COMBINING TEXTUAL AND VISUAL THESAURUS FOR A MULTI-MODAL SEARCH IN A SATELLITE IMAGE DATABASE

Sahbi BAHROUN[1,2], Nozha BOUJEMAA[2] and Ziad BELHADJ[1]

[1] Institut Sup'Com Tunis, Unité de Recherche en Imagerie Satellitaires et ses Applications, Route de Raoued km3.5, 2083 ElGhazala Ariana, Tunisie
[2] INRIA Rocquencourt, B.P. 105 78153 Le Chesnay Cedex, France
sahbi.bahroun@inria.fr, nozha.boujemaa@inria.fr, zied.belhadj@supcom.rnu.tn

1. INTRODUCTION

Satellite images are more and more numerous and with increasing resolution. Hence it is urgent to develop tools able to (semi-) automatically process such images in order to efficiently exploit them. When searching satellite images by visual content, in most of the cases the query image resides in the mind of the user as a set of subjective visual patterns, psychological impressions or "mental pictures". In this paper, we enrich the visual description by introducing to the user "the page zero" which is multiple visual patches summarizing the image database. Since there is no perfect description of visual content of images, most methods try to find a good compromise in balancing the image description between low and height level features. However, there exists an evident semantic gap between the demanding of user and the representation of low-level features [6]. Text-based methods are very powerful in matching context but do not have access to image content. The basic idea of our work is to combine high level and low level features to find a class of 'similar' images with similar keywords and with similar descriptors. In this paper, a novel content-based image retrieval framework is introduced. We propose three retrieval strategies by typing a keyword, selecting a visual thesaurus or by using the multi-modal description. We can prove that the multi-modal approach is able to retrieve complex concepts better than standards visual or semantical approaches taken separately.

2. REGION EXTRACTION AND DESCRIPTION

Satellite images are characterized by a great huge of content. The identification of the different regions located in remote sensing images is performed through the classical approach of image classification or cut it into overlapping patches. Clustering is a very hard task in the case of satellite images. Regions contours are never well defined especially for weak resolution images. So we decided to decompose satellite images into patches with size 64x64 pixels. This size was chosen experimentally. Smaller areas will probably provide more homogeneous textures and therefore could be better for fine textures like forest, fields and sea. Larger areas allows to have better estimates when they cover homogeneous fields, they may also capture textures made of large grains (cities) but increases the computational complexity for features. These patches are then used to compute visual features. Several recent studies and researches on features computation focused on the combination of global and fine local region description using points of interest. The main benefit of global approaches is that they can catch the global aspect of objects in images whereas points of interest [11] are located on parts of the image that have significant local photometric variation and are suitable to carry out salient details search. The visual feature that we used in our work is the Local Binary Pattern Correlogram (LBPC) that we introduced in [7]. The LBPC estimates a joint distribution of local and relational properties. The LBPC expresses how the spatial correlation of pairs of Local Binary Patterns [10] extracted around interest points changes with distance. The LBP feature performs well for local description around interest points. On the other hand, the correlogram [3] is good for global region description.

3. VISUAL THESAURUS GENERATION

Fauqueur and Boujemaa [2] introduced a new Query By Visual Thesaurus paradigm that enriches the Query By Visual Example by the creation of the"page zero". This paradigm stipulates that if the user has forgotten his query image or simply has a vague idea about it, he can compose his query by selecting several visual patches on a Visual Thesaurus (VT). VT is then a new query alternative that overcomes the absence of starting example image and offers the possibility to combine multiple visual patches in order to retrieve the target mental image. The Visual Thesaurus is obtained by means of relational feature clustering [1]. Relational feature clustering algorithms propose an alternative to partition a set of images into clusters based on their visual feature similarity.
It is a summary of all the regions in the database; each category contains similar regions according to low-level visual features. The Figure 1 shows an example of a visual database thesaurus (page zero) performed on the IKONA system [5].

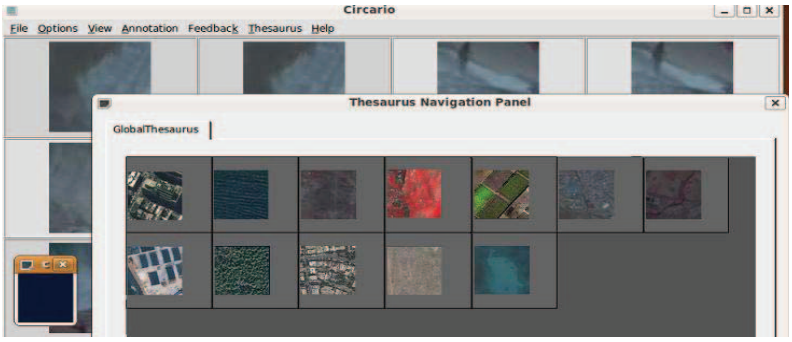

Circario
File Options View Annotation Feedback Thesaurus Help
Thesaurus Navigation Panel
GlobalThesaurus

Figure 1: Example of the visual thesaurus generation from our satellite image database.

In this application, finding the optimum number of clusters (representing visual thesaurus) is not critical as long as it is large enough to avoid lumping different profiles into one cluster. Here, we report the results when C=200. After clustering, we obtained a set of 12 visual thesaurus representing different image regions. The urban region, for example, is represented with different spatial resolutions because it is very difficult to find a similarity between two urban regions with a large difference in spatial resolution (1m/pixel and 20m/pixel).

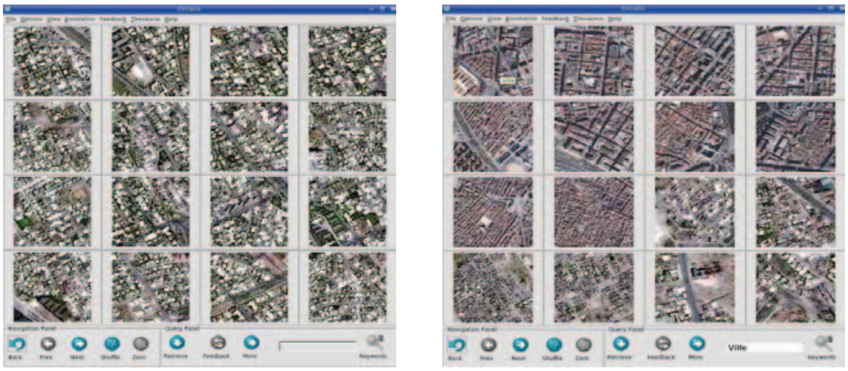
4. MULTIMODAL SEARCH AND IMAGE ANNOTATION

Image retrieval procedures can be divided into two approaches: text-based image retrieval and content-based image retrieval (CBIR) as we have seen in paragraph 2. In traditional text-based systems, images are manually annotated by human labelers and then searched using annotated keywords. The main disadvantage of traditional text based approach is that manually annotating large quantity of images is too tedious and time-consuming for common labelers. Although CBIR has been extensively studied since 1990s, the semantic gap [6] between low-level image features and high-level semantic concepts is still the key hindrance in the effectiveness of CBIR systems. Given the drawbacks of the visual and textual approaches taken independently, it became necessary to combine these two approaches [4]. In this work, we target at the problem of bridging the semantic gap in content-based image retrieval on satellite images. The multimodal research [9] uses conventional paradigms of queries by combining visual and textual information. This approach provides the user with new tools to explore an image. The user can perform his search by combining the text and the visual features into one feature. The visual and textual signatures are used in a single representation. This is called early fusion.
In this work, we target at the problem of bridging the semantic gap in content-based image retrieval on satellite images. The multimodal research uses conventional paradigms of queries by combining visual and textual information. This approach provides the user with new tools to explore an image. Therefore, content-based image retrieval has to be conducted with a unique feature composed by visual and textual features. The image annotation is performed semi-automatically with a controlled vocabulary and based on a set of manually annotated images (train database). The strategy of semi-automatic image annotation is better than manual annotation in terms of efficiency and better than automatic annotation in terms of accuracy.

The image annotation was achieved by combining the analysis of visual content of these images and the manually annotated training database. The training database is composed by 6 keywords clusters: city, field, sea, desert, forest and cloud (data given by CNES). We selected these regions because they are the most representative regions in our database. Each class has 100 textures thus we have 600 samples. We tried to transfer the human expertise to our system through supervised learning algorithm, the Support Vector Machine [8] that can interpret the human decision and automates the annotation of satellite images. In SVM, given a set of labeled examples selected from a finite set, an inductive procedure builds a function that (hopefully) is able to map unseen instances to their appropriate classes. The information provided by the visual descriptors and that provided by keywords are different and complementary. Visual information indicates the content of the image while the text information indicates the possible meanings for this content. In the annotation process, non labeled images are visually compared to the annotated images in order to obtain several annotations of visually similar images. For this purpose, we compute a set of low level features for the non annotated images. In our work, we set the visual feature as a concatenation of 7 color and texture features extracted from the IKONA system [5]. We tried to use the maximum of information to characterize the visual content of the non annotated images. We don't care about the computation time because the annotation process is performed off-line. Afterwards, the same 7 features were computed on the training set. To transfer the annotations from the training set to the non annotated images, we used a multi-class support vector machine (SVM) [8].

5. EXPERIMENTS

The objective of the present experiments was to evaluate the effectiveness of our proposed approach. The satellite image database is composed by about 50000 multispectral SPOT images with spatial resolutions ranging from 1m to 20m per pixel. First, each satellite image is decomposed into equal sized patches. Second, each region is characterized by the LBPC [7] texture feature. The total number of feature vectors representing all image regions was clustered using CARD [1] to extract the visual thesaurus. After extracting visual thesaurus, image regions will be annotated by keywords. A total of 6 words (city, field, sea, desert, forest and cloud) were used to annotate image regions. Each region was characterized by a total of 7 feature vectors from the IKONA system to enhance the annotation process. We exploit many global features and extract the most descriptive ones, which effectively capture the intensity, texture and color information from the satellite image content. All these visual descriptors are fused into one feature vector to annotate images by a multi-class SVM [8]. The textual feature set consists of a 6-Dim vector that indicates the presence/absence of each keyword. Therefore, in this work, we present a content based satellite image retrieval system where images are described by 12 visual thesaurus and 6 keywords. The user can initiate his query by selecting a visual thesaurus, selecting a keyword or by combining the textual and the LBPC into one feature. Feature computation and image searching was performed on the IKONA system [5] as shown in Figure 2.



Ville

Figure 2: Example of 2 queries. The first query (left image) is only using a visual approach. The second query (right image) is using the multimodal approach.

Figure 2 shows 2 examples of queries on the IKONA system. The first query is based only on the visual thesaurus. The second query uses the multimodal search with the keyword (ville). We can notice that in the second query, we can find images with different spatial and spectral resolutions (top left image). For example if the query is an urban area has spatial resolution 1m/pixel. It is very difficult with a visual feature to find result regions urban areas with spatial resolutions 10m/pixel or 20m/pixel. This was achieved by the use of the multimodal approach. The multimodal search allows a significant improvement in the quality of results. The performance was evaluated according to the averaged image retrieval accuracy versus a sequence of queries, which are applied in the following order: *keyword, visual thesaurus and multimodal.* The accuracy is defined as the ratio of relevant images in the top *T* retrieved images. The averaged accuracy is simply the average of the accuracies measured for the 1600 randomly selected test queries. The best retrieval accuracies were obtained by the multimodal approach. The obtained results showed that the multimodal approach outperforms the textual approach by 63% and outperforms the visual thesaurus approach by 71% in Mean Average Precision at top 160 best results. So, we can conclude that the multimodal approach is better in retrieval accuracy than the visual and the textual approaches taken independently.

6. REFERENCES

[1] H. Frigui, C. Hwanga, F. Chung-Hoon Rhee, Clustering and aggregation of relational data with applications to image database categorization. Pattern Recognition journal 40 (2007) 3053 – 3068.

[2] J. Fauqueur and N. Boujemaa. New Image Retrieval Paradigm: Logical Composition of Region Categories. In IEEE International Conference on Image Processing, September 2003.

[3] J. Huang, “Color-Spatial Image Indexing and Applications,” PhD thesis, Cornell University, 1998.

[4] M. Inoue, On the need for annotation-based image retrieval. In Workshop on Information Retrieval in Context (IRiX) Sheffield, UK, 2004.

[5] N. Boujemaa and al., Ikona: Interactive generic and specific image retrieval. In International workshop on Multimedia Content-Based Indexing and Retrieval, 2001.

[6] R. Zhao, & W .I. Grosky. 2002. Narrowing the Smantic Gap Improved Text-Based Web Document Retrieval Using Visul Features. Pages189_200 of: IEEE Transactions on Multimedia, vol. 4

[7] S. Bahroun, Z. Belhadj, N. Boujemaa, Texture based satellite image indexing by Local Binary Pattern Correlograms. In ACM International Conference On Internet Multimedia Computing and Servive. Kunming, Yun nan, China November 2009.

[8] S. Taylor Nello Critianini. An Introduction to Support Vector Machines. The Press Syndicate of the University of Cambridge, 2002.

[9] S. Tollari. Filtrage de l’indexation textuelle d’une image au moyen du contenu visuel pour un moteur de recherche d’images sur le web. In Actes d’ACM Confrence en Recherche d’Informations et Applications (CORIA’05), Grenoble, France, mars 2005.

[10] T. Ojala, M. Pietikäinen and T. Mäenpää, “Multiresolution gray-scale and rotation invariant texture classification with local binary patterns”, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002.Volume24, 971–987.

[11] V. Gouet and N. Boujemaa. Object-based queries using color points of interest. In Workshop on Content- Based Access of Image and Video Libraries (CVPR), 2001.