IJDAR (2008) 11:9–18
DOI 10.1007/s10032-008-0064-6

ORIGINAL PAPER

Document image characterization using a multiresolution analysis of the texture: application to old documents

Nicholas Journet · Jean-Yves Ramel · Rémy Mullot · Véronique Eglin

Received: 5 April 2007 / Revised: 4 April 2008 / Accepted: 30 April 2008 / Published online: 24 June 2008

Abstract In this article, we propose a method of characterization of images of old documents based on a texture approach. This characterization is carried out with the help of a multi-resolution study of the textures contained in the images of the document. Thus, by extracting five features linked to the frequencies and to the orientations in the different areas of a page, it is possible to extract and compare elements of high semantic level without expressing any hypothesis about the physical or logical structure of the analyzed documents. Experimentation based on segmentation, data analysis and document image retrieval tools demonstrate the performance of our propositions and the advances that they represent in terms of characterization of content of a deeply heterogeneous corpus.

Keywords Document image analysis · Texture features · Multiresolution · Digital libraries · Indexing

N. Journet (✉) · J.-Y. Ramel
LI, 64 Avenue Jean Portalis, 37200 Tours, France
e-mail: njournet@univ-tours.fr

J.-Y. Ramel
e-mail: jean-yves.rame@univ-tours.fr

R. Mullot
L3I, 17042 La Rochelle Cedex 1, France
e-mail: rmullot@univ-lr.fr

V. Eglin
LIRIS INSA de Lyon, Villeurbanne Cedex, France
e-mail: veronique.eglin@insa-lyon.fr

1 Introduction

The massive digitization of thousands of pages of documents, requires the implementation of computer resources allowing fast and pertinent access to the piece of information among all these pages. Several years of scientific work on the analysis of images of contemporary documents, have made it possible to generate efficient tools, allowing several forms of indexing by analysing the content (OCR, retro-conversion programs, etc.). However, the new framework representing the analysis of images of old documents does not allow the simple transposition of these tools initially designed for contemporary documents. The explanation is mainly in the very nature of the corpus of the treated images to be analyzed (cf. Fig. 4 for old document images).

This article addresses the characterization of the content of old documents considered as an alternative to the methods that so far have been mainly based on the segmentation of pages and on the interpretation of their structure. The original idea of this work is to use a multiresolution analysis of the textures in the images in order to avoid any assumption about the structure or content of the documents. This paper is based on three points. First, a study of the methods of document image characterization is proposed. Then detail of our proposal for characterizing the content of old documents is given. With the help of the calculation of texture features at different resolutions, we show that it is possible to characterize the content of images without expressing any hypothesis, neither about the structure nor about the characteristics of the treated images. In parts 4 and 5, the pertinence of our proposition with segmentation experiments and two image indexing tools based on an analysis of the texture feature is shown.

2 Methods of characterization of document image content

In the context of a collaboration with the Centre d'Etudes Supérieures de la Renaissance de Tours,[1] we accessed more than 100 pieces of work dating from the fifteenth and sixteenth century. One of the major characteristics of these old documents is the heterogeneity of the available material. The rudimentary character of the techniques and equipment used, the deterioration of the documents and the variety of the editorial rules are some of the reasons that explain the large range of such old documents. The images of the documents we accessed, encompass three centuries of printing and history. The complicated layout (several columns of irregular sizes), the use of specific fonts, (no longer in use today), the frequent use of embellishments, (illuminations, drop caps, frames), the small line spaces, or also the presence of non constant spacing between the characters and the words, the superimposing layers of pieces of information (noise, handwritten notes), are just so many specificities of old documents which make it difficult to characterize their content.

Should a digitized library let access to the information inside a piece of work (research on the text, looking for images, etc.) a manual indexing process consisting of associated keywords with every page (or part of page), cannot be carried out. The amount of work involved in keyboarding the keywords is a clear impediment. Research into automatic indexing by means of analysis of the content was then implemented so as to be able to index quickly and precisely a large number of images of old documents. The following section proposes a short survey and an evaluation of document imaging methods initially designed for contemporary document indexing.

2.1 Analysis of image documents: various approaches using a strong a priori

In most cases, the methods of structure analysis are based on the supposed repetitiveness of the structure of the documents in a corpus. They are also based on the study of "physical" characteristics through, for example, techniques of grouping or fusion of black pixels, of cut-out of images areas, of study of alignment of white pixels, etc. These methods of analysis of documents are mainly aimed at segmenting into blocks the different elements of the page content. The most known state of the art [9,21], break down these methods depending whether they are driven by the data or by the model. The approaches driven by the data are characterized by analysis which converges on an emergence of homogeneous blocks or areas.

The objective of old document indexing, requires one to be able to break down a page into homogeneous elements in order to classify them and make possible to index. It means, in most cases, to be able to locate the characters and the blocks of pictures. These methods often rely on the study of the separation between pixels. For example the Run Length Smearing Algorithm (RLSA) of [31] analyses the spaces between black pixels in order to merge characters into lines and paragraphs, while white space algorithms study [2,25] the white pixels area. Methods driven by the model are widely used in the work of segmentation of the documents, the cut-out algorithm in XY (XY-CUT) is one of them. Proposed by [23] in the 1980s, its principle is based on a cut-out. It consists of implementing the same algorithm on an area recursively, and this, until a condition about the separation of the objects (into paragraphs or lines), has been satisfied. Multi-resolution approaches for the segmentation of documents are also often used. In their respective articles, the authors of [8,28] use a pyramid with several levels of resolution to allow the recognition of the physical structure of the document they treat. For every image resolution, the authors extract different pieces of information linked to the collateral elements (size, position, etc.). An analysis of the obtained data to these different resolutions make it possible to define the labels of the elements and the way to merge them in order to obtain a text/graphics segmentation. After comparing several methods of segmentation (XY-CUT, RLSA, search for white space), the conclusions drawn by [27] are similar to ours. Thus, the XY-CUT and RLSA are sensitive to noise and not very resistant to skewed texts. The search algorithms for white space include complex stipulations which must be defined. From a general point of view, one may ponder over the variations of the sizes and font types of a corpus. In this context, it would be too complex to edit classification rules or to do a training given the variability of the content from one work to another. These tools are thus as well adapted to documents with formatted structures as to the contemporary documents.

2.2 Document image analysis: the texture approach

2.2.1 Spatial texture-based methods for document image segmentation

True alternatives to the methods described above, some tools for the image analysis process allow the extraction of information without any necessary know-how relative to the context, to the semantic, or to the physical characteristics of the studied image. Among the great classics of the statistical methods, it is impossible not to mention the works of Haralick and Laws. The gray level co-occurrence matrix (GLCM) was proposed by Haralick in [14]. The GLCM is a matrix which, in an image, shows the number of appearances of pairs of pixels with grayscale (i, j) according to a given

[1] http://www.bvh.univ-tours.fr.

direction. Attributes calculated with GLCM allow the characterization of regularity, repetitiveness and texture contrast. Another method of texture characterization based on the calculation of characteristics is that of [18]. This method consists of applying spatial convolution with pre-determined filters. Every convolution aims at a precise characteristic of the texture (presence of horizontal, vertical lines, etc.). The auto-correlation function which helps to get pieces of information about the characteristics of a texture can also be mentioned. So much so that, if the texture is “rough” (wide patterns), the function will slowly go down when the analysis distance goes up. On the contrary, if the texture is finer (small and little spaced out patterns), then the function goes down quickly [30].

According to [13], the approaches of statistic nature have the advantage of being simple to implement and their efficiency is no longer to be proved. However, it can be noted that these tools based on the statistical study of the gray levels, do not seem to be very appropriate to the old documents images. Even if most historical documents are digitized in grayscale or color mode, techniques of printing of the Renaissance give document mainly composed of black ink over white paper. Furthermore, the dataset we use did not containing color images. The only variations of gray levels are due to the digitization or to the deterioration of paper or ink. One is then very far from the grayscale variations that are to be found in natural images. This is why the Haralick and Laws attributes do not seem very appropriate to the segmentation or the characterization of old document images. The geometrical methods correspond to a characterization of the shapes and of their spatial relations making up a texture. In [29], the authors show that it is possible to segment textures with the help of geometrical moments. It is also possible to use a geometrical method to segment documents. Thus, the authors of [17] propose a method of separation text/graphics of Hebrew documents based on the building of horizontal histograms. In [7], the author analyses previously cut-out blocks in order to classify them, either as drawings or as text. The extracted texture standards come from an analysis of the results of the grayscale pixels according to different angles. The Markov Fields and the fractals are the two most commonly used tools of the texture algorithms based on models. The fractal dimension is used to measure texture roughness and repetitiveness of a pattern. In [5], the authors use the Power Law (Law of Zipf) to identify areas within a natural image. In [24], the authors use the HMM to segment images of handwritten documents into areas of labeled interest areas (text lines, scratches, note in the margin, etc.).

2.2.2 Frequency-based methods

The standard methods based on the use of primitive functions coming from signal processing methods are ideal to make the texture characterization possible. Actually, these tools allow the detection of frequencies and orientation characteristics. These tools work in the frequency domain. The Fourier or Gabor Transforms or the wavelets are widely used in the works about the indexing and the segmentation of natural images. In [20], the authors use the Gabor filters. The authors calculate the results of the Gabor filters at different resolutions. After every transformation, the average and the standard deviation of the calculated coefficients are extracted. The filter bank is made up of four resolutions and six orientations. The measurement of dis-similarity between two vectors is the total sum of the difference term to term of the averages and standard deviations of the vector. In [19] the authors present a supervised multi-class classifier based on Gabor filters that is used to classify the scripts, font-faces, and font-styles (bold, italic, normal, etc.) in an application where the classes are known. The method was applied to a variety of bilingual dictionaries to identify different scripts, and to classify Roman scripts into bold, italic and normal font-styles.

2.2.3 Conclusion

These methods generally rely on the fact that the text areas, due to the great number of ink/paper transitions, are characterized by high frequencies, while the images are generally made up of more extended homogeneous areas and then associated with low frequencies. The authors of [6,12], use Gabor filters or wavelets to segment their document images. In our opinion, the main advantage of the use of texture tools is in the greatest generalization ability that these tools can give. A significant number of a priori knowledge used by the method exclusively driven by the data or the model, become then useless. Among the other advantages, it can be said that in the most cases, these tools work on grayscale images. As a result binarisation is not systematically necessary. It must be noted that if these texture tools allow us to characterize the image content, they do not let us segment into blocks (paragraphs, graphics, titles, etc.). This objective has to be carried out only at the end of post-processing.

3 Characterization of the images

We propose a process based on an analysis of the textures within the images, without looking for or taking into account the a priori knowledge of the structure of the pages. Given that the use of the tool is directed to non-specialists in image processing, our method must not include any threshold, or models, or explicit structures of the analysis process. The global process consists then in characterizing precisely the contents of the pages of a piece of work, and this with the help of new extraction algorithms of texture features devoted to the analysis of documents. This characterization of the pixels is the basis of our global image analysis process. The texture features are calculated on a local level at different resolutions. With the help of a sliding window (whose size is the only parameter of our method), it is possible to associate these extracted texture attributes to each pixel of the image. This analysis is carried out at 4 different resolutions, giving finally 20 numeric values for each pixel. Once all the pages of a work have been analyzed, the extracted meta data are stored in a database.

3.1 Characterization of the orientations

The orientation is one of the main visual characteristics involved in the pre-attentive view. We were interested in this characterization in order to propose three texture features linked to the orientation information. We chose to use a non-parametric tool based on the auto-correlation function: the rose of directions (proposed by [4]). In [11] the authors propose a methodology of printed text characterization for document labeling using, among others, the autocorrelation function.

The definition of the auto-correlation function for a two-dimensional signal is defined by

$$C(k,l) = \sum_{k'=-\infty}^{+\infty} \sum_{l'=-\infty}^{+\infty} x(k',l') \cdot x(k'+k, l'+l).$$

The rose of directions is a polar diagram based on the study of the answer of the auto-correlation function when it is applied to an image. In his work, Eglin [10] defines the auto-correlation function as being the gathering of the whole values that one can get by doing the sum of all the products of the grayscale of the points in correspondence after translation of an image compared with itself. So, a point $C(k,l)$ of the auto-correlation function contains the value of the sum of the products of the grayscale $x(k',l')$ of the points in correspondence after a translation of vector (k,l). These different translations allow the inspection of an image according to its different directions. On the auto-correlation function, the data related to a same direction will be located on a same straight line, which also has direction and going through the origin. Figure 1b gives examples of an auto-correlation calculation applied to a square composed of a diagonal line. On these simple shapes, it is easily seen that this function allows the identification of the main orientation. The translation of the two horizontal straights lines in their own direction will lead to a high level of correspondence, which results in an important value of the auto-correlation function in the horizontal direction. It will be the same for the vertical and the diagonal lines. It will not be the case if the direction of the calculation is in another direction.

The rose of directions is a diagram allowing the analysis of the result of the auto-correlation function. Let (k,l) be the central point of the image after the computation of an auto-correlation and the line D_{origin} be the axis going through this point. Let θ_i be the studied orientation, one calculates then the straight line D_i so that any couple of points (a,b) respects the following relation: angle formed by the straight line (a,b) and going through $D_{\text{origin}} = \theta_i$. At least, for every orientation θ_i, one calculates the sum of the different values of the auto-correlation function $R(\theta_i) = \sum_{D_i} C(a,b)$, with $\arctan(b/a) = \theta_i$.

These values are later standardized in order to keep only one aspect relative to the contribution of every orientation : $R'(\theta_i) = \frac{R(\theta_i)-R_{\min}}{R_{\max}-R_{\min}}$ with $R_{\max} \neq R_{\min}$. Figure 1 sums up the different steps in the construction of the directional rose of an image composed with a square and a diagonal line. Figure 1c is rose of directions associated with the Fig. 1a. In this rose two information are important. First, it indicates that only horizontal, vertical and diagonal orientations are present in Fig. 1a. Then it also indicates that the main orientation are the horizontal and the vertical (the diagonal line does not touch the border of the rose). This two information are only available on the rose of orientation.

The Hough transform could give similar information to the rose of directions. Both are giving information about the pixel orientation of an image. We decided to use the rose of direction because, unlike the Hough transform, the autocorrelation computation relies on the frequency decomposition of the analyzed image with a Fourier transform (FFT). The Plancherel theorem is at the basis of this simplification. In practice, the autocorrelation function is efficiently computed with the image spectrum and not directly with the previous mathematical definition. In [15], we give more detail about the way it is computed and how this choice gives robust information, even if it is computed on noisy document images. We believe that, the autocorrelation function is the most interesting method to extract orientation information about old document images.

We will now study the behavior of the rose of directions on images of documents. As shown in Fig. 2, there is no precise pattern of rose. As regards for a rose computed on a text area,

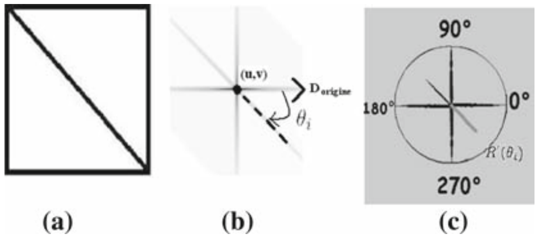

Fig. 1 Construction of a directional rose. **a** Image, **b** autocorrelated image, **c** rose of directions of the original image

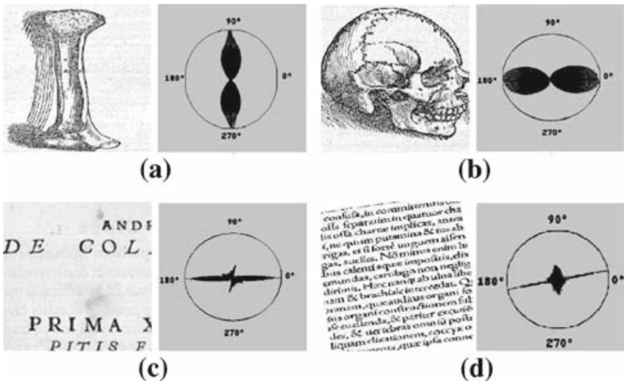

Fig. 2 Example of roses of directions computed on different images. **a** Vertical drawing, **b** skull, **c** four text lines, **d** skewed text

the shape of the rose depends on the number of lines, on the size of the characters, on the orientation of the text (Fig. 2c, d). Regarding the drawings, the same remark can be made. The great variety of the existing images does not allow the definition of an homogeneous model of roses (Fig. 2a, b). Nevertheless, the calculation of the rose makes it possible to extract the features which are very rich in information. We will see further that the main orientation, the shape and the intensity of the rose are characteristics and provide a very fine characterization of the content.

In document image analysis, the multi-resolution allows us to perceive structures of different sizes. Figure 3a shows the importance of a calculation of the rose on a text area at different resolutions of the image. Application of three different sizes of windows, will generate three different shape of roses. Nevertheless it is to be noted that the information concerning horizontality does not vary. Figure 3b shows the same principle, when the rose is calculated on a picture this time. Unlike calculation on text, the rose presents high variations if the window is of a different size. In the two examples, the size of the window (in pixels) are 32 × 32, 64 × 64 and 128 × 128, respectively. The two image sizes are 600 × 889.

We have then defined pertinent characteristics to describe the content of the document made from this tool. We have then decided to extract three features which permits the characterization of texture information relative to the orientations possible. The first extracted signature is the angle matching the main orientation of the rose of directions: Feature1^k $(i, j) = |180 - \text{ArgMax}(R'_{i,j})|$. So as not to have to manipulate circular data, this angle is standardized according to the deviation from the horizontal angle. So, at resolution k, for every pixel (i, j) of an image the texture attribute 1 is calculated. $R_{i,j}$ stands for computation of the rose centered on the pixel (i, j). The isotropy of the image is estimated according to the intensity of the auto-correlation function. So, at the main orientation found by the first equation, every pixel (i, j) will be characterized with the help of the following equation: Feature$2^k(i, j) = R(\text{ArgMax}(R'_{i,j}))$.

The last texture feature linked to the orientation, characterizes the global form of the rose. To do so, the standard deviation of the intensities of the rose is calculated, except for the orientation of maximal intensity : Feature$3^k(i, j) = \text{Std}_{\theta \in [0,\pi]}(R'_{i,j})$ with $\theta \neq \text{ArgMax}(R_{i,j})$. If the standard deviation is high, it means that the rose is deformed and that a large number of orientations are present in various proportions.

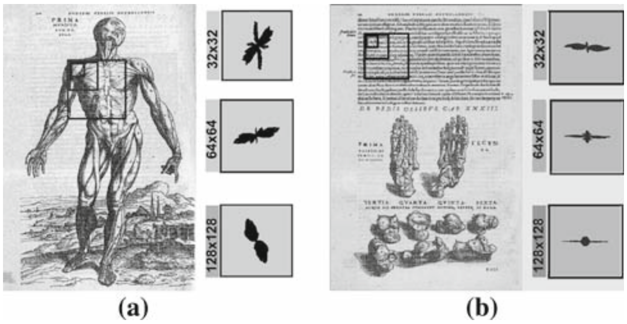

Fig. 3 Importance of a calculation at different resolutions. **a** Multi-resolution on text, **b** multi-resolution on image

3.2 Characterization of frequencies

In document image processing, the notion of "frequencies" is linked to the transitions between paper and ink. In order to characterize the frequency information, we chose to draw our inspiration from the work [1,26]. These authors detail how it is possible to characterize different styles of text or to separate the text from the images, by studying the properties of the grayscale pixel transitions.

The first feature allows us to characterize the transition between ink and paper. For every line of the analyzed area by means of the sliding window, the sum of the difference between a grayscale pixel and its neighbor on its left is done. Feature$4^k(i; j) = \text{Avg}_{i \in I'}(\sum_{j \in J'}(p_{ij} - p_{ij+1}))$ with I' and J' the size of the analysis window and p_{ij} the gray level of the pixel of coordinates (i, j), makes the calculation of a signature about the transitions in the studied area possible.

The last calculated texture feature is based on a characterization algorithm of the white spaces separating the collateral elements. Thus we look for a means to get pieces of information on the extent of the various background areas of the pages. A recursive approach was adopted. It consists of calculating four iterations of a recursive XY-CUT algorithm. To every iteration, the iteration which has just been analyzed is cut into four areas of identical size. For each of them the feature of the following equation Feature$5^k = \frac{\sum_{l \in J'} p^k_{il} + \sum_{h \in I'} p^k_{ih}}{2}$ with I' and J' the size of the window of analysis to the iteration k of the recursive algorithm. This feature is equal to the average of the sum of the grayscale per column and per line for every pixel.

The five texture features have been created without using any priori knowledge. The association of several texture characteristics to extract the different kinds information present in old document images is an original idea. In the next section different experiments and a discussion will be detailed in order to evaluate the quality of these five texture features.

A C-ANSI source code of the rose of direction and the five features are available,[2] thus readers will be able to implement the process we presented here.

4 Discussion about our multi-resolution texture extraction

4.1 Evaluation through pixel clustering

In this section, the quality of the categorization of the content is estimated through a clustering of the pixels on the basis of the 20 proposed texture features. Classifying the elements of content of the works make it possible to determine the objective of separation of the pixels into layers, when it is done on a complete work. Every image pixel has 20 values coming from the 5 texture features calculated at 4 different resolutions. Our objective is to group together the pixels corresponding to homogeneous areas, which amounts to grouping together characteristic vectors. It is a problem of non-supervised clustering. We used a clustering algorithm of mobile center kind, where only the number of clusters that one wishes to get is indicated. This clustering algorithm named Clustering LARge Applications (CLARA) is defined in [16]. CLARA performs for large scale data-bases by using a k-medoid algorithm (a medoid is a point such as the sum of the dis-similarity within the cluster is minimized). This algorithm is in $O(kn)$ where k is the number of clusters and n the number of points. Figure 4 shows the kind of results that one gets when clustering a complete work is done. Clustering on a complete work means that only one clustering operation is computed, all six pages are clustered together. Then if two pixels of two different images have the same color it means that they belong to the same cluster.

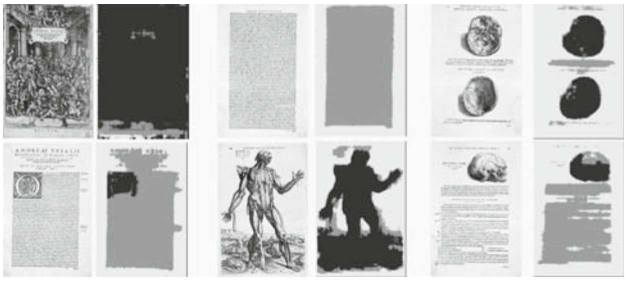

Fig. 4 Pixels Clustering of a work. Each cluster is symbolized by a *color*

These tests demonstrate the actual coherent separating power of the extracted signs. As regards the main limits of the proposed marking, they become localized on the level of the analysis of transition areas between texts and pictures. Because of that, a great part of the titles (isolated from the body of the text) are identified as being drawings. In order to measure the pertinence of the method, we propose a simple estimation of the abilities of our data, instead of giving a great quantity of visual results, we have then decided to estimate the ability of separation of the pixels into three classes: text/drawing/background. To do that, we have manually captured a ground truth with the help of an application that we developed and which allows the delimitation (with the mouse) of the outlines of the drawing areas and of the text areas (ground truth). A file has been created in order to store this ground truth so that they can be finally compared with the calculated classification. Our tests were done on 400 pages of old documents, extracted from 9 different works. Given, that we wanted to have an idea of the pertinence of the extracted signs, we have made up a corpus of test images with contents as varied as possible. The used texture features allow a good separation of the information layers of the documents. The rates of good classification is 83% for the drawings pixels and 92% for text pixels.

4.2 Comparative analysis with a Gabor filters approach

In order to assess the pertinence of our texture features, we have also compared our classification results with those obtained after categorizing of the content with the Gabor filters. These filters allow the characterization of the frequencies and orientations of the textures present in the images. We were inspired by the algorithms present in [3,6]. The document image segmentation is carried out by detecting the areas having specific characteristics of orientations and frequencies. The filter bank was tested with about 20 contemporary documents and about 20 old ones. These characteristics, combined with the recommendations found in the articles of reference, brought us to build a filter made up of five orientations $\theta_l = \{0°, 30°, 60°, 90°, 120°\}$ and of six frequencies $f_i = \{1, 2\sqrt{2}, 4, 32\sqrt{2}, 64\sqrt{2}, 128\sqrt{2}\}$. After application of the filters banks, every pixel is described with 30 characteristics. We then submitted these data to the same classification algorithm used in the previous tests. Figure 5a shows the results obtained with a contemporary document. They meet our expectations more particularly regarding a good detection of the text whatever its orientation.

We have applied bank of filters on to our images. Figure 5b shows the results of classification summing up the quality of

[2] http://njournet.free.fr.

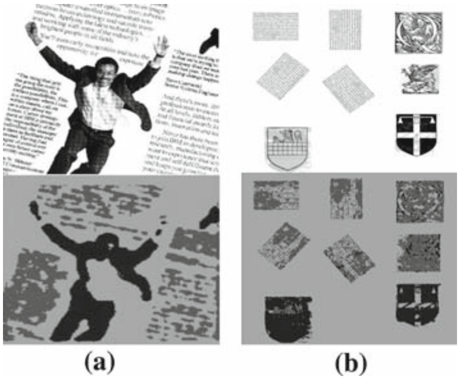

Fig. 5 Document image segmentation of different kind of documents using Gabor filters. Text (*gray*) and drawing (*black*)

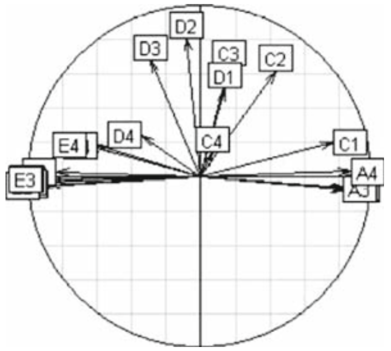

Fig. 6 Correlation between variables after PCA on the generated data

the obtained results. The recurrent problem is linked to the detection of drawings. Actually, if the detection of text areas (multi-directed) is not a problem, it is on the drawing level that the mistakes of classification are visible. It happens that the drawings are made up of a mass of small segments more or less close to each other according to the effect desired by the designer. This physical characteristic results in a large quantity of transitions between ink and background, which means high frequencies. As it seen in Fig. 5b, the two coats of arms (made up of few transitions) are on the whole well recognized as drawings while the two others drop caps are likened to text.

4.3 Importance of the multi-resolution texture extraction

Due to the large quantity of generated data, an analysis of this data is a compulsory step. A factorial analysis (PCA) of the data shows interesting pieces of information. The first is that the eigen values carried by the first four axis is always very good. It is then possible in all cases, to reduce our data from a dimension 20 to a dimension 4 so as to keep about 78% of the information. The second piece of information emerging from the analysis is that the features linked to the transitions ink/paper (fourth feature) and to the intensity of the correlation for the main orientation (second feature) are strong correlated. This is understandable, since the text is always horizontal and that it is always very present in our images. The feature linked to the transition is very sensitive to the strong black/white transition corresponding to the text lines. It happens that the feature linked to the intensity of the auto-correlation is also sensitive to the preferential orientations (here horizontal). It is then quite possible to calculate only four of the five proposed features in this article. Finally, the last important result is that it has not been possible to find stable combinations between variables. A linear combination is directly linked to the intrinsic composition of the page. Figure 6 shows, among others, the existing relations between the variables when they are projected on the first vectorial plane. As a clarity precaution, the projected features are symbolized by a letter [A: white spaces (fifth feature); B: ink/paper transitions (fourth feature); C: std of the rose (third feature); D: main orientation (first feature); E: value of the auto-correlation function second feature], every number stands for the resolution at which the feature was calculated.

Due to the large size of the images, the describers files (after calculation of the features) are thick. Because of that every operation on these data files is time consuming. The strong spatial correlation of the pixels, allows us to suppose that a sample can be done without damaging the analysis quality. We verify this hypothesis while comparing results of a PCA on sampled or non-sampled data. The pixels are chosen at random. The whole realizable tests showed the retained variance under projection is maximal (78% with 4 axis). The question is then to know whether if the fact of taking a sample of pixels will denature the relations between the variables. On the trials carried out on an image of dimensions 600 × 889, one must go down below 1% of the points to have correlation circles which differ. These tests show that a data selection has little impact on the calculation of the PCA. Because of this, one may suppose that every analysis treatment (hierarchical classification, PCA, etc.), which is difficult to carry out on such data sizes, will be quite feasible and exploitable on a data sample. Figure 7 shows the importance of a multiresolution features extraction. Experiments done with image Fig. 7a (image size: 370 × 548) shows that the classification results are better when the whole multiresolution features are used (Fig. 7b), than when only one resolution features are used (Fig. 7c–f). The same kind of experiments have been carried out with different image size (600×889 and $1,232 \times 2,257$). These experiments ended with the same conclusions.

Very few previous works on document images analysis propose evaluation method considering different resolutions for the images. The above study seems for us an other important contribution of our work because it allows to obtain a more generic analysis system.

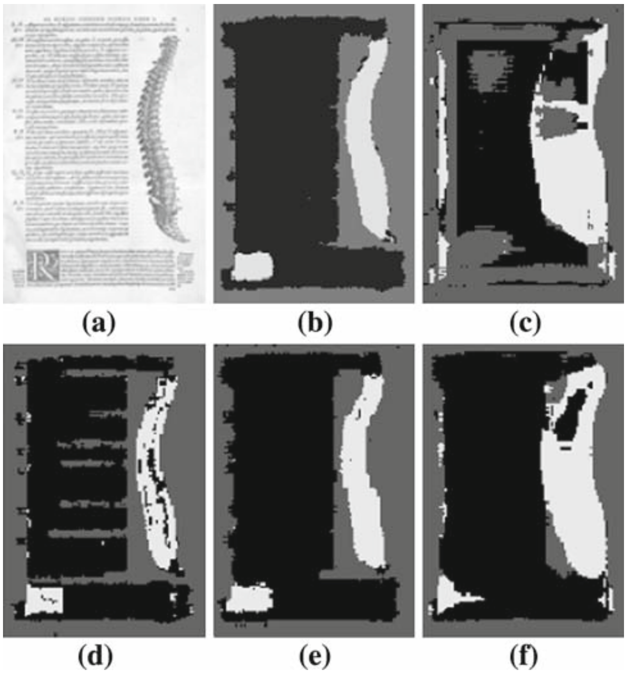

Fig. 7 **a** Original image. Pixel classification based on **b** multiresolution features, **c** only first resolution features, **d** only second resolution features, **e** only third resolution features, **f** only fourth resolution features

5 Features evaluation through retrieval applications

The tests that were carried out in this last section show the pertinence of the texture features and the advances that they represent in terms of characterization of content are detailed.

5.1 Comparison of pages

The first tests that we wish to carry out consists of the comparison of pages of old documents. These tests will help the comparison of their layout coming from texture pieces of information without segmenting in any way. We chose to characterize the pages by using the spatial organization of blocks of texts, of drawings and of background. On the basis of this definition, we propose the use of tools for comparison of partitions presented in [32]. In the framework of our research a partition is the result of a classification of pixels carried out on the basis of the generated texture features.

As it is detailed in the previous section, each pixel of a complete work is first associated with one of the three clusters (text/graphics/background). Thus each image does not have the same number of classes (i.e. a page can be composed of only text and background). As the images have the same size, it is possible to apply the following comparison. Let α and β be two images for which a classification of their pixels has been carried out ($L^{\alpha}(i) = u$ means that the class of the pixel i from the image α is u). It is then possible to build a contingency table $\overset{\alpha,\beta}{N}$ of these two images α, β: $\overset{\alpha,\beta}{N}_{uv\in p,q} = \sum_{i\in n} X^{i}_{uv}$ and $X^{i}_{uv} = 1$if $L^{\alpha}(i) = u$ et $L^{\beta}(i) = v$ else $X^{i}_{uv} = 0$.

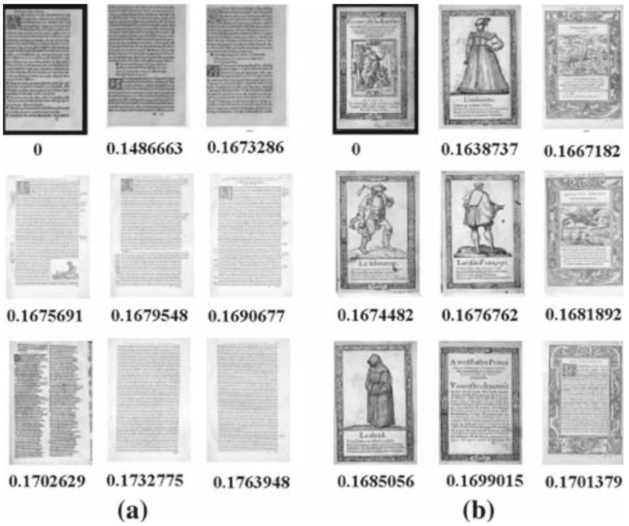

Fig. 8 Examples of document requests results. **a** Page with text and drop caps request, **b** page with borders request; *top left* is request; dissimilarity value is under each result

This table allows the comparison of two partitions in a reduced data space ($\overset{\alpha,\beta}{N}_{uv}$ is of dimension pXq with p the number of clusters of the image α and q the number of cluster of the image β). The building is in $O(n)$, with n the number of pixels of the analyzed images.

This contingency table is at the basis of a dis-similarity measure $S(\alpha, \beta)$ between two images: $S(\alpha, \beta) = \frac{\sum_u \sum_v N_{uv}^2 + \sum_u N_{uu}^2 - \sum_u N_{u.}^2 - \sum_v N_{v.}^2 + n^2}{n^2}$. To estimate the quality of the comparison of documents, we drew our inspiration from the works of [22]. We have then decided to separate the documents into five different classes: the pages with a frame which entirely surrounds the content, the pages made up of text only and justified on the right and on the left, the pages made up of text only but this time arranged into two columns, the pages made up of a drop cap only and the rest of the page made up of text only and finally the pages made up of drawings only. The results shown in the rest of the article were all carried out on the same database. We have then chosen nearly 400 pages out of 9 different works. Every test begins with the application of the classification algorithm for three clusters (text/graphics/backround).

Figure 8 shows the system's ability to detect visually similar pages in a large data-base. An image is given as request (the top left image) and the system provides the images which are the most similar to it.

Table 1 summarizes the rate of good answers obtained to five kinds of different requests. The results meet accuracy rates to a Top5, Top10 and Top15. An accuracy rate is

Table 1 Accuracy rate obtained for five styles of different requests

	Top5	Top10	Top15
Pages with borders	1	0.93	0.86
Text on two columns	0.93	0.76	0.78
Pages with drawings only	0.9	0.62	0.6
Pages with text only	0.74	0.56	0.50
Pages with text & drop caps	0.65	0.56	0.55

worked out by dividing the number of obtained good answers after request by the considered number of images (size of the studied Top), Rate $= \frac{\text{good results}}{\text{size of the top}}$. In the tests carried out, all the pages of the whole works are mixed. The used measure allows us to categorize very different structures visually from each other.

5.2 Comparison of textured images

The second experiment consists of doing a search for images by the content on a basis made up of historical drawings of old documents. We have made up a basis of tests containing more than 400 images. More than a third of the basis is made up of drop caps, the rest is divided into several categories: coats of arms, characters, emblems, skulls, various ornamental elements, etc. We wish to calculate a dis-similarity between two images according to the characteristic textures of which they are made up. For this the use of a metric system allowing the computation of a dis-similarity between two matrices of texture signs is proposed. The dis-similarity function $d(k,l) = \sqrt{\text{trace}((\overset{k}{C}_{i,j} - \overset{l}{C}_{i,j}).^{t}(\overset{k}{C}_{i,j} - \overset{l}{C}_{i,j}))}$ makes it possible to measure dis-similarity between two images k, l and $\overset{k}{C}$ the matrix describing the textures of the images k and l. In this section, the tests meet a search for images by the example. The main interest of this dis-similarity functions is the high speed computation due to the direct comparison of texture matrix attributes (the classification process is not performed here).

Figure 9 shows the good results obtained on the drawings databases. The requested image is the top left one. Below every output is indicated the measurement of dis-similarity between each image and the request one. After studying the results, we notice that the discrimination of the different categories of the basis meets the anticipations. On more than 100 tested drop caps, the majority of the obtained answers in a top 20 are drop caps.

To allow a global estimation of the requests made, we propose to implement the same procedure as the one which was used for the comparison of pages. An accuracy rate to a top 5, 10 and 15 at two different textures of the basis is calculated.

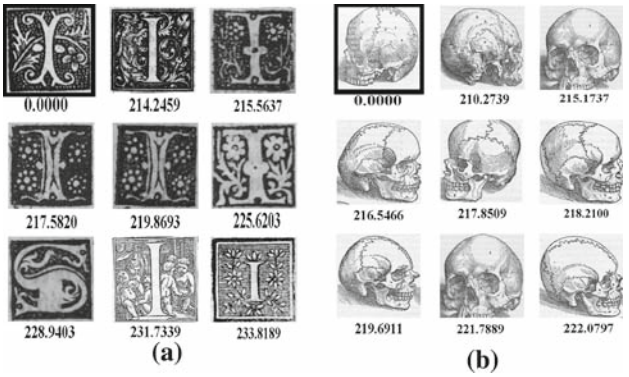

Fig. 9 Examples of image requests results. **a** Case where the request is a drop cap, **b** case where the request is a skull; *top left* is request; dis-similarity value is under each result

Table 2 Accuracy rates for different requests

	Top5	Top10	Top15
Drop caps	0.95	0.92	0.90
Characters	0.92	0.90	0.89
Skulls	0.91	0.86	0.79
Emblems	0.90	0.87	0.78
Coasts of arms	0.88	0.78	0.73

Table 2 sums up the obtained average rates. The obtained accuracy rates are encouraging. Good classification results were obtained. These experiments shows the accuracy of the proposed texture features.

6 Conclusion

This article presents our proposal for a characterization of document images without any a priori knowledge. The originality of our proposal is based on the fact that we do not try to segment or extract the structure of the analyzed documents. Thus, we describe how it is possible to characterize the content of documents by basing pieces of information on non-parametrical textures and with a multi-resolution approach. By extracting signatures linked to the frequencies and the orientations of the different parts of a range, it is possible to extract and to compare elements of content without putting forward a hypothesis about the physical or logical structure of the analyzed documents. We still need to study their integration into more complete indexing devices (CBIR systems for example). The first of these prospects that we set ourselves is then to finalize an indexing system able to produce automatically the descriptive meta-data of the old document images comprising our texture signs but also other pieces of information (linked to the colors, the shapes, the positions, etc.).

Acknowledgments We would like to acknowledge CESR-CNRS (France) for providing the images (http://www.cesr.univ-tours.fr). All images are CESR intellectual property right.

References

1. Allier, B., Emptoz, H.: Font type extraction and character prototyping using gabor filters. ICDAR **02**, 799–804 (2003). http://doi.ieeecomputersociety.org/
2. Antonacopoulos, A.: Page segmentation using the description of the background. Comput. Vis. Image Underst. **70**(3), 350–369 (1998). doi:10.1006/cviu.1998.0691
3. Basa, P., Sabari, P.S., Nishikanta, R.: Gabor filters for document analysis in Indian bilingual documents. Proc. Int. Conf. Intell. Sens. Inf. Process. **1**, 123–126 (2004)
4. Bres, S.: Contributions a la quantification des critFres de transparence et d'anisotropie par une approche globale. Ph.D. thesis, LIRIS, Université de Lyon (1994)
5. Caron, Y., Charpentier, H., Makris, P., Vincent, N.: Power law dependencies to detect regions of interest. Lect. Notes Comput. Sci. **2886**, 495–503 (2003)
6. Chan, W., Coghill, G.: Text analysis using local energy. Pattern Recognit. **34**(12), 2523–2532 (2001)
7. Chetverikov, D., Liang, J., Komuves, J., Haralick, R.M.: Zone classification using texture features. In: ICPR '96, vol. III–7276, p. 676. IEEE Computer Society, Washington, DC (1996)
8. Cinque, L., Lombardi, L., Manzini, G.: A multiresolution approach for page segmentation. Pattern Recogn. Lett. **19**(2), 217–225 (1998). doi:10.1016/S0167-8655(97)00169-4
9. Doermann, D.: The indexing and retrieval of document images: a survey. Comput. Vis. Image Underst. CVIU **70**(3), 287–298 (1998). http://citeseer.ist.psu.edu/doermann98indexing.html
10. Eglin, V.: Contribution a la structuration fonctionnelle des documents imprims. Ph.D. thesis, LIRIS (1998)
11. Eglin, V., Bres, S.: Analysis and interpretation of visual saliency for document functional labeling. Int. J. Doc. Anal. Recognit. **7**(1), 28–43 (2004). doi:10.1007/s10032-004-0127-2
12. Etemad, K., Doermann, D., Chellappa, R.: Multiscale segmentation of unstructured document pages using soft decision integration. IEEE Trans. Pattern Anal. Mach. Intell. **19**(1), 92–96 (1997). doi:10.1109/34.566817
13. Hall-Beyer, M.: Glcm texture: a tutorial. Technical report (2000). http://www.cas.sc.edu/geog/rslab/Rscc/mod6/6-5/texture/tutorial.html, GLCM
14. Haralick, R., Shanmugam, K., Dinstein, I.: Textural features for image classification. SMC **3**(6), 610–621 (1973)
15. Journet, N., Mullot, R., Ramel, J.Y., Eglin, V.: Ancient printed documents indexation: a new approach. In: ICAPR (1), pp. 580–589 (2005)
16. Kaufman, L., Rousseeuw, P.J.: Finding Groups in Data. Wiley, New York (1990)
17. Khedekar, S., Ramanaprasad, V., Setlur, S., Govindaraju, V.: Text–image separation in devanagari documents. In: ICDAR '03: Proceedings of the Seventh International Conference on Document Analysis and Recognition, vol. 2, p. 1265. IEEE Computer Society, Washington, DC (2003)
18. Laws, K.I.: Rapid texture identification. In: Image processing for missile guidance; Proceedings of the Seminar, San Diego, CA, July 29–August 1, 1980 (A81-39326 18-04) Bellingham, WA, Society of Photo-Optical Instrumentation Engineers, pp. 376–380 (1980)
19. Ma, H., Doermann, D.: Gabor filter based multi-class classifier for scanned document images. In: ICDAR '03: Proceedings of the Seventh International Conference on Document Analysis and Recognition, p. 968. IEEE Computer Society, Washington, DC (2003)
20. Maderlechner, G., Suda, P., Breckner, T.: Classification of documents by form and content. Pattern Recogn. Lett. **18**(11–13), 1225–1231 (1997). doi:10.1016/S0167-8655(97)00098-6
21. Mao, S., Rosenfeld, A., Kanungo, T.: Document structure analysis algorithms: a literature survey. SPIE **5010**, 197–207 (2003)
22. Marinai, S., Marino, E., Soda, G.: Tree clustering for layout-based document image retrieval. In: Proceedings of DIAL '06, pp. 243–253. IEEE Computer Society, Washington, DC (2006). doi:10.1109/DIAL.2006.44
23. Nagy, G., Kanai, J., Krishnamoorthy, M., Thomas, M., Viswanathan, M.: Two complementary techniques for digitized document analysis. In: DOCPROCS '88: Proceedings of the ACM Conference on Document Processing Systems, pp. 169–176. ACM Press, New York (1988). doi:10.1145/62506.62539
24. Nicolas, S., Kessentini, Y., Paquet, T., Heutte L.: Handwritten document segmentation using hidden Markov random fields. ICDAR **1**, 212–216 (2006)
25. Pavlidis, T., Zhou, J.: Page segmentation by white streams. ICDAR **2**, 945–953 (1991)
26. Ramel, J., Busson, S., Demonet, M.: Agora: the interactive document image analysis tool of the bvh project. DIAL **0**, 145–155 (2006). doi:10.1109/DIAL.2006.2
27. Shafait, F., Keysers, D., Breuel, T.M.: Performance comparison of six algorithms for page segmentation. In: Procedings of the Seventh IAPR Workshop on Document Analysis Systems (DAS) **3872**, 368–379 (2006)
28. Shi, Z., Govindaraju, V.: Multi-scale techniques for document page segmentation. ICDAR **0**, 1020–1024 (2005). doi:10.1109/ICDAR.2005.165
29. Tuceryan, M.: Moment-based texture segmentation. PRL **15**(7), 659–668 (1994). http://citeseer.ist.psu.edu/tuceryan94moment.html
30. Uttama, S., Ogier, J., Loonis, P.: Top-down segmentation of ancient graphical drop caps. GREC, pp. 87–95 (2005)
31. Wong, K.Y., Casey, R.G., Wahl, F.M.: Document analysis system. IBM J. Res. Dev. **26**(6), 647–656 (1982)
32. Youness, G., Saporta, G.: Une méthodologie pour la comparaison de partitions. Revue de Statistique Appliquée **52**, 97–120 (2004)