Indexing Multimedia Content for Textual Querying: A Multimodal Approach

Abdesalam Amrane[1], Hakima Mellah[1], Youssef Amghar[2] and Rachid Aliradi[1]

[1] Research Center on Scientific and Technical Information (CERIST), Algiers, Algeria
[2] University of Lyon, CNRS, INSA-Lyon, LIRIS, UMR5205, F-69621, Villeurbanne, France

Abstract. Multimedia retrieval approaches are classified into three categories: those using textual information, and those using low-level information and those that combine different information extracted from multimedia. Each approach has its advantages and disadvantages as well to improving multimedia retrieval systems. The recent works are oriented towards multimodal approaches. It is in this context that we propose an approach that combines the surrounding text with the information extracted from the visual content of multimedia and represented in the same repository in order to allow querying multimedia content based on keywords or concepts. Each word contained in queries or in description of multimedia is disambiguated by using the WordNet in order to define its semantic concept.

1 Introduction

Multimedia information retrieval approaches are classified into three categories that are: text-based retrieval, content-based retrieval and multimodal retrieval [1]. In text-based retrieval approach, multimedia content is described by a number of keywords. In content-based retrieval approach, various low-level features like color, texture and shape are extracted for describing image and video, or spectrum of a signal to describe audio content. In multimodal approach, high-level and low-level information are combined to improve description of multimedia content.

Some work has been done to combine ontologies with visual features [20], for example Hoogs et al. [11] linked ontologies and visual features by manually extending WordNet with tags describing visibility, different aspects of motion, location inside or outside, and frequency of occurrence. [10] A visual ontology contains general and visual knowledge from two existing sources: WordNet and MPEG-7. Bertini et al. [3] suggests a "pictorially enriched" ontology in which both linguistic terms and visual prototypes constitute the nodes of the ontology.

Multimedia annotation has become a very important task for multimedia content semantic description; it can be done either manually or automatically. One of the technical manual annotation is the collaborative tagging, Flickr and Facebook are examples of systems that allow users to annotate multimedia content [14]. However, this technique suffers from some disadvantages, we can mention:

Amrane A., Mellah H., Amghar Y. and Aliradi R..
Indexing Multimedia Content for Textual Querying: A Multimodal Approach.
DOI: 10.5220/0004576200030012
In *Proceedings of the 2nd International Workshop on Web Intelligence* (WEBI-2013), pages 3-12
ISBN: 978-989-8565-63-1

- Indexing subjectivity: two annotators do not produce systematically the same annotation for the same multimedia content.
- Language dependence: annotation is generally achieved in the language of the annotator.

A critical point in the progress of content-based retrieval is the semantic gap[1], where the meaning of an image is rarely self-evident [12]. Detection techniques of semantic concepts by classification methods based on supervised learning have been proposed to reduce this semantic gap [16].

To retrieve multimedia contents, several query languages have been proposed. In the work of Richard Chbeir [5], the query languages are classified into three generations: textual, graphical and visual languages. The first generation of languages was based on textual resources (keywords, free text). The latter allows the user to easily express his information needs and it is adapted to all media; while the second generation deal with the graphical languages, the most famous query languages of the latter is the QBE (Query By Example) language [7].

In order to improve the multimedia search and allow textual querying, while bridging the semantic gap between user's needs and content description, we propose a model for semantic search of multimedia content that combine contextual information with visual multimedia content.

In order to achieve the latter process, we have organized our paper as follows: In the second section, we have described the existing works proposed in multimodal retrieval systems. The proposed approach is presented in the third section, where our indexing techniques and querying method are detailed. Finally, section four describes the results of our experiments.

2 Related Works

Multimodal retrieval approach combines semantic and visual features. In addition to visual content, the semantic content (keywords, manual annotations ...) is analyzed and put under adequate representation. Most of the research works in text/image information retrieval have shown that combining text and image information even with simple fusion strategies, allows us to increase multimedia retrieval results [6]. Two methods of fusion are used: early fusion and late fusion. The early fusion method consists in concatenating both image and text feature representations [6]. The visual and textual information are taken into account simultaneously in the different treatments. The late fusion is consisted to separately treat the visual similarity and the textual similarity [23]. Two ordered lists of results are obtained and should be merged by a suitable method before presenting them to the user.

Several works have been proposed in the framework of information fusion textual and visual multimedia retrieval: Belkhatir et al. [2] have proposed the combination of textual image retrieval with query by example (QBE); Lemaitre et al. [15] have proposed the combination of visual and textual information for multimedia information retrieval; Tollari et al. [22] used a Forest of Fuzzy Decision Trees (FFDTs) to auto-

[1] The gap between the low level description and semantics of visual content.

matically annotate images with visual concepts to improve text-based images retrieval.

The visual query languages suffer from a major problem which is the ambiguities of communication between man and machine [8]. The disadvantage of the graphical languages that is the formulation of complex queries consisting of combinations of several criteria remains a weak point for these languages [5]. The advantage of textual query languages is that they give the user the possibility to pose queries in a high-level language in allowing him to express his information need easily [17]. Users often express their queries in a textual description representing high level concepts [4], In this case, the use of lexical ontology becomes necessary to align the user query and multimedia contents in the same repository.

3 The proposed Model

In the model MMSemSearch (Multimedia Semantic Search) that we propose, the type of media considered is the image and this restriction does not exclude its extension to other types of media such as video or audio. To address the problem of semantics, two types of ontologies are used: a lexical ontology and multimedia ontology. These two ontologyies are used in the extraction phase of semantic concepts.

The following figure (Fig. 1) describes the system of architecture:

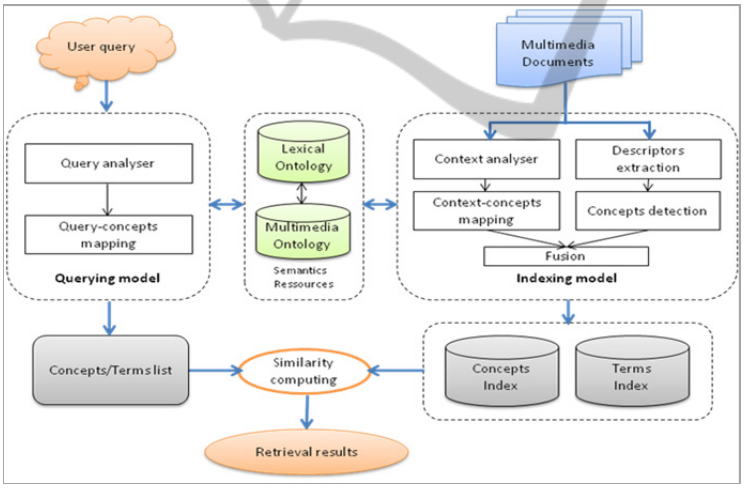

Fig. 1. Architecture of automatic indexing and semantic search of multimedia information.

In the proposed architecture, three main models are considered:

- Indexing model,
- Querying model,
- Matching model.

The proposed system for multimedia information indexing and retrieval use automatic extraction of semantic concepts combined with textual information that is extracted from surrounding text of multimedia resource.

3.1 Indexing Model

Indexing by Context. Two steps to indexing the surrounding text of multimedia are considered:

- Context analysis, which is to extract the words from the context of the image, removing stop words, stemming the words and the weighting of terms used,
- Context-concepts mapping, which consists to identify concepts in the lexical ontology by a disambiguation process, the terms are not identified as stored in the corresponding index.

In the work of Heng Tao Shen et al [9], four parts of the textual information (surrounding image) of the web page are listed and used to represent the image, these parts are:

- Image title: Image file title (simply image title) is a single word that basically indicates the main object that the image is concerned with.
- Image ALT: (alternate text). The image ALT tag in HTML document is a phrase that usually represents an abstract of the image semantics.
- Image caption: The image caption usually provides the most semantics about an image. It is the image's surrounding text in the HTML document. It can range from one sentence to a paragraph of text that contains many sentences.
- Page title: Since images are used for enhancing the Web page's content, page title is most probably related to the image's semantics. It is usually a short sentence that summarizes the Web page's content.

We have just used these four parts to represent image content, this information is projected on the lexical ontology. Terms having an entry in the ontology are taken afterward as the concepts describing image, against the terms in the ontology unidentified are kept in the index of terms. We have not ignored the terms that we have not find them corresponding concepts in the ontology as these terms may be important, such as proper names or neologisms (a neologism is the phenomenon of creating new words).

For automatic sense disambiguation of words we adapt in our experiments to use a simplified Lesk algorithm proposed by Adam Kilgarriff and Rosenzweig [13], where meanings of words in the text are determined individually, by finding the highest overlap between the sense dictionary definitions of each word and the current context.

Indexing by Content. The content is indexed by an automatic annotation process of semantic concepts; it is done in three steps:

- Extraction of low-level descriptors,
- The training of the classifiers (learning)

- Prediction of concepts.

Given the image descriptors, a classifier is applied to predict the classes of the test images. The parameters of the classifier are trained on the training data and tuned using the validation data [18]. Some examples of classes are: airplane, horse, person, motorcycle.

Currently, support vector machines (SVM) [24] are the most frequently used classifiers for the detection of concepts [21]. Therefore we chose the SVM classifiers as the classification method. For training classifiers and predicting concepts, we extract from all images, the SIFT (Scale Invariant Feature Transform) features under 16x16 scale.

Weighting of Concepts and Terms. Once the concepts/terms are extracted, the weighting step allows estimate the important concept and thus to classify images from their information.

We have used a variant of TF-IDF for weighting concepts. The concept weight is measured by its occurrence frequency in the context and in the image visual content, whereas terms weight depends only on the term frequency in the context of the image.

To calculate the degree of similarity of the images relatively to the user query, arithmetic sum of concepts and terms weights is used.

In our approach we call cf.idf (c, d) the weight of a concept c with respect to the image (the concept c is either identified in the lexical ontology or not), this weight is calculated by the following formula:

$$cf.idf_{c,d} = cf_{c,d} \times log\left(\frac{N}{df_c}\right) \quad (1)$$

Where $cf_{c,d}$ is the frequency of concept in the context of the image and in visual content,

N is the number of images in the database,

df_c is the number of images containing the concept c.

$cf_{c,d}$ is calculated by the following formula :

$$cf_{c,d} = \alpha.cf_{c,d}(content) + (1-\alpha).cf_{c,d}(context) \quad (2)$$

Where $\alpha \in [0..1]$, is a parameter for adjusting the weight assigned to each modality. It can play the role of confidence score assigned arbitrarily to context information and SVM classifiers, the value of α is determined by experimentation,

$cf_{c,d}(content)$ is the frequency of the concept in the visual content of the image (the concepts are detected by the SVM classifier, the weight of each concept can match the frequency of each word extracted from the visual image),

$cf_{c,d}(context)$ is the frequency of the concept in the context of the image (the concepts are derived from the mapping context with lexical ontology).

It happens that a concept appears simultaneously in the context and in the content of the visual image, in this case its weight (cf.idf) will be higher compared to other images containing that concept but only on one of the two modality context or visual content.

Concerning the terms that are not identified in the lexical ontology, the weighting formula derived from previous formulas (1) and (2), is as follows:

$$tf.idf_{t,d} = \alpha.tf_{t,d}(context) \times log\left(\frac{N}{df_t}\right) \quad (3)$$

Where N, df_t and α are the same parameters described in the above formulas.

3.2 Query Model

The multimodal retrieval approaches propose to users a query language that combines two modalities of information. To express their needs, users must provide for the system keywords or image as an example. Then, the system combines the two pieces of information to retrieve the images corresponding to the query. Our semantic retrieval model is based only on textual queries. The user formulates his query by simple keywords which will be analyzed and identified or not identified as ontological concepts.

Concepts or keywords contained in the user request will be used to build a new SQL query that uses two indexes proposed in our model.

We give an overview on the content of database that we have used in our experiments:

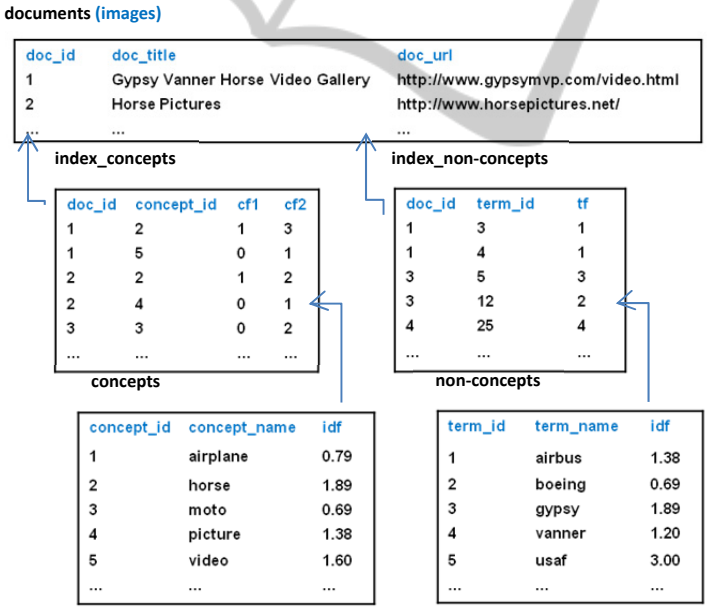

Fig. 2. Overview of database contents.

3.3 Matching Model

The proposed research system is based on semantic search; it allows the user to for

mulate his query with simple keywords. These keywords can identify semantic concepts in the ontology vocabulary or simple words. The matching between multimedia contents and the query image is established in order to classify the search result in order of relevance. The similarity score of an image based on a query q is calculated by the following formula:

$$Sim_{q,d} = \sum_{t \in q \cap d} tf.idf_{t,d} + \sum_{c \in q \cap d} cf.idf_{c,d} \quad (4)$$

Where t is a term and c a concept in the lexical ontology.

4 Experimental Evaluation

To evaluate the proposed indexing and retrieval model, we have used a collection consisting of 125 web pages containing a set of 200 images with contextual information for testing.

The learning database used for the detection of concepts consists of 500 images, as follows:

Table 1. Training and testing data for image classification.

Concepts	Aircraft	Motorcycle	Car	Person	Horse
Training	100	100	100	100	100
Tests	22	25	35	50	68
Total	122	125	135	150	168

In the indexing phase by the context, we have used the Porter algorithm [19], the Lesk algorithm [25] and WordNet ontology to give a semantic representation of images. Concepts are weighted by TF-IDF formula.

To index the image content, we calculated the SIFT (Scale-Invariant Features Transform) descriptors [4], which are invariant to scale changes and rotations.

Here is a query example executed by the realized system. The query consists of the keyword "horse", which is then translated into a SQL format, to be executed. The corresponding SQL query is:

```
SELECT i.doc_id, doc_title,
       sum((α*i.cf1+(1- α)*i.cf2)*c.idf)
FROM index_concepts i, documents d, concepts c
WHERE d.doc_id=i.doc_id and c.concept_id=i.concept_id and
c.concept_name in ('horse')
GROUP BY doc_id, doc_title
```

The results returned by the system are sorted by relevance and focuses on the context and visual content of images.

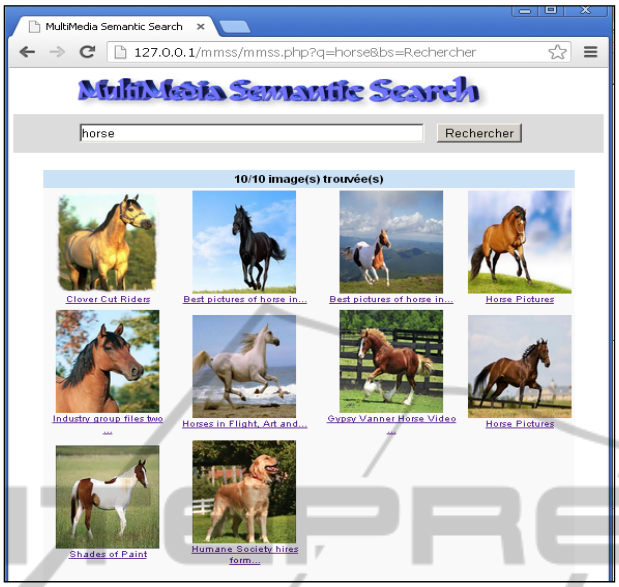

Fig. 3. Results returned by « horse » query.

To evaluate our model, in a first time of experimentations we tested a single modality context or concepts. In a second time, we combined the two methods. The search system uses two indexes built in the indexing process.

Several experiments were realized to set the value of the score of confidence α, the value which gave a better precision is the value α = 0.40, which attaches importance to information as context more than visual content for weighting concepts (the number of concepts detected is low compared to concepts extracted from visual content of images).

The evaluation measures calculated are precision and recall (Fig. 4).

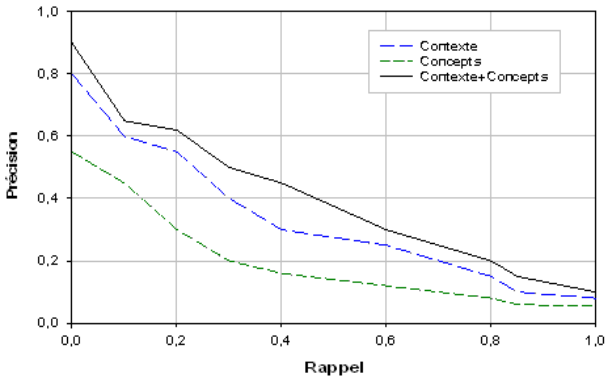

Fig. 4. Curves Recall / Precision for the three modality of retrieval.

In this graph it is clear that the combination of context information and semantic concepts slightly improves the image retrieval. However, to generalize the use of this method requires a lot of effort for the establishment of several classifiers.

5 Conclusions

The work presented remains in the area of multimedia information retrieval based on semantic concepts. A uniform representation of images and query content is done, to allow the user expressing his needs on a semantic level. Our contribution is mainly concerned with the following points:

- A conceptual representation of images,
- Semantic concept detection method based on SVM classifiers,
- A fusion model of textual and visual information,
- A contribution to an expressive textual query language.

Among the perspectives considered we quote:

- Improve the disambiguation technique of verbs in the sentence with the lexical ontology mapping,
- Diversify the number of semantic concepts detectable by the classification methods,
- Intend introducing the relevance feedback function in the retrieval process to refine the learning of concepts process.

References

1. Bannour, H., A Survey of Image Retrieval Approaches and their limitations, Report of Laboratoire Mathématiques Appliquées aux Systèmes, 2009.
2. Belkhatir, M., Mulhem, P. and Chiaramella, Y., A Conceptual Image Retrieval Architecture Combining Keyword-Based Querying with Transparent and Penetrable Query-by-Example, Springer-Verlag, pp. 528-539, 2005.
3. Bertini, M., Torniai, C. and Del Bimbo, A., Automatic video annotation using ontologies extended with visual information, ACM Multimedia, pp. 395-398, 2005
4. Brilhault, A., Indexation et recherche par le contenu de documents vidéos, Grenoble, France, 2009.
5. Chbeir, R., Modélisation de la description d'images : Application au domaine médical, 2001.
6. Clinchant, S., Ah-Pine, J. and Csurka, G., Semantic combination of textual and visual information in multimedia retrieval, Proceedings of ICMR, pp. 44-44, 2011
7. Del Bimbo, A., Visual Information Retrieval, Morgan Kaufmann, 1999.
8. Favetta, F. and Aufaure-Portier, M., About Ambiguities in Visual GIS Query Languages: a Taxonomy and Solutions, Proceedings of the Fourth International Conference on Visual Information Systems (VISUAL'2000), pp. 154-165, 2000.
9. Heng, T. S., Beng, C. O. and Kian, L. T., Giving meanings to WWW images, The eighth ACM international conference on Multimedia, p. 39–47, 2000.

10. Hollink, L. and Worring, M., Building a visual ontology for video retrieval, Proceedings of ACM Multimedia, pp. 479-482, 2005.
11. Hoogs, A., Rittscher, J., Stein, G. and Schmiederer, J., Video content annotation using visual analysis and a large semantic knowledgebase, IEEE Int'l Conf. on Computer Vision and Pattern Recognition, vol. 2, pp. 327-334, 2003.
12. Jonathon, S.H., Patrick, S., Paul, L., Kirk, M., Peter, E. and Christine, S., Bridging the Semantic Gap in Multimedia Information Retrieval, Workshop on Mastering the Gap, From Information Extraction to Semantic Representation, 2006.
13. Kilgarriff, A. and Rosenzweig, R., Framework and results for English SENSEVAL, Computers and the Humanities, p. 15–48, 2000.
14. Kim, H., Roczniak, A., Lévy, P. and El-Saddik, A., Social media filtering based on collaborative tagging in semantic space, Multimedia Tools Appl, pp. 63-89, 2012.
15. Lemaitre, C., Moulin, C., Barat C, C. and Ducottet, C., Combinaison d'information visuelle et textuelle pour la recherche d'information multimédia, GRETSI2009, 2009.
16. Liu, Y., Zhang, D., Lu, G. and Ma, W., Asurvey of content-based image retrieval with high-level semantics, Elsevier J. Pattern Recognition, no. 40, pp. 262-282, 2007.
17. Mulhem, P., Lim, J.H., Leow, W.K. and Kankanhalli, M., Advances in digital home photo albums, 2004.
18. Nowak, S., Hanbury, A. and Deselaers, T., Object and Concept Recognition for Image Retrieval, vol. The Information Retrieval Series, no. 32, 2010.
19. Porter, M. F., An Algorithm for Suffix Stripping, Program, vol. 14, no. 3, pp. 130-137, 1980.
20. Snoek, C. G. M., Huurnink, B., Hollink, L. and Rijke, M., Adding Semantics to Detectors for Video Retrieval, IEEE Transactions on Multimedia, vol. 9, no. 5, pp. 975-986, 2007.
21. Snoek, C. G. M. and Worring, M., Concept-based video retrieval, Foundations and Trends in Information Retrieval, p. 215–322, 2009.
22. Tollari, S., Detyniecki, M., Marsala, C., Tabrizi, A., Amini, M. and Gallinari, P., Exploiting Visual Concepts to Improve Text-Based Image Retrieval, ECIR '09 Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval, pp. 701-705, 2009.
23. Torjmen, M., Approches de Recherche Multimedia dans des Documents Semi-Structurés : Utilisation du contexte textuel et structurel pour la sélection d'objets multimedia, 2009
24. Vapnik, V. N., The Nature of Statistical Learning Theory, New York, 1999.
25. Vasilescu, F., Désambiguïsation de corpus monolingues par des approches de type Lesk, 2003.