A Multimodal Approach to Relevance and Pertinence of Documents

Matteo Cristani(✉) and Claudio Tomazzoli(✉)

University of Verona, Verona, Italy
{matteo.cristani,claudio.tomazzoli}@univr.it

Abstract. Automated document classification process extracts information with a systematical analysis of the content of documents. This is an active research field of growing importance due to the large amount of electronic documents produced in the world wide web and made readily available thanks to diffused technologies including mobile ones. Several application areas benefit from automated document classification, including document archiving, invoice processing in business environments, press releases and search engines. Current tools classify or "tag" either text or images separately. In this paper we show how, by linking image and text-based contents together, a technology improves fundamental document management tasks like retrieving information from a database or automatically routing documents. We present a formal definition of pertinence and relevance concepts, that apply to those documents types we name "multimodal". These are based on a model of conceptual spaces we believe compulsory for document investigation while using joint information sources coming from text and images forming complex documents.

1 Introduction

Nowadays the wide availability of electronic documents through the Internet or private business networks has changed the way people search for information. We deal with a huge quantity of knowledge which has to be organized and searchable to be utilized. Also for this reason in information technology research community there is an always growing interest in the field of automatic document classification. Although several innovative studies are produced every year, some topics are still to be deeply investigated. Among these, the problem of efficient classification and retrieval of documents containing both text and images has been treated in a non multidisciplinary approach. There are several publications of efficient information retrieval from text. There are also publications about information extraction from images and even text contained in images [1], but the joint analysis of text and image information from a complex document still lacks a well documented solution. For example, if a brochure from an isolated hotel in the Dolomites describes the hotel's features and includes maps and pictures of mountainous surroundings, the categorizer will automatically discover the content and link the text and the images together. Then someone searching

H. Fujita et al. (Eds.): IEA/AIE 2016, LNAI 9799, pp. 157–168, 2016.
DOI: 10.1007/978-3-319-42007-3_14

for an isolated mountain lodge within a certain price range would retrieve the brochure even if "isolated lodge in the mountains" were never mentioned in the actual text. The paper is organized as follows: Sect. 2 presents the areas in which automatic document classification is relevant; Sect. 3 summaries the main approaches existing in current literature; Sect. 4 presents the model and the approach of extracting joint textual and image information; finally in Sect. 5 we give the formal definition of the model, of *Pertinence* and *Relevance* and make some conclusions.

2 Motivations

Automatic document classification is an interesting process for a wide variety of application areas, due to the huge amount of electronic documents in which is stored the information a user can search for. Among these there are Web Mining, Press Survey, Scientific Research, Image Indexing.

Press Survey

Press Survey is the task of retrieving what has been "printed" and diffused on the mass media, usually newspapers, about a particular topic.
Politicians are interested in knowing who is writing about them or about a particular subject they are interested in. Firms are interested in knowing how the Press responded to a particular marketing event or a new product release. Most of this work is performed by humans, who scan the several sources for relevant information. As the Press are going to deploy on the Internet their former printed daily or weekly magazines, we can consider the sources of information to be digital, thus eligible for automatic elaboration. Due to the visual impact of images, articles are very often equipped with pictures which add informative content to the article itself. Articles are an example of documents in which textual and visual information are related and concur to form the meaning of the work. Therefore, a classifier able to use the joint information of both text and images can build a good tool in constructing collection of articles related to a specific topic, leading to more accurate and efficient surveys (Fig. 1).

3 State of the Art

During this research we develop a model that had significant previous references. In particular we employed techniques used in Text and Image Mining.

3.1 Text Mining

Text mining is about inferring structure from sequences representing natural language text, and may be defined as the process of analyzing text to extract information that is useful for particular purposes, such as extraction of hierarchical phrase structures from text, identification of keyphrases in documents,

Fig. 1. Magazine information is contained in both text and image

locating proper names and quantities of interest in a piece of text, text categorization, word segmentation, acronym extraction, and structure recognition. There are several text mining task; among the most frequently used are *Supervised Learning* (or Classification), *Unsupervised Learning* (or Clustering) and *Probabilistic Latent Semantic Analysis*.

3.2 Image Mining

Image search is traditionally obtained mainly through relational database search of caption or keywords [2]; the automatic classification is often achieved using content-based image retrieval (CBIR) systems [3]; in this topic research focus is divided between low-level (or visual) feature extraction algorithms and high-level (or textual) feature extraction, the latter used to reduce the so called 'semantic gap' between the visual features and the richness of human semantics. We identify five major categories of the state-of-the-art techniques in narrowing down the 'semantic gap' [4]:

(1) using object ontology to define high-level concepts;
(2) using machine learning methods to associate low-level features with query concepts;
(3) using relevance feedback to learn users' intention;
(4) generating semantic template to support high-level image retrieval;
(5) fusing the evidences from HTML text and the visual content of images for WWW image retrieval.

There are low-level features extraction algorithms which make use of text mining techniques above explained. Features like color, texture, shape, spatial relationship among entities of an image and also their combination are used for the computation of multidimensional feature vector [5]; *color*, *texture* and *shape* are known as primitive features. *Color* and *texture* are used as a base for image detection and classification using a support vector machine (SVM), where color is represented using HSV (hue, saturation, value) color model because this model

is closely related to human visual perception and texture is computed using the entropy of rectangular regions of the image in [6]. According to [7] *shapes* can be described textually using parts, junction line and disjunction line using XML language for writing descriptors of outline shapes. Thus, we can build a method for shape comparison and similarity measure which is computed directly from the textual descriptor.

There are high-level features extraction algorithms which make use of text close to the image. Text-based image retrieval (TBIR) first labels the images in the database according to text close to the image and then uses the database management system to perform image retrieval based on those labels [8], sometimes taking into account the extent to which a word can be perceived visually in images [9] exploiting a self-organizing neural network architecture [10] to extract labels or combining high and low level features [11], or using an ontology model that integrates both these information [12]. Other techniques make use of the 'bags of visual words' model, having images as documents, and categories as topics (for example, grass and houses) so that an image containing instances of several objects is modeled as a mixture of topics [13] or define a scene categorization method based on contextual visual words, and introducing contextual information from the coarser scale and neighborhood regions to the local region of interest based on unsupervised learning [14]. Images are classified through the surrounding text also with statistical methods, such as $TFIDF$ [15]: For a single piece of text, a word's *term frequency (TF)*, is the number of times that this word occurs in that text. For a category (such as all indoor images), the TF assigned to a word is the number of times that word occurs in all documents of that category. A word's *inverse document frequency (IDF)*, is the logarithm of the ratio of the total number of documents to the word's *document frequency (DF)*, which is the number of documents that contain that word; this measure remains constant independently of the particular document or category examined. There is also a wide documentation about the task of *Text Extraction from Images* in which images containing text are analyzed to automatically extract the included text [16], having to deskews the image, extracts text regions, segments text regions into text lines [17] or differentiating between region of text, graphics and background, using a neuro-fuzzy methodology [18], finally using local energy analysis for segmenting text [19] or Support Vector Machine [20]. The visual appearance of a document can be used as a feature to achieve clustering [21] where a statistical approach is used to characterize typical texture patterns in document images.

3.3 Text and Document Joint Information Retrieval

In [22] is proposed a method to learn the relationships between images and the surrounding text. For an image, a term in the description may relate to a portion of an image. If we divide an image into several smaller parts, called blocks or regions, we could link the terms to the corresponding parts. This is analogous to word alignment in a sentence aligned parallel corpus. Here the word alignment is replaced with the textual-term/visual-term alignment. If we treat the visual

representation of an image as a language, the textual description and visual parts of an image are an aligned sentence. The correlations between the vocabularies of two languages can be learned from the aligned sentences. First, images are segmented into regions using a segmentation algorithm (in [22] "Blobword" is used).

Finally, in [23] we face a paper which deals with document similarity extracting both textual and visual information, which are called "mode" of a document, so that the authors refer to them as "multimodal" documents. Image similarity is computed using a "bag of visual word" representation (Fisher vector) in which the visual vocabulary is obtained with a Gaussian mixture model (GMM) which approximates the distribution of the low-level features in images. The similarity measure between two images is then defined as the L1-norm of the difference between the normalized Fisher Vectors of the two images. Text similarity is computed with text being pre-processed including tokenization, lemmatization, word decompounding and standard stop-word removal. The authors in [23] define a global similarity measure between two multi-modal objects d and d_q using, for instance, a linear combination:

$$sim_{glob}(d, d_q) = \lambda_1 sim_{TT}(d, d_q) + \lambda_2 sim_{TV}(d, d_q) + \lambda_3 sim_{VT}(d, d_q) + \lambda_4 sim_{VV}(d, d_q)$$

4 The Model

We found the model described in [23] as a valuable starting point for our model; we will use accordingly the term "mode" of a document for both text and image and we will use the "bag of word" representation for features set of both modes, but we define those contributes in a more general sense than in [23]; we showed in [24] that the model has solid experimental ground truth and leads to computable algorithms. Then we apply "noise" and we define our general model, which will be used later in the framework.

4.1 Latent Semantic Analisys

The problem of classification can be considered the problem of properly attach tags (class names) to documents. Suppose we have n *documents* $\mathcal{D} = \{d_1, d_2, ..., d_n\}$ and m *tags* $\mathcal{T} = \{t_1, t_2, ..., t_m\}$.

The links between these n *documents* and the m *tags* are denoted by a $n \times m$ matrix A. The elements $A_{i,j} \in \mathbb{R}^{n \times m}$ of this matrix represent the *weight* of link, e.g., $A_{i,j} = 1$ if jth tag is assigned to ith document, or $A_{i,j} = 0$ otherwise. The goal is to construct a set of feature vectors $\{X_1, X_2, \ldots, X_n\}$ in a latent semantic space $\mathbb{R}^k$ to represent these multimedia objects in the form $A = U\Sigma V^T$. Here, U and V are orthogonal matrices such that $UU^T = VV^T = I$, and the diagonal matrix Σ has the singular values as its diagonal elements. By retaining the largest k singular values in Σ and approximating others to be zero, we can create an approximated diagonal matrix Σ^k with fewer singular values.

This diagonal matrix is used to approximate Σ as $A \cong U\Sigma^k V^T$. Then the matrix $X = U\Sigma^k$, $X \in \mathbb{R}^{n \times k}$ yields a new feature representation, each row of which is a k-dimensional feature vector of a document, $X = [X_1, X_2, \ldots, X_n]^T$.

4.2 The Model for Multimodal Documents

We are considering both textual and visual contributions to the meaning of a document. Details of this model and its motivations can be found in [24].

Suppose we are given a matrix Q of content links, where $Q_{i,j}$ can represent the similarity measurement between the ith document and the jth document. Recalling the works in latest literature [23] we have that documents can be described as *multimodal* when made of both text and visual content, each defined as "mode"; a repository that contains a set of multimodal documents is then $D = \{d_1, d_2, \ldots, d_n\}$ (Fig. 2).

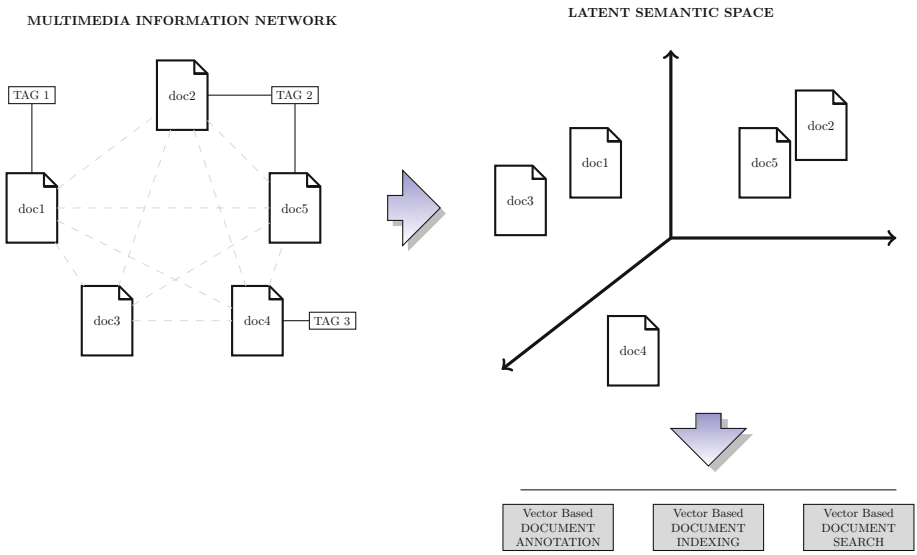

Fig. 2. Model for multimodal documents

We can define a global similarity measure between two multi-modal objects d and d_q using, for instance, a linear combination as in [23]:

$$sim_{glob}(d_i, d_j) = \lambda_1 sim_{TT}(d_i, d_j) + \lambda_2 sim_{TV}(d_i, d_j) + \lambda_3 sim_{VT}(d_i, d_j) + \lambda_4 sim_{VV}(d_i, d_j)$$

so we have that the elements in our matrix Q of similarity of multimodal content of documents can be

$$Q_{i,j} = \lambda_1 sim_{TT}(d_i, d_j) + \lambda_2 sim_{TV}(d_i, d_j) + \lambda_3 sim_{VT}(d_i, d_j) + \lambda_4 sim_{VV}(d_i, d_j) \quad (1)$$

We assume that the documents with stronger links ought to be closer to each other in the latent semantic space. Based on this assumption, we introduce the quantity Ω to measure the smoothness of documents in the underlying latent space.

$$\Omega(X) = \frac{1}{2} \sum_{i,j=1}^{n} Q_{i,j} \|X_i - X_j\|_2^2 = \frac{1}{2} \sum_{i,j=1}^{n} Q_{i,j} (X_i - X_j)(X_i - X_j)^T \quad (2)$$

where, $\|M\|_2^2$ is the l_2 norm of matrix M, and X_i and X_j are the ith and jth row of X. It is easy to see that by minimizing the above regularization term, a pair of documents with larger $Q_{i,j}$ will have closer feature vectors X_i and X_j in the latent space (Fig. 3).

Given D as the diagonal matrix with its elements as the sum of each row of Q and $L = D - Q$, with some matrix operations we obtain

$$\Omega(X) = trace(X^T L X) \tag{3}$$

using the factorization $X = U\Sigma^k$, defining H as $H = U\Sigma_k V^T = XV^T$ and knowing that $VV^T = I$ we have

$$\Omega(X) = trace(H^T L H) \tag{4}$$

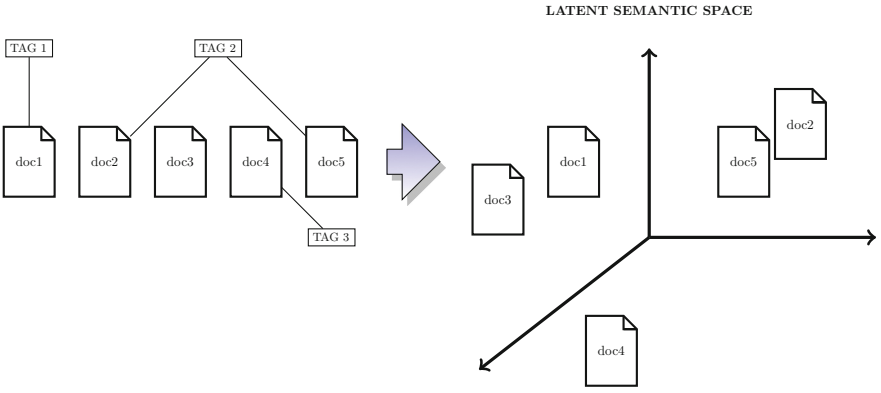

Fig. 3. Documents with stronger links will be closer

4.3 The Noisy Model

Due to the fact that we consider both textual and visual contribution to the meaning of a document, we have to consider the existence of noise in process so a noise term ε exist on the matrix Q such that $Q = H + \varepsilon$ where H is the matrix which denotes the noise-free tag links, after the noise ε has been removed. The goal is to make a correctly representative H of 'minimal rank'. The problem, as shown in [24] can be solved using the *nuclear norm* of a matrix M ($\|M\|_*$)

$$min\|Q - H\|_F + \gamma\|H\|_* \tag{5}$$

where $\|M\|_F$ is the squared summation of all elements in a matrix M (the Frobenius norm) and γ is a balancing parameter. Always in [24], a consistent solution to the problem is found to be

$$min\|Q - H\|_F + \gamma\|H\|_* = H_\gamma = U\Sigma_+^\gamma V^T \tag{6}$$

The difference with normal *Latent Semantic Indexing* is that it directly selects the largest k singular values of A where this Formulation subtracts something $(\frac{\gamma}{2})$ from each singular value and thresholds them by 0. Suppose the resulting noise free matrix H is of rank k, then the Support Vector Machine of H has form as $H = U\Sigma_k V^T$ where Σ_k is a $k \times k$ diagonal matrix. Similar with *Latent Semantic Indexing*, the row vectors of $X = U\Sigma_k$ can be used as the latent vector representations of documents in latent space. It is also worth noting that minimizing the rank of H gives a smaller k so that the obtained latent vector space can have lower dimensionality, and then the storage and computation in this space could be more efficient.

4.4 The Global Model for Multimodal Documents

Considering both contribution to the model we can make use of both Eqs. 4 and 5 so our problem can be completely described as finding

$$min\|Q - H\|_F + \gamma\|H\|_* + \lambda trace(H^T LH) \tag{7}$$

Here λ is another balancing parameter. In contrast to Formulation (4), Formulation (7) does not have an closed-form solution. Fortunately, this problem can be solved by the *Proximal Gradient method* known from literature which uses a sequence of quadratic approximations of the objective function in order to derive the optimal solution.

5 The Framework

5.1 The Matrix Q of Similarity for Multimodal Content

We have considered in [24] both textual and visual contributions to the meaning of a document. We defined matrix Q_t of content links, where $Q_t(i, j)$ can represent the similarity measurement between the text of the ith document and the text of the jth document. We defined matrix Q_v of content links, where $Q_v(i, j)$ can represent the similarity measurement between the image of the ith document and the image of the jth document. Following PLSA approach as above specified we have, respectively for textual and visual mode

$$Q_t \cong U_t \Sigma_t^k V_t^T \qquad Q_v \cong U_v \Sigma_v^k V_v^T \tag{8}$$

We have similarly, the dual representation of the visual and textual part as:

$$S_T = U_t \Sigma_t^k \qquad S_V = U_v \Sigma_v^k \tag{9}$$

We denote the textual part of d_j by $S_T(d_j)$ and its visual part $S_V(d_j)$ which are the jth rows of matrix S_T and S_V. Recalling that in our model in Eq. 1

$$Q_{i,j} = \lambda_1 sim_{TT}(d_i, d_j) + \lambda_2 sim_{TV}(d_i, d_j) + \lambda_3 sim_{VT}(d_i, d_j) + \lambda_4 sim_{VV}(d_i, d_j)$$

and *assuming that the both text and image part of a document shall define the same meaning for the document in the meaning space* we will use these partial latent semantic representations to define the single components of the equation above

$$sim_{TV}(d_i, d_j) = \|S_T(d_i) - S_V(d_j)\|_F \tag{10}$$

$$sim_{VT}(d_i, d_j) = \|S_V(d_i) - S_T(d_j)\|_F \tag{11}$$

$$sim_{TT}(d_i, d_j) = \|S_T(d_i) - S_T(d_j)\|_F \tag{12}$$

$$sim_{VV}(d_i, d_j) = \|S_V(d_i) - S_V(d_j)\|_F \tag{13}$$

This model benefits from two major aspects: it is simple to understand and it is simple to implement, both because it involves only measure of distance in a vector space. The main assumption is that there is *one* meaning space so that features in text and features in images all refers to a set of concepts or meanings which are the same but are expressed with words and with images.

When these meanings are expressed with words the dimensionality of the feature space is different than the dimensionality of the feature space coming from the images, but using a dimensionality reduction algorithm we can reduce these different dimensions to be the same, so that we could compute a distance. Experiments performed with a knowledge base of almost a million newspaper articles shows [24] that model and framework holds.

5.2 Pertinence and Relevance

We have that $H = U_H \Sigma_H^k V_H^T$ and $X = U_H \Sigma_H^k$ will be the our full latent vector representations of documents in latent space.

Definition 1. *We define the Pertinence of the text in a document informally as how near is the meaning of the text to the meaning of the whole document. This leads to the definition of a distance which in our vector space is*

$$P_T(d_i) = \|X(d_i) - S_T(d_i)\|_F \tag{14}$$

This definition can be used for other modes of a document, so for the image the *Pertinence* of the image in a document is how near is the meaning of the image to the meaning of the whole document, so we have

$$P_V(d_i) = \|X(d_i) - S_V(d_i)\|_F \tag{15}$$

Definition 2. *We define the Relevance of the text in a document informally as how important is the meaning of the text in defining the meaning of the whole document. This also leads to the definition of a distance in our vector space as*

$$R_T(d_i) = \|\frac{S_T(d_i)}{X(d_i)}\|_F \tag{16}$$

Fig. 4. Example of pertinent and not relevant

and as above for the image the *Relevance* of the image in a document is how important is the meaning of the image in defining the meaning of the whole document, so we have

$$R_V(d_i) = \| \frac{S_V(d_i)}{X(d_i)} \|_F \tag{17}$$

These definitions are sound with the fact that a meaning might be pertinent but not relevant or not pertinent and not relevant to a document but the same meaning can not be relevant and not pertinent, which reflects everyday life experience. These definitions are simple to understand and to implement, mainly because they follow the model above which is both simple and computable. In Fig. 4 we have and example with the concept of 'panda': the image of the car *Fiat Panda* is not pertinent and not relevant while *WWF* contribute is pertinent but not relevant whereas the article of the family of bears is both pertinent and relevant in defining the meaning of the document.

6 Conclusions and Further Work

The first part of this work was dedicated to point out the overview of the research and the problems and choices we got through during the path of this research. Then we focused on the model we would use to determine different contribution to classification of the text and image information of a document; we've given the details of the definition of a meaning space using Latent Semantics for multimodal documents including consideration and modeling of the possible noise that shall be considered in this process and how to deal with it. Then we focused on the definitions of the distances in the meaning space and we've given

the formal definition of *Persistence* and *Relevance* which will lead to a computable algorithm for our model, which will then enable a better understanding of semantic gap between the different parts, or "modes" of a document. This can be extended also to other kind of multimodal documents, such as videos, which have a spoken (i.e. text) and visual parts and the correlation with time can be explored as further research.

References

1. Ye, Q., Huang, Q., Gao, W., Zhao, D.: Fast and robust text detection in images and video frames. Image Vis. Comput. **23**(6), 565–576 (2005)
2. Kahn, C.: Dynamic inline images: context-sensitive retrieval and integration of images into web documents. J. Digit. Imaging **21**(3), 274–279 (2008)
3. Park, G., Baek, Y., Lee, H.-K.: Web image retrieval using majority-based ranking approach. Multimed. Tools Appl. **31**(2), 195–219 (2006)
4. Liu, Y., Zhang, D., Guojun, L., Ma, W.-Y.: A survey of content-based image retrieval with high-level semantics. Pattern Recogn. **40**(1), 262–282 (2007)
5. Schettini, R., Brambilla, C., Ciocca, G., Valsasna, A., De Ponti, M.: A hierarchical classification strategy for digital documents. Pattern Recogn. **35**(8), 1759–1769 (2002)
6. Seo, K.-K.: An application of one-class support vector machines in content-based image retrieval. Expert Syst. Appl. **33**(2), 491–498 (2007)
7. Larabi, S.: Textual description of shapes. J. Vis. Commun. Image Represent. **20**(8), 563–584 (2009)
8. Sagara, N., Sunayama, W., Yachida, M.: Image labeling using key sentences of HTML. Electron. Commun. Jpn. (Part III Fundam. Electron. Sci.) **89**(7), 31–41 (2006)
9. Fei, W., Han, Y.-H., Zhuang, Y.-T.: Multiple hypergraph clustering of web images by MiningWord2Image correlations. J. Comput. Sci. Technol. **25**(4), 750–760 (2010)
10. de Mello, R.F., Bueno, J.M., Senger, L.J., Yang, L.T.: Image indexing and retrieval using an ART-2A neural network architecture. Int. J. Imaging Syst. Technol. **18**(2–3), 202–208 (2008)
11. Shen, H.T., Zhou, X., Cui, B.: Indexing and integrating multiple features for www images. World Wide Web **9**(3), 343–364 (2006)
12. Wang, H., Liu, S., Chia, L.-T.: Image retrieval with a multi-modality ontology. Multimed. Syst. **13**(5), 379–390 (2008)
13. Bosch, A., Zisserman, A., Munoz, X.: Scene classification using a hybrid generative/discriminative approach. IEEE Trans. Pattern Anal. Mach. Intell. **30**(4), 712–727 (2008)
14. Qin, J., Yung, N.H.C.: Scene categorization via contextual visual words. Pattern Recogn. **43**(5), 1874–1888 (2010)
15. Sable, C.L., Hatzivassiloglou, V.: Text-based approaches for non-topical image categorization. Int. J. Digit. Libr. **3**(3), 261–275 (2000)
16. Zhao, M., Li, S., Kwok, J.: Text detection in images using sparse representation with discriminative dictionaries. Image Vis. Comput. **28**(12), 1590–1599 (2010)
17. Srihari, S.N., Tao, H., Geetha, S.: Machine-printed Japanese document recognition. Pattern Recogn. **30**(8), 1301–1313 (1997)

18. Caponetti, L., Castiello, C., Gorecki, P.: Document page segmentation using neuro-fuzzy approach. Appl. Soft Comput. J. **8**(1), 118–126 (2008)
19. Chan, W., Coghill, G.: Text analysis using local energy. Pattern Recogn. **34**(12), 2523–2532 (2001)
20. Chang, Y., Chen, D., Zhang, Y., Yang, J.: An image-based automatic arabic translation system. Pattern Recogn. **42**(9), 2127–2134 (2009)
21. Wen, D., Ding, X.-Q.: Visual similarity based document layout analysis. J. Comput. Sci. Technol. **21**(3), 459–465 (2006)
22. Lin, W.-C., Chang, Y.-C., Chen, H.-H.: Integrating textual and visual information for cross-language image retrieval: a trans-media dictionary approach. Inf. Process. Manage. **43**(2), 488–502 (2007)
23. Ah-Pine, J., Bressan, M., Clinchant, S., Csurka, G., Hoppenot, Y., Renders, J.-M.: Crossing textual and visual content in different application scenarios. Multimed. Tools Appl. **42**(1), 31–56 (2009)
24. Cristani, M., Tomazzoli, C.: A multimodal approach to exploit similarity in documents. In: Ali, M., Pan, J.-S., Chen, S.-M., Horng, M.-F. (eds.) IEA/AIE 2014, Part I. LNCS, vol. 8481, pp. 490–499. Springer, Heidelberg (2014)