Interactive Indexing and Retrieval of Multimedia Content

M. Worring *[1], A. Bagdanov[1], J. v. Gemert[1], J-M. Geusebroek[1], M. Hoang[1], A. Th. Schreiber[2], C.G.M. Snoek[1], J. Vendrig[1], J. Wielemaker[2], and A.W.M. Smeulders[1]

[1] Intelligent Sensory Information Systems, University of Amsterdam
Kruislaan 403, 1098 SJ Amsterdam, The Netherlands
[2] Social Science Informatics, University of Amsterdam
worring@science.uva.nl

Abstract. The indexing and retrieval of multimedia items is difficult due to the semantic gap between the user's perception of the data and the descriptions we can derive automatically from the data using computer vision, speech recognition, and natural language processing. In this contribution we consider the nature of the semantic gap in more detail and show examples of methods that help in limiting the gap. These methods can be automatic, but in general the indexing and retrieval of multimedia items should be a collaborative process between the system and the user. We show how to employ the user's interaction for limiting the semantic gap.

1 Introduction

The multimedia retrieval field is advancing rapidly, but the ease of use of multimedia content is still far behind compared to textual content. The reason for this is twofold. First, due to the huge datasizes, access to multimedia content requires high-speed networks and sophisticated system architectures. These important aspects related to multimedia delivery are not considered here. The second reason is the sensory nature of the data. Image, video, and audio content are all observations of some phenomenon in the world. The resulting data array bears all the information in an implicit form. This problem is known as the semantic gap [1]:

> *The semantic gap is the lack of coincidence between the information that one can extract from the (sensory) data and the interpretation that the same data has for a user in a given situation.*

In accessing multimedia data there are two related, but still distinct tasks. The first task is multimedia indexing, in which the data is annotated with indices describing the content. This task is performed by a professional indexer, or to

* This work is supported by the ICES-KIS MIA project

some extent by an automatic system. These indices form the starting point for the second step namely multimedia retrieval where a non-expert user is searching for relevant information in the multimedia database. Both of these steps suffer from the semantic gap. Different solutions, however, are required as the intended user has a different level of expertise and also has different intentions.

To be successful in multimedia indexing and retrieval, a system should bridge the gap. Given the "size" of the gap it is clear that this is a difficult and challenging task. Work in this direction is by its nature a collaboration between various disciplines. Computer vision, speech recognition, natural language processing, statistical pattern recognition, knowledge engineering, information visualization, human-computer interaction, and database technology are among the key fields to consider. In this paper we present results from the large-scale Multimedia Information Analysis[3] project (MIA) which involves various groups covering most of the above disciplines. We focus on how parts of this project have contributed in limiting the semantic gap. For work of others, good starting points are the two reviews we have published on content based image retrieval [1] and multimodal video indexing [2] respectively. Together they cover some 300 references. Furthermore, a very good review on pattern recognition, the basic tools on which many indexing methods are based, can be found in [3].

In the rest of this paper we will first consider multimedia indexing in section 2 from both the user side and the system side of the problem. In section 4 we consider how we can create synergy between the system and the user in the indexing process. We then move our attention to the user who is searching for multimedia items in an archive in section 5.

2 Multimedia indexing

In manual annotation of multimedia data the user is directly looking (or listening) to the data and provides a complete description of its content. Manual annotation seems not to suffer from the semantic gap. The user sees the data and can immediately add his own description. But, considered from the system point of view the gap is in fact very large as the system cannot understand anything of the user's annotations.

At least some automation of the annotation process is required as manual annotation is a very time-consuming task. Furthermore, the terminology used by the indexer might not be consistent over time. Also, the user is annotating for the task at hand, whereas later use might require different descriptions.

Computing features from the data like various color histograms, texture descriptions, and spatio-temporal activity is a first step in bringing system and user closer together. It is in this respect important that the features computed are closely related to the indexing task at hand. In particular, we make a distinction between the properties of an object in the image, versus the sensory conditions under which the scene is recorded.

[3] See http://www.science.uva.nl/~worring/mia.

At the user side we can limit the gap by using ontologies to provide a closed vocabulary for the domain at hand, making sure that terms are used in a consistent way.

For some terms in the ontology it might be feasible to compute the relevant terms from the data and thus provide an automatic index for the data. In some cases it is sufficient to consider one media in computing the index, but in many cases it is advantageous to use different media like visual and auditory information in deriving the index. This does, however, lead to more complex methods for automatic processing.

The above steps and their relation to the semantic gap and effort are illustrated in figure 1

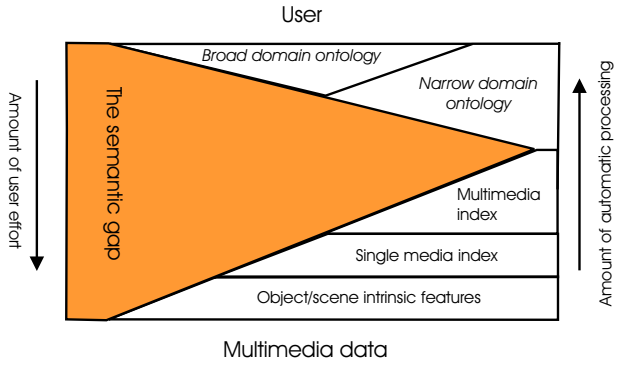

Fig. 1. *Illustration of the semantic gap and how indexing and knowledge based tools can help, as well as an indication of how the user effort and system effort increase when trying to limit the gap.*

2.1 Object/scene intrinsic features

Finding good features to describe multimedia data is an important and difficult task. This is also a hard task as it depends on both the data and the indexing task at hand. A key notion in finding the proper features is invariance [1]. It is a formulation of the notion that irrelevant information in the data should be reduced to the largest extent possible while retaining as much discriminatory power as possible. To be precise:

> *A feature f is invariant with respect to some unwanted condition W if and only if $f(t)$ yields the same value for any pair of multimedia items t_1 and t_2 which only differ in characteristics caused by W.*

To illustrate, consider an image containing an object against a uniform background which is described using the common RGB histogram. If the viewpoint

of the camera would be changed or the intensity of the light varied, RGB histogram features would show dramatic changes even though the object itself has not changed at all. Describing the image with features based on the histogram of the Hue component of the HSV color space would yield much better results in the above example as it is invariant to intensity changes. In general, the proper choice of the color space depends on the task. A full overview of different invariant color spaces and an indication when they should be used can be found in [4]. On the other hand, the variant properties do have a great influence on how a human perceives the image or sound. A loud piece of music might be far more annoying than hearing the same piece of music at normal volume.

In general, we observe that invariant features are related to the intrinsic properties of the objects in the image, video or audio. Thus, invariant features are good for classifying or recognizing the object. The remaining variation within an equivalence class is a good descriptor for the scene and recording related properties of the multimedia item. Decomposing the features into object and scene intrinsic features makes it easier to map concepts to the proper features.

2.2 Ontologies and domains

Invariant features reduce the irrelevant variance at the data side. When a user is annotating the data, especially when this is in the form of free text, there is also a large unwanted variance at the user side. Different terms might be used for the same concept and the level of description might also vary. A user might be annotating a picture with "a scene in Amsterdam" but will not include the description "a scene in the Netherlands". Depending on the search task, one of the two indices is needed.

To limit the variance at the user side an explicit ontology, being a closed vocabulary for a particular domain, is urged for [5]. In such an ontology based annotation, terms used are directly mapped to one concept, and the problem of the level of description is automatically taking care of as the terms are part of a concept hierarchy.

At this point we should make a distinction between broad and narrow domains for the purpose of relating descriptions and appearance. In the repertoire of multimedia items under consideration there is a gradual distinction between narrow and broad domains. At one end of the spectrum we have:

> *A narrow domain has a limited and predictable variability in all relevant aspects of its appearance.*

Usually, the recording circumstances are also similar over the whole domain. In the narrow domain of lithographs, for instance, the recording is under white light with frontal view and no occlusion. Also, when the object's appearance has limited variability the semantic description of the image is generally well-defined and, by and large, unique. Another example of a narrow domain is a set of frontal views of faces, recorded against a clear background. Although each face is unique and has large variability in the visual details, there are obvious

geometrical, physical and color-related constraints governing the domain. The domain would be wider had the faces been photographed from a crowd or from an outdoor scene. In that case variations in illumination, clutter in the scene, occlusion and viewpoint will have a major impact on the analysis. On the other end of the spectrum we have the broad domain:

> *A broad domain has an unlimited and unpredictable variability in its appearance even for the same semantic meaning.*

As is clear from the definition, for specific narrow domains there is a chance to bridge the semantic gap using automatic methods. For broad domains this is currently out of reach.

For narrow domains we have built an ontology based annotation system which provides a direct mapping between annotation terms and concepts in the ontology, by letting user select annotations directly from the ontology. Thus it prevents the use of annotations terms which are inconsistent, and more specific, or more general terms are induced automatically. The tool is illustrated in figure 2.

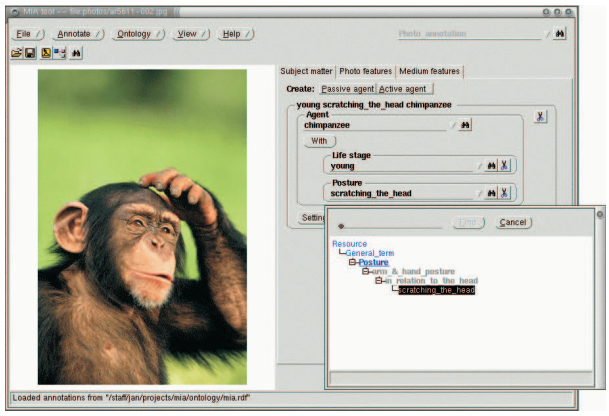

Fig. 2. *Screendump of the ontology driven annotation system (from [5]). The interface is generated automatically from the ontology description. Whenever the user chooses a description for a multimedia item it is related to all relevant topics associated with this description.*

3 Single media index

Having reduced the variance at both data and user side, we can make an effort to bring the ontology and the features together for narrow domains. For broad domains we should first decompose the dataset into different broad categories, which can then be considered narrow domains.

We consider three different, hierarchically ordered, broad categorizations:

– *Purpose*: set of multimedia items sharing similar intention;
– *Genre*: set of multimedia items sharing similar style;
– *Sub-genre*: a subset of a genre where the multimedia items share similar content;

As purpose is usually not directly reflected in the data, we do not consider methods for automatic categorization according to purpose.

We have considered genre classification in the context of black and white documents, either scanned or in electronic form. The style of a document is reflected in its layout in particular in its visual appearance. This in turn is determined by the distribution and size of white and black regions on the page. To that end, we have introduced rectangular granulometries to capture the visual appearance and hence style of the document [6]. Experiments on classifying scientific documents based on their publishing style using rectangular granulometries have shown its superior performance over existing methods in terms of precision and recall. A screendump of the system is shown in figure 3.

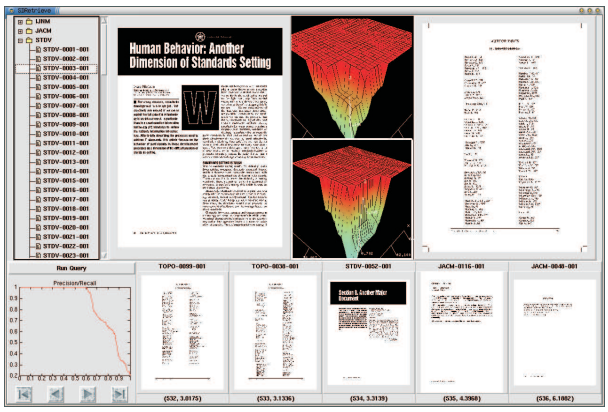

Fig. 3. *Screendump of the system for style based classification of documents described in [6]. The plots in the middle of the picture visualize the distribution and size of the black and white regions in the binary pictures. The graph in the bottom-left gives the precision and recall of the classification.*

We have also considered sub-genre classification in the context of scientific documents - in particular biomedical publications. These publications contain a large amount of pictures with associated captions. To group these according to sub-genre we consider two images to have similar content if they are obtained with the same imaging technique and two captions similar if they use similar terms. To find the image sub-genres we have selected a set of generic features derived from the color/luminance histogram and a set of texture measures [7]. These were computed for all images. A training set was labelled with the imaging technique namely brightfield microscopy, fluorescence microscopy, electron

microscopy, and gels. Then the system was trained to automatically classify an image with unknown type into the proper category based on a decision tree. An overall classification accuracy of 87.5 % was obtained. In addition to the indexing of the picture all associated captions are analyzed using Latent Semantic Indexing. This is an unsupervised methods which associates every caption with a concept built up out of a set of keywords found in the caption. The system is illustrated in figure 4.

Fig. 4. *Indexing of biomedical documents. It allows the user to find images based on the imaging technique and by selecting concepts found in the caption based on user keywords [7].*

In the classification of document and biomedical images, the features were used to classify an unknown multimedia item into a class. It relies on a sufficient amount of previously annotated multimedia items. It thus becomes a supervised classification problem, for which many techniques are found in literature [3]. Supervised classification is the valid mode for automatic annotation. Categories without a name, i.e. clusters found using unsupervised methods like the LSI approach, are only relevant when a user is retrieving information. For text this is feasible as the user can provide the system with a keyword.

3.1 Multimedia index

Indexing single media is sufficient in some cases, but in more cases the concept is reflected in all the modalities that constitute the complete multimedia item. For example a radiograph can be annotated by the radiologist using speech and a picture on a website has related information on the page on which it is placed. In video the multimodality is most prominent as we have a visual channel, an audio signal, and possibly text in the form of scripts for films and closed captions for

news broadcasts. Clearly using all modalities in conjunction yields better indices as the different information channels give complementary information.

For video we have written an elaborate review [2] where we aimed at a general framework fitting the methods in literature. This framework is shown in figure 5. Here we briefly recall the most important aspects of the review.

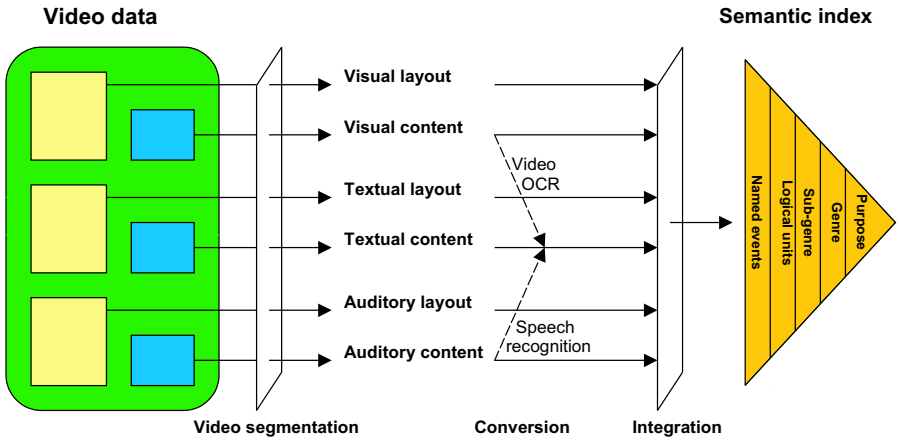

Fig. 5. *Role of conversion and integration in multimodal video document analysis [2].*

The process of multimodal indexing can be divided into three major stages. In the first stage each of the modalities is segmented into the layout e.g. the shots and edits for the visual channel and the content which includes detecting people, objects, and the setting. The second stage is the conversion of modalities into a more appropriate form. In particular, to the textual overlays in the visual channel VideoOCR can be applied to obtain the corresponding ASCII string. Furthermore, speech can be converted to text using speech recognition. It should be noted here that errors are often made in this process, but for later retrieval the quality does not have to be 100%. The final step is the integrated analysis of the different modalities. This is the crux of multimodal analysis and we will elaborate on this stage.

To achieve the goal of multimodal integration, several approaches can be followed. We have categorized existing approaches by their distinctive properties with respect to the *processing cycle*, the *content segmentation*, and the *classification method* used. The processing cycle of the integration method can be iterated, allowing for incremental use of context, or non-iterated. The content segmentation can be performed by using the different modalities in a symmetric, i.e. simultaneous, or asymmetric, i.e. ordered, fashion. Finally, for the classification we have found statistical and knowledge-based approaches. Considering existing methods along these lines brings structure in this new and challenging

field. It also reveals that most methods are still simple, non-iterated methods based on knowledge based, but often ad-hoc, methods. Currently the most successful multimodal indexing methods are based on Dynamic Bayesian Networks or Hidden Markov Models.

For the final index to obtain from a video we add two more levels to the semantic index hierarchy, purpose, genre, sub-genre. These new levels are related to parts of the content, rather than the whole video. They are:

- *Logical units*: a continuous part of a video document's content consisting of a set of named events or other logical units which together have a meaning;

Where named event is defined as:

- *Named events*: short segments which can be assigned a meaning that doesn't change in time;

In the reference, for all four levels of the hierarchy, we have given a complete overview of indices for which automatic indexing techniques have been defined in literature. It indicates that some interesting work has been done in this direction, but that the number of concepts for which methods exist is still far too small. More generic, rather than ad-hoc, techniques for indexing have to be developed over the coming years.

4 Interactive indexing

Up to this point we have considered fully automatic processing and fully manual annotation. This is still the case in most of the existing systems. When automatic methods are used, one either accepts the imperfections in the results or performs an expensive post-processing step to correct the mistakes made. If we accept the fact that human intervention is unavoidable we can better start off with designing indexing as a synergetic process between the user and the system.

We have developed the i-Notation system [8] in which the system and the user in a collaborative process assign labels to shots. In particular the systems aids the user in annotating the shot with the names of all the people present, based on the visual information and the script. The system uses intelligent shot selection to optimize the chance that the user can annotate all of the shots on the screen at the same time using one label. An illustration of the system in action is shown in figure 6.

For shot selection, the user's previous labelling actions are used, allowing for *adaptive* shot selection. The goal is to find the unlabelled shots most likely to have the target label, where we choose the target label to be the previously selected label.

Interaction information comprises both positive and negative information about shot labels. A user gives positive information when he selects a label and associated shots. Thus, negative information is given for the remaining shots because they are not associated with the selected label.

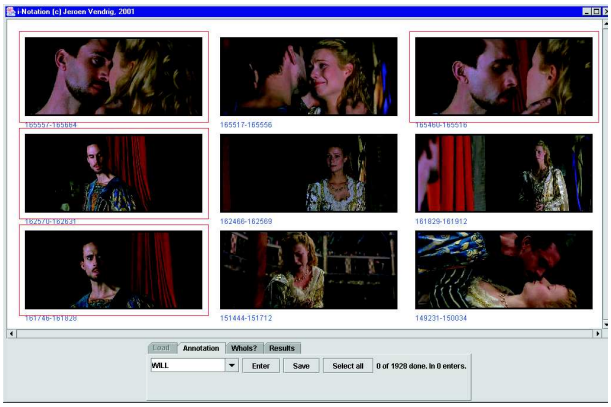

Fig. 6. *Screendump of the system assisted indexing of video material described in [8].*

Based on the various information sources, shots are ranked according to the likelihood they match the target label. To be precise, the likelihood is based on the following similarity scores:

– *Visual similarity* between already labelled shots and an unlabelled shot.

Visual similarity is based on positive feedback, where the background is used to compare shots. We use a combination of the hue-saturation histogram for the chromatic part of the color space and the intensity histogram for the achromatic part.

A shot is compared to all shots already labelled with the target label. The score for the most similar labelled shot is used as the final visual similarity score.

– *Visual dissimilarity* between shots not having the target label and an unlabelled shot.

Visual dissimilarity is based on negative feedback. A shot shown while the user was selecting the target label, but not selected itself, does not have the target label.

– *Label similarity* between target label and expected label for the unlabelled shot.

Label similarity measures to what extent the character names in the target label correspond to the names of the speaking characters in the shot. For each shot an expected label is extracted from the script. This label is compared to the target label for common names.

– *Person presence* similarity between target label and unlabelled shot.

Person presence similarity measures correspondence of the unlabelled shot's visual content to the target label's type. Due to the poor face detection performance on the complex visual scenes found in a film, it is restricted to measuring whether both shot and label contain people.

– *Temporal similarity* between unlabelled shot and shots known to have the target label.

Temporal similarity makes use of the movie characteristic that characters are more likely to reappear in close by shots.

The five similarity scores are combined into an overall similarity measure between the given label and the unlabelled shot, so that shots can be ranked. The top ranked shots are shown to the annotator in the form of key frames. The annotator selects the shots with a similar label as well as the label itself. The system computes a new ranking and the process iterates.

Evaluation based on an explicit user model showed that the proposed process outperforms fully manual and fully automatic indexing using post-processing to correct the automatic result.

5 Interactive retrieval

The user who is searching for information in a multimedia database will have the data as well as the indices generated in the previous sections as his/her disposal. All these indices are objective descriptions of the multimedia data. These descriptions can be features of the data, relations between multimedia items captured in a similarity function, or interpretations of the data.

When the indices are structured using a narrow or broad domain ontology the user's search terms can be mapped directly to the proper index term and multimedia items annotated with this term can be retrieved directly.

However, there will always be a remaining gap between the user and the system. Firstly, because the user searches for information which requires indices not foreseen by the annotator. Secondly, because the domain can be so broad that deriving all indices is not feasible. Thirdly, because the required answer can depend on the user's preferences, which are not known beforehand. Thus the information should be found in an interaction with the user, allowing subjective features, subjective similarities, and subjective interpretations to occur in the course of the interaction. The above is illustrated in figure 7.

5.1 Query space

Before looking at the retrieval task itself let us first consider the different goals a user can have. Commonly the following tasks are distinguished: *target search* which is finding one specific multimedia item which the user knows to exist, *category search* which is finding a set of multimedia items from a specific category, and finally *browsing* in which the user just wants to explore and encounter interesting findings.

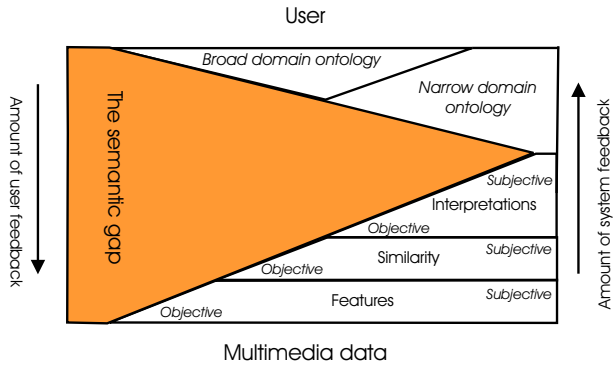

Fig. 7. *Illustration of the semantic gap in interactive retrieval.*

In [1] we have reviewed existing methods for interactive retrieval and structured the description by introducing the query space which is the four-tuple defined as:

The query space Q is the goal dependent 4-tuple $\{I_Q, F_Q, S_Q, Z_Q\}$

In which I_Q is the active set of multimedia items. F_Q is a set of goal dependent features. S_Q is a parameterized similarity function for measuring how much two multimedia items resemble each other. And finally, Z_Q is a set of labels or interpretations, where each element in I_Q has a probability associated with each label.

When a user is interactively accessing a multimedia dataset the system performs a set of five processing steps: initialization, specification, visualization, feedback, and output. Of these five steps the visualization and feedback step form the iterative part of the retrieval task. Using the notion of query space, we can define an interactive search session as follows:

An interactive query session is a sequence of query spaces $\{Q^0, Q^1, ..., Q^{n-1}, Q^n\}$ *where the interaction of the user yields a relevance feedback* RF_i *in every iteration* i *of the session.*

After initialization the user poses a query to the system. For specifying a query many different interaction methodologies have been proposed. A query falls into one of two major categories:

- *exact query*, where the query selects multimedia items if they fulfill a given set of predicates.
- *approximate query* where the system ranks the multimedia items in the dataset with respect to the query.

Within each of the two categories, three subclasses can be defined depending on whether the query relates to the spatial content of the image or video, to the global visual information, or to groups of multimedia items.

After the user has posed the query, the system can generate and visualize the first user defined query space Q_1 from where the interaction starts. Then, in the interactive part of the process, the user gives feedback on the basis of the visualization of the query space. The transition from Q^i to Q^{i+1} materializes the feedback of the user. In a truly successful session Q^n bounds the search goal and the output is precisely the target, the category, or the set of interesting items encountered.

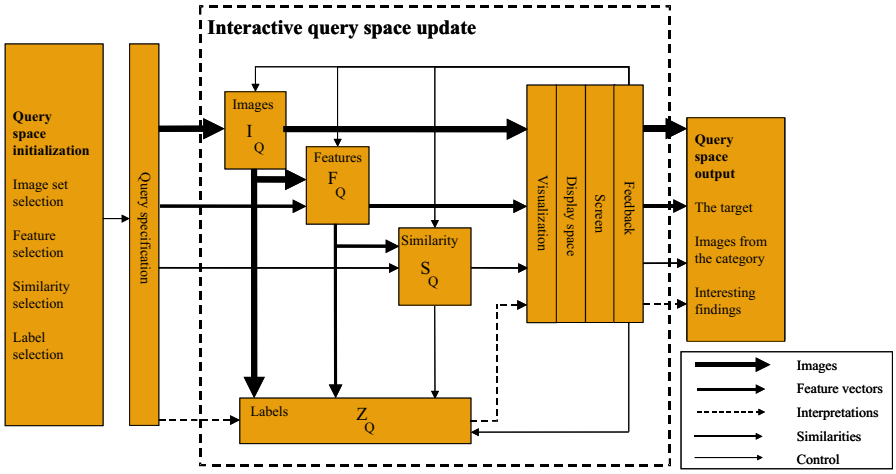

Fig. 8. *The processing steps in content based retrieval [1].*

In literature many methods for updating the query space based on user feedback are presented. They all update one of the components, either the multimedia items, the similarity or the interpretations. Few methods learn from manipulating the features, but these are likely to follow.

When multimedia content is described using features, grouped into genres, and indexed with domain knowledge driven terms using multimedia indexing, it creates a highly complex information space. It is important that this information space be visualized in such a way that the user can navigate through the space and understand its meaning, as otherwise relevance feedback to be used by the system cannot be given. Although general visualization is a rather mature field, the challenge posed by the complex information space generated by multimedia content is at the edge of current possibilities. Considering visualization, a distinction should be made between the display space which should relate to the intrinsic dimensionality of the query space, and the screen space which is always two dimensional. Few retrieval systems have gone beyond trivial visualization. In fact most methods consider screen space only, but some advanced techniques have been published, see [1] for an overview.

Looking at current systems, we observe that there are no systems which take an integral approach to the retrieval problem, i.e. visualizing and interacting with features, similarities, and interpretations rather than one of them in isolation. However, for most search tasks all elements are relevant in reaching the user goals. How the different elements of query space interact in reaching a goal is still an open issue.

6 Conclusion

Indexing and finding multimedia data is a challenging task which requires multiple scientific disciplines to collaborate. Only such an integrated approach can provide the means to bridge the semantic gap. In this paper, we identify difficulties and we have given an overview of our solutions. Clearly a lot of research effort will be needed to bridge the semantic gap and make multimedia data as easy to use as text.

References

1. Smeulders, A., Worring, M., Santini, S., Gupta, A., Jain, R.: Content based image retrieval at the end of the early years. IEEE Transactions on Pattern Analysis and Machine Intelligence **22** (2000) 1349–1380
2. Snoek, C., Worring, M.: Multimodal video indexing: a review of the state-of-the-art. Multimedia Tools and Applications (2002) to appear.
3. Jain, A., Duin, R., Mao, J.: Statistical pattern recognition: a review. IEEE Transactions on Pattern Analysis and Machine Intelligence **22** (2000) 4 – 37
4. Geusebroek, J.M., van den Boomgaard, R., Smeulders, A.W.M., Geerts, H.: Color invariance. IEEE Pattern Analysis and Machine Intelligence **23** (2001) 1338–1350
5. Schreiber, A., Dubbeldam, B., Wielemaker, J., Wielinga, B.: Ontology based photo annotation. IEEE Intelligent Systems (2001) 2–10
6. Bagdanov, A., Worring, M.: Granulometric analysis of document images. In: IEEE press, 16th International Conference on Pattern Recognition, Quebec City. (2002)
7. Geusebroek, J.M., Hoang, M., v. Gemert, J., Worring, M.: Genre-based search through biomedical images. In: IEEE press, 16th International Conference on Pattern Recognition, Quebec City. (2002)
8. Vendrig, J., Worring, M.: Interactive adaptive movie annotation. In: IEEE press, International Conference on Multimedia and Expo, Lausanne. (2002)