Dublin City University

School of Computing

CA4009: Search Technologies

Section 9: Multimedia Search

Gareth Jones

November 2019

Introduction

Much information is contained in media other than electronic text:

- spoken content - radio broadcasts, recordings of meetings and lectures
- printed items - legacy texts, e.g. books, newspapers, corporate documents
- static images - photos, cartoons
- video - sports, films
- music (recordings), music (notation)

It is thus logical to think in terms of extending information retrieval methods to multimedia information sources.

Introduction

- The information required to satisfy an information need can be contained in more then one media, e.g. a newspaper article and a radio broadcast.
  - In which case the choice of media may depend on the user's context or the user themselves, e.g. an audio response may be better for an "eyes busy" driver or for a partially sighted user
- Or it may only appear in one media, e.g. an historical handwritten document or video of a family party.
- Or the user may wish to locate relevant information in multiple media, e.g. when researching for an article on an historical event.
- Many items are actually multi-modal, e.g. web pages, and using all the information they contain can potentially enable improved search.

Introduction

- User interaction with information retrieval systems is an important topic in information retrieval.
- Exploration of interaction in information retrieval typically focuses on informational issues of:
  - query expression,
  - interpretation and comprehension of retrieval results,
  - satisfaction of information need.
- These issues are even more important in multimedia information retrieval which can present information in a combination of text, audio and visual forms.
  - with a temporal dimension for audio and video.

Introduction

However, the nature of human interaction with different media forms also needs to be considered. e.g.

- Text must be read, but can be annotated to highlight relevant words and phrases, e.g. using different fonts or colours.
- Information in an image can be taken in a glance by the user.
- Many images can be browsed very rapidly when displayed as thumbnails in a single panel.
- Spoken content is temporal and can only be accessed in real-time, since humans cannot interpret audio faster than this.
- Video too is temporal, but can often be represented in a summarised by meaningful individual keyframes.

Introduction

- Advances in interactive technologies mean that users increasingly have the option to search using non-text or multi-modal queries.
  - e.g. spoken queries on mobile devices, image examples to search for objects in images and video.
- Multimedia information retrieval can be taken to include such non-text querying of any archive.

Introduction

- Multimedia content and queries not only present new interactive challenges, but also the information itself may be expressed differently.
  - e.g. contrast the language used in carefully written text with spontaneously spoken speech.
- This may have implications for technical solutions, but also for the usefulness (as opposed to the absolute relevance) of the content in different contexts,
  - e.g. would a user want to read a verbatim transcription of a radio interview? - with all its “umms” and “errs” or prefer a carefully written article containing the same information.

Introduction

- Many of the information retrieval techniques developed for electronic text data can be applied to information held in other media, but multimedia information retrieval presents new challenges.
- A key challenge in multimedia information retrieval is the need to process the content to identify what it actually contains.
  - For documents this enables the content to be indexed for search and retrieval.
  - For queries it makes them useable for search.

Specifying Requests

While appearing simple or obvious, multimedia search requests are often complex to interpret and perform.

Examples of some basic multimedia search requests:

- Find me movies containing car chases.
- Find me goals in soccer matches.

Real requests are likely to be more specific:

- Find me tries by Brian O'Driscoll in his last 6 nations series.
- Find explanations of database normalisation in online lectures.

Specifying Requests

Within a specific video. more detailed browsing requests might include:

- Show me the car chase in this movie
- Show me the goals in this soccer match.
- Personalised summary - Show me a summary of this F1 race focusing on Lewis Hamilton.
- “Query by example” - Is the person in this picture in this movie?

Content Recognition

For speech data:

- we can user automatic speech recognition (ASR), but this can be difficult - see later.
- increasingly we can use crowdsourcing - where human volunteers transcribe the content in return for a small payment (how could we check the accuracy?), but much spoken data will be confidential, e.g. corporate meetings or presentations, and not available to volunteers.

For printed content:

- use optical character recognition (OCR) - again much content can be difficult to transcribe correctly.
- again we can use crowdsourcing.

Content Recognition

Images:

- How should we analyse an image?
- What should we try to identify? What features will be useful for search?

Videos:

The same as images, but also

- Multiple frames showing very similar scenes - use of multiple similar images can make analysis more reliable.
- Temporal dimension - find events along the timeline of the content.

Again, automatic analysis or crowdsource-based labelling.

Content Recognition

Music:

- What should we try to extract?
  - what will be useful to support users in search?
  - melody, chord progressions, lyrics, rhythm, structure of the composition
  - instruments playing
- Automatic audio feature extraction - this is difficult
- Analysis of printed notated music.
- How can we use computer representation of music, e.g. MIDI files.

Interacting with Retrieved Results

- How should we represent content to determine relevance?
  - images? keyframes from video?
  - speech/music - listening to whole item too time consuming.
  - audio snippets - how should these be created? excerpts from speech transcripts?
- How should we access the content in a relevant item to satisfy the user's information need?
  - listening to whole item may be the goal,
  - but finding a specific section by listening will be time consuming

Scope of Multimedia Search

Multimedia Search (or Multimedia Information Retrieval (MIR) is a very active research area which seeks to:

- better understand user needs, their cognitive abilities and preferences for media interaction,
- develop technologies to create effective multimedia search systems.

Solutions described here should not be regarded as definitive; new ideas are being proposed, developed and evaluated all the time.

Topics for Multimedia Search

- Searching Multimedia using Text
- Retrieval of Spoken Content
  - introduction to the principles of speech recognition
- Content-Based Retrieval of Visual Media
  - features for visual indexing
- MIR systems
  - search
  - browsing

Searching Multimedia using Text

- Multimedia content is often accompanied by text metadata which can be used search for relevant items without requiring analysis of the multimedia content.
  e.g. for a movie: title, actors, director synopsis.
- This can have limited value for temporal media since the content must be manually auditioned, potentially in its entirety, to determine relevance and/or access information contained within it.
- Metadata may be created manually by the content provider or potentially via crowdsourcing,
  - e.g. manual transcription of audio content, manual annotation of images (people, places, event, etc.).

Searching Multimedia using Text

- Metadata may be enriched using automatic expansion methods.
  - e.g. using a knowledge graph to link entities in the metadata to their attributes, e.g. to identify the year of release of a movie and add it to the metadata.

Searching Multimedia using Text

Some other possible sources of information for use in the retrieval process:

- Alternate text for image on a web page
- Image URL e.g. http://www.cars.com/alfa_romeo/alfa_156.jpg will have the terms cars, alfa, romeo, and 156 extracted.
- Paragraph text - use text near image as probably related to the image - boost weight of search terms nearer to the image.
- Document name or image “Title”
- Heading - most recent heading in a document prior an image.
- Anchor text - text pointing to the content via a hypertext link.
- Other terms - any term on the page with the image.

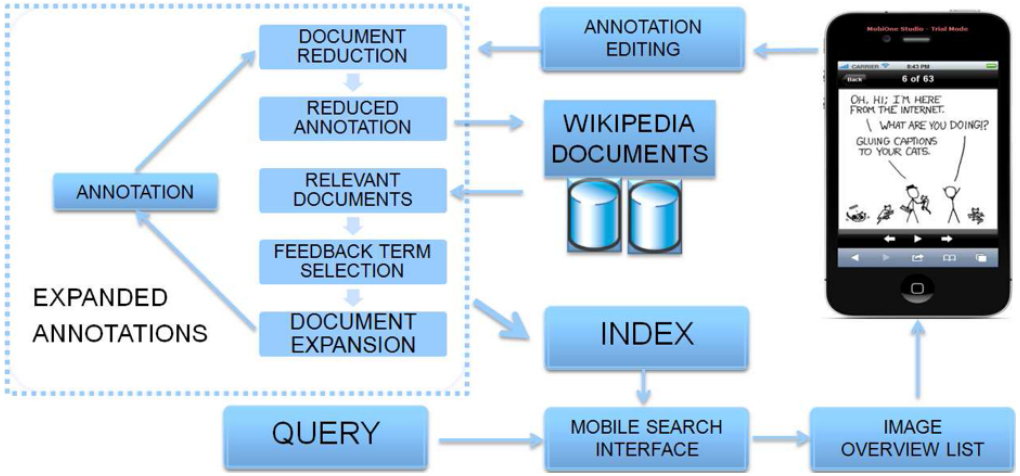
Searching Multimedia using Text

- Metadata to support search of multimedia content can often be derived from attributes or context data (see the *MediaAssist* example in Section 1: Introduction):
  - date and time of content creation/editing/access
  - GPS location of content creation/editing/access
- These can be expanded or described in terms that might appear in a search query as remembered by the seacher:
  - morning, midweek, weekend, summer
  - Dublin, Cork, Ireland
- Text IR can be used in combination with metadata filtering,
  - e.g. date ranges of image capture, location at which it was taken.

Searching Multimedia using Text

Textual metadata can be enriched in various ways.

- For example, user text annotation of images is often very limited, e.g. small number of text labels accompanying images uploaded to a photo sharing website, clip art collections.
- Annotations can be enriched using external knowledge resources such as Wikipedia.



Multimodal Information Retrieval

- While text-based retrieval of multimedia content often enables some level of information access:
  - it generally provides a very limited description of the content
  - it fails to identify the specific location of relevant content within the data - important in spoken audio and video
- It is important to consider how to search based on the actual multimedia content.
- As we will see, the effectives of multimodal search can often be improved by using content-based information in combination with text metadata.

Retrieval of Spoken Content

- Spoken content retrieval (SCR) can be based on:
  - text requests to search spoken data,
  - spoken requests to search text data,
  - spoken requests to search spoken data.

  The emergence of voice driven technologies means that all of these are now posssible for retrieval applications.
- From a technology perspective these scenarios are the most straightforward multimedia extension to text retrieval, since the queries and the documents are still textual.
- However, the temporal dimension of speech means that user interaction and context issues need careful consideration.

Retrieval of Spoken Content

An obvious first step is to :

- create a perfect transcription of the spoken data,
- then treat the content as text data and apply standard text IR methods.

But, perfect transcriptions are frequently not available:

- perfect automated transcription is not possible;
- full professional manual transcription is generally usually uneconomic, e.g. business meetings (huge amount of content, very limited audience);
- crowdsourcing may be available for non-confidential content, e.g. radio and TV content, but may suffer from issues of poor quality.

Retrieval of Spoken Content

- Information retrieval for text conventionally assumes that the search targets are easily defined in terms of individual documents.
- This will also apply for scripted or formal spoken content,
  - e.g. broadcast news stories.

  However, much spoken content,
  - e.g. recorded lectures or meetings,

  will be long and multi-topic.
- Long spoken content will be inefficient to audition.
- Ranking multi-topic content can be unreliable since it lacks focus.
- Questions arise: what should the retrieval units be? how should they be identified in the speech data?

Retrieval of Spoken Content

- Automatic speech recognition (ASR) systems can be used to generate imperfect index information for spoken content.
- Index errors arising from ASR errors will reduce the effectiveness of a retrieval system (consider how?), but overall retrieval is often good enough to be useful (consider why?).
- Speech recognition can be thought of as a decoding task.
  - specifically the decoding of air pressure wave signals into a written word signal.
- The speech signal is captured using a microphone and an audio amplifier, then sampled using an analogue-to-digital converter.

Retrieval of Spoken Content

The level of reduction in retrieval effectiveness for an SCR system using ASR can be explored using quantitative experiments to measure variations in metrics such as the *precision* and *recall*.

- For example, what is the change in MAP between use of a perfect transcript and an errorful transcript created using an ASR system?

While the usefulness of the search system to a potential user can be indicated by the results of quantitative experiments, it is also important to explore the user's response to the system via user evaluation.

- What does reduction in MAP mean in terms of the user experience?
- Is a system which has a 10% reduction in MAP acceptable to the user? Might 30% reduction in MAP be acceptable?

Automatic Speech Recogition

- Automatic speech recognition (ASR) is hard!
- Speech patterns of different people show significant variations.
- Every individual utterance of a word by every speaker is unique.
- Speech is continuous – often with no gaps between spoken words.
- Words are “smeared” together,
  - e.g. *you need to know the d-t* can be merged giving a phrase more like *you nee’to know*.
- So, in addition to not knowing what the words are:
  - an ASR system doesn’t know where the word boundaries are,
  - or even how many words there are!

Automatic Speech Recognition

- Some words are acoustically ambiguous.
  - For example,

    to, *too* and *two*

    are homophones which sound the same.
- Some words only have a very small acoustic distinction.
  - For example,

    bee and *pea*

    while not true homophones, can be highly confusable, especially

    when the initial consonant is not well articulated.

Speech Recognition Systems

ASR systems typically comprise two fundamental components:

- Acoustic models - statistical models of all the possible speech sounds that the recogniser will attempt to correctly recognise.
- A language model - a statistical model of the expected word (or other units to be recognised) sequences of the language.

These are combined in a process referred to as *decoding* to generate the most likely output sequence.

Lattice or N-best lists of alternative likely outputs can also be output.

Acoustic Modelling

All words can be broken down into a small set of constituent sounds, referred to as the *phones* of the language.

English has about 45 distinct phones.

- e.g. `speech` is composed of the phones `s p iy ch`

Each phone can be produced in a number of ways depending on its context, leading to slight variations in its sound.

Acoustic Modelling

- A *speaker independent* ASR system must recognise the speech of any individual. Acoustic models of speech must be rich enough to capture all of these variations.
- Speaker independent models are trained using gender balanced speech examples from many people with diverse regional and age groups.
- If sufficient training data is available better performing *speaker dependent* models can be used.
- *Speaker adaptation* enables speaker independent models to be adapted for an individual speaker using a small amount of speech data provided by from this speaker.
- A separate acoustic model is built for each acoustic unit, usually a word or subword phone unit.

Acoustic Pattern Matching

- For more than 30 years, the most successful approach to acoustic pattern matching to use statistical modelling using hidden Markov models (HMMs).
- Significant improvements have been achieved in recent years using *context-dependent deep neural networks (CDDNN-HMMs)*.
- ASR systems work at the subword phone level:
  - Enables ASR to be computationally tractable.
  - Makes best use of limited available training data.
  - All words are built from their constituent phones.
  - A phonetic dictionary is used to constrain the allowed phone sequences to allow only those of real words.

Acoustic Pattern Matching

- A very simple ASR task with a vocabulary of a small number of words can use a simple finite-state network with no language model relying only on the acoustic model.
- The word model with the highest matching score when compared to the acoustic input is the recogniser output.
  - which doesn't necessarily mean that it is the word that was spoken!.

Language Modelling

- Speech is too ambiguous to attempt ASR based only on acoustic models for more than very simple tasks using a small number of words:
  - would produce large number of recognition errors
  - use of large number of acoustic models in parallel too computationally expensive
- A *language model* has to be incorporated into the recognition process.
- Ideal language models would model human language processing.
- However, currently the most effective language models are based on simpler word-level models, referred to as *n-grams*.
- A general definition of an n-gram language model is as follows $Pr(w_n|w_1, \ldots, w_{n-1})$, but this model is impractical. Why?

Language Modelling

- Practical n-grams are bigrams, trigrams and 4-grams, $Pr(w_2|w_1)$, $Pr(w_3|w_1, w_2)$ and $Pr(w_4|w_1, w_2, w_3)$ respectively.
- $Pr(w_n|w_1, ..., w_{n-1})$ is estimated using the n-grams, e.g using higrams.

$$Pr(w_n|w_1, ..., w_{n-1}) \approx Pr(w_n|w_{n-1}) \times \cdots \times Pr(w_2|w_1) \times Pr(w_1|start)$$

- However, even most possible word pairs will not actually ever be observed in even a large corpus of training data.
  - We cannot estimate probabilities for unobserved n-grams directly from the training data.

Language Modelling

- So called “backed-off” models are used to give non-zero probabilities to unobserved events, e.g. $Pr(w_2|w_1) \approx Pr(w_1) \times Pr(w_2)$.
- But these models give very poor estimates to unobserved events.
- Word embedding methods (see Semantic Search notes) can be used to address this weakness.
- Semantic matching of related training sentences can be used to better estimate probability of semantic events.
- Unobserved word sequences get high probability in the language model if they are made of words that are semantically similar to words forming an observed word sequence.

Vocabulary of an ASR System

- The set of all words which can be recognised by an ASR system is its *vocabulary*.
- The total vocabulary available in a typical ASR system typically includes many words which are very rarely used, but since we never know what is going to be said, they still need to be in the vocabulary!
- Thus, we need to have as many words as possible in the vocabulary to ensure that most of the words are present.

Vocabulary of an ASR System

- *out-of-vocabulary (OOV)* words are words not present in the vocabulary of a particular ASR system.
- if they appear in the spoken input, OOV words cannot by definition be recognised correctly by an ASR system.
- OOV words are usually proper nouns, technical jargon or slang.

  This has implications for SCR, since many useful words for IR are proper nouns or domain specific items of vocabulary.
- Since ASR at the phone level means we do not have to explicitly build models of all words, new words can be added to the vocabulary easily if their phonetic structure is known.

  Adding new words to the language model is more difficult.

Read vs Spontaneous Speech

ASR accuracy varies great depending on the factors outlined previously.

In addition, the recognition accuracy greatly depends on the formality of the structure, level of spontaneity, and location of the speaker relative to the microphone.

Scripted read speech in a quiet environment with a suitable domain specific vocabulary with well trained acoustic and language models is now almost 100% accurate.

Unscripted spontaneous speech in a noisy environment where the speaker is moving around relative to the microphone might be lee than 50% correct.

One of the major issues is disfluencies, e.g. false starts, self corrections, “um”, “err”, etc, and poor linguistic structure.

Output of Speech Recognition

Ideally speech recognition systems would produce 100% accurate output. However, as already explained ASR systems make mistakes.

For transcription systems these errors can be classified into the following types:

- substitutions - where the word spoken is misrecognised as a different word
- deletions - where a word spoken is missing from the output, e.g. where 4 words are spoken and the transcribed output only consists of 3 words
- insertions - words not spoken are inserted into the output, e.g. when 3 words are spoken, the transcription system outputs 4 words

Output of Speech Recognition

Deletions and insertions can occur since the speech recognition system has to determine the number of words spoken, and does not always do this correctly!

All types of errors can occur where:

- the acoustic models fail to model the spoken speech input well;
- or the language model does not model the spoken word sequence well.

All types of errors can be made worse when an OOV word is spoken.

- OOV words cannot by definition be recognised correctly, but their presence can also make it difficult to correctly recognise adjacent spoken words.

Spoken Content Retrieval (SCR)

- Early SCR focused on spoken *document* search.
  - e.g. search of news stories in radio and television broadcasts.
- But, this task is quite easy:
  - Broadcast news content is “self describing”.
  - Document units clearly defined.
  - ASR system can be well trained for the news domain with large amounts of data
    - Large amounts of closely related text available for the information retrieval algorithms.

Spoken Content Retrieval (SCR)

More challenging situations may have:

- Content is not “self describing”
  - it does not contain details needed to support high accuracy search based on ASR
  - the participants do not speak the necessary content words.
  - text metadata needs to be included to support the search
    - e.g. giving details of the entities being discussed, and perhaps background information to give detail and context.

Spoken Content Retrieval (SCR)

- Long unstructured spoken audio files where there are no obvious document boundaries, e.g. meetings, lectures.
  - automatic location of ideal “jump-in” points to begin playback of content from unsegmented content
  - automatically segment into smaller focused retrieval units, e.g
    - fixed length segments - will divide semantically focused units,
    - topic-based segmentation using natural language processing based algorithms, but suitable boundaries are not clear and anyway these algorithms make mistakes.

    so neither solution is ideal.

Accessing and Browsing Spoken Content

- Due to the temporal nature of audio, browsing to determine relevance and access relevant information is a significant issue for multimedia data, particularly audio data.
- With text we can visually scan a large amount of data very quickly looking for relevant details - search terms can by highlighted in the text to make this even easier.
- However, with audio data the scanning process is much slower since only a single stream can be listened to at a time.
- Experiments have shown that speech can be played at around double speed (digitally to avoid pitch shifting) without loss of intelligibility, but no faster, but listening to this content is cognitively demanding, and the listener will rapidly get tired.

Accessing and Browsing Spoken Content

Since listening to complete audio recordings is highly inefficient:

- Retrieval process should identify suggested points to begin playback of potentially relevant content.
- Needs to allow sufficient context to make content understandable.
- Possibly begin playback before relevant content.

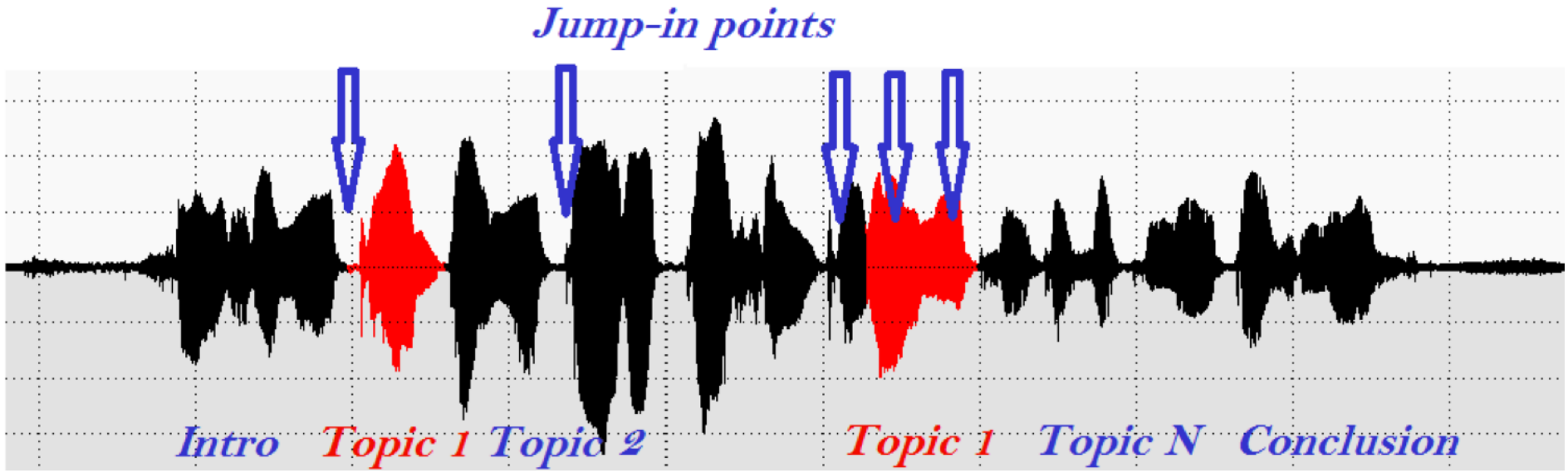
How can we anticipate the quality of suggested playback points that will be acceptable to the user and their behaviour when using the system?

Interactive access applications typically use graphical visualisation of content, possibly using noisy ASR transcripts.

Accessing and Browsing Spoken Content

Consider accessing relevant information in a long multi-topic recording:

- red = relevant content, black = non-relevant content
- Down arrows indicate some possible locations where the user might begin playback, as might be suggested by the system - “Jump-in points”
- What would be the best system suggested Jump-in point?



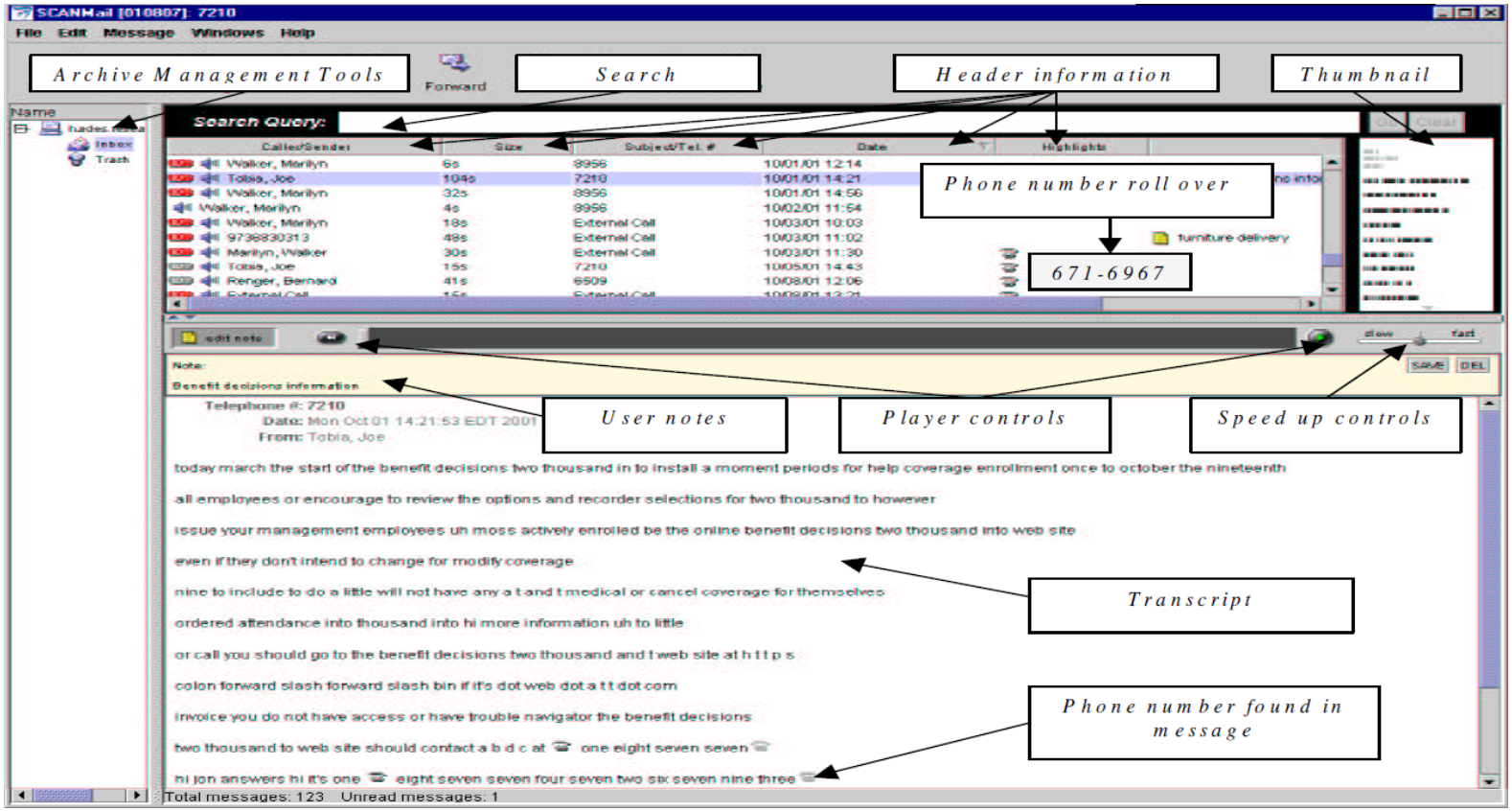
Accessing and Browsing Spoken Content

- Since we are unable to spot the interesting section of a document quickly by looking at it, it must be auditioned from start to finish to find the information required, which takes significant time.
- Visual tools for browsing spoken documents are thus potentially very important in maximising data access efficiency.
- Potentially interesting portions of the document can be identified by the user, and playback can be started at any point by selecting it in the interface.

Accessing and Browsing Spoken Content

- Since reading text is faster than listening, one option could be to present users with the ASR transcript of a spoken item.
- Scanmail is an example of a system that presents the user with the ASR transcript directly, in this case a voicemail message.
- A user study suggested that this feature was appreciated. However, there is a relationship between the error levels of ASR transcripts and their usefulness in the system, with higher error rates being less useful.
- A user study on the usefulness and usability of ASR transcripts for a web archive found that:
  - transcripts with WER $> 45\%$ were unsatisfactory,
  - while transcripts with WER $< 25\%$ were useful and usable.

Accessing and Browsing Spoken Content



ScanMail user interface

Accessing and Browsing Spoken Content

- It is important to keep in mind that an SCR system must not allow users to develop an unfounded trust in the ASR transcripts.
- One study showed that professional users were found to have significant confidence in the the transcripts and their own ability to work with them.
- This resulted in the users failing to seek relevant content not explicitly appearing the transcripts, meaning that they missed relevant material if this absence was caused by transcription errors.

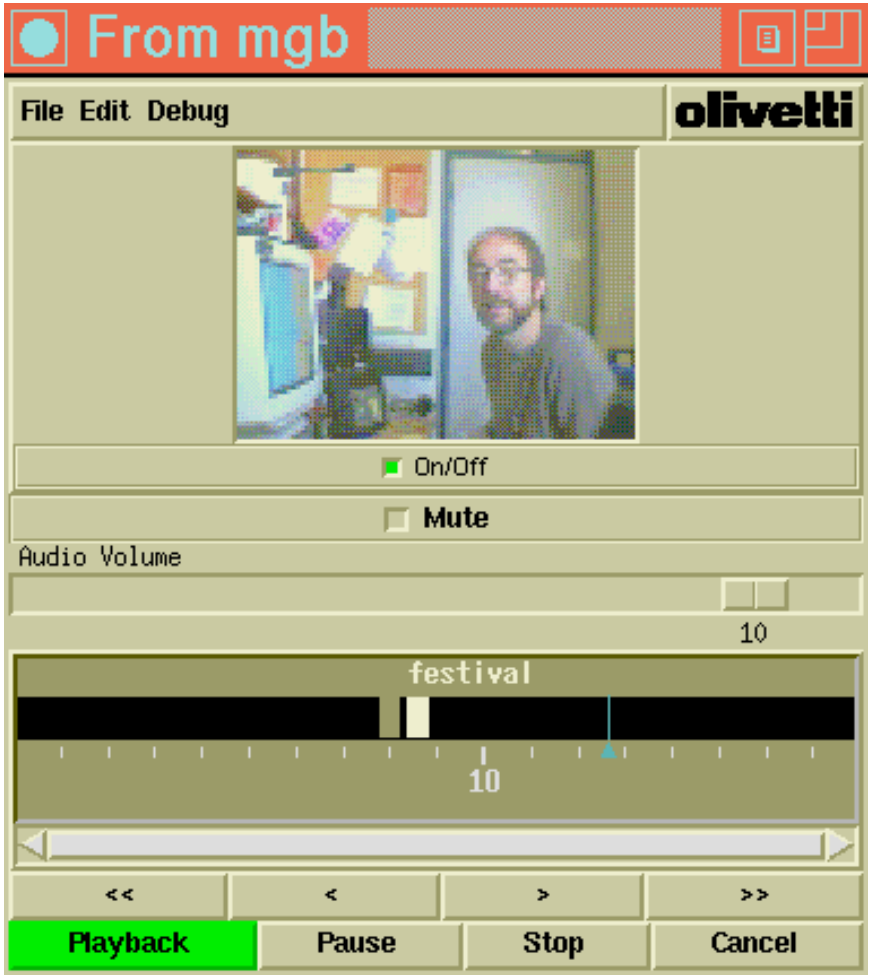
  The same effect was reported in users of Scanmail.

  Recall is more critical for voicemail search and misplaced trust in the ASR transcripts caused users in the study to miss crucial information that was not recognized by the ASR system.

Accessing and Browsing Spoken Content

- An alternative to use of transcripts is provided by visual tools for browsing spoken documents.
- Visual browsers typically attempt to represent the audio stream as a static image that can be viewed at a glance.
- Content can be represented on a horizontal left-to-right timeline, with the search word hypotheses displayed graphically along it.
- Time runs from left to right and events are represented proportionally to when they occur.
- Potentially interesting portions of the document can be identified and playback can begin at any point by selecting it.

Accessing and Browsing Spoken Content



Video Mail Retrieval (VMR) message browser

Accessing and Browsing Spoken Content

- The Video Mail Retrieval (VMR) browser shows words from the query along a timeline.
- The brightness indicates the confidence of the ASR system in the correctness of the word.

Accessing and Browsing Spoken Content

- The Dutch podcast search engine *Kunststofzuiger* developed at the University of Amsterdam illustrates typical strategies for SCR players.
- The player page has a query-independent representation of the episode, in the form of the podcast title, broadcast date and description, and also a term cloud of important words that has been extracted from the transcript of the podcast.
- It also has a query-biased representation of the episode in the form of the player, which contains markers pointing to the moments within the podcast at which the query word occurs.
- Clicking one of the markers moves the user to the point in the speech stream at which the query word is spoken.

Accessing and Browsing Spoken Content

Accessing and Browsing Spoken Content

Two variants on these strategies are:

- A term cloud spread out along the player to give the user a general idea of the topical development over the course of the speech media.
- A heat map display that uses shading or colour to reflect the relative likelihood of a position along the timeline being relevant to a query, rather than showing position of specific words.
  - This approach is adopted in the VMR Broadcast News browser.
  - In order to create this representation, the ASR transcript is divided into equal-length segments each of which is scored against the query.

Accessing and Browsing Spoken Content

VMR news browser

- Segments are scored against the query.
- Brightness of segment in display indicates the score against the query.
- User can click at any point on timeline to begin playback at this point.

Accessing and Browsing Spoken Content

- The player bar of the Radio Oranje application illustrates the use of a magnifying glass metaphor.
- The player displays the entire speech in a timeline, as well as a magnified view showing a window of 45 seconds around the current position of the cursor.
- Above the playbar, the transcript of the currently playing segment is displayed, with the query words in bold and a moving underline tracking the progression of the playback.
- The magnified view makes it possible to depict segmentation information for the entire programme in a compact space without losing detail.

Accessing and Browsing Spoken Content

Radio Oranje using magnifying glass metaphor.

Spoken Input to Applications

The increased computing power available via cloud-based services means that it is now possible to provide online robust ASR services.

The ubiquity of networked mobile computing devices such as smartphones and tablets which can easily be connected these services, means that ASR services can be accessed on such devices.

Use of speech-based input is often an attractive and efficient option on mobile devices due to the small size of these devices making text input less convenient than on conventional computers.

- However, it must be accurate and fast enough for users to find it more attractive as an option than typed text input.

Spoken Input to Applications

Speech input is not suitable in all situations:

- e.g., in public places where use of a speech interface would remove privacy or disturb other people.

An attractive application for voice input on mobile device is spoken input of search queries, referred to as *Search by Voice*.

Search by Voice

ASR for Search by Voice should be able to support any query that can be entered using text input.

ASR systems for Search by Voice have potentially huge amounts of training data available:

- Acoustic models:
  - train using manually labelled data (as for a standard ASR system) (referred to as *supervised* data),
  - but also using data from real spoken queries while system is operational (run ASR system on data and automatically extract words with high recognition conference as additional training data to update models) (referred to as *unsupervised* data).

Search by Voice

- Language models: use text collections (as for a standard ASR system) and user text queries.
- Use of location information is found to be important to recognition quality,
  - e.g. if the ASR systems knows that the query was spoken in Ireland, then Irish related vocabulary items can be favoured in the ASR output.

Analysis of a query logs from Yahoo! mobile searched showed:

- Ave length of spoken queries = 4.2 words
- Ave length of text queries = 3.2 words

Spoken queries have more function and question words (often these words are treated as “stop words” in IR systems).

Multi-modal Interaction on Mobile Devices

The large hi-res screens on smartphones and tablets enable use of high quality visual output.

In general, applications can support mulit-modal interaction:

- Input: text, speech, click
- Output: text, speech, graphics, image, video

Multi-modal Interaction on Mobile Devices

Advantages of visual over audio output:

- Visual output rich and efficient presentation - human visual bandwidth is much greater than acoustic.
- Lower time for users to search for and digest information.
- Reduced cognitive load for the user - speech is lost once is it spoken, information must be remembered by the listerner; visual content can be scanned repeatedly.

Example of effective multi-modal application: Google maps

Visual output not always best: applications where users are “eyes busy” on other activities, users with limited or no vision - e.g. use sat navs.

Content-Based Retrieval of Visual Media

Content-based retrieval of visual media is a much more challenging prospect than spoken content retrieval.

Fundamental questions include:

- What visual features should be extracted to describe the content to be retrieved?
- Can these features be extracted reliably?
- How should queries be posed:
  - text - using extracted feature descriptions?
  - image - query by similar examples?

Content-Based Retrieval of Visual Media

- Successful content-based retrieval systems for visual media have generally been *task* or *domain specific*,
  - i.e. they are developed for a specific activity such as spotting cars in a video or tracking football players in match.
  - They can often be adapted for use in different domains, e.g. to track aeroplanes or baseball players,
    i.e. they are not general purpose.
- Automatic understanding tools have been (to date) impossible to develop.
- Analysis of visual content is currently developing very rapidly with the use of machine learning methods which seek to overcome the limitations of established methods.

Content-Based Retrieval of Visual Media

- Content-based visual search systems are generally highly interactive:
  - User enters their best attempt at a text or visual query.
  - Collects the search results, and provides feedback on the relevance of the retrieved content.
  - Query is adapted via relevance feedback, and the user searches again.
  - Continues iteratively as the user steers the system towards relevant terms.
- Referred to as having a “human-in-the-loop”.
- The difficulty of describing a visual information need means that this model will probably persist as visual analysis technologies advance.

Content-Based Retrieval of Visual Media

- Humans (the searchers) should be given simple tasks which they can perform consistently using the search system.
- rather than complex ones which can be confusing and lead to variable outputs.
  - e.g. they could be asked to identify relevant features of retrieved images in a simple way in order to refine the query when performing multiple search stages.
- They should **not** be required to understand the science of image analysis or describe why an image or parts of it are relevant in a detailed and personally subjective way (which is very likely to produce inconsistent behaviour).

Issues in Visual Media Retrieval

- Multimedia objects have multiple dimensions.
- How we view an object depends on: what our task is or what we are looking for, and will be based on different properties of the media.
- To allow for the many different interpretations of different searchers, we must capture all these possible features when indexing the content.
  e.g. Is this a boat or a sunset?

Issues in Visual Media Retrieval

- Each feature used in a visual retrieval system generates a ranked listing of items.
- This ranking is based on a match between a query image and the image collection.
- The set of ranked lists created for each individual feature must be integrated into a single overall ranked list combining the evidence for each individual feature into an overall similarity value for this query for each image in the collection.
- Labelling of features (whether automatic or manual) will often contain errors, or in the case of manual annotations inconsistencies, which will affect ranking behaviour and reliability of effective retrieval.

Issues in Visual Media Retrieval

- We must also understand that query specifications will be incomplete,

  e.g. ‘find a picture “like” this one’

  What does the user mean by “like”?

  We can’t tell what the user’s interest in the image is, and we don’t know which of the, often many, interpretations of an image to concentrate on.

- Visual media is very content rich with many possible interpretations, and can thus often address many diverse information needs.

- Query specifications may be refined in cycles of relevance feedback - the “human-in-the-loop”.

The Semantic Gap

- The *semantic gap* refers to the gap between the contents of a multimedia data stream and its meaning as interpreted by human observers.
- it is relatively easy to extract low level features from visual images,
  - e.g. the colours present in an image,

  but very much harder to automatically describe images in the way that a human observer would.
- This difference between machine and human descriptions of visual media is referred to as the *semantic gap.*

The Semantic Gap

Consider the following examples of gaps in representation and interpretation:

- There is little gap between a table of salary numbers and their meaning.
- There is a larger gap between the words in a document and its overall information meaning.
- There is a much larger gap between the features of a video which can be extracted automatically (as described in these notes) and its meaning or semantic interpretation by a user.

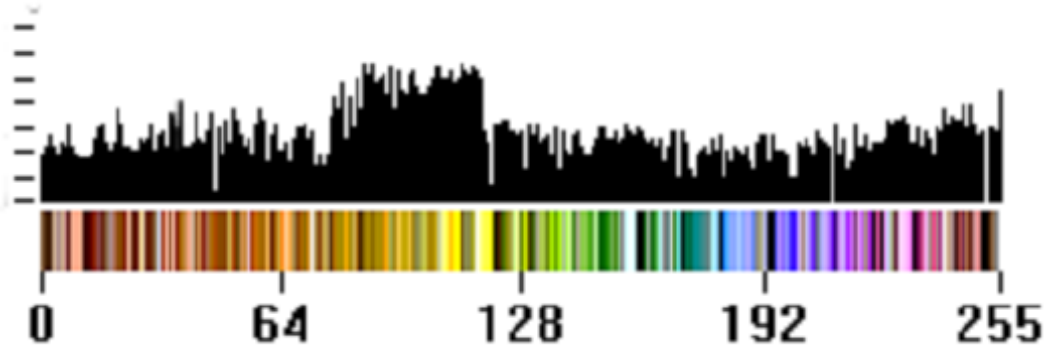
Content-Based Retrieval of Image data

- Relevant images can be located using a variety of attributes.
- These can include automatically extracted and manually assigned features.
- Depending on the type of feature, retrieval can be carried out on three different levels:
  - Level 1 - using image primitives based on extracted features.
  - Level 2 - based on derived attributes (iconography).
  - Level 3 - inferred abstract attributes (iconology).

Level 1 - extracted image primitives

The lowest level (pre-iconography), using image primitives based on extracted features such as:

- **colour** ... the full RGB spectrum or perhaps segmented into bands. An image containing blue/green, yellow and orange could be a sunset!



- This is a colour histogram showing a colour profile of the image.

Level 1 - extracted image primitives

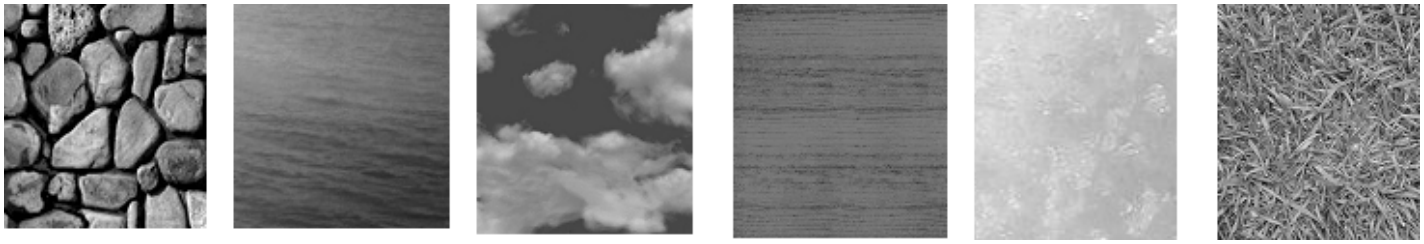
- **texture:**
  - Attribute representing the spatial arrangement of grey pixels (relates to texture NOT colour).
  - A measure of properties such as smoothness, coarseness and regularity.
  - Repetition of basic texture elements called *texels* which contain several pixels, whose placement could be periodic, quasi-periodic, or random.
  - Natural textures are generally random whereas artificial textures are often periodic.
  - Texture may be coarse, fine, smooth, granulated, rippled, regular, irregular or linear.

Level 1 - extracted image primitives

- A query could be:

  find images with regions of texture similar to grass

- Texture is a black and white feature, e.g.

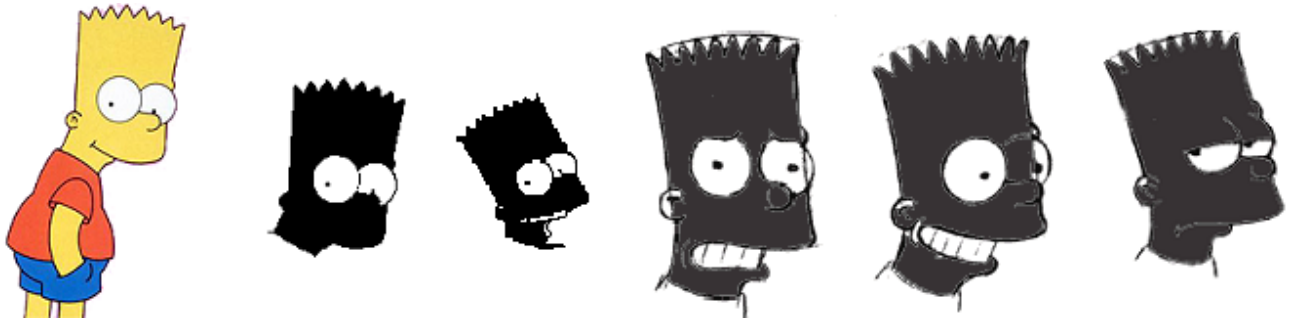


Textures & Shape

- **shape**: - geometric shapes, 2D

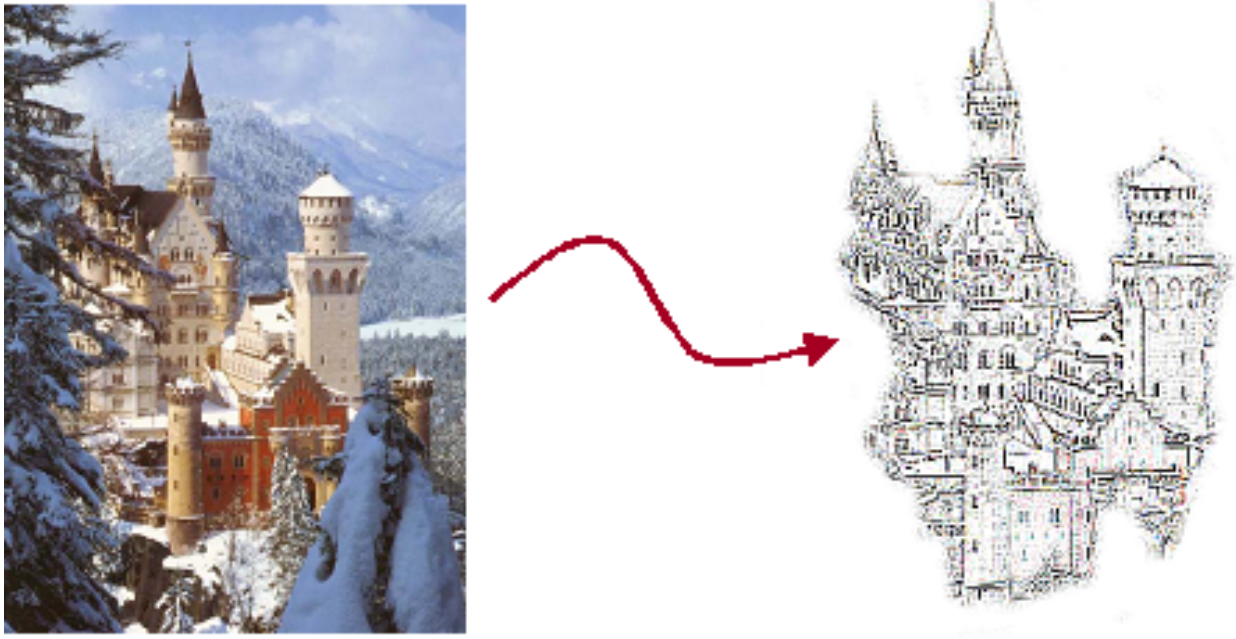
  find images containing a shape similar to this sketch

- example application: a Bart Simpson detector



Textures & Shapes

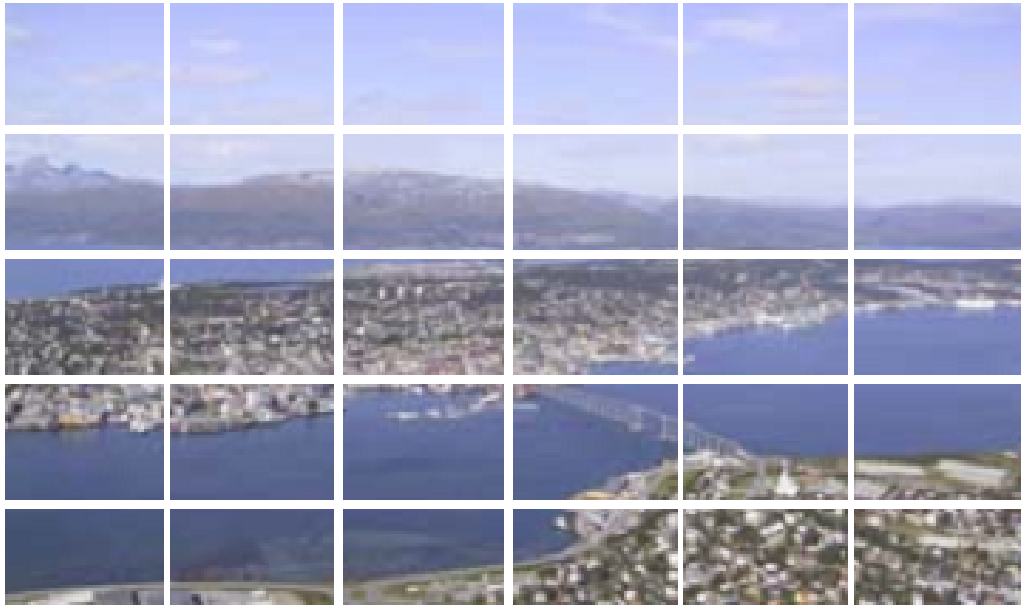
- complex shapes can be difficult to process



Level 1 - extracted image primitives

- **spatial placement**: - X-Y co-ordinates in a rectangle ... find images containing some object or interest in the top left corner...

  or image with sky texture in top of picture and water near bottom



Level 1 - extracted image primitives

Some other primitive feature examples:

- Text within images - captions in newscast frames
- Domain concepts - noses, eye colour in faces, depressions in weather maps, etc.
- Spatial relationships - X in_front_of Y, Y behind Z, etc.
- But the one which works generally is a combination of the above
  e.g. find images with yellow triangles arranged in a circle.

Level 1 - extracted image primitives

- Another example.

  Find images with sky texture and colour in the top half of the image, with a water texture in the bottom right and with an aeroplane shape in the middle of the image.

Level 2 - iconography

- We can also have derived attributes such as the presence of specific objects, e.g. chairs around a table, or named specific individuals.
- Retrieval based on derived attributes.

  (iconography: describing a picture's actual contents, or icons)
- This enables queries of the form:

  "find pictures of a train crossing a bridge"

  or

  "pictures of Bill Clinton meeting Gerry Adams."

Level 2 - iconography

“Picture of Bill Clinton meeting Gerry Adams”

“Picture of a wine car in front of a house”

Level 3 - iconology

- More abstract still, inferred abstract attributes which do not correspond directly to content in the image, but rather to some inferred attribute,

  e.g. if we have football players and a goalpost and a football in a picture then we have a “football match”.

- This is level 3 (iconology: describing a picture’s deeper artistic significance)

- This corresponds to queries like: *find images of a football match*,

  as opposed to : *find images of 2 footballers and a football*,

  as opposed to: *images of green with a grass texture and two splashes of black and white in a striped arrangement with a black/white circular pattern in the middle of the two splashes*.

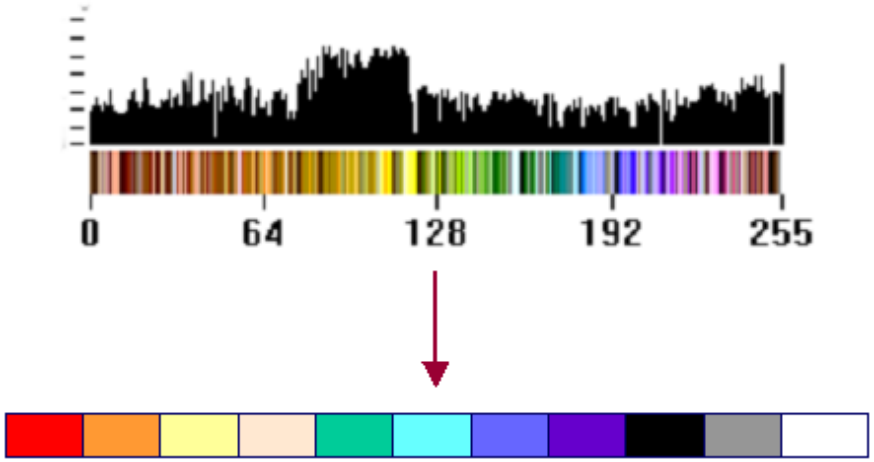
Level 3 - iconology

“Picture of a motor race”

“Picture of a storm over a city”

Using Colour Histograms

- A problem with retrieval based on colour is that small variations in colour can lead to non-matches.
- This problem is addressed by grouping perceptually similar colours together.
- Colour histogram based retrieval is the easiest and most obvious approach to doing this.



Using Colour Histograms

- Example of textual markup description of colour histogram.

```
<colour>
<red> 0.8 </red>
<orange> 0.4 </orange>
<yellow> 0.3 </yellow>
<green> 0.2 </green>
<blue> 0.1 </blue>
<indigo> 0.5 </indigo>
<violet> 0.4 </violet>
</colour>
```

- A separate description could be generated for each region in a divided image.

Using Colour Histograms

- Similar analysis can be carried out for all images in a collection and a query image.
- The query image is then compared to each image in the collection.
- There are many possible ways of computing a similarity measure between the query and each image.
- One simple measure is to compute the difference between each feature of the query and the each image, and then compute the sum of the differences.
- Images are then ranked in increasing order to sum of difference, i.e. the image with the smallest sum of differences is the most similar to the query.

Texture-Based Retrieval

- Example of textual markup description of texture.

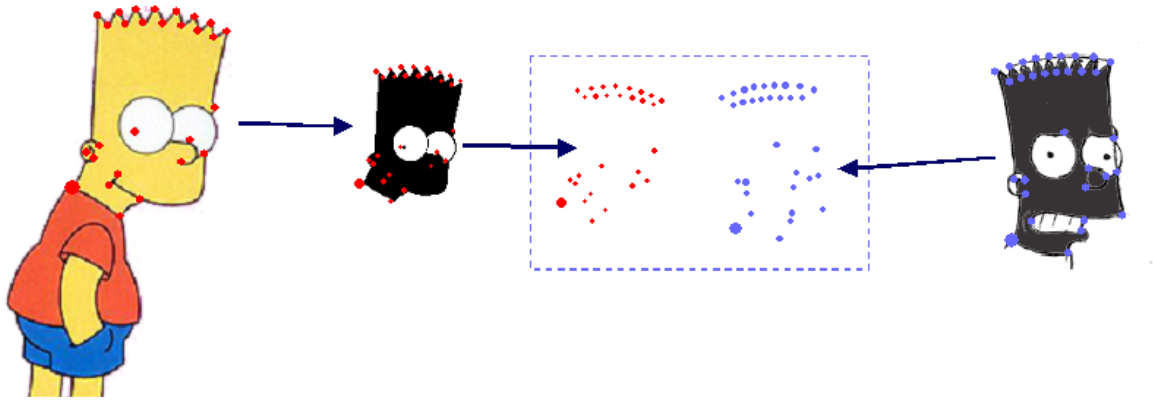
```
<texture>
<sea> 0.1 </sea>
<wood> 0.7 </wood>
<grass> 0.3 </grass>
</texture>
```

- Many other texture models can be defined for comparison, e.g. sky, concrete.

  Or the above one may be subdivided, e.g. into different types of wood.

Shape-Based Retrieval

- Important for meteorology, medicine, manufacturing, law, etc.
- To keep things simple look at 2D shapes only, but there are issues:
  - shapes are identified by determining boundary points and collecting these into feature descriptors / vectors.
  - similarity between shapes is measured in distance between feature vectors.
  - retrieval requires a query or sample shape as a starting point

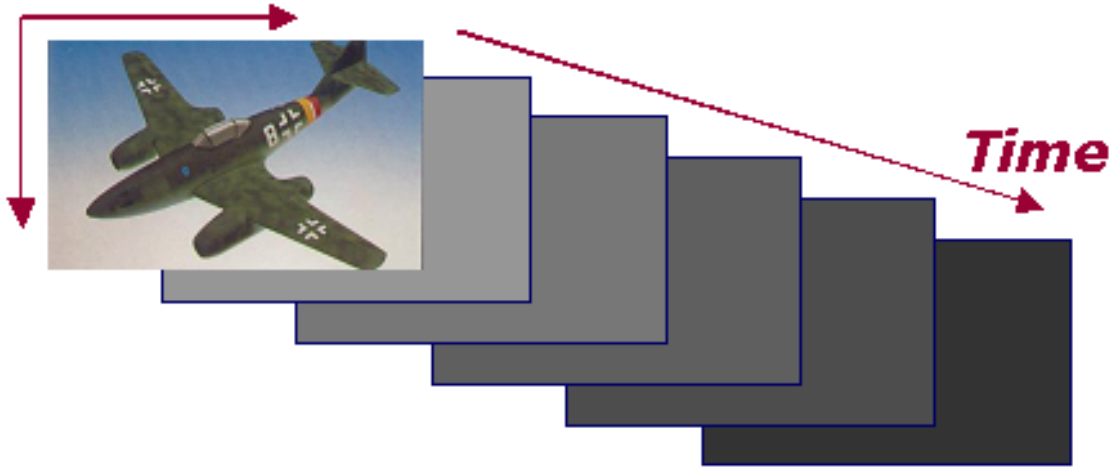


Content-Based Retrieval of Video

- Video is a continuous media.
- Events in video add a temporal dimension to those in single images.
- Video is usually combined with an audio soundtrack.
- For most effective access an application should generally use both the video and audio media synchronised for retrieval.
- Some searching on video can be achieved using metadata such as: date, title, director plus a textual description of movie contents, but this gives only very limited search functionality.
  - much more interesting is “content-based search”.

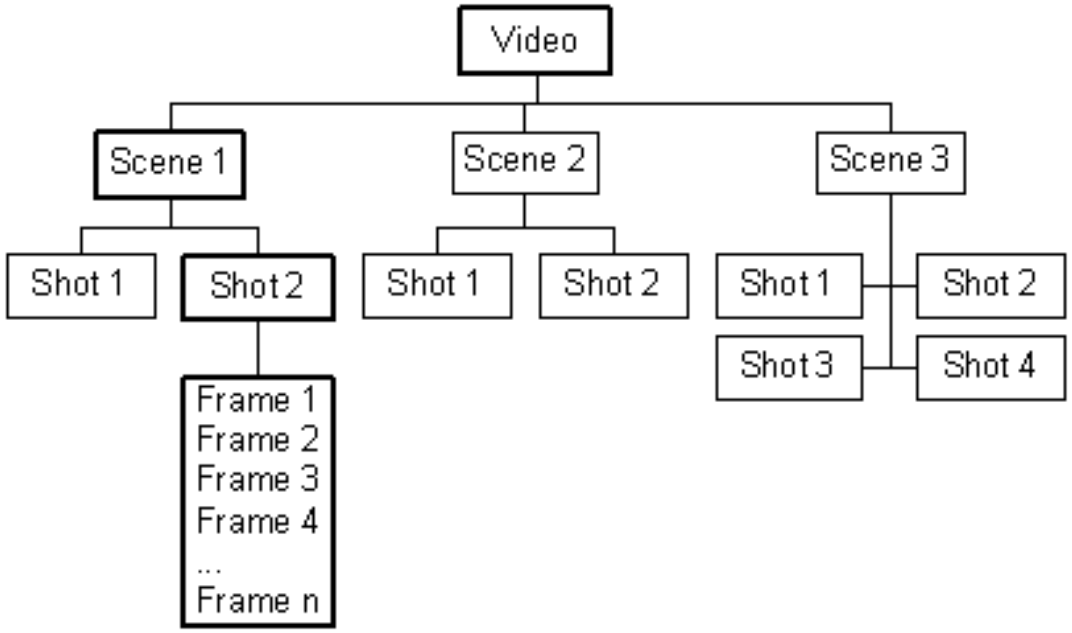
What is Video?

- Video is a sequence of individual shots combined together in some way.
- Video is of variable lengths and played as a continuous stream into a 2D window.
- Thus it has 3 dimensions: x, y and time.
- To do video retrieval we must identify the clips and then segment the video into a list of clips.



Scenes and Shots

- Video is typically composed of a series of scenes.
- Most scenes can be decomposed into a sequence of shots.
- A shot in video information is a sequence of continuous images (frames) from a single camera.

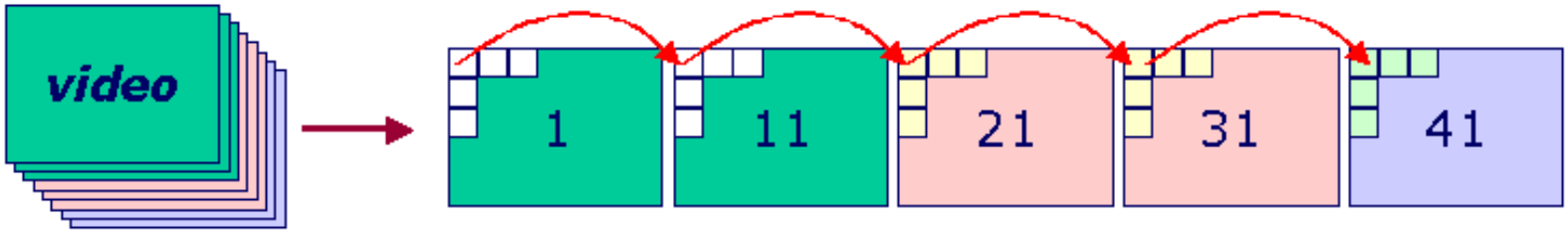


Shot Boundary Detection

- A shot boundary is crossed when a new camera is used, or a recording instance ends and a new one begins.
- Shot Boundary Detection automatically segments video into its constituent shots.
- Why do this?
  - Allows content-based operations over video at the granularity of shot units, e.g. browsing, searching.

Shot Boundary Detection

- How? By examining every X frames / adjacent frames to look for shot cuts.
  - Simple shot cuts are easy to process - Based on colour, texture, intensity/brightness, etc



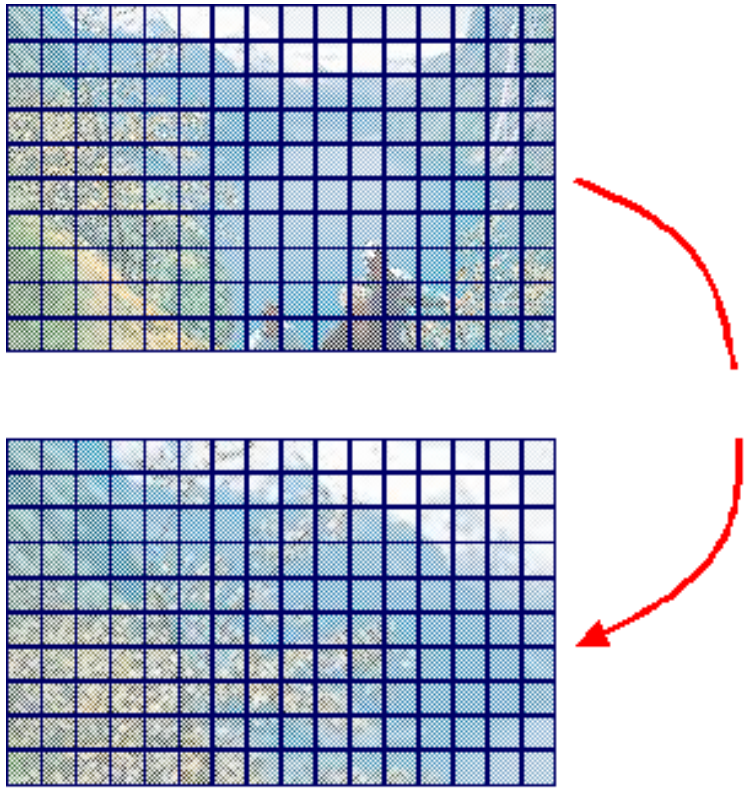
Shot Boundary Detection

But it can be quite difficult because of camera tricks:

- Dropped Shot Boundaries:
  - fade-in and fade-out
  - dissolving
  - morphing
  - wipes
  - many other chromatic effects
- False Shot Boundaries:
  - zooming and panning
  - tilting
  - booming and tracking
  - events in the content itself - camera flashes

Shot Boundary Detection

Example of “zooming”.



Browsing Digital Video

- Regular video has always been linear - video cassette.
- DVD/Blueray allow random access:
  - Based on chapters.
  - Which have been segmented manually - chapters are represented by keyframes or video clips.
- A problem is that browsing keyframes is still browsing through a lot of video (hours),
  but shot-level granularity and keyframe browsing is relatively easy to achieve.

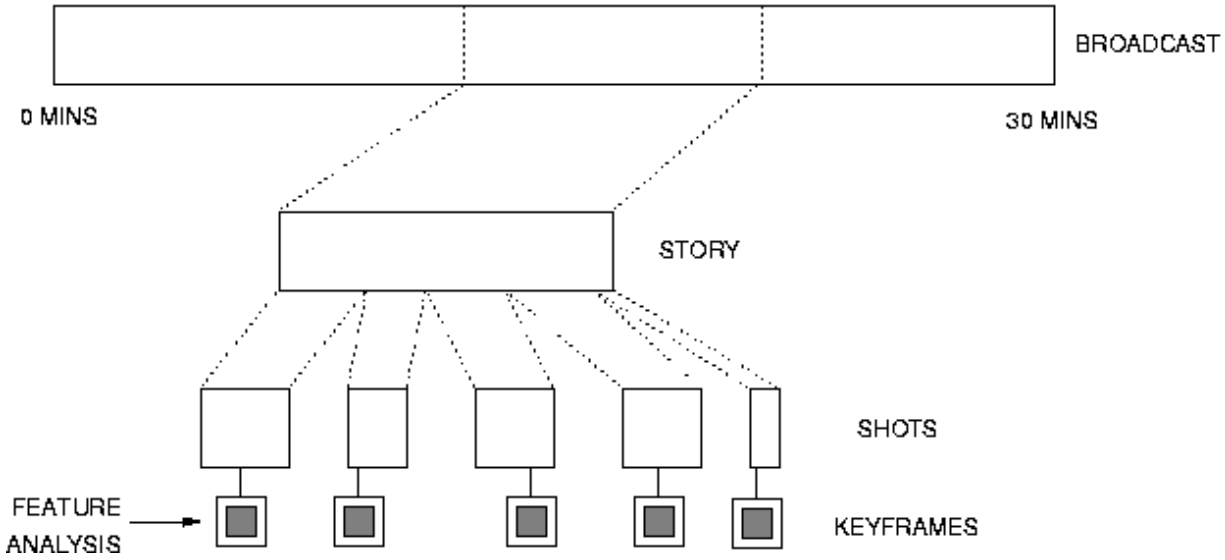
Keyframes

How do you identify a representative keyframe for a shot?

- Random frame is hit and miss.
- First or Last frame is also problematic.
- Choosing the middle frame from a shot is the easiest approach.
- Choosing the frame with the most average colour histogram from the shot.
- A virtual centroid - a frame that does not actually exist, but contains the average of all colour data in the shot.

This enables a video to be represented as a sequence of keyframes, one for each shot.

Structuring Broadcast News Content



Breakdown of a news broadcast into stories, shots and keyframes.

- Shots are defined by scene or camera changes.
- Ideally the keyframe is the single frame which “best” represents the visual information in the shot.
- Shots are indexed by analysing the visual features of the keyframe.

Multimedia Indexing

- Spoken content:
  - convert into text using speech recognition (or if available use close-captions subtitles) and then apply text based information retrieval.
- Extract features from representative keyframes from each shot:
  - Low-level features: colours, textures, edges.
  - Determine the presence of people, faces.
  - Classify images as: indoor, outdoor, landscape, cityscape.
  - Identify more difficult features (presence of objects, landscapes or people tracking).
  - Classify audio into: spoken data, music, silence.

Multimedia Indexing

- Keyframes can be examined to extract useful additional features.
- Depending on the domain of the data various useful features can be identified.
- For many applications a very useful feature is face detection.
- For structured content such as news data. Text appearing in the video images can be extracted using “video OCR”, and then added to the document text index data. This is effectively an additional source of metadata for the images.

The following examples are taken from the *Informedia* project at Carnegie Mellon University.

Informedia: Face and Text Detection

Informedia: Face and Text Detection

Detects faces and text in video keyframes.

- Requires models of typical faces, for example to find standard features such as eyes, nose and mouth.
- Ability to detect text in the frame and segment the text region.

In the example we can see examples of success and failure in object detection.

Failure to correctly identify a face or a piece of text typically results from difficulty in detecting the features due to difficulty in distinguishing them, e.g. finding all the faces in a crowd (top right), black and white straight lines on a building looking like text (bottom right).

Informedia: Face and Text Detection

Consider how we might address these problems?

- Look for flesh tones and textures to find faces.
- Look for stone texture as a building rather than text.

Informedia: OCR in Image Text

Informedia: OCR in Image Text

Video OCR from a single keyframe can be difficult, e.g. the fidelity of television images can be quite poor. So it is quite common to misrecognise individual characters.

How to address this problem?

One idea:

- Use multiple keyframes from the same section of video - will have the same letters on the screen.
- Perform Video OCR on each keyframe.
- Then take some vote of the majority for each letter. Hopefully the majority count for each letter will be more reliable than output of a single keyframe.

Informedia: OCR in Image Text

- Check spellings of video OCR with a dictionary, but many caption words are names or jobs titles which may often not be in a dictionary.

  How can we find examples of correct spelling in context of news story?

- One idea:
  - use the “most likely to be correct” character strings as a search query to a text search engine (e.g. online newspaper or newswire service);
  - then check for the presence of the words or similar ones, e.g. in this example we are likely to find stories about “Gerry Adams”, but not “CERRN ADAMS” as detected by the video OCR system in the example.

Video Search

Various forms of query can be used:

- Query by submitting text queries alone and returning ranked lists of shots.
- Query by submitting text queries and lists of required features:
  - e.g. an outdoor, landscape with text query “aircraft takeoff” - implement as a text search followed by Boolean shot filtering according to the specified criteria (e.g. must be outdoor shot).

Video Search

- Query by submitting sample keyframes or drawings of required content:
  - use image retrieval techniques to match queries to the shot keyframes.
- Query for named object, e.g. submit a face and get back shots containing this person.
- Query video specifically e.g. “a green car travelling towards the camera” - needs to include temporal information in the index data.

Video Search

- Do people want to browse or search?
  - Browsing is time consuming, but searching may not be accurate enough yet.
  - Best is a combination of searching to narrow the number of videos which need to be browsed to find useful content.

    Thus video retrieval is typically a two phase process.

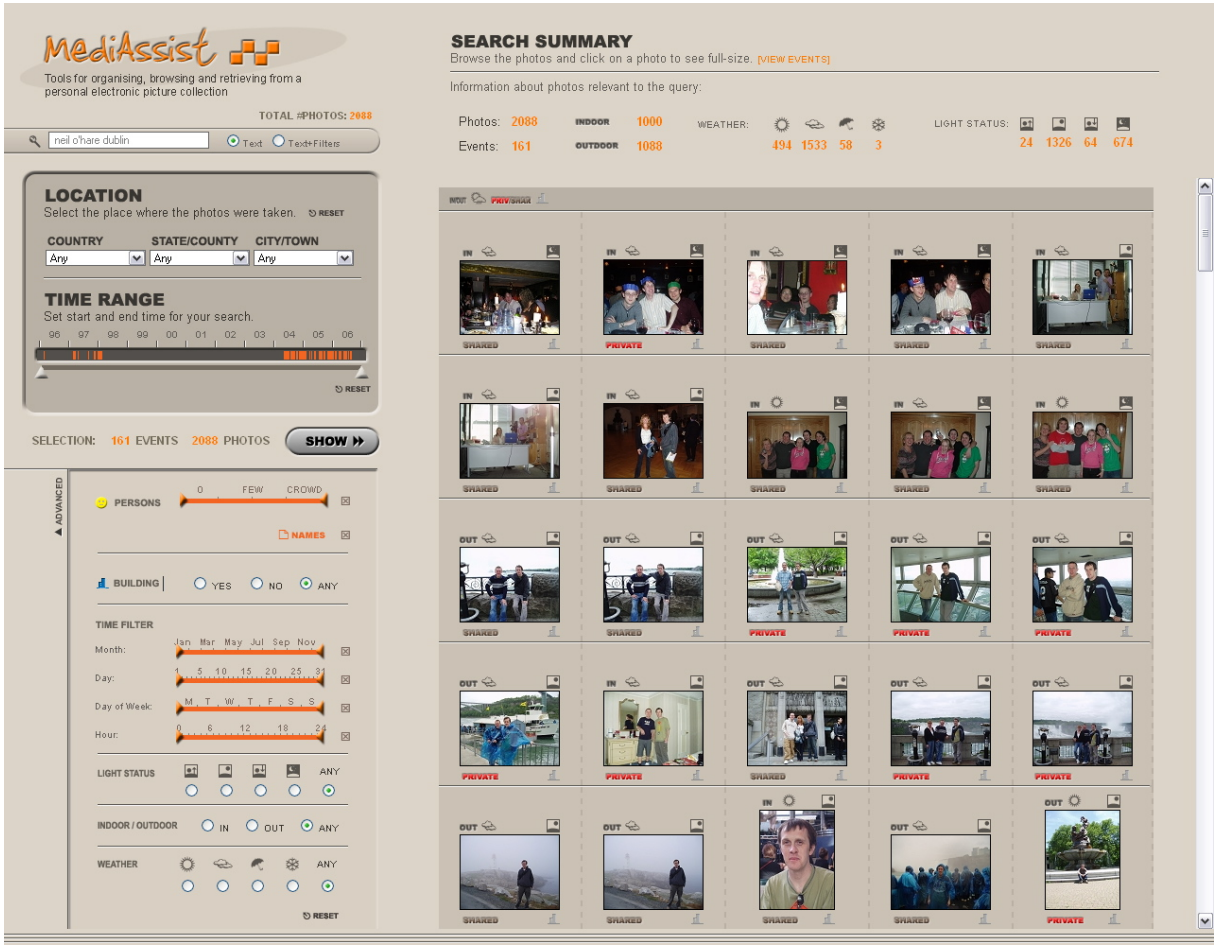
  The “human-in-the-loop” - interactively and gradually specifying and refining description of the information need.

Combining Content and Context

- Improved and increasingly complex multimedia content analysis is set to continue indefinitely with continuing search.
- However, functional and useful applications are already possible using available technologies.
- The key to the success of these applications is creative use of the technologies and information currently available.
- For example, textual metadata can be combined with relatively simple content analysis to provide highly effective search applications.

Personal Digital Photo Search with *MediAssist*

- MediAssist digital photo search with location and time.



Personal Digital Photo Search with *MediAssist*

- Designed for personal photo search.
- Photos uploaded into MediAssist photo search application.
- All images stamped with time and date, plus GPS location.

Personal Digital Photo Search with *MediAssist*

- Photos are grouped into “events” based on time and location.
  - “Event” is set of photos taken in roughly the same place in fairly quick succession - i.e. it’s a bit approximate, but is quite accurate.
- GPS location is looked up in a gazetteer of place names, e.g. DCU, Glasnevin, Dublin, Ireland.
- Time and date mapped to other values: name of month, day of the week, season (summer, autumn, etc).
- Ambient light condition calculated based on time, date and location - is it dark, light, dusk, etc.
- Prevailing weather conditions looked up online from weather station archive data.

Personal Digital Photo Search with *MediAssist*

Semi-automated annotation of names of people in photos.

- Faces detected in photos.
- Linked together in an event using “body patch matching”, e.g. looks for colours beneath a face - to link the same person between photos.
- Matches faces against models and names in a database.
- User can confirm or correct suggested names for people in photos.
- Corrections are used to update the parameters of the models. The objective being to improve the accuracy of the face identification over time, as more training examples are included.