Appendix A

Experimental setup

The ultimate performance test for a CBIR system is the degree to which it is found useful by its users. This is however not an easy property to test without conducting field tests with real users, and performing experiments with human subjects is usually a complex task. A significant number of subjects must be assembled for the results to be statistically significant, the experiments must be carefully designed to ensure that they do not bias the subjects to the desired responses, and only simple and relatively fast tests can be conducted if one expects to engage a large number of subjects. Furthermore, it is not possible to repeat an experiment when something goes wrong, or to modify the system parameters and try again.

The next best alternative is to define an objective criteria for performance evaluation that does not require human intervention. Because it is so much simpler than field testing, this has been the evaluation method of choice among the retrieval community. It must, however, be performed carefully if one is to avoid oversimplified scenarios that have small resemblance with reality. In this context, the two main free variables for the design of an automated testing strategy are the choice of databases and performance criteria. In this appendix, we discuss the reasons that motivated the experiments described in the thesis.

A.1 Databases

While CBIR systems are ultimately designed to deal with generic images, generic databases are not always the most suited to allow the understanding of the strengths and weaknesses of different retrieval approaches. This is due to various reasons. First, classifying a collection of generic images is usually a subjective task to which different people will give different answers. This ambiguity makes it difficult to establish the ground truth that is required for automated testing. Second, on generic databases it is difficult to determine exactly what are the properties of the recognition architecture that are responsible for particular successes and failures. E.g. one image representation may characterize better color, while other may characterize better texture, and a third may characterize better object shape, all leading to similar overall results. While combining the strengths of the three methods would lead to significantly better performance, it may be difficult to determine what those strengths are since it is difficult to tell what properties are most important for each image. Third, tests that require ground-truth for particular visual concepts (e.g. the presence of a given object in an image) can only be performed upon manual annotation of the entire database. This is particularly difficult in cases where it is not even clear what the objects of interest may be before testing begins, e.g. the evaluation of learning algorithms. Finally, because there is only a short history of evaluation with generic databases, it is difficult to compare results with previously proposed retrieval solutions.

For all these reasons, while it is imperative to, whenever possible, evaluate performance on generic databases, it is also useful to consider databases that 1) stress specific aspects of the retrieval problem, 2) have unambiguous classification ground truth, and 3) have a long usage history in the retrieval literature. This observation motivated us to, in addition to the generic Corel database of stock photography, also consider in many chapters of the thesis two specialized databases: the Brodatz texture database, and the Columbia object database. For all databases, we have also tried to identify the approaches that are known to give best retrieval results for the image properties captured by the database, implement those approaches, and compare results to those obtained with the recognition architecture now proposed. This is an unfair test in the sense that these approaches tend to be specific to the database domain (e.g. texture), while the architecture now proposed is generic. It was, nevertheless, necessary since one of the important goals of the thesis was to know how close a generic architecture can get to the performance of the specialized approaches in their domains of expertise. We next provide a more detailed description of each of the databases.

A.1.1 Standard databases

The Brodatz database is a set of images from 112 fairly homogeneous textures. Each of these 112 images was broken into 9 128 × 128 patches for a total of 1008 database entries [134]. This database was further divided into two subgroups: while the first image in each texture class was stored in a *query database* (112 total images), the remaining 896 images formed the *retrieval database*. Among the various approaches proposed in the texture recognition literature, independent tests conducted by different laboratories [134, 94, 96] have shown that the combination of 1) the coefficients of a least squares fit of the multi-resolution SAR (MRSAR) texture model [104] and 2) the Mahalanobis distance (MD) achieves the best performance on Brodatz. The performance of the MRSAR/MD combination is, therefore, usually considered a good benchmark for the evaluation of state of the art texture recognition algorithms.

The Columbia database is a set of images of 100 objects each shot in 72 different views obtained by rotating the object in $3D$ in steps of 5^o. Once again, the database was split into two subgroups, one containing the even and the other the odd images from each object. For computational simplicity, the retrieval database was further sub-sampled by four (9 views of each object separated by 40^o) and only the first image of each object was kept in the query database (100 total images). All images were converted from the original RGB to the YBR color space (as defined in the JPEG standard [128]). The Columbia database is similar in many aspects to the color database used by Swain and Ballard when they introduced histogram-based recognition [172], but significantly larger. While there as not been such an extensive evaluation of color-based retrieval techniques as in the case of texture, it is safe to say that the combination of color histograms with the histogram intersection (HI) metric [172] has so far become the de-facto standard in the area. We therefore rely on HI as a benchmark for the evaluation of color-based retrieval.

The Corel database is a set of $20,000$ images from 200 generic image classes (100 images in each class). While the groupings are of a semantic nature and there are several classes which are impossible to recover by the analysis of low-level properties such as color and texture (e.g. classes containing too much variety of visual stimuli like "China" or "Egypt," broad concepts like "nature scenes" that overlap with more specific ones like "North American wild flowers," concepts that require higher-level content understanding like the "spirit of Buddha," etc.), there are also various classes characterized by an amount of visual uniformity that makes the task feasible. For our experiments we selected 16 among these ("Arabian horses," "auto racing," "coasts," "divers and diving," "English country gardens," "fireworks," "glaciers and mountains," "Mayan and Aztec ruins," "oil paintings," "owls," "land of the pyramids," "roses," "ski scenes," "religious stained glass," "tigers") leading to a total of $1,600$ images.

In order to create the query database, we randomly selected 20% of the images in each class, leaving the remaining 80% in the retrieval database. All images were converted from the original RGB to the YBR color space. Note that, even though the classes are somewhat visually uniform, there is plenty of variation within them and retrieval is significantly harder than in the case of the two previous databases. In particular, it does not suffice to use color or texture attributes alone, but representations that can account for both color and texture. Thus, in addition to MRSAR/MD and histogram intersection , we considered two such approaches: color correlograms, and linear weighting of texture and color. These are discussed in Chapter 6.

A.1.2 Artificial databases

Automated performance evaluation is particularly difficult for local queries, since these involve image segmentation and it is infeasible to manually segment all the images in the database to establish ground truth. The problem is even worse when evaluating learning algorithms because, in this case, the objects or concepts to retrieve may themselves change during learning. A feasible alternative is to construct artificial databases where the ground truth is always known. In this thesis, we pursue this alternative exactly for the evaluation of local queries and short-term learning. In particular, all experiments performed in these areas were based on two artificial databases constructed from Brodatz and Columbia.

In each case, an artificial database was created from the retrieval databases described above, by randomly selecting 4 images at a time and making a 2 × 2 mosaic out of them. Figure A.1 shows two examples of these mosaics. We call these image sets the *mosaic* databases. They are representative of databases whose images do not consist of a single object or visual concept but are instead a composition of different visual stimulae.

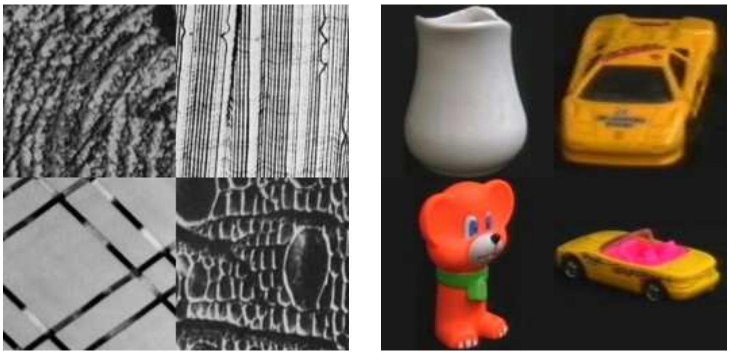

Figure A.1: Example mosaic images derived from the Brodatz (left) and Columbia databases (right).

A.2 Evaluation criteria

In addition to a database for which the classification ground truth is unambiguous, the objective evaluation of a retrieval system also requires a criteria for performance evaluation. The most commonly used criteria in the visual recognition literature is classification accuracy. This is obtained by performing several queries and measuring the fraction of these in which the top match belongs to the same class as the query. Despite the long history of its use, classification accuracy only provides a limited view of the abilities of a retrieval system. If there is an image in the database which is a close replica of the query (e.g. two pictures of the same scene taken a few minutes apart), any sensible retrieval approach will return that image as the best match. This, however, does not mean that the retrieval results are good. In fact, if the retrieval system is poorly designed and the remaining images of the same class are not near exact replicas of the query, these images may receive a very low rank in the list of returned images.

This issue has been discussed at length in the text retrieval literature, where it has long been agreed upon that a good evaluation criteria should consider more than the best match alone. The standard performance metric in text retrieval is *precision/recall* (PR), and consists of a mix of two different criteria. The basic idea is that, since users are likely to look only at the top matches, only a portion of database entries should actually be returned in response to a query[1]. *Precision* is the fraction of the returned images that are relevant to the query, and *recall* the fraction of the total number of relevant images that are returned. If $\mathbf{T}$ is the set of returned images and $\mathbf{R}$ the set of images that are relevant to the query, then

$$precision = \frac{|\mathbf{R} \cap \mathbf{T}|}{|\mathbf{T}|} \tag{A.1}$$

$$recall = \frac{|\mathbf{R} \cap \mathbf{T}|}{|\mathbf{R}|} \tag{A.2}$$

where $|\mathbf{A}|$ is the cardinality of the set $\mathbf{A}$. Since there is no optimal value for the cardinality of the set of retrieved images, results are usually presented in the form of a PR curve. Several levels of recall $\{l_1, \ldots, l_m\}$ are established and $\mathbf{T}_i$ is the smallest set of returned images that satisfy recall level l_i. Precision is then measured for each $\mathbf{T}_i$ originating a PR curve. Usually, the curve is averaged over several queries.

PR is a much more complete performance criteria than classification error, since it also provides information about the images that were not returned as the best match. For example, low precision at high recall indicates that the system has difficulty in capturing the *diversity* of the images in the class of the query. Since generalization is one of the most difficult problems in visual recognition (where a simple change of the imaging parameters, e.g. 3-D rotation, can lead to substantial changes of visual appearance), PR is a much better performance criteria than classification error for this domain. This has indeed become the prevalent view in the image retrieval community, where PR is the main tool for performance evaluation.

[1] Otherwise, users may feel overwhelmed and assume that the system is not sure about the results of the search.