Multimedia Indexing and Retrieval

1 Motivation

Devices for creating, storing, and transmitting multimedia information are increasing in prevalence and capacity, and are decreasing in price. With little prospect of such changes slowing in the foreseeable future, it is not hard to predict with some confidence that content-based access to multimedia information (indexing, retrieval, browsing, summarisation, etc.), is set to become a significant problem for web and enterprise search as well as for personal retrieval. The problem space of this topic is large because

1. the types of objects to be retrieved are varied ranging from collections of audio (e.g. speech, music), images (e.g. photographs, clip art, scanned documents), video (e.g. TV, camcorder, or security camera output), as well as less common objects (e.g. vector drawings, recorded pen strokes, and VRML);

2. the methods available for indexing and retrieving such objects vary significantly depending on their type, which has a strong impact on the forms of retrieval possible;

3. however, it is also clear that the forms of retrieval users will wish to conduct will vary for each media type.

It is apparent that a series of short term challenges exist, which may lead to a longer term grand challenge. Both are outlined here as well as the needs of the research community so that it can effectively investigate the topic. The impact of language modelling is also discussed.

2 Three to five year challenges

The current challenge in multimedia retrieval centers on indexing: given a non-text media object, the following options are available. Text may be associated with the object (e.g. captions, juxtaposed text); part of the object might be convertible to text (e.g. through speech recognition or OCR); metadata might be assigned manually or media specific features might be extractable.

1. Extracting good indexing features from most forms of multimedia is hard (except within restricted domains). As an alternative, fragments of text or pieces of metadata may be located and used for indexing. For certain media types, text may be easy to come by, for others, however, little or no text will be directly associated with media objects.

Automatic annotation may be one way of associating text with such media objects. This may involve learning from examples how text and media features are associated and then using that information for annotation.

2. The context of user activity when a media object is created may provide a good approach to indexing: for example, emails, diary entries or location-based information associated with
the time that a photo was taken could be examined for pertinent text. Context at retrieval
time will also be important: the location of the querier (a particular country; a particular
part of an office) or the type of device a user is using will inform the type or content of
objects to be retrieved.

3. One of the major reasons for the successes of (text) IR has been the application of formal
models for indexing and retrieval. Applying such models to the multimedia domain has
been challenging partly because the features which are often most useful do not easily
lend themselves to indexing. The extraction of appropriate features which can be used
for indexing is a challenge. So is the application of formal IR models to existing features.
Multimedia (image, video and audio) IR tasks need to be formulated which cannot be solved
using text alone but will require advances in finding such features and advances in applying
formal IR models to such tasks.

4. We also need to think of different kinds of tasks which involve data mining and retrieval
of time sequences of images, video and audio from other domains. Example: A few years
ago there was a paper on collecting time sequenced images of storms from radars and then
retrieving similar storms from a database of time sequenced images of storms to predict the
future track of the storm. This is basically a natural application of video retrieval applied
to this task.

5. To deal effectively with multimedia retrieval, one must be able to handle multiple query and
document modalities. In video, for example, moving images, speech, music audio, and text
(closed captions) can all contribute to effective retrieval. Integrating the different modalities
in principled ways is a challenge.

3 Needs

In order to assess research effectively in multimedia retrieval, task related standardized databases
on which different groups can apply their algorithms are needed. In text retrieval, it has been
relatively straightforward to obtain large collections of old newspaper texts because the copyright
owners do not see the raw text being of much value, however image, video, and speech libraries
do see great value in their collections and consequently are much more cautious in releasing their
content. While it is not a research challenge, obtaining large multimedia collections for widespread
evaluation exercises is a practical and important step that needs to be addressed. We suggest that
task related image and video databases with appropriate relevance judgements be included and
made available to groups for research purposes as is done with TREC. Useful video collections
could include news video (in multiple languages), collections of personal videos and possibly movie
collections. Image collections would include image databases (maybe on specific topics) along with
annotated text - the use of library image collections should also be explored.
4 Language modelling

The application of information retrieval and other statistical machine learning techniques, analogous to language modeling, may be useful in multimedia retrieval. Language modeling has been successful in text related areas like speech, optical character recognition and information retrieval. There is some evidence that some of these models may be useful in automatic annotation, combining image and text retrieval and image segmentation.

5 Grand challenges

If the challenges listed above can be solved, then the over-arching and outstanding task will be to bring together the strands of multimedia IR - different media types, different indexing approaches, multiple modalities - and then to incorporate context in order to produce a “grand challenge”: to create a context aware, transmedia IR system. With such a system, a user will enter a query, e.g. “Taj Mahal”, and if the user is at a desktop computer and has spent time earlier planning a conference trip to India (reading emails on the trip; examining travel web pages; placing entries in a diary), then the system will be aware of this context and will be more inclined to retrieve pictures and videos of the Indian mausoleum while music from the jazz band with the same name would be less likely to be retrieved. Not only would images indexed by the query text be presented, also featuring in the ranking would be images similar in visual content to the retrieved text-tagged images. The high bandwidth of the machine’s connection would also be taken into account when choosing images or videos to be delivered, the bandwidth may even be a factor in the retrieval algorithm. On the other hand, if a user were starting out on a long car trip (the system being aware of this context from diaries and location sensors) and the user has often requested audio in such a situation, then the music from the jazz band will be more likely to be retrieved. It is this context awareness, aware of the user, the current task, the history and future plans, location and device, and the fact that retrieval can be of any kind of digital information, regardless of query type or mode, that makes this a grand challenge.