Logo Spotting For Document Categorization

Viet Phuong LE[1], Muriel VISANI[1], Cao De TRAN[2], Jean-Marc OGIER[1]
[1]Laboratory L3I, Faculty of Science and Technology, La Rochelle University, France
[2]College of Information and Communication Technology, Can Tho University, Vietnam
{viet_phuong.le, muriel.visani, jean-marc.ogier}@univ-lr.fr, tcde@cit.ctu.edu.vn

Abstract

Logo spotting is of a great interest because it enables to categorize the document images of a digital library of scanned documents according to their sources, without any costly semantic analysis of their textual transcript. In this paper, we present an approach for logo spotting, based on the matching of keypoints extracted both from the query document images and a given set of logos (gallery) using SIFT. In order to filter the matching points and keep only the most relevant, we compare the spatial distribution of the matching keypoints in the query image and in the logo gallery. We test our approach using a large collection of real world documents using a well-known benchmark database of logos and show that our approach achieves good performances compared to state-of-the-art approaches.

1. Introduction

The last decades have seen the explosion of the amount of digitized document libraries. In order to properly index these documents, it is necessary to categorize them. In the first years, several document classification systems were investigated, based on OCR and analysis of text by natural language processing. However, OCR systems reach good performances only with typewritten and printed documents, and the natural language processing depends greatly on the context. On the other hand, graphical objects in the document images contain much important information. In particular, logos (as well as stamps) are commonly used in documents, especially in business and administrative documents. It allows us to determine the source of the documents quickly and accurately, without any textual transcription and at a low cost. As a consequence, logo spotting is of great interest to the field of document image analysis and recognition. It consists in matching a set of query document images (documents to be categorized) towards a set of known logos (logo gallery).

In the earlier works, Doermann *et al.* [4] presented an approach to logo recognition based on combination of text, shape and global and local affine invariants. G. Zul et al [6] presented an approach to logo detection and extraction in document images. They used a multi-scale boosting strategy. At a coarse image scale, a Fisher classifier provides an initial classification. Then, each logo candidate region is further classified at finer image scale by a cascade of simple classifiers. M. Rusinol and J. Llados [5] proposed a method for document categorization by logo spotting. The graphical logo and the query documents are described by a set of local features and matched using a bag-of-words model. In order to filter the matching keypoints and consider only the keypoints belonging to the logo in the query document, they consider clusters of keypoints.

This paper proposes a system for logo spotting inspired from the method in [5] and analyzing the distributions of keypoints on the object (see section 4) to integrate spatial relationships between keypoints in order to refine the matching procedure. This paper is organized as follows. In section 2, we describe the outline of our approach. Then, in section 3, we present the keypoint matching algorithm. In section 4, we show how the spatial distributions of keypoints on the logo are integrated in our model. Finally, we present the experimental results in section 5 and draw the conclusions in the last section.

2. Outline of the proposed approach

Our approach is based on matching pairs of keypoints between the query document images and each logo in the gallery. Then, the query document is assigned to the category corresponding to the logo having the maximum proportion of matching points.

The outline of our method is as follows (see Figure 1). After removing noise from the document image by pre-processing based on morphological operators, the interest points of the query document image and logo images are extracted and described using SIFT (Scale Invariant Feature Transform) descriptor [3]. Then, the keypoints of the query document image are matched towards the keypoints of all the logo images in the gallery (see section 3). The logo in the query document is segmented by clustering the matching keypoints using a density-based clustering algorithm, and the logo category is determined thanks to an accumulating histogram (see section 4).

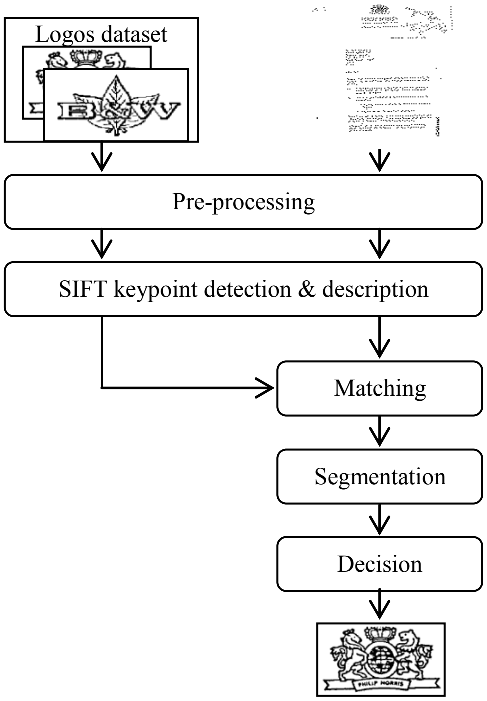


Figure 1: The outline of our approach

3. Extracting and matching keypoints

In our approach, we need to use local features to describe the logo and document images in order to match the query document towards all the possible logos in the gallery. The SIFT descriptor [3] is widely used for describing interest keypoints, because it is invariant towards scaling, rotation and partially invariant to affine transform. First, the interest keypoints are detected by DoG (Difference of Gaussians) filter at different scales, then they are described by the SIFT feature vector. Each SIFT feature vector is characterized by a 4x4 matrix of orientations of the intensity gradient in the sixteen 4x4 windows around the considered keypoint. The orientation being quantized over eight values, the resulting SIFT feature vector is 128-dimensional, $S_i=(s_{i1},..,s_{i128})$. In addition to these 128 features, the x- and y-position of the keypoints are also used for logo spotting (see section 4).

Matching is performed by considering, for each keypoint in a given query document image, its two nearest neighbours within each logo keypoint in the SIFT feature space.

$$d_1 = \min_k(\|S_q - S_k\|) \; and \; k^* = \operatorname{argmin}_k(\|S_q - S_k\|)$$
$$d_2 = \min_{k \neq k^*}(\|S_q - S_k\|)$$

where $||S_q - S_k||$ is the Euclidean distance between two keypoint SIFT descriptors q and k. The ratio r of d_1 over d_2 is then used for matching.

$$r = \frac{d_1}{d_2}$$

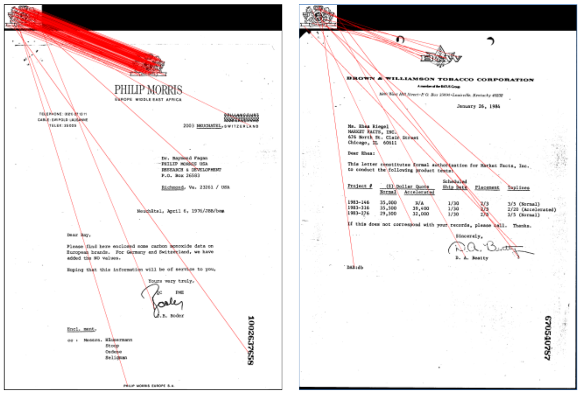
If r is greater than a given threshold t, then it means that the matching is not reliable, as there is a possible ambiguity between the two nearest neighbors. On the other hand, if r is lower than t, then the keypoint is representative enough to be considered. In practice, we settle the value of the threshold t to *0.6*, based on experiments.

After this stage, in [5] they use an accumulating histogram H for counting the number of matching keypoints corresponding to each logo and decide which is the logo in the query document. However, if we limit the strategy to this part, several matched keypoints are located outside the logo region, as shown in Figure 2a. We therefore decide to add an intermediate segmentation step, where the spatial density of the keypoints in the query document is taken into account, in order to filter keypoints and enhance our decision.

4. Logo segmentation and decision

As discussed above, not all the matching keypoints are located inside the logo region. Some of them are located in other regions of the document image. We consider these keypoints as noise and we do not want them to be taken into account in our final decision. As illustrated in Figure 2.(a), there is a high concentration of matched keypoints on the correct logo. On the other hand, the concentration decreases when considering the incorrect logo or when the logo is not in the gallery (see Figure 2(b)).

DBSCAN (Density-Based Spatial Clustering of Application with Noise) algorithm was proposed by Martin Ester *et al.* in [7]. DBSCAN's definition of a cluster is based on the notion of *density reach-ability.* Basically, a point belongs to a group if it is within a certain distance ε from any point of this group. A group forms a cluster if it has more than *MinPts* points, otherwise it is noise. DBSCAN can determine the number of clusters automatically (as opposed to the basic version of k-means for instance), and it can find arbitrarily shaped clusters (while the clusters provided by k-means are circular). Here, we use the Euclidean distance to compare the positions of the keypoints $(x,y)\in R^2$ and cluster them using DBSCAN. In practice, the values of ε and *MinPts* affect the clustering result, and may be settled by experiments. We therefore obtain clusters of matched keypoints as an output of the segmentation stage.



(a) (b)

Figure 2. Red lines show the matching between keypoints of the gallery logo (top left corner) and the query document (bottom) (a) high density of matching keypoints on the correct logo and (b) low density of matching keypoints on the incorrect logo.

Finally, we use an accumulating histogram to count the number of matching keypoints in the biggest cluster (determined by using DBSCAN). The matching logo is finally determined by searching the maximum *m* of the accumulating histogram *H* after normalizing each cell with the number of keypoints in the corresponding logo. If *m* is equal or greater than a threshold *T*, it means that the document image contains this logo; otherwise, it does not contain any logo from the gallery (or no logo at all).

5. Experiments

In our experiments, we use the Tobacco-800 dataset [1] which is a well-known public dataset for document analysis. The Tobacco-800 dataset has 1290 document images including 412 document images containing a logo and 878 document images containing no logo. There are 33 logos in the gallery, and for each logo, the number of document images containing this logo ranges from 1 to 84. We perform two series of experiments: the first one aims at evaluating the performances of our algorithm for logo recognition. The second experiment aims at comparing our approach with other state-of-the-art methods for logos detection.

According to [7], estimating automatically the values of parameters ε and *MinPts* of DBSCAN is very difficult. After experimenting numerous possible values of the DBSCAN parameters ε and *MinPts,* we select $\varepsilon=60$ and *MinPts=5.*

In first experiment, we consider the 15 logos from the database contained in at least 3 document images, coming up in total with 374 document images containing logos and all the 878 document images without any logo.

Table 1. Confusion matrix corresponding to document categorization (the ground-truth is in row)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	**2**														
2		**59**													
3			**49**	1							1				
4			1	**43**											
5	1				**17**										
6						**67**			1				1	2	
7							**17**								
8								**8**							
9			1						**5**		1				
10										**32**					
11											**5**				
12												**9**			
13													**4**		
14														**5**	
15															**3**

As a result of the first experiment*,* Table 1 shows the confusion matrix between the 15 logo classes. Additionally, we have to consider that the false negative rate (cases where a known logo is missed) is 10.42% and the false positive rate (cases where a document with no logo is matched to a gallery logo) is 3.75%. Second, we consider the recognition of each class. The average classification rate is 86.90% However, it should be noticed that some classes such as the fifth and ninth classes achieve the lowest accuracies of 61% and 56%, respectively (see Table 1). This may be explained by the fact that, in both cases, the size of the logo image in the gallery is small and therefore only a little number of keypoints can be considered for matching and classification.

In the second experiment, we compare the precision and accuracy measures of our approach with different state-of-the-art methods for logo detection. We consider all of the documents containing a gallery logo as positive and those containing no logo from the gallery as negative. We compare our approach with those of of G. Zhu and D. Doerman [6], Zhe Li *el al.* [8] and Pham *et al.* [9]. However, our algorithm relies on matching logos in the query document and in the gallery, so we can only use the subset of logos from the Tobacco-800 database which have at least two occurrences (which represents 19 logos, 378 query documents containing logos and all the 878 document images without any logo), while other authors in Table 2 use the whole Tobacco-800 database (which represents 412 query documents containing 33 different logos and all the 878 document images without any logo). In addition, we also evaluate the performance of our approach with and without the density-based clustering process. The comparison results presented in Table 2 show that our approach with the density-based clustering process is more accurate than previous methods, while precision remains very satisfying. However, our method is conceived for logo recognition more than logo detection, so we use a gallery of logos for matching, which is not the case of other methods and should be taken into account when analyzing these comparison results.

Table 2. Comparison of our approach with state-of-the-art methods using Tobacco-800 database. Because of the requirements of our method, we use only a subset of this database (34 documents are used for constituting our logo gallery)

ID	Approaches	Accuracy	Precision
01	G. Zhu and D. Doerman [6]	84.2%	73.5%
02	Zhe Li *el al.* [8]	86.5%	99.4%
03	Pham *el al.* [9]	91%	85%
04	Our appoach without DBSCAN	82.22%	80.72%
05	Our appoach with DBSCAN	94.04%	91.11%

6. Conclusion

In this paper, we present an approach for document categorization based on graphic logo spotting. We use local features for describing the query document images and the logos in the gallery. Interest points are extracted and described using SIFT. Keypoint matching is determined by the ratio of the distances to the two nearest neighbor keypoints in each logo. To filter the query document matching keypoints outside the logo region, we apply segmentation using density-based clustering of the logo keypoints before the final decision, which enhances our categorization results. Finally, the document category is determined based on an accumulating histogram.

We are currently investigating different strategies to better integrate spatial knowledge in the keypoint matching algorithm, in order to enhance the precision of our system and its performances in a multi-scale context.

References

[1] D. Lewis, G. Agam, S. Argamon, O. Frieder, D. Grossman, and J. Heard. *Building a test collection for complex document information processing*. In Proc. 29th Annual Int. ACM SIGIR Conference (SIGIR 2006), pp. 665-666, 2006.

[2] G. Zhu, Y. Zheng, D. Doermann, and S. Jaeger. *Multi-scale Structural Saliency for Signature Detection*. In Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR 2007), pp. 1-8, 2007.

[3] D. Lowe. *Distinctive Image Features from Scale-invariant Keypoints*. International Journal of Computer Vision, 60(2): pp. 91-110, 2004.

[4] D. Doermann, E. Rivlin & I. Weiss. *Logo Recognition Using Geometric Invariants*. International Conference on Document Analysis and Recognition (ICDAR), pp. 894-897, 1993.

[5] M. Rusinol and J. Llados. *Logo Spotting by a Bag-of-words Approach for Document Categorization*. ICDAR, pp. 111-115, 2009.

[6] G. Zhu and D. Doerman. *Automatic Document Logo Detection*. ICDAR, pp. 864–868, 2007.

[7] M. Ester, H.-P. Kriegel, J. Sander and X. Xu. *Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise*. In E. Simoudis, J. Han, & U. M. Fayyad (Eds,), Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining, pp. 226-231, 1996.

[8] Zhe Li, M. Schulte-Austum and M. Neschen. *Fast Logo Detection and Recognition in Document Images*. International Conference on Pattern Recognition (ICPR), pp. 2716-2719, 2010.

[9] T.-A. Pham, M. Delalandre and S. Barrat. *A contour-based method for logo detection*. ICPR, pp. 718-722, 2011