Short version: evaluation and the development of content-based retrieval techniques

Arjen P. de Vries
arjen@cs.utwente.nl
Centre for Telematics and Information Technology
University of Twente
The Netherlands

1 Literature review

There seem to be no satisfactory methods for measuring the effectiveness of multimedia search techniques. Precision and recall types of metrics have been used in some of the literature but are impractical due to the tedious process of measuring relevances. The process is even more complicated because of human subjectivity in tasks involving multimedia. Also, multimedia collections quickly grow very large, making evaluation expensive with respect to the required hardware. As a result, there seem to be no standard corpora or benchmark procedures.

topic	# papers	# projects
audio/speech	7	4
image features	13	13
image databases	20	9
video	7	6
web, multimedia	8	4

Table 1: Overview of amount of papers reviewed

We reviewed the evaluation approaches that have been taken in the literature about multimedia retrieval systems. In this report, we just present some overall conclusions of our review. Our review covers far from all related publications, but we believe the sample of papers studied is sufficiently large to draw valid conclusions. Table 1 summarizes the amount of research studied in the review. We categorized the papers dealing with image retrieval in two categories, one focused on feature representations themselves and one on the retrieval of images using these representations. The web/multimedia category consists of papers on research projects that attempt to address retrieval of multimedia documents rather than base objects of some particular new medium. The last field provides a rough estimate of the number of different research projects that are reported upon in the papers used.

A great deal of published multimedia retrieval research barely has an evaluation phase. The techniques are explained, and the results of a small set of example queries are given to convince

us that the techniques work. The lack of evaluation makes it very hard to say something useful about the performance of these approaches. Multimedia retrieval research is still in its infancy. Apparently, the introduction of new techniques does not yet have to be supported by thorough experimental evaluation. The community is still in the 'proof-of-concept' phase.

In all the papers studied in this review, the data sets are very small. Many collections are tailored to evaluate only a very specific low-level task, cf. the evaluation of texture algorithms based on the Brodatz collection. However, the results on such a collection are easily generalized for very different, high-level search tasks. Although some authors seem to realize the relative weaknesses in their evaluations, others are perfectly happy with the 'proof' derived from their experiments.

Most papers claim to present a novel, *better* approach to multimedia query processing. As proof that the novel approach really is 'better', results provided in an 'evaluation' section are vaguely based on concepts borrowed from the scientific evaluation methodology used in IR. The evaluation sections in most papers proof however mainly the authors' lack of understanding of that methodology:

- only a small number of queries is used (often one or two);
- only precision-recall measures for one cut-off point in a ranking are presented;
- a significance test is not applied;
- the data is divided in a small number of classes, and these are considered both the relevance judgments and the complete description of the user's information need;
- relevance judgments are considered completely objective, but usually made by the paper's authors and not by real users.

The inherent subjectivity of multimedia search is usually ignored completely. Almost all papers report experiments with multimedia search in which only the success on a task of object identification is tested: e.g. does the image contain a lion or not. The emotional and aesthetic values that play a role in the evaluation process of the user are overlooked. Or, even worse, the underlying techniques are 'improved' in a such a way that they are less sensitive to exactly those aspects that *are* important for such values.

We are (unfortunately) not convinced that an evaluation methodology for multimedia retrieval exists that can draw valid conclusions based on experiments without real users. The underlying problem is that there exists no objective ground truth in retrieval experiments involving multimedia data. Historical data of experiments with real users in common test sets may be crucial to allow comparison between different approaches.

Until we find a better approach to measure the performance of multimedia retrieval systems, it is very important that we realize the limitations of our experimental 'proof'. We should also realize that to the end users, multimedia retrieval often constitutes much more than 'just' the identification of objects of some particular class.