

Digital Libraries and Document Image Analysis Techniques: a Survey

Simone Marinai and Beatrice Miotti and Giovanni Soda

Abstract

Nowadays, Digital Libraries have become a widely used service to store and share both digital born documents and digital versions of works stored by traditional libraries. Document images are intrinsically non-structured and the structure and semantic of the digitized documents is in most part lost during the conversion. Several techniques related to the Document Image Analysis research area have been proposed in the past to deal with document image retrieval applications. In this chapter a survey about the more recent techniques applied in the field of recognition and retrieval of text and graphical documents is presented. In particular we describe techniques related to recognition-free approaches.

1 Introduction

Under a broad point of view a Digital Library (DL) can be seen like a more general document database. If all is known about the preservation, indexing, and retrieval of records belonging to structured and fielded data, maintenance and accessing to a full text document archive is a more challenging problem. Traditional relational databases store information in a structured way and each field of the records can be accessed, the queries can be formulated, and records retrieved by indexing the involved fields [14]. In Digital Libraries, data are in most cases made by digitized documents and as such they are less structured [76][78]. Therefore, it is not easy to define suitable queries to this kind of archive because of the difficulties to understand the semantics of the stored data. If it is not easy to retrieve information from scanned documents processed by Optical Character Recognition (OCR) or from digital born documents, because of their limited structured nature, it is even more complicated when documents are stored as images and it is not possible to use recognition-based

University of Florence e-mail: simone.marinai@unifi.it

approaches. Storing the whole documents by their images preserves users from losing important and useful information which cannot be suitably recognized by OCR such as graphics, pictures, and stylistic features like font and layout. In case of document images not processed by OCR the textual content is not explicitly available and this prevents users from performing queries on the basis of the text itself. Recent literature proves that this challenging problem has gained the interest of researchers and several approaches to automatic indexing and retrieval of document images have been proposed.

Nowadays, digital library technologies are well established and understood. This is proven by the large number of books and papers related to this topic and published in the last few years [40][64][70][78]. When DLs deal with scanned images of the works held in traditional libraries, Document Image Analysis and Recognition techniques (DIAR) can be applied to create, store, retrieve, and transmit electronic documents. When dealing with printed text the images can be processed by means of Optical Character Recognition systems to extract the textual content. These kinds of techniques are available, and perform well, on most document typologies, but a lot of information, e.g. related to layout or text style, is likely lost. Furthermore, depending on the nature of the data to be accessed, OCR systems do not perform well unless some ad hoc training has been carried out. For instance in the case of handwritten documents and ancient manuscripts, as in case of tables or mathematical formulae. A different approach with respect to the document recognition advises in the analysis of document images a relevant alternative especially in the cases which bring to a failure of OCR systems. In this case each document is seen as a set of pixels without any known relationship among them. The retrieval of a semantic query on a document collection is in this case translated to a different domain because it must be disguised as a set of image features such as color, shape, texture, and spatial relations. This broad category of approaches, called Document Image Retrieval (DIR), is an important research line. DIR techniques identify relevant documents relying only on image features and are the main subject of this chapter.

This chapter is organized as follows. In Section 2 the retrieval paradigms are presented. The focus is on approaches related to the recognition-free strategy; in particular the word spotting, the word shape coding, and the bag of words techniques are described. In the next sections the retrieval process summarized above is described step by step analyzing each phase and reporting the most commonly used approaches. An overview of the features commonly used in DIR to describe documents is reported in Section 3. The techniques used to represent these features are analyzed in the next Section. In Section 5 the similarity functions used to compare document image representations are reported. In Section 6 some clustering techniques used in a bag of words context are described. Finally, in Section 7 the matching approaches are described on the basis of the different feature representations and similarity functions previously described. Some concluding remarks are in Section 8.

2 Retrieval Paradigms

Several techniques which can be used to perform information retrieval from Digital Libraries have been proposed. The majority of these techniques follow a common paradigm: the documents are first stored and indexed in an offline phase; then the user formulates a query and the system evaluates its similarity with the stored documents and gives as output the ranked results. One important difference among the various techniques is the “level” at which the similarity computation occurs [44][75].

The simplest approach can be referred to as *free browsing*: a user browses through a document collection, looking for the desired information. In this case the similarity is evaluated by the user that visually identifies the most relevant documents.

The second approach is the *recognition-based retrieval* which relies on the complete recognition of the documents. According to it the similarity between documents is evaluated at the symbolic level and it is expected that a recognition engine can extract the full text from text-based documents or metadata from multimedia documents. The textual information is then indexed and the retrieval can be performed either by considering full text queries or by means of keywords provided by the user. The recognition-based approach has the advantage that the similarity computation and result ranking has a low computational cost. On the other hand it has some limitations when dealing with very noisy documents or containing multilingual texts printed with non-standard fonts and a variable layout, such as historical or damaged ones. Some of the earliest methods adopted for the recognition-based approach, and in particular for the OCR-based text retrieval, have been described in two comprehensive surveys [14][51]. Recently, some works have proposed to use a mixed approach where document image analysis techniques are used together with OCR engines and metadata extraction. For instance, in [6] Belaid et al. propose an indexing and reformulation approach for ancient dictionaries, where OCR engines are trained to classify additional classes such as ligatures, gothic characters and specific shapes. In [9] and [29] the OCR engine is used to recognize words and perform layout analysis, while the vector model approach, derived from information retrieval, is used to index the dataset and retrieve the documents.

The last technique is the *recognition-free approach* which is related to content-based image retrieval (CBIR). In this case the similarity is evaluated considering the actual content of the document images that is described by means of suitable features like color, texture, or shape. An advantage of a content-based retrieval approach is the possibility of looking for information without the need of specific domain knowledge. For example, users may not be able to perform correct textual queries if they have no knowledge about the indexed language, but can perform layout queries in a language independent manner. On the other hand, even the CBIR approach has some problems, especially regarding the selection of appropriate features to represent the indexed objects. Most systems work with low level features such as color, texture, and shape, while only few systems attempt to extract high level or semantic features. Examples of these techniques are reported in [2][34][39]

where keyword spotting techniques have been proposed considering a word-level representation on the basis of a set of low-level features.

Word spotting is one widely used approach to perform text retrieval in the recognition-free paradigm. Word spotting was initially proposed by Jones in the field of speech processing [23], while later this definition was adopted by several researchers for printed [12] or handwritten [43] document indexing. This approach permits to localize a user selected word in a document without any syntactic constraint and without an explicit text recognition or training phase. The word spotting technique draws all the word images belonging to indexed documents and returns a ranking of them according to a similarity measure with the query word image. This method is more closely related to CBIR than to word recognition because the matching is carried out considering image features only. Some word spotting methods are based on the clustering of words belonging to the document collections, in order to create the index. The clustering divides the word images into equivalence classes. In each class we expect to have several instances of the same word. By assigning to one representative word for each cluster an ASCII interpretation, it is possible to index the word occurrences in the indexed documents. One disadvantage of this approach is that it can be sensitive to the style and font used for the template word especially when dealing with handwritten text. One important element of word-spotting methods is the segmentation technique used to extract the textual items from the documents. The segmentation can be carried out at local or global level. In the first case each word is segmented into characters and therefore one crucial step is the splitting of scanned images in elementary objects. On the contrary in the global approach the recognition takes place considering the whole word without attempting to perform the character segmentation.

The approaches proposed to represent words in keyword spotting methods can be roughly divided into two groups. The first class methods analyze a word image by means of global image-level features, such as intensity autocorrelation and moments that are used to represent each image. These approaches are suitable for low quality documents and are language independent but require a training phase to identify the best feature combination. The second group of methods is based on word shape coding where each word image is encoded as a sequence of symbols roughly corresponding to characters. In most cases the symbol set has a lower cardinality with respect to the character set in the original language and it is easier to recognize. Each word is in this case represented by a symbol string. Because of the reduced number of symbol classes, usually there is no guarantee of a one to one correspondence between a symbol and a character and therefore a symbol string can be mapped to several words. The main advantage of these approaches is the simple query formulation and the absence of a training phase. However, they are language dependent and are not as robust as the first group in case of poor quality images.

A common model used in Information Retrieval to represent documents is the Bag of Words (BOW) approach, early referred by Zellig H.¹ in [20]. According to

¹ *And this stock of combinations of elements becomes a factor in the way later choices are made ... for language is not merely a bag of words but a tool with particular properties which have been fashioned in the course of its use.*

this schema a document is represented by the occurrences of words in it regardless their position in the document. The Bag of Visual Word (BOVW) approach has been introduced in a paper on object and scene video retrieval [67] as an extension of BOW to the case of images. The BOVW approach relies on three main steps: in the first a certain number of image keypoints or local interest points are automatically extracted from the image by means of an appropriate detector. Keypoints are salient image points rich of information content and for this reason, suitable to describe the whole image. In the second step keypoints, or in some cases shape descriptors evaluated on keypoints, are clustered and similar descriptors are assigned to the same cluster. Each cluster corresponds to a visual word that is a representation of the features shared by the descriptors belonging to that cluster. The cluster set can be interpreted as a visual word vocabulary.

In the last step each image is described by a vector containing the occurrences of each visual word in that image. The most critical points of this approach are the detection of the local interest points (e.g. by means of Scale-Invariant feature transform (SIFT) or corner points) and the choice of the most suitable description of the regions of interest. Regarding the clustering method, K-means, K-Nearest Neighbor algorithm (K-NN), probabilistic Latent Semantic Analysis (pLSA) and Support Vector Machine (SVM) are the most popular techniques [55].

In the rest of this chapter the main steps of the recognition-free systems are presented. In Section 3 several types of features are presented according to the image level in which they are computed. In Section 4 the main representation models are described. In Section 5 and Section 6 the different kinds of distances and clustering techniques are presented. In Section 7 the matching approaches are described considering the different features and descriptors used in the previous steps. Some conclusions are then reported in Section 8.

3 Features

In a document image retrieval system the identification of features is a crucial task since it significantly affects the overall performance. Broadly speaking, the features can be divided into two main groups: the first is related to *local* features, according to which one feature is extracted for each point in the input domain, in the second group *global* features are evaluated on sets of pixels (e.g. a word), on a region or even on the whole document.

3.1 Pixel level

When features are computed at a local level some values are obtained for each pixel. In [33] Leydier et al. propose a word-spotting method to access the textual data of medieval manuscripts. This approach does not require image binarization and

layout segmentation and is tolerant to low resolution and image degradations. The informative parts of the images are represented through a set of features provided by gradient orientation. In [24], Journet et al. propose a method for the characterization of pictures of old printed documents based on a texture approach. For each pixel of the image, five features are extracted at different resolutions for a total of 20 values. In particular, three features are related to the texture orientation and evaluated by means of an auto-correlation based approach while the other two are related to the properties of the pixels grey level transitions. The computation of some features in regions defined by a grid superimposed to the page is adopted in [22][73], while in [49] this zoning technique is applied to the connected components of symbols in the task of script recognition and writer identification. In the latter case for each cell the number of black pixels is computed. Something different is proposed by Delalandre et al. [13] regarding the retrieval of old printed graphics (initial letters) from a large database. A run length encoding algorithm is first used to compress the image and then a template matching between images is obtained evaluating the distance with a pixel to pixel comparison.

The methods in this category are in most cases very expensive from the computational point of view because of the high number of objects involved both for the feature extraction (during the indexing) and for the matching (during the retrieval).

3.2 Column level

Some approaches require a segmentation phase, such as the segmentation of words and characters. In this case a method based on the analysis of column pixels in segmented objects can be exploited. In [28], Khurshid et al. present a method for figure caption detection which is performed by wordspotting of figure labels. The segmentation of words and characters is done by finding the connected components and, for each pixel column of the character, a set of 6 features is calculated: vertical projection profile on the gray level image, upper character profile position, lower character profile position, vertical histogram, number of ink/non-ink transitions and middle row transition state. Similarly, in [11] words are analyzed column by column and the corresponding Hidden Markov Model is built.

3.3 Sliding window

One technique related to column level representation adopts a sliding window. In this case a fixed size window is moved across the word image and some features are evaluated for each position. This strategy is frequently used to obtain the input descriptors for supervised classifiers such as the MultiLayer Perceptron (MLP) neural network [46]. The sliding window approach is also used in [72], where Terasawa et al. propose a method for word spotting in historical handwritten documents without

performing word segmentation. In this case the text lines are scanned by a sliding window whose width depends on the character sizes and a low dimensional descriptor is generated for each slit image by applying the eigenspace method. A different task is addressed by Schomaker et al. in [66] where they propose a line strip retrieval on the basis of content similarity with respect to the query. Line strips are used as the basic objects for search and retrieval because they represent a good compromise between the reliability of segmentation and the recognition performance. The sliding window has a width defined on the basis of the query line and is moved along the indexed text lines. Some features related to the ink density, the connected component shapes and contours are computed at each step. In [36] Licata et al. develop a system to identify the provenance of ancient handwritten documents basing on the ink appearance similarity. The features considered in the sliding window are in this case first and second order statistical features such as histogram mean, skewness, contrast, entropy. Uttama et al. [74] exploit the sliding window approach to separate homogeneous and textured regions in images of historical documents. The features are based on the Gray-Level Co-occurrence Matrix.

3.4 Stroke and primitive level

When the objects in document images are complex and important spatial relationships among primitives are possible, such as in sketches [35] and in trademarks [77], one structural representation, which is able to represent the variety of connections, is essential. Liang et al. [35] and Wei et al. [77] propose the combination of structural and global features: the structural part describes the interconnections among primitives and the global features reflect the object as a whole. Liang et al. [35] exploit eight kinds of spatial relations between primitives: cross, half-cross, adjacency, parallelism, cut, tangency, embody and ellipse intersection. Global features are composed by seven types of descriptors: eccentricity, normalized distances of sketch centroid in major and minor axis orientations, average distance between centroids of sketch primitives and the number of primitives of each type as line, arc, and ellipse. Wei et al. [77] propose global features based on 15 Zernike moments of orders 0-4, the standard deviation of the curvature, the mean and the standard deviation of distance to centroids.

In the context of shape-based image retrieval, Wong et al. [79] propose a two step feature extraction process: the shape contour is first represented by the Freeman chain code as a connected sequence of straight line segments with specified lengths and directions. Then the relative spectrum is plotted as the normalized curve length with respect to the normalized geometric moment, where the normalized curve length is the length of the segment between two keypoints of a shape divided by the total curve length and the normalized geometric moment is the distance between a dominant point and the geometric center. On this spectrum four features are evaluated: the total normalized moment variance, the total normalized area cov-

ered by the spectrum, the cross-sectional normalized area and the cross-sectional normalized moment variance.

In [16] Fonseca et al. present a shape classification technique based on topological and geometrical descriptors. In this case the spatial organization is described by the relationships of inclusion and adjacency, while the geometry of shapes is described by some geometric attributes like area and perimeter. In [61] Rusinol et al. deal with symbol recognition starting from a vectorial representation of the image. Information about constraints between segments such as parallelisms, straight angles and overlap-ratios are analyzed.

Dealing with symbol-spotting, Rusinol et al. [62] propose to detect graphical symbols in large and complex document images by techniques which do not need neither a segmentation step nor a priori knowledge. For each symbol its primitives are extracted by means of connected component analysis and contour detection then the Cassinian ovals parameters are evaluated. In a two-center bipolar coordinate system a Cassinian oval is described by all the points such that the product of their distances from the two centers is a constant. In this case, the parameters that characterize the minimum Cassinian oval which encompasses the normalized shape contour, are used as shape features. Similarly Zhang et al. [81] propose to extract a simple set of features from the vectorial representations of the symbols: in particular they propose to evaluate the angle between two line segments to represent their mutual relationship. The relationship between a line segment and an arc and between arcs is represented considering the angle between arc starting points and arc centers. Finally, the relationship between a line segment and a circle is evaluated considering the tangency, intersection, and disjointness relations.

A similar approach can be even used in text document analysis. For instance in [10] Chellapilla et al. segment words in strokes and then the sequence of Chebyshev polynomial coefficients is evaluated for each segment. In [21] a word spotting for online handwritten documents is proposed. In this work strokes are sampled and for each sampled point three features are evaluated: the height of the sample point, the direction and the curvature of the stroke in that point.

3.5 Connected-component level

In the processing of handwritten or ancient printed documents, it is not always easy to segment a document and to identify text lines, words, and characters. Especially the character segmentation is a difficult task because of the variability of handwriting and the presence of touching characters. In [8] handwritten touching characters are addressed by identifying all the possible ligatures connecting two characters by heuristic analysis of the contour. In so doing a word image is divided in several pieces. Assuming that a character is made up by at most four consecutive pieces, a series of hypothesized character images are created. One way to segment text in objects, broadly corresponding to characters, is to find the connected components in the document image. Moghaddam et al. [52] present a line and word segmen-

tation free method to perform word-spotting on old historical printed documents. After the detection of the connected components, a set of six features: aspect ratio, horizontal frequency, scaled vertical center of mass, number of branch points, height ratio to line height, and presence of holes, are extracted from them. Similarly, Marinai [45] propose to use the connected component clustering as a first step in the indexing of early printed books. Relevant words are identified considering a modified Dynamic Time Warping (DTW) algorithm that includes the word width in the distance computation. Barbu et al. [5] work on graphical document images considering graph-based representations: connected components in the image are represented by graph nodes and rotation and translation invariant features, based on Zernike moments, are extracted.

3.6 Word level

In most word spotting applications it is possible to assume that the word segmentation in indexed documents is not problematic. In this case the retrieval is carried out considering the word as a whole. In [3] each word is described by profile-based and shape-based features, while in [57] and [58] Rath et al. present single value features based on Projection Profile, Word Profile, Background to Ink Transitions, and Grayscale Variance. In [80] Zhang et al. propose to use the Gradient-based binary features (GSC) evaluated under a 4×8 division of the word image. GSC features are based on the evaluation of the direction of the gradient, on structural information, and on concavity features. Structural and concavity features are evaluated by means of 12 rules applied to the image pixels and are based on the pixel density, on larger strokes in both horizontal and vertical directions, and on the direction of concavity each pixel belongs to. Similarly, in [31] and [27] the zoning technique is applied to word images and the density of the character pixels in each zone is evaluated. Subsequently a second group of features based on the area under the upper and lower word profiles, is considered. In [17] and [60] some sets of feature vectors are evaluated for non-overlapping windows on the query image: in [17] the features are based on pixel density, in [60] the Local Gradient Histogram (LGH) features, introduced by Rodriguez in [59], are used. Dealing with printed documents, Meshesha et al. [50] propose to describe words by means of word profiles, moments and transform domain representations. Similarly, Bai et al. [2] propose the extraction of seven features: character ascenders, descenders, deep eastward and westward concavity, holes, i-dot connectors and horizontal-line intersection. In [42] features are extracted for each word by means of the Left-to-Right Primitive String (LRPS) algorithm. This algorithm splits each word in primitives which can be described by means of two-tuples: the Line-or-Traversal Attribute (LTA) and the Ascender-and-Descender Attribute (ADA). In a more recent work [40], Lu et al. extend the previous approach to the case of printed document images captured by a digital camera. In this task the set of extracted features includes three perspective invariants: holes, water reservoirs, and character ascenders and descenders. Nakai et al. [53] propose

a mixed approach: they work at a lower level considering the centroids of connected components and some features related to the area of connected components while at a higher level analyze words and compute the word centroids.

3.7 *Line and Page level*

Tan et al. [71] deal with the script identification among three different on-line handwritten scripts: Arabic, Roman and Tamil. After the detection of text lines, they extract a set of features at line-level such as the horizontal and vertical interstroke direction, horizontal and vertical stroke direction, average stroke length, stroke density and the reverse direction.

Some features can be extracted at page level by means of geometric transformations. In [25] Joutel et al. develop a system for paleographers and literary experts, to support their work on manuscripts dating and authentication through different historical periods. The approach is based on the Curvelet transform to compose an unique signature for each handwritten page.

3.8 *Shape Descriptor*

Shape descriptors are frequently used in image analysis to compare 2D object silhouettes. Recently they have been adopted also in document image analysis to compare symbol images in a recognition-free approach. According to the object representation the shape descriptors are evaluated on, three main categories can be identified [32]. In the first category *contour based* descriptors are evaluated on the object contours; in the second, *image based* descriptors include the shape descriptors based on the overall image pixel values; in the last category, *skeleton based* descriptors are evaluated on the image skeletons. Since document image retrieval has to deal with images affected by scale and perspective changes, the shape descriptors must be invariant to similarity and affine transformations as well as rotation and scale. In most cases, descriptors are computed on keypoints, that are points of the image rich of information content. To reduce the complexity of the evaluation, images are usually preprocessed and the contour or the skeleton are detected. The interesting points are in this case extracted from the preprocessed image. An example of keypoints are corners (points with high curvature), which can be detected with the Harris-Laplace detector [63] or on the basis of the curvature variance [79]. In [54] Nguyen et al. propose the use of the Difference-of-Gaussian detector and the keypoints in an image are considered as the extrema in a scale-space pyramid built with DoG filters.

A first example of shape descriptor is the Scale Invariant Feature Transform (SIFT) proposed by Lowe [38]. SIFT are able both to localize keypoints and to evaluate a shape descriptor on these points according to the local gradient histogram. SIFT descriptors are proved to be invariant to rotation, scale changes and affine

Table 1: Features

Level	Data	Features	Ref.
Pixel	Medieval manuscripts	Gradient orientation.	[33]
	Old printed pictures	Auto-correlation based and pixels grey levels transitions.	[24]
Zoning	Printed	Pixel density.	[27][31][49]
	Layout	Row encoding.	[22][73]
Column	Historical printed	Projection profiles, vertical histogram, number of ink/non-ink transitions.	[33]
Sliding Window	Historical handwritten	A vector is generated for each slit image by applying the eigenspace method.	[72]
	Historical handwritten	Ink density, the connected component shapes and the contours.	[66]
	Historical handwritten	First and second order statistical features (e.g. histogram mean, skewness, entropy).	[36]
	Old printed pictures	Gray-Level Co-occurrence matrix and texture uniformity.	[74]
Stroke and primitives	Sketches/ on-line handwritten	Spatial relations between primitives (e.g. cross, adjacency, parallelism, tangency) and global features (e.g. eccentricity, number of primitives of each type).	[35][71]
	Trademarks	15 Zernike moments of orders 0-4, standard deviation of the curvature, mean and standard deviation of distance to centroids.	[77]
	Shape	Normalized moment variance and the total normalized area.	[79]
	Graphics	Cassinian ovals.	[62]
	Handwritten	Sequence of Chebyshev polynomial coefficients.	[10]
Connected components	Printed	Aspect ratio, horizontal frequency, scaled vertical center of mass, number of branch points, and presence of holes.	[52]
	Graphics	Zernike moments.	[5]
Word	Handwritten	Single value features: projection profiles, background to ink transitions, grayscale variance, gaussian smoothing and gaussian derivatives.	[57][58]
	Printed	Area under the projection profiles.	[27] [31]

Continued on next page

Table 1 – continued from previous page

Level	Data	Features	Ref.
Word	Printed	Word profiles, moments and transform domain representations.	[50]
	Printed	Character ascenders, descenders, deep eastward and westward concavity, holes, i-dot connectors and horizontal-line intersection.	[2] [40]
Page	Historical manuscript	Curvelet transform.	[25]

Table 2: Shape Descriptors

Shape Descriptor	Data	Features	Ref.
Keypoints	Contour or skeleton	Corners detected by means of the Harris-Laplace detector or considering the curvature variance.	[63][79]
	Contour	Interest points are the extrema in a scale-space pyramid built with Difference-of-Gaussian filters.	[54]
SIFT	Character	Modified SIFT algorithm.	[82][63]
Shape Context	Contour or skeleton	Descriptor captures the spatial distribution of points in proximity of the point where it is computed.	[37][48][54][63]

transformations. In [82] Zhang et al. propose the use of a modified version of SIFT descriptors applied to handwritten Chinese character recognition.

Another descriptor is the Shape Context, proposed in [7]. This descriptor is able to capture the spatial distribution of points in proximity of the point where it is computed. The descriptor is represented as a logarithmic polar mask centered on the point of interest and divided in bins in the polar space. Each cell is populated according to the position of other image points with respect to the center of the mask. Shape Contexts are invariant to translation and scale. In [37] shape context descriptors are used to represent words and are computed on points belonging to the skeleton of word images, while in [48] they are computed on points on the contour of mathematical symbols. In [54] shape contexts are computed on points of interest detected by means of Difference-of-Gaussian detector. In [63] Rusinol et al. propose the use of the SIFT and the shape context descriptors for the symbol spotting task and compare their performance with the results obtained with more simple descriptors such as geometric moments of steerable filters.

4 Representation

After a set of features has been extracted from the document image, it is essential to identify a suitable representation of their values. Some of the features presented in the previous section are naturally represented in the form of vectors. For instance in the case of zoning the features are extracted from each cell of a grid superimposed to the image and the resulting descriptor has the same dimensionality of the number of cells [17][22][49][73][82]. In case of a sliding window each slit is usually represented by a low dimensional descriptor [66][72], or by an high-dimensional descriptor such as in [36] where the intensity histogram statistics and co-occurrence statistics are concatenated. When the features are evaluated at pixel level, such as in [74], a feature vector of a certain dimension is created for any pixel of the input image. If the segmentation is performed at word level [3][27][31][80], some binary vectors are generated starting from the features related to the word profiles, while in [16] and [35] the same approach is proposed in case of shape classification where the features describe the spatial organization and structural characteristics of geometric shapes. Even in case of features computed at character or connected-component level, the vector representation is usually exploited. Some examples can be found in [8][21][47][52]. Regarding the automatic indexing and retrieval of graphical document images, Barbu et al. [5] describe each connected component related to a graphical object as a feature vector. A similar approach is used in [37][48][54], but in this case Shape Context descriptors are used to describe the shape of a word and each word is represented as a collection of vectors. Rusinol et al [63] exploit a vector representation for each type of shape descriptor. In the field of shape matching, Super et al. [69] propose to describe the shape contour as a vector in a $2n$ - dimensional shape space, where n is the number of contour sample points. An extension of the vectorial features representation is proposed in [81] where symbol signatures are represented in a matrix form and each bin of the matrix represents the relationships between strokes of the symbol, such as segments, arcs and circles.

When the number of documents in a database is large the retrieval process can be computationally expensive. Some techniques use a compressed data structure to represent the images in order to decrease their handling times. In particular Delalandre et al. [13] propose to use the run-length encoding algorithm to compress images. In [25], Joutel et al. propose an approach for retrieval of handwritten historical documents at page level based on the Curvelet transform to compose an unique signature for each page. In [10] words are segmented and each stroke is represented as the sequence of Chebyshev polynomial coefficients. Then each segment is processed by a Time Delay Neural Network (TDNN) to determine the probability the segment belongs to each possible character in the language. The TDNN outputs associate to each segment the most likely characters according to their membership probabilities.

When the vocabulary is limited, as in bank check recognition or in automatic handwritten mail sorting, probabilistic approaches can be used. In [11] Choisy et al. propose to model words by means of a Non-Symmetric Half Plane Hidden Markov Model (NSSP-HMM), while Rodriguez et al. [60] propose a system where typed

text is used as support for handwritten recognition. They use robust LGH features to describe the word shapes and Semi-continuous HMM (SC-HMM) for modeling the link between typed and handwritten words.

The character and word shape coding approaches are strictly connected to the vector representation of words or characters in case of word spotting. Generally character shape code encodes, in the form of a code string, the properties of each symbol such as whether or not the character in question fits between the baseline and the x-line, whether it has an ascender or descender and the spatial distribution of the connected components [2][68]. In [40] and [42] Lu et al. present a system based on a word image coding schema: the Left-to-Right Primitive String (LRPS), Line-or-Traversal Attribute (LTA) and the Ascender-and-Descender Attribute (ADA) features are used to assign a code to each primitive of the word images. Recently some extensions of the previous works have been presented by Lu et al. [39][41]. In [41] a word shape coding approach for documents in five different Latin languages has been presented. In this case each word is converted into a word shape code that is composed of two parts. In the first part, character extrema points, which are in the upward and downward text boundaries, are classified as belonging to one of three categories according to their position with respect to the base line. In the second part of the word code the number of horizontal word cuts is reported. A similar approach is proposed by Li et al. in [34] where they encode every word as a sequence of numbers and each number represents a character.

The vector space model was introduced by Salton et al. [65] in 1975. According to this model, documents and queries can be represented as vectors whose elements represent the frequency of each term in the document [41]. Some weighting schema can be applied to the vector model e.g. the tf-idf [15] and some modified versions of it [9][19][71]. The same approach can be extended to the case of vectors corresponding to occurrences of features [48][49][54][61][71].

Another kind of document representation is related to the graph-based approach. In [5] Barbu et al. describe a document image according to a graph-based representation at primitive level analyzing the relationships between connected components. Each connected component is a graph node and one arc between primitives exists only if they are spatially near. The same approach is used in [18] where Gordo et al. represent the document layout as a bipartite graph built considering the centroids of the regions on one side and the center of mass of all the regions on the other. In symbol spotting applications the graph based approach is quite frequent. In [26], Karray et al. develop a method for the analysis of initial letters which is based on the Attributed Relation Graph representation. In particular, after the segmentation step nodes represent regions and arcs express the relationships among regions. For the same task, in [74], Uttama et al. describe an initial letter by means of its signature according to two approaches: Minimum Spanning Tree and Pairwise Geometric Attributes. In [56] each symbol is described by means of a graph to capture the spatial and topological relationships between graphical primitives. Alajlan [1] propose to use the Curvature Tree hierarchical data structure which reflects the inclusion relationships among objects and holes.

Table 3: Representation

Representation	Features	Method	Ref.
Vectors	Zoning	Vector corresponding to a grid overlapped to the image.	[17][22][49][73][82]
	Sliding windows	Descriptors of low or high dimensionality.	[36][66][72]
	Pixel	One descriptor for each pixel.	[74]
	Word	Binary vectors generated from the profile features.	[3][27][31][80]
	Word	Binary vectors describe the spatial organization and structural characteristics of geometric shapes.	[16][35]
	Character or connected-component	One dimensional vector for each character or connected component.	[8][21][47][52]
	Shape descriptors	Describe the shape of a word. Each word is represented as a collection of vectors.	[37][48][54][63]
Matrix	Strokes	Each bin of the matrix represents the relationships between strokes of the symbol such as segments, arcs and circles.	[81]
Probabilistic	Word	Non-Symmetric Half Plane Hidden Markov model and Semi-continuous HMM.	[11][60]
Character and Word shape coding	Strokes	A code is assigned to each primitive of the word image.	[34][39][40][42][41]
Vectorial model	Symbol	Vector elements represent the frequency of visual terms in the documents and queries.	[48][49][54][61][71]
Graph-based	Connected component or strokes	Each connected component is a graph node; arcs connect near primitives.	[5][18][26][56][74]

5 Similarity measure

In document image analysis and recognition, the retrieval of a query word can be performed considering the similarity or distance between two images: the reference image, given by the user, and the dataset images representing the indexed words. According to the different models used to represent keywords, a different similarity

measure may be used. Moreover, in case a clustering phase is performed, the more appropriate distance to compare features has to be chosen such as Euclidean distance [48] or cosine distance [54].

When keywords are represented by feature vectors, the most common way to compare them is the Euclidean distance [16][35][77] or the L1 distance [27][31].

In case of template matching the distance among images is computed at pixel level [13] by a simple value comparison or considering a specific dissimilarity function [33]. When documents and keywords are represented by means of the vector space model in analogy with the vector model in Information Retrieval, to evaluate the similarity it is frequently computed the cosine of the angle between two vectors. This approach is used in [15][39][54]. In [48] a modified version of the cosine similarity is proposed. Taking account of the properties of the clustering algorithm used to group features, Marinai et al. introduce an additional term to the formula which is necessary to deal with inexact match. In [71] the similarity is computed by means of the Chi-square distance which involves the distribution of tf-idf vectors, while in [74] the Bhattacharyya distance between two histograms is used. To deal with a word shape coding representation, some distances may be used. In particular in [15][34] the use of the Edit distance is proposed. The Edit distance between two strings is given by the minimum number of operations needed to transform one string into the other, where an operation is an insertion, deletion, or substitution of a single character. In [52] an enhanced version of the edit distance is proposed where stroke width is used as a priori information. In [40] Lu et. al propose the Hamming distance to compare word shape codes.

6 Clustering

In computer vision applications a family of methods based on the *bag of visual words* framework has been recently proposed [4][5][30][48][54]. These methods extend the *bag of words model* used in textual information retrieval, that represents documents considering the number of occurrences of words, regardless of their position in the text. In the bag of visual words approach a visual vocabulary is constructed by clustering the feature vectors that represent symbols. Each cluster can be considered as a visual word and all the feature vectors belonging to that cluster can be represented by its centroid. The clustering can be performed by means of a semi-supervised learning, or by means of an unsupervised learning approach. To this second class of methods belongs the K-means algorithm used in [58][54] and its revisited version called k-medoids [5]. In [36] Licata et al. propose to use the K-Gaussian clustering where the number of clusters is determined using Minimum Description Length. In this case each feature vector is assigned to a cluster by selecting the component that maximizes the posterior probability. This algorithm may be more appropriate than K-means clustering when clusters have different sizes. Another approach to perform clustering is the Self Organizing Map (SOM) applied to the feature vector quantization [45][47][48][52]. The SOM map has the property

Table 4: Similarity

Similarity measure	Representation	Ref.
Euclidean distance	Vectors/ Vector model	[77][35][16]
L1 distance	Vectors	[27][31]
Comparison	Pixels	[13][33]
Cosine similarity	Vector model	[39][15][54][48]
Chi-square distance	Vector model	[71]
Bhattacharyya distance	Vector model	[74]
Minimum Edit distance	Word coding	[52][34][15]
Hamming distance	Word coding	[40]

that more similar patterns are usually grouped in closer clusters. In [52] the SOM has been used to cluster the feature vectors belonging to connected components while in [48], the shape context descriptors, evaluated for each symbol, are clustered by means of a SOM and then each symbol is represented by the occurrences of shape contexts assigned to a particular centroid according to the Euclidean distance.

7 Matching

Feature matching deals with measuring the similarity between the feature representation of the query image, that is based on feature vectors, graphs or statistical models and the database images. The choice of the feature matching technique is essential for the good performance of the system. As a matter of fact an inappropriate approach may lead to bad results even though the considered feature representation is the more appropriate for that task. When the queries and documents are represented as feature vectors, the matching is carried out comparing the vectors according to some similarity measure. In [17], the query word feature vectors are compared with the corresponding vectors of the database by applying a matching which is constrained by the regions of interest where the features have been computed. The similarity measure is in this case based on the Euclidean distance between vectors. The matching by means of the Euclidean distance is proposed also in [16][35][77]. A similar approach is proposed in [27] and in [31], but in this case the similarity is evaluated by means of the L1 distance to reduce the computational costs. In [69], Super et al. propose to perform the matching between shapes, considering their normalized contours as vectors. To face the different shape sizes, the similarity between two vectors is evaluated by the Euclidean distance normalized by the squared average of the arc lengths of the two poses. In [81] the matching is performed considering how many common relationships at primitive level the symbols share and choosing the shapes that share the most. Also in [74] the matching is performed comparing feature vectors but in this case the Bhattacharyya distance between two vectors is used. This distance exploits the correlation between vector

Table 5: Clustering

Clustering	Representation	Number of clusters	Ref.
K-mean	Feature vectors assigned to the nearest cluster.	1-20, 20	[54] [58]
K-medoids	A more robust version of the K-mean.	16	[5]
K-Gaussian	Feature vectors assigned to clusters by selecting the component that maximizes the posterior probability.	200	[36]
SOM	Feature vectors assigned to the nearest cluster according to the Euclidean distance. Closest neurons in the map are the most similar.	10, > 100	[45][47][48][52]

contents to obtain the similarity measure. A related approach is proposed by Joutel et al. in [25], where the matching between two document signatures is performed by means of the normalized correlation similarity.

When the document images are represented by means of the vector model, some other techniques can be applied for the matching, in particular a common matching schema is the cosine similarity measure. In [41], Lu et al. propose a word shape coding technique for document retrieval and the vector representation of the strings. The similarity between two document representations is evaluated considering the angle formed by the two vectors. The same matching approach is used in [15][39][48][54].

Different matching techniques have been proposed when documents are represented by means of word or character shape coding. In [2][21][57][58], the Dynamic Time Warping is considered to compare two sequences of data points. This method aligns the feature vectors of the query and of each word in the database using a dynamic programming-based algorithm. The algorithm computes distance scores for matching points by means of the Euclidean distance. At the end of the process, similar shapes have a lower distance than different ones. The same approach is used in [50] and [3], but a DTW-based partial matching technique that takes care of word form variations in the beginning and at the end of the word is also proposed. The DTW technique is also used in [18] to perform matching on layout vector representations while in [45] a modified version on DTW that takes into account both the clusters similarity and the estimated widths of alternative sub-words is proposed. In [42] and [39] the DTW algorithm is considered after a pruning step by means of coarse matching that is used to reduce the number of dataset elements the query code has to be compared to.

Another matching technique is proposed at word level in [34] and in [28]. In these cases the matching among code strings is performed by means of the minimum edit distance. The edit distance between two strings is given by the minimum number of operations needed to transform one string into the other, where an operation is an insertion, deletion, or substitution of a single character.

When the document representation is either probabilistic or probability-based, the model obtained in the feature extraction phase is used to evaluate the recognition score [11][60]. On the other hand when the objects are represented in a graph-based approach some ad hoc methods should be used. In particular Qureshi et al. [56] propose to use a graph matching routine where sub graphs are matched against model graphs using polynomial bound greedy algorithm. The output of this process is the score similarity. This technique is error-tolerant and works well in case of under or over segmentation of symbols.

8 Conclusions

In this chapter we have described in a comprehensive survey the main document image analysis and recognition techniques which have been proposed in recent years to perform document image retrieval. We have focused our analysis on recognition-free approaches that do not explicitly recognize the document content (e.g. with OCR tools) but work at various levels with a symbolic or sub-symbolic representation of the document image. The comparison of the different techniques has been organized along the main steps of a general retrieval process.

In the first step the features are extracted from the document images that are represented at various levels, from pixel to primitive, character, word or even zones. In each case different features can be considered according to the task (e.g. wordspotting or trademark retrieval) and to the typology of processed documents (e.g. printed vs handwritten or modern vs historical). In the second processing step the features previously extracted are encoded in suitable representations that depend both on the nature of the features and on the subsequent processing steps. Eventually, the feature representations are used to compute the distance between objects. The distance can be used to cluster indexed objects to speed up the subsequent processing or can be employed in similarity measures to compare query items with indexed ones.

The research on document image retrieval is particularly dynamic in the last few years. To give a measure of it, about half of the papers cited in this chapter have been published after our previous survey on the field appeared in 2006 [44].

Two main areas should be addressed in the future in order to increase the widespread use of the techniques presented in this chapter. From one side approaches that have been tested only at a prototype level should be extended to larger datasets in order to study their effective scalability towards real size DLs. On the other side, a deeper integration of the recognition-free techniques analyzed in this chapter in the common architecture of contemporary Digital Libraries should be addressed and solved. Current Digital Libraries are mostly based on relational database to store the information of interest. The use of this architecture is based on the assumption that the information stored is certain. It is therefore not easy to handle errors or ambiguities in DIR that might be obtained by using recognition-free approaches. Concerning the techniques described in this chapter, we believe that more work is still needed to deal with cursive text and also for the retrieval of

Table 6: Matching

Representation	Similarity	Method	Ref.
Feature vectors	Euclidean distance	Word feature vectors are compared with the database feature vectors.	[77][35][16]
	Constrained euclidean distance	Word feature vectors are compared with the database feature vectors constrained by the regions where features have been computed.	[17]
	L1 distance	All word feature vectors are compared with the corresponding feature vectors of database images.	[27][31]
	Normalized correlation	Word feature vectors are compared with the database feature vectors.	[25][74]
Vector model	Cosine similarity	The similarity is evaluated considering the angle formed by the two document vectors: if it is close to zero, the documents are similar.	[41][39][15][54][48]
Word encoding	DTW	Aligns the query and the word feature vectors using a dynamic programming-based algorithm.	[21][57][58][2][18][42][39]
	DTW	DTW-based partial matching technique based on word form variations in the beginning and at the end of words.	[50][3][45]
	EDT	Minimum number of operations needed to transform one string into the other.	[28][34]
Probabilistic	NSSP-HMM SCHMM	The probabilistic model is used to estimate the recognition scores.	[11][60]
Graph	Polynomial bound greedy algorithm	Use a graph matching routine where sub graphs are matched against model graphs.	[56]

other graphical objects. Also in this case, the integration of techniques for graphical items retrieval into existing DL architectures should be addressed. Another related research field whose application in the area of DLs should be explored is the Camera-Based Document Image Analysis that aims at developing recognition and retrieval techniques starting from images captured with mobile devices such as cameras and smart phones.

References

1. Alajlan, N., Kamel, M.S., Freeman, G.H.: Geometry-based image retrieval in binary image databases. *IEEE Trans. on PAMI* **30**(6), 1003–1013 (2008)
2. Bai, S., Li, L., Tan, C.: Keyword spotting in document images through word shape coding. In: *Proc. Int'l Conf. on Document Analysis and Recognition*, pp. 331–335. IEEE Computer Society (2009)
3. Balasubramanian, A., Meshesha, M., Jawahar, C.: Retrieval from document image collections. In: *Proc. IAPR Int'l Workshop on Document Analysis Systems*, pp. 1–12 (2006)
4. Banerjee, S., Harit, G., Chaudhury, S.: Word image based latent semantic indexing for conceptual querying in document image databases. *Proc. Int'l Conf. on Document Analysis and Recognition* **2**, 1208–1212 (2007)
5. Barbu, E., Héroux, P., Adam, S., Trupin, É.: Using bags of symbols for automatic indexing of graphical document image databases. In: *Proc. Int'l Workshop on Graphics Recognition*, pp. 195–205 (2005)
6. Belaid, A., Turcan, I., Pierrel, J.M., Belaid, Y., Hadjamar, Y., Hadjamar, H.: Automatic indexing and reformulation of ancient dictionaries. In: *Proc. Int'l Workshop on Document Image Analysis for Libraries*, pp. 342–354. IEEE Computer Society, Washington, DC, USA (2004)
7. Belongie, S., Malik, J., Puzicha, J.: Shape matching and object recognition using shape contexts. *IEEE Trans. on PAMI* **24**(4), 509–522 (2002)
8. Cao, H., Bhardwaj, A., Govindaraju, V.: A probabilistic method for keyword retrieval in handwritten document images. *Journal of Pattern Recognition* (2009)
9. Cao, H., Govindaraju, V.: Vector model based indexing and retrieval of handwritten medical forms. In: *Proc. Int'l Conf. on Document Analysis and Recognition*, vol. 1, pp. 88–92 (2007)
10. Chellapilla, K., Piatt, J.: Redundant bit vectors for robust indexing and retrieval of electronic ink. In: *Proc. Int'l Conf. on Document Analysis and Recognition*, vol. 1, pp. 387–391 (2007)
11. Choisy, C.: Dynamic handwritten keyword spotting based on the NSHP-HMM. In: *Proc. Int'l Conf. on Document Analysis and Recognition*, pp. 242–246. IEEE Computer Society, Washington, DC, USA (2007)
12. Curtis, J.D., Chen, E.: Keyword spotting via word shape recognition. In: *Proc. SPIE - Document Recognition II*, pp. 270–277 (1995)
13. Delalandre, M., Ogier, J.M., Lladós, J.: A fast cbr system of old ornamental letter. In: *Graphics Recognition. Recent Advances and New Opportunities, Lecture Notes in Computer Science*, vol. 5046, pp. 135–144. Springer-Verlag, Berlin, Heidelberg (2008)
14. Doermann, D., Doermann, D.: The indexing and retrieval of document images: A survey. *Computer Vision and Image Understanding* **70**, 287–298 (1998)
15. Fataicha, Y., Cheriet, M., Nie, Y., Suen, Y.: Retrieving poorly degraded ocr documents. *Int'l Journal of Document Analysis and Recognition* **8**(1), 1–99,999 (2006)
16. Fonseca, M.J., Ferreira, A., Jorge, J.A.: Generic shape classification for retrieval. In: *Proc. Int'l Workshop on Graphics Recognition*, pp. 291–299 (2005)
17. Gatos, B., Pratikakis, I.: Segmentation-free word spotting in historical printed documents. In: *Proc. Int'l Conf. on Document Analysis and Recognition*, p. 271. IEEE Computer Society (2009)
18. Gordo, A., Valveny, E.: A rotation invariant page layout descriptor for document classification and retrieval. In: *Proc. Int'l Conf. on Document Analysis and Recognition*, pp. 481–485. IEEE Computer Society (2009)
19. Govindaraju, V., Cao, H., Bhardwaj, A.: Handwritten document retrieval strategies. In: *Proc. of Workshop on Analytics for Noisy Unstructured Text Data*, pp. 3–7. ACM, New York, USA (2009)
20. Harris, Z.: Distributional structure. *Word* **10**(23), 146–162 (1954)
21. Jain, A.K., Namboodiri, A.M.: Indexing and retrieval of on-line handwritten documents. In: *Proc. Int'l Conf. on Document Analysis and Recognition*, p. 655. IEEE Computer Society, Washington, DC, USA (2003)

22. J.Hu, R.Kashi, G.Wilfong: Comparison and classification of documents based on layout similarity. *Information Retrieval* **2**(2/3), 227–243 (2000)
23. Jones, G., Foote, J., Sparck Jones, K., Young, S.: Video mail retrieval: the effect of word spotting accuracy on precision. In: *Int'l Conf. on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 309–312 vol.1 (1995)
24. Journet, N., Ramel, J.Y., Mullot, R., Eglin, V.: A proposition of retrieval tools for historical document images libraries. In: *Proc. Int'l Conf. on Document Analysis and Recognition*, pp. 1053–1057. IEEE Computer Society, Washington, DC, USA (2007)
25. Joutel, G., Eglin, V., Bres, S., Emptoz, H.: Curvelets based queries for CBIR application in handwriting collections. In: *Proc. Int'l Conf. on Document Analysis and Recognition*, pp. 649–653. IEEE Computer Society, Washington, DC, USA (2007)
26. Karray, A., Ogier, J.M., Kanoun, S., Alimi, M.A.: An ancient graphic documents indexing method based on spatial similarity. In: *Proc. Int'l Workshop on Graphics Recognition*, pp. 126–134. Springer-Verlag, Berlin, Heidelberg (2008)
27. Kesidis, A., Galiotou, E., Gatos, B., Lampropoulos, A., Pratikakis, I., Manolessou, I., Ralli, A.: Accessing the content of greek historical documents. In: *Proc. of Workshop on Analytics for Noisy Unstructured Text Data*, pp. 55–62. ACM, New York, USA (2009)
28. Khurshid, K., Faure, C., Vincent, N.: Fusion of word spotting and spatial information for figure caption retrieval in historical document images. In: *Proc. Int'l Conf. on Document Analysis and Recognition*, pp. 266–270. IEEE Computer Society (2009)
29. Kise, K., Wuotang, Y., Matsumoto, K.: Document image retrieval based on 2D density distributions of terms with pseudo relevance feedback. In: *Proc. Int'l Conf. on Document Analysis and Recognition*, pp. 488–492. IEEE Computer Society, Washington, DC, USA (2003)
30. Kogler, M., Lux, M.: Bag of visual words revisited: an exploratory study on robust image retrieval exploiting fuzzy codebooks. In: *Proc. Int'l Workshop on Multimedia Data Mining, MDMKDD '10*, pp. 3:1–3:6. ACM, New York, USA (2010)
31. Konidakis, T., Gatos, B., Ntzios, K., Pratikakis, I., Theodoridis, S., Perantonis, S.J.: Keyword-guided word spotting in historical printed documents using synthetic data and user feedback. *Int'l Journal of Document Analysis and Recognition* **9**(2), 167–177 (2007)
32. Latecki, L.J., Lakämper, R., Eckhardt, U.: Shape descriptors for non-rigid shapes with a single closed contour. In: *IEEE Computer Society Conf. in Computer Vision and Pattern Recognition*, pp. 424–429 (2000)
33. Leydier, Y., Lebourgeois, F., Emptoz, H.: Text search for medieval manuscript images. *Journal of Pattern Recognition* **40**(12), 3552–3567 (2007)
34. Li, L., Lu, S.J., Tan, C.L.: A fast keyword-spotting technique. In: *Proc. Int'l Conf. on Document Analysis and Recognition*, pp. 68–72. IEEE Computer Society, Washington, DC, USA (2007)
35. Liang, S., Sun, Z.: Sketch retrieval and relevance feedback with biased SVM classification. *Pattern Recognition Letters* **29**(12), 1733–1741 (2008)
36. Licata, A., Psarrou, A., Kokla, V.: Revealing the visually unknown in ancient manuscripts with a similarity measure for IR-imaged inks. In: *Proc. Int'l Conf. on Document Analysis and Recognition*, pp. 818–822. IEEE Computer Society (2009)
37. Lladós, J., Sánchez, G.: Indexing historical documents by word shape signatures. In: *Proc. Int'l Conf. on Document Analysis and Recognition*, pp. 362–366. IEEE Computer Society, Washington, DC, USA (2007)
38. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision* **60**, 91–110 (2004)
39. Lu, S., Li, L., Tan, C.L.: Document image retrieval through word shape coding. *IEEE Trans. on PAMI* **30**(11), 1913–1918 (2008)
40. Lu, S., Tan, C.: Keyword spotting and retrieval of document images captured by a digital camera. In: *Proc. Int'l Conf. on Document Analysis and Recognition*, pp. 994–998. IEEE Computer Society, Washington, DC, USA (2007)
41. Lu, S., Tan, C.L.: Retrieval of machine-printed latin documents through word shape coding. *Journal of Pattern Recognition* **41**(5), 1816–1826 (2008)

42. Lu, Y., Zhang, L., Tan, C.L.: Retrieving imaged documents in digital libraries based on word image coding. In: Proc. Int'l Workshop on Document Image Analysis for Libraries, pp. 174–187. IEEE Computer Society, Washington, DC, USA (2004)
43. Manmatha, R., Han, C., Riseman, E.M.: Word spotting: A new approach to indexing handwriting. In: IEEE Computer Society Conf. in Computer Vision and Pattern Recognition, vol. 0, pp. 631–637. IEEE Computer Society, Los Alamitos, CA, USA (1996)
44. Marinai, S.: A Survey of Document Image Retrieval in Digital Libraries. In: Laurence Likforman Sulem (ed.) Actes du 9ème Colloque International Francophone sur l'Ecrit et le Document, pp. 193–198. SDN06 (2006)
45. Marinai, S.: Text retrieval from early printed books. In: Proc. of Workshop on Analytics for Noisy Unstructured Text Data, pp. 33–40. ACM, New York, USA (2009)
46. Marinai, S., Gori, M., Soda, G.: Artificial neural networks for document analysis and recognition. IEEE Trans. on PAMI **27**(1), 23–35 (2005)
47. Marinai, S., Marino, E., Soda, G.: Font adaptive word indexing of modern printed documents. IEEE Trans. on PAMI **28**(8) (2006)
48. Marinai, S., Miotti, B., Soda, G.: Mathematical symbol indexing using topologically ordered clusters of shape contexts. In: Proc. Int'l Conf. on Document Analysis and Recognition, pp. 1041–1045 (2009)
49. Marinai, S., Miotti, B., Soda, G.: Bag of characters and SOM clustering for script recognition and writer identification. In: Proc. Int'l Conf. on Pattern Recognition, pp. 2182–2185 (2010)
50. Meshesha, M., Jawahar, C.V.: Matching word images for content-based retrieval from printed document images. Int'l Journal of Document Analysis and Recognition **11**(1), 29–38 (2008)
51. Mitra, M., Chaudhuri, B.: Information retrieval from documents: A survey. Information Retrieval **2**(2/3), 141–163 (2000)
52. Moghaddam, R., Cheriet, M.: Application of multi-level classifiers and clustering for automatic word spotting in historical document images. In: Proc. Int'l Conf. on Document Analysis and Recognition, pp. 511–515. IEEE Computer Society (2009)
53. Nakai, T., Kise, K., Iwamura, M.: Real-time retrieval for images of documents in various languages using a web camera. In: Proc. Int'l Conf. on Document Analysis and Recognition, pp. 146–150. IEEE Computer Society (2009)
54. Nguyen, T.O., Tabbone, S., Terrades, O.R.: Symbol descriptor based on shape context and vector model of information retrieval. In: Proc. IAPR Int'l Workshop on Document Analysis Systems, pp. 191–197. IEEE Computer Society, Washington, DC, USA (2008)
55. Perronnin, F.: Universal and adapted vocabularies for generic visual categorization. IEEE Trans. on PAMI **30**(7), 1243–1256 (2008)
56. Qureshi, R.J., Ramel, J.Y., Barret, D., Cardot, H.: Spotting symbols in line drawing images using graph representations. In: Proc. Int'l Workshop on Graphics Recognition, pp. 91–103. Springer-Verlag, Berlin, Heidelberg (2008)
57. Rath, T.M., Manmatha, R.: Features for word spotting in historical manuscripts. In: Proc. Int'l Conf. on Document Analysis and Recognition, pp. 218–222. IEEE Computer Society, Washington, DC, USA (2003)
58. Rath, T.M., Manmatha, R.: Word spotting for historical documents. Int'l Journal of Document Analysis and Recognition **9**(2), 139–152 (2007)
59. Rodriguez, J.A., Perronnin, F.: Local gradient histogram features for word spotting in unconstrained handwritten documents. In: Proc. Int'l Conf. on Handwriting Recognition (2008)
60. Rodriguez-Serrano, J., Perronnin, F.: Handwritten word-image retrieval with synthesized typed queries. In: Proc. Int'l Conf. on Document Analysis and Recognition, pp. 351–355. IEEE Computer Society (2009)
61. Rusiñol, M., Lladós, J.: Symbol spotting in technical drawings using vectorial signatures. In: Proc. Int'l Workshop on Graphics Recognition, pp. 35–46 (2005)
62. Rusiñol, M., Lladós, J.: A region-based hashing approach for symbol spotting in technical documents. In: Proc. Int'l Workshop on Graphics Recognition, pp. 104–113. Springer-Verlag, Berlin, Heidelberg (2008)

63. Rusiñol, M., Lladós, J.: Word and symbol spotting using spatial organization of local descriptors. In: Proc. IAPR Int'l Workshop on Document Analysis Systems, pp. 489–496. IEEE Computer Society, Washington, DC, USA (2008)
64. Rusiñol, M., Lladós, J.: *Symbol Spotting in Digital Libraries: Focused Retrieval over Graphic-rich Document Collections*. Springer Publishing Company, Incorporated (2010)
65. Salton, G., Wong, A., Yang, C.S.: A vector space model for automatic indexing. *Commun. ACM* **18**, 613–620 (1975)
66. Schomaker, L.: Retrieval of handwritten lines in historical documents. In: Proc. Int'l Conf. on Document Analysis and Recognition, vol. 2, pp. 594–598 (2007)
67. Sivic, J., Zisserman, A.: Video Google: A text retrieval approach to object matching in videos. In: Proc. Int'l Conf. on Computer Vision, vol. 2, p. 1470. IEEE Computer Society, Los Alamitos, CA, USA (2003)
68. Smeaton, A.F., Spitz, A.L.: Using character shape coding for information retrieval. In: Proc. Int'l Conf. on Document Analysis and Recognition, pp. 974 – 978 (1997)
69. Super, B.J.: Retrieval from shape databases using chance probability functions and fixed correspondence. *Int'l Journal of Pattern Recognition and Artificial Intelligence* **20**(8), 1117–1138 (2006)
70. Tahmasebi, N., Niklas, K., Theuerkauf, T., Risse, T.: Using word sense discrimination on historic document collections. In: Proc. Joint Conf. on Digital Libraries, pp. 89–98. ACM, New York, USA (2010)
71. Tan, G., Viard-Gaudin, C., Kot, A.: Information retrieval model for online handwritten script identification. In: Proc. Int'l Conf. on Document Analysis and Recognition, pp. 336–340. IEEE Computer Society (2009)
72. Terasawa, K., Nagasaki, T., Kawashima, T.: Eigenspace method for text retrieval in historical documents. In: Proc. Int'l Conf. on Document Analysis and Recognition, pp. 437–441 (2005)
73. Tzacheva, A., El-Sonbaty, Y., El-Kwae, E.A.: Document image matching using a maximal grid approach. In: Proc. SPIE Document Recognition and Retrieval IX, pp. 121–128 (2002)
74. Uttama, S., Loonis, P., Delalandre, M., Ogier, J.M.: Segmentation and retrieval of ancient graphic documents. In: Proc. Int'l Workshop on Graphics Recognition, pp. 88–98 (2005)
75. Wan, G., Liu, Z.: Content-based information retrieval and digital libraries. *Information Technology & Libraries* **27**, 41–47 (2008)
76. Waters, D., Garrett, J.: Preserving digital information. report of the task force on archiving of digital information. Tech. rep., The Commission on Preservation and Access (1996)
77. Wei, C.H., Li, Y., Chau, W.Y., Li, C.T.: Trademark image retrieval using synthetic features for describing global shape and interior structure. *Journal of Pattern Recognition* **42**(3), 386–394 (2009)
78. Witten, I.H., Bainbridge, D.: *How to Build a Digital Library*. Elsevier Science Inc., New York, USA (2002)
79. Wong, W.T., Shih, F.Y., Su, T.F.: Shape-based image retrieval using two-level similarity measures. *Int'l Journal of Pattern Recognition and Artificial Intelligence* **21**(6), 995–1015 (2007)
80. Zhang, B., Srihari, S., Huang, C.: Word image retrieval using binary features. In: SPIE, Document Recognition and Retrieval XI, pp. 45–53 (2004)
81. Zhang, W., Liu, W.: A new vectorial signature for quick symbol indexing, filtering and recognition. In: Proc. Int'l Conf. on Document Analysis and Recognition, pp. 536–540. IEEE Computer Society, Washington, DC, USA (2007)
82. Zhang, Z., Jin, L., Ding, K., Gao, X.: Character-sift: a novel feature for offline handwritten chinese character recognition. In: Proc. Int'l Conf. on Document Analysis and Recognition, pp. 763–767. IEEE Computer Society (2009)