Technical Report

Pattern Recognition and Image Processing Group
Institute of Computer Aided Automation
Vienna University of Technology
Favoritenstr. 9/1832
A-1040 Vienna AUSTRIA
Phone: +43 (1) 58801-18351
Fax: +43 (1) 58801-18392
E-mail: hanbury@prip.tuwien.ac.at
URL: http://www.prip.tuwien.ac.at/

PRIP-TR-102 May 16, 2006

Review of Image Annotation for the Evaluation of Computer Vision Algorithms

Allan Hanbury

Abstract

In the field of computer vision, automated image annotation and object recognition are currently important research topics. It is hoped that these will lead to improved general image understanding which can be usefully applied in Content-based Image Retrieval. Three approaches to image annotation are reviewed: free text annotation, keyword annotation and annotation based on ontologies. An analysis of the keywords which have been used in automated image and video annotation research and evaluation campaigns is then presented. The outcome of this analysis is a list of 525 keywords divided into 15 categories. Given that this list is collected from existing image annotations, it could be used to check the applicability of ontologies describing entities which are portrayable in images.

1 Introduction

The usual reason to annotate data (i.e. add metadata to it) is to simplify access to it. This is particularly important for the semantic web. The metadata added to documents or images allow for more effective searches. The problem with adding metadata manually is that it is an extremely labour-intensive and time-consuming task. In the field of computer vision, automated image annotation and object recognition are currently important research topics [2, 4, 6, 19, 29]. This automatic generation of image metadata should allow image searches and Content-Based Image Retrieval (CBIR) to be more effective. For example, 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, 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 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 can 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 [18]. 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 online searching for them, semantic richness is an advantage. For images, one can create complex ontologies allowing the specification of objects and actions. For example, in [23], such an ontology is created for annotating photographs of apes. One can specify the type of ape, how old it is and what it is doing. 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 as a human annotator. The current state-of-the-art in automated annotation tends to operate at an extremely low level — for example, there is still no algorithm that can make an error-free distinction between images of cities and images of landscapes, or which can make an error-free decision as to the presence or absence of human faces in an image.

Evaluating the abilities of current algorithms requires a rather low level of annotation. For example, the TRECVID 2005 high-level feature detection task tested automatic detection of only 10 concepts. The IBM MARVEL Multimedia Search Engine[1] extracts only six concepts in the online image retrieval demo version[2] (face, human, indoor, outdoor, sky, nature). Carbonetto et al. [4] use a vocabulary of at most 55 keywords. The largest number of keywords have been used by Li and Wang [19], who assigned 433.

Three types of annotation: free-text annotations, keyword annotations and classifications based on ontologies are described in Section 2. A good way of collecting keywords which would be useful in an ontology describing images is to analyse the vocabularies used in the ground truth

[1]Information is available here: `http://www.research.ibm.com/marvel`

[2]The demo is available for download here: `http://www.alphaworks.ibm.com/tech/marvel`

of image annotation and object recognition tasks. In this way, one can find out which words are important in applications and which words correspond to objects which can be detected using current state-of-the-art image understanding algorithms. We analyse the annotations which have been used in image and video understanding publications and evaluation campaigns in Section 3. Section 4 concludes.

2 Annotation approaches

Different types of information can be associated with images or videos. They are [7]:

- *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 the Houses of Parliament or London 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. [24]. Content-descriptive metadata can be specified using one or more of the following approaches [14], listed in order of increasing structure:

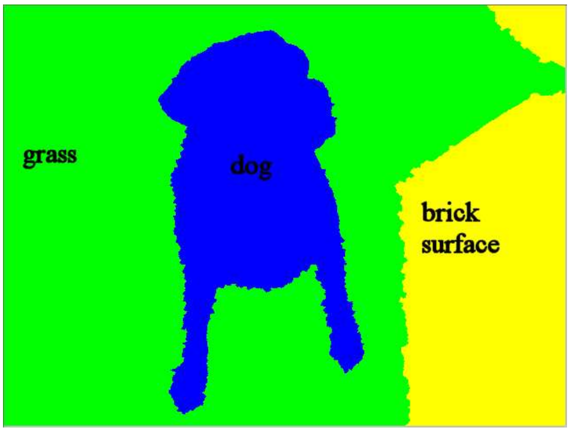
Free text descriptions: No pre-defined structure for the annotation is given.

Keywords: Arbitrarily chosen keywords or keywords chosen from *controlled vocabularies*, i.e. limited 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.

(a) outdoors, dog, grass, brick surface



(b) outdoors

Figure 1: Examples of image annotation: (a) Whole image annotation – the listed keywords are associated with the image. (b) Segmentation and annotation – keywords are associated with each segment. Keywords describing the whole image can also be used (shown below the image).

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 in two levels of specificity:

1. A list of keywords associated with the complete image, listing what is in the image (see Figure 1a for an example).
2. A segmentation of the image along with keywords associated with each segment (region of the segmentation). In addition, keywords describing the whole image can be provided (see Figure 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.

Keyword lists are currently widely used in annotating image archives. For example, an extensive one is in use at the GETTYIMAGES archive[3]. While the full list of keywords does not seem to be available, parts of this list divided into different categories are available in the *Keyword Guide*[4]. Many of the keywords given here, such as “Body concern”, “Futility”, “Greed” and

[3] `http://www.gettyone.com`

[4] Available for download here: `http://corporate.gettyimages.com/marketing/m01/PDF/Keyword_UK_1_Jan_05.pdf`

“Wolf in sheep’s clothing” are of limited use for evaluating current automated image retrieval or annotation algorithms. However, in many usage scenarios, motivating users to annotate images correctly is difficult. One of the main application areas is in simplifying access to personal multimedia collections (photo collections, etc.). In this area, it is difficult to motivate users to annotate the images at all [15], and hence impractical to request that they use a controlled vocabulary.

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

1. Different image collections are annotated using different keyword sets and differing annotation standards.

2. A naive user does not necessarily know the list of keywords which has been used to annotate an image collection. This makes searching by text input more difficult.

Forcing the user to choose from a list of keywords is a solution, but this makes the search task more frustrating. As a solution to both the above problems, the GETTYIMAGES search engine uses a thesaurus to extend the list of search words entered by a user. A more sophisticated approach is to extend one’s knowledge or 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 [10], in the biomedical abstract retrieval domain by Doms and Schroeder [8], and in the image retrieval domain by Kutics et al [16].

As there exist so many studies and evaluation campaigns using different sets of keywords, we present an overview and analysis of keywords for describing images in Section 3.

2.2 Annotations based on ontologies

An ontology is a *specification of a conceptualization* [11]. It basically contains concepts (entities) and their relationships and rules. Adding a hierarchical structure to a list of keywords produces a *taxonomy*, which is an ontology as it encodes the relationship “is a” (a dog is an animal).

Ontologies are important for the Semantic Web[5], and hence a number of languages exist for their formalisation, such as OWL[6] and RDF[7]. Developing ontologies to describe even very limited image domains is a complicated process, as can be seen in the papers by Schreiber et al. [23], 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.

ICONCLASS[8] 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

[5] `http://www.w3.org/2001/sw/Activity`
[6] `http://www.w3.org/TR/owl-features/`
[7] `http://www.w3.org/RDF/`
[8] `http://www.iconclass.nl`

correlate). For example, the code 47D31 refers to “windmill” and translates into the following hierarchy:

4 Society, Civilization, 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

A lot of 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?

The use of the WordNet lexical database[9] is increasing in the computer vision community. WordNet is an online lexical reference system which organises English nouns, verbs and adjectives into synonym sets, each representing one underlying lexical concept [20]. Barnard et al. [3] gave the full WordNet vocabulary to people producing the ground truth for their recognition evaluation dataset. This involved labelling segments on 1014 manually segmented images. The annotators were also 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 [3]). Zinger et al. [30] 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 is currently underway to develop a more focused ontology for broadcast video. In the LSCOM *Large Scale Concept Ontology for Broadcast Video* [13], it is intended to find 1000 concepts in broadcast news video that can be detected and evaluated.

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 [23]: "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 [17] 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 [25] also provides this option.

The ImageCLEF 2004 [22] 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

[9] `http://wordnet.princeton.edu/`

Figure 2: The annotation of one of the images in the IAPR-TC12 dataset (from [12]).

- Views of St. Andrews cathedral by John Fairweather
- Men in military uniform, George Middlemass Cowie
- Fishing vessels in Northern Ireland

The full list of titles in all 13 languages is available for download[10].

The IAPR-TC12 dataset of 25 000 images [12] 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 is also stored. An example showing the annotation of one of the photos is given in Figure 2.

3 Analysis of Keywords used in Annotation Experiments

In this section we analyse the keywords that have been used in image annotation, categorisation and object recognition experiments and evaluation campaigns. To begin, a brief discussion on the difference between annotation and categorisation is presented in Section 3.1. Some methods currently used for collecting manual annotations of images are listed in Section 3.2. We then present an analysis of the keywords that have been used in image annotation experiments. The analysis was carried out in two steps. The first step consisted of creating a list combining all the keywords used in the experiments, datasets and evaluations considered and removing the unsuitable words (Section 3.3). The second step was the categorisation of keywords (Section 3.4). From a practical point of view, it is useful if the keywords are sorted into categories. When one is annotating images, this simplifies the choice of a word from the keyword list — one can select

[10] `http://ir.shef.ac.uk/imageclef2004/adhoc.html`

the category that the image belongs to in order to reduce the choice of keywords. The result of this analysis is a list of 525 keywords assembled from various sources and divided into 15 categories.

3.1 Annotation and Categorization

There are two approaches to associating textual information with images described in the 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 [5]. This can range from more general two category classification, such as *indoor/outdoor* [26] or *city/landscape* [27] to more specific categories such as *African people and villages*, *Dinosaurs*, *Fashion* and *Battle ships* [5]. Categorisation can be used as an initial step in image understanding in order to guide further processing of the image. For example, in [28] 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. Recognition would distinguish between images of two structurally distinct cups, while categorisation would place them in the same class [6]. Recognition also has its uses in annotation, for example in the recognition of family members in the automatic annotation of family photos.

Categorisation can be considered as annotation in which one must choose from a fixed number of keywords (the categories) and one is limited to assigning one keyword to each image. The discussion of annotation and categorisation is therefore combined in this section.

3.2 Manual annotation collection methods

The manual annotation of images is a very labour-intensive and time-consuming task. Various systems to simplify the collection of image annotations or to receive input from a large number of people have been set up.

An interesting experiment is taking place on the *Gimp-Savvy Community-Indexed Photo Archive* website[11]. This archive contains more then 27 000 free photos and images, and the users of the site are requested to annotate the images using keywords which they are free to choose (tips on choosing keywords are made available[12]). 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 list[13] as well as the over-annotation (assignment of too many keywords) of many of the images. On the *Flickr*[14] photo archive, people 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, so this database represents a good example of the annotation practice of amateur photographers on their own images.

An innovative approach to collecting annotations of images by keywords has been developed

[11] `http://gimp-savvy.com/PHOTO-ARCHIVE/`
[12] `http://gimp-savvy.com/PHOTO-ARCHIVE/tips_on_indexing.html`
[13] `http://gimp-savvy.com/cgi-bin/masterkeys.cgi`
[14] `http://www.flickr.com`

by von Ahn and Dabbish [1]. In their ESP game[15], 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 "taboo" list shown for that image, and can 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[16] 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[17].

An online annotation application aimed at collecting keywords for image regions is the LabelMe tool[18] by Bryan C. Russell at MIT. 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, as can be seen by examining the keyword statistics on the webpage[19]. This problem is solved by a verification step by the database administrators. At present[20], there are 101 verified keywords, the majority of which are shown in Table 2. The incentive to annotate the images is that the annotator is then allowed to download the latest annotations.

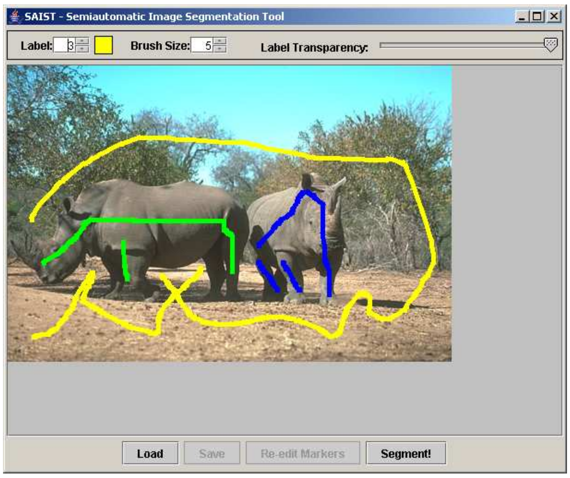
There are a few tools available to aid in image annotation. The Freiburg University Annotation Tool for assigning keywords to images has the disadvantage that its output is in a non-standard format (not in XML format) and that it imposes some constraints on keyword grouping (into the three groups "Events", "Objects" and "Static Scene"). The MATLAB annotation software written in the PASCAL NoE[21] only allows rectangular regions to be selected and requires that the keywords are selected from a pull-down menu, which is not suitable for large vocabularies. The semi-automatic image segmentation tool (SAIST)[22] uses a marker-based watershed segmentation. The user draws in the markers, as shown in Figure 3a, which leads to the segmentation shown in Figure 3b. This process can be iterated by adding or removing markers (Figure 3c) until the required segmentation is obtained (Figure 3d).

3.3 Overview of Visual Keywords

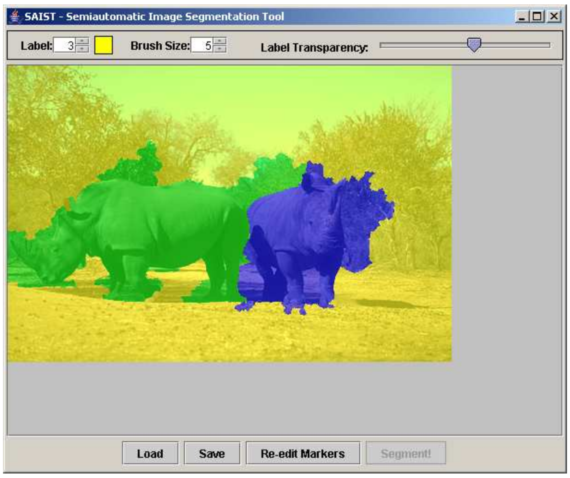
We present a collection of groups of keywords which have already been used for testing automated image annotation algorithms or in automated image and video annotation evaluation campaigns.

---

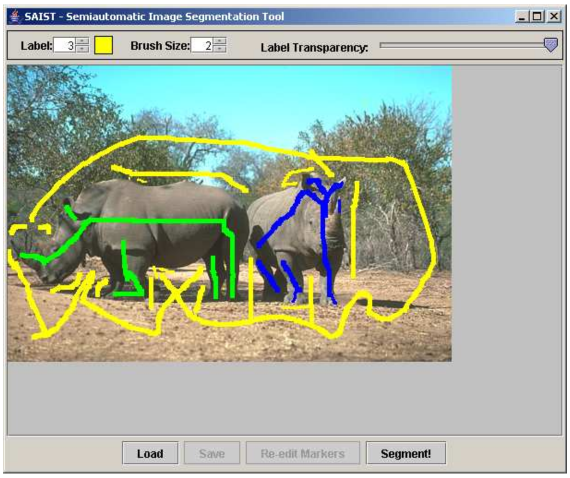
[15] `http://www.espgame.org`
[16] `http://www.peekaboom.org`
[17] `http://www.captcha.net/esp-search.html`
[18] `http://people.csail.mit.edu/brussell/research/LabelMe/intro.html`
[19] 400 keywords on the 29th of July 2005.
[20] 27 July 2005
[21] Downloadable from `http://www.pascal-network.org/challenges/VOC/`
[22] `http://muscle.prip.tuwien.ac.at`



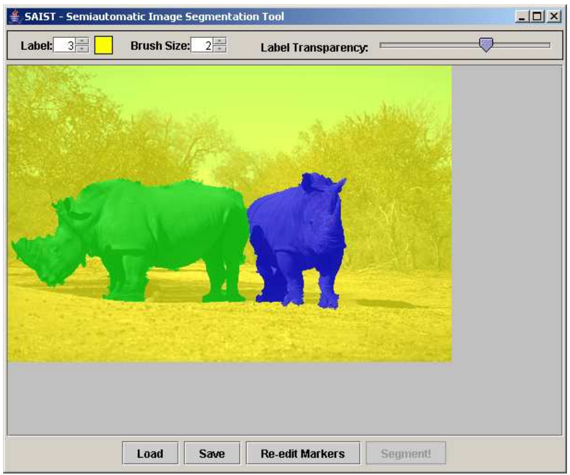
(a)



(b)



(c)



(d)

Figure 3: Use of SAIST. (a) Initial markers. (b) Segmentation resulting from the markers in (a). (c) Additional markers. (d) Segmentation resulting from the markers in (c).

The 10 features which were tested in the TRECVID 2005 high-level feature detection task are described in Table 1. All 40 news concepts defined for TRECVID 2005 are available for download[23] (they are part of the LSCOM creation task [13]).

Two categorisation tasks are part of the ImagEVAL[24] campaign: for the general image description task, the hierarchically organised global image categories shown in Figure 4 will be tested. There is also an object detection task, although the list of objects to be tested has not been finalised yet. The examples given are car, tree, chair, Eiffel Tower and American Flag.

The PASCAL Visual Object Classes Challenge 2005 consisted of classification and detection tasks for four objects: motorbikes, bicycles, people and cars. However, in the database collection

[23] `http://www-nlpir.nist.gov/projects/tv2005/LSCOMlite_NKKCSOH.pdf`
[24] `http://www.imageval.org`

Keywords	Segment contains video of ...
People walking/running	more than one person walking or running
Explosion or fire	an explosion or fire
Map	a map
US flag	a US flag
Building exterior	the exterior of a building
Waterscape/waterfront	a waterscape or waterfront
Mountain	a mountain or mountain range with slope(s) visible
Prisoner	a captive person, e.g., imprisoned, behind bars, in jail, in handcuffs, etc.
Sports	any sport in action
Car	an automobile

Table 1: The 10 features which were tested in the TRECVID 2005 high-level feature detection task.

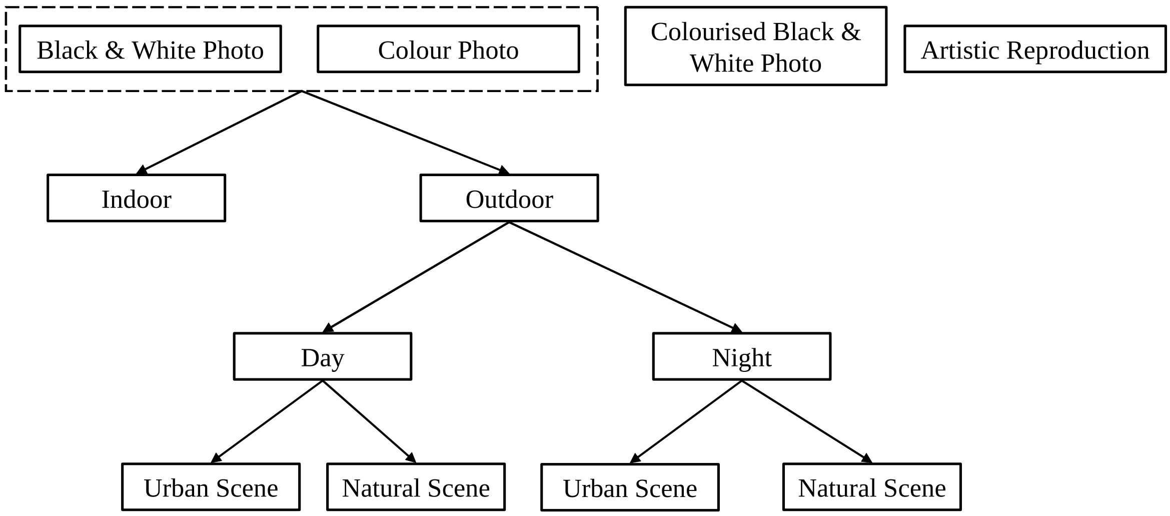


Figure 4: The hierarchy of keywords used in the global image characteristics task of ImagEVAL.

set up as part of this challenge[25], five databases are provided with standardised ground truth object annotations. The keyword list arising from this standardisation is shown in Table 2.

As part of the EU LAVA project[26], a database consisting of 10 categories of images was made available[27]. These categories are: bikes, boats, books, cars, chairs, flowers, phones, roadsigns, shoes and soft toys.

[25] `http://www.pascal-network.org/challenges/VOC/`
[26] `http://www.l-a-v-a.org`
[27] `ftp://ftp.xrce.xerox.com/pub/ftp-ipc/`

aeroplaneSide	apple	background	bicycle	bicycleSide
bookshelf	bookshelfFrontal	bookshelfPart	bookshelfSide	bookshelfWhole
bottle	building	buildingPart	buildingRegion	buildingWhole
can	car	carFrontal	carPart	carRear
carSide	cd	chair	chairPart	chairWhole
coffeemachine	coffeemachinePart	coffeemachineWhole	cog	cow
cowSide	cpu	desk	deskFrontal	deskPark
deskPart	deskWhole	donotenterSign	door	doorFrontal
doorSide	face	filecabinet	firehydrant	freezer
frontalWindow	head	keyboard	keyboardPart	keyboardRotated
light	motorbike	motorbikeSide	mouse	mousepad
mug	onewaySign	paperCup	parkingMeter	person
personSitting	personStanding	personWalking	poster	posterClutter
pot	printer	projector	roadRegion	screen
screenFrontal	screenPart	screenWhole	shelves	sink
sky	skyRegion	sofa	sofaPart	sofaWhole
speaker	steps	stopSign	street	streetSign
streetlight	tableLamp	telephone	torso	trafficlight
trafficlightSide	trash	trashWhole	tree	treePart
treeRegion	treeWhole	walksideRegion	wallClock	watercooler
window				

Table 2: The keywords in the PASCAL Object Recognition Database Collection (the prefix “PAS” has been removed from each keyword).

Chen and Wang [5] classified images into 20 categories: African people and villages, Beach, Historical buildings, Buses, Dinosaurs, Elephants, Flowers, Horses, Mountains and glaciers, Food, Dogs, Lizards, Fashion, Sunsets, Cars, Waterfalls, Antiques, Battle ships, Skiing and Deserts.

Two databases have been released by Microsoft Research in Cambridge[28]. The “Database of thousands of weakly labelled, high-res images” contains images divided into the following 23 categories: aeroplanes, cows, sheep, benches and chairs, bicycles, birds, buildings, cars, chimneys, clouds, doors, flowers, forks, knives, spoons, leaves, countryside scenes, office scenes, urban scenes, signs, trees, windows, miscellaneous. 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 following 23 labels: building, grass, tree, cow, horse, sheep, sky, mountain, aeroplane, water, face, car, bicycle, flower, sign, bird, book, chair, road, cat, dog, body, boat. The majority of the images are roughly segmented, although accurate segmentations of some of the images are available.

[28]Downloadable here: `http://www.research.microsoft.com/vision/cambridge/recognition/default.htm`. Version 1 of the pixel-wise labelled image database has been ignored here, as it forms a subset of version 2.

Faces	Faces easy	Leopards	Motorbikes	accordion	airplanes
anchor	ant	barrel	bass	beaver	binocular
bonsai	brain	brontosaurus	buddha	butterfly	camera
cannon	car side	ceiling fan	cellphone	chair	chandelier
cougar body	cougar face	crab	crayfish	crocodile	crocodile head
cup	dalmatian	dollar bill	dolphin	dragonfly	electric guitar
elephant	emu	euphonium	ewer	ferry	flamingo
flamingo head	garfield	gerenuk	gramophone	grand piano	hawksbill
headphone	hedgehog	helicopter	ibis	inline skate	joshua tree
kangaroo	ketch	lamp	laptop	llama	lobster
lotus	mandolin	mayfly	menorah	metronome	minaret
nautilus	octopus	okapi	pagoda	panda	pigeon
pizza	platypus	pyramid	revolver	rhino	rooster
saxophone	schooner	scissors	scorpion	seahorse	snoopy
soccer ball	stapler	starfish	stegosaurus	stop sign	strawberry
sunflower	tick	trilobite	umbrella	watch	water lilly
wheelchair	wildcat	windsor chair	wrench	yin yang	

Table 3: The 101 categories used by Fei-Fei et al. [9].

It is, of course, possible to greatly extend the number of categories if one is recognising specific objects, such as in the Caltech 101 category database[29] [9], which contains images of objects in the categories shown in Table 3.

If one restricts oneself to such specific categories, it is obviously possible to create many thousands. A set of 16 broader categories has been defined for the 15 200 images in the CEA-CLIC database [21]. These are shown in Table 4.

A number of papers on automatic image or image region annotation have also been published. The following three all use parts of the Corel image database along with keywords usually extracted from the annotations accompanying the Corel images. The 55 keywords used by Carbonnetto et al. [4] are given in Table 5. Li and Wang [19] used the largest number of keywords. They defined 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", the images in the "Lion" category were assigned the keywords "lion, animal, wildlife, grass" and the images in the "eagle" category were assigned the keywords "wildlife, eagle, sky, bird". The 433 keywords used by Li and Wang [19] are shown in Table 8 in Appendix A. The 323 keywords used by Barnard et al. [2] are shown in Table 9 in Appendix A.

[29] `http://www.vision.caltech.edu/Image_Datasets/Caltech101/Caltech101.html`

Category	Description
Food	Images of food, and meals.
Architecture	Images of architecture, architectural details, castles, churches, Asian temples.
Arts	Paintings, sculptures, stained glass, engravings.
Botanic	Various plants, trees, flowers.
Linguistic	Images containing text areas.
Mathematics	Fractals.
Music	Images of musical instruments.
Objects	Images representing everyday objects such as coins, scissors, etc.
Nature & Landscapes	Landscapes, valley, hills, deserts, etc.
Society	Images with people.
Sports & Games	Stadiums, items from games and sports.
Symbols	Iconic symbols, roadsigns, national flags (real and synthetic images)
Technical	Images involving transportation, robotics, computer science.
Textures	Rock, sky, grass, wall, sand, etc.
City	Buildings, roads, streets, etc.
Zoology	Images of animals (mammals, reptiles, bird, fish).

Table 4: The 16 categories in the CEA-CLIC image database and their descriptions [21].

airplane	astronaut	atm	bear	beluga	bill	bird
boat	building	cheetah	church	cloud	coin	coral
cow	crab	dolphin	earth	elephant	fish	flag
flowers	fox	goat	grass	ground	hand	horse
house	lion	log	map	mountain	mountains	person
pilot	polarbear	rabbit	road	rock	sand	sheep
shuttle	sky	snow	space	tiger	tracks	train
trees	trunk	water	whale	wolf	zebra	

Table 5: The 55 keywords used by Carbonetto et al. [4].

3.4 Analysis of Visual Keywords

The aim of this analysis is to create a list of keywords which reflect the current interest in automated image annotation with keywords. These keywords could then serve as an initial controlled vocabulary for re-annotating the image collections used in previous experiments and for annotating new image collections. The use of a keyword list generated in this way has the following advantages:

- As the keywords represent a fusion of those from many experiments, the generated list is challenging for automated annotation systems.
- It is certain that the keywords in the new list are applicable to the many thousands of existing images used for automated image annotation research. As many of the existing images are poorly annotated, it would make sense to re-annotate them.

3.4.1 Creation of a combined keyword list

The first step of the analysis consisted of creating a list combining all the keywords and categories used in the experiments, datasets and evaluations covered in Section 3.3. We then removed words which were considered to be unsuitable. These include place names, such as “Australia”, “Boston” and “New Zealand”, which, even for a human, are very difficult to assign to images for which one has no supplementary information. Confusing keywords, such as “history” and “north”, and keywords requiring too high a level of a priori semantic information, such as “landmark” and “rare animal” were also removed. We have not yet collected statistics on how often a single keyword appears in different lists.

3.4.2 Categorisation of keywords

From a practical point of view, it is useful if the keywords are sorted into categories. When one is annotating images, this simplifies the choice of a word from the keyword list — one can select the category that the image belongs to in order to reduce the choice of keywords. The 16 categories of the CEA-CLIC database [21], with some minor changes, turn out to be well-suited to grouping the combined list of keywords.The changes are:

- the fusion of the “Architecture” and “City” categories to form an “Architecture / City” category. This was done as it is often difficult for an annotator to decide between these two categories.
- the addition of an “Abstract / Global” category to contains words such as “female” and “exterior”.
- the removal of the “Mathematics” category, which has no members in the list of keywords collected.
- the removal of the “linguistic” category, as this is an image category and not a keyword category.

#	Category	Description
0	Abstract / Global	Words which describe the whole image or which are applicable to more than one class of objects.
1	Food	Food and meals.
2	Architecture / City	Architecture, architectural details, castles, churches, Asian temples, buildings, roads, streets, etc.
3	Arts	Paintings, sculptures, stained glass, engravings.
4	Botanic	Plants, trees, flowers.
5	Objects	Everyday objects such as coins, scissors, etc.
6	Nature & Landscapes	Landscapes, valley, hills, deserts, etc.
7	Society	People, groups of people, activities undertaken by society (celebrations, parades, war, etc.).
8	Sports & Games	Stadiums, items from games and sports.
9	Symbols	Iconic symbols, roadsigns, national flags
10	Technical	Transportation, robotics, computer science.
11	Textures	Words which describe a texture.
12	Zoology	Animals (mammals, reptiles, birds, fish).
13	Anatomy and Medicine	Biological organs, anatomical diagrams, etc.
14	Music	Musical instruments.

Table 6: The 15 categories of the combined keyword list and their descriptions. The first column contains a category number.

- the addition of the "Anatomy and Medicine" category, which at present includes one keyword, but can be expanded later.

The list of categories and their descriptions are given in Table 6.

We assigned each of the keywords in the combined list to at least one category. A few keywords were assigned to two categories, for example, "grass" appears in the "Texture" and "Nature and Landscapes" categories. A table showing the keywords assigned to each category is given in Appendix B. A histogram of the number of keywords per category is shown in Figure 5.

One can see from this histogram that the categories "Objects", "Nature and Landscapes" and "Zoology" contain the most keywords, which could be an indicator that these categories have received the most attention in past research on automated image annotation and categorisation. This could be because of the image databases used — the Corel databases, for example, appear to contain a high proportion of natural and animal images. The man-made objects appear to be more prevalent in the databases designed for object categorisation experiments.

Lower level keywords can be extracted from the PASCAL Object Recognition Database Collection keywords. These are words such as "Side" and "Rear" that can be added to most of the keywords to give more detail about which part of an object is visible (e.g. Cow - side). There are two types of such keywords: *view* and *action* keywords, which are shown in Table 7.

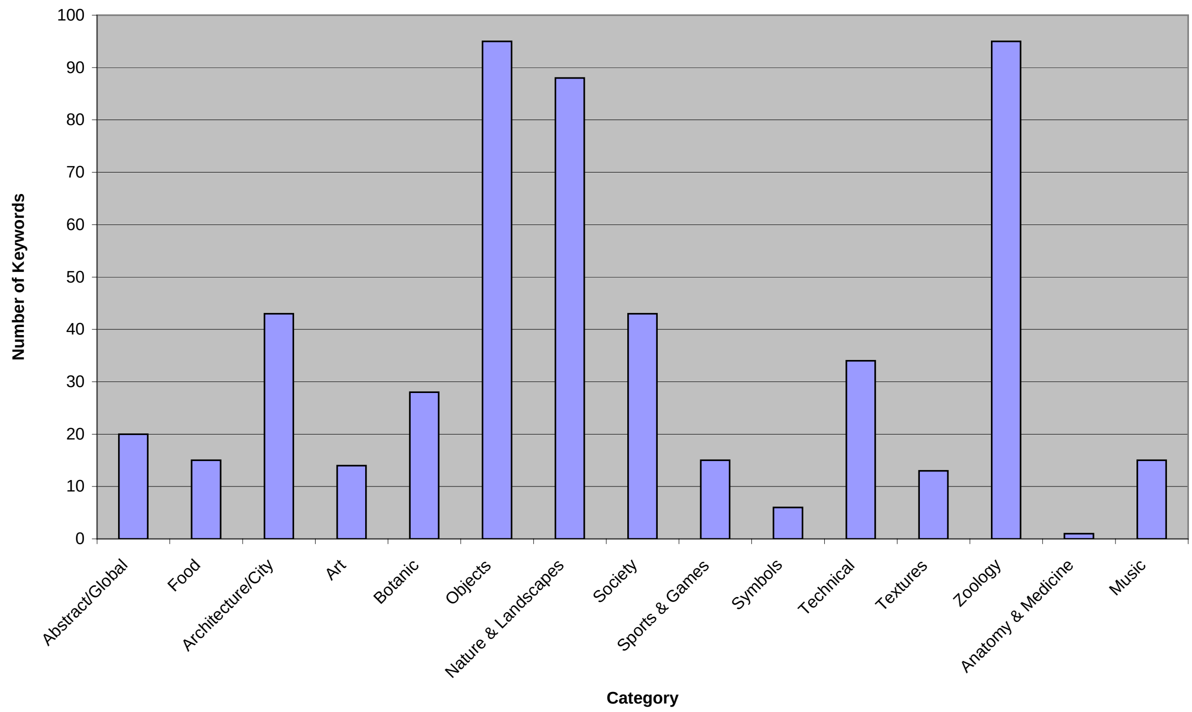


Figure 5: The number of keywords in each category.

View Keywords				
side rear	front rotated	part clutter	whole	region

Action Keywords				
sitting	standing	walking		

Table 7: The view and action keywords from the PASCAL Object Recognition Database Collection.

4 Conclusion

We give an overview of three different types of image annotation: free text annotation, keyword annotation and annotation using ontologies. We then analyse the keywords which have been used to annotate images in a number of image retrieval publications and evaluation campaigns. A significant contribution is the creation of a keyword list based on these keywords, where the keywords are divided into 15 categories.

From this analysis one can see that the main automated annotation effort has been directed at images of everyday objects; nature and landscapes; and animals (zoology). As these keywords were extracted from annotations of existing image datasets, they should be well-suited to a more precise re-annotation of these same datasets. For the same reason, they are also suited to verify the applicability of newly developed image ontologies intended to represent portrayable entities and objects.

A disadvantage is that while the keywords in this list certainly correspond well to the images used in image annotation experiments so far, there is no guarantee that these images are representative of all possible electronic images. It would therefore be useful to compare this collection of keywords to an ontology constructed in a more rigorous way, such as the ontology of portrayable objects based on WordNet [30]. This should provide a useful link between possible portrayable objects and those that are often found in images, or that are of interest to image understanding researchers.

A Comprehensive Keyword Lists

The following papers on automatic image annotation used keyword lists of a few hundred keywords. They are shown in this appendix due to their length.

A.1 The Li and Wang Keywords

The Li and Wang [19] keywords are available for download in a format showing the keywords assigned to each of the 60 categories (i.e. keywords are repeated) [30].

abstract	Africa	agate	agriculture
Alaska	ancestor	animal	antelope
antique	architecture	Arizona	art
Asia	Asian	Australia	autumn
aviation	Bali	ballet	balloon
barbecue	barnyard	bath	battle
beach	bead	Belgium	Berlin
Bhutan	bike ads	bird	black and white
blue	boat	bonsai	Boston
botany	Brazil	British Columbia	builder
building	bus	business	butterfly
cactus	California	camel	Canada
candy	canyon	car	card
Caribean	carve	castle	cat
cave	child	China	Christmas
church	city	close	cloth
cloud	coastal	college	color
Colorado	communication	compete	Costarica
cougar	couple	coyote	craft
Croatia	cruise	crystal	cuisine
cyber	Czech Republic	dawn	death valley
decoration	decoy	desert	design
dessert	Devon	dining	dinosaur
dish	dog	dogsled	doll
door	drawing	drink	dusk
eagle	earth	Easter egg	Egypt
elephant	engine	England	environment
estate	Europe	everglade	exploration
fabric	face	Far East	farm
fashion	fauna	feast	female
festival	fight	Finland	fire

[30] from `http://wang.ist.psu.edu/docs/related/`

firearm	firework	fish	fitness
flag	flora	Florida	flower
flowerbed	foliage	food	forest
fountain	fowl	fox	fractal
France	front door	frost	fruit
fun	Galapago	game	garden
gem	glacier	glamour	goat
golf	graffiti	Grand canyon	grape
grass	Greece	green	group
guard	Guatemala	gun	hairstyle
Hanover	harbor	Hawaii	hawk
herb spice	highway	historical building	history
holiday	Holland	home	Hong Kong
horse	house	ice	ice_frost
image	India	Indonesia	indoor
industry	insect	interior	Ireland
isle	Italy	item	Jamaica
Japan	jewelry	Kenya	kitchen
Korea	kungfu	Kyoto	lake
landmark	landscape	leaf	leisure
life	light	lighthouse	lion
lizard	location	London	machine
male	mammal	man	man-made
marble	maritime	market	mask
medicine	Mesoamerica	Mexico	micro image
Middle East	mineral	modern	molecule
Monaco	Montreal	monument	mosaic
moth	motorcycle	mountain	mural
museum	mushroom	music ads	Namibia
nation	natural	nature	nautical
nest	New Guinea	New Mexico	New York
New Zealand	night	no fear	north
Nova Scotia	occupation	ocean	ocean animal
office	old	orange	orbit
orchid	Oregon	Ottawa	owl
painting	palace	parade	paradise
Paris	park	pastoral	pathology
pattern	penguin	people	perenial
Peru	pet	Philadelphia	photo
pill	pioneer	plane	planet
plant	play	polo	pomp and pageantry
Portugal	poster	power	Prague
predator	primate	produce	public sign

Pyramid	Quebec	R Beny	race
rafting	rail	rare animal	recreation
red	reflect	relic	religion
reptile	river	Riviera	road
road sign	rock	rock form	rockies
rodeo	Rome	rose	royal
royal guard	ruin	rural	rural England
rural France	Russia	sacred	sail
Samer	San Diego	San Francisco	scene
science	Scotland	sculpture	sea
season	seed	shape	shell
shimmer	ship	show	shuttle
Silkroad	Singapore	ski	skin
sky	skyline	snow	South Pacific
space	Spain	speed	sport
stamp	star	steam	still life
Stmoritz	studio	sub sea	success
summer	sun	sunset	supermodel
surf	surf side	SW US	Swiss
tallship	technology	textile	texture
Thailand	thing	things	tiger
tissue	tool	Toronto	toy
train	transportation	travel	tree
tribal	tropical	Tulip	Turkey
turtle	up	US	Utah
valley	vegetable	Vietnam	vineyard
Virginia	volcano	Wales	war
Washington	Washington DC	water	waterfall
wave	way	west	wet
wild	wildcat	wildlife	wind
wind surf	winter	woman	women
work	works	world	worship
yellow	Yellowstone	Yemen	Yosemite
young animal	youth	yuletide	Zimbabwe
Zion			

Table 8: The 433 keywords used by Li and Wang [19].

A.2 The Barnard et al. Keywords

The Barnard et al. [2] are available for download, along with other data used in the paper[31].

anemone	angelfish	animal	animals	antlers
arch	arches	architecture	arctic	art
background	baby	bay	beach	bear
bears	beetle	bengal	bighorn	bills
bird	birds	black	boat	boats
bobcat	bottles	branch	branches	bridge
building	buildings	bull	bulls	bush
bushes	butterfly	cactus	candy	canoe
canyon	car	caribou	cars	carvings
castle	cat	caterpillar	chairs	cheetah
church	city	cliff	close-up	closeup
clouds	coast	columns	coral	costume
costumes	cougar	courtyard	coyote	crop
crystal	crystals	cubs	currency	dall
deer	desert	design	designs	detail
display	diver	dock	dog	door
doors	doorway	dress	dunes	eagle
elephant	elephants	elk	entrance	f-16
f-18	face	fan	farm	feline
fence	field	fish	flag	flags
flight	floor	flower	flowers	foal
foals	food	forest	formation	formula
fox	frost	frozen	fruit	fungus
furniture	garden	gardens	giraffe	glass
goats	grapes	grass	grizzly	ground
guard	gun	guns	harbor	hat
hats	hawk	hawks	head	helicopter
herd	hills	hillside	hippo	hippos
horizon	horns	horse	horses	hotel
house	houses	hunter	hut	ice
iceburg	iguana	indian	insect	island
jaguar	jet	kauai	kayak	kitten
lake	landscape	leaf	leaves	leopard
lichen	light	lights	lion	lizard
locomotive	log	lynx	man	mane
mare	market	meadow	military	model
money	mosque	moss	mountain	mountains

[31] from `http://vision.cs.arizona.edu/kobus/research/data/jmlr_2003`

museum	mushroom	mushrooms	nest	night
ocean	orchid	outside	owl	paintings
palace	palm	paper	parade	park
path	pattern	patterns	peaks	penguin
people	perch	petals	pillar	pillars
plain	plane	plants	polar	prototype
pumpkin	pumpkins	pyramid	rabbit	race
railroad	rapids	reef	reefs	reflection
relief	reptile	restaurant	rhino	river
road	rock	rocks	rodent	roofs
rose	ruins	runway	saguaro	sail
sailboats	sails	sand	scotland	sculpture
sea	seals	shadow	shadows	sheep
ship	ships	shop	shops	shore
shrine	sign	signs	ski	skis
sky	skyline	slope	smoke	snake
snow	sponge	sponges	squirrel	stairs
statue	statues	stem	stems	stone
stones	street	sun	sunset	tables
tail	temple	textile	texture	tiger
tower	town	tracks	train	tree
trees	trunk	tulip	tulips	tundra
turn	valley	vegetable	vegetables	vegetation
vehicle	vehicles	village	vineyard	wall
walls	water	waterfall	wave	waves
white-tailed	wildlife	window	windows	wine
wings	wolf	woman	wood	woodland
woods	zebra			

Table 9: The 323 keywords used by Barnard et al. [2].

B Combined Keyword List

The following table lists the combined keyword list. It is a simple two-level hierarchy, with 15 headings at the top level (in bold). Note that some words are repeated under more than one heading.

Abstract / Global				
background	black	black_and_white	blue	color
exterior	female	fractal	green	group
indoor	interior	male	nature	orange
outdoor	pattern	red	shadow	yellow

Food				
apple	cuisine	dessert	drink	feast
food	fruit	grapes	herb_spice	orange
pizza	pumpkin	strawberry	vegetable	wine

Architecture / City				
arch	architecture	building	castle	chimney
church	city	college	column	courtyard
dock	fountain	harbor	historical_building	hotel
house	hut	industry	kitchen	market
minaret	monument	mosque	museum	office
pagoda	palace	park	pillar	restaurant
roof	ruin	shop	skyline	stairs
statue	street	studio	temple	tower
town	village	window		

Art Objects				
art	carving	decoration	design	drawing
graffiti	mosaic	mural	painting	photo
poster	sculpture	statue	still_life	

Botanic

apple	bonsai	botany	branch	bush
cactus	flower	foliage	fungus	grapes
leaf	lichen	log	moss	mushroom
orchid	palm	perenial	petal	plant
pumpkin	rose	seed	strawberry	sunflower
tree	tulip	water_lily		

Objects (man-made everyday)				
anchor	antique	atm	balloon	barbecue
barrel	bath	bead	bench	bicycle
binoculars	book	bookshelf	bottle	camera
can	candy	card	cd	cellphone
chair	clock	cloth	coffee_machine	cog
coin	cup	currency	decoration	desk
dish	dogsled	doll	door	dress
Easter_egg	fabric	fan	fence	file_cabinet
fire_hydrant	firearm	firework	flag	floor
freezer	furniture	glass	gun	hat
headphones	horn	jewelry	keyboard	lamp
light	map	marble	mask	medicine
money	mousepad	mug	paper	paper_cup
parking_meter	pill	pot	printer	projector
relic	scissors	screen	shelves	shoe
sink	sofa	speaker	sponge	stamp
stapler	table	telephone	textile	tool
toy	traffic_light	trash	umbrella	wall
watch	watercooler	wheelchair	wood	wrench

Nature and Landscapes				
agriculture	autumn	barnyard	bay	beach
canyon	cave	cliff	cloud	coast
coral	crop	crystal	dawn	desert
dune	dusk	earth	farm	field
flowerbed	forest	frost	frozen	garden
gem	glacier	grass	ground	hill
ice	iceberg	island	lake	landscape
maritime	meadow	mountain	night	ocean
pastoral	path	peak	plain	planet

polar river rural shrine spring summer tropical volcano wind	pyramid road sail sky star sun tundra wall winter	rapids rock sand smoke steam sunset valley water woodland	reef ruin shell snow stone surf vegetation waterfall	reflection runway shore space sub_sea tree vineyard wave

Society				
astronaut builder couple fight head man pilot science work	baby business diver glamour holiday model pomp_and_pageantry travel worship	ballet child face graffiti home occupation religion tribal youth	barbecue Christmas fashion guard hunter parade royal war	battle costume festival hand leisure person sacred woman

Sports and Games				
fitness play rodeo	football polo ski	game race sport	golf rafting tennis	kungfu recreation wind_surfer

Symbols				
public_sign sign_yield	road_sign	sign_do_not_enter	sign_stop	sign_oneway

Technical				
aeroplane bridge communication	aviation bus engine	balloon cannon ferry	battle_ship canoe helicopter	boat car highway

jet molecule runway tallship	lighthouse motorcycle sailboat train	locomotive pathology ship transportation	machine railroad space_shuttle vehicle	military road street

Textures				
fabric ice textile	fire marble texture	glass sand wood	grass skin	ground stone

Zoology				
anemone	angelfish	animal	ant	antelope
antlers	bear	beaver	beetle	bird
bobcat	bull	butterfly	camel	caribou
cat	caterpillar	cheetah	coral	cougar
cow	coyote	crab	crayfish	crocodile
cub	deer	dinosaur	dog	dolphin
dragonfly	eagle	elephant	elk	feline
fish	flamingo	foal	fowl	fox
giraffe	goat	hawk	hedgehog	herd
hippopotamus	horn	horse	iguana	insect
jaguar	kangaroo	kitten	leopard	lion
lizard	llama	lobster	lynx	mammal
moth	mouse	nest	ocean_animal	octopus
owl	panda	penguin	pet	pigeon
polar_bear	predator	primate	rabbit	reptile
rhinoceros	rodent	rooster	scorpion	seahorse
seal	sheep	skin	snake	sponge
squirrel	starfish	tiger	turtle	whale
wildcat	wildlife	wolf	young_animal	zebra

Anatomy and Medicine				
brain				

Musical Instruments				
accordion	cello	double_bass	electric_guitar	guitar
horn	mandolin	piano	piano_grand	saxophone
trombone	trumpet	tuba	viola	violin

References

[1] Luis von Ahn and Laura Dabbish. Labeling images with a computer game. In *Proc. ACM CHI*, pages 319–326, 2004.

[2] Kobus Barnard, Pinar Duygulu, Nando de Freitas, David Forsyth, David Blei, and Michael I. Jordan. Matching words and pictures. *Journal of Machine Learning Research*, 3:1107–1135, 2003.

[3] Kobus Barnard, Quanfu Fan, Ranjini Swaminathan, Anthony Hoogs, Roderic Collins, Pascale Rondot, and John Kaufhold. Evaluation of localized semantics: Data, methodology, and experiments. Technical Report TR-05-08, Computing Science, University of Arizona, 2005.

[4] Peter Carbonetto, Nando de Freitas, and Kobus Barnard. A statistical model for general contextual object recognition. In *Proceedings of the ECCV 2004, Part I*, pages 350–362, 2004.

[5] Yixin Chen and James Z. Wang. Image categorization by learning and reasoning with regions. *Journal of Machine Learning Research*, 5:913–939, 2004.

[6] Gabriella Csurka, Christopher R. Dance, Lixin Fan, Jutta Willamowski, and Cedric Bray. Visual categorization with bags of keypoints. In *Workshop on Statistical Learning in Computer Vision (at ECCV)*, 2004.

[7] Alberto del Bimbo. *Visual Information Retrieval*. Morgan Kaufmann Publishers, Inc., 1999.

[8] Andreas Doms and Michael Schroeder. Gopubmed: Exploring pubmed with the geneontology. *Nucleic Acids Research*, 33, 2005.

[9] L. Fei-Fei, R. Fergus, and 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*, June 2004.

[10] Evgeniy Gabrilovich and Shaul Markovitch. Feature generation for text categorization using world knowledge. In *Proceedings of The Nineteenth International Joint Conference for Artificial Intelligence*, Edinburgh, Scotland, 2005.

[11] Thomas R. Gruber. A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5(2):199–220, 1993.

[12] Michael Grubinger, Clement Leung, and Paul 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*, pages 17–23, Vienna, Austria, September 2005.

[13] Alexander G. Hauptmann. Towards a large scale concept ontology for broadcast video. In *Proceedings of the Third Intl. Conf on Image and Video Retrieval*, pages 674–675, 2004.

[14] Eero Hyvönen, Avril Styrman, and Samppa Saarela. Ontology-based image retrieval. In *Proceedings of XML Finland Conference*, pages 51–27, 2002.

[15] Jack Kustanowitz and Ben Shneiderman. Motivating annotation for digital photographs: Lowering barriers while raising incentives. Technical Report ISR 2005-55, ISR, University of Maryland, 2004.

[16] Andrea Kutics, Akihiko Nakagawa, Shoji Arai, Hiroyuki Tanaka, and Sakuichi Ohtsuka. Relating words and image segments on multiple layers for effective browsing and retrieval. In *Proceedings of the International Conference on Image Processing*, pages 2203–2206, 2004.

[17] Yves Lafon and Bert Bos. Describing and retrieving photos using RDF and HTTP. W3C Note, http://www.w3.org/TR/photo-rdf/, April 2002. Last accessed: 15 May 2005.

[18] Clement H. C. Leung and Horace Ho-Shing Ip. Benchmarking for content-based visual information search. In *Proceedings of the 4th International Conference on Advances in Visual Information Systems*, pages 442–456, 2000.

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

[20] George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller. Introduction to wordnet: An on-line lexical database. *Inetrnational Journal of Lexicography*, 3(4):235–244, 1990.

[21] Pierre-Alain Moëllic, Patrick Hède, Gregory Grefenstette, and Christophe Millet. Evaluating content based image retrieval techniques with the one million images clic testbed. In *Proceedings of the Second World Enformatika Congress, WEC'05*, pages 171–174, 2005.

[22] C. Peters, P. Clough, J. Gonzalo, G.J.F. Jones, M. Kluck, and B. Magnini, editors. *Multilingual Information Access for Text, Speech and Images*, volume 3491 of *LNCS*. Springer, 2004.

[23] A. Th. (Guus) Schreiber, Barbara Dubbeldam, Jan Wielemaker, and Bob Wielinga. Ontology-based photo annotation. *IEEE Intelligent Systems*, 16(3):66–74, 2001.

[24] 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*, 22(12):1349–1380, December 2000.

[25] J. R. Smith and B. Lugeon. Visual annotation tool for multimedia content description. In *Proc. SPIE Vol. 4210, p. 49-59, Internet Multimedia Management Systems*, pages 49–59, 2000.

[26] M. Szummer and R. W. Picard. Indoor-outdoor image classification. In *Proc. IEEE International Workshop on Content-based Access of Image and Video Databases*, pages 42–51, 1998.

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

[28] James Z. Wang, Jia Li, and Gio Wiederhold. SIMPLIcity: Semantics-sensitive integrated matching for picture libraries. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(9):947–963, 2001.

[29] J. Winn, A. Criminisi, and T. Minka. Object categorization by learned universal visual dictionary. In *Proceedings of the International Conference on Computer Vision(ICCV)*, 2005.

[30] S. Zinger, C. Millet, B. Mathieu, G. Grefenstette, P. Hède, and P.-A. Moëllic. Extracting an ontology of portrayable objects from WordNet. In *Proceedings of the MUSCLE/ImageCLEF Workshop on Image and Video Retrieval Evaluation*, pages 17–23, Vienna, Austria, September 2005.