

Chapter 4

Data Design and Engineering

The last chapter has reported on the lack of evaluation for VIR from generic photographic collections (*i.e.* containing everyday real-world photographs akin to those that can frequently be found in private photographic collections as well, *e.g.* holiday pictures or photos of sporting events). In particular, the following issues were raised:

- copyright restrictions hinder the redistribution of image collections and limit the reproducibility of results for other researchers;
- the search topics are often not representative of real-world information needs and are often either too difficult or too easy for the state-of-the-art image retrieval methods;
- most of the collections are static and cannot be easily adapted to different requirements or changes as far as evaluation goals are concerned;
- there are no evaluation efforts that are concerned with the retrieval from generic photographic collections.

This chapter is concerned with the first of these four issues and reports on the creation of a parametric image collection which was specifically designed and implemented to fill that particular gap, with the subsequent Chapters 5, 6 and 7 taking on the other three aspects. Most of this chapter is taken from [146, 147].

4.1 Introduction

This section provides an overview of the remaining sections of this chapter and explains the main motivation as well as the specifications and requirements for the design and development of a representative document collection (*i.e.* images and their semantic descriptions) for the evaluation of retrieval from generic photographic collections.

4.1.1 Motivation and Overview

A core component of any TREC-style benchmark is a set of documents (*e.g.* texts, images, or videos) that is representative of a particular domain [267]. Finding such resources for general use, however, is often difficult, not least because of copyright issues which restrict the distribution and future accessibility of data. This is especially true for visual resources that are, in general, often more valuable than written texts and hence subject to limited availability and access for the research community.

Although there are some image collections that have recently been acquired for VIR evaluation purposes (see Section 3.2), there is still a lack of more general image collections in order to cater for the growing research interest in information access to generic photographic collections. Such real-life collections are inherently multilingual (especially online photo collections such as *FlickrR*, see Section 3.2.7), thus logical image representations in multiple languages are essential for TBIR.

This chapter describes such a generic photographic collection which we specifically designed and implemented with the following aims in mind:

- to provide a realistic, real-life collection of generic photographs suitable for a wider range of evaluation purposes;
- to provide images with associated written information representing typical textual meta-data that can be used to explore the semantic gap;
- to provide semantic image descriptions in multiple languages;

- to provide a data set that is free of charge and copyright restrictions and therefore available to the general research community.

First, Section 4.1.2 discusses the requirements for such a benchmark document collection together with its specification. Section 4.2 then introduces our methodology for creating the image collection and describes how we carried out this creation process, together with sample images, statistics and distribution details. Section 4.3 later on describes and analyses the multilingual semantic descriptions of the images and introduces our methodological framework to ensure consistency within the descriptions. Section 4.4 finally summarises in general and points out the benefits of the image collection presented in this chapter in particular.

4.1.2 Requirements and Specification

This work is based on a set of recommendations for an image retrieval benchmark [234] that followed an initiative by *Technical Committee 12 (TC-12)* of the *International Association of Pattern Recognition (IAPR)*, see Section 4.2.1). Apart from suggestions for relevance judgments and performance measures, these recommendations included aspects such as the general evaluation scope, benchmark parameters, the size of the collection, image characteristics and whether the speed of retrieval should be evaluated or not.

Evaluation Scope

While many benchmarks in other areas of computing are preoccupied with speed and response time (*e.g.* TPC, SPEC), these measures should not play a central role in our benchmark. Following the TREC methodology, the main focus of this research concentrates on the evaluation of an algorithm's ability to identify relevant images from a generic photographic collection, rather than the algorithm's ability to carry out efficient search (in terms of retrieval speed).

Benchmark Parameters

The benchmark collection should be *parametric*; it should allow the specification of a number of parameters that may be adjusted according to different requirements.

Only by this means can the benchmark suite be geared to supply a variety of different needs and be adapted to changing evaluation goals: reasons for such changes can be due to more powerful retrieval systems or to changing interests in the research community (expressed by, for example, participants' feedback at an evaluation event).

Size of the Benchmark Database

The size of the database should not be too small in order to be useful for a benchmark: results would not be representative if the database is too small, and in order to be considered as a “database”, the collection also has to consist of a significant number of items. The initial benchmark should therefore consist of at least $N = 1,000$ images and should be scaled up to larger orders of magnitude in subsequent developments [234].

Having too large a database, on the other hand, would be impractical too; retrieval by image contents will require some degree of indexing and, as a consequence, the costs of indexing the database could be immoderate because the proper indexing of an such an image can require up to 10 minutes per image [232]. Thus, the indexation of collections with more than $N = 20,000$ images is felt to be not feasible unless one has access to unlimited manpower and resources [234].

Image Characteristics

The benchmark database should contain images that exhibit the following characteristics:

1. *Content Range*. The image collection should contain a variety of images that are representative of generic photographic collections (*i.e.* photos of holidays, events, family, *etc.*).

2. *Multi-object images.* The general character of the images should be object-based photographs (with the majority of the images comprising multiple objects).
3. *Relationships.* There should be diverse relationships between the objects (*e.g.* actions).
4. *Attributes.* Objects and/or relationships should include a variety of qualifying attributes.
5. *JPEG files.* The image format should be JPEG because it is a widely used format and easy to handle in terms of storage requirements.
6. *Royalty free.* The image collection should be free of charge and without copyright restrictions that would hinder the redistribution of the collection for large-scale evaluation events.

Retrieval Speed

Retrieval speed and response time should not play a central role in this benchmark as it targets the evaluation of retrieval precision rather than retrieval speed. Although retrieval speed is generally considered as an essential factor for the usability of a system and also tests the algorithm's ability to carry out efficient search, it is also often dependent on extraneous factors such as network connection, disk bandwidth or processor speed. Since usability evaluation is already carried out in separate evaluation events such as [136, 137], it is not covered within the scope of this research.

4.2 Collection Creation

This section reports on the creation of an image collection representative of generic photographic collections and suitable for a wider range of evaluation purposes. This includes the evolutionary development of the image collection (Section 4.2.1), the origin of the images (Section 4.2.2) and how the relevant images were chosen according

to specific image selection rules (Section 4.2.3). Then, a few representative sample images of some chosen categories are exemplified and the visual diversity of the collection is illustrated (Section 4.2.4). Finally, statistical data on a range of attributes that characterise the image collection are provided (Section 4.2.5) and the distribution of the collection (or rather the access to the collection) is explained (Section 4.2.6).

4.2.1 Collection Development

Figure 4.1 provides an evolutionary overview of the stages this project had to undergo.

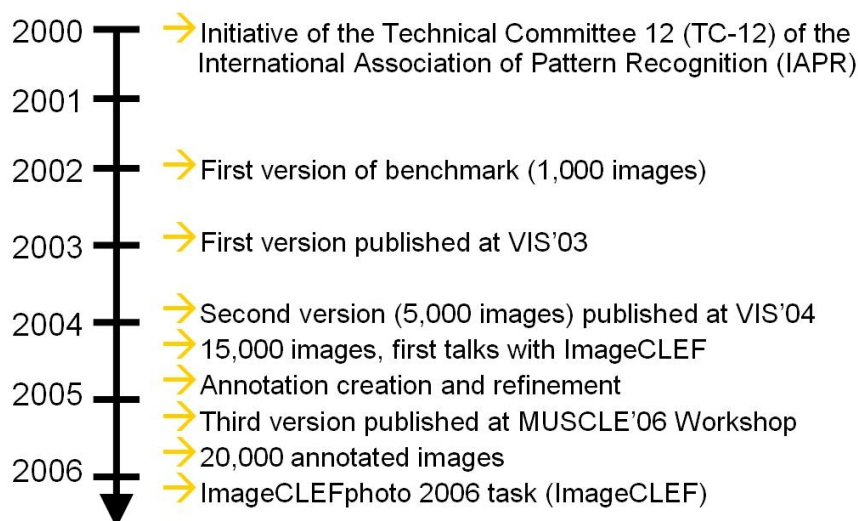


Figure 4.1: Collection development overview.

In 2000, *TC-12* of the *IAPR* recognised the need for a standard benchmark for multimedia retrieval and began an effort to create a freely available database of images with associated alphanumeric representation files. This started by developing a set of recommendations and specifications for an image benchmark suite [234].

Based on the criteria in the specifications, we set up a first version of a benchmark consisting of 1,000 multi-object colour images, 25 search requests (queries) and a collection of performance measures in 2002 [144] and published it in 2003 [147]. We called this benchmark the *IAPR TC-12 Benchmark* as it had grown out of the aforementioned initiative by TC-12 of the IAPR, and it was further promoted

as such on the web page of the IAPR¹ as well as in subsequent publications such as [329].

The initial size of the benchmark seemed rather small, and since the development of a benchmark is universally considered to be an incremental and ongoing process, we refined and improved the *IAPR TC-12 Benchmark* and soon (2004) extended the number of images to 5,000 using a custom-built benchmark administration system to support the incremental benchmark development [148]. At the end of that year, an independent travel organisation (see Section 4.2.2) provided access to around 10,000 of their images including multilingual semantic representations of varying quality in several languages (English, German, Spanish, Portuguese). This increased the total number of images in the benchmark to 15,000.

Of course, a benchmark is not beneficial unless it is actually used by the research community. Therefore in 2005, we began discussions with *ImageCLEF* about a possible involvement of the *IAPR TC-12 Benchmark* as part of an ad-hoc retrieval task at *ImageCLEF* (see Chapter 7), and I was appointed as an *ImageCLEF organising committee member* for its general ad-hoc retrieval tasks soon afterwards.

With 10,000 additional images from the travel organisation, the total number of available images had soon risen to 25,000 images [150], but we subsequently reduced it to 20,000 images annotated in three languages to maintain the high quality of the collection.

Finally, we began to use the *IAPR TC-12 Benchmark* for *ImageCLEF's* photographic ad-hoc retrieval task in 2006 (called *ImageCLEFphoto*) and will continue using it for future tasks [60].

4.2.2 Image Origin

The image collection originates from two different sources: *viventura* and myself.

The majority of the images have been provided by *viventura*², an independent travel company organising adventure and language trips to South America.

¹www.iapr.org/committees/techco.php

²<http://www.viventura.de/>

Travel guides accompanying the tourists maintain a daily online diary including photographs of the trips made and general pictures of each location. Furthermore, the guides provide general photographs of tourist attractions, accommodation facilities and ongoing social projects.

The minor but by no means less significant part of the images consists of photos taken by myself. These photos comprise pictures of holidays and (mainly sporting) events and do not only add to the diversity of the image collection, but were also included to enhance and complement the photos provided by *viventura*. As an example, I specifically selected sports photos and added them to the collection in order to increase the percentage of action photos in the benchmark and to create a more balanced spectrum of relationship levels between objects (which is one of the image selection criteria specified in the next section).

4.2.3 Image Selection

Not all images provided by *viventura* were suitable for a benchmark; as a consequence, images had to undergo a selection process before they were added to the collection. In order to ensure the quality of the benchmark and to achieve a representative cross-section of real-world photographic images in generic collections, we set up several image selection rules according to which we carefully selected the benchmark images. Some of the selection criteria include the quality of the images, the number and level of “repetitiveness” of the objects in the image, and the number of relationships between these objects.

Image Quality

The first and most obvious rule in order to guarantee a high level of quality for the benchmark database was to only select high quality images. The following criteria should therefore apply to any image candidate:

- *High resolution.* Images had to exhibit at least the minimum resolution of 480 x 360 pixels to be selected for the benchmark database. Photos captured or scanned using a lower resolution were not relevant for selection.

- *Clarity.* Blurry photos due to camera movement or any other reason at the time of capturing the image were not included (although these images might frequently occur when capturing actions). Only clear images were selected.
- *Contrast.* Solely photos with a good level of contrast were selected for the benchmark. Photos that were taken in poor photographic positions and were therefore too dark due to underexposure (or that were too bright due to overexposure respectively) were not included. However, if taken properly, night shots and photos of sunsets were considered for selection.

One might argue that not all the images in real-world collections are of high quality. While this is certainly true, it is easily possible to automatically lower the quality of images in a collection (*e.g.* by decreasing the resolution, blurring the photo, *etc.*) if this is required for a specific evaluation - the converse is not possible.

Number of Objects

The second rule for image selection is concerned with the number of objects in an image³. Let \mathcal{O} denote the set of objects and $N_{\mathcal{O}} = |\mathcal{O}|$ the number of objects in an image. Ideally, the image collection would exhibit a spectrum ranging from images with only one object to images with an almost uncountable number of objects.



(a) $N_{\mathcal{O}} = 2$.

(b) $N_{\mathcal{O}} = 9$.

(c) $N_{\mathcal{O}} > 100$.

Figure 4.2: Number of objects.

Figure 4.2 illustrates examples of images with a diverse number of objects: Figure 4.2(a) shows an image with only two objects (bird, rock), Figure 4.2(b) an image

³Only the main objects in the foreground are considered, as it is quite impossible to count all the objects in the background.

with a few objects (seven birds and a small fence around a tree), and Figure 4.2(c) image with an uncountable number of objects (birds, rocks).

Therefore in the selection process, we made sure that the images in the database would not repeatedly comprise an identical number of objects to guarantee a fair distribution of object cardinality in the collection. In doing so, the image collection can subsequently cater for a broad range of evaluations, ranging from object recognition tasks (that, at this stage, rather involve images with a low number of objects, especially as regards training data) to retrieval evaluation of complex image contents (requiring a higher number of objects). To further add to this diversity, objects should also include a variety of qualifying attributes such as colour, size, shape, or condition.

Repetitiveness of Objects

The images in the benchmark collection should not just show a variation in the number of objects, but also provide an even distribution of “repetitiveness” between these objects. Yet, in order to be able to observe such object repetitiveness during the image selection process, it is necessary to establish a measure in order to quantify the level of repetition of objects for an entire image.

There are certainly multiple ways to arrive at such quantification; in its simplest form, the repetition could be expressed by the number of repeated objects in an image. Such an absolute value is, however, not very representative and unfeasible for subsequent collection statistics. We therefore define the object repetitiveness ρ in an image as the relation of the actual entropy H of the objects in the image divided by the maximum entropy H_{max} possible in such an image with $N_{\mathcal{O}}$ objects:

$$\rho := 1 - \frac{H}{H_{max}} \quad (4.1)$$

The definition of ρ in (4.1) implies that $0 \leq \rho \leq 1$. The actual entropy H of an image with $N_{\mathcal{O}}$ objects of N_T different object types amounts to

$$H = - \sum_{i=1}^{N_T} p_i \log_2 p_i \quad (4.2)$$

with $N_T \leq N_{\mathcal{O}}$ and p_i denoting the likelihood of randomly selecting the i^{th} object type in the image,

$$p_i = \frac{N_{\mathcal{O}}(i)}{N_{\mathcal{O}}} \quad (4.3)$$

with $N_{\mathcal{O}}(i)$ denoting the number of objects of type i , whereas the maximum entropy in an image is achieved when all the objects are unique: $N_T = N_{\mathcal{O}}$ and $p_i = 1/N_{\mathcal{O}}$, and therefore

$$H_{max} = - \sum_{i=1}^{N_{\mathcal{O}}} \frac{1}{N_{\mathcal{O}}} \log_2 \frac{1}{N_{\mathcal{O}}} = - \log_2 \frac{1}{N_{\mathcal{O}}} \quad (4.4)$$

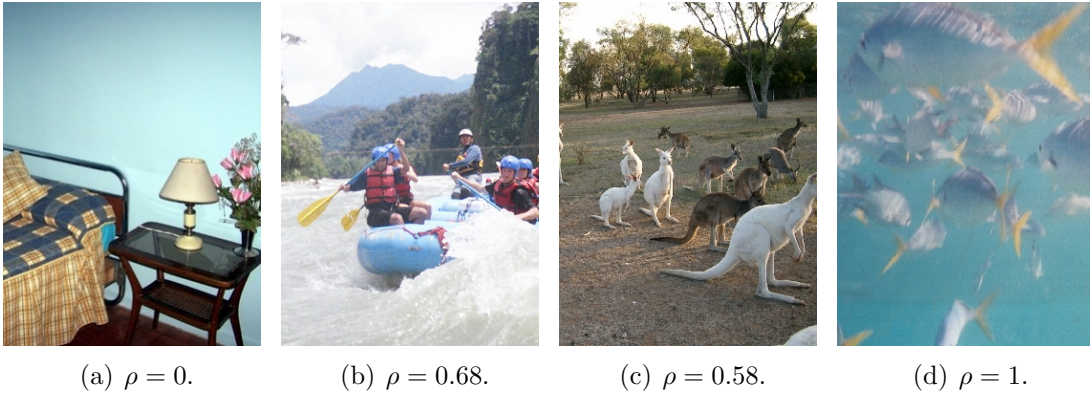


Figure 4.3: Level of repetitiveness.

To achieve a good cross-section of various levels of object repetitiveness, we subsequently selected images like in Figure 4.3(a) in which all the objects (bed, pillow, blanket, table, lamp, flowers, vase, *etc.*) were of unique types, images like in Figure 4.3(b) in which some objects (paddles, helmets, life-vests, *etc.*) appeared more often, others like in Figure 4.3(c) in which some objects (kangaroos) appeared more often but with different attributes (brown, white), and images like in Figure 4.3(d) that contained only one type of object (fish) repeated several times.

Number of Relationships

The last image selection rule observes the number of binary⁴ relationships (N_R) in an image. As generic photographic collections would contain images without any

⁴Only binary relationships, *i.e.* relationships between two objects, are considered. Any relationship involving three or more objects can always be broken down to several binary relationships between these objects.

relationships between their images (*e.g.* static landscape shots) as well as photos depicting a certain level of interaction between their objects (like, *e.g.*, in many sport images), we would also expect the images in the benchmark collection to cover a cross-section of the number of relationships in the following range:

$$0 \leq N_R \leq N_{Rmax} \quad (4.5)$$

where N_{Rmax} is the maximum number of binary relationships,

$$N_{Rmax} = \binom{N_{\mathcal{O}}}{2} \quad (4.6)$$

and $N_{\mathcal{O}}$ denotes the number of objects in an image, with relationships between any two objects only being counted once.

Identifying and observing N_R in an image collection is crucial insofar as it allows for the subsequent generation of subsets to cater for the evaluation of retrieval from collections with complex image contents, *i.e.*, containing multiple objects (agents) and relationships (actions).

Hence, in the image selection process we considered images with absolutely isolated objects ($N_R = 0$) like the landscape shot in Figure 4.4(a), images with a few relationships ($0 < N_R < N_{Rmax}$) such as the one in Figure 4.4(b), and images like in Figures 4.4(c) and 4.4(d) in which all objects are somehow related to each other ($N_R = N_{Rmax}$).

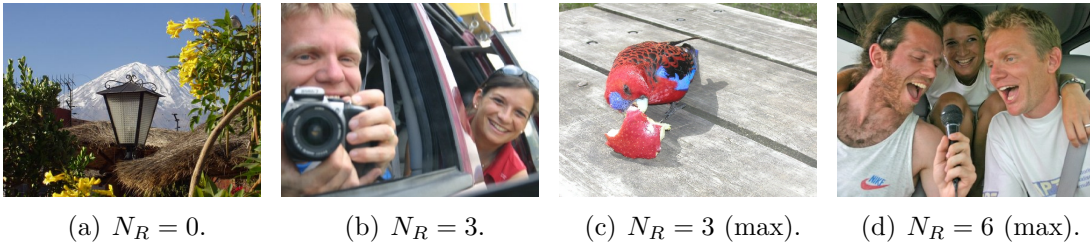


Figure 4.4: Number of relationships.

For example, the image in Figure 4.4(b) consists of four ($N_{\mathcal{O}} = 4$) objects (man, woman, camera, car) and three ($N_R = 3$) relationships (man *is sitting* in car, man *is holding* camera, woman *is sitting* in car), but there is, for instance, no relation between the camera and the woman.

Images with only three objects ($N_{\mathcal{O}} = 3$) often show the maximum number of relations $N_{Rmax} = 3$: in Figure 4.4(c), for instance, all three objects *table*, *bird* and *apple* are related to each other: *bird is sitting on table*, *apple is lying on table*, *bird is eating apple*.

In rare cases, images with four objects ($N_{\mathcal{O}} = 4$) also exhibit the maximum number of relations $N_{Rmax} = 6$, like the image in Figure 4.4(d): all the four objects *man-left*, *man-right*, *woman* and *microphone* have direct relations with each other: *man-left is holding and singing into microphone*, *man-right is singing into microphone*, *woman is singing into microphone*, *woman is hugging man-left*, *woman is hugging man-right*, *man-left is looking at man-right*).

Images with five or more objects are, in general, unlikely to reach N_{Rmax} relationships.

4.2.4 Image Examples

This section provides sample images to illustrate the nature, range and diversity of the images within the *IAPR TC-12 Benchmark*.

Selected Image Categories

The image collection includes pictures of a range of sports and actions, photos of people, animals, cities, landscapes and many other aspects of contemporary life (see Figures 4.5 to 4.10).

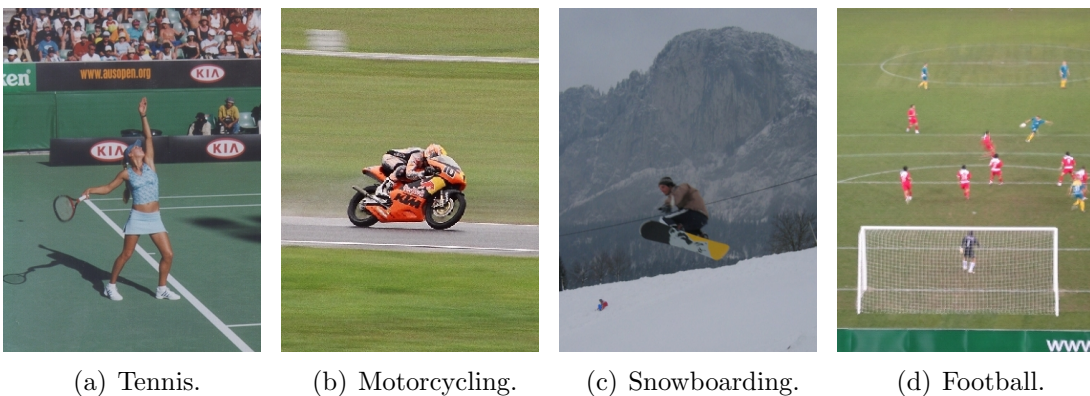


Figure 4.5: Examples of sports photos.



(a) Pushing.

(b) Celebrating.

(c) Drinking.

Figure 4.6: Examples of action pictures.



(a) Sydney.

(b) Paris.

(c) Las Vegas.

Figure 4.7: Examples of city pictures.



(a) Peruvian kids.

(b) Korean guards.

(c) Russian singers.

Figure 4.8: Examples of people shots.



(a) Whale.

(b) Kangaroos.

(c) Tortoise.

(d) Boobies.

Figure 4.9: Examples of animal photos.

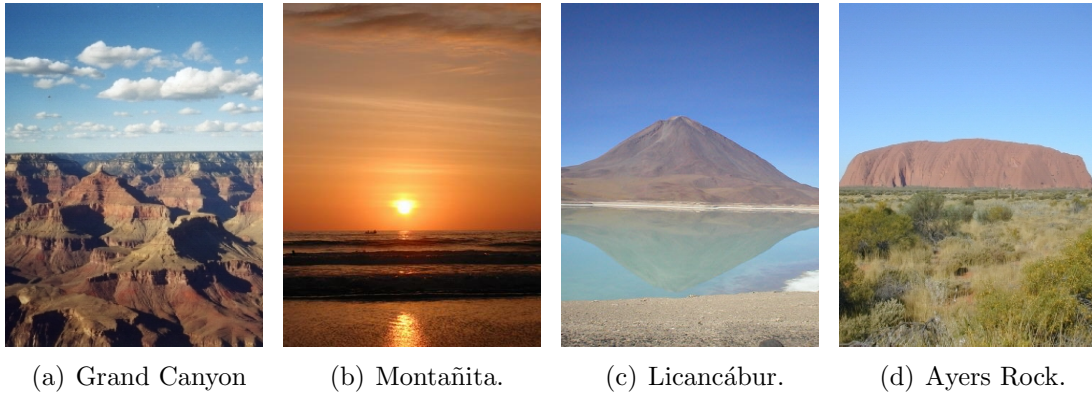


Figure 4.10: Examples of landscape photos.

Collection Diversity

The *IAPR TC-12 photographic collection* contains many images of similar visual content, but varying illumination, viewing angle and background. This is because most of the tours offered by the travel organisation are repeated on a regular basis and have fixed itineraries. Thus, the tours always visit the same tourist destinations where the guides usually take photos of tourists in varying poses (Fig. 4.11) and/or tourist attractions with varying viewing angles (Fig. 4.12), weather conditions (Fig. 4.13) or at different times of the day (Fig. 4.14). These varying characteristics make the benchmark also well-suited for content-based retrieval tasks as it allows a range of prototypical searches to explore retrieval effectiveness with these varying settings.



Figure 4.11: Tourists in varying poses at the Salt Lake of Uyuni in Bolivia.



(a) Right.

(b) Left.

(c) Front.

Figure 4.12: The Cathedral of Cusco, Peru, in different viewing angles.



(a) Bright sunshine.

(b) Overcast and cloudy.

(c) Foggy and rainy.

Figure 4.13: The Inca ruins of Machu Picchu in varying weather conditions.



(a) At night.

(b) In the morning.

(c) During the day.

Figure 4.14: A cyclist riding a racing bike at different times of the day.

4.2.5 Image Collection Statistics

This section provides information on a range of attributes which characterise the image collection, *e.g.* the size and colour of images, image formats, and the temporal and geographic extent of the collection.

Image Sizes

The photographs in the collection exhibit the following differences based on the technology used to capture the images: photographs taken with digital cameras have a 4:3 relation of width to height (96x72 pixels for the thumbnails, and 480x360 pixels for the larger versions), and photos taken with a non-digital (or traditional) camera which have been subsequently scanned have a 3:2 relation of width to height (92x64 pixels for the thumbnails, and 480x320 pixels for the larger versions).

The choice of these specific dimensions is solely based on the fact that *viventura* provided their images in 480x360 pixels, which were included in the benchmark without any further processing to guarantee the highest resolution possible for the collection. Images not provided by *viventura* were then re-sampled to the same dimensions (or to a dimension with the same length for the scanned photos respectively).

As for the file sizes, thumbnails require between 2 and 10 KB each (with an average size of 5.69 KB); the larger versions range from 20 to 200 KB (with an average size of 85.25 KB), depending upon their content and colour composition. The total size of the image collection is 1.66 GB (and 111 MB for the corresponding thumbnails). All images are stored in the JPEG format.

Colour Variation

The majority of the images (99.63%) in the *IAPR TC-12 Benchmark* are colour images of high quality (see also Figure 4.15). This is because the photographs in the collection have mainly been taken with digital cameras in recent years (see also Section 4.2.5). However, the collection also contains a few images in black-and-white (0.37%).

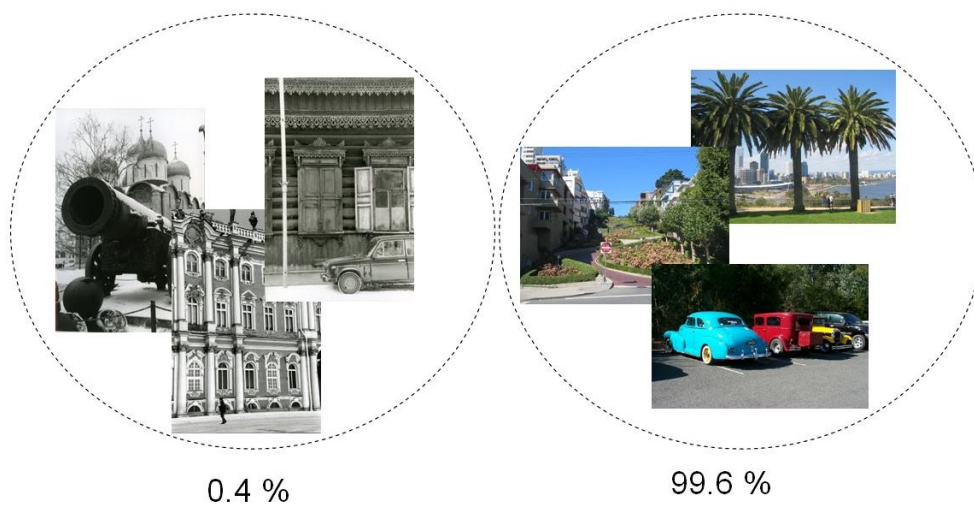


Figure 4.15: Colour variation.

Figure 4.15 shows sample images for both colour and black-and-white images and their distribution in percent.

Temporal Range

Figure 4.16 shows the temporal distribution of images in the collection between 2001 and 2005. Most photographs have been taken since 2001; the earliest photo in the collection dates back to 2000, and the most recent one was taken in July 2005. The mean date is June 2003, the standard deviation is 1.12 years and the median is January 2004.

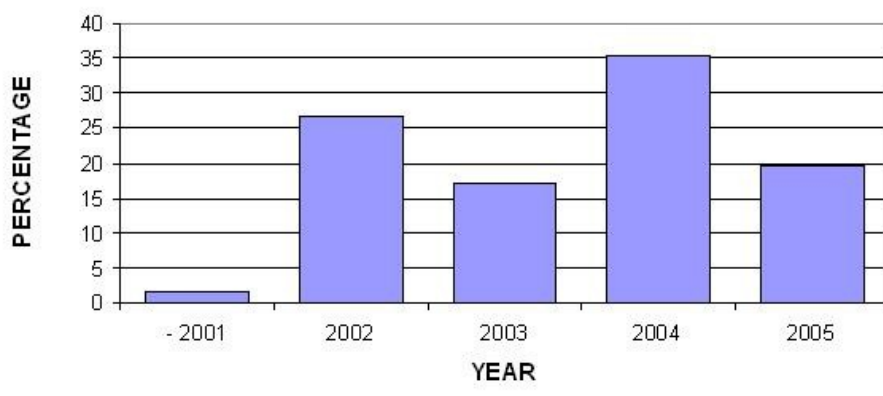


Figure 4.16: Temporal range.

Geographical Range

The image collection of the *IAPR TC-12 Benchmark* is spatially diverse, with pictures taken in more than 30 countries worldwide, including Argentina, Australia, Austria, Bolivia, Brazil, Chile, Colombia, Ecuador, France, Germany, Greece, Guyana, Korea, Peru, Russia, Spain, Switzerland, Taiwan, Trinidad and Tobago, Uruguay, USA and Venezuela.

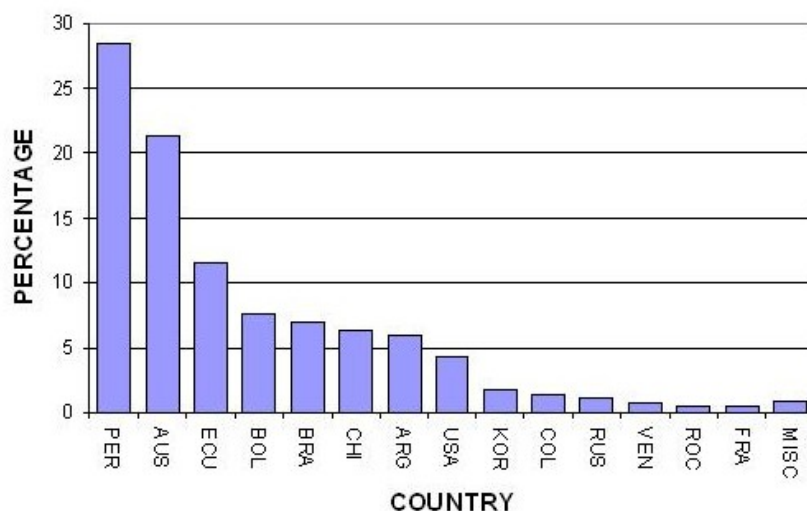


Figure 4.17: Geographical range.

Figure 4.17 shows the proportion of the images taken in these countries⁵. Most of the images originate from Peru (28.4%), Australia (21.3%) and Ecuador (11.6%), reflecting the geographic location and contributors. The collection comprises a total of 11 countries contributing more than 1% of the collection, and 14 countries with at least 100 images or 0.5% of the collection.

Number of Objects

Figure 4.2.5 provides an overview of the distribution of the number of objects within an image, showing a good cross-section of object cardinality within the collection.

About 30% of the images contain between one and six predominant objects in the

⁵The countries are represented by their international three letter code following the specification in ISO 3166-1 alpha-3.

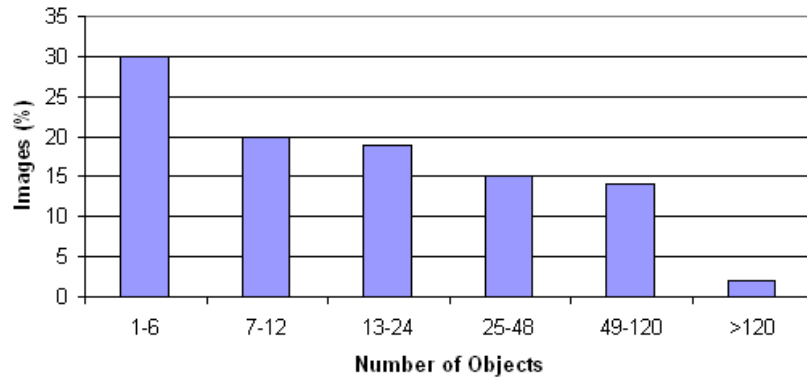


Figure 4.18: Number of objects.

foreground (and 7% of the entire collection only contain one object), which makes the collection well-suited for automatic image annotation or classification tasks for the state-of-the-art methods in object recognition. Roughly half of the collection contains images with up to a dozen objects, which are often in relationships to each other (see below); 19% of the images contain between one and two dozen, another 15% between two and four dozen and 14% between four and ten dozen of objects. Only rarely (2%) do images contain more than 120 objects in the foreground.

Repetitiveness of Objects

Figure 4.19 shows a very even distribution of object repetitiveness within the images of the collection. While most of the objects only occur once in 21% of the images

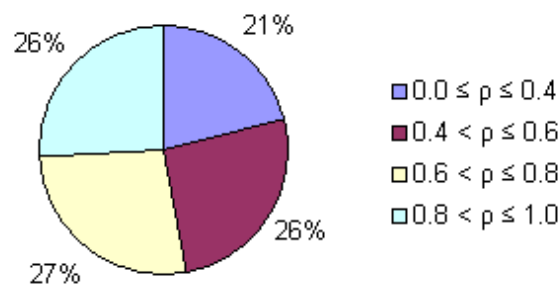


Figure 4.19: Repetitiveness of objects (ρ).

(13% of the images only contain unique objects), predominant objects can also be found in a rather repetitive way in 26% of the images (with 11% of the images

containing only one type of objects but repeated many times). About half of the images in the collection show a medium level of object repetition, *i.e.* some of the objects in the images are repeated while others are unique.

Relationships

Figure 4.20 illustrates the distribution of binary relationships between the objects in the images of the data collection. About one quarter of the images in the collection

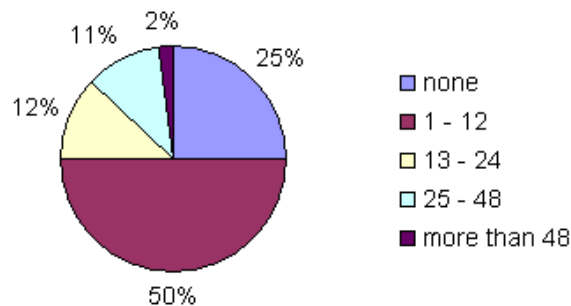


Figure 4.20: Number of relationships.

does not contain any relationships. Most of these static images are, for example, landscape shots or photos of facilities such as accommodation or language schools. Roughly 50% of the images contain up to a dozen relationships, with another 12% exhibiting between one and two dozen, and 11% between two and four dozen relationships. Only in rare cases (2% of the images) do images show more than four dozen relationships between their objects.

4.2.6 Distribution of the Collection

The entire image collection of the *IAPR TC-12 Benchmark* is available for download⁶ in one archive which is made available to *ImageCLEF* participants as well as to other researchers after having signed a provisional end-user agreement⁷. This archive consists of the large versions of the images, the corresponding thumbnails and logical (alphanumeric) representation files.

⁶http://eureka.vu.edu.au/~grubinger/IAPR/TC12_Benchmark.html

⁷See Appendix B for the complete agreement.

Images and Thumbnails

The benchmark archive contains 20,000 image files (jpg) which are stored in subdirectories of the `images` directory. The subdirectory `images/00/` contains all the images with a unique identifier between 0 and 999; `images/01/` subsequently contains the images with a unique identifier between 1000 and 1999, and so on.

The corresponding 20,000 thumbnail files (jpg) are stored in subdirectories of the `thumbnails` directory. The subdirectory `thumbnails/00/` contains all the thumbnails of the images with a unique identifier between 0 and 999; `thumbnails/01/` contains all the thumbnails of the images with a unique identifier between 1000 and 1999, and so on.

Multilingual Semantic Descriptions

Each image file is accompanied by at least one corresponding (text) file containing the semantic descriptions of that particular image. These text files are stored in subdirectories of the `annotations` directory. The subdirectory `annotations/00/` contains all the semantic descriptions of the images with a unique identifier between 0 and 999; `annotations/01/` contains all the semantic descriptions of the images with a unique identifier between 1000 and 1999, *etc.*

Extension	Text file with semantic description in...
.eng	English
.ger	German
.spa	Spanish
.fra	French (for future use)
.ita	Italian (for future use)
.por	Portuguese (for future use)
.rnd	a randomly selected language

Table 4.1: Text file extensions for semantic descriptions.

Table 4.1 provides an overview of the file extensions for the text files containing the semantic descriptions of the images. The language is thereby denoted by the file extension using the international three-letter representation for language names⁸.

⁸ISO 639-2 Alpha-3.

4.3 Image Semantics and their Specifications

In order to successfully evaluate algorithms that extract semantic information from images in a benchmark collection, a reliable ground-truth is necessary. Such ground-truth can be achieved by a semantically rich description of the objects, actions and background settings in an image. Semantic richness is obviously a benefit as it facilitates the categorisation of and the search for images and at the same time eases the query creation process for evaluation events.

Logical image representations, however, are not only an advantage for internal collection management purposes, they are still a vital part of image retrieval algorithms as well and therefore an indispensable component of an benchmark image collection. Currently, the state-of-the-art methods in CBIR (or automatic image annotation respectively) tend to operate on an extremely low level. We are a long way from bridging the semantic gap (compare Section 2.1.3) using CBIR approaches only, and semantic search requests can still only be successfully processed by the inclusion of text representations. A lot of current research therefore deals with CBIR combined with TBIR [60, 62, 118, 161]. It is expected that, as CBIR methods evolve and improve, the importance of text representations will decrease.

For the benchmark, as a consequence, the goal was to carefully and systematically create logical (alphanumeric) image representations that are as semantically rich as possible (within a feasible time-frame) to allow a reasonable evaluation for the current state-of-the-art methods of image retrieval from generic photographic collections. Although complete and consistent image representations are typically not found in practice, the goal of creating this resource was to provide a general-purpose collection which could subsequently be used for a variety of research purposes. For example, these alphanumeric representations could be offered to participants with a lower level of “completeness” to keep up the challenge for improved retrieval methods in the future (see also Chapter 7); it is easily possible to generate a less semantically rich subset for specific evaluation tasks if a benchmark is designed with enough flexibility (*i.e.* parametric benchmark architecture, see Chapter 6).

However, this process is obviously not possible vice versa (*i.e.* to automatically increase the semantic richness of the logical image representations). It is therefore felt that the semantic alphanumeric representations for image benchmarks should be created as precisely as possible.

This section describes the creation of multilingual image representations in English, German and Spanish for the *IAPR TC-12* image collection presented in Section 4.2. In particular, it explains where the original semantic information of the benchmark images came from (Section 4.3.1), introduces the methodology that we established in order to guarantee the consistency and high quality of the semantic information (Section 4.3.2), illustrates how these rules were carried out in the image annotation process itself (Section 4.3.3) and finally provides very comprehensive analyses and statistics of the completed representations (Section 4.3.4).

4.3.1 Original Image Captions

Most of the original captions in the document collection were created by *viventura*'s employees, and especially by their tourist guides who accompany the tourists and maintain a daily online diary to record the adventures and places visited by their customers (see also Section 4.2.2). The guides are also supposed to add a short description for each image they include with their diaries. These descriptions comprise a title for and a short description of the image as well as the location and the date of creation. Most image descriptions are thereby created in German as *viventura* targets the German-speaking market. However, in some cases, guides also use Spanish, Portuguese or English.



Figure 4.21: Original image caption.

Figure 4.21 shows an example image with a mixed-language original caption in Portuguese and German. The Portuguese title provides a brief statement about the image contents (in this case, the name of the beach: “Flamingo Beach”); the description of the image is in German and provides further detail (“Flamingo Beach is considered as one of the most beautiful beaches of Brazil!”). Both the location (“Salvador, Brazil”) and the date (“October 2nd, 2004”) are expressed in German language and format.

The quality of the German captions (and also their detail) varies tremendously, since most of the tour guides are local employees from South America and therefore native Spanish or Portuguese speakers. Hence, the captions have to undergo a revision and refinement process in order to provide a consistent set of semantic image representations for the benchmark collection.

4.3.2 Image Annotation Rules

A consistent set of semantic image representations, however, can only be achieved if a set of rules is established *before* the image annotation process is started. This should guarantee the best possible semantic description of images for the image collection, which is essential for image benchmarks: one can, if necessary, always decrease the quality of the semantic descriptions when the collection is used at subsequent evaluation events, but it is not possible vice versa. Thus, we established the following rules for the semantic description of the visual contents of the images.

Representation Format

The logical image representations should contain a semantic description of the image contents; in other words, they should describe in short sentences and noun phrases what can be recognised in an image without any prior information or knowledge. Keywords alone should not be used as they are not very precise due to the lack of syntax [437], and studies show that users, when unconstrained from a retrieval task, tend to create short narratives to describe images [198, 324].

Number of Sentences

Obviously, there is no limit to how semantically rich one could make the visual description of an image. However, in order to be able to complete the annotation process in a reasonable amount of time, the number of sentences to describe an image should not exceed six. The minimum number of sentences (or noun phrases) is one, as no image should be without an alphanumeric representation.

Sentence Order

The semantic descriptions of the image should follow a certain priority pattern: the first sentence (or the first sentences respectively) should describe the most obvious semantic information. The latter sentences should be used to describe the surroundings or settings of an image, like smaller objects or background information. For instance, the first sentence of the last example in Figure 4.22 describes

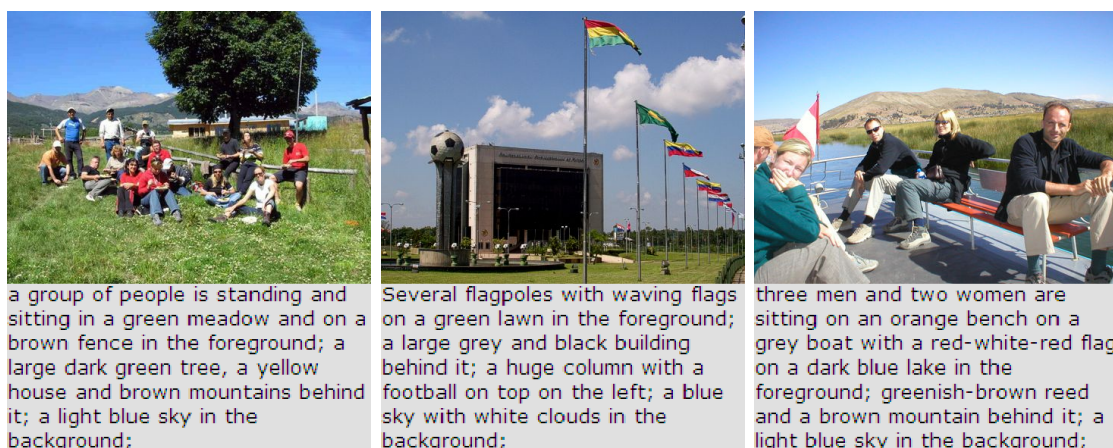


Figure 4.22: Sentence order examples.

the significant information of that image (*i.e.* that there are men and women sitting on a bench on a boat), while the latter sentences describe the settings and background information (reed, mountain, blue sky). This prioritisation of the representation sentences allows for the systematic reduction of representation quality in subsequent evaluation events. As retrieval methods improve, textual representations might become less influential for retrieval success, and having such systematic

representations provides the possibility for additional evaluation scenarios like retrieval from an image collection with gradually reduced background (or foreground) information.

Grammatical Number and Numerals

The representation sentences should describe the image as precisely as possible. This includes the specification of qualifying attributes for the objects as well as the exact use of their grammatical number (singular, plural). The following rules apply: *singular* must be used if there is only one occurrence of an object in an image (like “*a llama* is eating grass”); *plural* must be used if there are two or more occurrences of an object in an image (like “*llamas* are eating grass”)⁹. If the exact quantity of nouns can be determined, then a *numeral* should be used to describe that quantity (like “*eight llamas* are eating grass”). Figure 4.23 shows sample



eight brown llamas are grazing on a green meadow in the foreground; a valley with dark green trees, some water and brown mountains in the background;

(a) Eight llamas.



nine light blue canvas chairs and three sun shades in front of a brown wall at a light brown sandy beach in the foreground; dark green palms behind it; a blue sky in the background;

(b) Nine chairs.



a white alpaca is standing between two white llamas on a green meadow in the foreground; a bald slope with grey rocks in the background;

(c) One alpaca and two llamas.

Figure 4.23: Sample images for grammatical number and numerals.

images and their semantic descriptions in English to highlight the correct use of the grammatical number of objects. Consider Figure 4.23(b) more closely: while the number of chairs (nine) and the number of sun shades (three) can be determined, it is not possible to determine the exact numeral for the number of palm trees beyond doubt and thus, no numeral for the plurality of palm trees has been specified.

⁹Most of the currently spoken languages only use singular and plural. However, there are exceptions like Slovene that also use dual, trial or paucal grammatical numbers. These exceptions are not considered in the benchmark at this stage.

Terminology

The precise description of an image does not only involve the exact grammatical number but also implies the use of the appropriate terminology. This is due to the fact that people tend to use a specific language when they search for a specific image. The use of correct terminology involves two separate aspects: the appropriate



cricket players (one bowler, two batsmen, one wicket-keeper, two fielders) are playing on a green field with a brown pitch; there are two umpires on the left and spectators on a two-storey grandstand in the background;

(a) Cricket.



a female tennis player in a white and yellow dress is hitting a backhand on a green hard court; the net and a ball boy in the background;

(b) Tennis.



two golfers are standing on a putting green; one is putting, the other one is leaning on his putter and is watching; and a two-storey, light brown building with red flowers, green bushes and green hedges behind it;

(c) Golf.

Figure 4.24: Sample images for terminology.

description of objects and of actions. Figure 4.24(a) provides a good example of the use of the correct terminology in the sport of cricket: rather than just saying “cricket players”, the specific positions of the players are specified: “bowler, batsmen, wicket-keeper, fielders”. Figures 4.24(b) and 4.24(c) are good examples of the correct use of terminology of actions: “female tennis player is hitting a backhand” is more appropriate than just saying “person playing tennis”, and the same applies to “golfer putting on putting green” rather than to “people playing golf”.

The trend, again, is to make the logical image representations as precise as possible, because *hypernyms* can always be generated automatically, but this is not possible vice versa for the generation of *hyponyms*: if it is known that an image shows a “person putting on a green”, it can be concluded that that person is playing golf (*hypernym*). But if an image is annotated with “a person playing golf”, it is not possible to automatically determine whether this person is putting, driving the ball off the tee or playing out of a bunker (possible *hyponyms*).

No Interpretation

The semantic representation of images must be free from any interpretations; the annotator is only allowed to describe the visual content of the image and must not carry out the annotation process based on assumptions. For example, “a duck is



Figure 4.25: Sample images for interpretation rules.

swimming in a pond” for the first image in Figure 4.25 would present an inappropriate representation because the image just shows a bird swimming in water without any indication that this body of water is a pond. Any representation that is more specific than “water” (like a pond, a lake, a river, a billabong, *etc.*) as well as annotations on iconological level such as the expression of emotional or symbolic meanings would only be based on an assumption and should therefore be avoided. The same applies for the other two examples given in Figure 4.25.

Grammatical Tense

Any image that contains an action shows an incomplete action in progress at the time of capturing the image. This grammatically corresponds to the present tense with a continuous aspect to express that the action is incomplete and happening at that very moment.

Hence in English, the grammatically correct tense is the *present continuous tense* and should therefore be used for the English representations to describe the actions and situations in images. The auxiliary verb for English *to be* is optional in noun phrases of the logical image representations and can be omitted.

There is no continuous aspect in standard German; here, the *Präsens* (present tense) describes an action in progress and should therefore be used in the German image representations.

In Spanish, the continuous is constructed much as in English, using a regularly conjugated form of the verb *estar* together with the *gerundio* of the main verb. Thus, the construct *estar + gerundio* should be used for the description of the actions in the Spanish image representations. As with the English version, the auxiliary verb for Spanish (*estar*) is also optional for the Spanish noun phrases in the logical image representations and can be omitted.

Adjectives

As with the number of representation sentences, there is obviously no limit to how much detail each object could be described through the use of adjectives. In order to keep the annotation effort to a reasonable extent, we established the following general guideline: the fewer objects there are in an image, the more adjectives should be used to describe such an object (and vice versa).

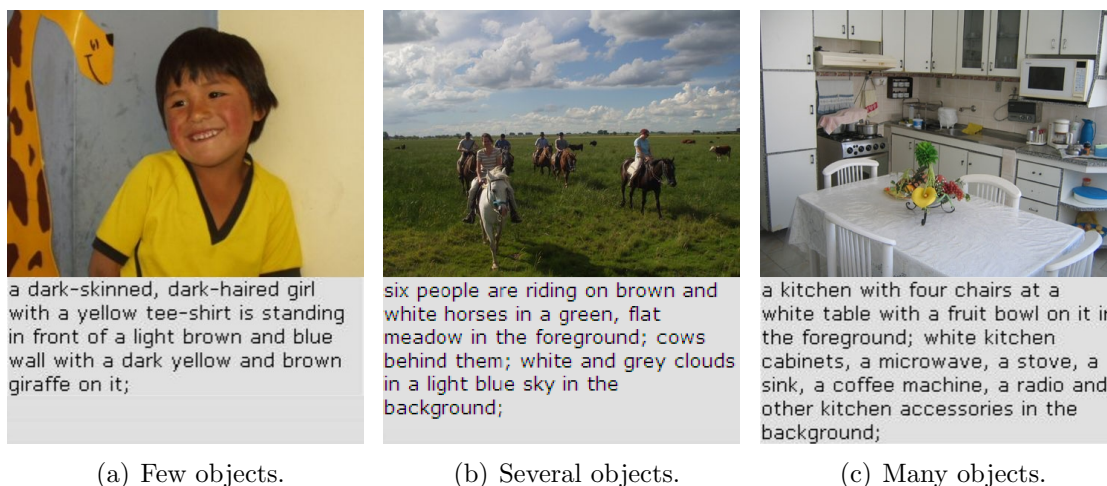


Figure 4.26: Sample images for adjectives.

This rule is also based on a log file analysis of a search interface used by the employees of *viventura* (see Section 5.4.1), which showed that people looking for only one object rather tend to use qualifying attributes with their search than people

looking for more objects or actions. Figure 4.26(a) illustrates a sample image with only a few objects (child, wall) but many adjectives describing that object. In contrast, Figure 4.26(c) shows an image with a large number of objects (chairs, table, cabinets, microwave, stove, sink, coffee machine, radio, *etc.*) that are not further specified by the use of attributes.

Colour Attributes

If the sentence pattern is not too complicated, then each of the annotated objects should receive at least one colour attribute. However, the use of colour attributes in the logical image representation is not as trivial as it might seem; the colour value of a pixel is usually stored using 24 bits in the RGB colour space, which means there are 16,777,216 possible colour values for each pixel. Although the perceptual ability of humans allows a much lower level of granularity for the visual differentiation of colour, there exists an immense number of colour names for ever so slightly different shades, saturations or intensities of colours (see *Coloria*¹⁰ for a very impressive list and representation of many colour names in several languages).

As a consequence, the more colour names are used in textual image representations, the smaller the difference between the colour names and therefore the harder it will be to provide a consistent use of colour attributes among all the representations. This is further complicated by the fact that one and the same colour can appear to be different in many images due to different surrounding colours.

A study [26] has shown that significant differences exist between naming colours in different languages and cultures. For example, a kind of *sea green* called “*aoi*” in Japanese is generally regarded as a *shade of green* in English, while what an English speaker would identify as *green* can be regarded as a different *shade of sea green* in Japanese. However, the study has also shown that there are substantial regularities in naming colours across many languages and that a concept of the following basic colour terms has been identified: black, grey, white, pink, red, orange, yellow, green, blue, purple and brown. All other colours can be considered to be variants of these

¹⁰<http://www.coloria.net/bonus/colornames.htm>

basic colours.

Hence, colour attributes in the logical image representations should just use the aforementioned eleven basic colour terms. Variations in intensity should only be expressed by adding the labels *light* and *dark* (like “a *dark* green palm tree”). The suffix *-ish* should be used if the colour is similar to one of the basic colours (“a *greenish* palm tree”). Objects with a colour between two basic colour terms should be described with a combination of the two (like “a *yellowish-orange* drink”).

4.3.3 Annotation Process

In order to provide a consistent set of logical image representations for evaluation purposes, we manually checked, corrected and completed the original captions (see Section 4.3.1) of the images that we had selected for inclusion in the *IAPR TC-12 Benchmark* following the prior established image selection rules (see Section 4.2.3). The annotation process was carried out in several iterations using a custom-built benchmark administration system (see Section 6.3):

1. The first iteration saw the elimination of all the images that were either copyrighted or did not have any information about their originator (and therefore no information about potential copyright restrictions). We also eliminated all the images with insufficient information about their location or creation date.
2. In the second iteration, we double-checked the titles of the original captions. This included a spell and grammar check as well as the check for appropriateness and correctness of the title. We also translated the mainly German titles into English and Spanish, and in case of doubt, we verified the title and location of an image with *viventura* and/or the guide that had captured it.
3. The third iteration was the most time-consuming step: the creation of the German semantic descriptions of the visual contents of the images plus non-visible additional information (if available) and their translation into English.

4. Finally, we created the Spanish translations in the fourth iteration¹¹. At the same time, the German and English image representations were double-checked once again and statistical information was added to each image.

4.3.4 Image Semantics Analysis and Statistics

After having completed the annotation process, each image was assigned a semi-structured representation that consists of the following seven fields: (1) a unique identifier, (2) a title, (3) a free-text description of the semantic and visual image contents, (4) notes for additional information, (5) the name of the photographer, (6) the date when and (7) the location where the photo was taken.

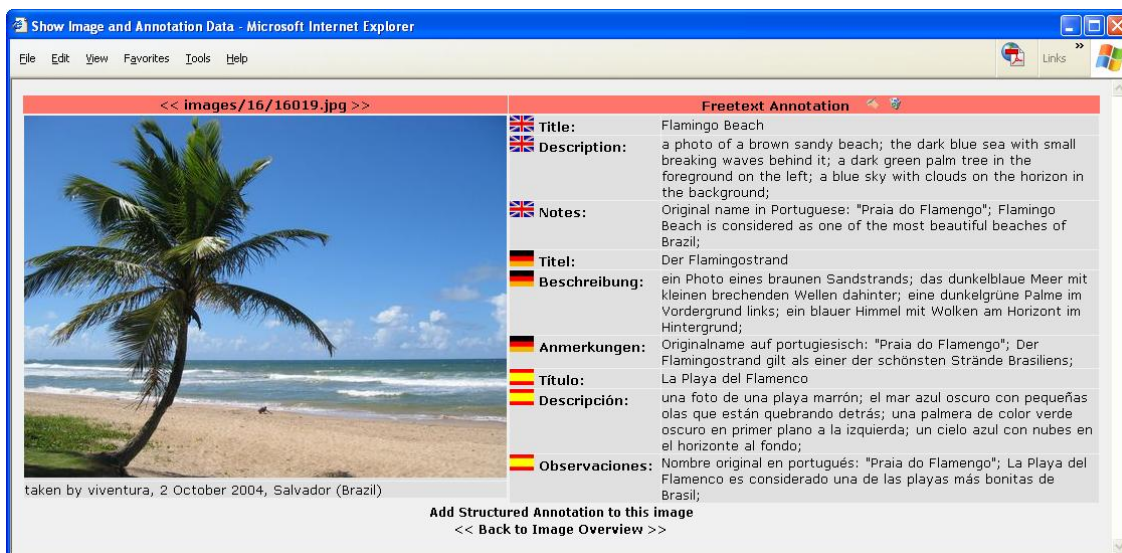


Figure 4.27: Complete caption for image 16019.jpg.

These seven fields are stored in a MySQL database and managed by a benchmark administration system (see Chapter 6 for a detailed description). Figure 4.27 shows the final semantic image representation for the sample image in Section 4.3.1.

Unique Image Identifiers

Each image is assigned a unique identifier. For instance, the unique identifier of the example in Figure 4.27 is *16019*, which also determines the filename of the image

¹¹The Spanish description fields are currently being verified by a native speaker and therefore not yet in a release status.

(16019.jpg) and of the alphanumeric representation files (16019.eng for English, 16019.ger for German, 16019.spa for Spanish).

Title

The *title* field contains a short statement about the general contents of an image. This can include proper names like “Flamingo Beach”, general noun phrases like “cyclist at night” or a combination of both such as “llamas at Machu Picchu”. The title can also be a short sentence like “Max is surfing in Torquay”.

This title field is equivalent to the descriptive captions found in many generic photographic collections (*i.e.* captions that typical users might add to their own photographs in private collections). In most cases, the title field is not very different from the original captions. The average length of the title field for English is 5.35 words, with a standard deviation of 2.37 words. The shortest title consists of one word; the longest comprises 17 words. Table 4.2 displays the statistics for the titles in English, German and Spanish.

Number of words	English	German	Spanish
Average	4.85	5.35	5.97
Standard deviation	2.10	2.37	2.68
Minimum	1	1	1
Median	5	5	6
Maximum	14	17	19

Table 4.2: Word statistics for the *title* field.

The German titles are on average shorter in length (and Spanish titles longer) than the English titles. This does not necessarily mean that the Spanish titles are more complex than the German ones; it is more likely due to the fact that composite nouns that can be described in one word in German (*e.g.* “Flamingostrand”) are often expressed by two words in English (“Flamingo Beach”) and by three words in Spanish (“Playa del Flamenco”).

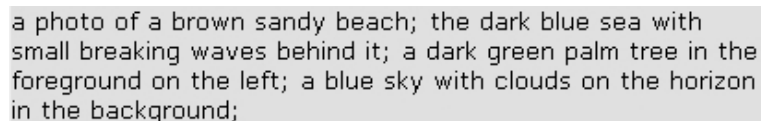
Description

The *description* field contains a semantic description of the image contents; it describes in short sentences and noun phrases (terminated by semicolons) what can be recognised in an image without prior information or extra knowledge and was created in compliance with the annotation rules described in Section 4.3.2. These free-text representations exist in three languages: English¹², German¹³ and Spanish.

Number of words	English	German	Spanish
Average	23.06	18.92	27.46
Standard deviation	10.35	8.38	12.72
Minimum	2	2	2
Median	22	18	25
Maximum	85	74	93

Table 4.3: Word statistics for the *description* field.

The average length of the English description field is 23.06 words (with a standard deviation of 10.35 words). The shortest description comprises two words; the longest is 85 words, with a median of 22 words (see Table 4.3). Again, the German descriptions use, on average, fewer and the Spanish descriptions more words than the English version.



a photo of a brown sandy beach; the dark blue sea with small breaking waves behind it; a dark green palm tree in the foreground on the left; a blue sky with clouds on the horizon in the background;

Figure 4.28: Description field for image 16019.jpg.

Figure 4.28 shows an example description field with four representation sentences. Many of these sentences or noun phrases follow a specific linguistic pattern P (or a more difficult combination based on these) as illustrated in Table 4.4.

Any of the patterns P mentioned in Table 4.4 are used in both foreground and

¹²We used Australian English for the English version as the image annotation process was undergone in Melbourne, Australia. We did, in cases of doubt, ask local native speakers for translations or vocabulary.

¹³The German version uses Austrian German vocabulary and spelling because the logical image representations were created by an Austrian citizen (myself).

Pattern P	Example
S	a red rose
S-V	a boy is singing
S-TA	a boy at night
S-PA	a boy in a garden
S-PA-TA	a boy in a garden at night
S-V-TA	a boy is singing at night
S-V-PA	a boy is singing in a garden
S-V-PA-TA	a boy is singing in a garden at night
S-V-O	a boy is kissing a girl
S-V-O-TA	a boy is kissing a girl at night
S-V-O-PA	a boy is kissing a girl in a garden
S-V-O-PA-TA	a boy is kissing a girl in a garden at night

Table 4.4: Linguistic patterns of the *description* field.

background information and can further be specified with respect to their spatial location within the image (see Table 4.5).

Pattern P	Example
P-PA	P on the left
P-BG	P in the background
P-FG	P in the foreground
P-BG-PA	P in the background on the left
P-FG-PA	P in the foreground on the right

Table 4.5: Linguistic patterns of the *description* field.

Table 4.6 provides an overview and a description of the symbols used in Tables 4.4 and 4.5.

Symbol	Description
S	subject(s) with or without adjective(s)
V	verb(s) with or without adverb(s)
O	object(s) with or without adjective(s)
PA	place adjunct(s) including place preposition(s)
TA	time adjunct(s) including time preposition(s)
P	any pattern or combination of patterns
FG	in the foreground
BG	in the background

Table 4.6: Pattern symbols.

Notes

The *notes* field contains additional free-text information about the images in English, German and Spanish and, unlike the description field, does not follow any underlying patterns or annotation rules. It can include information like original names in other languages (Figure 4.29(a)), historical information (Figure 4.29(b)), eventual results of sports events (Figure 4.29(c)) or any description that is not visible in the image and requires prior or deeper knowledge of the image contents.

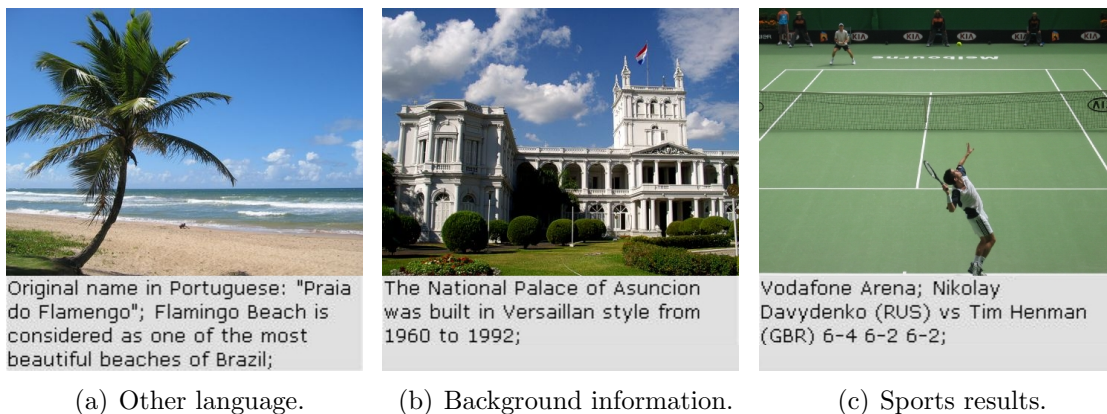


Figure 4.29: Sample images for notes.

Not all images have notes fields. In fact, just 10.3% of the images hold additional, non-visible information, with an average length of 11.88 words per notes field and a standard deviation of 7.99 (for the English representations). The longest notes field contains 55 words, the shortest just one, with a median of eleven words (see Table 4.7).

Number of words	English	German	Spanish
Average	11.08	10.84	12.05
Standard deviation	7.99	7.26	8.72
Minimum	1	1	1
Median	11	9	13
Maximum	53	59	62

Table 4.7: Word statistics for the *notes* field.

Dates

The *date* field contains the date when the image was taken, with each language having its own version and format: English (*e.g.* “2 October 2004”), German (*e.g.* “2. Oktober 2004”) and Spanish (*e.g.* “2 de octubre de 2004”).

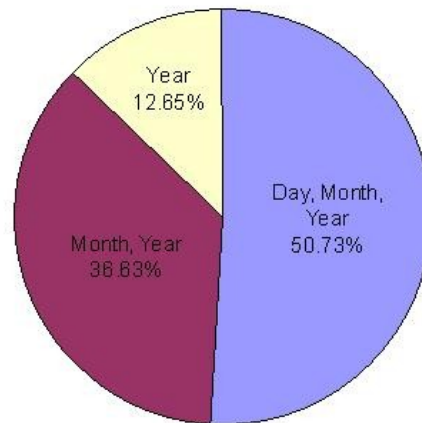


Figure 4.30: Temporal granularity.

Figure 4.30 shows the three different time granularity levels: 51.7% of the images have a complete date (day, month, year), 36.6% have month and year, and 12.7% of the logical image representations just state the year.

Locations

The *location* field describes the place where the image has been taken and is divided into two parts:

- the exact location (*e.g.* “Salvador”);
- the country to which that location belongs to (*e.g.* “Brazil”).

Some images (2.35%) only have country information in cases where the exact location in that country could not be verified.

Location names are stored in three different languages (English, German and Spanish). However, the question of whether place names are to be translated or not is a special challenge *per se* as there is no general answer: while most countries

do have their own version in each of the three languages, like “Brazil” (English), “Brasilien” (German) and “Brasil” (Spanish), there is no pattern as to whether city names are translated or not. In many cases, it is true that the more unknown a place is, the less likely it will be translated into a foreign language.

This rule of thumb is not always applicable though. Consider the two locations Rome and Buenos Aires, both big and famous cities: the Argentine capital is the same in all the three languages (“Buenos Aires”), whereas the Italian capital has a different version in each of the languages: “Rome” in English, “Rom” in German and “Roma” in Spanish. Hence, since there is no universal rule, we had to check each location individually whether there exists an official translation or not, no matter how big or famous the location.

4.4 Summary

This chapter reported on one of the major scientific contributions of this research: the creation of an image collection called the *IAPR TC-12 image collection*, which was specifically designed and implemented to deal with the lack of resources for evaluation of visual information retrieval (VIR) from generic photographic collections.

While most other benchmark collections were originally created for purposes other than VIR evaluation, the novel (and only) goal for the development of the image collection for the *IAPR TC-12 Benchmark* was to provide a generic photographic collection which could be used for a variety of research and evaluation purposes. This first involved the definition of requirements and specifications, before we established the design methodology that (1) subsequently built the foundation for the image selection and annotation processes and (2) allowed for the strict control over consistency and quality within all aspects of collection creation. As a consequence, the newly created evaluation resource exhibits the following benefits:

- *Quality.* All the photos are high-quality colour photographs with excellent levels of resolution and contrast.

- *Variety.* The generic image collection contains real-life photographs from a range of subjects and settings.
- *Multilingualism.* High-quality multilingual semantic image representations make the collection suitable for the evaluation of a range of retrieval tasks.
- *Royalty-Free.* The collection is available freely and without any fees or charge.
- *Copyright.* There are no copyright restrictions that would hinder the large-scale redistribution of the collection for evaluation purposes¹⁴.

The last two benefits, in particular, provide a significant contribution since high-costs, as well as copyright restrictions, have hindered the progress for large-scale VIR evaluation events for a long time.

The successful establishment of the *IAPR TC-12 image collection* completes the creation of a representative image collection - the first key component of the *IAPR TC-12 Image Benchmark*. In the next chapter, we will report on the creation of its second key component: a balanced set of query topics representing realistic user information needs for retrieval from the *IAPR TC-12 image collection*.

¹⁴The conditions of an end-user agreement apply. Please see Appendix B for more details.