

# Annotation-based Image Retrieval

XIN-JING WANG, LEI ZHANG

Microsoft Research Asia, Beijing, China

## Definition

Given (i) a textual query, and (ii) a set of images and their annotations (phrases or keywords), annotation-based image retrieval systems retrieve images according to the matching score of the query and the corresponding annotations. There are three levels of queries according to Eakins [7]:

- Level 1: Retrieval by primitive features such as color, texture, shape or the spatial location of image elements, typically querying by an example, i.e., “find pictures like this.”
- Level 2: Retrieval by derived features, with some degree of logical inference. For example, “find a picture of a flower.”
- Level 3: Retrieval by abstract attributes, involving a significant amount of high-level reasoning about the purpose of the objects or scenes depicted. This includes retrieval of named events, of pictures with emotional or religious significance, etc., e.g., “find pictures of a joyful crowd.”

Together, levels 2 and 3 are referred to as semantic image retrieval, which can also be regarded as annotation-based image retrieval.

## Historical Background

There are two frameworks of image retrieval [6]: annotation-based (or more popularly, text-based) and content-based. The annotation-based approach can be tracked back to the 1970s. In such systems, the images are manually annotated by text descriptors, which are used by a database management system (DBMS) to perform image retrieval. There are two disadvantages with this approach. The first is that a considerable level of human labor is required for manual annotation. The second is that because of the subjectivity of human perception, the manually labeled annotations may not converge. To overcome the aforementioned disadvantages, content-based image retrieval (CBIR) was introduced in the early 1980s. In CBIR, images are indexed by their visual content, such as color, texture, shapes. In the past decade, several commercial products and experimental prototype systems were developed, such as QBIC, Photobook, Virage, VisualSEEK, Netra, SIMPLiCity.

However, the discrepancy between the limited descriptive power of low-level image features and the richness of user semantics, which is referred to as the “semantic gap” bounds the performance of CBIR. On the other hand, due to the explosive growth of visual data (both online and offline) and the phenomenal success in Web search, there has been increasing expectation for image search technologies. For these reasons, the main challenge of image retrieval is understanding media by bridging the semantic gap between the bit stream and the visual content interpretation by humans [3]. Hence, the focus is on automatic image annotation techniques.

## Foundations

The state-of-the-art image auto-annotation techniques include four main categories [3,6]: (i) using machine learning tools to map low-level features to concepts, (ii) exploring the relations between image content and the textual terms in the associated metadata, (iii) generating semantic template (ST) to support high-level image retrieval, (iv) making use of both the visual content of images and the textual information obtained from the Web to learn the annotations.

## Machine Learning Approaches

A typical approach is using Support Vector Machine (SVM) as a discriminative classifier over image low-level features. Though straightforward, it has been shown effective in detecting a number of visual concepts.

Recently there has been a surge of interest in leveraging and handling relational data, e.g. images and their surrounding texts. Blei et al. [1] extends the Latent Dirichlet Allocation (LDA) model to the mix of words and images and proposed a Correlation LDA model.

## Relation Exploring Approaches

Another notable direction for annotating image visual content is exploring the relations among image content and the textual terms in the associated metadata. Such metadata are abundant, but are often incomplete and noisy. By exploring the co-occurrence relations among the images and the words, the initial labels may be filtered and propagated from initial labeled images to additional relevant ones in the same collection [3].

Jeon et al. [5] proposed a cross-media relevance model to learn the joint probabilistic distributions of the words and the visual tokens in each image, which are then used to estimate the likelihood of detecting a specific semantic concept in a new image.

## Semantic Template Approaches

Though it is not yet widely used in the techniques mentioned above, Semantic Template (ST) is a promising approach in annotation-based image retrieval (a map between high-level concept and low-level visual features).

Chang and Chen [2] show a typical example of ST, in which a visual template is a set of icons or example scenes/objects denoting a personalized view of concepts such as meetings, sunset, etc. The generation of a ST is based on user definition. For a concept, the objects, their spatial and temporal constraints, and the weights of each feature of each object are specified. This initial query scenario is provided to the system, and then through the interaction with users, the system finally converges to a small set of exemplar queries that “best” match (maximize the recall) the concept in the user’s mind.

### Large-Scale Web Data Supported Approaches

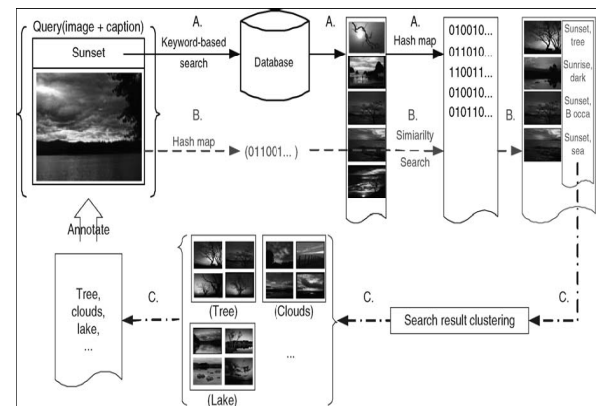
Good scalability to a large set of concepts is required in ensuring the practicability of image annotation. On the other hand, images from the Web repositories, e.g. Web search engines or photo sharing sites, come with free but less reliable labels. In [9], a novel search-based annotation framework was proposed to explore such Web-based resources. Fundamentally, it is to automatically expand the text labels of an image of interest, using its initial keyword and image content.

The process of [9] is shown in Fig. 1. It contains three stages: the text-based search stage, the content-based search stage, and the annotation learning stage, which are differentiated using different colors (black, brown, blue) and labels (A., B., C.). When a user submits a query image as well as a query keyword, the system first uses the keyword to search a large-scale Web image database (2.4 million images crawled from several Web photo forums), in which images are associated with meaningful but noisy descriptions, as tagged by “A.” in Fig. 1. The intention of this step is to select a semantically relevant image subset from the original pool.

Visual feature-based search is then applied to further filter the subset and save only those visually similar images (the path labeled by “B.” in Fig. 1). By these means, a group of image search results which are both semantically and visually similar to the query image are obtained. Finally, based on the search results, the system collects their associated textual descriptions and applies the Search Result Clustering (SRC) algorithm to group the images into clusters.

(Abridged)

Figure 1:



### Recommended Reading

1. Blei D. and Jordan M.I. Modeling Annotated Data. In Proc. 26<sup>th</sup> Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, 2003, pp. 127–134.
2. Chang S.-F., Chen W., and Sundaram H. Semantic Visual Templates: Linking Visual Features to Semantics. In Proc. Int. Conf. on Image Processing, Vol. 3. 1998, pp. 531–534.
3. Chang S.-F., Ma W.-Y., and Smeulders A. Recent Advances and Challenges of Semantic Image/Video Search. In Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, 2007, pp. 1205–1208.
4. Eakins J. and Graham M. Content-based image retrieval, Technical Report, University of Northumbria at Newcastle, 1999.
5. Jeon J., Lavrenko V., and Manmatha R. Automatic Image Annotation and Retrieval Using Cross-Media Relevance Models, In Proc. 26th Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, 2003, pp. 119–126.
6. Liu Y., Zhang D., Lu G., and Ma W.-Y. A survey of content-based image retrieval with high-level semantics. Pattern Recognition., 40(1):262–282, 2007.
7. Long F., Zhang H.J., and Feng D.D. Fundamentals of content-based image retrieval. In Multimedia Information Retrieval and Management, D. Feng (eds.). Springer, 2003.
8. Rui Y., Huang T.S., and Chang S.-F. Image retrieval: current techniques, promising directions, and open issues, J. Visual Commun. Image Represent. 10(4):39–62, 1999.
9. Wang X.-J., Zhang L., Jing F., and Ma W.-Y. AnnoSearch: Image Auto-Annotation by Search, Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition, 2006, pp. 1483–1490.
10. Zhuang Y., Liu X., and Pan Y. Apply Semantic Template to Support Content-based Image Retrieval.

210 In Proc. SPIE, Storage and Retrieval for Media  
Databases, vol. 3972, December 1999, pp. 442–449.

**Answer the following questions:**

- 1) Describe *retrieval by primitive features*.
- 2) What is meant by *abstract attributes* in the context of retrieving images?
- 3) What is meant by *semantic image retrieval*? Why is it called *semantic*?
- 4) What is *annotation* and what are its disadvantages?
- 5) What are the machine learning tools good for in image retrieval?
- 6) Describe the term of *relation exploring approach*.
- 7) What is the generation of *semantic template* based on?

**Match the following terms and their definitions:**

- 1) semantic gap
  - 2) SVM
  - 3) metadata
  - 4) ST
- 
- a) a machine learning approach
  - b) data about data
  - c) the discrepancy between the limited descriptive potential of low-level image features and the richness of user's description
  - d) a set of icons or example scenes/objects denoting a personalized view of concepts

**Mark the following statements as *true* or *false*:**

- 1) *Retrieval by derived features* can be based on color of the picture.
- 2) *The content-based approach* uses descriptors to retrieve images.
- 3) SVM works on low-level features.
- 4) Images from the Web repositories come with highly reliable labels.
- 5) *Search Result Clustering* refers to grouping information about images.

# Vocabulary

---

**abundant** [ə'ʌn.dənt] – hojný, překypující

**query** ['kwɪə.ri] – dotaz

**annotation** ['æn.əʊ.teɪt] – anotace

**classifier** ['klæ.sɪ.faɪə] – klasifikátor

**co-occurrence** [kəʊə'kə.rəns]

– společný výskyt

**descriptor** [dis'kriptə] – descriptor, popis

**discriminative** [dis'krɪmɪnətɪv]

– rozlišující, schopný rozlišovat

**exemplar** [ɪg'zemplər] – typický příklad, vzor, model

**explosive** [ɪk'spləʊ.sɪv] – explozivní, výbušný

– explozivní, výbušný

**hence** [henʃ] – tudíž (formální)

**incomplete** [ɪn.kəm'pli:t] – nekompletní

**likelihood** ['laɪ.kli.hud] – pravděpodobnost

**machine learning** [mə'ʃi:nɪŋ] – strojové učení

– strojové učení

**metadata** ['metədeɪtə] – metadata, data popisující jiná data

**noisy** ['nɔɪ.zi] – obsahující šum: hlučný

**phenomenal** [fə'nɒm.i.əl] – úžasný, výjimečný

– úžasný, výjimečný

**pool** [pu:l] – úložiště, zásoba; bazén

**relational** [ri'leɪʃənəl] – relační, vztahový

**retrieval** [ri'tri:v] – získávání, vyhledávání

**scenario** [sɪ'næ:ri.əʊ] – scénář

**semantic** [sɪ'mæn.tɪk] – sémantický, významový

**subset** ['sʌb.set] – podmnožina

**token** ['təʊ.kən] – znamení, znak, symbol

**to annotate st** ['æn.əʊ.teɪt] – anotovat něco, vybavit anotací

**to bridge st** [brɪdʒ] – překlenout něco

**to leverage st** ['li:vər.ɪdʒ] – využívat k užtku

– využívat k užtku

**to overcome st**, overcame, overcome [əʊ.və'kʌm]

– překonat něco

**to propagate** ['prɒp.ə.geɪt] – rozšiřovat, množit

– rozšiřovat, množit

**to retrieve** [ri'tri:v] – získávat, vyhledávat

**visual** ['vɪʒ.u.əl] – vizuální