

# Multimedia information technology and annotation of video

Arnold W.M Smeulders, University of Amsterdam, ISLA, [smeulders@science.uva.nl](mailto:smeulders@science.uva.nl)

Franciska de Jong, TNO/University of Twente, CTIT, [fdejong@ewi.utwente.nl](mailto:fdejong@ewi.utwente.nl)

Marcel Worring, University of Amsterdam, MediaMill, [worring@science.uva.nl](mailto:worring@science.uva.nl)

Sponsored by MultimediaN & DELOS

The state of the art in multimedia information technology has not progressed to the point that a single solution is available to all reasonable needs from documentalists and users of video archives. In general, we do not have an optimistic view on the usability of new technology in this domain, but in the area of video archiving quite a few tables will turn enabled by the digitization and digital power. 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 to learn selected concepts from, which 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 machine learning.

## 1. Multimedia

There are at least three different interpretations of *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 became then a synonym for the computerized version of information and knowledge. In a bookstore, the multimedia department will display encyclopedias and interactive courses, possibly computer games. Traditionally separated forms of information carriers like paper, audio, and tutoring have converged into a single one: digital productions interactively delivered through a computer window. And, it is obvious that neither the convergence nor the flexibility in which information is employed have reached their end.

For one, the form of delivery in 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 mode of entry 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*.

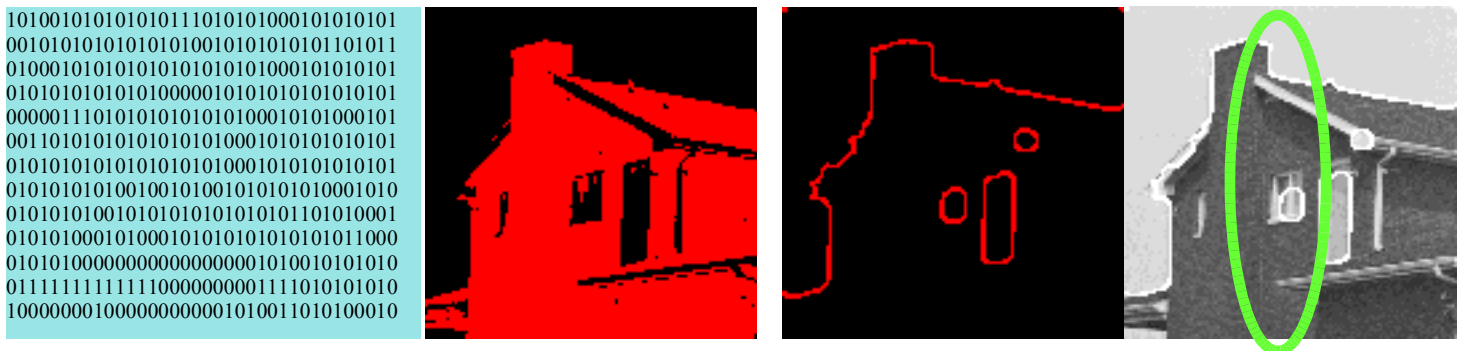
The *convergence* of hitherto different media demands universal solutions for file formats and intellectual property rights but such is a matter of time. And, convergence and interactivity of media rely heavily on meta-data. They require a detailed description of the content of the message to meet the 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. *Multimedia* of the future encompasses both broadcasting as well as 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 for a channel to survive.

The point we want to make here is that whatever the technological advances are in digital television and internet, more detailed metadata and knowledge of the target

audience is needed to be able to match user profiles to the metadata of the archived content. So, digital media will increase the need of detailed meta-data.

The dominant type of information in information systems still is 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 equal to managing bit patterns. Multimedia information systems are distinctively different. In this context, *multimedia* refers to visual information, audio information or textual information, either or not in combination. They require elaborate information analysis of the content. Digital multimedia information is immediately visible, audible or readable to the user, 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*.



The semantic gap (a) A small part of an image as perceived by a computer. (b) Display of a most simple approach to distinguish automatically the darker foreground in the image from its lighter background as an essential step in the interpretation of the image. Note that the result erroneously indicates the windows and the gutters are part of the background. Note also that the human eye easily restores the proper segmentation result. Humans cannot escape the identification of a house in spite of the distorted grey values and in spite of the fact they have never seen such a house before. (c) An alternative approach shows the most salient local transitions in the image usually but by far not always indicating where the foreground – background transitions in the image are. (d) Overlay of the second approach over the original gray tone image. Had the house be made of grey stone, then the result in both approaches would have been much more difficult to achieve. Consider for example the chimney as the foreground figure and 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 program) to describe what makes a chimney simply because the visual evidence is meager at best. Semantic interpretation requires a 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 long the performance of automatic annotation when measured against manual annotation quality will appear to be lousy at best. But in the end, 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 an integrated work process of man and machine exists where one can effectively find a target. In this paper, we review the state-of-the-art in multimedia analysis for the purpose of fitting it in a process of automatic annotation for video content.

## 2. The challenge

The prime motivation for introducing automation in the metadata generation is that an all-digital recording process and post-process will command faster reuse. At the same time, the scope of reuse 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 more topics to cover, 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 to meet present requirements.

We put forward 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 to stay the same nor will the exchange of information in a search. 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 when moving to an all digital environment.

First, the widening horizon of the archive induces at a *perceived loss of accuracy*. In the foreground, a larger part of the archive is more readily available. In the background by the increased connectivity in the world by internet, conflicting archival codes between archives that grew in separation (e.g., thesauri, ontologies) will demand conversion and merging of coding systems. This can only happen at the expense of a perceived loss of accuracy of the user confronted with other than the usual code systems.

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 thereof. 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 system 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 and from there earlier access from well-posed questions to more open-ended browsing by the user. Next 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 require different search patterns including acceptance of working with different code systems. Archives will be under pressure of better performance, however abstract the search is formulated at the start.

Fourth, computer-aided archival systems will put pressure on the *user's expectation*. As we argue in the sequel, automatic generation of an archive that is equally complete and accurate as the manually generated counterpart is out of the question. But the user may expect precisely that, as all information is "in the computer". The question is what to do with that expectation: to combat it, to give in on accuracy or to only accept automation 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 anticipate that change.

To avoid frustration with archivists or users, there are some additional challenges ahead in the development of automatic systems:

1. There is the need to design and implement systems which fit a daily work process.  
Note that the fit with a daily process may seem a void addition but experience in many other application areas of information systems has sadly learned this lesson.
2. To that end, deliver computerized archiving systems that are fast and accurate in their retrieval results.  
Note for a computerized system the 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.  
Experience with early systems has lead to considerable cynicism and misunderstandings of the general applicability, due to the fact that systems have only been tested for a small set of perfect data.

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

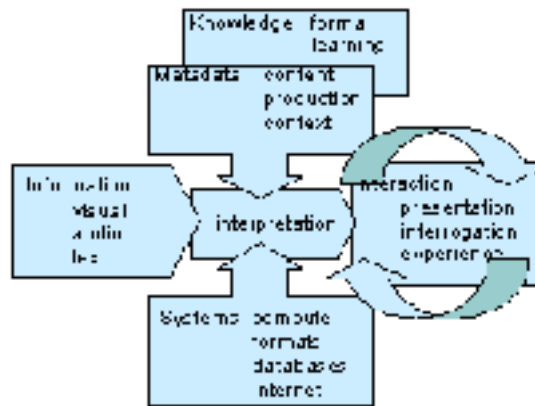
### 3. Ingredients of video archival systems

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 one has an emphasis on image in – image out processes, whereas computer vision studies the interpretation of static or dynamic scenes. Apart from 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 heavily relies 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 much easier to understanding than the visual part. Visual information is so rich in content and variety, even for one single object, that it appears hard to cover by automatic analysis. As a consequence of the difference in progress in these fields, their practices are rather distinct. But where 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 like 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 datasets are the better the result will be. More importantly, when learning from realistic datasets the result is also more robust against 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 ease of exchange as well as proper storage of added value. Formats will develop into more abstract formulations leaving more flexibility for early adapters as well as nuisance for the ones coming late.

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 an enormous compute and storage power to handle a stream of video data. A text stream is relatively condense in its semantic content, but learning facts from texts streams require large datasets which in turn require a large compute power. Analysis of the audio signal requires more power but real time or near real time processes of the visual component is most demanding. Compute power is an important consideration in practical video analysis for some time to come. The solution to the storage and compute powers needed for archival and learning is in *grid computing*, the 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 answering. Any interaction requires carefully designed *presentation* of the result, which in the case of video requires summarization of various kinds as the screen offers only so much space. The interactive component of systems will only be useful when they remember the preferred behavior as well as the preferred presentation in the interaction *experience*, learned by the system form previous sessions. With high-speed wireless technology at the doorstep there is enough opportunity to insert the meta-data at the production site.

#### 4. Interacting with video archives

Interaction is an essential ingredient of 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 as a digital archive will eventually learn from the pattern of interaction of the users, as well as from user annotations of the data.

For assistance of the archivist, the aim is to limit the time needed for the annotation. 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, which 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 better perform more elaborate annotation, sequentially.

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

- Query from a controlled vocabulary

In this query mode, the user inputs query terms from the controlled vocabulary used by the archivist for annotation of the data. Specification of the query should in this case 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 to the words in the document. When the text in the archive is the result of speech recognition from the audio channel this is still a feasible approach, but fuzzy matching techniques have to be used as 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. Required are descriptors of the data, which summarize and emphasize specific characteristics. This is difficult to decide if the purpose is not known beforehand. Hence, it is often limited to rather general descriptors like 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 contain limited information to the system. Only in context such queries can 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 either. 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 a full piece of text, computer retrieval has a better chance to be on target. Similarly, not just one picture should be presented in a query but more than one is to be preferred. And, not just positive pictures should be presented but it is best in computer search to include positive as well as negative images as they bring about much better the intention of the user. The same point also helps in full text retrieval. When negative example texts maybe inserted in the query, the computer may decide much quicker on the proper response.

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 the query example is in practice not annotated, thus the system can only search for similar items based on the content descriptors described above. When the example is internal similarity can in addition 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 need. 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 compute, storage, and display power 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 in turn requires comparing the descriptors of all elements in the dataset to 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 lack of context the computer misses. 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 positive as well as negative examples. By using *relevance feedback* and *visual presentation* of the best results, see the figure, current content-based retrieval systems scratch only the surface of what is to be expected in the near future.

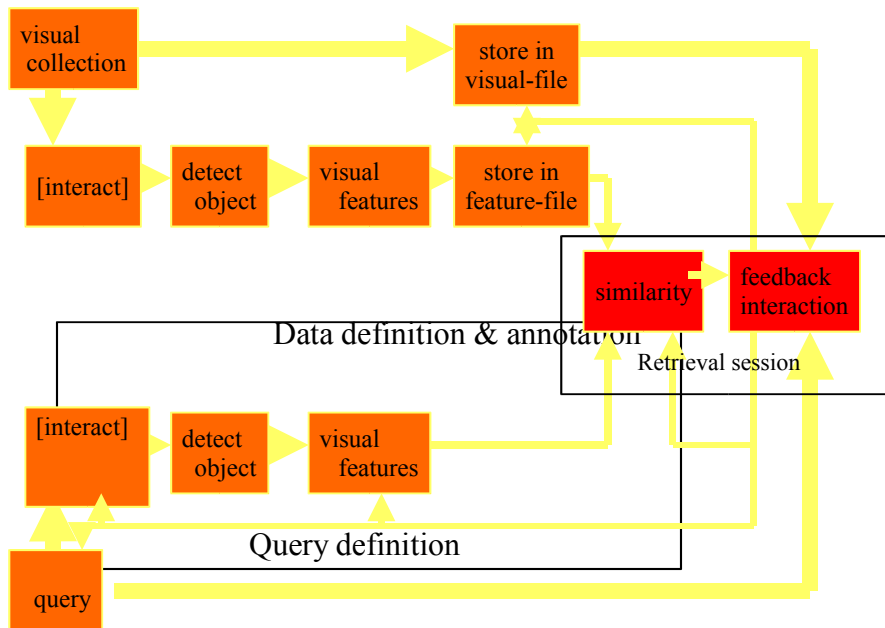
## 5. Progress in multimedia analysis

In this paragraph 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 incidental space or medical images. Processing was concentrated on large computers. In the early nineties personal computers became big enough to hold a digital image, popularizing picture computations. 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 unreachable. There is evidence that even humans divide the image in object and their names only when needed. To identify a

scene, recognition of some details may suffice. A typical example is an orange circle somewhere in the middle of a picture signifying a sun when setting. Another typical example on texture is a patch of stripped 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 generalize their success into a general capability of recognizing items in an image [Fergus 2003].



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

In computer vision, *large volumes of data* are just recently the case. Until the mid nineties computer vision programs were tested on less than hundred images as opposed to the thousands being used today. As a byproduct, test data are no longer perfect. Hence, computer programs are more robust against many sources of variation. For video archives nevertheless larger test collections are needed still as archives typically contain millions of single frames.

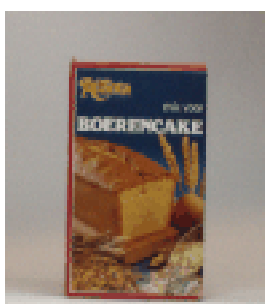
Computer vision starts with *good features*, capable of describing the semantics of the scene and the object and ignoring the irrelevant circumstances of the recording. An object comes in a million different appearances. This is known as the sensory gap in addition to the semantic gap discussed before. 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 coming with media content, such as text and speech. The role of speech recognition is the focus of the next paragraph. 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 part of the broadcast item proper such as subtitle files for productions in a foreign

language, 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. At the current state of affairs, it may contribute to comprehension of the content of the text. As textual elements have a link to the temporal structure of the video, they can therefore be used to generate a time-coded index that allows for the searching of video fragments. As is the case with text indexing in general, morphological normalization (stemming), stop word removal and disambiguation are optional techniques to enhance the generation of the indices, which may improve 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, which have proven to be effective by information retrieval research.



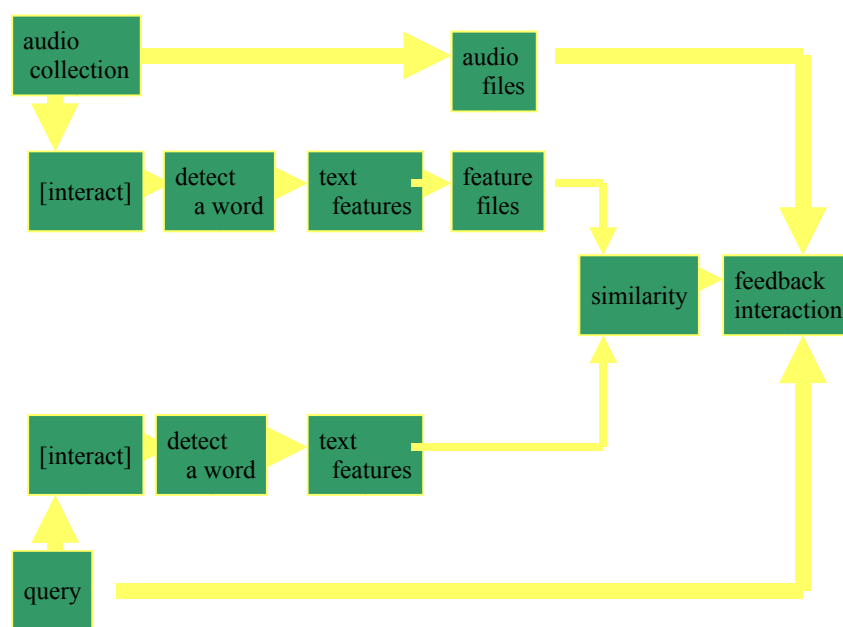
The sensory gap in computer vision: Different versions of the appearance of a single object in (a) are easily recognized by humans whether they are recorded in the dark (b), in blue light (c), in occlusion (d), or under a different viewing angle (e). Good, invariant features describing the object should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

Other examples of text processing techniques that can be employed for more advanced disclosure of media archives are: automatic topic classification, automatic topic segmentation, automatic clustering of documents, automatic summarization, named entity recognition, 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, as this is gaining interest for the linking of archives and collections that have been functioning in isolation for decades.

*Audio processing* to support automated audiovisual disclosure 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. For automated video annotation, there is little use in the performance and functionality of speech technologies that have a longer history, e.g., spoken dialogue systems and dictation technology. Dialogue systems typically operate online but in a narrow domain. Dictation requires training of speaker characteristics and hence would be applicable for rapid subtitling of news broadcasts, but not for general video speech understanding.

In the context of audio disclosure the main technology of interest is speech transcription. Transcription technology in principle detects which words were spoken in what order and at what point in time. Because of the time information, transcripts are the basis for the generation of a time-coded index and therefore 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 require large amounts of textual training data and as a consequence, the volume of available a sets determine the success of the statistical language models. The more variation is absorbed in the model the better the proper word combinations can be sieved out of all candidate word combinations suggested by the acoustic models.

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

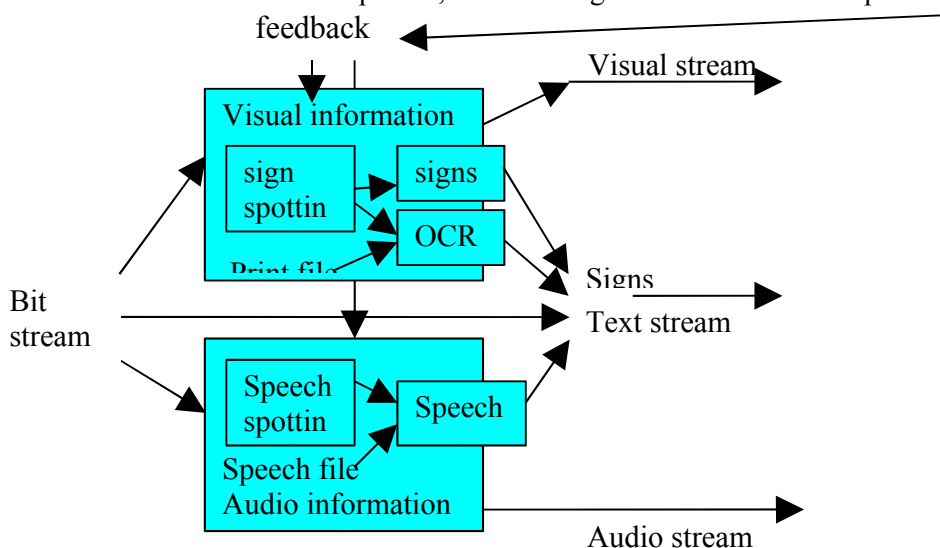
Another ingredient for content-based search is *machine learning* and the hinc-et-nunc version thereof: *interaction*. Interaction has absorbed user relevance feedback, interactive visualization of the result of a query, and adaptable similarity measures [Worrying 2001], yet it requires a major step in tools and machine power to profit in full of interaction as a means for interactive learning experience.

Application of machine learning techniques takes away the incidental variations of within a concept. New techniques use learning in feature space distances, or Gaussian mixture models. Another successful line of learning concepts is to combine weakly performing classifiers into stronger ones. All of these approaches have brought a substantial improvement of the machine learners' capabilities to recognize concepts.

All of the above is getting better all the time, except the amount of data. More data demands more effort in annotation until the point where 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 increasingly many computations is in place, 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 paragraph we discuss recognition, defined as the unambiguous, context-free denotation of signs. The visual representation *A* in all practical circumstances 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 in a pdf-file. It requires a computer function known under the generic name of optical character recognizer, OCR, to convert the printed version of a text to a stream of characters. OCR's are in wide-use and built in many search programs to the effect that paper scans and texts printed in files are now well-accessible. Depending on the quality of the scan data, the quality of the method of OCR-program, and its ability to recognize the font of the text the 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 - least of all - at the proper interpretation of the text block sequence. For example, it requires only a slight misinterpretation to miss a footnote and where it fits in the text. OCR programs rely heavily on built in knowledge on the structure of the texts,

conventions behind letters and the structure of books. OCR programs for print in languages where individual characters are frequently annotated with accents (as in Turkish), or where characters change form as part of a word (as in Arabic), or when there are many characters and compound characters (as in Chinese) will be much harder to decode. Where the conversion of facsimile of texts to character codes is nearly perfect for standard text in main-stream 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. Where it is hard for humans to not see a text in an image, a computer must recognize the distinct pattern of stripes as a sign of language. Text can be of different sources: it can be added to the picture in the later stages of production such as captions and headers. These texts are relatively easy to detect, as they will appear in one style and font, usually at a standard position on 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 of an OCR and see whether it has detected some readable text with some certainty. 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 of language, script, font, depicted size to expect, 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. Reading the text on a billboard, scripts on a t-shirts or a demonstrations often carry most of the message of a photograph but they remain 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 the development of this technology for several languages is and will remain lagging behind.

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

As signs are well-defined symbols outside of context, task-independent performance figures on the recognition performance of signs are obtainable from a large enough data set, which is representative for the data quality? For text spotting and text recognition, a modern recognizer will generate a certainty on detection typically specified, as *for a detection rate of 95% the recognizer will falsely detect 10% of all signs*. In sophisticated recognizers alternative interpretations of each detected character are presented along with their certainty. Certainty of recognition as well as the alternative are a necessary component of robust recognizers at the expense they occasionally introduce confusion, of course.

Up to this point, the recognition of visual and audio data to text was depicted as an exclusively forward process of interpretation. But this can only be a superficial impression of the definitive system as quite likely feedback from the interpretation is inevitable in the

recognition. When the conversion to text yields non-sense, are we spotting text at all? Is the OCR properly tuned to detect the peculiarities of the script? Are we framing the right language; is it Japanese rather than Chinese? From the examples it is clear feedback from interpretation is important in human recognition and so it is with machines especially when the demand on quality will rise. And hence, the availability of certainty and alternatives is important in recognition for visual or audio signals alike.

## 7. Interpretation

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

An essential bottleneck in automatic interpretation is the *semantic gap*. As discussed above, it 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 and no experience what part of the image corresponds to one object in the real world. There is simply no general rule telling 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, see above, computer vision will not solve that problem without learning to recognize them one by one. And that will take a while.

At the state of the art, it is important to grasp the difference between *broad* versus *narrow search domains*. In a narrow domain the data set has well-defined proportions, whereas a broad domain can only be described 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 exposure time for that logo during the Super Bowl, the domain no longer is narrow. The image of the logo is distorted by a skew viewing angle, partially occluded from sight, with changing illumination and cast by shadows, 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, a 1000 different ways to occlude the logo and still recognize it, and 10 different magnification, yielding some million 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 domain.

|   |  |
|---|--|
| 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 <i>visual</i> 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. State of the art: lousy detectors, useful when combined with other lousy ones. |

In the state of the art of computer vision, topics which are hard to do and not so anymore.

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   |

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

As is well-known among archivists, reduction of a video to keywords and key features imply a severe information reduction of the message, brought into practice 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 minimized set of features. In lack of an automatic understanding of context, larger sets of features will carry information about the context, which is implicit in manual search.

In similar vein, what is similar 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. In the table a few major differences in similarity are given. The degree and measure of similarity is an essential part of the query definition.

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

Types of similarity important in computer-aided search.

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

## 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 metadata generation is that an all-digital recording process and post-process will command faster reuse. Automatic analysis is an essential ingredient to meet 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 as both sides will yield unworkable practices. We put forward the importance of understanding some of the peculiarities of the current practice as well as of the current machine performance to design a reasonable process.

Whereas humans make an instant and precise semantic assessment of a scene, machine cannot and will not. Not for visual information, not for audio information. Text information may stand some change of automatic information provided it has been acquired as text and not from visual or audio information. The annotation of machines will neither be precise nor perfect for quite some time to come. And, as we have argued above, as they lack insight in context it is essential that the computer analysis of multimedia is broad. Hence, their analysis may be sloppy on individual items whereas they are still precise in the identification of target. 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. Where annotation is now in the hands of 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 of 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 which can reconstruct the information need 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 for the machine to learn from the user. A long-term goal also is to present the information with high density in a natural and bilateral dialogue with the user. Research is done on almost all topics, yet in isolation. For years to come there is room for improvement of video handling systems, and developers in several IT domains will be keen to collaborate with media archives. We discuss two highly promising areas from which already interesting results can be obtained, both organized around international benchmark events: topic clustering and video retrieval.

As mentioned in section 5, topic clustering is an information disclosure task 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 [Wayne 2000].

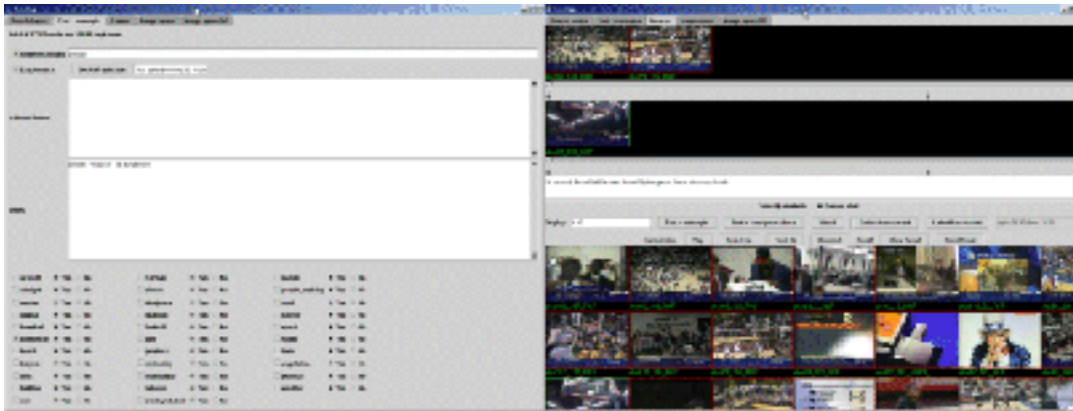
In combination with automatic classification, topic clustering can help to organize large archives, and to build tools that allow users to browse through information dossiers with items in a variety of formats. For example, all newspaper articles, tv-news items and radio broadcasts on the eruption of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

The technology is applicable to textual archives and dynamic news streams, but also to transcribed speech. A task 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 ground truth on 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, VideoOCR to read overlay text, automatic speech recognition, identification of a very limited number of speakers, style recognition, face detection (but no face recognition as it performs very poorly yet), shot length, camera distance, weak segmentation using invariant color descriptors and more [Snoek 2004]. They are used in turn to derive higher-level concept detectors like *boat/ship*, *Bill Clinton*, *Madeleine Albright*, *people walking or running*, and *physical violence*.

All basic detectors function with a different reliability ranging from poor to high quality. In spite of their sometimes weak performance they all are 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 (next to the longer existing ontologies for text).



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

To consider the performance of search, TRECVID 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 *bicycles rolling along*. To indicate the performance, per search NIST considers the precision and recall figures of 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. A top ranking performance [Snoek 2004] indicates that an expert user, combining keyword search and query by similarity with a set of 32 automatically detected high-level concepts, can yield in 15 minutes the following scores:

| Topic   | precision | recall |
|---|-----------|--------|
| <i>people walking with their dogs</i>                       | 28%       | 42%    |
| <i>tennis player contacting the ball</i>                    | 10%       | 19%    |
| <i>bicycles rolling along</i>                               | 41%       | 59%    |
| <i>Bill Clinton with at least part of a US flag visible</i> | 35%       | 36%    |

Automatic video annotation is still a hard 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 for practice, but the *yearly* progress is considerable. Even poor quality descriptors help in the automatic annotation and they will be improving by learning from larger data sets. When combining automated analysis with interaction, a useful new search paradigm is about to emerge.

In this paper we have indicated where progress is to be expected from automated analysis, and which solutions are much further away. We have done so at the risk to be ridiculed by our fellow researchers as painting a too simplistic view. Nevertheless, as is always the case in such scouting of technology, the answer of the oracle will give an answer to a different question than raised. The answer is more complicated than desired, but such is inevitable as the frontier of progress follows its own internal logic.

Nevertheless, we hoped to wet your appetite for the future computer-aided annotation will bring. We look forward to communicate with you where our vision of the modern archive needs mending.

## 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] A.W.M. Smeulders, M. Worring, S. Santini, A. Gupta, R. Jain: 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 disclosure 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 different needs and involving innovative methodologies.