Review: Automatic Image Annotation for Semantic Image Retrieval

Hasna Abioui(✉), Ali Idarrou, Ali Bouzit, and Driss Mammass

IRF-SIC Laboratory, Ibn Zohr University, Agadir, Morocco
hasna.abioui@gmail.com, {ali.idarrou,a.bouzit,mammass}@uiz.ac.ma

Abstract. Nowadays, the number of digital data sets grows exponentially. Hence, the need to conceive efficient and powerful image indexation and retrieval systems grows as well. Automatic image annotation was adopted by several research as the emerging trend in image retrieval area. Actually, it is considered as the best solution that combines the content-based techniques by using low-level image features and text-based techniques exploiting textual annotations, associated to the image. In this way, the semantic gap between low-level image features and high-level semantics will be reduced. This paper presents a review of image retrieval approaches, by focusing especially on the automatic image annotation methods, in order to analyse the impact of annotations and associating semantics to the visual data for an image retrieval process.

Keywords: Image retrieval · Automatic image annotation
Semantic gap

1 Introduction

Due to the vigorous growth of the Web as well as digital technologies added to computer and mobile devices, huge amounts of data are generated. Even document contents changed and tend towards multimedia contents especially visual data which overpass textual data in terms of expressiveness, as said, a picture is worth a thousand words. All these reasons explain the enormous number of pictures that are created, stored and shared over web applications, social networks, and mobile devices. [1] classifies visual data into three main categories: personal pictures presenting individuals and their families, specific domain pictures describing one domain such as medical or panoramic ones; and finally, the web pictures published and shared through social networks and blogs. Hence, this image proliferation imposes an urgent intervention by researchers. So, a multitude of techniques and approaches aiming to retrieve and manage images are considered, since indexation and image retrieval (IR) remain challenging tasks. In the literature, images are traditionally retrieved by using the content-based image retrieval (CBIR) approaches, that are based on the low-level image characteristics like colour, shape and texture [2,3]. In fact, what is more challenging

A. Mansouri et al. (Eds.): ICISP 2018, LNCS 10884, pp. 129–137, 2018.
https://doi.org/10.1007/978-3-319-94211-7_15

concerning this category of approaches is the semantic gap between the image content as a visual data and its semantic interpretation [4]. Then the second category of approaches was appeared, which are keyword-based image retrieval approaches. These approaches treat the textual information that was previously attached manually to the image as annotations. Also here, two issues mainly occur: the impossibility of manually annotating large amounts of images, and the quality of human annotations that may be subjective and seen only from the annotator perspective. In order to overcome each category's issues, research in this area focuses on the combination of both low-level image features and high-level semantics. So, they opt for the automatic image annotation (AIA) as a typical solution providing semantic annotations by using machine learning techniques [5]. In this paper, we present a review on image retrieval approaches the pros and cons of each category, and focusing on the automatic image annotation since it represents the emerging trend for image retrieval. The rest of the paper is organized as follows: in Sect. 2 we will be presenting the image retrieval background and fundamental concepts, then we will compare various automatic image annotation techniques in Sect. 3, before discussing the presented survey in Sect. 4 in order to come out with a synthesis at the end.

2 Fundamentals of Image Retrieval

2.1 Image Representation: Features and Metadata

In image retrieval area, image is defined as a combination of physical attributes set, referring to the image content, and metadata set referring to its context [6]. Regarding the first category, the image is treated as an array of pixels providing the low-level features. Features that are commonly used by the content-based image retrieval (CBIR) community are: the image colours [7], the shape [8] and the texture features [9]. As for the second category, it presents attributes of meaning or metadata which define the image context, and qualified as high-level features [10]. We distinguish two types of image contextual properties: internal metadata extracted from the image itself, and external metadata collected from the text surrounding the image. These are all textual information reflecting the meaning of the image, and associated with it, in order to make it more meaningful and to show its implicit semantics according to the context in which it appears.

2.2 Image Retrieval Generic Processes

In this section, we will describe generic processes adopted by research community at the aim of retrieving images. But before, it is worthwhile to distinguish between the three types of image queries, that each generic process of the two that will be described below can exploit. An image query can be expressed by textual words which are more intuitive and natural for users to express their needs and expectations [11]. Also, it can be an image given by the user as an entry to the retrieval system, then it is obvious that only visual features will be

taken as comparison criterion. Another type of queries is expressed by sketches or designs of objects must be included in the retrieved image. Similar to queries in the form of an image, sketches queries impose that the comparison must be done according to low-level features with some tolerance related to them. Retrieval image processes are categorized into two main categories: the first one which accepts image and sketch type queries, known as content based image retrieval (CBIR) process. The second type accepts textual queries, so named: key-words based image retrieval process (KBIR).

CBIR Process: It consists of comparing – by referring to the low-level features extracted from pixels – the image provided by the user as a query with images in the collection. Generally, CBIR processes consist of two phases: indexation and retrieval. As for the first phase – which is an offline phase –, the objective is to represent each image from the collection as a set of visual feature vectors forming the image signature. Thus, images will be compared based on similarity measures using those signatures. So, two images are similar if they have similar signatures according to the used similarity measure [12]. The research phase – or the online process – consists of calculating the query image signature in accordance with the same adopted methods in the offline phase. This way, the query image will be compared to each image from the database by comparing their signatures based on dissimilarity measures [13].

KBIR Process: The keyword-based process has the same two steps: indexation and retrieval steps. As they are represented by using textual descriptors formed through collecting keywords, terms or any other metadata associated with the image as a description aiming to make it semantically interpreted during retrieving step. Retrieval step consists of comparing the textual query given by the user with textual descriptors of indexed images. Table 1 shows the advantages and disadvantages of each process.

Table 1. Advantages and disadvantages on image retrieval techniques

IR technique	CBIR process	KBIR process
Advantages	- The most practical for indexing and retrieving large amount of images	- Users express their queries easily by using keywords or sentences
	- Reduce ambiguities related to the textual indexation	- Most accurate, as semantic concepts are used to interpret images
	- Less time-consuming	
Disadvantages	- Low-level features are not able to describe and interpret semantically the image context	- Time consuming, expensive, and subjective
	- Unserviceable for general users, as they are required to provide query in the form of images	- Impractical when it comes to annotating large-scale image databases

For CBIR processes, image descriptors adopted in order to extract signatures are not universal. So, choosing to compare colours, shapes or even textures of images depends on the objective of the comparison. In addition, this latter itself is far from being able to semantically interpret the image. Otherwise, retrieving images based on their visual features is seen as the practical solution overcoming limits of keywords based image retrieval depending on manual annotation of images especially when it comes to annotate large-scale image databases, in that case, either manual or semi-automatic annotation of images become impractical and expensive, as they require human intervention. So, to overcome the limits of each process while benefiting from both, image retrieval community tends towards the image automatic annotation. The next section presents a study and comparison of the image annotation techniques.

3 Automatic Image Annotation for Image Retrieval

Automatic image annotation (AIA) is the process aiming to assign words to image based on its visual features. It combines both of image analysis and machine learning techniques, by taking advantage of the text-based annotation and CBIR, in order to reduce the semantic gap between low-level features and high-level semantics. The AIA basic principle consists in extracting semantic concept models automatically from image samples. Afterwards, the extracted models will be used to annotate new images. In this way, annotated images are retrieved by using textual queries and keywords. In the literature, there are several reviews studying and comparing image annotation techniques. However, each review is done according to specific purposes or needs leading to annotating images. [14] presents a study of AIA techniques used specifically for medical images. Authors insist when studying techniques to compare their automation level that can be automatic or semi-automatic, extracted visual features as they are the key entry point, while they offer a local, global and pixel by pixel comparison of medical images, methods used for the classification and the image type taken into account for each technique, for instance, radiology, MRI and X-RAY images. In [15], the study is oriented towards the device capturing the image. So, it gives more importance to images captured using mobile phones, which provide and improve the quality of contextual information used at the moment of annotation. In this paper, we shade lights on the essential techniques used to annotate images in order to simplify their retrieval process. As we described before, an image retrieval process can be based on visual features as it can be based on the associated metadata. The same when annotating an image, two main categories are distinguished according especially to the type of the image collections: the first one is based on the visual content to produce annotations. The second category can ignore any visual information that can be provided by the image itself, and exploit other metadata included in the collection. Generally, works coming within this category are interested more in semantics and context by identifying different interpretations and meanings extracted from the associated text.

3.1 Visual Features-Based Automatic Image Annotation

As stated before, AIA – in the case of collections that contain only images – relies on visual features of images. Hence, three methods are used to exploit those visual features [16]. (i) Global methods using the whole-image characteristics in order to provide a global distribution of image visual subjects. (ii) Local methods dividing the image into subsets or regions, and by using these methods, image visual features are separately extracted according to each subset. (iii) Hybrid methods are the combination of the two previous methods; so, it combines the advantages of both of them in terms of precision of objects extraction and detection. Once the visual content is identified through one of the three methods mentioned above, what remains is to analyse low level features and provide a semantic understanding of the content, by extracting the relationship between visual descriptors and identified classes from the image. Basically, approaches aiming to learn this relationship are either supervised or unsupervised learning techniques:

(i) **Supervised learning approaches** classify semantic concepts representing image annotations into defined and previously-known classes. Consequently, visual characteristics taken into account are set and limited before the learning phase which means that at the moment of the creation of the class. In the following, we cite the commonly used algorithms for supervised approaches. We start by the most popular classification algorithm especially for shape recognition, Support Vector Machine – SVM introduced by [17]. Works [18–20] use SVM as supervised learning technique for image classification and annotation purpose. [21] opts for the K-nearest neighbour algorithm to classify image regions. [22] uses Artificial Neutral Network which is able to make decisions about multiple classes at a time contrary to the previous algorithms which learn only one class at a time. [23] proposes a classifier designed using Decision Trees (DT) and Rough Sets (RS) to annotate images.

(ii) **Unsupervised learning approaches** which propose an important number of classes and concepts compared to the supervised learning approaches. Generally, those techniques use probabilistic models to define the relationship between visual features and textual information. [24] is one of the works adopting this approach, by proposing an image categorisation model discovering the image semantic content using probabilistic latent semantic analysis.

3.2 Metadata-Based Automatic Image Annotation

For images coming with text descriptions as the case of images in Web pages; external information are exploited and used in order to annotate the image. Several methods are proposed to annotate image from the web, some of them use only the image associated text, others combine external information and visual features in order to improve the accuracy of annotation. In [25], the proposed

system is based on at least one keyword and one image example as the entry of the process enabling on the one hand, text-based retrieval to search images that are semantically similar by comparing their descriptions provided by image titles surrounding texts..., on the other hand, the content-based retrieval aiming to rank and return visually similar images. [26] proposes also I-Tag, a system combining both visual and textual descriptors to annotate automatically image. In addition, the title of the image is used to recommend tags for the image for more accuracy and relevance. In [27] weighted nearest neighbour models are proposed. The idea is to rank a set of images that are visually similar, annotate them according to their content characteristics, then predict for each image the term relevance given by a weighted sum of previously- affected annotations. These latter are extracted from capturing measures of local shape descriptors and global colour histograms. Works that are based only on the image associated text, make use of metadata provided by the same digital support including the image. According to [15], annotation approaches using contextual factors are convivial and can be easily applied with satisfactory retrieval accuracy compared to content-based annotation approaches. [1] comes within the same context, as it proposes an image annotation process aiming to identify keywords that can precisely describe a given query-image. In order to do so, for each image spatial, temporal and spatio-temporal filters are applied to retrieve similar images based on their tags. Then different descriptors are merged within a probabilistic model to define terms that effectively describe each query-image. The majority of works treating the image annotation through contextual factors, exploits external vocabularies and domain ontologies which reinforce the quality and precision of the annotations. [28] use WordNet lexicon to refine the annotations by calculating the similarity between affected words. This way, words which are similar to each other are belonging to the same higher level semantic. [29] also proposes a new approach for image semantic annotation especially in cinema field. The approach uses RDF patterns in conjunction with an OWL ontology. Basically, knowledge extraction and language processing are required for such cases treating high level semantics.

4 Discussion and Synthesis

We have presented above, a review on the state of art of the AIA approaches. As we mentioned before the challenging task concerning AIA is the semantic gap between the image content provided by visual features and semantic information extracted from metadata or textual descriptors. The reason why the majority of works tend towards the combination of both low-level features and high semantics when annotating an image. In the same context, many works related to AIA are still trying to improve and give the best of obtained results. To do so, automatic learning methods linking image visual features and semantic concepts are performed. In fact, extracting concepts from image samples, leads to treating the image content either globally or locally, by opting for supervised learning or unsupervised one. In one side, global methods have the advantage of being

simple and easy to use as they don't require a segmentation of the image into regions or sub-segments, which means that they offer a low computational cost in terms of characteristics extraction. However, treating the whole image doesn't provide the best interpretation, as the internal content and complex objects are ignored or at least weakly detected. In the other side, local methods dividing the image into regions, then extract from each region its features. This way, the extraction will be more precise and explicit, especially, when looking for specific and complex objects. But this precision cost highly in terms of computational complexity. So as a solution, researchers opt for the combination of the two. Hybrid methods even if they tend to be complex, but they are more suitable to use at the aim of extracting multiple semantics from one image. Although AIA methods have certain advantages, they come up with several limitations: (i) It's compulsory to have an already image annotated base as a reference in order to perform annotations for other images. Particularly, the image database must be large enough to be able to return for each image-query its similar images. (ii) The number of concepts used for the annotation is limited, since the learning phase depends on concepts provided by images in the database. In the same context, using supervised learning has the disadvantage of being limited in terms of visual features taken into account for the learning per each previously defined class. On the other hand, unsupervised learning needs a huge number of data for the learning process. (iii) Concepts provided by the annotated images in order to describe image-query contextually, are not able to define semantics and interpretations behind the image. Faced with all these limits of AIA images, researches are recently more interested in associating semantics to images by using the image context extracted from web pages, such as textual information, metadata describing information of localisation and time. Based on that, many works propose approaches aiming to describe images with relevant terms. Moreover, works exploiting vocabularies and domain ontologies show more accuracy and relevance comparing to others approaches using only textual descriptors.

5 Conclusion

In this paper, we presented a study on image retrieval particularly, automatic image annotation. We focused the review on the main classes of AIA methods which are summarized on the content-based annotation and context-based annotation of images. We presented learning strategies using the content-based process and the way the image was exploited either globally or locally. Through this study, we outlined the pros and cons of each method, in order to combine the best ideas overcoming the limitations presented by this field.

References

1. Mitran, M.: Annotation d'images via leur contexte spatio-temporel et les métadonnées du Web. Université de Toulouse, Université Toulouse III-Paul Sabatier (2014)

2. Islam, M.M., Zhang, D., Lu, G.: Automatic categorization of image regions using dominant colour based vector quantization. In: Proceedings of the Digital Image Computing: Techniques and Applications, pp. 191–198 (2008)
3. Ben Ismail, M.M.: A survey on content-based image retrieval. Int. J. Adv. Comput. Sci. Appl. **4**(8) (2013)
4. Smeulders, A.W., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-based image retrieval at the end of the early years. IEEE Trans. Pattern Anal. Mach. Intell. **22**(12), 1349–1380 (2000)
5. Olszewska, J.I.: Semantic, automatic image annotation based on multi-layered active contours and decision trees. Int. J. Adv. Comput. Sci. Appl. **4**(8), 201–208 (2013)
6. Enser, P.: Visual image retrieval: seeking the alliance of concept-based and content-based paradigms. J. Inf. Sci. **26**(4), 199–210 (2000)
7. Sakhare, S.V., Nasre, V.G.: Design of feature extraction in content based image retrieval (CBIR) using color and texture. Int. J. Comput. Sci. Inform. **1**(II) (2011)
8. Zhang, D., Lu, G.: Review of shape representation and description techniques. Pattern Recogn. **37**(1), 1–19 (2004)
9. Tamura, H., Mori, S., Yamawaki, T.: Textural features corresponding to visual perception. IEEE Trans. Syst. Man Cybern. **8**(6), 460–473 (1978)
10. Westerveld, T.: Image retrieval: content versus context. In: Content-Based Multimedia Information Access, vol. 1, pp. 276–284 (2000)
11. Zhang, C., Chai, J.Y., Jin, R.: User term feedback in interactive text-based image retrieval. In: Proceedings of the 28th Annual International Conference on Research and Development in Information Retrieval, pp. 51–58 (2005)
12. Gouet-Brunet, V.: Recherche par contenu visuel dans les grandes collections d'images (2005)
13. Liu, H., Song, D., Rüger, S., Hu, R., Uren, V.: Comparing dissimilarity measures for content-based image retrieval. In: Li, H., Liu, T., Ma, W.-Y., Sakai, T., Wong, K.-F., Zhou, G. (eds.) AIRS 2008. LNCS, vol. 4993, pp. 44–50. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-68636-1_5
14. Khademi, S.M., Pakize, S.R., Tanoorje, M.A.: A review of methods for the automatic annotation and retrieval of medical images. Int. J. **4**(7) (2014)
15. Pagare, R., Shinde, A.: A study on image annotation techniques. Int. J. Comput. Appl. **37**(6), 42–45 (2012)
16. Bouyerbou, H., Oukid, S., Benblidia, N., Bechkoum, K.: Hybrid image representation methods for automatic image annotation: a survey. In: International Conference on Signals and Electronic Systems, pp. 1–6 (2012)
17. Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. **20**(3), 273–297 (1995)
18. Qi, X., Han, Y.: Incorporating multiple SVMs for automatic image annotation. Pattern Recogn. **40**(2), 728–741 (2007)
19. Cusano, C., Ciocca, G., Schettini, R.: Image annotation using SVM. In: Internet Imaging V, vol. 5304, pp. 330–339. International Society for Optics and Photonics (2003)
20. Goh, K.S., Chang, E.Y., Li, B.: Using one-class and two-class SVMs for multiclass image annotation. IEEE Trans. Knowl. Data Eng. **17**(10), 1333–1346 (2005)
21. Guo, Y.T., Luo, B.: An automatic image annotation method based on the mutual K-nearest neighbor graph. In: Proceedings of the 6th International Conference on Natural Computation, Yantai, Shandong, pp. 3562–3566. IEEE (2010)

22. Zhao, Y., Zhao, Y., Zhu, Z., Pan, J.-S.: A novel image annotation scheme based on neural network. In: Proceedings of the 8th International Conference on Intelligent Systems Design and Applications, pp. 644–647 (2008)
23. Patil, M.P.: Automatic image annotation using decision trees and rough sets. Int. J. Comput. Sci. Appl. **11**(2), 38–49 (2014)
24. Olaode, A.A., Naghdy, G., Todd, C.A.: Unsupervised image classification by probabilistic latent semantic analysis for the annotation of images. In: International Conference on Digital Image Computing: Techniques and Applications (2014)
25. Wang, X.-J., Zhang, L., Jing, F., Ma, W.-Y.: AnnoSearch: image auto-annotation by search. In: Proceedings of the CVPR06, vol. 2, pp. 1483–1490 (2006)
26. Barai, S., Cardenas, A.F.: Image annotation system using visual and textual features. In: Proceedings of the 16th International Conference on Distributed Multimedia Systems, pp. 289–296 (2010)
27. Verbeek, J., Guillaumin, M., Mensink, T., Schmid, C.: Image annotation with tagprop on the MIRFLICKR set. In: Proceedings of the International Conference on Multimedia Information Retrieval, pp. 537–546 (2010)
28. Jin, Y., Khan, L., Wang, L., Awad, M.: Image annotations by combining multiple evidence and Wordnet. In: Proceedings of the ACM Multimedia (2005)
29. Jaouachi, R., Torjmen, M., Hernandez, N., Haemmerlé, O., Jemaa, M.B.: Vers une annotation semantique des images web fondee sur des patrons RDF (2015)