Available online at www.sciencedirect.com
ScienceDirect
Journal of Visual Languages and Computing 19 (2008) 617–627
Journal of Visual Languages & Computing
www.elsevier.com/locate/jvlc

A survey of methods for image annotation ☆

Allan Hanbury*

Pattern Recognition and Image Processing Group (PRIP), Institute of Computer-Aided Automation, Favoritenstraße 9/1832, A-1040 Vienna, Austria

Received 17 May 2006; received in revised form 25 May 2007; accepted 14 January 2008

Abstract

In order to evaluate automated image annotation and object recognition algorithms, ground truth in the form of a set of images correctly annotated with text describing each image is required. In this paper, three image annotation approaches are reviewed: free text annotation, keyword annotation and annotation based on ontologies. The practical aspects of image annotation are then considered. We discuss the creation of keyword vocabularies for use in automated image annotation evaluation. As direct manual annotation of images requires much time and effort, we also review various methods to make the creation of ground truth more efficient. An overview of annotated image datasets for computer vision research is provided.

Keywords: Image annotation; Object recognition; Computer vision; Ontology; Algorithm evaluation

1. Introduction

The usual reason to annotate data (i.e. add metadata to it) is to simplify access to it. This is one of the key ideas behind the semantic web. The metadata added to documents or images allow for more effective searches. In the case of images, if they are completely described by a textual annotation, then many image searches can be done effectively by text search techniques. The problem with adding metadata manually is that it is an extremely labour-intensive and time-consuming task. Many World Wide Web image search engines attempt to automate this task by using text from the image filename and text near the image on a webpage. However, search results using this method usually contain many irrelevant images. With the aim of improving the automated metadata generation for images, automated image annotation and object recognition are currently important research topics in the field of computer vision [1–8]. This automatic generation of image metadata should allow image searches and content-based image retrieval (CBIR) [9] to be more effective. In the following image retrieval scenario, use is made of these techniques: an image database could be annotated offline by running a keyword annotation algorithm. Every image containing a cup would then have the keyword "cup" associated with it. If a user wishes to find images of a specific cup in this database, e.g. for an on-line shopping task, he/she would select a region containing the target cup from an image. An object recognition algorithm could then categorise the selected region as a cup and a text search could be carried out to find all

☆This work was supported by the Austrian Science Foundation (FWF) under Grant SESAME (P17189-N04), and the European Union Network of Excellence MUSCLE (FP6-507752).
*Tel.: +43 1 58801 18351; fax: +43 1 58801 18392.
E-mail address: hanbury@prip.tuwien.ac.at

1045-926X/$ - see front matter
doi:10.1016/j.jvlc.2008.01.002

images in the database with an associated keyword "cup". This would significantly reduce the number of images in which it would be necessary to attempt to recognise the specific cup selected by the user.

To measure progress towards successfully carrying out this task, evaluation of algorithms which automatically extract this sort of metadata is required. For successful evaluation of these algorithms, reliable ground truth is necessary. This ground truth should be a semantically rich description of the objects in an image [10]. There is obviously almost no limit to how semantically rich one could make the description of an image. Indeed, for manual annotation of such documents destined to aid in on-line searching for them, semantic richness is an advantage. Nevertheless, it should be borne in mind that the automated content description and annotation algorithms being developed cannot yet be expected to perform at the same level of detail as a human annotator. As demonstrated by the results of the recent ImagEVAL campaign [11], algorithms providing global annotations, such as distinguishing between city and landscape images or between images acquired indoors and outdoors, have a higher success rate than algorithms attempting to detect specific objects, such as cars, cows and sunglasses. Automatic recognition of activities, events and abstract or emotive qualities in images currently performs rather poorly.

In this paper we review the annotation of images for evaluation purposes. Three types of annotation: free text annotations, keyword annotations and annotations based on ontologies are described in Section 2. We pay particular attention to the creation of vocabularies for image annotation and to methods which have been applied for reducing the amount of effort required for image annotation in Section 3. An overview of available annotated image datasets for computer vision research is also provided. Section 4 concludes.

2. Annotation approaches

Different types of information can be associated with images or videos. They are [12]:

- *content-independent metadata* is related to the image or video content, but does not describe it directly. Examples are: author's name, date, location, cost of filming, etc;
- data which directly refers to the visual content of images can be divided into two types:
  - *content-dependent metadata* refers to low/intermediate-level features (colour, texture, shape, motion, etc);
  - *content-descriptive metadata* refers to content semantics. It is concerned with relationships of image entities with real-world entities or temporal events, emotions and meaning associated with visual signs and scenes.

Except in very rare cases, for example extracting the location as "London" from an image including well-known landmarks such as the Houses of Parliament or Tower Bridge, the content-independent information cannot be extracted from the image. Content-dependent metadata is easy to extract—with enough computation time, one can extract huge feature vectors containing colour histogram features, texture features calculated by different algorithms, etc. [9,12,13]. Annotation by content-descriptive metadata is the focus of this paper—this is the type of annotation which is most challenging to automate and which requires extensive testing to evaluate the performance of annotation algorithms. Content-descriptive metadata can be specified using one or more of the following approaches [14], listed in order of increasing structure:

Free text descriptions: No pre-defined structure for the annotation is given.

Keywords: Arbitrarily chosen keywords or keywords chosen from *controlled vocabularies*, i.e. restricted vocabularies defined in advance, are used to describe the images.

Classifications based on ontologies: Ontologies—large classification systems that classify different aspects of life into hierarchical categories [14]—are used. This is similar to classification by keywords, but the fact that the keywords belong to a hierarchy enriches the annotations. For example, it can easily be found out that a "dog" is a subclass of the class "animal".

These approaches are discussed in the following subsections.

2.1. Annotation using keywords

Each image is annotated by having a list of keywords associated with it. There are two possibilities for choosing the keywords:

(1) The annotator can use arbitrary keywords as required.

(2) The annotator is restricted to using a pre-defined list of keywords (a *controlled vocabulary*).

This information can be provided at two levels of specificity:

(1) A list of keywords associated with the complete image, listing what is in the image (see Fig. 1a for an example).
(2) A segmentation of the image along with keywords associated with each region of the segmentation. In addition, keywords describing the whole image can be provided (see Fig. 1b for an example). Often the segmentation is much simpler than that shown, consisting simply of a rectangular region drawn around the region of interest or a division of the image into foreground and background pixels.

If one is searching within a single image database that has been annotated carefully using a keyword vocabulary, then one's task is simplified. Unfortunately in practice, the following two problems arise:

- Different image collections are annotated using different keyword vocabularies and differing annotation standards.
- A naive user does not necessarily know the vocabulary which has been used to annotate an image collection. This makes searching by text input more difficult.

Forcing the user to choose from an on-screen list of keywords is a solution to the second problem, but this makes the search task more frustrating if the number of keywords is large. As a solution to both of the above problems, a thesaurus can be used to extend the list of search words entered by a user. A more sophisticated approach is to extend the annotation of a document by using ontologies and other information available on the World Wide Web. This has been done in the text retrieval domain by Gabrilovich and Markovitch [15], in the biomedical abstract retrieval domain by Doms and Schroeder [16], and in the image retrieval domain by Kutics et al. [17].

As there exist a large number of studies and evaluation campaigns using different sets of keywords, we present an overview of keyword vocabulary creation for describing images in Section 3.

2.2. Annotations based on ontologies

An ontology is a *specification of a conceptualisation* [18]. It basically contains concepts (entities) and their relationships and rules. Adding a hierarchical structure to a collection of keywords produces a *taxonomy*, which is an ontology as it encodes the relationship "is a" (a dog is an animal). An ontology can solve the problem that some keywords are ambiguous. For example, a "leopard" could be a large cat, a tank, a gecko or a Mac operating system. Ontologies are important for the Semantic

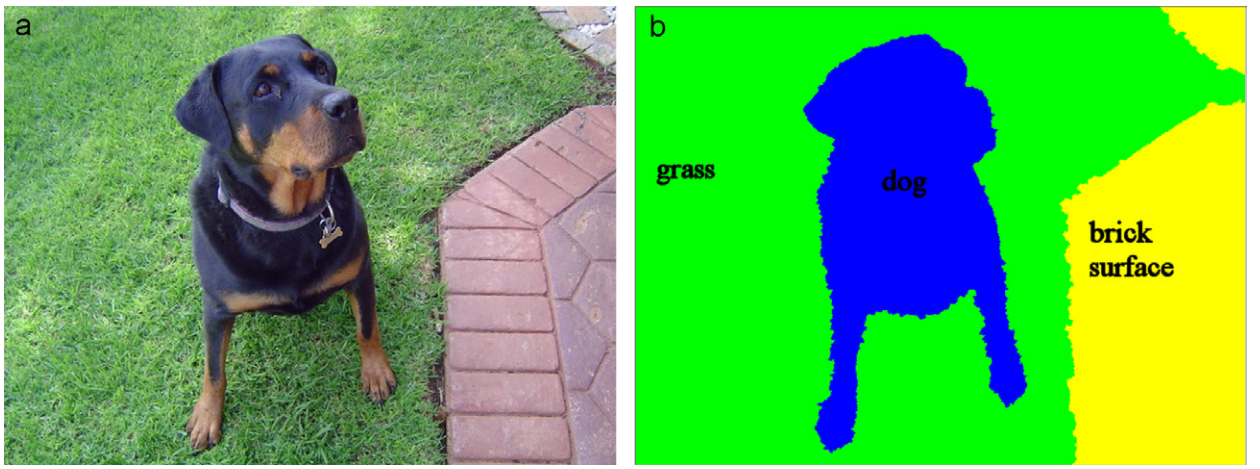

Fig. 1. Examples of image annotation: (a) whole image annotation—the listed keywords are associated with the image (outdoors, dog, grass, brick surface); (b) segmentation and annotation—keywords are associated with each region of the segmentation. Keywords describing the whole image can also be used (outdoors).

Web, and hence a number of languages exist for their formalisation, such as OWL and RDF.

Work on the development of ontologies which aim to arrange all the concepts in the world into a hierarchical structure is not new. One of the first comprehensive attempts was made by Wilkins [19] in 1668. One of the main problems is that there are many possible logical ways of classifying concepts, which also depend for example on the influence of culture [20]. Developing ontologies to describe even very limited image domains is a complicated process, as can be seen in the work by Schreiber et al. [21], who develop an ontology for describing photographs of apes, and by Hyvönen et al. [14], who develop an ontology for describing graduation photographs at the University of Helsinki and its predecessors.

In the domain of image description, ICONCLASS [22] is a very detailed ontology for iconographic research and the documentation of images, used to index or catalogue the iconographic contents of works of art, reproductions, literature, etc. It contains over 28 000 definitions organised in a hierarchical structure. Each definition is described by an alphanumeric code accompanied by a textual description (textual correlate). For example, the code 47D31 refers to "windmill" and translates into the following hierarchy:

4 society, civilisation, culture
47 crafts and industries
47D machines; parts of machines; tools and appliances
47D3 machine driven by wind
47D31 windmill

Note that this is distinct from the concept of "windmill in landscape" which, falls into a completely different category. It has the code 25I41, which translates into:

2 nature
25 earth, world as celestial body
25I city-view, and landscape with man-made constructions
25I4 factories and mills in landscape
25I41 windmill in landscape

Many very specific events are also encoded in the hierarchy, for example, the code 11H(GEORGE)65 corresponds to:

1 religion and magic
11 Christian religion
11H saints
11H (...) male saints (with NAME)
11H(GEORGE) the warrior martyr George (Georgius); possible attributes: banner (red cross on white field), (red) cross, dragon, (white) horse, broken lance, shield (with cross), sword
11H(GEORGE)6 martyrdom, suffering, misfortune, death of St. George
11H(GEORGE)65 St. George is torn apart by horses

As can be seen, this is a very complete ontology, which contains much more information than can currently be extracted from images using automated methods. The assignment of its classes is also open to interpretation—for the windmill example given above, is it a landscape containing a windmill, or is the windmill the focal point?

2.3. Free text annotation

For this type of annotation, the user can annotate using any combination of words or sentences. This makes it easy to annotate, but more difficult to use the annotation later for image retrieval. Often this option is used in addition to the choice of keywords or an ontology. This is to make up for the limitation stated in [21]: "There is no way the domain ontology can be complete—it will not include everything a user might want to say about a photograph". Any concepts which cannot adequately be described by choosing keywords are simply added in free form description. This is the approach used in the W3C *RDFPic* software [23] in which the content description keywords are limited to the following: Portrait, Group-portrait, Landscape, Baby, Architecture, Wedding, Macro, Graphic, Panorama and Animal. This is supplemented by a free text description. The IBM VideoAnnEx software [24] also provides this option.

The ImageCLEF 2004 [25] bilingual ad hoc retrieval task used 25 categories of images each labelled by a semi-structured title (in 13 languages). Examples of the English versions of these titles are:

- portrait pictures of church ministers by Thomas Rodger;
- photos of Rome taken in April 1908;
- views of St. Andrews cathedral by John Fairweather;

Fig. 2. The annotation of one of the images in the IAPR-TC12 dataset (from [27]).

- men in military uniform, George Middlemass Cowie;
- fishing vessels in Northern Ireland.

The IAPR-TC12 dataset of 20 000 images [26] contains free text descriptions of each image in English, German and Spanish. These are divided into "title", "description" and "notes" fields. Additional content-independent metadata such as date, photographer and location are also stored. Fig. 2 shows the annotation of one of the photos.

3. Image annotation in practice

The practical aspects of image annotation are covered in this section. This includes the creation of keyword vocabularies and methods for making manual annotation more efficient.

There are two approaches to associating textual information with images described in the computer vision literature: *annotation* and *categorisation*. In annotation, keywords or detailed text descriptions are associated with an image, whereas in categorisation, each image is assigned to one of a number of predefined categories [28]. This can range from more general two category classification, such as *indoor/outdoor* [29] or *city/landscape* [30] to more specific categories such as *African people and villages*, *Dinosaurs*, *Fashion* and *Battle ships* [28]. Categorisation can be used as an initial step in image understanding in order to guide further processing of the image. For example, in [31] a categorisation into textured/non-textured and graph/photograph classes is done as a pre-processing step. *Recognition* is concerned with the identification of particular object instances. *Object recognition* would distinguish between images of two structurally distinct cups [4], while *category-level object recognition* [32] would place them in the same class. Recognition also has its uses in annotation, for example in the recognition of family members in the automatic annotation of family photos. Category-level object recognition can at present be seen as annotation using a small keyword vocabulary. This is because current category-level object recognition algorithms tend to be capable of recognising only a few objects, for example from 10 categories in the PASCAL visual object classes (VOC) 2006 challenge [33] to 101 categories in [34]. As object recognition algorithms improve, it is to be expected that the vocabulary sizes will increase.

The best, but also the most labour intensive, method for creating ground truth for algorithm evaluation is to first create the required keyword vocabulary, and then to manually annotate the images using these keywords (or preferably annotate segmented images so that keywords are associated with specific areas of the image). For tasks with a small vocabulary such as many categorisation tasks, this approach is more feasible, as is demonstrated by the categorisation of the 15 200 images in the CEA-CLIC dataset [35], divided into 16 main categories each containing up to 15 sub-categories. An overview of research

toward the creation of vocabularies is given in Section 3.1. Due to the extensive effort needed to do manual image annotation, various approaches have been developed to simplify the task. These are described in Section 3.2. An overview of annotated image datasets available for computer vision research, including the size of the keyword vocabulary used to annotate each dataset and the annotation methods used is presented in Section 3.3.

3.1. Creating a vocabulary for image annotation

While a number of ontologies and vocabularies are available, they tend to suffer from at least one of the following disadvantages listed in [36]:

- The vocabularies or ontologies developed for commercial purposes, such as those belonging to CORBIS and Getty Images, are proprietary competitive tools and are not available for public use.
- The vocabularies or ontologies developed for specific areas of application, such as the Iconclass ontology described in Section 2.2, while containing a wealth of terms, are concentrated on too narrow a domain to be useful for annotating general collections of images.

There are a number of criteria that affect the construction and usefulness of a vocabulary. One is the range of terms to be included [36]. This is tied closely to the planned use of the vocabulary and the specification of which information should be included in an image annotation. A vocabulary including a wide range of terms, ranging from names of objects to emotions provoked by an image is applicable in a wide range of situations. However, annotating an image with all the expressive capability of such a vocabulary will most likely be time-consuming. If the annotated images are to be used to evaluate object recognition algorithms, then some of the annotation will exceed the requirements of the task. Solutions are to use an extensive vocabulary with additional annotation guidelines which restrict the parts of the vocabulary to be used, or to create a restricted vocabulary containing only keywords suitable to the task at hand. A further design criteria to be considered is how to impose a suitable hierarchical (or other) structure on the vocabulary. As there exist a large number of acceptable logical ways to group keywords (see Section 2.2), the hierarchy should also be designed to simplify the finding of the correct keyword during the manual annotation process.

It is possible to use WordNet as a vocabulary, thereby including an extremely wide range of terms. WordNet is an on-line lexical reference system which organises English nouns, verbs and adjectives into synonym sets, each representing one underlying lexical concept [37]. For example, Barnard et al. [38] gave the full WordNet vocabulary along with a set of annotation guidelines to people producing the ground truth for their recognition evaluation dataset. WordNet has also been used as the basis for creating a more restricted vocabulary. Zinger et al. [39] construct an ontology of *portrayable objects* by pruning the WordNet tree. They began with the subclass "object" of the class "entity" and extracted a tree with 102 nodes in the level below "object" and 24 000 words describing portrayable objects in the leaf nodes of the tree.

An effort, described in [36], was begun to create a vocabulary of 12 000 to 15 000 terms for general collections of images. This was done in a first stage by gathering a large number of terms from existing vocabularies for image classification followed by the merging of vocabulary lists created by a number of participants. The expansion of the vocabulary in the second stage was done by examining sources of images such as multilanguage visual dictionaries and specialised reference works. Unfortunately the work on this vocabulary seems to have been abandoned. Development of a more focused ontology for broadcast video is currently underway. In the LSCOM *Large Scale Concept Ontology for Broadcast Video* [40], it is intended to find 1000 concepts in broadcast news video that can be detected and evaluated. Version 1.0 of this ontology [41] contains 856 concepts.

Researchers often do not pay much attention to the development of a good vocabulary, and are often restricted to using annotations which are already available due to having limited resources to expend on manual annotation.

3.2. Ground truth annotation collection methods

The manual annotation of images is a very labour-intensive and time-consuming task. This usually has the effect that comprehensively annotated datasets contain few images, while datasets with more images are more "lightly" annotated. An example of the former is the Sowerby database [42], which contains 250 images with manually corrected

segmentations and a keyword assigned to each region of the segmentation. The images are all of rural or urban outdoor scenes, and the 85 word vocabulary is limited to this subject matter. Barnard et al. [38] created a larger set of manually labelled segmented images: the regions on 1014 manually segmented images were labelled. WordNet was used as a controlled vocabulary, and the annotators were provided with a set of annotation guidelines. The guidelines dealing with WordNet are:

- words should correspond to their WordNet definition;
- the sense in WordNet (if multiple) should be mentioned as word(i), where i is the sense number in WordNet except if $i = 1$ (e.g. tiger(2));
- add the first synonym given in WordNet as an additional entry (e.g. building edifice).

Other guidelines deal with the words (should be lowercase and singular), what to label as "background", etc. (the full set of guidelines is available in [38]). The regions were labelled by 1297 keywords, as well as two special keywords "unknown" and "background". A dataset containing a large number of manually annotated images, but without information on the relations between words and locations in the image, is the IAPR-TC12 dataset [26], which contains 20 000 images comprehensively annotated with free text.

To avoid the overhead of manual annotation, annotations which "already exist" have often been converted into a form useful for the evaluation of keyword annotation. In [43], free text annotations are converted to keyword annotations using a part of speech tagger allowing certain parts of speech to be retained, and WordNet for sense disambiguation. This resulted in a vocabulary of 3319 words. The annotations of groups of 100 Corel images have also been used, although this batch annotation does not always provide sensible annotations for individual images. Further discussion on the use of the Corel database can be found in [44].

Various approaches and systems to simplify the collection of image annotations or to receive input from a large number of people have been set up. The simplest is to get a group of people together to create the annotations—the PASCAL VOC challenge [45] organises a yearly "annotation party" where a group annotates intensively over 3–4 days. This is found to be more effective than distributed asynchronous annotation.

The collaborative potential of the World Wide Web is widely used to obtain image annotations, as the following annotation approaches show. Users of the *Gimp-Savvy Community-Indexed Photo Archive* website,[1] an archive containing more then 27 000 free photos and images, are requested to annotate the images using keywords which they are free to choose (tips on choosing keywords are made available). That this "free annotation by all" approach has not been totally successful can be seen by the extremely large number of "junk" keywords on the master keyword list as well as the over-annotation (assignment of too many keywords) of many of the images. On the *Flickr*[2] photo archive, users who upload photos may also assign keywords to them. These are then used to search for images. Other users may add comments to the images. There is no standardised keyword list and no restriction on which language is to be used, so this database represents a good example of the annotation practice of amateur photographers on their own images. It can also be seen here that false keywords are often added to images, which affect the search results.

An on-line annotation application aimed at collecting keywords describing image regions for object recognition evaluation is the LabelMe tool [46]. Here, the user clicks the vertices of a polygon around an object and then enters a keyword describing the object. As the vocabulary is not controlled, multiple keywords and misspelled keywords often occur. This problem is solved by a verification step by the database administrators. The incentive to annotate the images is that the annotator may then download the latest annotations.

An innovative approach to collecting annotations of images by keywords has been developed by von Ahn and Dabbish [47]. In their ESP game, they aim to make the annotation of images enjoyable. Players access the ESP game server and are paired randomly. They have no way of communicating with each other. Pairs of players are shown 15 images during the game, with the aim being for both players to type in the same keyword for an image so as to advance to the next. This is an intelligent way of avoiding the problem of "junk" keywords, as the pairs of players verify the keywords. Keywords which are typed often for an image are added to a

[1] http://gimp-savvy.com/PHOTO-ARCHIVE/.
[2] http://www.flickr.com.

"taboo" list shown for that image, and may no longer be entered as keywords by the players. The keywords entered correspond to the whole image, although the authors have discussed implementing, for example, a "shooting game", where the players have to click on the requested object. The Peekaboom game [48] from the same research group is of this type. An image search engine based on the keywords collected from the ESP game for about 30 000 images is accessible on the web.

An alternative approach is to start with keywords and collect images illustrating these keywords. For example, in [49] the results of a text-based Google image search are post-processed using a combination of manual and automated methods to weed out false images. While such an approach is most likely useful for collecting data for the evaluation of object recognition tasks having a small vocabulary, it is still demanding if used for larger vocabularies. Issues in dataset creation for object recognition evaluation are discussed in [50].

Table 1 summarises the methods presented above, lists the decisions needed for each method, as well as the advantages and disadvantages of each method.

3.3. Annotated image datasets for computer vision research

Datasets of annotated images are widely used as ground truth in object recognition and automated image annotation research. Table 2 lists papers describing research and evaluation campaigns that have created such annotated datasets (and which have been made available on-line). The number of keywords used in each dataset and the annotation method used (cf. Table 1) are also listed.

Among the datasets aimed at object recognition evaluation, the largest vocabularies are used in those by Fei-Fei et al. [34] and from the PASCAL VOC Challenge 2005 [45]. The latter consisted of classification and detection tasks for four objects: motorbikes, bicycles, people and cars. However, in the dataset collection created as a part of this challenge, five datasets are provided with standardised ground truth object annotations. The 101 keywords referred to in Table 2 correspond to this dataset collection. In the PASCAL VOC Challenge 2006, 10 objects were to be recognised [33]. As part of the EU LAVA project, a dataset consisting of 10 categories of images was made available. The

Table 1
Summary of image annotation methods

Method	Associated decisions	Advantages	Disadvantages
Direct manual annotation	–	Generally good annotations due to world knowledge of the annotators	Time-consuming and labour-intensive (proportional to level of detail required)
			Possible inconsistency between annotators
Intensive group annotation ("annotation party")	Event logistics	Could be more effective than distributed asynchronous annotation	Event organisation overhead
			Possible inconsistency between annotators
Convert existing annotations	How to convert to the format required	No manual annotation necessary	The existing annotations may not meet the requirements for the new task
Collaborative annotation over the www	System design	Large group of potential annotators	Possible inconsistency between annotators
	Motivation for annotators		Possible "junk" annotations
			Motivating the annotators
Search the www for images illustrating keywords	Design of a methodology and system	Large pool of illustrative images	Possible copyright restrictions on the images found
			Checking for false positive images found by the system

Table 2
Annotated image datasets available for computer vision research

Source	# Keywords	Annotation method
PASCAL VOC challenge 2005 databases [45]	101	Manual annotation
PASCAL VOC challenge 2006 dataset [33]	10	"Annotation party"
EU LAVA Project [4,51]	10	Manual annotation
Chen and Wang [28]	20	Corel annotations
Microsoft Research Cambridge Databases [8,52]	35	Manual annotation
Fei-Fei et al. [34]	101	www image search and then manual filtering
Carbonetto et al. [3]	55	Corel annotations
Li and Wang [6]	433	Corel annotations
Barnard et al. [1]	323	Corel annotations
University of Washington Ground Truth Image Database	392	Manual annotation

The left column gives the source and references, the middle column gives the number of keywords used and the right column describes the annotation method used. Some of the datasets are discussed in more detail in the text.

version of the dataset used in the cited papers [4,51] has only 7 categories, including a face category which is not available for download. Two datasets have been released by Microsoft Research in Cambridge.[3] The "Database of thousands of weakly labelled, high-res images" contains images divided into 23 categories. Some of these are divided into sub-classes, such as different views of cars. The "Pixel-wise labelled image database" contains 591 images in which regions are manually labelled using the 23 labels. Combining the keyword lists results in 33 unique keywords.

The papers by Carbonetto et al. [3], Li and Wang [6] and Barnard et al. [1] on automatic image or image region annotation use parts of the Corel image dataset along with keywords usually extracted from the annotations accompanying the Corel images. Li and Wang [6] used 600 categories of image, and to each category assigned on average 3.6 keywords. Each of the 100 images in each category was then assigned the same keywords associated with the category. For example, all images in the "Paris/France" category were assigned the keywords "Paris, European, historical building, beach, landscape, water" and the images in the "Lion" category were assigned the keywords "lion, animal, wildlife, grass". The University of Washington Ground Truth Image Database,[4] which is used by Hardoon et al. [53] and Frigui and Caudill [54], contains publicly available images that have been manually annotated with an average of 5 keywords per image.

4. Summary and conclusion

We discuss the use of image annotation in the creation of ground truth for the evaluation of object recognition and automated image annotation algorithms. We give an overview of three different types of image annotation: free text annotation, keyword annotation and annotation using ontologies.

Creating an annotation of a set of images can be done in a various ways. However there are always a set of decisions to be made before beginning the annotation process. These decisions and the sections of the paper covering them are summarised here:

(1) Annotation method (Section 3.2 and Table 1).
(2) Type of annotation: free text, freely chosen keywords, keywords from a controlled vocabulary or terms from an ontology (Section 2).
(3) If a controlled vocabulary or ontology is required, how it should be chosen or created (Section 3.1).
(4) Whether the annotation will be for the whole image or specific to sub-regions of the image. For the latter, decisions on the form of the sub-regions and the method for creating them must be taken.

The best method for creating ground truth is first to create a keyword vocabulary based on the requirements of the evaluation task and then to use this vocabulary in the manual annotation of images. As this approach is time-consuming and labour-intensive, various methods to reduce the manual annotation effort have been used. Promising new annotation methods make use of either the collaborative potential or the current search capabilities of the www to annotate images more efficiently.

[3]Version 1 of the pixel-wise labelled image dataset has been ignored here, as it forms a subset of version 2.

[4]http://www.cs.washington.edu/research/imagedatabase/groundtruth.

References

[1] K. Barnard, P. Duygulu, N. de Freitas, D. Forsyth, D. Blei, M.I. Jordan, Matching words and pictures, Journal of Machine Learning Research 3 (2003) 1107–1135.

[2] K. Barnard, P. Duygulu, R. Guru, P. Gabbur, D. Forsyth, The effects of segmentation and feature choice in a translation model of object recognition, in: Proceedings on Computer Vision and Pattern Recognition, 2003, pp. II:675–682.

[3] P. Carbonetto, N. de Freitas, K. Barnard, A statistical model for general contextual object recognition, in: Proceedings of the European Conference on Computer Vision (ECCV), 2004, pp. I:350–362.

[4] G. Csurka, C.R. Dance, L. Fan, J. Willamowski, C. Bray, Visual categorization with bags of keypoints, in: Workshop on Statistical Learning in Computer Vision (at ECCV), 2004.

[5] R. Datta, W. Ge, J. Li, J.Z. Wang, Toward bridging the annotation-retrieval gap in image search by a generative modeling approach, in: Proceedings of the ACM Multimedia Conference, 2006.

[6] J. Li, J.Z. Wang, Automatic linguistic indexing of pictures by a statistical modeling approach, IEEE Transaction on Pattern Analysis and Machine Intelligence 25 (9) (2003) 1075–1088.

[7] J. Li, J.Z. Wang, Real-time computerized annotation of pictures, in: Proceedings of the ACM Multimedia Conference, 2006.

[8] J. Winn, A. Criminisi, T. Minka, Object categorization by learned universal visual dictionary, in: Proceedings of the International Conference on Computer Vision (ICCV), 2005, pp. 1800–1807.

[9] A.W.M. Smeulders, M. Worring, S. Santini, A. Gupta, R. Jain, Content-based image retrieval at the end of the early years, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (12) (2000) 1349–1380.

[10] C.H.C. Leung, H.H.-S. Ip, Benchmarking for content-based visual information search, in: Proceedings of the 4th International Conference on Advances in Visual Information Systems, 2000, pp. 442–456.

[11] C. Fluhr, P.-A. Moëllic, P. Hede, Usage-oriented multimedia information retrieval technological evaluation, in: MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval, 2006, pp. 301–306.

[12] A. del Bimbo, Visual Information Retrieval, Morgan Kaufmann, Los Altos, CA, 1999.

[13] B. Manjunath, P. Salembier, T. Sikora (Eds.), Introduction to MPEG-7: Multimedia Content Description Interface, Wiley, New York, 2002.

[14] E. Hyvönen, A. Styrman, S. Saarela, Ontology-based image retrieval, in: Proceedings of XML Finland Conference, 2002, pp. 27–51.

[15] E. Gabrilovich, S. Markovitch, Feature generation for text categorization using world knowledge, in: Proceedings of the 19th International Joint Conference for Artificial Intelligence, Edinburgh, Scotland, 2005, pp. 1048–1053.

[16] A. Doms, M. Schroeder, GoPubMed: exploring PubMed with the gene ontology, Nucleic Acids Research 33.

[17] A. Kutics, A. Nakagawa, S. Arai, H. Tanaka, S. Ohtsuka, Relating words and image segments on multiple layers for effective browsing and retrieval, in: Proceedings of the International Conference on Image Processing, 2004, pp. 2203–2206.

[18] T.R. Gruber, A translation approach to portable ontology specifications, Knowledge Acquisition 5 (2) (1993) 199–220.

[19] J. Wilkins, An Essay Towards a Real Character and a Philosophical Language, London, 1668.

[20] G. Lakoff, Women, Fire, and Dangerous Things: What Categories Reveal About the Mind, University of Chicago Press, 1987.

[21] A.T.G. Schreiber, B. Dubbeldam, J. Wielemaker, B. Wielinga, Ontology-based photo annotation, IEEE Intelligent Systems 16 (3) (2001) 66–74.

[22] H. van de Waal, et al., Iconclass, An Iconographic Classification System (17 volumes), North-Holland, Amsterdam, 1985.

[23] Y. Lafon, B. Bos, Describing and retrieving photos using RDF and HTTP, W3C. Note ⟨http://www.w3.org/TR/photo-rdf/⟩, 2002 (accessed: 17.5.2006).

[24] C.-Y. Lin, B.L. Tseng, J.R. Smith, Videoannex: IBM MPEG-7 annotation tool for multimedia indexing and concept learning, in: Proceedings of the IEEE International Conference on Multimedia and Expo (ICME), 2003.

[25] C. Peters, P. Clough, J. Gonzalo, G. Jones, M. Kluck, B. Magnini (Eds.), Multilingual Information Access for Text, Speech and Images, Lecture Notes on Computer Science, vol. 3491, Springer, Berlin, 2004.

[26] M. Grubinger, P. Clough, H. Müller, T. Deselaers, The IAPR TC-12 benchmark—a new evaluation resource for visual information systems, in: Proceedings of the International Workshop OntoImage'2006, 2006, pp. 13–23.

[27] M. Grubinger, C. Leung, P. Clough, The IAPR benchmark for assessing image retrieval performance in cross language evaluation tasks, in: Proceedings of the MUSCLE/ImageCLEF Workshop on Image and Video Retrieval Evaluation, 2005, Vienna, Austria, pp. 17–23.

[28] Y. Chen, J.Z. Wang, Image categorization by learning and reasoning with regions, Journal of Machine Learning Research 5 (2004) 913–939.

[29] M. Szummer, R.W. Picard, Indoor–outdoor image classification, in: Proceedings of the IEEE International Workshop on Content-based Access of Image and Video Databases, 1998, pp. 42–51.

[30] A. Vailaya, M.A.T. Figueiredo, A.K. Jain, H.-J. Zhang, Image classification for content-based indexing, IEEE Transactions on Image Processing 10 (1) (2001) 117–130.

[31] J.Z. Wang, J. Li, G. Wiederhold, SIMPLIcity: semantics-sensitive integrated matching for picture libraries, IEEE Transactions on Pattern Analysis and Machine Intelligence 23 (9) (2001) 947–963.

[32] J. Ponce, M. Hebert, C. Schmid, A. Zisserman (Eds.), Toward Category-Level Object Recognition, Lecture Notes in Computer Science, vol. 4170, Springer, Berlin, 2006.

[33] M. Everingham, A. Zisserman, C. Williams, L. Van Gool, The Pascal visual object classes challenge 2006 (VOC 2006) results, Technical Report, PASCAL Network of Excellence, 2006.

[34] L. Fei-Fei, R. Fergus, P. Perona, Learning generative visual models from few training examples an incremental Bayesian approach tested on 101 object categories, in: Proceedings of the Workshop on Generative-Model Based Vision, 2004.

[35] P.-A. Moëllic, P. Hède, G. Grefenstette, C. Millet, Evaluating content based image retrieval techniques with the one million images clic testbed, in: Proceedings of the Second World Enformatika Congress, WEC'05, 2005, pp. 171–174.
[36] C. Jörgenson, P. Jörgenson, Testing a vocabulary for image indexing and ground truthing, in: Proceedings of Internet Imaging III, 2002, pp. 207–215.
[37] G.A. Miller, R. Beckwith, C. Fellbaum, D. Gross, K. Miller, Introduction to WordNet: an on-line lexical database, International Journal of Lexicography 3 (4) (1990) 235–244.
[38] K. Barnard, Q. Fan, R. Swaminathan, A. Hoogs, R. Collins, P. Rondot, J. Kaufhold, Evaluation of localized semantics: data, methodology, and experiments, Technical Report TR-05-08, Computing Science, University of Arizona, 2005.
[39] S. Zinger, C. Millet, B. Mathieu, G. Grefenstette, P. Hède, P.-A. Moëllic, Extracting an ontology of portrayable objects from WordNet, in: Proceedings of the MUSCLE/ImageCLEF Workshop on Image and Video Retrieval Evaluation, 2005, Vienna, Austria, pp. 17–23.
[40] A.G. Hauptmann, Towards a large scale concept ontology for broadcast video, in: Proceedings of the 3rd International Conference on Image and Video Retrieval, 2004, pp. 674–675.
[41] L. Kennedy, A. Hauptmann, M. Naphade, J. R. Smith, S.-F. Chang, Lscom lexicon definitions and annotations version 1.0, Technical Report ADVENT #217-2006-3, Columbia University, March 2006.
[42] D. Collins, W.A. Wright, P. Greenway, The Sowerby image database, in: Proceedings of the 7th International Conference on Image Processing and its Applications, vol. 1, 1999, pp. 306–310.
[43] K. Barnard, P. Duygulu, D. Forsyth, Clustering art, in: Proceedings of Computer Vision and Pattern Recognition, 2001, pp. II:434–441.
[44] H. Müller, S. Marchand-Maillet, T. Pun, The truth about Corel—evaluation in image retrieval, in: Proceedings of the Conference on Image and Video Retrieval (CIVR), 2002, pp. 38–49.
[45] M. Everingham, A. Zisserman, C. Williams, L. Van Gool, M. Allan, C. Bishop, O. Chapelle, N. Dalal, T. Deselaers, G. Dorko, S. Duffner, J. Eichhorn, J. Farquhar, M. Fritz, C. Garcia, T. Griffiths, F. Jurie, D. Keysers, M. Koskela, J. Laaksonen, D. Larlus, B. Leibe, H. Meng, H. Ney, B. Schiele, C. Schmid, E. Seemann, J. Shawe-Taylor, A. Storkey, S. Szedmak, B. Triggs, I. Ulusoy, V. Viitaniemi, J. Zhang, The 2005 PASCAL visual object classes challenge, in: Selected Proceedings of the 1st PASCAL Challenges Workshop, Springer, Berlin, 2006.
[46] B.C. Russell, A. Torralba, K.P. Murphy, W.T. Freeman, LabelMe: a database and web-based tool for image annotation, Technical Report AIM-2005-025, MIT AI Lab, September 2005.
[47] L. von Ahn, L. Dabbish, Labeling images with a computer game, in: Proceedings of ACM CHI, 2004, pp. 319–326.
[48] L. von Ahn, R. Liu, M. Blum, Peekaboom: a game for locating objects in images, in: Proceedings of ACM CHI, 2006.
[49] T.L. Berg, D.A. Forsyth, Animals on the web, in: Proceedings of the Conference on Computer Vision and Pattern Recognition, vol. 2, 2006, pp. 1463–1470.
[50] J. Ponce, T.L. Berg, M. Everingham, D. Forsyth, M. Hebert, S. Lazebnik, M. Marszalek, C. Schmid, B. Russell, A. Torralba, C.K.I. Williams, J. Zhang, A. Zisserman, Dataset issues in object recognition, in: J. Ponce, M. Hebert, C. Schmid, A. Zisserman (Eds.), Toward Category-Level Object Recognition, Springer, Berlin, 2006, pp. 30–50.
[51] F. Perronnin, C. Dance, G. Csurka, M. Bressan, Adapted vocabularies for generic visual categorization, in: Proceedings of the European Conference on Computer Vision (ECCV), 2006, pp. IV:464–475.
[52] J. Shotton, J. Winn, C. Rother, A. Criminisi, TextonBoost: joint appearance, shape and context modeling for multi-class object recognition and segmentation, in: Proceedings of the European Conference on Computer Vision (ECCV), 2006, pp. I:1–15.
[53] D.R. Hardoon, C. Saunders, S. Szedmak, J. Shawe-Taylor, A correlation approach for automatic image annotation, in: Proceedings of 2nd International Conference on Advanced Data Mining and Applications, 2006, pp. 681–692.
[54] H. Frigui, J. Caudill, Building a multi-modal thesaurus from annotated images, in: Proceedings of the International Conference on Pattern Recognition ICPR, vol. 4, 2006, pp. 198–201.