based algorithm focus on the image contents instead of the text information. Several approaches are summarized in a survey. Spitz maps alphabetic characters to a small set of character shape codes (CSC) which can be used to compile search keys for ASCII text retrieval [2]. CSC's can also be obtained from text images based on the relative positions of connected components to baselines and x-height lines, as used by Spitz for word spotting in document image. Doermann, etc. al. [1] extend the application of CSC's to document duplicate detection by constructing multiple indexes using short sequences of CSC's extracted from the first line of text of sufficient length. Hull and Cullen [3] have proposed a method to detect equivalent document images by matching the pass codes of document. Yu and Tan [4] presented a retrieval method for Chinese document images based on stroke density code of each character object. These methods have shown their advantages on fast and light-weight solutions for normal document images. However, these methods are inherently relying to some extent on the text line, word or even character segmentation results. Low document image quality or images with complex layout can inhibit accurate recognition and segmentation. What is more, image features are usually represented as numerical feature vector and the similarity between images are measured via vector distance metric. The high-dimensional indexing technique is usually chosen for fast accessing the similar images from large image database. However, there are problems of vector space solution on the ambiguities of distance measure and scalability issues.

To overcome the challenges of both the image feature extraction problems caused by image resolutions, complex layout, different languages, distortion and the indexing and distance measure problems caused by the vector space model, we propose a novel framework which is a marriage of our efficient full text index engine and the novel image feature symbolization technique.

The remainder of this paper is organized as follows. In Section 2, we will present our approach. In Section 3, the experimental results are presented. Finally, conclusion is drawn in Section 4.

2. Our Approach

In this report, we propose a novel approach which is directly addressing the limitations of the document image retrieval problems. Fig. 1 illustrates the system diagram. The distinguished module is the "Image2Symbols" which play an important role in our system. There are four major steps in this module:

- Making the image projection profile
- Data normalization: envelope filter the projection profile
- Feature extraction - landmark identification
- Feature presentation - symbolization process

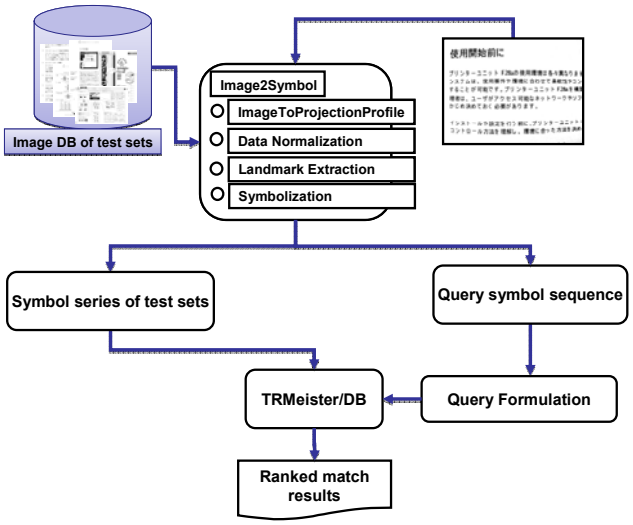

Fig.1 System diagram.

Given an image, we choose to use the primitive pixel histograms (projection profile) based on the binary image because the flipping foreground and background pixels with equal probability has only small effect on the distribution of white and black pixels in a row or column. There are mainly two types of document images, one is of the multiple text lines, like the publications, technical reports, etc.., the other is of the pictorial style, such as diagrams, drawings, etc.. In the case of the paper-style document images, to enable us to search the target page images based on small patch of the target copies, we choose to use the quick and dirty line segmentation before making the projection profiles. After segmenting the target images into multiple line-style images, the vertical projection profile along the line image is generated for the subsequent feature extraction. While, for those pictorial style images, to reflect the 2D structure of the entire image, both the horizontal and vertical projection profile are taken into consideration. The goal is to relate the results of this analysis to the structure of the images which is reflected by the image foreground pixel histogram; the projection profile is a sequence of non-negative integer value.

2-1 Data Normalization

Data normalization process is to overcome the noise and lengthy effects in the raw projection profile. The foreground pixel projections can be viewed as a high frequency signal superimposed on a low frequency one. It is this underlying low frequency coarse scale signal that is highly characteristic of the image structure. To minimize the effect of noise, a low pass filter is adopted in our approach.

2-2 Landmark Extraction

In psychology and cognitive science, more and more evidence have shown that human and animals depend on landmarks to arrange spatial memory [5]. Landmarks representation for time series is proposed by Chang-Shing Perng to capture the most significant evidence of the time series data. In our approach, similar to the landmark idea, instead of using the entire low pass filtered data series, feature points (called landmarks) along it are extracted and used to be the representative information. In the data structure of each landmark containing two elements, one is the attribute (hill and valley) and the other is the magnitude value; for the landmark point with hill attribute, it indicates that this point goes though the change from up to down. Conversely, the landmark with valley attribute indicate that it goes though the change from down to up. The detection algorithm performs one pass through the entire data sequence. The attribute of each landmark will be used in the symbolization process to assign different symbolic sets for the landmark points with different attributes.

2-3 Symbolic Representation

In this section, we will introduce how to symbolize the landmarks which are extracted in the last process. Every landmark will be assigned a symbol in the symbolic process. The symbolization scheme called "*ContextSYM*" is illustrated in Fig 2. Each landmark will be assigned a

symbol following the rule illustrated in Fig 2. In Fig 2, vertex "*CV*" with attribute valley and vertex "*CH*" with attribute hill are symbolized by choosing a symbol from different symbolic sets. For example, the ASCII code of the symbolic sets for valley vertices range from 48 to 55, while the ASCII code of the symbolic sets for hill range from 56 to 63. For a particular code, the three neighboring vertices with the same attribute will be used to determine its symbol. In the case of "*CV*" in the left-to-right symbolic process, three right neighboring vertices are "*RV_1*", "*RV_2*" and "*RV_3*" respectively. Comparison results between these three vertices and the vertex "*CV*" are categorized into two types, less-equal and greater, which are represented as "0" and "1" respectively. Vertex "*CV*" illustrated in the figure got the code "*111*" and thus '7' is assigned on "*CV*". Similarly, vertex "*CH*" is symbolized to be ";". In the conventional symbolization approaches proposed in [6, 7, 8], usually, coarse symbol definition involves partitioning the range of the original observations into a finite number of regions [9]. Each region is associated with a specific symbolic value and each original measurement is thus uniquely mapped to a particular symbol depending on the region in which the measurement falls. The SAX model proposed by Keogh is one well-cited symbolic method. The symbolization process in SAX is basically following the way described above. Symbols are allocated based on mean value of equal sized frames along the entire data series. However, the SAX model will not be suitable in the case of analyzing the projection profile of document image because it may cause a possibility to miss some important patterns. Moreover, the coarse symbol definition in SAX involves partitioning only the amplitude range (projection value) of the original observations into a finite number of regions without considering the phase range (the position of the foreground pixel histogram from left to right). However, the local variation in a particular phase may possibly have the distinguished pattern for matching operation. What is more, the symbolization results may be even worse in SAX model when the query series are not synchronized with the target series in the collection; in our approach, the landmark feature point derived from projection profile is encoded via magnitude relationship of the local context within a phase range. Such kind of design enables us to overcome the limitations of the conventional symbolic algorithms.

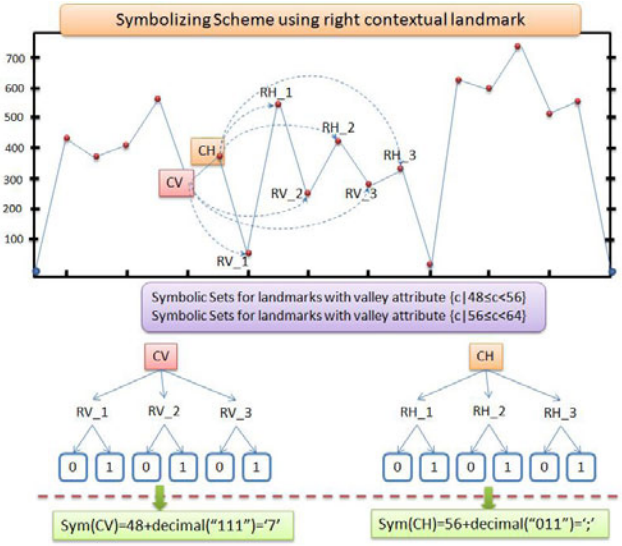

Fig.2 Symbolization illustration.

2-4 Searching

After symbolization, a document image is converted into a symbol series. Since the image features are finally represented as a symbol sequence, it enables us to choose the conventional full text search technology which has merit on the high scalability and effective probabilistic ranking function; An N-gram model based full text search is used to find the relevant images for a query from the database. The ranking function used in our system is OKAPI BM25 [10].

$$score(q;d)=\frac{tf(t_i)\times\log\left(\beta\times\frac{N}{df(t_i)}+1\right)}{tf(t_i)+\alpha\times(1-\gamma)+\gamma\times\frac{docLen}{aveDocLen}}$$

Where the parameter sets, α, β and γ are set to be 1.0, 0.2 and 0.9 respectively.

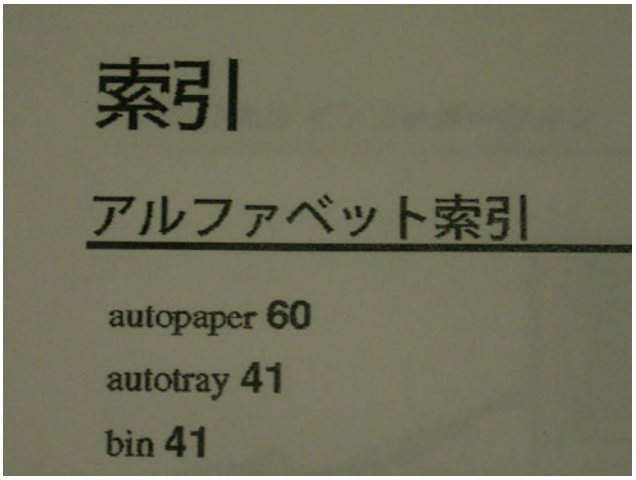

Fig.3 Query example in the language independence test.

3. Experiment

To verify the effectiveness of our approach on dealing with the document image retrieval task: we have done experiments:

- Language Independence Test
- Pictorial style image search with test samples having different image characteristic, such as resolutions, color and binarization status

For the evaluation metrics, we use the most commonly recall precision (RP) measure. R-precision de-emphasizes the exact ranking among the retrieved documents and more useful when there are a large number of relevant documents. A relevant judgment file will be created based on associating the query name with its corresponding resource document image. RP at top 1, 5, 10, 30, denoted as R@1, R@5, R@10, R@30, are calculated respectively. That is, system is evaluated by the percentage of queries which find their resource at top N return sets with total number of queries.

3-1 Language Independence Test

To investigate the effectiveness of our approach on searching a query with different language from the images in the database, a test corpus which contains document images with various languages (14 languages) is prepared and summarized in Table 1. Fig. 3 is an example of a query image which is captured by Ricoh Caplio camera. There are around 2653 page images with the multiple-text line style. 90747 line images are generated from these page images after the line segmentation process. All of the line images of the test pages are converted into symbols sequence by our approach and then inserted into the TRMeister's DB [11]. Queries are prepared by capturing a patch from the entire page image by Ricoh Caplio camera. 33 queries with English languages are made for the experiment. The same as the process done for the images in the database, the 33 query images are also separated into 443 line images, each query length is of 13 lines in average. To get the relevant target for a query, each line image of a query is converted into symbols and pools the top 100 relevant page ID with score in the search. Finally, the top 100 relevant targets are produced from the pool. The performance is shown in Table 2;

Table 1 Multi-lingual test set.

	#page	#lines	Language
Newsweek	61	4946	En
Multi-lingual	351	24856	Dutch, French.. etc..
MFP spec.	2241	60945	Jp

Table 2 Search performance of language independence test.

	R@1	R@2	R@5	R@10	R@30
RP at top K	93%	93%	93%	93%	96%

3-2 Pictorial- Style Image Search

A test corpus which is made of the power-point slides is prepared for the evaluation. There are 1993 colorful

images and 1993 binary version with the same image content generated through Open-Office [12] software.

The query images are prepared by choosing samples from test corpus randomly and then making the query images by print-and-scan. To investigate the robustness and insensitivity of our approach to the image colors and various image transformations, two test runs are performed. Color-color test: in this case, the search target is the 1993 color images while the queries are randomly chosen from them and then make copies by print-and-scan2color with 200 dpi. There are 30 queries prepared for tuning and 37 queries prepared for the test. In the same manner, for the binary-binary test run: the search target is the 1993 binary version images while the queries are randomly chosen from the binary collection and then make copies by print-and-scan2binary with 200 dpi. Similarly, there are 30 queries prepared for tuning and 37 queries prepared for the test. The important parameters involved in the system are the window size of the envelope filter and N-gram search unit. First of all, 30 queries are used to tune the system parameters by which the system obtained the best accuracy of R@1. And then based on the tuning parameters, experiments on the 35 test queries were performed. In the case of the color-color test run, the results for the 30 tuning queries are shown in Table 3 and the results for the 35 test queries are shown in Table 4. The parameters, filter_size, which is the window size of the low pass filter for the data normalization process, is set to be 7 for test images in the DB and 18 for the query images. N-gram unit is set to 4-gram. In the case of the binary-binary test run, the result for the 30 tuning queries are shown in Table 5 and the results for the 35 test queries are shown in Table 6.

Table 3　Color-color test run for the 30 tuning Queries.

	R@1	R@2	R@5	R@10	R@30
RP at top K	63%	86%	96%	96%	100%

Table 4　Color-color test run for the 35 test queries.

	R@1	R@2	R@5	R@10	R@30
RP at top K	72%	97%	97%	97%	100%

Table 5　Binary-binary test run: for the 30 tuning queries.

	R@1	R@2	R@5	R@10	R@30
RP at top K	43%	60%	73%	73%	83%

Table 6　Binary-binary test run: for the 35 test queries.

	R@1	R@2	R@5	R@10	R@30
RP at top K	62%	78%	89%	89%	91%

The experimental results confirmed that the high search performance is reached based on our proposed approach.

4. Conclusion

In this report, we propose a novel approach for content-based document image retrieval using symbolic image features. Novel models for symbolization are proposed as well. Through the experiments, we found the most effective way to reach the surprisingly good accuracy. Experiment results show that our method is very promising to be used in the related area.

Reference

1) D. Doermann, The indexing and retrieval of document images: A survey. Computer Vision and Image Understanding: CVIU, 70 (3), pp. 287-298, 1998.

2) A. L. Spitz, Duplicate document detection. in the Proceedings of the SPIE - International Society for Optical Engineering, Document Recognition IV, San Jose, pp. 88-94, 1997

3) J. J. Hull and J. F. Cullen, Imaged document text retrieval without OCR. in Proc. of the 4th Int'l Conf.

on Document Analysis and Recognition, pp. 308-312. ICDAR, 1997

4) C. L. Tan, W. Huang, Z. H. Yu and Y. Xu, Imaged document text retrieval without OCR. IEEE Trans. Pattern Analysis and Machine Intelligence, 24 (6), pp. 838-844, 2002
5) C.-S. Perng, D.S. Parker, K. Leung, Representing time series by landmarks, in Proceedings of the CIKM, 1999.
6) B. G. Hbrail, Symbolic representation of long time-series, in Symbolic representation of long time-series. Proc. of Applied Stochastic Models and Data Analysis Conference (ASMDA'2001), 2001
7) S. Lawrence, C. L. Giles, and A. C. Tsoi, Noisy time series prediction using symbolic representation and recurrent neural network grammatical inference, Technical Report UMIACS-TR-96-27 and CS-TR-3625, 1996.
8) J. Lin, E. Keogh, S. Lonardi, B. Chiu, A symbolic representation of time series, with implications for streaming algorithms, in proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, 2003.
9) C. Daw and C. Finney, A review of symbolic analysis of experimental data, Review of Scientific Instruments, Volume 74, Issue 2, pp. 915-930 (2003).
10) S. E. Robertson, Steve Walker, and M. H. Beaulieu, Okapi at TREC-7, in Proceedings of the Seventh Text Retrieval Conference. Gaithersburg, USA, November 1998.
11) TRMeister, http://www.ricoh.com/about/technology/showcase/trmeister.html
12) Open-Office Software, http://www.openoffice.org/