TEMPLATE FOR REGULAR ENTRY
(ENCYCLOPEDIA OF DATABASE SYSTEMS)

TITLE OF ENTRY
Automatic Image Annotation

BYLINE
Nicolas Hervé and Nozha Boujemaa, INRIA Paris-Rocquencourt, IMEDIA project, France.
http://www-rocq.inria.fr/imedia/

SYNONYMS
Multimedia Content Enrichment, Image Classification, Object Detection and Recognition, Auto-annotation

DEFINITION
The widespread search engines, in the professional as well as the personal context, used to work on the basis of textual information associated or extracted from indexed documents. Nowadays, most of the exchanged or stored documents have multimedia content. To reduce the technological gap so that these engines still can work on multimedia content, it is very convenient developing methods capable to generate automatically textual annotations and metadata. These methods will then allow to enrich the upcoming new content or to post-annotate the existing content with additional information extracted automatically if ever this existing content is partly or not annotated.

A broad diversity in the typology of manual annotation is usually found in image databases. Part of them is representing contextual information. The author, date, place or technical shooting conditions are quite frequent. Some semantic or subjective annotations, like emotions that flow out from images, can be found. Some other annotations could be related to the visual content of images. They provide information on a given image such as indicating whether it is a drawing, a map or a photograph... For photographs, the global aspect is often specified (vertical/horizontal, color/black and white, indoor/outdoor, landscape, portrait ...), as well as the presence of remarkable objects or persons.
The aim of automatic image annotation approaches is to provide efficient methods that extract automatically the visual content of pictures allowing semantic labeling of images. This is generally achieved by learning algorithms that, once being trained on annotated sub-corpora, are able to suggest keywords to the archivist through object detection/recognition and image classification methods.

HISTORICAL BACKGROUND

The exploration of visual content databases and their querying to retrieve some specific content usually rely on textual annotations that have been previously provided manually by human operators. The outcome of the tremendous improvements in digitization and acquisition devices is the availability of exponentially growing content. Usual annotation techniques then became more and more difficult to apply because they are time and cost consuming. Moreover, manual annotations are far from being perfect. They are often focused on the context, subjective, partial and driven by the needs of the end-users at the time they are produced. As these needs are evolving, part of the existing annotations becomes irrelevant and others are missing. This is especially true with the arising of Internet and the availability of all kind of databases online. An other issue lies in the lack of controlled vocabularies for most of the databases making difficult for the end-user to guess what query words he has to use in order to retrieve the content he has in mind.

Visual content indexing and retrieval community have achieved significant progress in the recent years [1] toward efficient approaches for visual features extraction and visual appearance modeling together with developing advanced mechanism for interactive visual information retrieval. One of the major issues was and remains the semantic gap [2, 3].

Two main types of images databases could be distinguished. Specific databases are focused on a given restricted field. In the scientific domain, one can cite satellite images for weather forecast or cultivation study, medical images or botanical databases for species recognition. They are also found in the cultural heritage domain (eg. paintings databases) or the military and security domain (eg. fingerprints and faces databases). On the other side, generic databases contain very different images, without any *a priori* on their content. This is usually the case for professional news agencies, illustration photo stock collections and personal family and holiday photo albums. We will only address methods for generic content databases labeling.

By analyzing automatically images and characterizing them with low-level features (mainly colors, textures and shapes) CBIR systems [4] provided new query paradigms that enable users to express their needs. The main one is “query by example” where the system retrieves images of the database that are the most similar to a given example. The scientific community has been facing the well known semantic gap problem for a while which
remain the major concern of the research community. Since the late 90's, relevance feedback mechanism is one of possible solutions to this difficult problem.
The early papers on automatic annotation that have been published tackled image orientation detection or the classical indoor vs. outdoor and city vs. landscape classifications of photographs [5, 6]. Recently, relevance feedback allows moreover helping for interactive mass-annotation of image collections. This approach is often referred to as semi-automatic image annotation.

SCIENTIFIC FUNDAMENTALS
Despite some of its drawbacks, the query by keyword is still very useful and quite natural for the end-user [7]. Automatic annotation generates such keywords to enrich the images semantic descriptions and ease further querying. Because of the computational costs of all current approaches, the existing systems are always composed of two parts. An offline part is in charge of indexing the visual content and generating the annotations. Eventually, a human operator can help the system during the process or after it to validate/invalidate the produced annotations. In such cases, we rather talk of semi-automatic annotation systems. The second part, online and real-time, is a query by keywords module.

As the main purpose is to describe the visual content, we prefer using the term “visual concept” than keyword to describe the labels a system has to discover in images. We have already mentioned that these visual concepts could be related to either global appearance of the image or presence of some objects. Objects detection can also be refined in generic object class detection or specific object instance detection. For example, one can ask a system to label only the “vehicle” concept, or more precisely to distinguish cars, motorbikes, boats and airplanes, and, at a very specific level, being able to recognize different makes of cars. This is the same problem with annotating persons. Being able to detect the presence of a person in an image is a different procedure and result than recognizing him. As face recognition is a well studied problem which is tackled by a specific research community. The ability to generalize from a few examples and to reach higher abstraction levels is natural for humans but it is very challenging task to achieve with current state-of-the-art's annotation systems [8, 9, 10, 11].

One of the fundamental hypotheses of automatic annotation is that what looks similar is probably semantically similar. Most of the approaches rely on this assumption. The main generic steps of automatic annotation are described below. First, visual features are extracted automatically from images in order to obtain representations in a visual space. The second step is to build models that will link the visual concepts to the relevant information in the visual space. When new content is proposed, models are then able to predict the corresponding visual concepts.

The performances evaluation of such methods may rely on the usage of the annotations by the final users. As in most of information retrieval systems, precision and recall measures are used. Precision emphasizes the retrieval of relevant documents earlier and recall focuses on the retrieval of the full set of relevant documents. Precision and recall are complementary to judge the quality of a system. But for some applications, precision is the only important measure. This is especially the

case when a huge image database is available (like Internet). When doing a query, a user is more interested in the first satisfying results than in the complete relevant result set.

1. Images description with low-level features

The visual description of images is of great importance as it is the raw material on which further models are built. There is not a universally good low-level features extractor. In specific databases, *a priori* on the content of images can be used to extract specialized features that will better describe their special nature. For example, numerous features can be found in the literature for faces or fingerprints description. In generic databases, compromises have to be made between exhaustiveness, fidelity to the content, ability to generalize and different invariance degrees (illumination changes, rotations, scales, occlusions ...). The use of inappropriate features leading to poor performances of a system has often been described as semantic gap. In this case, one rather faces the numerical gap, meaning that the visual information is present in images but it has not been extracted correctly.

Due to their ability to generalize to content in different conditions, statistical features are often used. They gather color, shape and texture information in histograms, separately or jointly. Color histograms are among the first features used to describe images. They vary depending on the underlying color space that is used, the quantization parameters, different weighting schemes or the use of co-occurrences of colors. Shapes can be described by properties of edges found in images like their types, orientations or lengths. Textures are focusing on the analysis of frequencies in images. They often rely on Fourier transform, Gabor filter banks or wavelets. Some features also combine different types of information, mixing for example color and texture in a single representation. Typically, these visual features are represented by vectors in high-dimensional spaces (generally between a few tens and a few hundreds dimensions) [6, 12].

Initially, the features were extracted over the full image. This approach is well suited to describe the global aspect of the content but is too coarse to represent small details and objects. Features need to be extracted locally. First, a support region has to be determined. Once its location, shape and size are known, features are computed on this small portion of the image. These features can be of the same types as those extracted at a global level or they can be specialized according to the nature of the support regions. Several strategies are used to select the support regions. Segmentation algorithms try to find the boundaries between homogeneous regions in images [13]. Segmentation is a difficult problem in itself that is not well defined. Unfortunately, the general trend has always been to focus on segmentation that detects objects, which is already a highly semantic task and, thus, not really achievable through automatic processes. Alternative approaches consist on sliding windows and fixed grid, with varying sizes and spacing, are common ways of obtaining dense sampling of the visual content [5, 12]. Another popular region selection approach is based on local features detectors via point-of-interest. They were originally designed for image registration. These detectors are generally attracted to specific areas of images that have high variation in the visual signal, such as the vicinity of edges and corners of regions. They allow the selection of a very small proportion of image locations having the highest visual variance [14, 15]. Typically, when using dense sampling, point-of-interest or when mixing them, between a few hundreds and a few thousands features are extracted per image. The computational cost is then much higher than with global features. Some representations also try to carry other information, like geometrical relations between features locations or contextual information [16].

With global features, the image representation is straightforward. However, even when local features are to be used, learning algorithms may sometimes require a global image representation that encompasses all the local visual information. The bag-of-visual-words representation, very much inspired by the classical bag-of-words representation for text, is one of the most popular for images. A visual vocabulary composed of visual words (some representative features) is generated. An image is then represented by a coordinate vector, each value of which expresses the degree of importance of a feature with respect to the image and/or the database as a whole. The creation of a visual vocabulary is an important step in the full process. The selected visual words (a few hundreds to a few thousands) have to be representative of the database content as they will serve as a basis for further representation. Creating a good vocabulary will avoid the loss

of too much local information. Generally, clustering algorithms either supervised or not, are used with a sample of the database. This step can be seen as a quantization of the local features.

Whatever the selected representation, the similarity between images is measured by a distance functional in the visual space. A broad variety has been developed: classical Euclidean distance (L2), L1, earth mover distance (EMD), chi-squared (χ^2), vector angle, histogram intersection, ...

2. Learning and models

Although several formulations have been proposed, the main purpose of building models for visual concepts is to associate them with the visual space regions that best represent them. This problem is at the cross-roads of computer vision, data mining and machine learning. Usually, the models are built through a supervised learning process. For given visual concepts, an algorithm is fed with a training dataset containing both positive and negative images regarding the concepts to learn. This algorithm has to find the discriminant information from the visual space that best models the concepts. Generally, the available annotations for the training set are not localized. The presence of a visual concept for an image is known, but its exact location is not provided. This is the case for almost all professional and personal databases. In the same way, annotating new images does not require to locate exactly the visual concept, but only to predict its presence. This is the main distinction that can be made with object detection tasks.

Two main learning algorithm families are used :

- generative: the system tries to estimate density distribution of concepts in the visual space or other hidden variables [13]. Popular examples include Gaussian Mixture Models, Hidden Markov Models, Bayesian networks, Latent Semantic Analysis and translation models from the text processing community. The Expectation-Maximization algorithm is often used to train these models.
- discriminative: instead of trying to model the distributions, discriminative approaches are focusing on detecting the boundaries between classes. For each visual concept, the annotation process is then often formalized as a two-class classification problem (present/not present). Although appearing to be a little bit more effective, discriminative approaches do not have the elegance of generative ones. They act more as black boxes and relationships between the different variables are not explicit, and thus difficult to analyze. The most famous algorithms are Support Vector Machine (SVM) [12, 14], boosting (eg. adaboost) [15] and all flavors of discriminant analysis (linear - LDA, biased - BDA, multiple - MDA, Fisher - FDA).

Both approaches may use global or local representations. Methods using bag-of-words representations are also called “multiple instance learning”. Sometimes, pre-processes may also be used to prepare the data in order to enhance the performances or to reduce the computational costs : feature selection, dimensionality reduction, scaling or normalization. Current systems are often composed of several components, using different low-level features, combining them according to different schemes and training models with multiple learning strategies.

Once the models have been learned, they can be used to predict the visual concepts. Two types of predictions are possible. Hard decision simply indicates the presence or absence of the concept. Soft decision also provides a degree of confidence in the prediction, allowing ranking more easily the results when answering an end-user query and thus improving the retrieval of pertinent images earlier.

3. Current results

The different methods are actually mature enough to predict global visual concepts like image types and scene categories. Regarding local concepts, huge improvements still need to be made in order to provide useful applications to real users. Both dense sampling and point-of-interest have shown to perform quite well on research databases, but results on real databases are quite poor[10, 12]. A good indication of state-of-the-art performances can be obtained in the results of official benchmark campaigns like ImagEval, Pascal VOC, Imageclef or Trecvid. In all cases, the contextual information (visual or from the existing metadata) has shown to be of great importance

in the results.

When the availability of correctly annotated images for a given visual concept is not guaranteed, the offline learning approach is not possible. One of the solutions is then to interact with a user through relevance feedback, also called interactive learning. In a few iterations, the user will provide the system positive and negative examples and guide it to recognize the visual concept. Part of the mechanism involved are the same as offline learning, but the training labels are provided online by a user. As an example, the following screen captures from Ikona [4] are showing the IAPR-TC12 database, used for the Imageclef benchmark. The first screen displays all the images annotated with the “tennis” keyword. The user is only interested in pictures where a tennis court is visible. He indicates positive (green border) and negative (red border) examples. After two iterations, one can see on the second screen that a lot of tennis court pictures have been retrieved. None of them was annotated with the “tennis” keyword. After a few more iterations, when no more correct pictures are retrieved from the database, the user is able to annotate massively all the pictures gathered through the iterations that were kept in a specific basket (third screen).

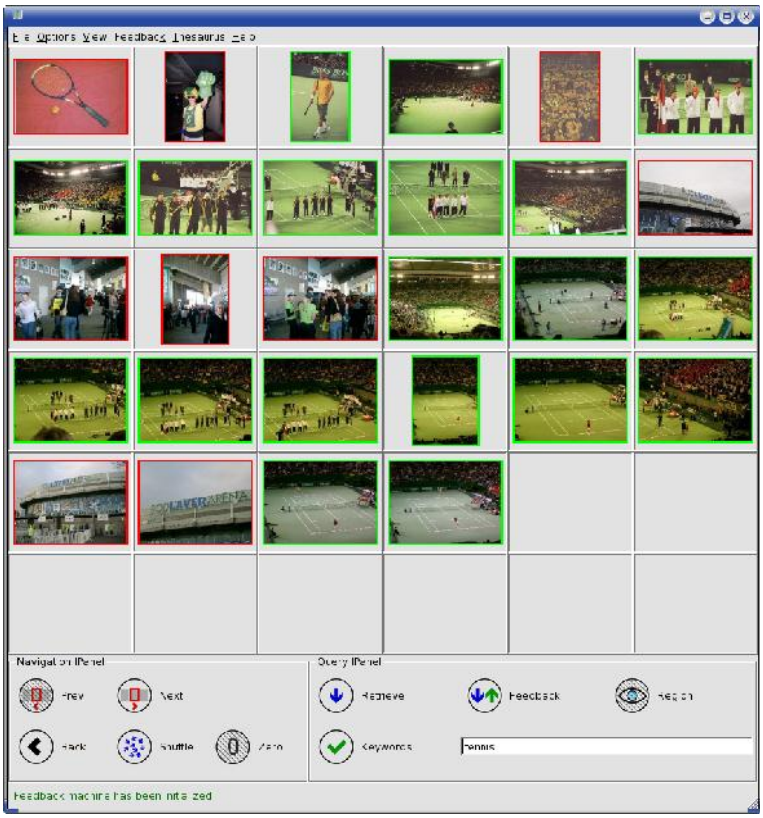

Fig 1 - Pictures annotated with the "tennis" keyword

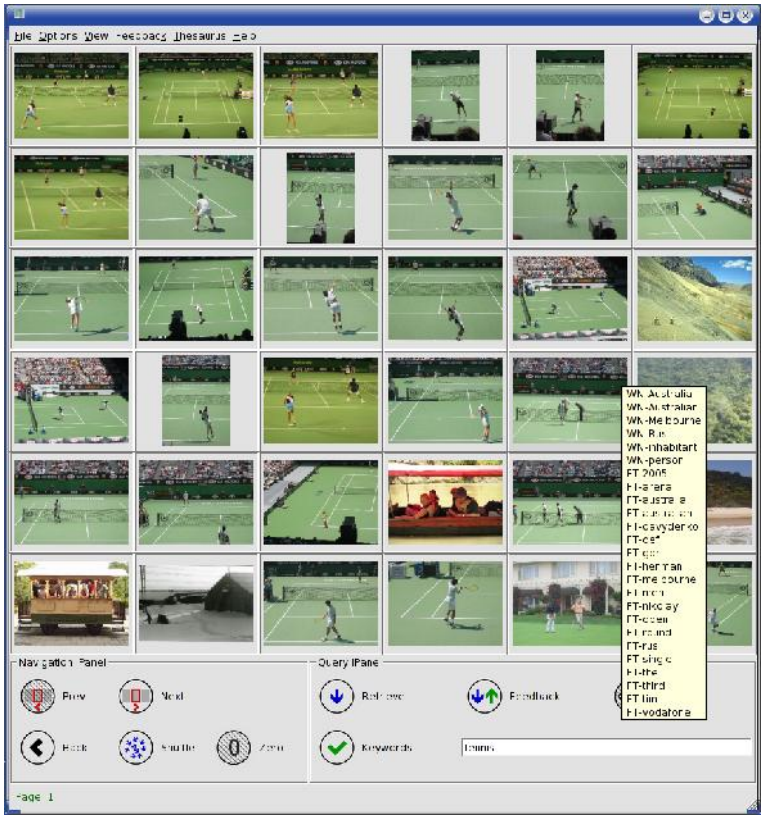

Fig 2 - Pictures displayed after two iterations

Fig 3 - All positive pictures basket allowing mass-annotation

4. Key issues and future research

The lack of generalization ability for both visual features and learning algorithms has to be compensated by a huge number of training examples. Depending on the complexity of the visual concept, a good estimation is around a hundred positive examples and ten times more negatives examples for the training set. Paradoxically, despite the tremendous amount of images available nowadays, finding content that has been reliably annotated for training dataset is hard.

The computational complexity is also still too high for real-time annotation when dealing with several thousands of visual concepts. Research is made on scalability issues in machine learning and is linked to existing high-dimensional data indexing structures.
Progresses for better description and integration of all types of available information need to be achieved.
There are also some questions arising: will we solve the problem with more computational power when we will be able to process images at every scale and location in real-time? Are massive collaborative annotation websites, like Flickr, going to change the annotation paradigm by transforming Internet in a giant common repository? What is the impact of GPS metadata, and more generally all the new information captured directly when a photograph is taken?

KEY APPLICATIONS

- Professional content owners: post-editing
- Personal family and holiday photo albums
- Web image search
- Searching into poorly human-made annotated corpora enhancing the quality of search results

CROSS REFERENCES

CBIR, Image database, Object Detection, Object Recognition

RECOMMENDED READING

[1] Arnold W.M. Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta and Ramesh Jain. *Content-Based Image Retrieval at the End of the Early Years*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, 2000.

[2] Nozha Boujemaa, Julien Fauqueur and Valérie Gouet. *What's beyond query by example ?* Technical report, INRIA, 2003.

[3] Jonathon S. Hare, Paul H. Lewis, Peter G. B. Enser and Christine J. Sandom. *Mind the Gap : Another look at the problem of the semantic gap in image retrieval*. In Multimedia Content Analysis, Management, and Retrieval, SPIE-IS&T, 2006.

[4] Nozha Boujemaa, Julien Fauqueur, Marin Ferecatu, François Fleuret, Valérie Gouet, Bertrand Le Saux and Hichem Sahbi. *Ikona: interactive specific and generic image retrieval*

[5] Martin Szummer and Rosalind W. Picard. *Indoor-Outdoor Image Classification*. Workshop in Content-based Access to Image and Video Databases, CAIVD98. Bombay, 1998.

[6] Aditya Vailaya, Anil Jain and Hong-Jiang Zhang. *On Image Classification : City Images vs. Landscapes*. Pattern Recognition Journal, 1998.

[7] Peter G.B. Enser, Christine J. Sandom and Paul H. Lewis. *Automatic Annotation of Images from the Practitioner Perspective*. International Conference on Image and Video Retrieval, CIVR05, Singapore, 2005.

[8] Michael S. Lew, Nicu Sebe, Chabane Djeraba and Ramesh Jain. *Content-based multimedia information retrieval : State of the art and challenges*. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP), vol. 2, pages 1-19, 2006.

[9] Ritendra Datta, Jia Li and James Z. Wang. *Content-Based Image Retrieval - Approaches and Trends of the New Age*. ACM SIGMM international workshop on Multimedia information retrieval, 2005.

[10] Jean Ponce, Martial Hebert, Cordelia Schmid and Andrew Zisserman, editors. *Toward category-level object recognition*. Springer-Verlag Lecture Notes in Computer Science, 2006.

[11] Alan Hanjalic, Nicu Sebe and Edward Chang. *Multimedia Content Analysis, Management and Retrieval : Trends and Challenges*. In Multimedia Content Analysis, Management, and Retrieval, SPIE-IS&T, 2006.

[12] Nicolas Hervé and Nozha Boujemaa. *Image annotation : which approach for realistic databases ?* ACM International Conference on Image and Video Retrieval, July 2007.

[13] Kobus Barnard, Pinar Duygulu, David Forsyth, Nando de Freitas, David M. Blei and Michael I. Jordan. *Matching Words and Pictures*. Journal of Machine Learning Research, vol. 3, pages 1107-1135, 2003.

[14] Jianguo Zhang, Marcin Marszalek, Svetlana Lazebnik and Cordelia Schmid. *Local features and kernels for classification of texture and object categories: A comprehensive study*. International Journal of Computer Vision, vol. 73, No 2, pages 213-238, 2007.

[15] Andreas Opelt, Axel Pinz, Michael Fussenegger and Peter Auer. *Generic Object Recognition with Boosting*. Pattern Analysis and Machine Intelligence, vol. 28, issue 3, pages 416-431, march 2006.

[16] Jaume Amores, Nicu Sebe and Petia Radeva. *Context-Based Object-Class Recognition and Retrieval by Generalized Correlograms*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 29, no. 10, october 2007.