Extraction of light and specific features for historical image indexing and matching

Mickaël Coustaty, Surapong Uttama, Jean-Marc Ogier

Extraction of light and specific features for historical image indexing and matching

Mickael Coustaty
L3i Laboratory
University of La Rochelle
La Rochelle, France
mcoustat@univ-lr.fr

Surapong Uttama
School of Information Technology
Mae Fah Luang University
Chiang Rai, Thailand
usurapong@gmail.com

Jean-Marc Ogier
L3i Laboratory
University of La Rochelle
La Rochelle, France
jmogier@univ-lr.fr

Abstract

This paper proposes a new feature extraction technique for indexing and matching historical images. The complexity of these historical images troubles the existing indexing approaches due to their line patterns. Our indexing method relies on the global segmentation integrated with the knowledge of edge density to deal with the line patterns and eliminate over-segmented regions. Then the historical images are represented by a set of ROI descriptors and the retrieval is performed by matching sorted features. Experiments and system evaluations with and without ground truth are presented and discussed.

1. Introduction

Line pattern images were broadly used in ancient printed documents especially in medieval age. These old images are, on the one hand, precious in terms of cultural heritage. On the other hand, they are challenging in field of document image analysis and recognition because of their complex patterns. Their composition is the combination of different line patterns to form shapes and backgrounds, and often letters (see figure 1). Digital libraries attempt to provide accessibility to these digitalized images for historians and publics. An example is the Virtual Humanistic Libraries in Tours [2] that prepares searchable database of several digitalized ancient documents including line pattern images.

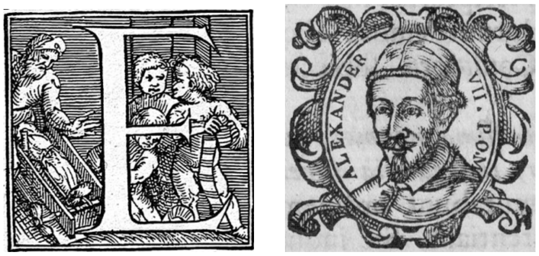

Figure 1. Ancient Line Pattern Images

Unfortunately the mentioned image retrieval system is keyword-based. If users need to look for an image, they have to input keywords in specified categories such as pattern, letter, font, etc. Therefore, in case users have an image and wish to search for similar images from database. They must understand and be able to describe the characteristics of this image which will be real difficulty with ancient images.

A content-based image retrieval system (CBIRS) is an interesting proposal to this condition. It generally comprises of two parts: indexing and retrieval. Indexing is mainly for feature extraction and description while retrieval always relies on feature matching.

Due to the complexity of line pattern images and the incomplete ground truth, not many related works are found. Indexing of ancient line pattern images was proposed in [6, 3, 4] where Zipfs law is applied to index and classify these images into predefined categories according to their patterns. This approach requires no segmentation. [4] focused more on recognition of a component in line pattern images by using decomposition method. These techniques provided promising results but are more suitable for classification or recognition than retrieval. A complete indexing and retrieval of line pattern images were found in [7] and [8]. The authors proposed a bag of strokes approach whose results were time consuming in the first article, while the second contribution used a full segmentation algorithm based on texture feature which requires high computing cost and induces lag in retrieval.

For our case, after reviewing several approaches in [5] it is observed that neither indexing with full segmentation nor without segmentation is the best option. The first must deal with the complex image patterns, lack of ground truth and computing cost. The latter always depends on keypoint detections i.e. edges, corners or junctions. These keypoints would not represent salient points in line pattern images due to its complicated patterns. Therefore, a partial segmentation becomes our first option.

The major contribution of this paper is on the proposal of the novel indexing technique based on edge density to partially segment line pattern images for CBIRS. Our approach is unsupervised and content-based which targets to balance the retrieval accuracy and speed.

This paper is organized as follows: the next section presents the outline of the proposed method. Section 3 focuses on the image indexing and Section 4 explains the image retrieval. Section 5 presents the experimental setup and results. The conclusions and discussions are mentioned in Section 6.

2. Outline of the Approach

Our working scheme is two phases: indexing and retrieval. This format is nearly identical for all CBIRS. figure 2 illustrates the global process of our approach. Indexing is done offline and in advance. Images are segmented and all extracting features are kept in the database. Then the query image is processed with the same indexing algorithm but its features are used for matching with the existing images to find the best match and ranked results.

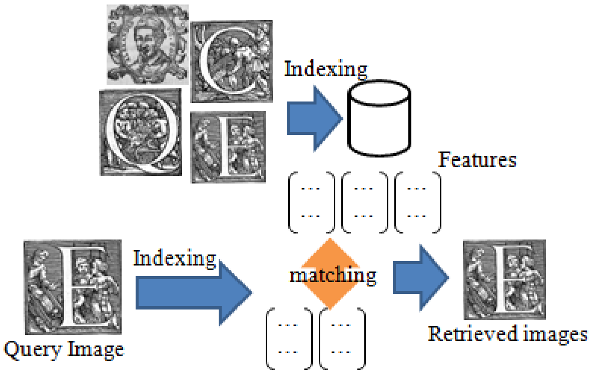

Figure 2. Approach Overview

3. Indexing

As mentioned earlier, our indexing technique is based on partial image segmentation. Then the feature extraction is applied for segmented regions.

3.1. Segmentation

It is necessary to mention that we do not apply full image segmentation. The focus of this paper is not to get the most accurate segmentation result which requires high computing complexity. But we need a partial segmentation algorithm that can produce stable segmented outputs for feature extraction.

We propose a segmentation concept based on edge density. Generally if we apply the region-based segmentation directly to the line pattern image, the result will be oversegmented because regions in such image have no exact boundary. The idea is to weaken the line patterns so that we can get the rough segmentation with smaller number of oversegmented regions. Then we find the correspondence of these segmented regions with edges. Any region having more edges should be more significant than those with fewer edges because they represent the patterns in our images.

Our segmentation method comprises of the following steps:

- First we decompose an image with Meyer decomposition [1] to filter out some strong textures and noises. We take only the so-called shape layer [4] for further process (figure 3 left).
- We blur it by Gaussian lowpass filtering (7x7 mask, $\sigma = 5$).
- We suppress all maxima (height = 50). At this stage we will have weaker line patterns.
- We segment it with watershed algorithm (figure 3 right)
- We compute the distance transform of each region and normalize it. The distance transform will emphasize the importance of edges. The edges near the border of regions are considered less significant than those which are closed to the region center.
- Finally we detect edges of the filtered image (first step) using Canny edge detection (figure 4 left).

Thus the initial outputs of segmentation are two images: distance transform image and edge image.

Then we map each segmented region (distance transform image) with edges in its region (edge image) to calculate the edge density (ED_R) defined by:

$$ED_R = \sum_{i \in R} \frac{D_i}{P_R}, \qquad E_i \neq 0 \qquad (1)$$

where R is the segmented region of interest (ROI). D and E are distance transform image and edge image. $R \subset D$ and $R \subset E$. P_R is the perimeter of the region R.

As earlier mentioned, the previously segmented image is usually over-segmented due to the rough segmentation. A segmented ROI should be a region having enough edge density. Otherwise it should be only a non-significant and over-segmented region. Therefore, we set a threshold to eliminate these regions (experimentally 0.45) to get our final segmentation. figure 4 (right) demonstrates the segmented image. Comparing to rough segmentation in figure 3 (right), it is clearly observed that there are some deleted segmented regions (white areas of figure 4 right) due to the insufficient edge density.

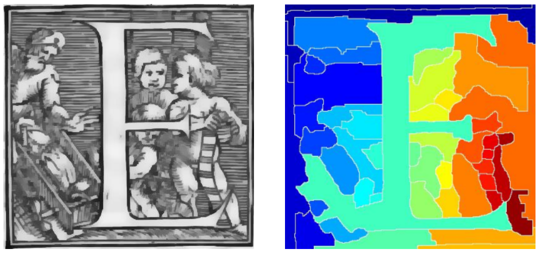

Figure 3. Filtered Image and Watershed Segmentation Result

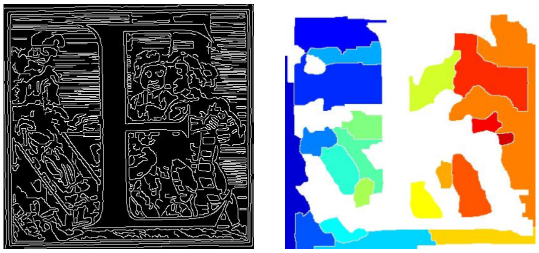

Figure 4. Edge Image and Final Segmentation

3.2. Feature Extraction

The feature extraction is performed on the segmented regions. To ensure that we have sufficient features, we select edge density and extract another region property called eccentricity. It is defined by the ratio of distance between foci to major axis length of an eclipse surrounding the segmented ROI. In summary, two features are extracted from each segmented ROI *i.e.* edge density and eccentricity. Then a given line pattern image L_i is represented by its n_i feature regions:

$$L_i = \{(ED_k, EC_k)\}, \qquad k \in \{1 \ldots n_i\} \qquad (2)$$

where EDk and Eck are edge density and eccentricity of the kth segmented region.

4. Retrieval

The classic bag-of-visual-words model inspires our retrieval algorithm where an image is presented by a group of unordered and clustered features. However, we adjust this model by not clustering the features and defining order of features. Firstly, the features must be sorted by the most significant feature i.e. edge density for our case. Then one-by-one comparing of sorted features does the matching between a query image and each line pattern images in the database. The matching score (M) is defined by:

$$M(Q_i, L_j) = \sum_{k=1}^{2} \sum_{i=1,j=1}^{N} (F^q_{i,k} - F^l_{j,k}) \qquad (3)$$

where Q_i and L_j are query image and local image. $F^q_{i,k}$ is a k^{th} feature of i^{th} order of query image and $F^l_{j,k}$ is a k^{th} feature of j^{th} order of local image. N is $max\{N_q, N_l\}$. N_q and N_l are number of ROI of image Q_i and L_j respectively. Therefore, if $N_q \neq N_l$, the matching score will have penalties and increases.

We performed matching between query image and all local images. The matching scores were then sorted in ascending order. The less the matching score is, the closest of the images becomes.

5. Experiments and Evaluations

Our experiments were performed by using 441 ancient line pattern images including 358 graphical drop caps and 83 portraits (see figure 1) provided by the Virtual Humanistic Libraries in Tours [2]. All these images were extracted from digitized ancient documents and some are slightly different. The image sizes vary from 400x400 pixels to 800x1000 pixels and all are in grayscale.

Unfortunately due to the lack of ground truth and relevant information of images, it is not possible to evaluate our method with standard tool such as precision-recall curve. Therefore, we develop our own ground truth and split the test into two categories.

5.1. Precision at 1 (P@1)

The first experiment is to retrieve a query image from indexing database and observe only the highest rank. If

Measure	Number of tested queries	Min (%)	Max (%)
P@1	441	100	100
P@n	104	44.2%	76.4%

Table 1. Evaluation

the first retrieval image matches with the query image, precision is 1 or 0 otherwise:

$$P@1_i = \begin{cases} 1 & \text{if } L_i = Q_i \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

where $P@1_i$ is the precision at 1 of an image i. L_i and Q_i are retrieved image and query image respectively. Our experiment on all images showed that the average $P@1$ is 1 or 100%. In other words, our method always returns the first correct match.

5.2. Precision at n (P@n)

In reality, the average P@1 is not sufficient to evaluate the CBIRS. It does not present the robustness of the retrieval because the second to last top ranks are not taken into account. Thus we propose another evaluation called precision at n (P@n):

$$P@n_i = |\{RV_i\} \cap \{RT_{i,n}\}| / |\{RT_{i,n}\}| \quad (5)$$

where $P@n_i$ is the precision at n of an image i. RV_i is the relevant images. $RT_{i,n}$ is the retrieved image at rank n^{th}. n is the minimum number of top rank images containing all relevance. Usually $|RV_i| \leq |RT_{i,n}|$.

To have relevant set of images, we found that in the database of line pattern images, there are 16 sets of nearly similar images (figure 5). They are just different only from the slight distortion and translation. The minimum and maximum numbers of these similar images in sets are 7 and 15 respectively. The total number of these images is 104. The experiments revealed that the P@n has the broad range starting from 44.2 to 76.4. We do not compute the average P@n due to the different value of n for each query. It is observed that the low P@n values are related to the complexity of line patterns of images.

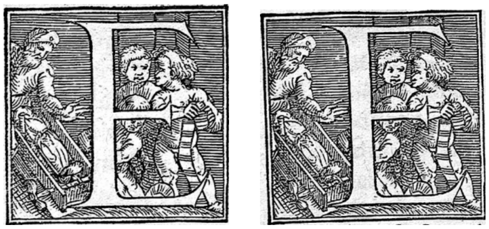

Figure 5. Two slightly different images

6. Conclusions and Discussions

In this paper we present a method for indexing and retrieval of ancient line pattern images. The indexing challenge is the complex line patterns that prevent the segmentation-free algorithm such as edges, lines, corners or junctions detection. So our proposal relies on the global segmentation and fine tuning of the segmented regions by using edge density. Then we extract features from ROI and implement CBIRS. The retrieval experiments present the effectiveness of our method especially for the first match. For rank retrieval, the precision at n drops and has a wide range. This could be further observed through the precision-recall curve for each query. In addition, we found the close relationship between the precision and line pattern complexity. Most of the query images with complex patterns show average to low precision. We expect that the main cause is at the watershed segmentation process and could be further improved. The future work should also include the improvement of retrieval algorithm, trying more features, and experiments on larger number of images.

References

[1] J.-F. Aujol, G. Gilboa, T. Chan, and S. Osher. Structure-texture image decomposition - modeling, algorithms, and parameter selection. *International Journal of Computer Vision*, 67(1):111–136, 2006.

[2] BVH. Les bibliothèques virtuelles humanistes - centre d'etude supérieur de la renaissance - *http://www.bvh.univ-tours.fr/*.

[3] H. Chouaib, F. Clopet, and N. Vincent. Graphical drop caps indexing. In *Eighth IAPR International Workshop on Graphics Recognition*, pages 179–185, La Rochelle, 2009.

[4] M. Coustaty, R. Pareti, N. Vincent, and J.-M. Ogier. Towards historical document indexing: extraction of drop cap letters. *IJDAR*, 14(3):243–254, 2011.

[5] Y. Liu, D. Zhang, G. Lu, and W.-Y. Ma. A survey of content-based image retrieval with high-level semantics. *Pattern Recogn.*, 40:262–282, 2007.

[6] R. Pareti and N. Vincent. Ancient initial letters indexing. In *ICPR '06: Proceedings of the 18th International Conference on Pattern Recognition*, pages 756–759, 2006.

[7] J.-M. O. Thi Thuong Huyen Nguyen, Michal Coustaty. Bags of strokes based approach for classification and indexing of drop caps. In *ICDAR 2011*, pages 349–353, 2011.

[8] S. Uttama, P. Loonis, M. Delalandre, and J.-M. Ogier. Segmentation and retrieval of ancient graphic documents. *LNCS*, pages 88–98, 2006.