Haar-like-features for query-by-string word spotting

Adam Ghorbel, Jean-Marc Ogier, Nicole Vincent

Haar-like-features for query-by-string word spotting

Adam GHORBEL [a, b], Jean-Marc OGIER [b], Nicole VINCENT [a]
[a] *LIPADE-SIP , Paris Descartes University*
75006, Paris, FRANCE
[b] *L3i, La Rochelle University*
17042, La Rochelle, France
adamghorbel@hotmail.com
nicole.vincent@mi.parisdescartes.fr
jean-marc.ogier@univ-lr.fr

Abstract. This paper addresses the problem of word spotting in handwritten documents. The method is segmentation-free and follows the query-by-string paradigm. In the paper, we focus on the first step of the whole bio-inspired process that is based on two filtering steps, which are a global filtering followed by a more local filtering after a change of observation scale. The contribution of this approach is the use and the generalization of the Haar-Like-Features for the analysis of the document images, inspired from the famous visual perception principle. Different pieces of information are extracted from the whole image before drawing a conclusion, after a process of accumulation of votes. The method is evaluated using the IAM Handwriting Database.

1. Introduction

The automatic study of handwritten documents is a difficult task because of the very high variability of representation of the information. Indeed, the access to the content of these documents is linked to text recognition. The performance of Optical Character Recognition (OCR) engines is still poor, especially for handwriting recognition. One way to recognize the information within the document image is to look for some characteristic of the different words within it. For example a document may contain some significant words (e.g. information, request, and subscription), allowing the classification of the document without deciphering the totality of the document words. OCR does not present a complete solution to the problem because of its limitations in dealing with handwritings. In fact, OCR techniques cannot be accurately achieved because character recognition systems are not well suited for handwritings in an open vocabulary context. For that, word spotting is considered as an alternative to traditional OCR for different applications such as indexing and retrieval in digitized document collections. In the literature, ancient documents are mainly concerned by these word spotting questions, even a few trials on modern writings have been done.

In the literature, word spotting approaches have been applied to various scripts such as Latin, Arabic, Greek, etc. Word spotting approaches have been divided into different categories in multiple ways by document analysis researchers. For instance, they can be divided into two main categories based on matching techniques which are respectively image based matching techniques and feature based matching techniques (J.L Rothfeder, S. Feng, T.M. Rath., 2003). The former includes methods that compute word distances directly on image pixels using the correlation for the query matching. On the other side, the latter compute certain features for word images and then those features are matched. Another classification can be found in (J. Lladós, M. Rusiñol, A. Fornés, D. Fernández and A. Dutta, 2012) where two main approaches of word spotting exist depending on the representation of the query. These two types of approaches are based on Query-by-string (QBS) and on Query-by-example (QBE). The QBS methods (H. Cao and V. Govindaraju., 2007) use character sequences as input. They typically require a large amount of training materials since characters are *a priori* learnt, basically in HMM or NN models, and the model for a query is built at runtime from the models of its constituent characters. In QBE methods (R. Manmatha, C. Han ,E. M. Riseman, 1996) the input is one or several exemplary images of the queried word. This is addressed as an image retrieval problem. Therefore, it does not require any training stage, but collecting one or several examples of the queried word. Another popular categorization technique divides the methods into either segmentation based methods or segmentation-free methods as in (T. Adamek, N.E. O'Connor, A.F. Smeaton, 2007) or in (B. Gatos and I. Pratikakis., 2009).

Based on the literature, we take into consideration the classification presented by the (J. Lladós, M. Rusiñol, A. Fornés, D. Fernández and A. Dutta, 2012) and we integrate it into the classification presented by both (T. Adamek, N.E. O'Connor, A.F. Smeaton, 2007) and (B. Gatos and I. Pratikakis., 2009).

In this paper, the aim is to find some words that are independently chosen from the document content. Particularly, if the processed documents do not contain many words, then a query by example is not possible. For that, we have chosen to express the query by a sequence of characters that is constructed by a keyboard input. This allows using our system in all circumstances, even if the word query is not present in any document images. In short documents, knowledge of word style could take too much time to be known, so the search for a word becomes a challenging problem. Furthermore, the aim is to avoid a training phase on a database to recognize graphemes or other entities in order to implement a word spotting system which is not only dedicated to one type

of documents. The genericity of the system requires designing a system that is capable of adaptation and self-learning. Thus, our method does not rely on a set of characteristics *a priori* fixed but on a family of features among which some will be selected in order to simultaneously fit the search term and the document properties wherein the research is performed.

The remainder of the paper is as following. In section 2, we introduce the family of operators applied in our work. Then in section 3, the proposed approach is detailed. Finally, section 4 is dedicated to the experimental results achieved with IAM database.

2. Generalized Haar-like-features

Looking at a word, it is recognized that human perception first considers to global view of the shape, before focusing on detailed parts. This depends on the observation scale. For example, the word "adam" that may be modeled by a global appearance looking like the pattern presented in Figure 1(f).

A word is characterized by intrinsic and exogenous characteristics. For instance, the number of letters, the position and presence of ascenders and descenders characterize the shape of each word. Besides, the size (width and height) of a character, which plays a major part in the construction of the different patterns, depends on the writing style in the documents to be processed. It is estimated by a size optimization using the Haar like feature of figure 1(l).

In a first step, we try to characterize the different patterns that may occur. This pattern characterization step is considered as an essential step of our approach because the final results depend on it. According to words in the document images, we can model a global view of each word by several patterns appearing simultaneously. Some of these patterns are illustrated in Figure 1. For example, Figure 1(g) represents a double ascender presence following by lowercase characters. Figure 1(b) can help in finding words similar to "adam".

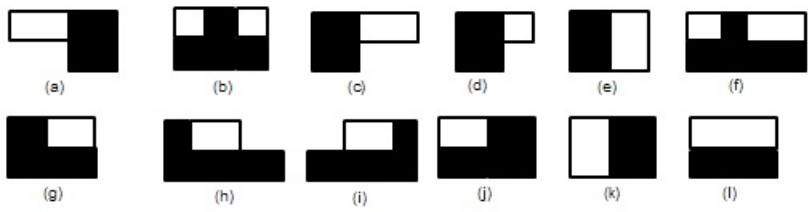

Figure 1. A few number of patterns applied in our work.

This approach is quite similar to Viola & Jones (P. Viola, M. Jones, 2001) approach that starts from a family of patterns and uses Adaboost to select some of them. It is more easy to use than Haar wavelet coefficients used in different document applications such as document text extraction (S. Audithan, 2009) or script identification (P. S. Hiremath, S. Shivashankar, 2008). In our case, the selection of the right patterns has to be done according to the word in order to retrieve it and according to the characteristics of the documents. Yet, we have generalized the patterns used in the Haar approach to fit the global shape of words written in Latin alphabet. The number of patterns and their shapes will define the query word and discriminate it from the other words. Furthermore, the more patterns applied, the better results are.

The patterns can be searched in the document using a convolution product between the image and a kernel containing 1, -1 or 0 values. The computation complexity is not too high when integral image is used. The operator enabling to detect a pattern P that is applied on the whole image will give a new image I_P, where the presence of the pattern is characterized by a high grey level. The filter can be applied in a blind way on the whole document and will automatically select the text lines where the answer to the operator will be high. At a lower level, similar patterns may be used to distinguish between an 'o' and a 'c', making evident the concavities.

This tool is the core of the approach we are proposing.

3. Proposed Approach

To process documents that present (i) a wide variability of style, (ii) ancient or modern with fragmented characters due to the non-homogeneity of the ink, (iii) crossing of lines, (iv) variability in writing style and (v) an above overlap of components such as components belonging to several lines of the text because of the presence of ascenders and descenders, we propose to consider the following major constraints:

- No layout segmentation: the query is directly compared to the whole document image components as it is very difficult to perform an accurate line, word or even character segmentation.
- No binarisation is required, which permits to avoid losing data in the pre-processing of document images.

Our approach globally looks at document images without any use of a word segmentation step, which is often assumed in current methods of Word Spotting. We introduce in our proposed approach two major phases:

- In the first one, at document level, in a global way, the search space is reduced to Zone of Interests (ZOI's) which are considered to contain Candidate Words (CW's).

- In the second one, a refining step enables to retain only the very similar CW's to the query.

The implementation of these two major phases relies on the same process, which can be considered as a filtering step applied sequentially at two different scales. These two bio-inspired steps (simulating the famous focusing perception principle) are based on the application of the Haar-like-features defined in previous section.

3.1 Document level

At document level, some zones of interest are selected; they are defined as the accumulation of specific patterns, in a limited and relative space. Each pattern, then each filter can be considered as a Viewpoint. The presence of these patterns is detected in the image I_P associated with the corresponding operator. The presence of the pattern is linked to a binarisation of I_P , where I is the original image in grey level. In fact as several patterns are associated with a word, several I_P images are available. The patterns associated with a word are in a limited area but their simultaneous presence is highlighted if the I_P images are translated according to the properties of the query word. Fusion of the filtered images is achieved by accumulation of the translated images. A binarisation process gives some hints at the position of words with same shape as query word. The position and area of patterns lead to the definition of the CWs. This Global analysis step represents the major and original step of our approach. It helps to limit the number of CW's and to obtain the CW's, which are the most similar to the query.

3.2 Word level

The overall filtering presented in previous subsection results in a large number of CW's that may correspond to the global shape of the query. This number is greater than the real number of the occurrences of the query in the document. In order to reduce the number of CW's, we introduce a second phase that is based on a refining filtering. Thus, we are going to change the observation scale. We concentrate our work at a lower scale, the word scale, rather than the previously used document scale. The candidates have approximately the same size (number of characters) as the query. However, we apply other new FW's on the selected ZOI's but handled by gradually changing the observation scale. This step aims at refining the results and improving the accuracy of the results. Finally, we obtain only the query occurrences existing in the processed document images.

The flowchart of our word spotting approach is highlighted in Figure 2.

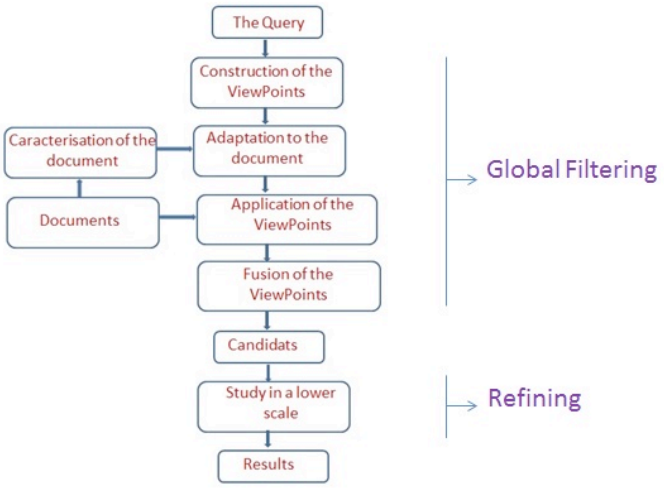

Figure 2. The flowchart of the proposed approach.

4. Experimental Results

Our approach has been evaluated on the IAM handwriting database consisting of 1539 pages written by 657 writers (U. Marti and H. Bunke., 1999) (U. Marti and H. Bunke., 2002). In our experiments, we have not worked on the isolated words but on the document images themselves. So, we work at document level. We randomly selected some document images and worked only on the handwritten texts. The performance is measured by using Precision and Recall criteria. Precision P is the percentage of the retrieved words that are relevant to users. Besides, Recall R is the percentage of the words that are same as the query and are successfully retrieved from the IAM Handwriting database.

$$R = \frac{TotalSameWords\,\mathrm{Re}trieved}{TotalSameWordsExisting} \times 100 \quad (1)$$

$$P = \frac{TotalSameWords\,\mathrm{Re}trieved}{\left(SameWords\,\mathrm{Re}trieved + FalsePositives\right)} \times 100 \quad (2)$$

The evaluation of our work is based on Quantitative and Qualitative studies. We illustrate the results on three scripts written with three different styles extracted from the IAM database and shown in Figure 3. Here, the query word is "the", a rather short, so difficult word to be spotted. We notice that the occurrences of the query have various lengths and widths in tested documents.

From our experiments, our approach is capable of finding all the instances of the queries in the tested document images. Thus, in most cases, the recall of our system is 100%. This extraction step looses very few positive answers. Furthermore, we notice that our approach detects other CWs for the query word that have almost the same shape or begin with letters having the same shape as the query. The precision strongly depends on the writing style. In Figure 3(c) the precision is 10% , in figure 3(a) it is about48%, as a whole from 40 pages we have 21%. These low precisions can be improved by increasing the number of applied patterns.

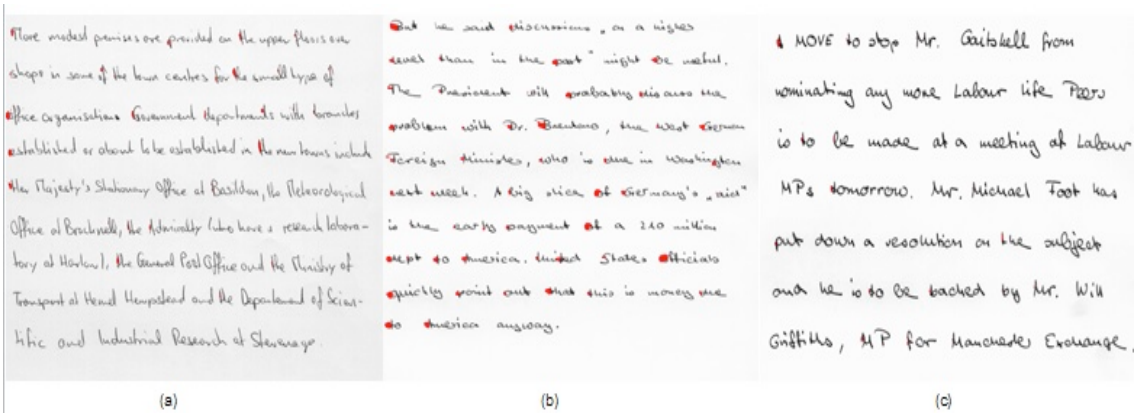

Figure 3. Obtained results of some tested images.

5. Conclusion

In this paper, we have presented a word spotting method that does not rely on any previous segmentation step. This approach can be used in heterogeneous collections containing both handwritten and typewritten documents. We proposed new generalized Haar-Like filters and apply them to word modelling and spotting. The presented work is applied on the IAM Handwriting Database. In future works, we will focus on proposing a refining filtering phase in order to increase the accuracy of our word spotting approach.

References

B. Gatos and I. Pratikakis. (2009). Segmentation-free Word Spotting in Historical Printed Documents. *In Proc. of the 10th Int. Conf. on Document Analysis and Recognition*.

H. Cao and V. Govindaraju. (2007). Template-free Word Spotting in LowQuality Manuscripts. *In 6th Int'l Conf. on Advances in Pattern Recognition.*

J. Lladós, M. Rusiñol, A. Fornés, D. Fernández and A. Dutta. (2012). On the influence of word representation for handwritten word spotting in historical documents. *International Journal of Pattern Recognition and Artificial Intelligence*.

J.L Rothfeder, S. Feng, T.M. Rath. (2003). Using corner Feature Correspondences to Rank Word Images by similarity. *Conference on Computer Vision and Pattern Recognition Workshop*, pp. 30–35.

P. S. Hiremath, S. Shivashankar. (2008). Wavelet based co-occurrence histogram features for texture classification with an application to script identification in a document image. *Pattern Recognition Letters 29(9)*, 1182–1189.

R. Manmatha, C. Han ,E. M. Riseman. (1996). Word Spotting: A New Approach to Indexing Handwriting. *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 631–637.

S. Audithan. (2009). Document Text Extraction from Document Images Using Haar Discrete Wavelet Transform. *European Journal of Scientific Research ISSN 1450-216X*, *36*, 502–512.

T. Adamek, N.E. O'Connor, A.F. Smeaton. (2007). Word Matching Using Single Closed Contours for Indexing Handwritten Historical Documents. *International Journal on Document Analysis and Recognition*.

U. Marti and H. Bunke. (1999). A full English sentence database for off-line handwriting recognition. *In Proc. of the 5th Int. Conf. on Document Analysis and Recognition*, 705–708.

U. Marti and H. Bunke. (2002). The IAM-database: An English Sentence Database for Off-line Handwriting Recognition. *Int. Journal on Document Analysis and Recognition*, *5*, 39–46.