Towards Semantic Image Retrieval Using Multimodal Fusion with Association Rules Mining

Raniah A. Alghamdi and Mounira Taileb

Faculty of Computing and Information Technology,
King Abdulaziz University,
Saudi Arabia, Jeddah
{ragalghamdi,mtaileb}@kau.edu.sa

Abstract. This paper proposes a semantic retrieving method for an image retrieval system that employs the fusion of the textual and visual information of the image dataset which is a recent trend in image retrieval researches. It combines two different data mining techniques to retrieve semantically related images: clustering and association rule mining algorithm. At the offline phase of the method, the association rules are discovered between the text semantic clusters and the visual clusters to use it later in the online phase. To evaluate the proposed system, the experiment was conducted on more than 54,500 images of ImageCLEF 2011 Wikipedia collection. The proposed retrieval system was compared to an online system called MMRetrieval and to the proposed system but without using association rules. The obtained results show that our proposed method achieved the best precision and mean average precision.

Keywords: Image Retrieval, Multimodal Fusion, Association Rules Mining, Clustering.

1 Introduction

Today, a huge amount of images exists in electronic formats on the Web and in different information repositories; and their size is exponentially growing day after another. Thus, we need for an efficient Image Retrieval system (IR) to get access to these images. IR could rely purely on textual metadata which may produce a lot of garbage in the results since users usually enter that metadata manually which is inefficient, expensive and may not capture every keyword that describes the image. On the other hand, the Content-Based Image Retrieval (CBIR) could be used to filter images based on their visual contents such as colors, shapes, textures or any other information that can be derived from the image itself which may provide better indexing and return more accurate results. At the same time, these visual features contents extracted by the computer may be different from the image contents that people understand. It requires the translation of high-level user views into low-level image features and this is the so-called "semantic gap" problem. This problem is the reason behind why the current CBIR systems are difficult to be widely used for

S. Yamamoto (Ed.): HIMI 2014, Part I, LNCS 8521, pp. 407–418, 2014.

retrieving Web images. A lot of efforts have been made to bridge this gap by using different techniques. In [1], the authors identified the major categories of the state-of-the-art techniques in narrowing down the 'semantic gap' one of them is to fuse the evidences from the text and the visual content of the images. Fusion in IR is considered as a novel area, with very little achievements in the early days of research [2]. Truly, we live in a multimodal world, and there is no reason why advantage should not be taken of all available media to build a useful semantic IR system. This paper tries to narrow down this gap and enhance the retrieval performance by fusing the two basic modalities: text and visual features. To determine the appropriate fusion method, it is important to answer the following basic questions: what is the suitable level to implement the fusion strategy? And how to fuse the multimodal information?

The proposed method is a Multimodal Fusion method based on Association Rules mining (MFAR). It is considered as a late fusion. This method combines two different data mining techniques: clustering and Association Rules Mining (ARM) algorithm. It uses ARM to explore the relations between text semantic clusters and image visual features clusters by applying *Apriori* algorithm. The method consists of two main phases: offline and online. The offline phase identifies the relations among the clusters from different modalities to construct the semantic Association Rules (ARs). On the other hand, the online phase is the retrieval phase. It uses the generated ARM to retrieve the related images to the query.

The rest of the paper is categorized as following. The next section will review the current information fusion approaches and how they fused different modalities. Section three gives the required background about ARM algorithm. Then section four describes the proposed method in detail. The experiment and the conclusion are presented at sections five and six respectively.

2 Related Work

Information retrieval community found the power of fusing various information sources on the retrieving performance [3]. Information fusion has the potential of improving retrieval performance by relying on the assumption that the heterogeneity of multiple information sources allows cross-correction of some of the errors, leading to better results [4]. In literature, the fusion of the visual and the textual features was performed in different levels of the retrieval process which are early fusion, late fusion, trans-media fusion and at re-ranking level.

2.1 Early Fusion

This method first extracts the low level features of the modalities using the suitable feature extractor. Then, the extracted vectors are concatenated into one vector to form one unique feature space. The advantage of this strategy is that it enables a true multimedia representation for all the fused modalities where one decision rule is applied on all information sources. Early fusion could be used without feature weighting such in [5]; they concatenated the normalized feature spaces of the visual

and the textual features. On the other hand, feature weighting was used in different works in order to provide more weight for specific features. In [6] and [7] as part of ImageCLEF 2006 and 2007 respectively, they presented a novel approach to weight features using support vector machines. The main drawback of early fusion is the dimensionality of the resulting feature space which is equal to the sum of all the fused subspaces which leads to the well-known problem the "curse of dimensionality" [8]. Also, increasing in the number of modalities and the high dimensionality make them difficult to learn the cross-correlation among the heterogeneous features [9].

2.2 Late Fusion

Late fusion (or decision level) strategies do not act at the level of one representation for all the modalities features but rather at the level of the similarities among each modality. The extracted features of each modality are classified using the appropriate classifier; then, each classifier provides a decision. Unlike early fusion, where the features of each modality may have different representation, the decisions usually have the same representation. As a result, the fusion of the decisions becomes easier. The main disadvantage of this strategy is that it fails to utilize the feature level correlation among modalities. Also, using different classifiers and different learning process is expensive in term of time and learning for each modality.

Late fusion is used widely in image retrieval systems, and there is a diversity in the proposed methods. The most widely used technique is a rule-based method [10-16]. In [16], web application called MMRetrieval is proposed which has an online graphical user interface that brings image and text search together to compose a multimodal and multilingual query. The modalities are searched in parallel, and then the results can be fused via several selectable methods. Fusion process consists of two components: score normalization and combination. It provides a combination of scores across modalities with summation, multiplication, and maximum.

2.3 Trans-media Fusion

In this method, the main idea is to use first one of the modalities (say image) to gather relevant documents (nearest neighbors from a visual point of view) and then to use the dual modalities (text representations of the visually nearest neighbors) to perform the final retrieval. Most proposed methods under this category are based on adopted relevance feedback or pseudo-relevance feedback techniques as in [17]. The authors in [17] used the pseudo-relevance feedback to gather the N most relevant documents from the dataset using some visual similarity measures with respect to the visual features of the query or, reciprocally, using a purely textual similarity with respect to the textual features of the query, then aggregate these mono-modal similarities.

2.4 Image Re-ranking

In image re-ranking level, we need first to perform the search based on the text query. Then, the returned list of images is reordered according to the visual features similarity.

In [18], the cross-reference re-ranking strategy is proposed for the refinement of the initial search results of text-based video search engines. While [18] method deals with clusters of the modalities, [19] proposed a method that construct a semantic relation between text (words) and visual clusters using the ARM algorithm. They proposed Multi-Modal Semantic Association Rules (MMSAR) algorithm to fuse key-words and visual features automatically for Web image retrieval.

MFAR in this paper is considered as a late fusion method. There are three main differences between the method of [19] and MFAR proposed method: (1) MFAR uses ARM algorithm to explore the relations between text semantic clusters and image visual feature clusters; (2) the fusion method in MFAR is used at the retrieval phase not for re-ranking the results; (3) it is possible in MFAR to make a query by example image. In literature, there are several attempts to couple image retrieval and association rules mining algorithm. First, it is used as a preprocessing strategy for a preliminary reduction of the dimensionality of the pattern space to improve the global search time for CBIR system as in [20]. Second, as mentioned earlier, ARM has been used in image re-ranking process [19].

The next section will present the required background about ARM algorithm, which helps to understand the proposed method.

3 Basics of Association Rules Mining Algorithm

ARM is a data mining technique useful for discovering interesting relationships hidden in large databases. The classical example is the rules extracted from the content of the market baskets. Items are things we can buy in a market, and transactions are market baskets containing several items. The collection of all transactions called the transactions database. Besides the market basket data, association rules mining are applicable for different applications of other domains such as bioinformatics, medical diagnosis and Web mining.

The problem of mining association rules is stated as following: $I=\{i_1 , i_2 , \dots , i_m\}$ is a set of items, $T=\{t_1 , t_2 , \dots , t_n\}$ is a transaction database or a set of transactions, each of which contains items of the itemset I. Thus, each transaction t_i is a set of items such that $t_i \subseteq I$. An association rule is an implication of the form: $X \rightarrow Y$, where $X \subset I$, $Y \subset I$ and $X \cap Y = \emptyset$. X (or Y) is a set of items, called itemset. If an itemset contains k items, it is called k-itemset. It is obvious that the value of the antecedent implies the value of the consequent. The process of mining association rules consists of two main steps. The first step is to identify all the itemsets contained in the data that are adequate for mining association rules. To determine that the itemset is frequent, it should satisfy at least the predefined minimum support count. To measure the support for an itemset, the following formal definition is used:

$$Supp(X) = \frac{count\ (X)}{N} \tag{1}$$

Where N is the total number of transactions in the transaction database T i.e. $N = count(T)$. The second step is to generate rules out of the discovered frequent itemsets.

For doing so, a minimum confidence has to be defined. The formal definition to calculate the rule confidence is given by the following equation:

$$Conf(X \longrightarrow Y) = \frac{count\ (X \cup Y)}{count\ (X)} \quad (2)$$

The confidence of the rule $X \rightarrow Y$ is a measurement that determines how frequently items in Y appear in transactions that contain X. Different algorithms attempt to allow efficient discovery of frequent patterns and for strong ARs such as the famous *Apriori* algorithm [21] which will be used later in MFAR.

4 Methodology

MFAR consists of two main phases: online phase and offline phase. The next subsections describe in details the inputs, the outputs and the steps of each phase.

4.1 Offline Phase

The input of this phase is the image dataset which contains two modalities: the images and their associated text. First, the visual and the textual features are extracted to run the clustering algorithm independently over them. Then, the modified ARM algorithm will identify the relations among the clusters from each modality to construct the ARs (see figure 1.a).

For visual features extraction, we used a set of generic MPEG-7 descriptors [22]. The features are selected to balance the color and the edge properties of the images. After extracting the visual features, images of the dataset are represented separately as objects in multidimensional space models for each visual feature. For textual features, they were obtained by applying the standard Bag-of-Words technique which needs to perform several linguistic preprocessing steps (tokenization, removing stop words, and stemming). Then, each document is described by a vector of constituent terms that represents the frequency occurrence of each term in the document which construct the vector-space model.

The large quantity of images and the high dimensionality of the visual descriptors need for an efficient clustering (or indexing) algorithm. The high dimensional index technique called NOHIS (Non Overlapping Hierarchical Index Structure) [23] is used for the indexing process which generates the NOHIS-tree. Then, an adapted k-nearest neighbors search is used for retrieving. On the other hand, K-means algorithm will be used for the textual features.

To apply the ARM algorithm, we need first to determine the items set I and the transaction database T. In our case, the items set is the generated images clusters based on the text (denoted by Ct_i) and based on the visual features (denoted by Cc_j for color-based clusters and Ce_k for edge-based clusters) where i, j and k are the identifiers of the clusters in each modality. After quantifying the features space of each modality, we aim to associate the text clusters and the visual feature clusters.

Thus, we need to construct the transaction database T first to run the ARM algorithm over it.

Each transaction in T contains the similar clusters from different modalities. Similarity here means the overlapping degree between the clusters. If the cardinality of the common images set is not zero, the clusters combine at the same transaction. It is possible to represent that in the following example: If $| Ct_i \cap Cc_i | > 0$, then add $\{Ct_i, Cc_i\}$ to T. The hypothesis in constructing T is that similar clusters tends to be semantically related; therefore, they are combined at the same transaction. We are interested in the association between text clusters and visual feature clusters only. Each transaction contains a text cluster and at least one visual cluster. The following are examples of the obtained transactions: $\{Ct_0, Cc_{111}\}$, $\{Ct_0, Ce_{206}\}$, $\{Ct_0, Cc_{111}, Ce_{173}\}$.

Two different reasons let us adjust the formal definitions of support and confidence (definitions (1) and (2)). First, using the standard support/confidence definition for the semantic rules, which is calculated for the entire T, will affect the generated rules because their support is extremely low. Second, the calculation of support and confidence is restricted within the result set of the text clusters because we are testing the semantic relations between the text clusters and visual clusters. Thus, we define the support and the confidence of the rule $Ct_i \rightarrow Cv_j$ (where Cv represents the visual cluster) as follows:

$$Supp(Ct_i \rightarrow Cv_j) = \frac{count(Ct_i, Cv_j)}{count(Ct_i)} \tag{3}$$

$$Conf(Ct_i \rightarrow Cv_j) = \frac{count(Ct_i, Cv_j)}{\max_k(count(Ct_i, Cv_k))} \tag{4}$$

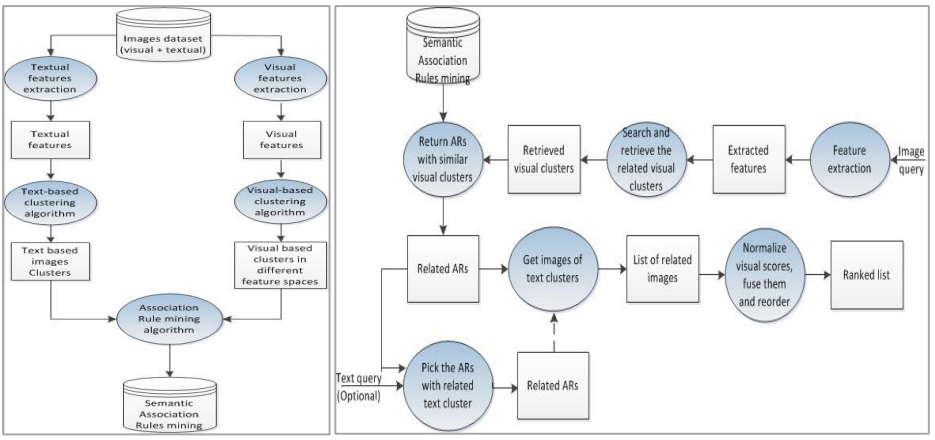
Where *count(A)* is the number of itemsets that contain A in T. Similarly in case there is more than one item at the right hand side of the rule is given by (5) and (6):

$$Supp(Ct_i \rightarrow \{ Cv_j | j = 1, \ldots, m\}) = \frac{count(Ct_i, \{ Cv_j | j = 1, \ldots, m\})}{count(Ct_i)} \tag{5}$$

$$Conf(Ct_i \rightarrow \{ Cv_j | j = 1, \ldots, m\}) = \frac{count(Ct_i, \{ Cv_j | j = 1, \ldots, m\})}{\max_k(count(Ct_i, Cv_k))} \tag{6}$$

We need to use a modified version of frequent itemsets mining algorithm based on *Apriori* algorithm with definitions (5) and (6) of support and confidence to discover all frequent patterns of the association between text clusters and visual feature clusters. The algorithm is in table 1. The algorithm do not start from 1-itemsets; that because we want to construct the relationships between text clusters and visual clusters; and in case starting from 1-itemsets, it is possible to build relations among visual clusters since they will be treated equally. The minimum support threshold should be given to run the algorithm.

Here, *apriori-gen* function is used to perform three main operations: (1) candidate generation; (2) candidate pruning; and (3) insuring that each candidate itemset should have one text cluster. The *subset* function is used to determine all the candidate itemsets in C_k that are contained in each transaction t. A transaction t is said to contain an itemset X if X is a subset of transaction t.



(a) **(b)**

Fig. 1. The offline (a) and online phase (b) of MFAR

Table 1. Frequent itemsets mining algorithm based on Apriori

Input:
a) The transaction database T
b) *minsup* threshold

Output:
The list of frequently itemsets L

1) $L_2 = \{(Ct_i , Cv_j) \mid \text{where } | Ct_i \cap Cv_j | > 0 \;\&\&\; (Ct_i , Cv_j).\text{supp} \geq minsup\}$; //Find all frequent 2-itemsets
2) for ($k = 3$; $L_{k-1} \neq \emptyset$; k++) do begin
3) C_k = apriori-gen(L_{k-1}); // New candidates with k-itemset with only one text cluster in it and a // combination of frequent sets from L_{k-1}
4) for all transactions $t \in T$ do begin
5) C_t = subset(C_k , t); // Identify all candidates that belong to t
6) for all candidates $c \in C_t$ do
7) $c.count$++;
8) end
9) $L_k = \{ c \in C_k \mid c.supp \geq minsup\}$
10) end
11) Return $\cup L_k$;

To generate strong ARs, the generated frequent itemsets L and the minimum confidence threshold value *minconf* should be used as input to the generating algorithm. The ARs in our case have one text cluster in the left hand side and one or multiple visual cluster(s) at the right hand side. There is no need to find all possible subsets of the large itemset L as in the original *Apriori* algorithm. For example, if $l = \{Ct_1, Cc_3, Ce_1\}$ is a frequent itemset, candidate rule is $Ct_1 \rightarrow \{Cc_3, Ce_1\}$. If the calculated confidence of the candidate rule using (6) is greater than or equal *minconf*,

then the rule is strong; otherwise, it is discarded. Finally, all the generated ARs are stored in the database along with the values of support and confidence for each rule which is the final output of this phase.

4.2 Online Phase

This phase uses the generated ARM of the offline phase. The main processes are illustrated in figure 1.b. The basic query model used here is the query by example image since when image is used as query, all the information it contains is provided to the system. Using a keyword as a query is optional. It could be provided to the system to support the results that generated by the image query. For the query image, we need to extract the same visual features that have been extracted from the image dataset. For the optional keyword query, we used one keyword and simple text matching to simplify this step.

We need to use the same index NOHIS-trees of the offline phase to retrieve the relevant clusters to the query image for each visual descriptor. In our case, we have two different NOIHS-trees for two different feature spaces. For each feature, we calculated the top 500 nearest neighbors and returned their clusters. The search should be conducted on the trees in parallel. The output of this process is a list of visual clusters from different feature spaces.

Then, the next process "retrieve ARs with similar visual clusters" gets the list of the related visual clusters as input; and then it uses them to make a search in the ARM to find the rules that contain these clusters. If the keyword query was provided, the retrieved rules should be filtered to pick the rules which contain text clusters that have similar term to the text query. Then, the images' scores in those text clusters should be increased. The dashed arrow in figure 1.b indicates that it is an optional path.

For all the retrieved ARs, we need to get the images of the text-based clusters. For each image, the relevant score to the query image q should be calculated if the image is not from the top 500 images for each visual feature. Regarding score normalization, we used Zero-One linear method which maps the scores into the range of [0, 1] [24]. The normalized scores of different modalities should be fused using CombSum method [24]. Then, if there is a keyword query as input, the fused score of each image that correlated to term similar to the keyword query should be incremented by one. Finally, the fused list will be reordered based on the fused scores.

5 Experiment

5.1 Experimental Setup and Tools

MFAR has been evaluated using ImageCLEF 2011 Wikipedia collection. It consists of 50 topics and 237,434 Wikipedia images along with their user-provided annotations in three different languages [25]. Since some images in the dataset do not have English description and others do not have a description at all, only images with English description are considered. Thus, the used dataset is a subset of ImageCLEF 2011

Wikipedia which contains more than 54,500 images. Some example topics of the dataset along with their titles, the used text query, the number of image queries in the topic and the number of relevant images in the collection subset are given in table 2.

For visual features extraction, the two MPEG-7 descriptors: Color Structure Descriptor (CSD) and Edge Histogram Descriptor (EHD) are extracted from the dataset using the tool given in [26]. For textual features extraction and K-means clustering, Text-Garden software is used[1]. To cluster the extracted visual features, NOHIS algorithm library is provided by the author of the algorithm. The system prototype is developed in C#.NET Framework with simple GUI for experiment purpose only (see figure 2). Based on different experiments, we set *minsupp* and *minconf* to be 2% and 70% respectively.

MFAR was compared to our system without using ARs and to the online system MMRetrieval[2] [16]. Since MMRetrieval system supports different fusion methods, the well-known method CompSum with MinMax normalization is selected. We used the example images of all the dataset topics. For our system without ARs, the queries are only images. On the other hand, for MFAR and MMRetrieval, the query can be either image only or image with keyword. The text query is restricted to be one word.

Table 2. Information of some topics of the subset collection

Topic ID	Topic Title	Text query	No.# of query images	No# of relevant images
85	Beijing bird nest	Volkswagen	5	8
95	photo of real butterflies	Butterfly	5	37
107	sunflower close up	Sunflower	5	4
111	two euro coins	Euro	5	30
115	flying bird	Flying	5	46

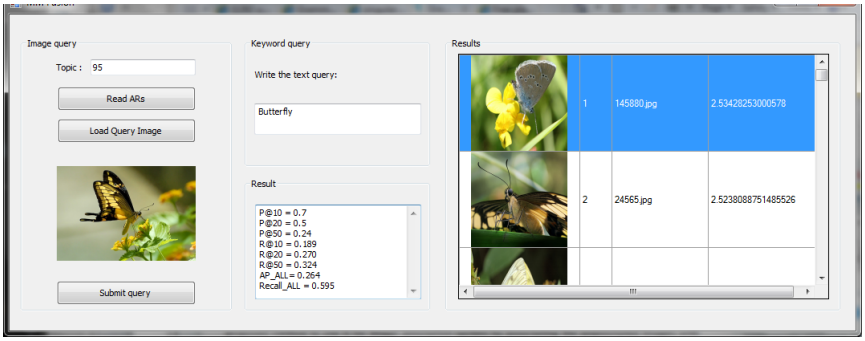


Fig. 2. Main GUI of MFAR

[1] Text-Garden – Text Mining Software Tools. http://www.textmining.net
[2] http://mmretrieval.nonrelevant.net

5.2 Experimental Results and Discussion

For evaluation, we used the Precision (P) at fixed rank (10 and 20), and the Mean Average Precision (MAP) [27]. The values of P@10, P@20 and Average Precision (AP) of five random categories (with different difficulty levels; and they are not the best results) are given in table 3. Each value in the table represents the average of the precisions for the five example images contained in the topic. In addition, table 4 shows the overall values of P@10, P@20 and MAP for all topics of the dataset. The results show that MFAR with composite query (image + keyword) performs better precision and MAP than the other two systems. Furthermore, the proposed system and MMRetrieval system have been evaluated with an image query only without using text; and the proposed system performs acceptable semantic results comparing to MMRetrieval system and provides better precision results than MMRetrieval. The precision values with image query mode in both systems are lesser than the systems with composite query.

We examined the retrieved ARs for different visual queries to study the relations between the image query and the rules. One example is an image from topic 107 with title "sunflower close up". Text cluster Ct_{645} is classified based on different words one of them is "sunflower". The retrieved ARs for the query in the two query modes: query by image only and the composite query contain rules that associate Ct_{645} text cluster to visual clusters consists of sunflower pictures. That means by using the visual features of the query image, it is possible to reach the text cluster which is semantically related.

In addition, we found that by using MFAR the search operation concentrate on the images subset that included in the retrieved ARs of the submitted query which increases the chance of retrieving a semantically related results.

Table 3. P@10, P@20 and AP of 5 different topics in: (1) Sys.1: our system without ARs (visual), (2) Sys.2: MMRetrival system (visual + text), and (3) Sys.3: MFAR (visual + text)

Topic	P@10			P@20			AP		
ID	Sys.1	Sys.2	Sys.3	Sys.1	Sys.2	Sys.3	Sys.1	Sys.2	Sys.3
85	0	0	0.2	0	0.013	0.13	0.001	0.035	0.282
95	0	0.72	0.66	0	0.62	0.56	0.004	0.366	0.234
107	0	0.28	0.3	0	0.15	0.15	0.004	0.468	0.658
111	0	0.38	0.4	0.01	0.23	0.47	0.018	0.236	0.350
115	0.12	0.14	0.22	0.07	0.11	0.18	0.031	0.033	0.047

Table 4. The overall values of P@10, P@20, and MAP of our system without ARs, MMRetrival system, and MFAR

Sys. without ARs			MMRetrieval			MFAR		
P@10	P@20	MAP	P@10	P@20	MAP	P@10	P@20	MAP
0.011	0.009	0.010	0.205	0.164	0.210	0.240	0.175	0.244

6 Conclusion and Future Work

In this proposed method, we used association rules mining algorithm in our image retrieval system to construct semantic relations between image clusters based on the visual features and the image clusters based on textual features for the same dataset. From information fusion perspective, we have used late fusion technique. The online phase uses the constructed ARs from the offline phase. Then, the retrieval process requires an example image query to start. The method gives the ability to retrieve images that are semantically related by using the extracted visual features of the query image and by exploring the related ARs from the constructed ARM. To support the results, it is possible to use a keyword query. The results show that the precision value of our proposed system is better than MMRetrieval system and the system without association rules.

The future work will involve using different clustering algorithm to improve the accuracy of the text clusters. The system with image query mode without keyword query needs for further improvements. Using pseudo-relevance feedback technique is one suggested solution. The correlated terms of the top retrieved ARs could be used to make feedback text query. Also, it is possible to generalize the proposed method to use it for image annotation system by associating the unannotated images with the semantically related text cluster.

References

1. Liu, Y., Zhanga, D., Lua, G., Ma, W.-Y.: A survey of content-based image retrieval with high-level semantics. Pattern Recognition 40(1), 262–282 (2007)
2. Datta, R., Joshi, D., Li, J., Wang, J.Z.: Image retrieval: Ideas, influences, and trends of the new age. ACM Computing Surveys (CSUR) 40(2), 1–60 (2008)
3. Wu, S., McClean, S.: Performance prediction of data fusion for information retrieval. Management Information Processing 42(4), 899–915 (2006)
4. Müller, H., Clough, P., Deselaers, T., Caputo, B.: ImageCLEF. The Springer International Series on Information Retrieval, vol. 32, pp. 95–114. Springer (2010)
5. Ferecatu, M., Sahbi, H.: TELECOM ParisTech at ImageClefphoto 2008: Bi–modal text and image retrieval with diversity enhancement. In: Working Notes of CLEF 2008, Aarhus, Denmark (2008)
6. Deselaers, T., Weyand, T., Ney, H.: Image retrieval and annotation using maximum entropy. In: Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (eds.) CLEF 2006. LNCS, vol. 4730, pp. 725–734. Springer, Heidelberg (2007)
7. Gass, T., Weyand, T., Deselaers, T., Ney, H.: FIRE in ImageCLEF 2007: Support vector machines and logistic models to fuse image descriptors for photo retrieval. In: Peters, C., Jijkoun, V., Mandl, T., Müller, H., Oard, D.W., Peñas, A., Petras, V., Santos, D. (eds.) CLEF 2007. LNCS, vol. 5152, pp. 492–499. Springer, Heidelberg (2008)
8. Bellman, R.: Adaptive Control Process: A Guided Tour. Princeton University Press (1961)
9. Atrey, P.K., Hossain, M.A., Saddik, A.E., Kankanhall, M.S.: Multimodal fusion for multimedia analysis: A survey. Multimedia Syst. 16(3), 1432–1882 (2010)

10. Lau, C., Tjondronegoro, D., Zhang, J., Geva, S., Liu, Y.: Fusing visual and textual retrieval techniques to efectively search large collections of wikipedia images. International Journal of Business Intelligence and Data Mining, 345–357 (2007)
11. Frigui, H., Caudill, J., Ben Abdallah, A.: Fusion of multi-modal features for effiecient content-based image retrieval. In: IEEE World Congress on Computational Intelligence, pp. 1992–1998 (2008)
12. Bartolini, I., Ciaccia, P.: Scenique: a multimodal image retrieval interface. In: Proceedings of the Working Conference on Advanced Visual Interfaces, pp. 476–477. ACM, Italy (2008)
13. Zhou, X., Depeursinge, A., Müller, H.: Information fusion for combining visual and textual image retrieval. In: Proceedings of the 20th International Conference on Recognizing Patterns in Signals, Speech, Images, and Videos, pp. 1590–1593 (2010)
14. Tong, H., He, J.R., Li, M.J., Zhang, C.S., Ma, W.Y.: Graph-based multi-modality learning. In: Proceeding of the ACM Int. Conf. on Multimedia, pp. 862–871 (2005)
15. Escalante, H.J., Hernandez, C., Sucar, E., Montes, M.: Late fusion of heterogeneous methods for multimedia image retrieval. In: Proceeding of MIR, pp. 172–179. ACM, Vancouver (2008)
16. Zagoris, K., Arampatzis, A., Chatzichristofis, S.A.: www.MMRetrieval.net: a multimodal search engine. In: Proceedings of the Third International Conference on SImilarity Search and Applications, Turkey (2010)
17. Ah-Pine, J., Bressan, M., Clinchant, S., Csurka, G., Hoppenot, Y., Renders, J.: Crossing textual and visual content in different application scenarios. Multimedia Tools and Applications 42(1), 31–56 (2009)
18. Wei, S., Zhao, Y., Zhu, Z., Liu, N.: Multimodal Fusion for Video Search Reranking. IEEE Transactions on Knowledge and Data Engineering 22(8), 1191–1199 (2010)
19. He, R., Xiong, N., Yang, L., Park, J.: Using multi-modal semantic association rules to fuse keywords and visual features automatically for web image retrieval. In: International Conference on Information Fusion (2011)
20. Kouomou-Choupo, A., Berti-Equille, L., Morin, A.: Multimedia indexing and retrieval with features association rules mining. In: IEEE International Conference on Multimedia and Expo (ICME 2004), pp. 1299–1302 (2004)
21. Agrawal, R., Imielinski, T., Swami, A.: Mining association rules between sets of items in large databases. In: Proceedings of ACM SIGMOD, pp. 207–216 (1993)
22. Manjunath, B.S., Salembier, P., Sikora, T.: Introduction to MPEG-7: multimedia content description interface. Wiley (2002)
23. Taileb, M., Lamrous, S., Touati, S.: Non Overlapping Hierarchical Index Structure. International Journal of Computer Science 3(1), 29–35 (2008)
24. Wu, S.: Data Fusion in Information Retrieval. ALO, vol. 13. Springer, Heidelberg (2012)
25. Tsikrika, T., Popescu, A., Kludas, J.: Overview of the Wikipedia Image Retrieval Task at ImageCLEF 2011. In: Working Notes of CLEF 2011, Amsterdam, The Netherlands (2011)
26. Bastan, M., Cam, H., Gudukbay, U., Ulusoy, O.: BilVideo-7: An MPEG-7 Compatible Video Indexing and Retrieval System. IEEE MultiMedia 17(3), 62–73 (2010)
27. Müller, H., Clough, P., Deselaers, T., Caputo, B.: ImageCLEF. The Springer International Series on Information Retrieval, vol. 32, pp. 81–92. Springer (2010)