2009 10th International Conference on Document Analysis and Recognition

Seal detection and recognition : An approach for document indexing

Partha Pratim Roy
Computer Vision Center
UAB, Barcelona, Spain
partha@cvc.uab.es

Umapada Pal
CVPR Unit, ISI
Kolkata, India
umapada@isical.ac.in

Josep Lladós
Computer Vision Center
UAB, Barcelona, Spain
josep@cvc.uab.es

Abstract

Reliable indexing of documents having seal instances can be achieved by recognizing seal information. This paper presents a novel approach for detecting and classifying such multi-oriented seals in these documents. First, Hough Transform based methods are applied to extract the seal regions in documents. Next, isolated text characters within these regions are detected. Rotation and size invariant features and a Support Vector Machine based classifier have been used to recognize these detected text characters. Next, for each pair of character, we encode their relative spatial organization using their distance and angular position with respect to the centre of the seal, and enter this code into a hash table. Given an input seal, we recognize the individual text characters and compute the code for pair-wise character based on the relative spatial organization. The code obtained from the input seal helps to retrieve model hypothesis from the hash table. The seal model to which we get maximum hypothesis is selected for the recognition of the input seal. The methodology is tested to index seal in rotation and size invariant environment and we obtained encouraging results.

1. Introduction

Indexing of documents having certain seals or logos is an efficient way to arrange documents with specified source. Undoubtedly, seal detection and recognition is an important stage to follow this approach. The information obtained from seals could be used for efficient storage and retrieval of the documents. It allows to identify the document sources reliably. But, detection and recognition of seal is a challenging task. Because, seals are generally unstable and sometimes contain unpredictable patterns due to imperfect ink condition, surface contact, and noise [14]. Overlapping with text/signatures and occlusion are typical problems. Seals generally bear some constant character strings to convey the information about the owner-organization and its locality. Besides, in many instances seal contains some variable fields [6] for e.g., date. A seal instance may be affixed on any position within the document, which requires detection to be carried out on the entire document. Also a seal may be placed in arbitrary orientation. All these conditions described above make the seal detection and recognition process complex. See Fig.1(a), where a part of historical document containing seal is shown. The seal is overlapped with part of signature and other regions and some information of this is missing.

There are some research papers on detection and recognition of seal images in documents. Different functionalities of Document Image Analysis (DIA) have been applied by various researchers for this purpose. A prior knowledge of seal-shape structure is helpful to localize them in documents. Zhu et al. [14] have proposed a seal detection method based on elliptical structure recognition. Due to the fact that, in many instances seals and text are in different colors, some researchers have studied the detection process based on color analysis. Frisch [3] presented a fuzzy integral based technique that selectively extracts color clusters in images and uses color information to detect seal. Also Ueda et al. [12] proposed a method for extraction and verification of seals and signatures in Japanese bank-cheques using color analysis. Sometimes seal has been treated as a symbol and the methodology due to symbol recognition have been applied for seal recognition. Chiang et al. [2] used the Delaunay tessellation for segmentation. The image is segmented into a mesh of triangles and the distribution of the triangular area was used as the primary feature. A symbol spotting based approach is also carried out to rank symbols due to its property of fast retrieval without the requirement of segmentation. Rusinol et al. [10] and Tombre et al. [11] made use of this method for matching certain discriminative features of a given query image and to get the ranking of the output images according to voting in the accumulation array. Some researchers have used the text information in image for recognition. Nourbakhsh et al. [6] considered the template matching of fixed string field with the model shape for recognition. A set of seal orientation

978-0-7695-3725-2/09 $25.00 © 2009 IEEE
DOI 10.1109/ICDAR.2009.128

were carried out in their process. Mao et al. [5] used the recognition result of fixed-pitch text characters for recognition of IATA flight coupons.

When imprinted, seal produces a combination of graphical symbol and text components by its nature. It contains text characters in different rotations and a geometrical shape (circular, rectangular, etc.) to cover these text characters. This paper presents a novel idea for detecting and classifying such multi-oriented seals in documents. We combine methods of graphical shape recognition and text character recognition and use them as the basis of our approach. The recognized text characters within seal are used as high label key-descriptor in our method. Inspired from Generalized Hough Transform [1] a voting architecture is employed for classification of seals. Here, first Hough Transform based methods are applied to extract the seal regions in documents. Next, isolated text characters within these regions are detected. Rotation and size invariant features and a Support Vector Machine based classifier have been used to recognize these detected text characters. Next, for each pair of character, we encode their relative spatial organization using their distance and angular position with respect to the centre of the seal, and enter this code into a hash table. Given an input seal, we recognize the individual text characters and compute the code for pair-wise character based on the relative spatial organization. This code retrieves model hypothesis from the hash table. The seal model to which we get maximum hypothesis is selected for the recognition of the input seal.

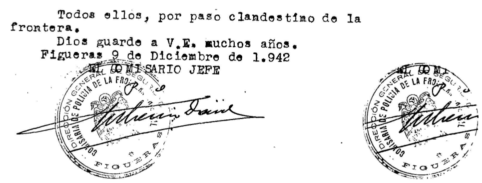
The rest of the paper is structured as follows. In Section 2, we explain the seal detection procedure in brief. Character extraction and recognition procedure is discussed in Section 3. In Section 4, we describe the representation of the seal models and their recognition process. The experimental results are presented in Section 5. Finally conclusion is given in Section 6.

2. Seal Detection

Seals generally consist in geometric shapes which have an analytical model such as circular or rectangular shapes. Our work was focused on detecting such specific shape of seal. Hough Transform (HT) is used to detect and verify circles and rectangles in documents. Due to the fact that, noise and other structures can degrade the accuracy of estimation in Hough Space, Run-Length Smoothing Algorithm ($RLSA$) is used to remove non-seal regions. On this smoothed image we apply component labeling to get individual blocks. Based on the property that, seal contains many small components compared to that of other text portion, we extract the blocks which surround smaller components [7]. Next we detect and verify the shapes using HT as discussed below.

Standard Hough Transform is applied here to detect line segments in the document. Since, C(ρ, θ) represents the number of edge points satisfying the linear equation $\rho = x\cos\theta + y\sin\theta$, a simple way to find peaks of the HT is to extract all points satisfying $C(\rho,\theta) \geq T_C$ (i.e., all straight lines with T_C or more pixels would be retrieved). When a straight line is placed in a parallel position in image space, only a proportional displacement along the θ -coordinate takes place in parameter space. If a straight line is rotated in image space, its associated peak in the accumulator space is shifted along the axis. Based on the peaks of the Hough image a rectangle is detected using certain geometric conditions as proposed by Jung et al.[4].

We have used the traditional Circle Hough Transform (CHT) [13] for circle detection in our application. It is done by first computing the gradient of the original image (Sobel gradient) and accumulating each non-zero pixels from the gradient image. The edge pixels that lie along the perimeter of a circle of the given radius, contribute a vote at the center of the circle. Hence, peaks in the transformed image correspond to the centers of circular features of a given range in the original image. Circular seal extraction is shown in Fig.1(b).



(a) (b)

Figure 1. (a) A part of historical document containing seal. (b) Extraction of circular seal in document

3. Character Extraction and Recognition

After detecting the location of seal, we extract the region from the original image, placing a mask of the corresponding shape at that location. Now the text characters within this region are separated and recognized as follows.

Text character separation: There may exist long lines due to form/table present in document or strokes from signature which can touch/overlap the text characters in the seal, as we can see in Fig.1(b). We call these long lines as $Graphical\ Lines$. To separate the text characters from the graphical lines, we have used the algorithm proposed in [9]. Here, the aim is to separate them into two layers mainly, $Text$ and $Graphics$ layers. Assuming the length of character segments are smaller compared to that of graphics, a histogram analysis on the size of connected components is done for separation of text parts. By a threshold selection, obtained dynamically from the histogram, large graphical parts touching with text are detected, leaving the smaller text components. Next, the skeleton information of these large components are analyzed to check the presence of long segments. For this purpose, all the segments of the skeleton are decomposed at the intersection point. Then the segments having longer length are removed. The remaining portion are considered as parts of text layer. Text layer extraction result of our algorithm when applied on Fig.1(b) is shown in Fig.2.

Figure 2. Long graphical segments of Fig.1(b) are removed

Feature extraction and recognition: Segmented text components are analyzed for feature extraction. To make the system rotation invariant, the features are mainly based on angular information of the external and internal contour pixels of the characters. Circular ring and convex hull ring based concept have been used to divide a character into several zones to compute features [8]. Given a sequence of contour pixels, the change of angle of each pixel is calculated from the neighbour pixels. The angles obtained from all the contour pixels of a character are grouped into 8 bins corresponding to eight angular intervals of 45°. For a character, frequency of the angles of 8 bins will be similar even if the character is rotated at any angle in any direction.

To get local features, a set of circular rings is defined as the concentric circles considering center as the center of minimum enclosing circle (MEC) of the character and the minimum enclosing circle as the outer ring. Similarly, convex hull rings are also constructed from the convex hull of the shape of the character. The radii of the rings are in arithmetic progression. In Fig.3 we show four circular and convex hull rings. MEC is circular in nature for all objects. Whereas, convex hull of an object is a closed polygon and its shape differs depending on the shape of objects. Convex hull ring may not be circular and it is better adjusted to the shape of the object.

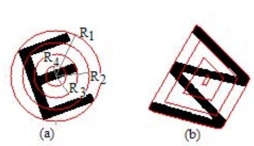


Figure 3. (a) Circular rings and (b) Convex hull rings.

The features are distributed in more small zones to obtain higher accuracy. Each ring is divided into few segments based on a reference line which is invariant to character rotation corresponding to that character. The reference line is detected based on the background part of the character using convex hull property [8]. Finally, combining circular and convex hull features, we have 512 dimensional feature vector for classification. The feature vector is normalized to obtain size invariance.

We have reported the character recognition algorithm using support vector machine in our earlier paper [8]. Both uppercase and lowercase alpha-numeric characters were considered for our experiment. Because of shape similarity of some characters/digits due to arbitrary orientation, we group them in 40 classes. To get an idea of training data quality for SVM classifier, we have shown some samples in Fig.4. Given a connected component image of a seal, we use SVM to compute the recognition confidence to obtain the corresponding character class. If recognition confidence is higher than a threshold, we consider that component for our seal recognition method.

Figure 4. Training Data from our database

4. Seal Recognition Architecture

Our seal recognition scheme is inspired from the Generalized Hough Transform (GHT) [1]. GHT uses edge information to define a mapping from the orientation of an edge point to a reference point of the shape and detect the shape based on the accumulation of mapping to that reference point. We will describe the architecture of our method with three key-parts namely, representation, model generation and compilation of hash table and hypothesis voting and recognition.

Representation: Our representation model for a seal is based on the recognition of good text characters (to be recognized with high confidence) within seal region. Since, we want to take care of occlusion and overlapping, we do not expect to obtain complete list of individual characters from the input seal image, but only a subset of characters. Again, text characters are too local and very generic to be used for matching primitives. To extract the knowledge about the seal, we encode each pair of recognized text characters with simple basic features: relative angle and their distance. See Fig.5, where a character pair ('A' and 'D') is shown in a seal. The attributes used in our approach are described below.

Reference Point: The centre of the minimum enclosing circle (MEC) of the seal is considered as the reference point (R). Note that, R will be fixed for each class of seal shapes in size and rotation invariant way.

Character Pair code: We use a Look Up Table (LUT) to find a suitable code CP for each pair of text characters. If the number of distinct character class is n, then the length of LUT is $(n \times n)/2$. As we have considered multi oriented text character class in our recognition classifier (we used SVM classifier [8]), our system produces 40 classes of character shapes (discussed earlier). So, in our case, the length of LUT is 800.

Relative Angle: We considered two angles for each character pair. We compute the CG (centre of gravity) of each character for this purpose. For each pair of character we can consider a triangle by their CGs with R (see Fig.5). Angle α_r is defined as the angle formed by the CGs of the character pair at the reference point. The other angle, α_n is the next angle of the triangle considering anti-clockwise direction from α_r. In Fig.5, these angles are "$\angle DRA$", "$\angle RAD$" respectively.

Distance: The distance of character pair is defined as the Euclidean distance D of the CGs ("AD" in Fig.5) between them. The distance is normalized by the diameter of the MEC of that seal. Hence, the value of distance will be between 0 and 1.

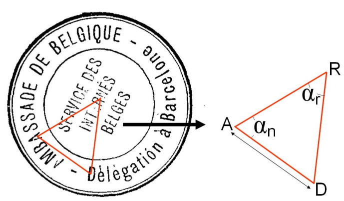


Figure 5. A character pair is shown in a seal. Here, R is the reference point.

We make our feature descriptor of seal using the positional information of the character pair using character-pair code, angle information and length. These positional information are encoded into a 4-tuple key (k) as given below.

$$k = \{CP, \alpha_r, \alpha_n, D\} \quad (1)$$

k is used as keys into a hash table, where we record the corresponding model class information as entries. Hash table, is used to coarsely quantize the n-dimensional descriptor space, and quickly retrieve all the key-points having a similar description. A hash function H projects a value from a set of members to a value from a set with a fixed number of (fewer) members. They are useful to find an entry in a database by a key value. The robustness is that, as there exist few text characters in a seal and the positional information among them is also fixed, so their relative position is invariant to scale and rotation, it is good enough to handle them using our above representation for quick access.

Model generation and hash table compilation: As mentioned above, we represent our model images with the distance and relative angle of each character pair with respect to reference point. For hash table generation, we manually remove the graphical (non text regions) portions and variable field (like date information) from each class of seal images to be used as model. The remaining text characters are recognized by SVM. Then for each pair of text characters we find out the angle and distance between them. We encode these information in 4-tuple key which describes the primitives of the class model and enter them in the hash table. If we consider of total m models in our database, then, each class model C_i is thus represented by a list of high level descriptors K_i where,

$$K_i = \{k_{i1}, k_{i2}..k_{ip}\}$$

where, p is the total number of character pairs found from model C_i and k_{ij} describes the encoded 4-tuple key of jth character pair using (1). Finally, all the character-pair descriptors of each class model are stored in the hash table. See Fig.6, where we have shown the model generation and seal recognition architecture in a diagram. Here, the hash table is compiled using the key descriptors of 4 different model classes of seal.

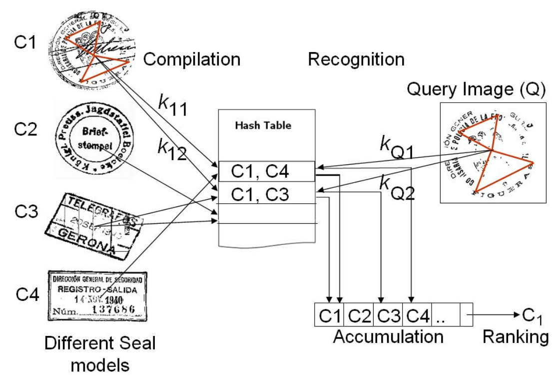


Figure 6. Model generation and Seal recognition architecture

Hypothesis voting and recognition: By using indexing for the hypothesis generation process, we select a candidate models that are likely to be similar to the input seal. Given an input seal image, we used our text graphics separation method (as discussed above) to extract text component from the seal. Next, SVM classification process is employed to recognize multi-sized and multi-rotation text characters present in the seal. The character components having good recognition accuracy are selected for matching.

For all pair of these selected components, we compute the relative angle and distance. These are encoded in a 4-tuple index key (K_Q) for the query image Q (see Fig.6). Using this encoded key, we check our hash table and find out possible class entries. The index key indicates the presence of high level character-pair descriptors in the model classes, retrieved from the hash table. As the seal is formed by several text characters, high voting to a model class points out the presence of many high level character-pair descriptors in the model. The more we will find the matching of character pair descriptor to a model, the more probable that model be the similar. We make an accumulator of size m, comprising of all class models and initialize it with 0s. We increment the counter of class whenever we find a matching entry in the hash table. Finally, the classes are ranked in descending order with the number of voting and the model having maximum accumulation is selected as the recognized seal.

5. Experimental Results

For the experiment, documents containing seal were collected from different sources like postal letters, official documents from universities etc. Seals of circular and rectangular shapes were considered for the experiment. The documents were digitized by a flatbed scanner at 200 dpi. We constructed the database with 127 documents containing seals of English text characters to test the seal recognition approach. There were total 12 classes. For index key generation of each character pair extracted from seal, the value of angles are quantized to 2 degree and the normalized distance is also quantized into 5 division. The quantizations are done based on experimental result. We have obtained 92.03% accuracy on taking the top-choice of voting. To get the idea of our seal recognition results, we show the result using different top choices (see Table.1). The recognition accuracy is affected mainly due to broken text characters in the seal and the presence of long graphical lines over text regions. If broken components of a character could not be joined through preprocessing, the character recognition algorithm does not perform well. We have also checked the robustness of our method from the number of selected text character components in the query seal. It is noticed that, our system can recognize the seal even if only 55% of non-variable text characters of seal are extracted. The more are the number of retrieved text characters, the higher the recognition probability of the seal.

Table 1. Seal recognition result based on different top choices

Number of top choices	Recognition rate
Only 1 choice	92.03%
Only 2 choice	93.81%
Only 3 choice	98.23%
Only 4 choice	99.1%
Only 5 choice	100%

6. Conclusion

We have presented an indexing approach of documents based on seal information retrieval and recognition. The contribution in this paper is two-fold. We have used a combination of symbol recognition and text recognition to take care of complex shaped multi-oriented and multi-sized seals. The recognition result of character components within the seal region are used to generate high-level descriptor to classify the seal. Relative positional information of text string characters are used for this purpose and hypothesis were generated based on that. The result shows that, our approach is efficient even if we do not extract all the text characters in a seal.

7. Acknowledgments

This work has been partially supported by the Spanish projects TIN2006-15694-C02-02, TIN2008-04998 and CONSOLIDERINGENIO 2010 (CSD2007-00018).

References

[1] D. Ballard. Generalizing the Hough transform to detect arbitrary shapes. *Pattern Recognition*, 13(2):111–122, 1981.

[2] J. Y. Chiang and R. Wang. Seal identification using the Delaunay tessellation. In *Proc. Natl. Sci. Counc. ROC (A)*, volume 22, pages 751–757, 1998.

[3] A. S. Frisch. The fuzzy integral for color seal segmentation on document images. In *Int. Conf. on Image Processing*, volume 1, pages 157–160, 2003.

[4] C. R. Jung and R. Schramm. Rectangle detection based on a windowed hough transform. In *Proceedins of the XVII Brasilian Symposium on Computer Graphics and Image Processing (SIBGRAPI04)*, pages 113–120, 2004.

[5] J. Mao, R. A. Lorie, and K. M. Mohiuddin. A system for automatically reading IATA flight coupons. In *ICDAR*, pages 153–157, USA, 1997.

[6] F. Nourbakhsh, P. B. Pati, and A. G. Ramakrishnan. Automatic seal information reader. In *Int. Conf. on Computing: Theory and Applications*, pages 502–505, 2007.

[7] K. Roy, S. Vajda, A. Belaid, U. Pal, and B. B. Chaudhuri. A system for indian postal automation. In *ICDAR*, pages 1060–1064, USA, 2005.

[8] P. P. Roy, U. Pal, J. Lladós, and F. Kimura. Convex hull based approach for multi-oriented character recognition from graphical documents. In *ICPR*, Tampa, USA, 2008.

[9] P. P. Roy, E. Vazquez, J. Lladós, R. Baldrich, and U. Pal. A system to segment text and symbols from color maps. *GREC. Selected Papers*, pages 245–256, 2007.

[10] M. Rusinol and J. Lladós. Word and symbol spotting using spatial organization of local descriptors. In *In Proceedings of Document Analysis Systems*, pages 489–496, 2008.

[11] K. Tombre and B. Lamiroy. Graphics recognition - from re-engineering to retrieval. In *ICDAR*, pages 148–155, 2003.

[12] K. Ueda, T. Mutoh, and K. Matsuo. Automatic verification system for seal imprints on japanese banckchecks. In *ICPR*, pages 629–632, 1998.

[13] H. Yuen, J. Princen, J. Illingworth, and J. Kittler. Comparative study of Hough transform methods for circle finding. *Image and Vision Computing*, 8:71–77, 1990.

[14] G. Zhu, S. Jaeger, and D. Doermann. A robust stamp detection framework on degraded documents. In *SPIE Conference on Document Recognition and Retrieval*, 2006.