Automatic Seal Information Reader

Farshad N, Bhavna Antony, Peeta Basa Pati and A G Ramakrishnan
Department of Electrical Engineering
Indian Institute of Science
Bangalore, INDIA – 560 012.

Abstract

Seals contain a lot of vital information about the document. Its detection is one of the important steps to gain access to these information. In the present paper, we present a correlation based technique which exploits the existence of some constant character strings and their topology for detection of seals. We also present a technique which separates the false positives from the actual seals. We achieved an accuracy of about 98% for extraction of such kind of seals. Once the seal is extracted from a page image, an OCR has been designed and employed to read the contextual information present in the seal. A knowledge based post-processing step is employed to enhance the accuracy of the recognized text strings.

Key Words: Connected Component Analysis (CCA), document image, seal detection, triangulation.

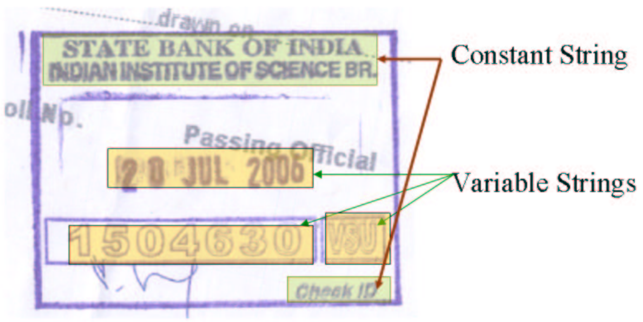


Fig. 1: A sample seal for processing of bank documents. Here, the name of the bank is specified in the first line, the second line contains information about the location of the bank and the date of processing is a variable string field.

1. Introduction

Seals represent a person or an organization. Their presence proves the authenticity besides providing some vital information about the document. Thousands of documents are processed in various offices, everyday. Each office has a seal which is unique to the office and its use. Some seals are used for presenting vital information about the document. Since there are thousands of such documents being processed out of a given office, the the pile of documents over a period of time is quite huge. Automatic processing of such documents necessitates the automatic processing of such seals. The information obtained from such seals could be used for efficient storage and retrieval of the documents.

Seals, generally, contain some constant strings which provide information about the owner-organization and its locality. Besides, it could contain some variable fields. Typical example includes seals that are used for processing of bank cheque or postal mails in India. These seals contain the name and the geographical location of the bank/post-office as a constant string and the date as a variable string field (see figure 1). Such kind of seals could also be seen in various governmental offices, being employed various purposes.

Generally, a document contains one such seal, and hence considered as single page document. However, there are cases of multi-page documents where the seal could be present in any one or more than one number of pages of the document. Besides, they could appear in any orientation and any location of the document page. In such a situation, it is important to select the page containing the seal. Once the page is selected, we need to localize the seal with appropriate orientation estimation so that we could extract the information contained in the seal.

Document analysis has been actively pursued by various researchers. A lot of work has been reported on labeling text and non-text areas (like images, flowcharts and handwritten areas) in a document. However, a little work is done for detection and recognition of seals. Frisch [1] presents a technique where a fuzzy integral based technique is used to selectively extract color clusters in images and uses this technique for isolation of seals. Ueda *et. al.* [2] have proposed a technique for isolation and verification of seals and signatures in color Japanese bank-cheques. Both the above techniques use color images for input. The scheme proposed by Ueda *et. al.* has 3 steps to achieve the desired result. In the first step they convert the RGB color input image to HSV color space and in the following step the cluster the HSV space for isolation of identical color components in the input image. Here, they extract the seals imprints and signatures. Subsequently, they eliminate the noise-elements and base lines. Such a removal helps in verification of the genuineness of the extracted seals.

Roy *et. al.* [3], [4]present a technique for automation of Indian Postal System. After a Run Length Smoothing (RLS) of the input image, they divide the image to smaller blocks. Based on the ON-pixel density and the number of CC's present in a block, they decide to separate the seals and logos from the written part of the mail-document image. Here, some prior information is employed about the position of the seal. Chang *et. al.* [5] present a point matching based seal identification technique. They assume that the seals have been localized and a recognition process is proposed.

Of the works reported on the detection and reading of seals in document images, some researchers have tried to be generic in detection of any and every kind of seals while others have tried to exploit the positional, structural, contextual information of the seals to localize the seals. Though some works report using the seal for identification of the owner organization, most of these work are carried out keeping in mind that seal is a natural hurdle for text recognition and hence should be removed from the document image. To our knowledge, no one has tried to localize seals for extracting information from the variable fields of the seals. In the present paper, we propose to localize a seal and read its contents to extract meaningful information.

Our database contains 305 each of seal-positive and seal-negative images Each seals contain 3 constant strings and at least a variable field. Our aim is to detect and localize the seal based on its constant string fields and extract the information contained in the variable field. Our collected images are scanned at 200 dpi resolution and are stored in one-bit depth black & white text format.

The following sections of the paper has been organized as mentioned below. We present a description of the system in section 2 while some results are presented in section 3. Sec. 4 concludes the paper with some discussions and pointers to wards the future explorations of such a work.

2. System Description

In this paper, we present a scheme which is designed for extraction of information present in the variable string fields of the seal. We assume that a small amount of skew is present after the seal is corrected for its orientation[1]. Figure 2 demonstrates the system (system for the information extraction from seals present in document images) in a block schematic fashion.

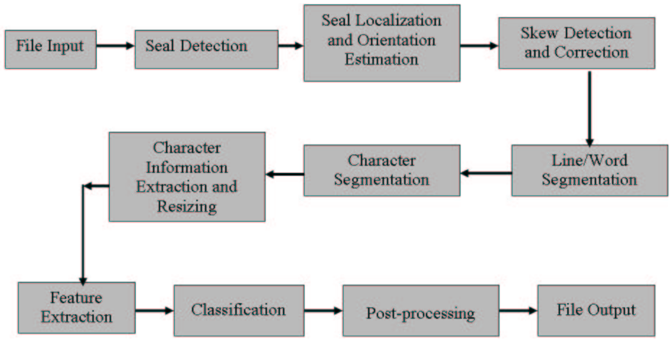


Fig. 2: Block schematic representation of the scheme to extract information from a seal present in a document image.

A. Detection System

Following observations were made for detection and localization of the fields.

- A seal occurs only once in the documents – single page or multi-page, however, there could be more than one seal in a given page image,
- the seal could occur in any one of the pages in multi-paged documents,
- the seal could occur in any position of the page, *i.e., a priori* locational information is unavailable,
- the seal could occur in any orientation – it mostly appears aligned to one of the four orientations of 0, 90, 180 and 270 degrees,
- there could be some skew present in the seal with respect to the free-flowing text of the document page,
- the seal has a specific font size and style,
- the seal has a some constant character strings and some variable character/numeral strings,
- the variable strings are the ones which contain the vital information, and
- the seal could be isolated, touching the text blocks or overlap with the general text in the document image.

The proposed technique targets to (i) detect the presence of such a seal, (ii) localize it's position in the detected page, and (iii) estimate its orientation from the four orientations mentioned above.

1) Block Division: A run length smoothing algorithm (RLSA) [6] is applied to bridge the intra-character and inter-character gaps. The threshold for this is so chosen that all the gaps existing between words in a line of text and between lines of text in a paragraph are closed. Thus, we divide the input image into smaller blocks, typically paragraphs and call them as the building blocks. Our seal has a constant size and dimensions. We employ the knowledge of these dimensions to see if a building block satisfies this dimension criteria, consisting of (i) total number of ON-pixels present, (ii) height of the block, and (iii) the width of the block, to be considered for the next level of processing. Any block which passes this test, is considered to be a possible candidate for seal.

2) Template Cross-correlation: We have selected three constant character-string fields(CCF's) to form templates. Because the seal contains more than 3 CCF's, we have selected the three based on the (i) number of characters present in a CCF, (ii)the distance between the CCF's (the relative positional topology of the selected CCF's should be constant), and (iii) the size of the characters present in the CCF. Each of the building block, passed in the dimensionality test mentioned above, is taken for cross-correlation with one of the constant string images. This correlation is performed for all the four orientations, mentioned above, of the template. A threshold T has been selected based on the study of autocorrelation of the template and its different rotated versions as shown in figure 3. In the correlated image if any pixel is found to be above the threshold, T, the CCF is declared to be present in the text block. This process is repeated for all the four orientations

[1] Orientation is defined to be the major angle with which the seal is present with respect to the original positioning of the page image. It is considered to be one of the following four: (i) normal orientation or aligned along with the text of the page — 0^o, (ii) up-side down orientation — 180^o, (iii) right orientation — 90^o, and (iv) left orientation — 270^o.

for all the three CCF templates. The cross-correlation of the building block with $CCF_1,\ CCF_2\ and\ CCF_3$ yields the correlated images $C_1,\ C_2\ and\ C_3$, respectively. images

The cross-correlated images are binarized with the thresholds $T_1,\ T_2,\ and\ T_3$, corresponding to the three CCF's. The centroid of the ON-pixel blobs are considered as the point of location of the CCF. These centroids are further taken for a topology test, to confirm the position of locality of the seal in the building block image. From figure 3, we have taken a value of 0.8, for T_1, keeping in mind that the system could tolerate a skew up to $\pm 3^o$. We have taken the value of 0.7 for both T_2 and T_3 based on similar studies on CCF_2 and CCF_3.

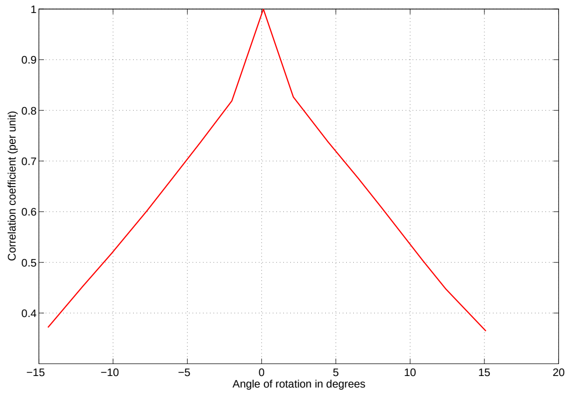


Fig. 3: The Graph demonstrates the auto-correlated values of CCF_1 at various rotation angles. The correlation coefficients have been normalized with the number of ON-pixels of CCF_1 at 0^o rotation.

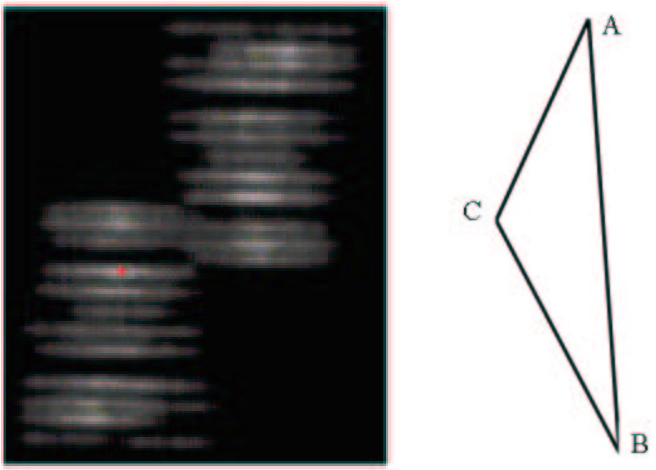
3) Topology Test: The relative position and the distance between the CCF's in the seal under consideration is our prior knowledge. This order of positioning with the known distances is known as the positional topology of the seal.

If a building block contains all the three CCF's, then the block is taken for what we call as a **Topology Test**. In this test, a triangle is formed by the three centroids, considering that each vertex of the triangle is derived from a different CCF. This has been amply demonstrated in figure 5. In this figure, the triangle ABC is formed by the vertices A, B and C. Here A corresponds to the centroid of the blob obtained by the binarization of the cross-correlated image of a building block with the CCF_1. The vertices B and C are derived similarly by using the CCF_2 and CCF_3 respectively.

Once the triangle is formed, it is tested for its consistency with the positional topology of the seal. We declare the triangle, ABC, to be consistent with the seal topology if the relative positions of the vertices are same as that of the positioning of the corresponding CCF's in the seal and distances are within a pre-specified tolerable limit. In this case the triangle ABC is said to pass the topology test and represent the seal.

B. False Positive Detection

While trying to detect the seals present in the document image, we came across cases where the non-seals were detected as seals. These are generally referred to as false positives. We observed that the rate of false positive occurrence is about 5.6%. This leads to higher inaccuracies at the system level. So, to improve the system accuracy we need to separate the false positives from actual seals. We have taken a transform based approach for this purpose.



(a) (b)

Fig. 4: (a) The cross-correlated image C_1 obtained by correlating a building block with CCF_1. (b) A triangle considered for topology test. In (b) the points A, B and C correspond to the centroid of the blobs obtained by binarizing the cross-correlated images $C_1,\ C_2,\ and\ C_3$.

In the training phase of this separation, we consider 35 seal images with another 35 false positives as our dataset. A linear sub-space is created using KL-transform for decomposition of the seal candidate images [7]. Using this basis, $\mathcal{B}$, which is a set of the eigen vectors coming from the covariance matrix of the image vectors, we decompose all the 70 images present in the dataset. Now each image is represented by a coefficient vector called a feature vector. Thus the training set contains 70 feature vectors.

At the testing stage, a prospective image is decomposed to its feature vector by projecting the image into this linear subspace. This vector, called a test vector, is compared against the training set using a neighborhood based classifier. This classifier declares the test image to be seal or non-seal.

C. Recognition System

We have assumed that a small amount of skew could be present even after the extracted seal image is orientation corrected. This is because we tolerate some amount of skew (up to $\pm 3^o$), as mentioned in the 2.A.2 section above, in the detection process. We employ a entropy based skew detection technique which detects skew with an accuracy of 0.1^o. The detected skew is corrected using a bilinear interpolation based matrix rotation technique [8].

The skew corrected image is segmented into lines, words and characters, in that order, as has been reported in [9]. The segmentation of the seal is into its constituent lines of text is based on the horizontal projection profile of the seal image. Each line of text is further segmented into words and characters using the vertical projection profile of each line of text. Once the character matrices are obtained from the segmentation module, the characters are size normalized to a pre-specified dimension and thinned. The matrix containing the thinned characters is bisected equally in horizontal and vertical directions. This generates 4 equal sectors. Each of these sectors is 2D discrete cosine transformed. 14 low-frequency coefficients, excluding the DC coefficient, in a zig-zag manner, are selected from each of these transformed sectors. The coefficients from each sector are appended to form a 56 dimensional feature vector, representing the character.

A set of feature vectors, called as training set, is formed by selecting a number of patterns from each class. The selection of these patterns is done in a class-representative manner. The feature vectors, evaluated from the test patterns, are compared against this training set using the nearest neighbor principle. This determines the class label for the test pattern. Some field specific post-processing rules are applied to enhance the accuracy of character recognition.

3. Experimental Results

Our Dataset contains 305 seal positive images with another 305 seal-negative images. Seal positive images are those in which at least one seal is found. Seal negative images are the ones without any seal (seal of the prescribed format).

We made two studies: (i) the accuracy of detection of the seal without the false positive separation (as has been described in section 2.B) – names Case-I, and (ii) the accuracy of the seal detection with the false positives being separated from the actual seals – case-II. If any of the seals present in an input image, in cases where more than one seal is present, we consider that the seal is correctly detected. Any non-seal being recognized as a seal is called a false positive. Reject points to the non-detection of seals when seals are actually present in the document image. The results of these experiments are presented below.

Table 1: Seal Detection and Localization Accuracy before false positive separation (Case-I) and after its employment (Case-II). The ”Seal Positive” and ”Seal negative” images refer to input images with and without a valid seal, respectively.

	No of Images	Image Category	Reject	False +ve
Case-I	305	Seal positive	6	15
	305	Seal Negative	–	19
Case-II	305	Seal positive	6	0
	305	Seal Negative	–	1

It can be noted in the table below that False positive detection system, described in section 2.B, has led to a substantial increase in the results. The amount of false positives present in Case-I is 34 which is reduced only one by this proposed system. Thus, considering the efficiency of Case-II, we have only 7 mistakes in 610 images. This gives us an accuracy of 98.8%.

We also considered the accuracy of the recognition system. After the employment of knowledge based post-processing rules, the system is evaluated to operate with a character level accuracy of 98.5%. But since the aim of the project is to read the information present in the specified fields, it makes sense to see how many were correctly evaluated. An independent evaluation of the system on a much larger dataset, containing about 1200 images, puts the field level accuracy at 82%.

4. Conclusion & Discussion

Here, we present a scheme for seal information extraction. The first part of the scheme deals with detection and localization of the seal with its orientation estimated to an integer multiple of 90^o. This has been seen to be working with a tolerable skew of $\pm 3^o$. Subsequent character recognizer algorithm employed extracts the information with about 98% accuracy. This whole process takes about 23 seconds in an average when tested on our two previously specified datasets.

The field level accuracy of 82% is fairly large. But we obseved that a lot of recognized fields exist with a mis-recognition of a single character. Many times, we miss the fields because of the use of a sub-optimal rule-set. This rule-set is supposed to detect the keywords associated with the fields and point to the location of the information string. An improved version of this rule-set is expected to further the accuracy.

The rejection of seals, in the seal detection part, is mostly due to the inherent loss of information in the CCF's of the seals. A closer look at the rejected seals revealed that the print of those seals was lighter than the regular ones and, hence, the characters were thin. This led to correlation values of the CCF's not reaching the set threshold values, leading to failure to detect these seals. The other reason for failure to detect a valid seal was the improper positioning of the seal. For example, in a case it was observed that the different CCF's of the seal were oriented in different directions, exceeding the specified skew limit, thus, leading to the failure.

References

[1] A. Soria Frisch, "The fuzzy integral for color seal segmentation on document images," in *Proc. of Intl. Conf. on Image Processing*, 2003, vol. 1, pp. 157–160.

[2] Katsuhiko Ueda, Takeshi Mutoh, and Ken'ichi Matsuo, "Automatic verification system for seal imprints on japanese bankchecks," in *Proc. of Fourteenth Intl. Conf. on Pattern Recognition*, 1998, vol. 1, pp. 629–632.

[3] K. Roy, S. Vajda, U. Pal, and B.B. Chaudhuri, "A system towards indian postal automation," in *Proc. of 9th Intl Workshop on Frontiers in Handwriting Recognition*, 2004, pp. 580–585.

[4] K. Roy, S. Vajda, U. Pal, B.B. Chaudhuri, and A. Belaid, "A system for indian postal automation," in *Proceedings of the International Conference on Document Analysis and Recognition*, 2005, vol. 2, pp. 1060–1064.

[5] John Y. Chiang and R. C. Wang, "Seal identification using the delaunay tessellation," *Proc. Natl. Sci. Counc. ROC(A)*, vol. 22, no. 6, pp. 751–757, 1998.

[6] R C Gonzalez and R E Woods, *Digital Image Processing*, Addison Wesley Publishing Co., New York, 1993.

[7] Thomas Vetter and Tomaso Poggio, "Linear object classes and image synthesis from a single example image," *IEEE transaction on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 733–742, 1997.

[8] K Mahata and A G Ramakrishnan, "Precision skew detection through principal axis," in *International Conf. on Multimedia Processing and Systems*, IIT Chennai, 2000.

[9] Aparna K. G. and A. G. Ramakrishnan, "A complete tamil optical character recognition system," in *LNCS-2423*, 2002, pp. 53–57.