Automatic generation of natural language descriptions for images.

Patrick Hède (1), Pierre-Alain Moëllic (1), Joël Bourgeoys (1), Magali Joint (1), Corinne Thomas (2)

(1) C.E.A. Fontenay-Aux-Roses (SCRI/LIC2M). (2) E.S.I.E.A.
patrick.hede@cea.fr, pierre-alain.moellic@cea.fr

Abstract

Image annotation is frequently used in image base management. Unfortunately, manual keyword indexing is costly and not exempt of errors for large image bases. In this article, we present a method for automatic image description in natural language for images without problems of occlusion. This method relies on a double expertise in image indexation and natural language processing and generation.

Keywords : image processing, segmentation, indexation, linguistic processing, natural language.

1. Introduction

How should images be automatically described ? Several studies have tried to answer that question in general or considered specific problems such as image description for blind people or the automatic generation of captions. When considering automatic image descriptions, decisions must be taken on: *1*- the nature of the description (keywords list or natural language text) ; *2* – the use of a global automatic method or a semi-automatic method with a user intervention (learning process).

Today, most of the methods reduce their scope to the generation of a keywords list, leaving out descriptions in natural language. From a semantic point of view, this choice implies a weak description but is explained by today's *Internet culture* which has popularized the document search by keywords (seldom hierarchical). Google™ is a perfect example of this kind of search engine. Nevertheless, the improvements in data mining, natural language processing and computer vision will necessarily make this tendency evolve into requests in natural language, which have a stronger semantic content.

Many methods rely on a semi-automatic process. Feedbacks like in [1] are used to improve the relation between low level data (from image treatments) and high level semantics (image databases already indexed by keywords). Unfortunately, feedbacks are usually not efficient enough to significantly improve the performance of the systems, especially if the techniques are not even able to treat simple semantics. Moreover, we can be skeptical about the use of semi-automatic methods by a common user. More and more systems opt for automatic algorithms, for instance by considering a probabilistic method [2], that is to say the probability of linking words to a region of the image. For instance, [2] and [3] use several more or less complex models based on the co-occurrence model. The segmentation techniques in computer vision still do not manage to treat rapidly and above all efficiently complex images (heterogeneous objects, occlusion, shadows, etc.), consequently, the generation of key-words with classical segmentation methods is not enough accurate above all for huge database (global results of [2] using Blobworld).

This article presents an automatic method for the generation of natural language descriptions of non-complex images. Describing an image with sentences requires an extraction of characteristics such as the main color, the shape and the texture of some regions (grass, rock, etc.). These criterions are extracted and characterized thanks to methods of images indexation by content like PIRIA [4] , QBIC [5] or VisualSeek

[6]. The criterions properly chosen can be linked to a strong semantic such as *red*, *grass*, *circle*. So, we have an automatic generation of keywords from picture characteristics. A higher level of semantic is reached by using a dictionary of objects indexed by few keywords translating concepts or objects which would be difficult (or even impossible) to extract by simple image processing. This data melting from the dictionary and the image processing enables the generation of sentences in natural language (grammatical coherence, no repetition, etc.) via linguistic techniques.
First of all, we present the database and the dictionary of objects we use for the tests. Then, we describe the methods of segmentation and indexation of images. In the third part, we present the linguistic processing which enables the generation and the analysis of the sentences. We conclude by explaining future works for processing more complex images.

2. Test base and general principle of the algorithm.

2.1 Images used, dictionary of objects.

Our test base consists of about 4000 images (Fig. 1). The images were shot with a digital camera. The background is always uniform but of different colors (black, white or gray) and the images represent different combinations of more or less complex objects (toys, pencils, etc.). Those objects build the dictionary: 100 objects shot in 5 different angles at most, that is to say a dictionary composed of about 400 images. The goal here is to have a dictionary of limited size : the over-representation of an object often implying more noise than useful information. Hence the importance to find a coherent number of representations of the object according to its complexity (above all its dissymmetry) and according to the capacities of discrimination of the indexing technique (Cf. part 4).

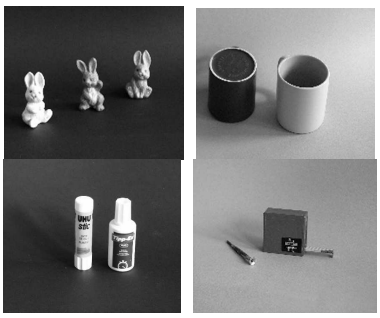


fig.1 : Four examples of treated images.

Each object is indexed by an *image signature* (from color and texture characteristics) and two keywords: name and category (fig. 2). There is no problem of occlusion in the images. Thanks to this constraint, a good segmentation is obtained, giving a good knowledge of the number of objects and their position.

SIGNATURE	
Image signature : *color + texture*	Keywords : `Toy, Rabbit`

fig. 2 : Five objects of the dictionary. Signature of the object.

2.2 Principle of the algorithm.

The first step of the algorithm is the image segmentation (Cf. fig.3), which isolates the different objects from the background. At the end of this step, the number of objects is known, as well as their absolute and relative position in the picture. The second step corresponds to the indexation of the segmented objects. For each isolated object, a query by example is launched in order to find the most similar object in the dictionary. This step is composed of two parts. First, the signature of each object is created according to criterions linked to shape, color and texture. Then the system searches for similar objects in the dictionary by comparing the signatures. The third step is the generation of the text description of the image using the keywords associated to the images of the dictionary.

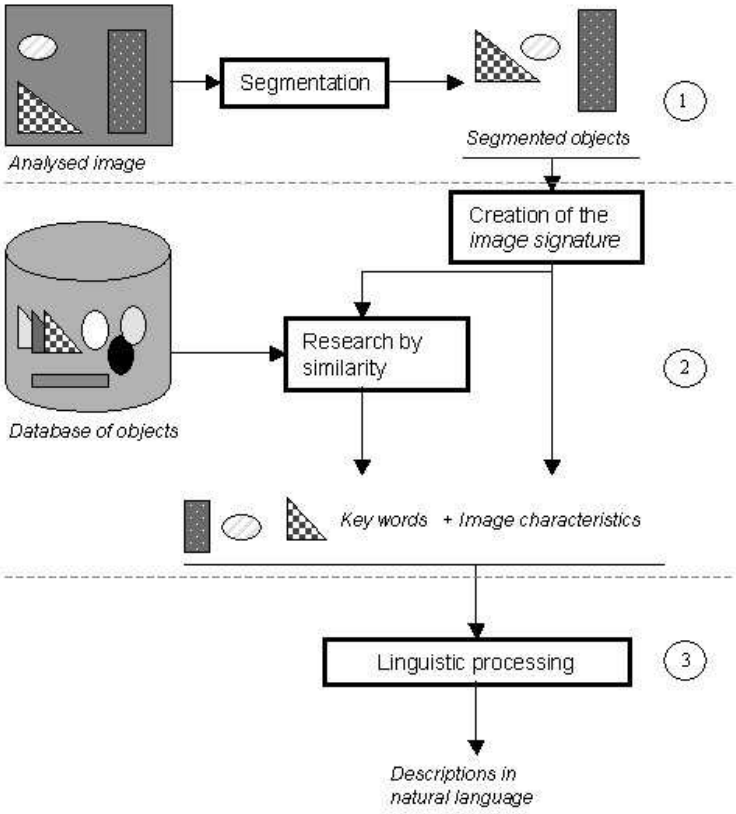


fig. 3 : The three steps of the algorithm.

3. Image segmentation.

3.1. Algorithm.

One of the objectives is to use a time effective method to segment the objects, considering the fact that the complexity of the images treated here is globally weak. The general idea is to binarize the image considering its background, then, by complementarity to isolate the objects (Cf. fig.4). We suppose that the image background has heterogeneities that are not too important. Different colorimetric criterions (homogeneity, continuity, etc.) are used by propagation from the four sides of the image. The background takes a 0 value and is removed from the image. Only the binary masks of the objects remain. The product of the binary mask and the original image gives the bounding box of each object. The area and the position of the image are computed for each object.

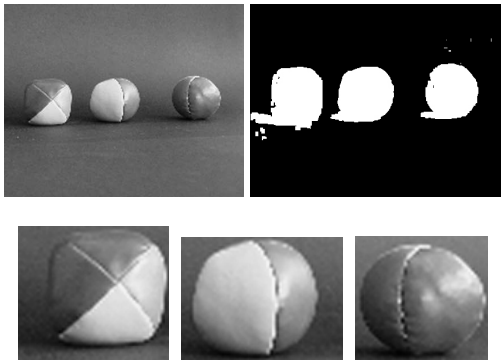


fig.4 :Up, segmentation of the image. Down, the three segmented balls.

3.2. Useful information.

At the end of the segmentation process[1], three types of information are available:

- the number of objects that compose the image,
- the localization of each object : absolute and relative position.
- the area of the object (the visual importance in the image).

From this information we can build a description of the *composition* of the image. The object with the largest surface area is considered as the main object of the image. We know its position perfectly (in the center, on the left, on the right, etc.) and the relative positions of the other objects which composed the image are also known.

4. Indexation of the segmented objects.

The quality of the results depends, for the most part, on the indexation method. Therefore, it is very important to have a method that is precise, fast and robust in order to have a good appropriateness between the textual description and the treated image. The choice of the method and therefore the choice of the representation of the objects have an influence on the size of the dictionary. The method described here matches objects from the image and the dictionary which have a difference in the shot of 30°. The mean error rate is very low for complex objects and null for objects which have rotation invariance.
The indexation of images is generally done considering three characteristics of the image: color, shape and texture. Here, the indexation is done thanks to the color and the texture characteristics. However, we will see that the analysis of the texture is done with the edges of the image and therefore with shape information.

4.1. Color characteristic.

Global histogram [7] is a very efficient method used extensively for color characterization [8]. When there is a large number of color spaces, the most difficult task is to find which color space is the best for the image processing. The model which uses brightness, saturation and hue (HSV model) has a better semantic impact. When a user describes a color, he gives a hue, an adjective for the brightness (bright or dark) and an adjective for the saturation (vivid). According to a European study [9] we normally define the color of different objects using the following eleven hues: *Red*, *Orange*, *Yellow*, *Green*, *Blue*, *Purple*, *Pink*, *Brown*, *Gray*, *Black* and *White*. The color division is used to obtain a strong semantic content: the color name. To refine the description, qualifiers are added: a “chromatic” color can either be *bright* (high brightness), *dark* (low brightness), *vivid* (high saturation) or *medium* (medium saturation and brightness). For an achromatic color, gray, black and white, only gray is dark, medium or bright. We have eight chromatic colors and four adjectives for each one, so we have thirty seven categories and we obtain a color and semantic quantization. We can then characterize an image by computing a color histogram H_{col}

[1] European C.E.A. patent. Inventors : Corinne Thomas, Patrick Hède, Hassane Essafi. Ref C.E.A. BD 1369. Date of registration : 7 June 2001. Registration number : 0107480.

normalized by the total number of pixel. H_{col} composes the image signature and is invariant by rotations, translations and scale transformation.

4.2. Texture information.

The indexation of the object according to the texture considers this one as the local relation between pixels. Ojala and Piettikaimen [10] define a descriptor (LBP descriptor) which analyses the local spatial structure of the texture. Here another descriptor is used, Local Edge Pattern (LEP [11]) which characterizes the texture in a same way but using the edge image therefore with a more *shape-based* aspect. First, we compute the edge image by thresholding the image result of a Sobel convolution and then by applying a 3x3 binomial filter.

-1	0	1
-2	0	2
-1	0	1

Sobel Gx

1	2	1
0	0	0
-1	-2	-1

Sobel Gy

1	2	4
8	256	16
32	64	128

Binomial filter

Let P(x,y) a pixel of the edge image and FB the binomial filter. We have the following inequalities (for a 256^3 colors image) :

$$0 \leq \sum_{i=-1}^{+1} \sum_{j=-1}^{+1} P(x-i, y-j).FB(i+1, j+1) \leq 512$$

So, it is easy to compute a texture histogram composed of 512 classes : H_{tex}. In fact, two histograms are computed whether the pixel is a part of an edge or not.

4.3. Signature and distance indexation.

The image signature is composed of two histograms. The distance used to compute the similarity between two objects is simply a histogram comparison. Let H_1 and H_2 two histograms (with Nc classes), the distance of similarity $d(H_1,H_2)$ will be :

$$d(H_1, H_2) = \sum_{i=0}^{Nc} \left| H_1(i) - H_2(i) \right|$$

So, the distance between two objects O_1 and O_2 will be:

$$d(O_1,O_2) = \alpha_{col} d(H^1_{col}, H^2_{col}) + \alpha_{tex} d(H^1_{tex}, H^2_{tex})$$

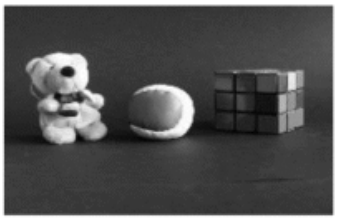
α_{col} and α_{tex} weight the importance of the texture and the color. We fix α_{col} et α_{tex} to 1.

5. Text processing

5.1 Automatic generation of descriptions

Online companies generally use keywords to describe the images they are selling. Unfortunately, a list of keywords brings ambiguity. For example, the keyword list definition of an image *blue, car* and *sky* is ambiguous. Is the car blue or is it the sky? This kind of keyword index, missing links between keywords, returns a lot of noise (non pertinent results). In a sentence, relations between words are explicit. In the sentence "*a car with a blue sky"*, it is clear that two objects have to be found, a car and the blue sky.
The fact that the recognition is based on the object makes for a non ambiguous identification (at the linguistic level). If certain preoccupations are taken into consideration to name the objects with words having only one meaning or unambiguous expressions, we increase precision at the time of the natural language interrogation. Moreover, if the dictionary is multilingual, it is possible to generate a description in each language without the problem of polysemic words.

The creation of sentences in natural language offers a clearer interpretation of the description supplied by the indexing system. The sentences are systematic, built using the information from the image processing.



"Photo horizontale de dimension 384 par 307 pixels ayant pour couleur dominante une couleur sombre. Le fond de cette photo est noir. Elle est composée de trois objets. Cette image a pour thème principal Jouet. L'objet principal est un nounours marron clair qui se trouve à gauche de la photo. Nous distinguons, au centre de l'image, une balle verte et blanche. Nous notons un cube multicolore à droite."[2]

fig.5 : Example of generated text in French.

In the example (fig. 5) the underlined words are determined by the information from the image processing. Basically, we have three categories of descriptors. The first category is the object names. For example, in fig.5, we have *nounours* (teddy bear), *balle* (bal) and *cube* (cube). The second category is the color information. In fig.5 we have *sombre* (dark), *noir* (black), *marron clair* (light brown), *verte et blanche* (green and white) and *multicolore* (multicolor). In the third category, we find spatial information. In fig.5 there is *horizontale* (horizontal), *384 par 307 pixels*, *à gauche* (left), *au centre* (center), *à droite* (right). Finally, the number of objects and the general theme of the image are added.

The types of information describing the images were not chosen randomly. A research [12] on the nature of the information used in the manual description of images in natural language showed that in 80% of the time, the color and the name of the object were used. Also, in every spoken language, prepositions exist to allow the location objects. Spatial localization is a strong semantic characteristic. It is interesting to indicate where an object is located when we are unaware of its name or we have little information on it.

Linguistic processing is essential to build syntactically correct sentences. Systematic construction of sentences often leads to errors. For example, the structure colour + noun usually gives good results as in the "*yellow banana"* but sometimes leads to phrases that should be avoided as in "*the orange orange"*. Moreover, the main object is chosen according to its surface, we have to avoid errors of construction like "*the table is under the bottle"* which should really be "*the bottle is on the table"*.

Our research group has built a new generation of indexing systems and multilingual querying that is based on, and improves significantly, the previous generation known as SPIRIT. The cross lingual aspect of the research was elaborated during the first European project between 1990 and 1994 (ESPRIT EMIR - European Multilingual Information Retrieval - project). At the present time, the system works on 6 languages (French, English, German, Spanish, Arabic and Mandarin Chinese) and will be upgraded in 2004 with Italian and Russian.

The system is based on a deep syntactic parsing and on named entity recognition. A statistical model is used to group documents by class and sort them by pertinence. The cross lingual aspect of the system is not necessary in the case of an indexation by image processing. However, in the case of large databases, it can be useful to have a textual description in only one language to save space and then use the cross lingual system to process queries in other languages. There is of course a risk that precision will be diminished.

[2] *"Horizontal picture of 384 by 307 pixels with a dark main color. The background of the picture is black. It is composed of three objects. The picture has Toy for principal subject. The principal object is a light brown teddy bear and is located on the left side of the picture. We distinguish, in the center, a green and white ball and on the right a multicolor cube. "*

The natural language querying of images is essential to find images that will serve as a query for other similar images. This second query is executed with the comparison of the signatures of the query image and the images in the database.

6. Evaluation

Evaluation on the retrieval of objects in images shows that images having restrictions mentioned previously (homogeneous background, objects without occlusions) gives good results. Testing with 500 images as query in a 7200 images database, we get a precision of 0,90 and a recall of 0,64. The errors always appear in the object recognition phase. No errors were found in the generation of color or the localization information.

An overall evaluation was also done. We gave a set of 100 images from our image base to 5 individuals and asked them whether the text generated was good, acceptable or wrong. In this evaluation, two aspects are verified: the relevance of the text in relation to the images and the quality of the textual description. A *good* result is a syntactically correct text which describes the objects present in the image as a common user would have done. We say that a result is *acceptable* if syntactic errors are present but the text is readable, understood and represents the image. Finally, by *wrong* result we mean a text which is totally incoherent and/or doesn't describe the correct images (confusion between different objects). From the 100 images, 88 were characterized good results, 3 were acceptable and 9 were wrong results.

We are aware that the evaluation needs to be completed by "recall and precision" on a larger image set. We are also aware that the evaluation of textual generation must consider less subjective criteria like the types of errors. In future evaluations, subjectivity will be kept to a minimum. However, these first results are very encouraging.

7. Future work

7.1. General complexity of the images.

Images can be arranged according to their visual complexity in four categories:
Category 1 : Binary images or artificial images. No shadow, no reflected or refracted lights.
Category 2 : Photographs or cliparts with homogeneous background.
Category 3 : Photographs with objects or persons well differentiated with few problems of occlusion.
Category 4 : Any complex images : landscapes, group scenes, paintings, etc.

fig.6 : From top to bottom and left to right, two images of category 1,2,3 and 4.

As we have seen, the segmentation of the images from categories 1 and 2 (3 in some cases) raises few problems because the objects are the complementary part of a background which is easily characterized. The complexity of the images of category 3 or 4 (heterogeneous objects and background, reflects, shadows, occlusions, etc.) makes the use of common segmentation techniques (edge or region based) difficult. Most of these techniques use criterions which refine iteratively the segmentation (regions growing, split and merge, snakes, etc.), these criterions usually give an *over-detailed* segmentation (one object is decomposed in several ones) or a fusion of different objects.

7.2. Strategies for complex images.

Today, one of the limitations of our system is not being able to process very complex images (category 4 and some images from category 3). We think it will be possible to correctly process all images from category 3 thanks to an efficient method of segmentation [4]. The generalization of category 4 goes through a different strategy notably by the generation of simple semantics ("*an orange triangular shape on a blue background*") which allows to put forward hypothesis ("*blue background*" implies *sea* or *sky*).

8. Conclusion

In this article we have presented a method for automatic generation of natural language descriptions for images. This method is based on image processing for segmentation and indexing. The robustness and precision of these methods are essential for the quality of the final results. The two techniques described offer very good results for images with restricted characteristics. The merger between these techniques and natural language processing techniques manages to generate a coherent and rich description in natural language which corresponds to what a user might find more intuitive. Furthermore, the indexing in natural language sentences improves the quality of the results by reducing the ambiguity present in a keywords index. Finally, this work represents a solid base for future research on more complex images.

9. References

[1] Xingquan Zhu, Liu Wenying, Hongjiang Zhang, Lide Wu, "An image retrieval and semi-automatic annotation scheme for large image databases on the Web", Microsoft Research China.
[2] J.Jeon, V. Lavrenko, R Manmatha, (2003), "Automatic Image Annotation and Retrieval using Cross-Media Relevance Models", *SIGIR '03*.
[3] Y.Mori, H. Takahashi, R.Oka (1999), "Image-to-word transformation based on dividing and vector quantizing images with words", *MISRM'99*, *First International Workshop on Multimedia Intelligent Storage and Retrieval Management*.
[4] M.Joint, P.A. Moëllic, P. Hède et P. Adam. (2004) "PIRIA: A General Tool for Indexing, Search and Retrieval of Multimedia Content". *SPIE.Electronic*
[5] M. Flickners, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker. (1995) "Query by image and video content: the QBIC system". *IEEE computer*, September.
[6] J.Smith, S. Chang. (1997) Querying by color regions using the visualseek content-based visual query system. *Intelligence Multimedia Information Retrieval AAAI Press*.
[7] M.J Swain and D.H Ballard. (Sept 1990). "Indexing via color histograms". *Image Understanding Workshop*, pp. 623-630.
[8] M.J. Swain and D.H. Ballard (1991). "Color indexing". *Journal of Computer Vision*, 7(1).
[9] J.P. Changeux (2002). L'homme de vérité. *Ed. Odile Jacob*.
[10] T. Ojala, M. Pietikainen. (1999) "Unsupervised texture segmentation using features distributions". *Pattern Recognition* 32, 477-486.
[11] Ya-Chun Cheng, Shu-Yuan Chen, (2003) "Image Classification using color, texture and regions", *Image and Vision computing* 21, pp. 759-776.
[12] C. Thomas, Thèse (PhD, sept. 2001) : Accès par le contenu à des documents numérisés contenant du texte et de l'image. C.E.A. et Université Paris VII, sept.