# Multimedia information technology and the annotation of video

Arnold W.M Smeulders, University of Amsterdam, ISLA, smeulders@science.uva.nl
Franciska de Jong, TNO/University of Twente, CTIT, fdejong@ewi.utwente.nl
Marcel Worring, University of Amsterdam, MediaMill, worring@science.uva.nl
Sponsored by MultimediaN & DELOS

The state of the art in multimedia information technology has not progressed to the point where a single solution is available to meet all reasonable needs of documentalists and users of video archives. In general, we do not have an optimistic view of the usability of new technology in this domain, but digitization and digital power can be expected to cause a small revolution in the area of video archiving. The volume of data leads to two views of the future: on the pessimistic side, overload of data will cause lack of annotation capacity, and on the optimistic side, there will be enough data from which to learn selected concepts that can be deployed to support automatic annotation. At the threshold of this interesting era, we make an attempt to describe the state of the art in technology. We sample the progress in text, sound, and image processing, as well as in machine learning.

## 1. Multimedia

There are at least three different interpretations of the term *multimedia*. It is interesting to review the interpretations here from the standpoint of meta-data.

The word *multimedia* was first interpreted as everything in the domain of digital information that wasn't text, and then became a synonym for the computerized version of information and knowledge. In a bookstore, the multimedia department will display encyclopedias and interactive courses, and possibly computer games. Traditionally distinct forms of information carrier, such as paper, audio, and tutoring, have converged into a single form: digital productions delivered interactively through a computer window. And, it is obvious that neither this convergence, nor the flexible way in which information is employed, have reached their limit.

For one thing, the form of delivery of digital productions is still very close to the original forms. The web pages of CNN started out having the layout of a newspaper; now they are genuine multimedia pages with various modes of access to multimedia content. Digital encyclopedias have almost the same structure as their paper precursors. Only computer games are distinctly different. This picture supports the view that *new technology is always first accepted in the old idiom*.

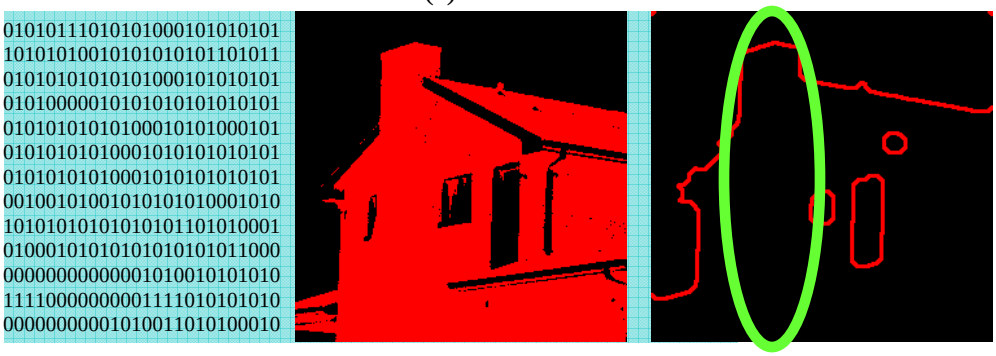
The *convergence* of hitherto different media demands universal solutions for file formats and intellectual property rights, but this will take time. And, convergence and interactivity of media rely heavily on meta-data. They require a detailed description of the content of the message if they are to meet the user's expectation of ready availability even when the context is unknown or open-ended.

The word multimedia can also be read with the emphasis on media. It then alludes to the multiplicity of channels by which we can deliver a message to the public. Multi*media* of the future encompasses both broadcasting and narrowcasting. In fact, the success of television in covering a broad audience has led to more channels. And, in turn, the multitude of channels has led to the need for differentiation and narrow casting if a channel is to survive.

The point we want to make here is that whatever technological advances are made in digital television and internet, more detailed meta-data and knowledge of the target audience are needed to make it possible to match user profiles to the metadata of the archived content. So digital media will increase the need for detailed meta-data.

The dominant type of information in information systems is still of the numerical and coded type. These information systems are successful because the message is directly encoded in the bit patterns. Hence, data processing is equivalent to managing bit patterns. Multimedia information systems are distinctly different. In this context, *multimedia* refers to visual information, audio information, or textual information, whether or not in combination. Multimedia information systems require elaborate information analysis of the content. To the user, digital multimedia information is immediately available (visible, audible, or readable) , and most often also understandable - but not to the machine. The discrepancy between the digital encoding and its semantic interpretation is known as *the semantic gap*.

(a)




(b) (c) (d)

The semantic gap. (a) Orgininal image (b) A small part of an image as perceived by a computer. (c) Display of a most simple approach to distinguish automatically the foreground in the image from its lighter background as an essential step in the interpretation of the image. Note that the result erroneously indicates that the windows and the gutters are part of the background. Note also that the human eye easily restores the proper segmentation result. Humans cannot help but identify a house in spite of the distorted grey values and in spite of the fact they have never seen such a house before. (d) An alternative approach shows the most clear contrast transitions in the image. Where the house is relatively simple to identify by the contrast to the background, the chimney in this example is hardly distinguishable as the foreground figure against the house as the background. The semantic interpretation is easily made for humans whereas it is incredibly hard to come up with a set of general rules (or computer programs) to describe what makes a chimney, simply because the visual evidence is meager at best. Semantic interpretation requires lifelong experience with meaning.

Because of the semantic gap, a completely automatic multimedia analysis cannot be expected. One can wait for high quality and complete coverage before starting to use

automatic aids in annotation, but that will take quite a while. For a long time yet, the performance of automatic annotation when measured against manual annotation quality will appear to be abysmal at best. But in any case, copying the manual annotation process is not the ultimate goal of a computer–assisted search. And hence, manual annotation is not a good performance indicator for machine annotation.

What does matter is whether it is possible to create an effective methodology in which man and machine work together in an integrated way to successfully find a target. In this paper, we review the state-of-the-art in multimedia analysis to show how the latter can contribute to a process for the automatic annotation of video content.

## 2. The challenge

The prime motivation for introducing automation in the generation of metadata is that an all-digital recording process and post-process will enable faster re-use. At the same time, the scope of re-use will be much broader than current practice. And, computer networks will permit extension of the archive with other virtual archives. This requires annotation of a much larger volume of data as well as extending the number of topics to be covered, while at the same time the anticipated response time decreases. In other words, the archive is under pressure from all sides. Automatic analysis is an essential ingredient in meeting present requirements.

We argue that automatic or computer-aided annotation cannot be seen as separate from the work practices in which it will function. It has to be part of a complete process of storing, enriching, and delivering multimedia information. All of these elements will change when the archive becomes all-digital. For digital archives, one cannot expect the flow of items into the archive, nor the exchange of information in a search, to stay the same. The point to decide in a multilateral view is what needs to be done to achieve a proper workflow around the digital archive, or for that matter, an effective digital system around the archivist.

The following aspects of video archiving environments will inevitably change as the result of moving to an all-digital environment.

First, the widening horizon of the archive induces a *perceived loss of accuracy*. In the foreground, a larger part of the archive is more readily available. In the background, resulting from the increased connectivity in the world provided by internet, conflicts between the archival codes of archives that developed in isolation (e.g., thesauri, ontologies) will demand conversion and merging of coding systems. This will inevitably induce a perception of loss of accuracy in the user confronted with code systems different from those to which he or she is accustomed.

Second, where one was already used to heterogeneity in data sources, computer-aided search will emphasize *variety and integration of data types*. The target content can be distinguished by type: visual (stills, photographs, graphics, logo), audio (speech, music, noise), and text (scripts, summaries, transcripts, reviews, letters, instructions, literature), and combined versions of these. As almost all subtypes can be combined with one another, the list of integrated information objects to be analyzed is beyond complete formalization, but will require awareness.

Third, computer-aided systems will stimulate *differentiation in search patterns*. The new search facilities will be well outside the paradigm of key-word search. Interactive systems will allow faster response, leading to an earlier transition from well-posed questions to more open-ended browsing by the user. In addition to precise target search, the user will frequently conduct an open-ended browsing search and use different kinds of interaction and

presentation techniques to view the result. Searching through larger and more heterogeneous, possibly remote, archives requires different search patterns including the acceptance of working with different code systems. Archives will be under pressure to provide better performance, however abstract the initial formulation of the search.

Fourth, computer-aided archival systems will put pressure on the *user's expectation*. As we argue later on, there is no realistic possibility of achieving the same completeness and accuracy in the automatic annotation of an archive as in the manually generated counterpart. But that is precisely what the user will expect, since all information is "in the computer". The question is what to do with that expectation: to combat it, to compromise on accuracy, or to accept automation only when it delivers the same quality. The practice of use will change, and hence inevitably the practice of archiving. The good news is that there is still ample time to prepare for that change.

If the frustration of archivists and users is to be avoided, there are some additional challenges ahead in the development of automatic systems:

1. There is the need to design and implement systems that fit a daily work process.
   Note: the fit with a daily process may seem to have no value, but experience in many other information systems application areas has frequently taught this lesson.
2. To that end, to deliver computerized archiving systems that are fast and accurate in their retrieval results.
   Note: for a computerized system, accuracy in the result is not necessarily the same as accuracy in the annotation.
3. And, to do so with robust methods for automatic understanding of non-ideal data.
   Note: experience with early systems has led to considerable cynicism and misunderstanding of the general applicability, due to the fact that systems have been tested for only a small set of perfect data.

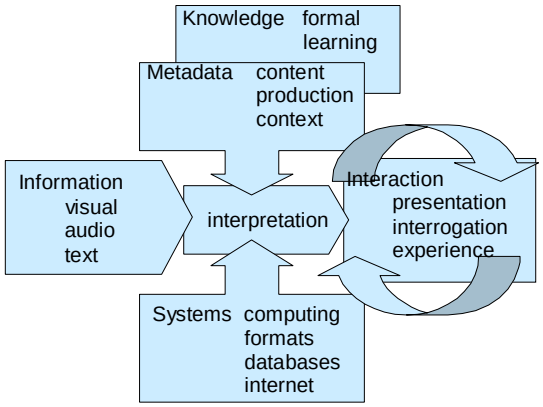
In our view, the practices of users and system designers, as well as of archivists, will change considerably before effective systems are introduced.

## 3. Constituent elements of video archival systems

The interpretation of multimedia requires attention from a wide variety of disciplines, currently usually operating separately. The analysis of visual information is studied in the areas of *image processing* and *computer vision*. The first of these has an emphasis on image in/image out processes, whereas computer vision studies the interpretation of static or dynamic scenes. As well as for speech recognition, audio signals are studied for *music recognition*. *Natural language processing* aims to deliver an interpretation of the content of a text.

By the nature of the information it processes, natural language processing starts from semantically meaningful units, namely words. So, it is no surprise that understanding a multimedia object relies heavily on the success of the interpretation of the linguistic elements, either written or spoken. The latter requires detection of speech and conversion to text as an intermediate step, but still lends itself better to understanding than the visual part. Visual information is so rich in content and variety, even for one single object, that it appears difficult to deal with it using automatic analysis. As a consequence of the difference in progress in these fields, their practices are quite distinct. But whereas they have grown in separation for twenty or thirty years, current progress is fastest when based on interdisciplinary cooperation.

Automatic interpretation of visual, audio, or textual information is greatly helped by detailed understanding of the content when the description is based on ontologies or other formal domain descriptions. Automatic interpretation is also supported if general background knowledge is available on things such as word combinations, pronunciation, faces, shouts, and their admissible variations, such as the morphological variants of words, the variety in visual appearances, and the variations in the background.




Knowledge can be acquired by formalization, but more success has been achieved by learning rules from large datasets. In effect, a general rule of machine learning is: the more specific, the larger, and the more reliable the datasets, the better the result. More importantly, when learning from realistic datasets, the result is also more robust, being able to cope with non-ideal circumstances. Modern natural language processing frequently uses techniques from the area of *information retrieval* to capture the content of the message. And, modern computer vision frequently uses *machine learning* techniques and *statistical pattern recognition* to understand the content of a scene.

When designing real systems, a few aspects of the state of the art in system technology need to be considered.

Proper choice of formats guarantees added value in the ease of exchange as well as in proper storage. Formats cast a long shadow into future as new systems have to adapt to the old formats to be useful. Therefore, the selection of a new format has to be done with care (but even then the predictability which formats will become popular is limited). Databases are useful in not losing information while delivering optimal handling speed. Truly multimedia databases with integrated formal knowledge descriptors of multimedia are a hot topic of research.

Computer-aided video archives demand enormous computing and storage capacity to handle a stream of video data. A text stream is relatively condensed in its semantic content, but learning facts from text streams requires large datasets, which in turn require large computing power. Analysis of the audio signal requires more power; but real time or near real time processing of the visual component is the most demanding. Computing power will continue to be an important consideration in practical video analysis for some time to come. The solution to the storage and computing capacity needed for archiving and learning lies in *grid computing*, internet based distributed processing power.

Interaction is the key to the user and hence to the system. Interaction is still poorly developed. *Interrogation* encompasses solicitation of the search either by specification, browsing, analogy, or by question and answer. Any interaction requires carefully designed

*presentation* of the result, which, in the case of video, requires various kinds of summarization since the screen offers only limited space. The interactive component of systems will be useful only when they become able to remember the preferred behavior as well as the preferred presentation in the interaction *experience* learned by the system from previous sessions. On the threshold of high-speed wireless technology, there is enough opportunity to insert the meta-data at the production site.

## 4. Interacting with video archives

Interaction is an essential ingredient in any video archival system. It can serve both the video archivist in annotating the wealth of information as well as the user accessing the archive. In the future these functions will merge, since a digital archive will eventually learn from the pattern of interaction of the users, as well as from user annotations of the data.

To assist the archivist, the aim is to limit the time needed for the annotation work. The major assumption underlying tools for this purpose is that similar video content is likely to have the same annotation. Hence, after the archivist has provided some initial annotations, the system can provide collections of similar items that have a high probability of having the same annotation. By manually filtering out the small percentage of incorrectly labeled items, the archivist can completely annotate collections of items. This strategy for limiting annotation time is particularly suited for simple bulk annotations. An expert can perform more elaborate annotation better, one at a time.

We turn to the information needs of the user. There are various types of exchange of information, leading to various types of query:

- Query from a controlled vocabulary

In this query mode, the user inputs query terms from the controlled vocabulary used by the archivist for the annotation of the data. In this case, specification of the query should be aided by a visual representation of the metadata model used in annotation. When multimedia analysis tools are employed to automatically index the video with a set of controlled terms from the metadata model, this approach can still be followed, with the essential difference that, in the interaction, both the system and the user should be aware that annotations have an associated probability of correctness.

- Query by keywords or descriptors

It is impossible to foresee all possible annotations on which a user might query the archive. Hence the user should also have the possibility to query on the content of the archive directly. For text this is a simple comparison of the word the user has provided with the words in the document. This is still a feasible approach when the text in the archive is the result of speech recognition from the audio channel, but fuzzy matching techniques have to be used, since errors are frequently found in the speech recognition result. For audio and video data it is clear that one will not query for a specific set of sample or pixel values, as they don't make sense to the user. Descriptors of the data are required, which summarize and emphasize specific characteristics. It is difficult to decide what these should be if the purpose is not known beforehand. Hence, query by descriptors is often limited to rather general descriptors such as pitch value or average volume for audio, and color / texture and motion distributions for video.

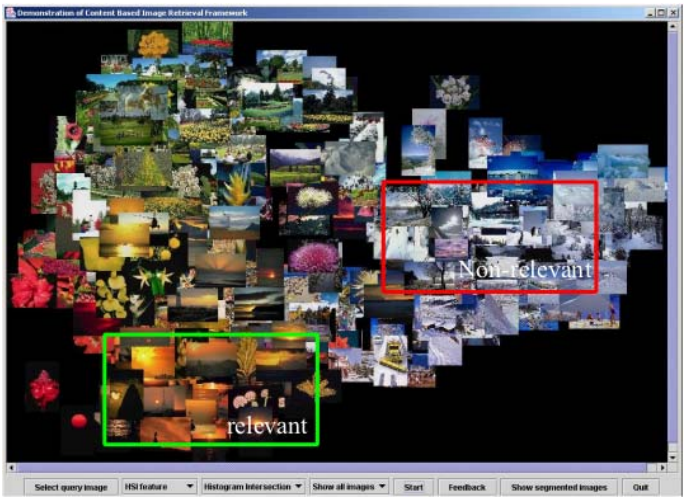
- Query by full text, full audio, or full visual examples

Keywords or descriptors entered by the user provide the system with only limited information. Only in context can such queries lead to the desired information. The computer does not understand the context by itself, nor does it have experience unless programmed, nor

does it have a good feel for purpose. Therefore, computer search profits from more information in the query. One way to achieve this is by giving examples of similar items. So, when the query is an item of full text, computer retrieval has a better chance to be on target. Similarly, several pictures should be presented in a query rather than just one. And it is best in computer search to include counter examples, as they help to convey the intentions of the user much better than just positive examples. The same point also helps in full text retrieval. When texts are included as counter examples in the query, the computer may be able to determine the proper response much more quickly.

For query by example, a distinction should be made between external examples brought in by the user and internal examples where the user has selected an item from the database. When the example is external, in practice the query example is not annotated, so the system can only search for similar items on the basis of the content descriptors described above. When the example is internal, similarity can also be based on the annotations of the items.

In practice the user will not get the answer directly from one of the above query types, but will engage in an interactive session with the system where advanced visualization and relevance feedback from the user are iteratively used to bring the user closer to the desired information. Ideally, the system is actively participating in finding the best solution by posing the most informative questions or showing the most informative results to the user.




Example of an advanced visualization tool where the user gives feedback to the system by indicating relevant and non-relevant items.

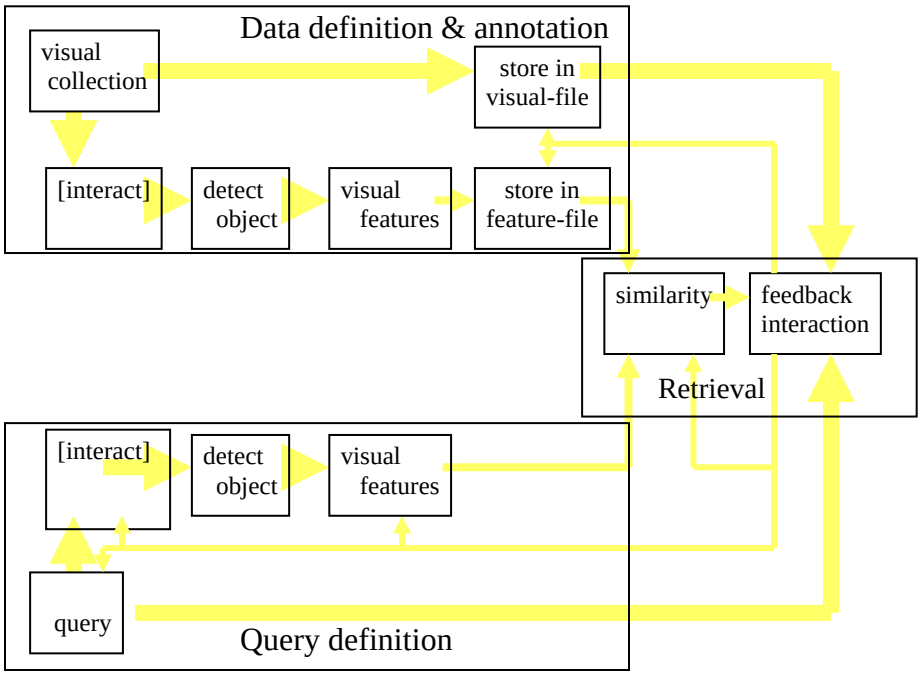
Interactivity poses heavy demands on the computing, storage, and display capacity of the system. Users want immediate feedback on their queries, but this might require computing a large set of relevant descriptors if external examples are used, and then requires comparing the descriptors of all elements in the dataset with the query. Advanced database techniques are required to limit the search. In addition, interactive search stretches the functionality of the presentation devices to the limit. Nevertheless, interactivity compensates for the inability of the computer to take account of context. In a full interaction scheme, not only the query may be modified but also what is to be considered similar, and what are to be considered good examples and counter examples. By using *relevance feedback* and *visual presentation* of the best results (see the figure), current content-based retrieval systems only scratch the surface of what is to be expected in the near future.

## 5. Progress in multimedia analysis

In this section we review the state of the art in multimedia information analysis disciplines: computer vision, text processing, and audio processing, followed by interaction and machine learning.

*Computer vision* started in the sixties with occasional pictures of space and medical images. Processing was concentrated on large computers. In the early nineties, personal computers became sufficiently powerful to hold a digital image, popularizing picture computation. Digital storage of pictures, and family communication with pictures through the internet, followed later. Digital image sensors are now found in many devices. It is estimated that more than half of all new cameras are digital as well as a quarter of all family video devices. Hence, computer vision has developed from an esoteric science to a necessary ingredient of the information society in just 15 years.

An essential step forward was the recognition that *precise segmentation* of an object in the foreground against the background is unattainable. There is evidence that even humans break down images into named objects only when necessary. To identify a scene, it may be sufficient to recognize just a few details. A typical example is an orange circle somewhere in the middle of a picture signifying a setting sun . Another typical example relating to texture is a patch of striped skin immediately identifying the presence of a tiger or a zebra. And a typical example of a characteristic spatial arrangement is a face. Now it can be understood why Hawaiian sunsets, faces, and tigers are frequently used in demonstrations of video search systems. But, it requires more progress to develop their success into a general capability of recognizing items in any image [Fergus 2003].




Sketch of the flow of information in a system for interactive visual annotation and query by example.

In computer vision, *large volumes of data* have only recently become an issue. Until the mid nineties, computer vision programs were tested on fewer than 100 images as opposed to the thousands being used today. As a byproduct, test data is no longer perfect. Hence, computer programs are more robust, and are better able to cope with many sources of variation. Nevertheless, for video archives, still larger test collections are needed since archives typically contain millions of single frames.

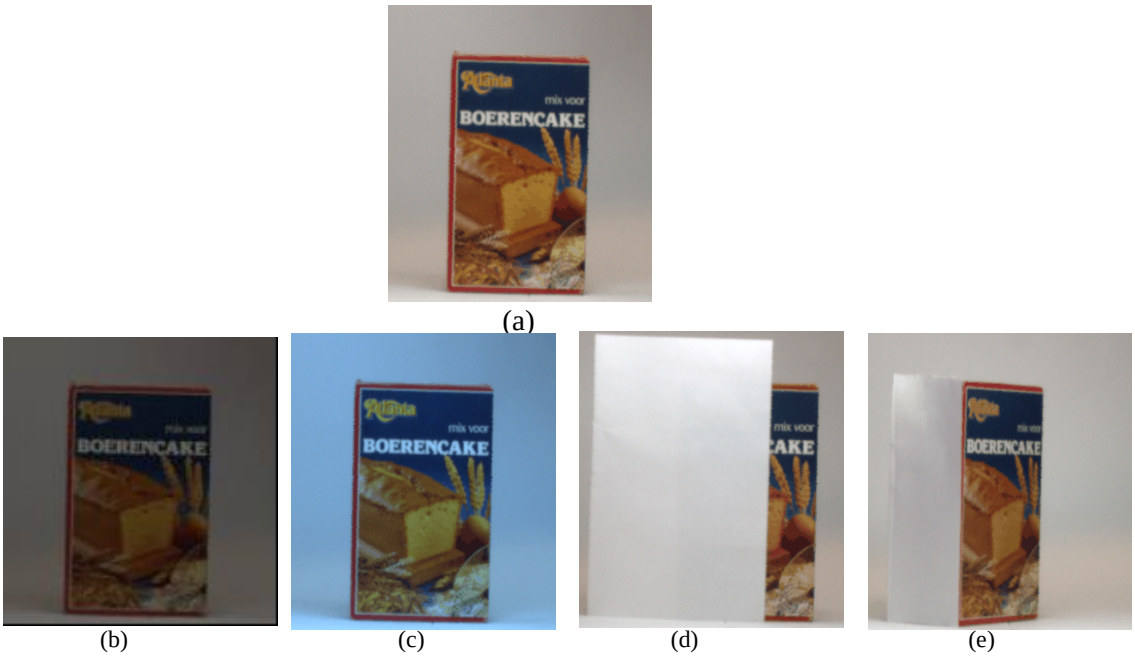
Computer vision starts with *good features*, capable of describing the semantics of the scene and the object, and of ignoring the irrelevant circumstances of the recording. An object comes in a million different appearances. This is known as the sensory gap, which comes on top of the semantic gap discussed before. Good features are invariant to accidental conditions of the recording, while they accurately record the semantically relevant differences in the objects. [Smeulders 2000, Schmid 2004].

Language is the most direct carrier of semantic content. Hence, for the generation of metadata, there is always a strong interest in the deployment of linguistic material, such as text and speech, accompanying media content. The role of speech recognition is the focus of the next section. Here we describe the potential contribution from the field of natural language processing (NLP) for the processing of textual elements in media archives.

There are various ways in which video archiving can benefit from natural language processing. In order to describe the various roles, we should distinguish between textual material (such as subtitle files for productions in a foreign language) that are part of the broadcast item proper, manually generated transcripts and the like, collateral texts (such as reviews, scripts, and other production files), and related sources such as newspaper articles.

The role of natural language processing in the processing of subtitle files and transcripts is straightforward. In the current state of affairs, it may contribute to comprehension of the content of the text. Since textual elements have a link to the temporal structure of the video, they can be used to generate a time-coded index that allows for the

searching of video fragments. As is common practice in natural language processing these days, reducing words to the stem, stop-word removal, and disambiguation are techniques to enhance the generation of indices, which usually improves the result depending on the nature of the text. Cross-language retrieval, i.e., searching in language A for information in language B, can be offered when translation functionality is built in [DeJong 2000]. These are examples of the language processing facilities that have been proved to be effective by information retrieval research.




(a) (b) (c) (d) (e)

The sensory gap in computer vision: Different versions of the appearance of the single object in (a) are easily recognized by humans whether they are recorded in the dark (b), in blue light (c), in occlusion (d), or under a different viewing angle (e). Good, invariant features describing the object should be capable of ruling out the unwanted variations in the scene while retaining the ability to discriminate among truly different objects.

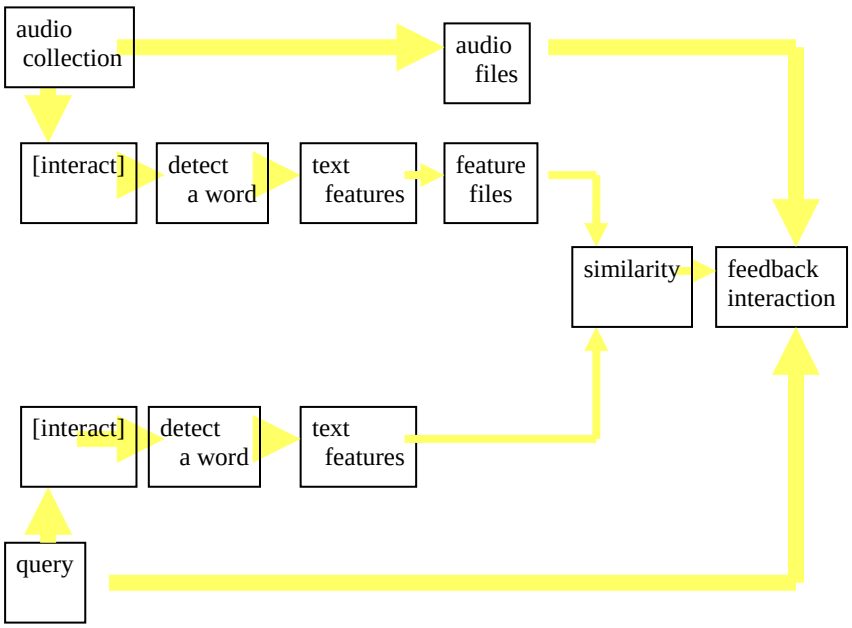
Examples of other text processing techniques that can be employed for more advanced access to the content of media archives are: automatic topic classification, automatic topic segmentation, automatic clustering of documents, automatic summarization, named entity recognition, and information extraction. Many of these techniques rely heavily on statistical language models.

The recent application of domain models for search tasks, such as ontologies and thesauri, is expected to be of importance in the media domain as well. This is not just for a mere conceptual search. The use of domain models is also important to enable cross-media search, since interest in this is increasing for the linking of archives and collections that have been functioning in isolation for decades.

*Audio processing* to support automated audiovisual access to the content has been a topic of active study since the early nineties. Contrary to what is often assumed, speech recognition is not a (nearly) solved problem. The task can be viewed as the conversion of recorded speech into a textual transcription. The confusion about the difficulty of speech processing is that there are many very different tasks of varying complexity that are all labeled as speech recognition. The performance and functionality of speech technologies that have been in existence for some time, e.g., spoken dialogue systems and dictation technology, is of little

use in automated video annotation. Dialogue systems typically operate online but in a narrow domain. Dictation requires training of speaker characteristics, and would therefore be applicable for rapid subtitling of news broadcasts, but not for general video speech understanding.

In the context of audio access, the main technology of interest is speech transcription. In principle, transcription technology detects which words were spoken in what order and at what point in time. Because of the time information, transcripts are the basis for generating a time-coded index, and therefore provide a good basis for spoken document retrieval: the search of audio or video fragments on the basis of the spoken content [Renals, 2005].




Sketch of the flow in querying by audio example.

The models applied in speech transcription have to capture various aspects: recurring variations in the acoustics of speech, the set of sounds for a specific language, the combinations of sounds (syllables, words), and the possible combinations of words. The latter requires large amounts of textual training data and, as a consequence, the volume of the available sets determines the success of the statistical language models. The more variation that is absorbed in the model, the better can the proper word combinations be sieved out of all candidate word combinations suggested by the acoustic models.

Current focus in the development of transcription technology is on tuning the existing methods to more difficult domains and conditions, such as spontaneous speech, non-native speakers, and spoken content that is less dense than news.

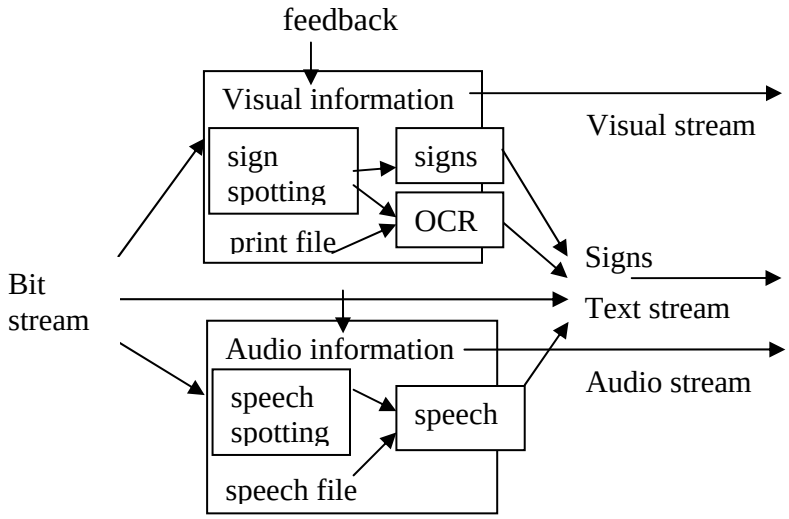
Another ingredient for content-based search is *machine learning* and the *in promptu* version of it: *interaction*. Interaction has absorbed user relevance feedback, interactive visualization of the results of a query, and adaptable similarity measures [Worring 2001], yet a major advance in tools and machine power is required to benefit fully from the interaction.

The application of machine learning techniques overcomes the incidental variations within a concept. A successful line of a machine learning concepts is to combine many weakly performing classifiers into stronger ones. All of these approaches have brought a substantial improvement in the capabilities of machine learners to recognize concepts.

The situation is improving all the time in all the above respects, except in terms of the amount of data. More data demands more effort in annotation, until the point at which the data set gets so big that annotation is no longer feasible. Annotating thousands and eventually hundreds of thousands of pictures is hard to achieve. Where the machine power to do increasing numbers of computations is available, the manpower for annotation will become the bottleneck.

## 6. Recognition

In this paper, we make a distinction between visual information, audio information, and textual information. In this section we discuss recognition, defined as the unambiguous, context-free denotation of signs. In all practical circumstances, the visual representation *A* refers to the first letter in the alphabet, so *A* is recognized rather than interpreted.




Textual information may take a visual form when it is printed on paper or held in a pdf-file. It requires a computer function known under the generic name of optical character recognition, OCR, to convert the printed version of a text to a stream of characters. OCR is in wide-use, and is built in to many search programs, with the result that paper scans and texts in computer files are now easily accessible. Depending on the quality of the scan data, the quality of the method of the OCR program, and its ability to recognize the font of the text, OCR will deliver near-perfect results. However, a guarantee that all information is understood correctly is hard to give, not at the level of single characters, not at the level of words, and - most of all – not at the proper interpretation of the text block sequence. For example, it requires only a slight misinterpretation to miss a footnote and its proper position in the text. OCR programs rely heavily on built in knowledge of the structure of texts, conventions behind letters, and the structure of books. OCR programs for print in languages in which individual characters are frequently annotated with accents (as in Turkish), or in which characters change form as part of a word (as in Arabic), or in which there are many characters and compound characters (as in Chinese) will be much harder to decode. Whereas the conversion of facsimiles of texts to character codes is nearly perfect for standard text in mainstream languages, there is quite some ground to cover in roughly scanned text or when the font, script, or language is non-standard.

Much harder than recognizing texts is to spot the presence of text in a photograph or a video stream. Whereas it is hard for humans to overlook a text in an image, a computer must recognize the distinct pattern of stripes that are a sign of language. Text can be from

different sources: it can be added to the picture in the later stages of production (for example, captions and headers). Such texts are relatively easy to detect, as they will appear in one style and font, usually at a standard position in the screen. Video edits indicating the topic usually appear somewhere in the lowest part of the screen, but not at the bottom. A basic strategy for text spotting is to do a trial run with an OCR program and to see whether it has detected some readable text with some degree of reliability. In the more general case, where text is an integral part of the picture, text is much harder to detect, as there is no information available on language, script, font, depicted size to be expected, nor on the distortion of the font due to the arbitrary viewpoint of the camera. Arbitrary camera positions depict characters in the scene in a skewed view, ruling out the use of standard OCR to read the script. The text on a billboard, a script on a t-shirt, or a banner at a demonstration often carry most of the message of a photograph, but this remains invisible to a computer interpretation of the picture.

For a better understanding of speech recognition it is crucial to distinguish between the various processing steps. Audio detection is relatively easy. The next step is audio segmentation to identify the audio segments where speech recognition is to be applied. Assuming that the language is known, spoken audio segments can then be input to a transcription module.

State of the art performance in broadcast news transcription is around 20% word error rate in international benchmarks. Word error rate depends on speaker and speaking style, ranging from 1-2% to over 50%. Recognition error rates for content words are better than for function words. Estimated retrieval performance with current word error figures: average precision is above 50%, which is sufficient for audio fragment retrieval. Comparable results have been reported for major languages (English, French, Mandarin, German, Italian, Spanish), but for several languages the development of this technology is and will remain lagging behind.

An alternative to the ‘full transcription’ approach to spoken document retrieval is word spotting: searching on the basis of the sound pattern of terms. This approach is feasible only for limited numbers of search terms.

Since signs are well-defined symbols that do not depend on context, given a data set that is representative of the quality of the data and is large enough, it is possible to obtain task-independent performance figures on the recognition of signs. For text spotting and text recognition, a modern recognizer will generate a figure indicating the certainty of detection. An example of the specification of such a certainty figure could be: *for a detection rate of 95% the recognizer will falsely detect 10% of all signs*. With sophisticated recognizers, alternative interpretations of each detected character are presented together with their certainty. This presentation of certainty of recognition in combination with alternatives is an essential component of robust recognizers, in spite of the fact that they occasionally introduce confusion, of course.

So far in this paper, the recognition of visual and audio data and conversion to text has been conceived as a forward process of interpretation, that is without feedback and interaction. But such a system can provide only the skeleton for the definitive system because feedback from the interpretation is probably essential to recognition. When the conversion to text yields nonsense, are we spotting text at all? Is the OCR properly tuned to detect the peculiarities of the script? Are we analyzing on the basis of the right language; is it Japanese rather than Chinese? From the examples it is clear that feedback from interpretation is important in human recognition, and so it is with machines, especially as the demands on quality rise. And hence, the availability of certainty and alternatives is important in recognition for visual and audio signals alike.

## 7. Interpretation

In this section we discuss the possibilities and problems of interpretation. We focus on key concepts determining the performance of automatic interpretation: the semantic gap, narrow versus broad domains, the keyword funnel, and similarity.

An unavoidable bottleneck in automatic interpretation is the *semantic gap*. As discussed earlier, this is the discrepancy between the digital encoding and its semantic interpretation. What is immediate and practically flawless for humans is very hard for machines to decide. How can the purpose of an object be derived from its appearance? To what class does a visual object or subject belong? And, what part of the picture makes up one entity in reality? A machine has no means of telling, and no experience of, what part of the image corresponds to one object in the real world. There is simply no general rule telling it how objects appear. One can only discriminate objects in a scene by learning them one by one in the course of one's life, by bumping into them, and later by identifying them as moving coherently on the retina. Also hampered by the sensory gap referred to earlier, computer vision will not solve that problem without learning to recognize them one by one. And that will take a while.

At the current state of the art, it is important to grasp the difference between *broad* and *narrow search domains.* In a narrow domain, the data set has well-defined proportions, whereas a broad domain can be described only in general, associative terms. The broadest domain around is the set of all information accessible through the Internet. An example of a very narrow domain is a logo recorded by scanning a document: the view is frontal and the illumination is perfect.

When searching for logos in a general video, for example to record the exposure time of the Coca Cola logo during the Super Bowl, the domain is no longer narrow. The image of the logo is distorted by a skew viewing angle, partially occluded from sight, with changing illumination and in shadow, and with varying magnification. So the repertoire of images admissible as countable Coca Cola logos is magnified enormously. At least 100 easily detectable viewing angles, a similar number of realistic illumination patterns, 1000 different ways to occlude the logo and still recognize it, and 10 different magnifications, yielding some millions of views of one well-defined and simple object. In general, in automatic analysis, the chances of success are better in systems working in narrower domains.

Consider the following list of narrow versus broad visual domains.

| | |
|---|---|
| Trademark detection in letters | standard camera, standard illumination recognition success rate: reasonable |
| Station identification in video (edits) | standard camera, noisy background recognition success rate: good |
| Trademark search in stadium | skew view, shadow, occlusion, fixed objects recognition success rate: state of the art |
| Face detection | frontal view, well-determined object class recognition success: good depends on pose. |
| VIP identification | well-lit conditions, skew view, abundant data of widely varying class; hard problem. |
| Face identification | any condition, very large class & minute *visual* differences among the members of the class: extremely hard problem. |
| Object retrieval (this train) | any recording condition, relatively narrow class, success depends on learned properties, state of the art. |
| Object class retrieval (a train) | for most object classes poorly defined: a broad class. Poor detectors, useful when combined with other ones. |

Topics that are difficult and those that are no longer difficult at the current state of the art of computer vision

The distinction between broad versus narrow domains also exists for speech recognition tasks. Consider the following examples:

| | |
|---|---|
| Speaker identification | feasible with studio quality, prepared speech, known acoustic profile of speaker, quiet background, standardized intonation |
| Speaker recognition | requires classification of acoustic profile and language use; allows speaker tracking; |
| Large vocabulary recognition | requires language models with broad lexical coverage; poorly-defined background |
| Recognition of read vs. spontaneous speech | possibly overlapping speech makes recognition hard |
| Speaker independent recognition | unknown speakers; training for acoustic profiles not feasible |
| Distorted voice | dialects, non-native speakers, covert speech |
| Music detection vs. classification | complex rhythms, quiet background |

At the current state of the art in audio processing, what is relatively easy to process and what is not.

As is well known among archivists, the reduction of a video to keywords and key features implies a severe information reduction in the message, implemented at a time when the archival codes had to be small. This is the *key-word* or *key-feature funnel*. In computerized systems there is no real need to go for the minimal set of features. In the absence of an automatic understanding of context, larger sets of features will carry information about the context, which is implicit in manual search.

In the same way, the recognition of similarities is almost automatic to humans. For computers, however, similarity is a mystery until it is fully specified. In fact, *similarity* is a complex notion requiring detailed analysis. A few major differences in similarity are indicated in the following table. The degree and measure of similarity is an essential part of the query definition.

| | | |
|---|---|---|
| literal similarity<br>literal<br>perceptual similarity | nearly identical appearance | <br>same station logo<br>same painting |
| object / subject similarity<br>same person / picture<br>same story | similar regardless appearance | <br>Bill Clinton<br>High jacking flight 203 |
| genre<br>the same subgenre<br>the same genre | same class | <br>soccer, weather, dialogue<br>sports, game show |
| semantically similar<br>the same logical unit<br>the same topic | identical meaning | <br>anchor presents highlights<br>politicians discuss |

Types of similarity important in computer-aided search.

For all media types, literal or nearly-literal computerized search is solved. Genres come second, well before object similarities. For visual and audio, object and subject similarity lags behind because of the huge variety possible in the appearance of an object. Obviously, semantic similarity is currently the hardest, but context may provide some clue here.

## 8. Discussion

At the end of this journey through the landscape of multimedia information analysis, we summarize the main issues.

The prime motivation for introducing automation in the generation of metadata is that an all-digital recording process and post-process will enable faster re-use. Automatic analysis is an essential factor in meeting present requirements.

Our view is that new technology is always first accepted in the old idiom. Computer-aided systems should not strive for a completely automatic imitation of the current manual process, nor should they strive towards a system designed in splendid isolation, since both these approaches will yield unworkable methods. We put forward the importance of understanding some of the peculiarities of the current methods as well as the importance of current machine performance in designing a reasonable process.

Whereas humans make an instant and precise semantic assessment of a scene, machines cannot and will not be able to do so in the foreseeable future - neither for visual information nor for audio information. Text information may stand some chance of automatic annotation provided it has been acquired as text and not from visual or audio information. It will be a long time before machine annotation achieves precision or perfection. And, as we have argued above, since machines lack insight into context, it is essential that the computer analysis of multimedia is broad. Hence, their analysis may be sloppy on individual items while their identification of the target may still be precise. This is a radical move away from the current practice where sloppy indices are a nuisance.

There are enough signs that computer-aided handling of video will bring annotation and search much closer to each other than the current practice. Whereas annotation is now in the hands of the experts and search in the hands of the users, annotation is likely to differentiate in levels of accuracy, from instant annotation by users supplemented by sloppy probabilistic annotation by machines, to precision annotation by experts. Interactive search may involve ad-hoc annotation and ad-hoc machine learning. In the new archive, a mark of quality for each annotated item is an important asset.

A long-term goal in querying is a system that can reconstruct the information needs of the user by building up experience with users, by semantic understanding of the content of the archive, and by generating the most informative question to enable the machine to learn from the user. Another long-term goal is to present the information with high density in a natural and bilateral dialogue with the user. Research is being done on almost all topics, but, as yet, in isolation. There will be room for improvement in video handling systems for many years to come, and developers in several IT domains will be keen to collaborate with media archives. We shall discuss two highly promising areas, topic clustering and video retrieval, both organized around international benchmark events, from which interesting results can already be obtained:.

As mentioned in section 5, topic clustering is an information access to its content that organizes news items in clusters corresponding to the topics discussed. The result can be regarded as a partition of the corpus in which each news item is assigned to a ‘dossier’ representing a topic. The state-of-the-art is demonstrated at the annual Topic Detection and Tracking meeting, a benchmark event organized by the National Institute of Standards and Technology, NIST [Wayne 2000].

In combination with automatic classification, topic clustering can help to organize large archives, and to build tools that allow users to browse through information dossiers containing items in a variety of formats. For example, all newspaper articles, TV news items, and radio broadcasts on the eruption of a particular volcano.

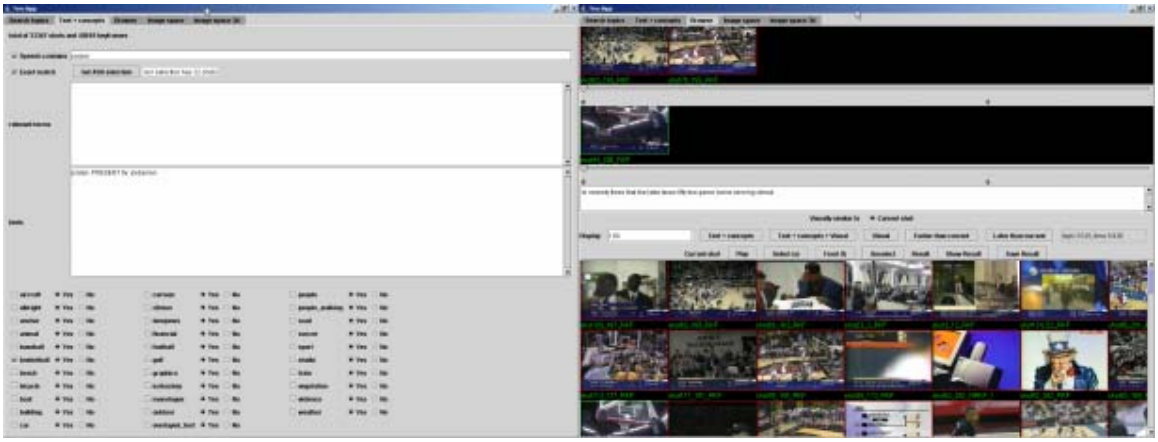
The technology is applicable to textual archives and dynamic news streams, but also to transcribed speech. A technique recently taken up in the Topic Detection and Tracking

evaluation program is *hierarchical* topic clustering. The aim is to organize a collection of unstructured news data in a structure that reflects the topics discussed, ranging from rather coarse category-like nodes to fine singular events. With this technique, browsing can be supported at levels of granularity that can be tuned to user needs [Trieschnigg 2005].

The state of the art in video retrieval is best represented by the Video benchmark TRECVID, also organized by the National Institute of Standards. This benchmark evaluates various components required for retrieval of video shots from an archive of 184 hours of news video. Tasks range from shot segmentation to story segmentation, concept detection, interactive search, and automatic search. Teams from around the world submit their detection and retrieval results. These are then manually judged by a set of experts providing the underlying facts against which the individual systems and approaches can be compared.

In typical modern systems competing in TRECVID, several methodologies are employed to build basic detectors. Natural language processing is used to read in the text stream and Video OCR to read overlay text, and these are coupled with automatic speech recognition, identification of a very limited number of speakers, style recognition, face detection (but no face recognition as it performs very poorly as yet), shot length, camera distance, weak segmentation using invariant color descriptors, and other techniques [Snoek 2004]. They are used in turn to derive higher-level concept detectors such as *boat/ship, Bill Clinton, Madeleine Albright, people walking or running,* and *physical violence*.

The reliability of the various basic detectors ranges from poor to high quality. In spite of their sometimes-weak performance, they are all of help in searching a digital video archive. Recent additions to the basic and high-level detectors include the detection of concepts by machine learning from large data sets, and a set of detectors ordered in an ontology of visual key elements (in addition to the established ontologies for text).



The interface for the interactive retrieval system [Snoek 2004]. The lefthand screen is used to define a query based on keywords and concepts. Results are presented on the righthand screen and can be used as visual examples in query by example.

To evaluate search performance, TRECVID has defined an interactive search task based on 25 topics. Users were given 15 minutes to find as many relevant items as possible. Typical examples of a search include *people walking with their dogs, congressman Henry Hyde, people moving a stretcher, Benjamin Netanyahu,* and *moving bicycles*. To determine the performance, for each search NIST considers the precision and recall figures of the best 100 results returned by the system. The precision is defined as the number of correct items divided by 100, and the recall as the number of correct items divided by the total number of relevant items. For a top ranking performance [Snoek 2004], in 15 minutes, an expert user,

combining keyword search and query by similarity with a set of 32 automatically detected high-level concepts, can yield the following scores:

| Topic | precision | Recall |
|---|---|---|
| *people walking with their dogs* | 28% | 42% |
| *tennis player contacting the ball* | 10% | 19% |
| *moving bicycles* | 41% | 59% |
| *Bill Clinton with at least part of a US flag visible* | 35% | 36% |

Automatic video annotation is still a difficult problem, and varies between very poor on some topics to reasonable on others. And, not all topics of this year's competition may be equally relevant in practice, but the progress made *each year* is considerable. Even poor quality descriptors help in automatic annotation, and they will improve through learning from larger data sets. When automated analysis is combined with interaction, a useful new search paradigm will emerge.

In this paper we have indicated where progress is to be expected in automated analysis, and which solutions are much further away. We have done so at the risk of being ridiculed by our fellow researchers for painting a too simplistic view. Nevertheless, as is always the case at the frontiers of technology, you often gets answers to the questions that you haven't asked. The answer is more complicated than desirable, but this is inevitable as the leading edge of progress follows its own internal logic.

Nevertheless, we hope we have been able to whet your appetite for the future that computer-aided annotation will bring. We look forward to communicating with you on where our vision of the modern archive needs amendment.

## 9. References


[Fergus 2003] R. Fergus, P. Perona, A. Zissermann: Object class recognition by unsupervised scale invariant learning. Proc. CVPR 2003, IEEE Press.

[DeJong 2000] F.M.G. de Jong, J.-L. Gauvain, D. Hiemstra & K. Netter. Language-Based Multimedia Information Retrieval. In: *Proceedings RIAO 2000: Content-Based Multimedia Information Access*, Paris, April 2000, ISBN 2-905450-07-X, 713-722.

[NIST] TREC Video retrieval evaluation, 2001-2004. http://www-nlpir.nist.gov/projects/trecvid/

[Renals 2005] S. Renals, J. Goldman, F.M.G. de Jong et. al: Accessing the spoken word**,** to appear in: International Journal on Digital Libraries.

[Schmid 2002] K. Mikolajczyk, C. Schmid: Scale and affine invariant interest point detectors. Nt. Journ. Comp. Vis 63 – 86, 2004.

[Snoek 2004] C.G.M. Snoek, M. Worring, J.M. Geusebroek, D.C. Koelma, and F.J. Seinstra: The MediaMill TRECVID 2004 Semantic Video Search Engine. In: Proceedings of the 13th Text Retrieval Conference (TREC), Gaithersburg, USA, November 2004.

[Smeulders 2000] Content-based image retrieval at the end of the early years, IEEE transactions PAMI, 1349 – 1380, 2000.

[Trieschnigg 2005] D. Trieschnigg, W. Kraaij: Hierarchical topic detection in large digital news archives. In Proceedings of the 5th Dutch Belgian Information Retrieval workshop (DIR), 2005.

[Wayne 2000] C. Wayne: Multilingual Topic Detection and Tracking: Successful Research Enabled by Corpora and Evaluation. In: Proceedings of Language Resources and Evaluation Conference (LREC), pages 1487-1494, 2000.

[Worring 2001] M. Worring, A. Bagdanov, J.C. VanGemert, J.-M. Geusebroek, Minh A. Hoang, T. Augustus, G. Schreiber, C.G.M. Snoek, J. Vendrig, J. Wielemaker, A.W. M. Smeulders: Interactive Indexing and Retrieval of Multimedia Content, Proc. SOFSEM, Springer-Verlag LNCS 2540, 135--148, 2002, http://www.science.uva.nl/~mark/pub/2002/WorringSofSem02.pdf}

Arnold W.M. Smeulders is full Professor of Multimedia Information at the University of Amsterdam. He heads the ISIS research group of 25 concentrating on theory, practice, and implementation of multimedia information analysis with an emphasis on computer vision, machine learning, and semantic annotation. He is scientific director of the MultimediaN national initiative for multimedia research and application in the Netherlands.

Franciska de Jong is full Professor of Language Technology at the University of Twente. She is also affiliated to TNO-TPD in Delft. Her main research interest is in the field of multimedia indexing, semantic access, cross language retrieval, and the access to the content of spoken audio archives. She is frequently involved in international program committees, expert groups, and review panels, and has initiated a number of EU projects.

Marcel Worring is associate Professor at the University of Amsterdam. His research interests are in semi-automatic video indexing and retrieval. He is co-founder of MediaMill, an application center for multimedia solutions. He has developed several demonstrators for multimedia applications catering for different needs and involving innovative methodologies.