ISSN: 2321-7782 (Online)
Volume 3, Issue 7, July 2015
International Journal of Advance Research in Computer Science and Management Studies
Research Article / Survey Paper / Case Study
Available online at: www.ijarcsms.com

Data Mining Approach for Image Retrieval in Multimodal Fusion Using Frequent Pattern Tree

Priyanka B. Kamdi[1]
2nd year Mtech CSE, Vidarbha Institute of Technology, Nagpur, India

Pravin Kulurkar[2]
Assistant Professor, Vidarbha Institute of Technology, Nagpur, India.

Abstract: ***In this project utilizes the fusion of the images (textual and visual) for retrieving images. It combines two different data mining techniques to retrieve semantically related images: clustering and Frequent Pattern Tree Based Algorithm. In the offline phase, the relations among the clusters will be identified from different modalities to construct the frequent pattern rules. On the other hand, the online phase (retrieving phase) uses the generated Tree pattern, to retrieve the related images of the query. The World Wide Web is regarded as the largest global image repository. Image mining is technique that can automatically extract semantically meaningful information from image data. The fundamental challenge in image mining is to determine how low-level, pixel representation contained in a raw image or image sequence can be efficiently and effectively processed to identify high level spatial objects and relationships. The inputs of this phase are the images 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 Frequent Pattern algorithm will identify the relations among the clusters from each modality to construct the tree. By using frequent pattern tree we can easily find out the related images quickly. In this project made the comparison between Apriori and Frequent Pattern Tree. From this result found to be frequent pattern tree is more affected and get quick response.***

Keywords: Clustering, frequent pattern tree, Apriori, retrieving

I. INTRODUCTION

Today, Information Technology (IT) is presented in virtually all areas. This leads to raise the importance of computers for remote monitoring which is the basis for control systems. Large volumes of databases, diversity and heterogeneity of data sources require a new philosophy of treating them. In this case, data mining looks to discover implicit knowledge in a dataset based on different techniques that can be implemented independently or coupled. These techniques aim to explore data, to describe their contents and extract the information more meaningful. Because much of the information that exists in organizations is informal and unstructured, these techniques are not limited to digital data and evidence, but must also address the textual and multimedia. In the recent years the acquisition, generation, storage and processing of data in computers and transmission over networks had had a tremendous growth. This changed radically with the appearance of the Internet and the first web browser which revolutionized the distribution of information. The ease of information exchange incited millions of people to create their own web pages and to enrich it with images, video, audio and music. Due to this rapid development in the domain of digital and informational data people now live in a multimedia world. More and more multimedia information is generated and available in digital form from varieties of sources around the world, this expansion presents new challenges.

Image mining refers to set of tools and techniques to explore images in an automated approach to extract semantically meaningful information. The retrieval process represents a visual query to the system and extracts the images based on the user request such mechanism referred to as query-by-example and it requires the definition of an image representation a set of descriptive features and of some similarity metrics to compare query and target images. The additional mechanisms have been introduced to achieve better performance and relevance feedback proved to be a powerful tool to iteratively collect information

from the user and transform it into a semantic bias in the retrieval process. RF increases the retrieval performance and it enables the system to learn what is relevant or irrelevant to the user across successive retrieval-feedback cycles. RF approaches critical issues yet unsolved. And user interaction is time consuming and tiring, and it is desirable to reduce as much as possible the number of iterations to convergence

It is mostly low level features (color, texture, etc.). And in general, there is no direct link between the high-level concepts and the low-level features.

However, this gap can be reduced by using techniques from Neural Networks, Genetic Algorithms and Clustering research areas that provide powerful learning tools for supervised or unsupervised learning. The nice part about these tools is that they can be used for an o -line processing step, during which a connectivity network (knowledge representation network) is created to point out the links between deferent low-level features and the high-level concepts. In this paper we consider the use of decision trees methodology and a variation of k-means clustering to automatically explore the data extracted from images, and mappings between low level color features and high level textual descriptors.

In the supervised approach, by using decision tree methodology, images are clustered based on their initial keywords assignment in such a way to minimize the classification error. The rules are derived from the resulting induced trees (one tree/keyword) and combined into a comprehensive rule base. The second approach, partially unsupervised, is performed in two stages: First, a k-means clustering step is executed for organizing the image collection into a hierarchy of clusters. Through this step, images with “similar” color content are assigned to the same clusters. Then, for each cluster, a mapping function is determined which is based on two things: cluster semantics and statistical properties of the low-level features of the images assigned to the cluster.

The organization of the paper is as follows. There are two methods used to encode the color information of images. One captures the global color information while the other captures the localized color distribution in the images. These representations are used to organize the images into clusters by the two clustering techniques. The k-means based technique is used to build a hierarchy of clusters that will be afterwards mapped into their optimal textual characterization. The decision tree based clustering technique is used mappings between textual descriptors (keywords) and global color distribution in a given image. Section 3 provides a brief exposition of decision tree and k-means clustering methodologies and describes the proposed scheme for mappings between low-level features and textual features associated with an image. The performance of the two approaches is described in Section 4. Section 5 presents a summary of the work and some concluding remarks.

The multimodal fusion method for image retrieval are already exist but by using apriori algorithm, but there are some limitations of apriori algorithm like, Apriori algorithm is very slow and bottleneck in candidate key generation, Needs several iterations of the data, Some competing alternative approaches focus on partition and sampling

To overcome these limitations, we propose a new method in which we use frequent pattern tree and k-means algorithm for clustering.

II. Related Literature

In literature, in this method, used association rules mining algorithm for Web image retrieval system to construct a semantic relations between images clusters based on the visual features and the images clusters based on textual features for the same dataset. After constructing the ARs in the offline phase, the retrieving process should starts with example image query in the online phase. 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 mining. To support the results.[1]

In this paper introduced a novel FP-array technique that allows using FP-trees more efficiently when mining frequent item sets. Our technique greatly reduces the time spent on traversing FP-trees, and works especially well for sparse data sets. By

incorporating the FP-array technique into the FPgrowth method, the FPgrowth* algorithm for mining all frequent itemsets was introduced. [2]

A new re ranking method that combines multimodal features via a cross-reference strategy. It can handle the initial search results independently in various modality spaces. Specifically, the initial search results are first divided into several clusters individually in different feature spaces. Then, the clusters from each space are mapped to the predefined ranks according to their relevance to the query. Given the ranked clusters from all the feature spaces, the cross-reference strategy can hierarchically fuse them into a unique and improved result ranking. Experimental results show that the search effectiveness, especially on the top ranked results, is improved significantly. [3]

III. Problem Definition

Before existing system there is no technique to search taking image as input and get efficient result by matching features of the image. Currently, most Web based images search engines rely purely on textual metadata. That produces a lot of garbage in the results because users usually enter that metadata manually which is inefficient, expensive and may not capture every keyword that describes the image.

In existing system search the images but taking lot of time and only limited images get displayed.

IV. Project Objectives

The objective of proposed techniques is

- To search the images
- To retrieve images and text and get multiple related search as faster way
- Reduce the complexity
- Extract semantically meaningful information from image data
- Comparison between Apriori and Frequent Pattern Tree
- Find frequent pattern on multiple images.
- Improve the result of existing system.
- Improve the time of the existing system.
- Form efficient cluster.

V. Proposed Approach

As discussed in literature, many techniques have been proposed. In existing system also search the images but required much more time and displayed limited number of images .

5.1 Overview:

The proposed method is a Multimodal Fusion method based on Frequent Pattern Tree Mining. It is considered as a late fusion.

Methodologies:

To achieve the objective, we have proposed following techniques;

- This method combines two different data mining techniques for retrieving: clustering and frequent pattern tree mining (FPTM) algorithm.

» It uses FPTM algorithm to explore the relations between text semantic clusters and image visual features clusters building a decision tree in the space of frequent patterns as an alternative for the two phases approach:

Offline and online phase:

» In the offline phase, the relations among the clusters will be identified from different modalities to construct the frequent pattern rules.

» On the other hand, the online phase (retrieving phase) uses the generated Tree pattern, to retrieve the related images of the query.

5.2 Algorithms/Technology Used

» Frequent pattern tree algorithm

» K-means algorithm for clustering

5.2 Algorithm:

Frequent pattern tree

Input: A database DB, represented by FP-tree constructed according to Algorithm 1, and a minimum support threshold.

Output: The complete set of frequent patterns.

Method: call FP-growth (FP-tree, null).

Procedure FP-growth (Tree, a) {

(01) if Tree contains a single prefix path then // Mining single prefix-path FP-tree {

(02) let P be the single prefix-path part of Tree;

(03) let Q be the multipath part with the top branching node replaced by a null root;

(04) for each combination (denoted as ß) of the nodes in the path P do

(05) generate pattern ß ∪ a with support = minimum support of nodes in ß;

(06) let freq pattern set(P) be the set of patterns so generated;

}

(07) else let Q be Tree;

(08) for each item ai in Q do { // Mining multipath FP-tree

(09) generate pattern ß = ai ∪ a with support = ai .support;

(10) construct ß's conditional pattern-base and then ß's conditional FP-tree Tree ß;

(11) if Tree ß ≠ Ø then

(12) call FP-growth(Tree ß , ß);

(13) let freq pattern set(Q) be the set of patterns so generated;

}

(14) return(freq pattern set(P) ∪ freq pattern set(Q) ∪ (freq pattern set(P) × freq pattern set(Q)))

Detailed Implementation Procedure (Step Wise)

The most important feature of the project is to use Frequent Pattern Tree Algorithm to search within the clusters and get relevant results i.e. images related to the provided labeled image. I have also implemented the Apriori Algorithm to prove that proposed algorithm is better in every way than Apriori algorithm.

Flow of Project

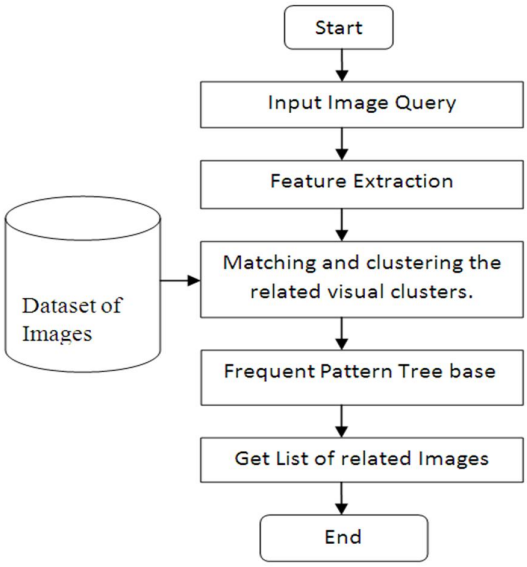
- Upload Image
- Enter Keyword
- Now apply the required algorithm to find relative images

Flow according to pages and classes:

- Load 'search.jsp'
- Browse and upload the image using multipart form-data
- This uses the servlet 'Upload.java' where we have code to upload different types of files. This servlet uploads the file and returns to the respective pages, in this case back to 'search.jsp'.
- Enter keywords related to your search in the provided text field.
- Press submit button. This will redirect you to 'searchResponse.jsp'
- Now 'searchResponse.jsp' will give you two options either apply FPT or Apriori. Select which ever you want. Both will redirect you to their respective search pages i.e. 'fpt.jsp' and 'apr.jsp' respectively.
- Both pages will call their respective methods in the common class 'SearchAgent.java'. Both the methods will take parameter as the input filename and the entered keywords.
- Both the methods i.e. 'searchApr()' and 'searchFpt()' uses three common classes namely 'InfoExtractor.java' to extract information related to the uploaded image, 'DataEntry.java' to perform operations related to database and 'ImageExtractor.java' which will extract images for the final keywords given by any of the algorithm.
- Method '**searchFpt()**' also calls 'FPT.java' where the Frequent Pattern Tree Mining Algorithm is defined.
- Method '**searchApr()**' also calls 'Prediction.java' class which defines the functionality of Apriori Algorithm.
- Both the algorithms returns the set of relative keywords which in turn we give input to 'ImageExtractor.java' for the extraction of related set of images.

The relative keywords that are returned by both the algorithms differ. Complexity of both the algorithm differs too. By analyzing the outputs, the time taken and the code complexity of each algorithm we have proved that FPT is better than that of Apriori.

5.3 Flow Diagram



5.4 K-means clustering algorithm:

1. Choose k number of clusters to be determined
2. Choose k objects randomly as the initial cluster center
3. Repeat
   - 3.1. Assign each object to their closest cluster
   - 3.2. Compute new clusters, i.e. Calculate mean points.
4. Until
   - 4.1. No changes on cluster centers (i.e. Centroids do not change location any more) OR
   - 4.2. No object changes its cluster (We may define stopping criteria as well)

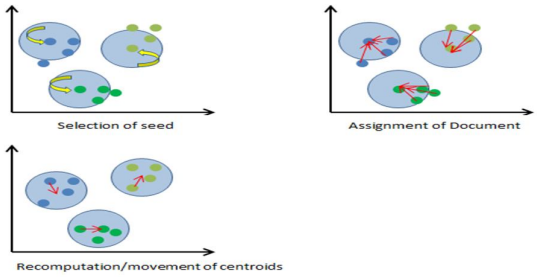


Fig. Selection of seed, assignment of document and recomputation of centroids

.5.3.3 Preprocessing

Preprocessing of text documents is necessary to clean data and to provide algorithms only the required data. The preprocessing techniques used in our system are described below:

5.3.3.1 Removal of Stop word

Stop words are the words which we used more frequently in Natural Languages. They are frequently used common words in any natural languages like preposition, pronoun, conjunction, etc. for example: am, is, was, where, etc. Such frequently appearing words are not very much important in mining usage. So, it can be skip to reduce the document size for mining and for data cleaning usage as well. But, one problem mainly found with the removal of stop words is that removal of stop word is

domain (language) specific and thus the stop word in one domain may not be the stop word in other. We maintained a stop word dictionary having all possible stop words. We scan our documents to find such stop words and remove it as well as we maintained the separate removed stop word list to keep the record for number of stop words found in particular document.

5.3.3.2 Stemming

Today, in information retrieval system and NLP, the word stem is of great importance as it facilitates the indexing of documents. Usually, the prefixes and suffixes are removed to form the stem. By using the stemming technique we can increase the number of retrieved documents. That means we can ultimately increases the recall rate without any affect on the fetched precision. But, there is also a problem of stemming errors for example: over stemming and under stemming which may affect the recall and precision rates. After stop word removal, we performed stemming of words. We maintained indexed stems. For first index position we kept the original stem, and then we scan the document to make the stems. For example: bail / bailed / bailing. So, if we found any word like bailed or bailing then we replace these words as bail.

5.3.3.3Indexing

The process of expressing the main subject or the theme of a text in document is called indexing. Text headings are often taken as indexes. There are primarily two categories of indexing:

» Classification

» Co-ordinate

With classification indexing, or classifying, the texts are included an appropriate class (one or several) depending on their content. All texts with basically the same semantic content are brought together. The index number of this class is assigned to each text within it and the number is then serves as it's search specification.

In coordinate indexing, the basic semantic content of the text is expressed by a list of significant words selected either from the text itself or its headings or from a special normative dictionary. In the first instance, such lexical units are termed key words, and in the second descriptors. Each key word or descriptors designates a class that potentially includes all the texts that have the word in the basic semantic content.

5.3.3.4 Data Collection

We collect our dataset used for the proposed system from various sources. It is a data to be considered as real time police investigation reports. The documents we collected is in various formats like doc, docx, pdf, etc. Also, it is not necessary to use only pre maintained dataset rather we can use any dataset on runtime. For example: the dataset from external devices like pen drives and other.

5.3.3.5 Synonyms

For better results, we maintained a synonym dictionary. If we don't gets accurate word matching then these synonyms could help us to create the related clusters. For example: bail, warranty, surety, bond, guarantee, and warrant. Our system finds any of word and considers it as similar word so that it places these words in same category.

We put a text field to search any query by forensic analysts. There is no need to scan and manually check the cluster of interest. Instead, one can search for the interested clusters by entering any keyword or the query. We maintained the indexing of keywords and the files in which the keywords can be found. We retrieve all these files and then the above preprocessing steps are applied on these files. Thus, we get the keywords found in all files. We then input these result to three different algorithms i.e. K-means, K-medoid and K-representative. We find Jaccard coefficient as given below to calculate the similarity distance between two keywords. Thus, formed the clusters having similar group of words.

Huffman tree, and match distances are compressed with another tree. The trees are stored n a compact form at the start of each block.

VI. Experimental Results And Analysis

This is the main page of Project.

It has three fields

» Update Files Dataset

» Input Image

» Search Dataset

Fig. Snapshot of main page

For searching dataset click on search dataset and get following window

In that there is browse button so first upload the image and the type keyword in next text box and submit it. This is the fusion (i.e. text + image as input)

Fig. Snapshot of image with Keyword as input

For selecting algorithm following window get appear

After that we get following window in that we choose the algorithm, which we want

» Frequent Tree Pattern Mining Images

» Apriori Mining Images

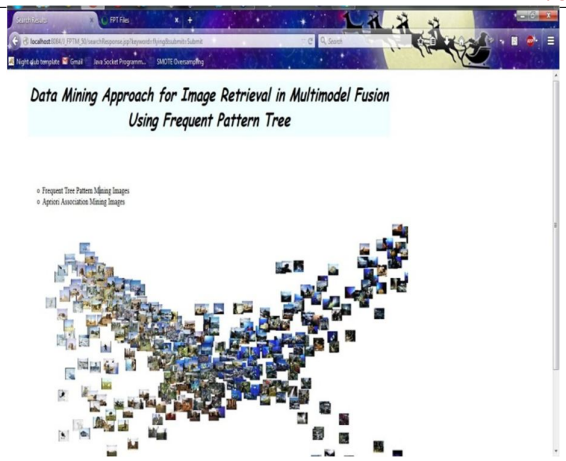


Fig.5 Snapshot of selecting algorithm

This result for eg. Image is arrice.jpg upload and keyword is flying so following window appear for frequent pattern tree

Fig. Output of result using FPT

If we click simultaneously on both the algorithm

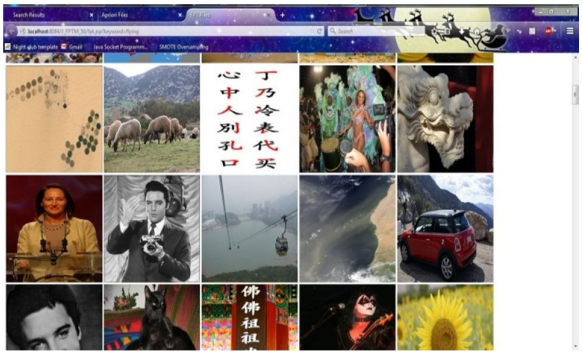


Fig. (conti..)Output of result using FPT

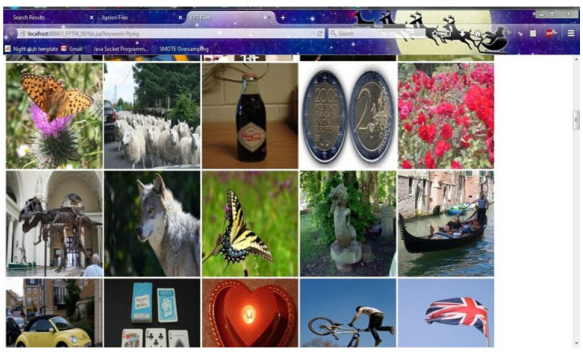


Fig. (conti..)Output of result using FPT

Comparison of Frequent Pattern Tree and Apriori

Keyword(Image+keyword)	FP(w.r.t time)	Apriori(w.r.time)
airrace.jpg+air	3.327	3.598
Cake.jpg+car	3.245	5.234
wolf.jpg+water	5.987	6.098

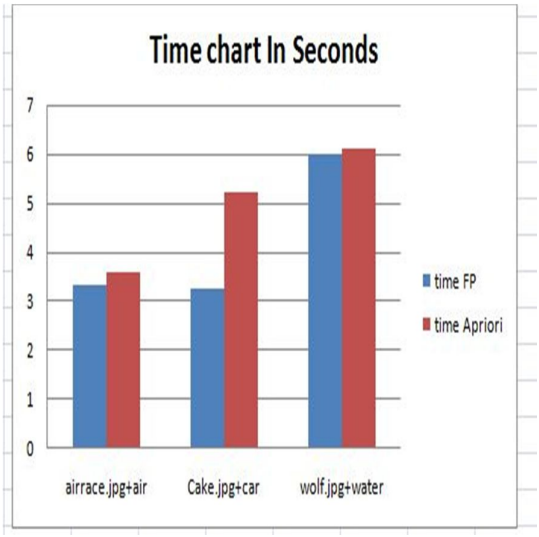
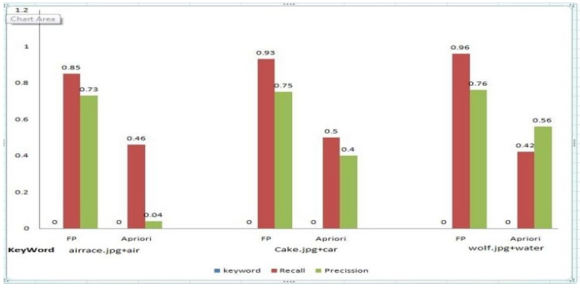


Fig. Comparison of Apriori vs. FPT in terms of time



Comparison of Apriori Vs FPT in terms of recall and precision

VII. Conclusion And Future Scope

In the proposed method, we used frequent pattern tree algorithm in our Web image retrieval system to construct a semantic relations between images clusters based on the visual features and the images clusters based on textual features for the same dataset. In this proposed system show the comparison between frequent pattern tree and apriori and found that frequent pattern tree get quickly result and show maximum search related to image and text. In this project we propose a novel image retrieval approach which combines text, content and interactive based retrieval. The accuracy is higher in comparison to using the techniques separately. We designed a hybrid image retrieval system with the method proposed, which successfully achieves the demands with respect to the system requirements (i.e., allows the users to retrieve their desired images based on the text and/or sample image query). A new refining search algorithm has been provided, which optimizes the search results. The experiments on the sample data sets prove the effectiveness of the system.

Further it is proposed to integrate and classify the techniques and add low level features, like texture, shape to the system. For clustering instated of k- means k-medoid algorithm will used because It could be more robust to noise and outliers as compared to *k*-means because it minimizes a sum of general pair-wise dissimilarities instead of a sum of squared Euclidean distances. For indexing used Inverted indexing for better performa

References

1. Raniah A. Alghamdi, Mounira Taileb, Mohammad Ameen, Faculty of Computing and Information Technology, King Abdulaziz University, Saudi Arabia – Jeddah, "A New Multimodal Fusion Method Based on Association Rules Mining for Image Retrieval", 17th IEEE Mediterranean Electrotechnical Conference, Beirut, Lebanon, 13-16 April 2014
2. Goasta Grahne, Jianfei Zhu Grahne, Member, IEEE, and Jianfei Zhu, Student Member, IEEE "Fast Algorithms for Frequent Item set Mining Using FP-Trees" IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 17, NO. 10, OCTOBER 2005
3. S. Wei , Y. Zhao , Z. Zhu , N. Liu, "Multimodal Fusion for Video Search Reranking", IEEE Transactions on Knowledge and Data Engineering, 2010, v.22 n.8, p.1191-1199.
4. Adrien Depeursinge, Samuel Duc, Ivan Eggel and Henning M¨uller " Mobile Medical Visual Information Retrieval" IEEE TRANSACTIONS ON INFORMATION TECHNOLOGY IN BIOMEDICINE, VOL. 17, NO. 11, OCTOBER 2011
5. M. Bastan, H. Cam, U. Gudukbay, and O. Ulusoy, "BilVideo-7: An MPEG-7 Compatible Video Indexing and Retrieval System", IEEE MultiMedia, 2010, vol. 17, no. 3, pp. 62-73.
6. R. Agrawal, T. Imielinski, and A. Swami, "Mining Association Rules between Sets of Items in Large Databases," Proc. ACMSIGMOD Int'l Conf. Management of Data, pp. 207-216, May 993.
7. S. Wu and S. McClean, "Performance prediction of data fusion for information retrieval". Information Processing, Management, 2006. 42(4): p. 899-915.
8. Ernst, R.: "Co design of embedded systems: status and trends", Proceedings of IEEE Design and Test, April–June 1998, pp.45–54
9. H. Müller, P. Clough, Th. Deselaers, B. Caputo, "ImageCLEF" (ser. The Springer International Series on Information Retrieval), vol. 32, pp.95 -114, 2010, Springer-Verlag.
10. M. Ferecatu and H. Sahbi, "TELECOM ParisTech at ImageClefphoto 2008: Bi–modal text and image retrieval with diversity enhancement". In: Working Notes of CLEF 2008, Aarhus, Denmark.
11. T. Gass, T. Weyand, T. Deselaers, and H. Ney. "FIRE in ImageCLEF 2007: Support vector machines and logistic models to fuse image descriptors for photo retrieval". In: CLEF 2007 Proceedings. Lecture Notes in Computer Science (LNCS), 2007, vol 5152. Springer, pp 492–499.
12. R. Bellman, "Adaptive Control Process: A Guided Tour". Princeton University Press, 1961.
13. P. K. Atrey , M. A. Hossain , A. E. Saddik and M. S. Kankanhall "Multimodal fusion for multimedia analysis: A survey", Multimedia Syst., vol. 16, no. 3, pp.1432 -1882, 2010.
14. C. Lau, D. Tjondronegoro, J. Zhang, S. Geva, and Y. Liu, "Fusing visual and textual retrieval techniques to effectively search large collections of wikipedia images" 2007, p. 345-357.
15. H. Frigui, J. Caudill, and A. Ben Abdallah. "Fusion of multimodal features for efficient content-based image retrieval.", IEEE World Congress on Computational Intelligence, pp. 1992-1998, 2008.
16. Y. Liu, D. Zhanga, G. Lua, and W-Y. Ma , "A survey of content-based image retrieval with high-level semantics", Pattern Recognition, Vol. 40, No. 1. (2007), pp. 262-282.
17. R. Datta, D. Joshi, J. LI, and J. Z. Wang, "Image retrieval: Ideas, influences, and trends of the new age", ACM Computing Surveys (CSUR), April 2008, 40(2):1-60.
18. T. Deselaers, T. Weyand, and H. Ney, "Image retrieval and annotation using maximum entropy", in Evaluation of Multilingual and Multi modal Information Retrieval, 2007, pp. 725-734.
19. I. Bartolini and P. Ciaccia., "Scenique: a multimodal image retrieval interface", in Proceedings of the working conference on Advanced visual interfaces, 2008, ACM, Italy. pp. 476-477.

