TREC 2002 Video Track Experiments at MediaTeam Oulu and VTT

Mika Rautiainen[†], Jani Penttilä[‡], Dmitri Vorobiev[†], Kai Noponen[†], Pertti Väyrynen[†], Matti Hosio[†], Esa Matinmikko[†], Satu-Marja Mäkelä[‡], Johannes Peltola[‡], Timo Ojala[†] and Tapio Seppänen[†]

[†] MediaTeam Oulu
P.O.BOX 4500, FIN-90014 University of Oulu, Finland
{firstname.lastname@ee.oulu.fi}

[‡] VTT Technical Research Centre of Finland
P.O. Box 1100, Kaitoväylä 1, FIN-90571 Oulu, Finland
{firstname.lastname@vtt.fi)

Abstract

In TREC 2002 Video Track MediaTeam Oulu and VTT Technical Research Centre of Finland participated jointly in semantic feature extraction, manual search and interactive search tasks. In the semantic feature extraction task, we sent results for semantic categories of cityscape, landscape, people, speech and instrumental sound. Spatio-temporal correlation of oriented gradient occurrences was used with example shots to detect shots containing people, cityscape or landscape. The audio signal features consisted of various statistical measurements and were used to detect shots containing speech or instrumental sound. Our video browsing and retrieval system, VIRE**,** was used for manual and interactive search tasks. Our system offers two techniques for video retrieval: 1. Multi-modal indexing based on self-organizing feature maps with semantic filtering. 2. An interactive navigating tool that combines two inter-shot properties, temporal coherency and metric similarities, into a view where database shots are presented in a lattice structure. We tested our interactive navigating tool with eight persons to obtain results for 25 pre-defined search topics. In this paper we give an overview of the approaches and a summary of the results.

1 Introduction

In this paper, we first describe our work in semantic feature detection. Then we introduce our video browsing and retrieval system VIRE and employ it in manual and interactive search tasks. In the end of the paper concluding remarks are given.

2 Semantic Feature Detection

2.1 People, Cityscape and Landscape

We evaluated the efficiency of visual features in the detection of people, cityscape and landscape from video shots. Since the video material contained a lot of narration in the audio signal, audio features were omitted from this test.

A well known strategy in city vs. landscape classification of images is to use image's edge gradients [19]. The same approach can be used to discriminate natural structures from non-natural. For example, man-made structures can be distinguished from natural views by computing edge histograms that cumulate in specific orientations when structures are non-natural containing many straight edges. Natural views have more evenly distributed histogram of different orientations.

Our Temporal Gradient Correlogram (TGC) feature computes local correlations of specific edge orientations producing an autocorrelogram, whose elements correspond to probabilities of edge directions occurring at particular spatial distances. The feature is computed over 20 video frames sampled evenly over the duration of video shot. Due to temporal sampling, autocorrelogram is able to capture also temporal changes in spatial edge orientations. From each sample frame, the edge orientations are quantized into four segments depending on their orientation being horizontal, vertical or either of the diagonal directions.

To make the feature vector more discriminative, we used detection of skin-colored local regions to generate four values describing the relative size and structure of consistent skin areas in a video shot frame. First value was *relative amount*, which was simple discrete value between 1 and 5 describing the relative amount of skin-colored local regions. Next values indicated number of *long*, *medium* and *short zero runs* (adjacent non-skin blocks) between skin-colored local regions measuring the uniformity of skin structured areas. To detect skin-colored regions, we marked manually skin areas into 40 key frames selected from the shots in feature development collection and trained a self-organizing map into a skin detector using localized HSV color histogram feature. The histogram was localized into a sector area that covers the typical colors of skin in HSV color space. The sampled local regions of 10x10 pixels were used in localized histogram computation. Degraded quality of test videos was prominent: some of them appeared closely monochromatic and there were large color variances between different videos. Due to this, the prognosis for the success of skin detector was initially set low. However, the feature values were normalized and joined with TGC to examine the discriminative power of less-than-adequately performing skin detector.

To find the shots consisting of a certain semantic concept, we selected sets of example shots from the collection of video data for training semantic feature development, 13 example shots were selected for people, and 10 for both cityscape and landscape. These example TGC and TGC+skin feature vectors were compared against the vectors computed from the shots in the test video collection. The dissimilarity between the features in the test collection and in the training set was computed with L1 norm. The resulting set of most similar shots was pruned from duplicates and rank-ordered to create the final list of shots most probable to contain the semantic feature in question.

2.2 Speech and Instrumental Music

Features for detecting speech and music were developed by researchers in VTT Technical Research Centre of Finland. The classification of audio signal between speech and music is widely studied [1][2][4]. Our approach was based on kNN classification of discrete audio samples with the k value set to 3.

The extraction of confidence for a shot to contain speech or music was derived from the weighted average of speech/music classification results for the discrete 3 second portions of the signal.

The classification between speech and music was computationally very inexpensive, and was determined using only four power-related features. A three-second window of the signal was divided into frames of 50 ms overlapping by 10 ms, and the power inside every frame was calculated. The four used features were the variance of the frame-by-frame power, and the variance of the first and the second order differentials of the power, and finally, low energy ratio [5], which is computed as the percentage of 50ms frames with RMS power less than the threshold-percentage of the mean RMS power. A threshold level of 20% for low energy ratio was found to give best results, and the spread of the four features was increased by log transformations. In the training stage the features were normalized to zero-average and unit standard deviation. The translation and scaling parameters for each feature were stored and used in the classification stage to normalize the test signal.

The audio database used to train the system was assembled by sampling music from a vast assortment of CD's and by using a digital recorder to sample speech from Finnish radio broadcasts. All the samples were then converted to 22050 Hz mono. Both conversational speech and single speaker sections were used. The database also contained both male and female speech, sampled from several speakers. Music from various

styles and genres was also included in the training set. All samples were 15 seconds long and the length of the whole database was about 20 minutes for speech and 40 minutes for music.

The classification results of 3 second segments were presented as low-pass filtered time series. Low-pass filtering reduced the effect of separate classification errors and smoothed the transition points between longer segments of speech and music. In addition, mixed signals (containing both speech and music) that would produce a fluctuating series of classification results with a traditional binary decision classifier, are now presented as 'gray' areas that belong to both classes. The new trail of annotation labels shows the degree of certainty of belonging to either class for each three-second audio segment at a time. The numerical results were scaled between 0 and 1, and the weighted mean of the classifications inside each shot was used as the relevance measure for instrumental sound detection. The relevance for speech detection was determined as the inverse of the measure for instrumental sound, so that the sum of these values was always 1.

2.3 Results

The results show that TGC seemed to be most efficient in detecting cityscapes, which may be a result of more structured type of imagery in typical cityscape scenes. Another observation is the degrading effect of poor skin feature detection in the results. Even the detection of people was better using plain TGC, which points out the challenges of using color information in low-saturated and monochromatic videos. The detection of speech and instrumental sounds from audio signal had an averaged precision of 0.641 which was over two times higher than the average precision of 0.246 in people, city and landscape categories. The average precisions for detecting semantic features are shown in Table 1.

Table 1. Average precisions for different semantic features

	TGC	**TGC+SKIN**	**AUDIO**
People	0.248	0.168	-
Cityscape	0.299	0.197	-
Landscape	0.193	0.128	-
Speech	-	-	0.645
Inst. Sound	-	-	0.637

3 Manual and Interactive Search Tasks

3.1 Video Browsing and Retrieval System - VIRE

Our video browsing and retrieval system VIRE was used in manual and interactive search experiments. It is based on Java code and can be run on both SUN Solaris and Windows 2000 operating systems. System uses J2SE, QuickTime 6 for Java, WordNet dictionary and MySQL JDBC. The system consists of server and client applications, where server controls the querying through self-organizing feature index maps and provides the client query results. Client software offers two views, one for constructing video queries and another for browsing the database shots. References to physical media data and additional feature information are stored in MySQL-database. Figure 1 depicts the architecture of the VIRE system. The browsing view of the client offers content-oriented parallel navigation through database videos. The browsing procedure would be exhausting without efficient indexing structure. By utilizing self-organizing feature maps we were able to unravel the problems with computational requirements.

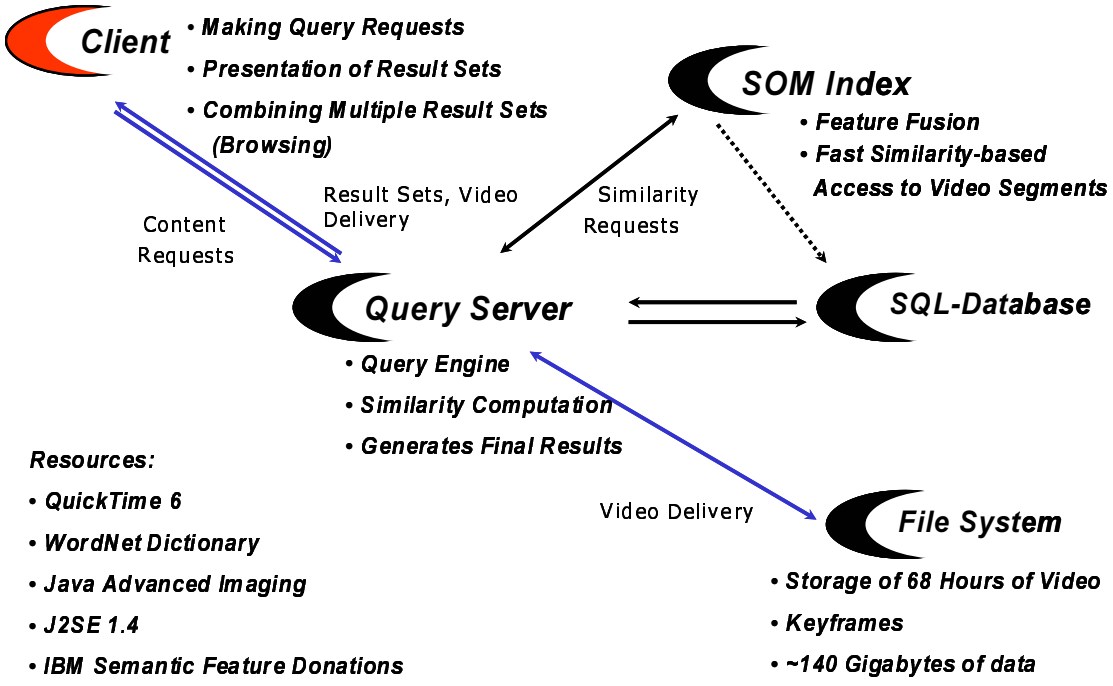

Figure 1. The overall architecture of VIRE, video browsing and retrieval system

3.1.1 Selected Features

Our system uses variety of features, which are used as the components for query. Features are based on different video modalities.

Generic Features
Generic features are measured from the physical properties of video data. Generic feature vectors of query and database shots are compared using distance metrics to measure dissimilarities. Different generic features can be used jointly to measure various physical properties simultaneously.

Color properties of a video shot were described with the Temporal Color Correlogram (TCC). Its efficiency against other color descriptors has been proven in [6][7]. TCC captures the correlation of HSV color pixel values in spatio-temporal neighborhoods. The feature captures temporal dispersion or congregation of color clusters unlike static color features. The parameters we used for the feature are described in [7]. TCC is computed over 20 video frames that are sampled evenly over a video shot.

Motion is a prominent property in video frame sequence. We computed features describing motion activity, based on definitions in MPEG-7 Visual standard [8]. Motion Activity descriptor defines following properties of motion: intensity, direction, spatial and temporal distribution of motion activity. Our system uses the following subset of those features:

- Intensity of motion is a discrete value where high intensity indicates high activity and vice versa. Intensity is defined as the variance of motion vector magnitudes normalized by the frame resolution and quantized in the range between 1 and 5.
- Average intensity measures the average length of all macroblock flow vectors.
- Spatial distribution of activity indicates whether the activity is scattered across many regions or in one large region. This is achieved measuring the short, medium and long runs of zeros that provide

information about the size and number of moving objects in the scene. Each attribute is extracted from the thresholded flow vectors obtained from the video data. All feature values are normalized so that they can be used jointly with other features in self-organizing indexes.

Audio contains a lot of information about the video shot content. And not just the spoken lexicon, but sounds and noises pinpoint details about video's semantic content. The real challenge here is however to choose meaningful parameters which somehow reflect the response to the various properties of sounds in the human aural perception. There has been extensive research on this topic (see [9] and references therein) primarily in connection with speech recognition systems [10]. We have selected following features to construct a generic audio feature descriptor [12][13]:

- Zero-crossing rate (ZCR), one of the most commonly used audio features, gives information about the spectral content of the signal. Taking into account relative spectral deficiency of the sound tracks of the movies in the VideoTREC database (which is obviously the result of their relatively old age) our measurements showed us that the bandwidth of a typical sound track was about 8 kHz. We feel that it is safe to assume that ZCR tracks the fundamental frequency f0 with quite high precision [9] due to the absence of high-frequency components. Some researchers believe [13] that ZCR is one of the most indicative and robust measures to discern unvoiced speech.
- RMS energy measures mean signal energy, which is what the human ear interprets as a notion of the acoustic volume, but it does not carry any information about the presence of transient sounds.
- Maximal and minimal energy are two parameters that correspond to the idea of "still" and "loud" sound, respectively. If the minimal energy is large and close to the value of the maximal energy, one can attribute the property of acoustic loudness to the entire sound clip. On the other hand, low maximal energy is a reliable indicator of a silent sound clip and hence it is useful for quick and easy silence detection.
- Mean and standard deviation sample energy gives a measure for scattering of the sample energies about their mean value. [11]
- Percentage of samples whose energy is below 50% of the mean energy of the sound clip has been used for the purpose of speech/music discrimination [14]. The fact is that the speech signal contains more "quiet" frames so this value will be higher for speech than for music.
- Percentage of samples whose energy is below 10% of the mean energy of the sound clip might carry some information about transient sounds and the purpose of using this feature was to improve scene detection. If the mean energy is significantly greater than that of the majority of sound samples, it might indicate the presence of short-term shooting-like acoustic events.
- Harmonicity ratio describes the proportion of harmonic components in the spectrum. The algorithm output is 1 for purely periodic signal and 0 for the white noise. A useful feature derived from harmonicity ratio was the length of the comb filter, which is an estimate of the delay that maximizes the autocorrelation function.
- Upper limit of harmonicity loosely defines the frequency beyond which the spectrum has no harmonic components. These features were calculated in 2048 sample windows that overlapped by 1024 samples. All audio files had sampling frequency of 22050.
- Spectral centroid is the center of gravity of the power spectrum. For audio signals that have clearly much energy in lower or higher parts of the spectrum this feature is useful.
- Spectral spread is the RMS deviation of the spectrum centroid, and thereby it describes if the spectrum is widely spread out or concentrated around its centroid. The used values were median values in one second window. These features were calculated in non-overlapping windows of 1024 samples.

To compute the dissimilarities of generic feature vectors, we have used L1 norm. Each generic feature vector is normalized prior inserting into self-organized index.

Semantic Features

Semantic features are single lexical concepts exist in a shot with certain confidence. Our system utilizes these concepts as binary filters to create subsets from results obtained with fuzzy generic feature dissimilarities. Different combinations of concepts can be used to create filter sets. In this case, the resulting subset will be the intersection of existing concepts in the initial result set.

We used following IBM donated semantic features: face, people, indoors, outdoors, instrumental sound, speech, landscape and text. These features had a confidence value that was thresholded into binary filter rules. The threshold value was decided upon criteria where approximately one third of the shots were assigned with the concept. Following threshold values were used: 0.62 for face, 0.75 for indoors, 0.35 for instrumental sound, 0.7 for landscape, outdoors, and people, 0.4 for speech, and 0.3 for text.

Text Features

Text features are derived from the automatic speech recognition (ASR) data that was made available for all participants by CLIPS-IMAG. The textual information was not used as a feature vector. Instead, it was used like semantic features as a filter for the results acquired from SOM-index structure. It was used to eliminate shots that did not contain indicated textual terms. To accomplish this, the ASR transcripts were indexed into a database treating the words as single-word terms. A stop word list was used to exclude grammatical and otherwise undiscriminating words that would have led to poor resolution. Qualified words were then lemmatized using the morphological processing features of the WordNet [15].

Because of the effects of lexical semantic phenomena such as homonymy or polysemy on the relevancy of the documents retrieved, techniques of word sense disambiguation (WSD) have been found useful in information retrieval to mitigate the effects of expressive power of natural language. Potentially relevant documents containing close synonyms of the query word such as ‘dog’ and ‘canine’ and hyponyms such as ‘Malamute’ will be missed unless queries are expanded by synonyms and hyponyms of the terms used in original user requests. [16]

To evaluate queries, we used relevance metrics to measure which shots were suitable given a set of topic related query words that were synonym expanded. The ranking of the shots was computed using Term Frequency Inverse Document Frequency (TDIDF) [17] based classification method that pinpoints relevant single-word terms occurring in the ASR transcripts. First, the given query words were automatically reduced to their base form. Second, the lemmatized query words were expanded with their synonyms using the WordNet [15]. Third, the relevance metric was computed for every shot that contained at least one of the words in the expanded query set. Finally, the neighbors of suitable shots were also included into the results within a time frame of 4 seconds. This was done for the sake of the temporal locality of the topic that spans over short shots that can hardly contain any ASR information.

Our approach encounter challenges as the video test material consisted of degraded audio quality, which seemed often lead to false detections of words. To use such data in a rather restricting filter yields inflexibility in the cases of erroneous interpretation of the spoken words.

3.1.2 Multi-modal Indexing Based on Self-organizing Maps

To avoid exhaustive searching on the server side, we used self-organizing index maps that are capable of finding metrically closest matches with any feature combination within tight timing requirements set by interactive video browsing that requires much parallel query processing. Our index structure consists of seven self-organizing maps (SOMs). Three of the SOMs are based on the primary generic features: color, audio and motion. Next three SOMs compose of joint features from the primary generic features: color&audio, color&motion and audio&motion. Last SOM uses all primary features: color&audio&motion. Each of the SOMs is generated with the SOM parameters set in the SOM Toolbox [18].

All the examples in a single query are processed individually and access to specific index maps is directed according the selected set of features. Each individual example launches best matching nodes –search returning a set of closest shots from the appropriate SOM.

Next these intermediate results are filtered using semantic feature or text description filters. Multiple filters can be selected, for example 'Outdoors', 'People' and text 'red Chevrolet'. The result sets are evaluated based on the existence of these filter terms and only the shots fulfilling the criteria are selected. Then, these sets of results are combined with fuzzy Boolean OR operator to form the final ranked result set. Different weights can be set to examples to change the order in the final, combined results. The selection of the amount of best matching nodes is guided by the source of the query. In browsing, the speed is emphasized over retrieval precision, so the number of best matching nodes is small. We have used 3, 10 and 70 best matching nodes in fast browsing, more precise browsing and manual querying, respectively. Figure 2 illustrates the indexing structure.

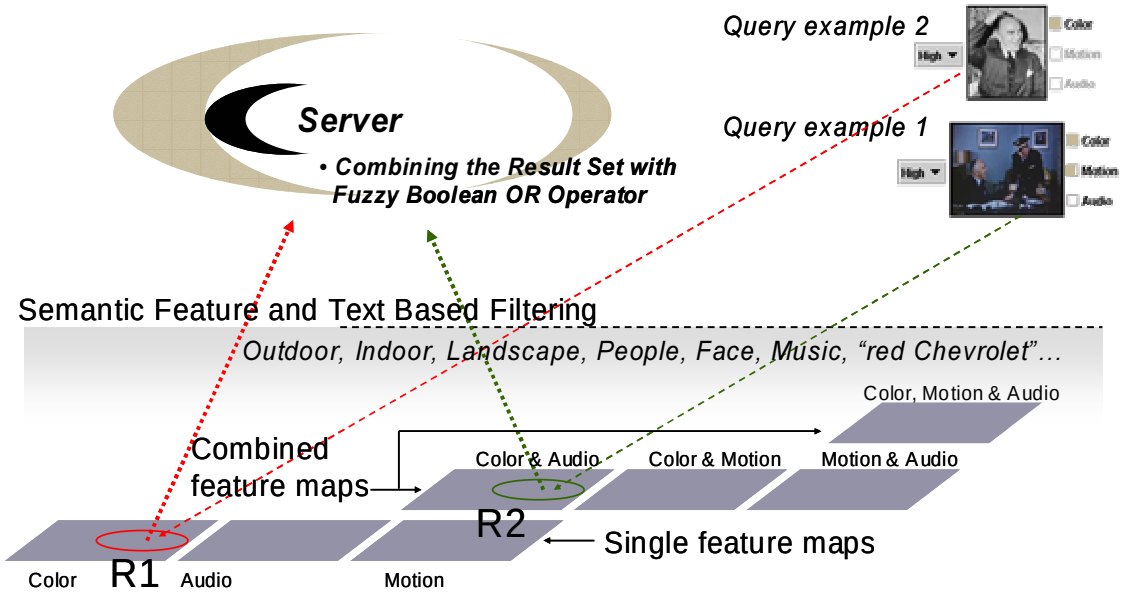

Figure 2. Self-organizing index structure organizes database shots into various maps.

3.1.3 Manual Query Interface

Figure 3 shows the query interface devised for manual search. The view offers selections of various search features and filters in the left side and resulting shots ranked by their similarity with the selected query

attributes are displayed in the right side view. Query topics can be changed from the Topic menu. This will change the example shots or images provided by the topic description to the lower left panel.

User can select any combination of three generic features (color, motion or audio) for any example shot, but for example image only color feature is supported. User can enable any set of semantic filters in addition to the selected generic features. Additionally, a lexical word filter can be constructed by selecting a set of words in the upper left panel of the interface. At least one example shot with one generic feature enabled is the minimum requirement for submitting a query. From the result set, user can select any interesting shot as a start point for navigation in the browsing interface.

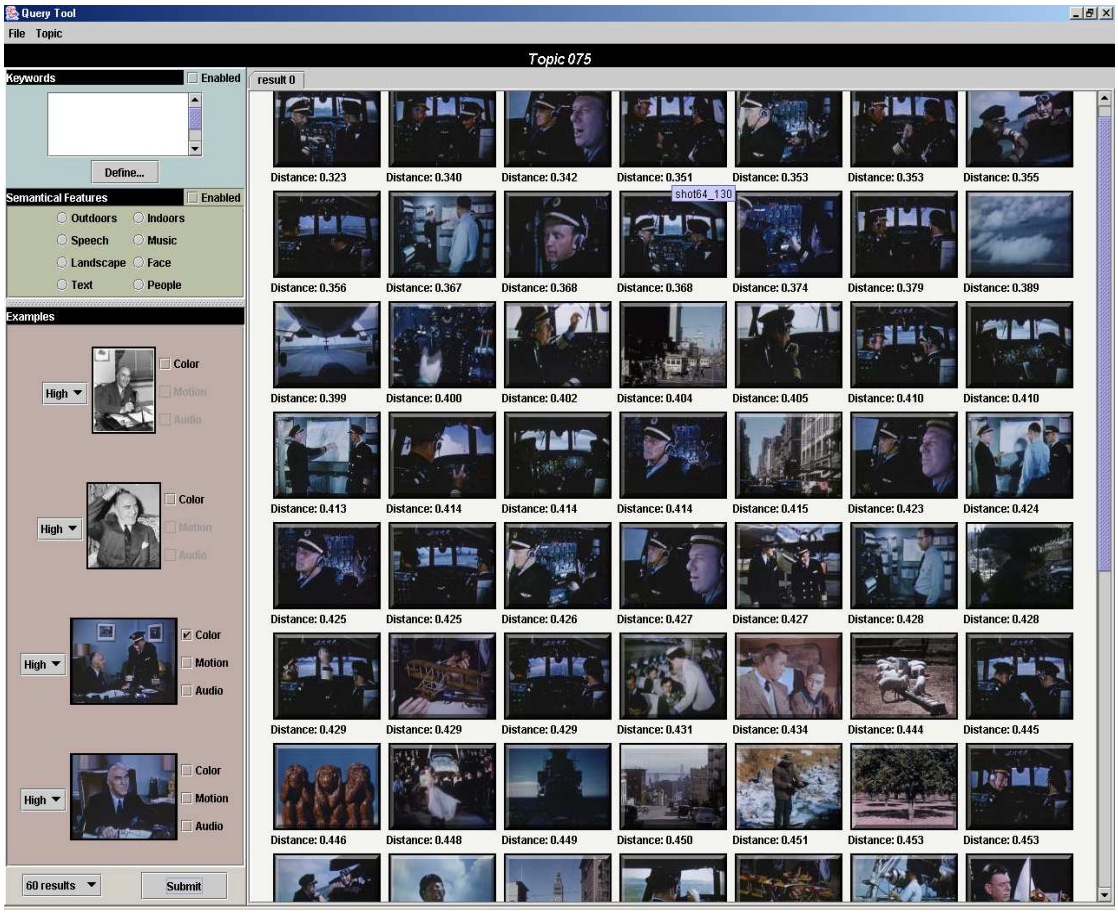

Figure 3. Manual query interface

3.1.4 Interactive Browsing Interface

The novelty of our approach in the interactive search task relies on two aspects: interface design and content-based feature processing. The motivation is to reduce the effect caused by ambiguous results that are usually obtained from a traditional content-based example search. Currently, two dominant approaches are used to realize video searching and browsing. Systems either select content-based presentation of video items or rely on more traditional time-line based organization into temporally adjacent items. The disadvantages of the approaches are in their incapability to associate computed features with the user's information need (ambiguity of content-based approaches) and to provide a holistic view over the linear temporal presentation (inefficiency of the time-line based browsing).

Our approach combines both inter-video similarities and local temporal relations of video shots in a single interface. In interactive search, users want the computer to act as a 'humble servant', providing enough cues and dimensions for users to navigate through the vast search space towards the relevant objects. In VIRE, users can perform *content-oriented browsing* that combines timeline presentation of videos with content-based retrieval. Content-oriented browsing implies that the video content is not utilized alone, but in conjunction with temporal video structure.

Figure 3 illustrates the browsing interface. The panel showing the first row of key frame images displays sequential shots from a single video in a chronological time-line. At any time, user can scroll through the entire video shot sequence to get an overview of the video content. The leftmost key frame in the top row shows always the first shot and name of the video. The first shot may contain initial setup for the entire video, so by viewing it, user can get instant idea about the semantic setting for the rest of the video shots.

The lower right panel gives user another content-oriented view, but this time from the entire database. The columns below the topmost shots show the most similar matches organized in top-down rank-order. The columns generate a similarity lattice that provides linkage to other database videos. The similarity is measured based on the features selected in the lower left panel. User can select a single feature or any combination depending on what properties they want to browse with.

Additionally, user can decide whether he want to include shots from the query video to be shown on a similarity lattice. When user locates interesting shots from the lattice, he can open the video in the topmost row so that the interesting shot is located in the middle column. After updating the shots in the topmost row, system re-computes the similarity lattice. At any time, user can update the current lattice using other feature combinations. The features that can be used in browsing are color, motion, audio and texture.

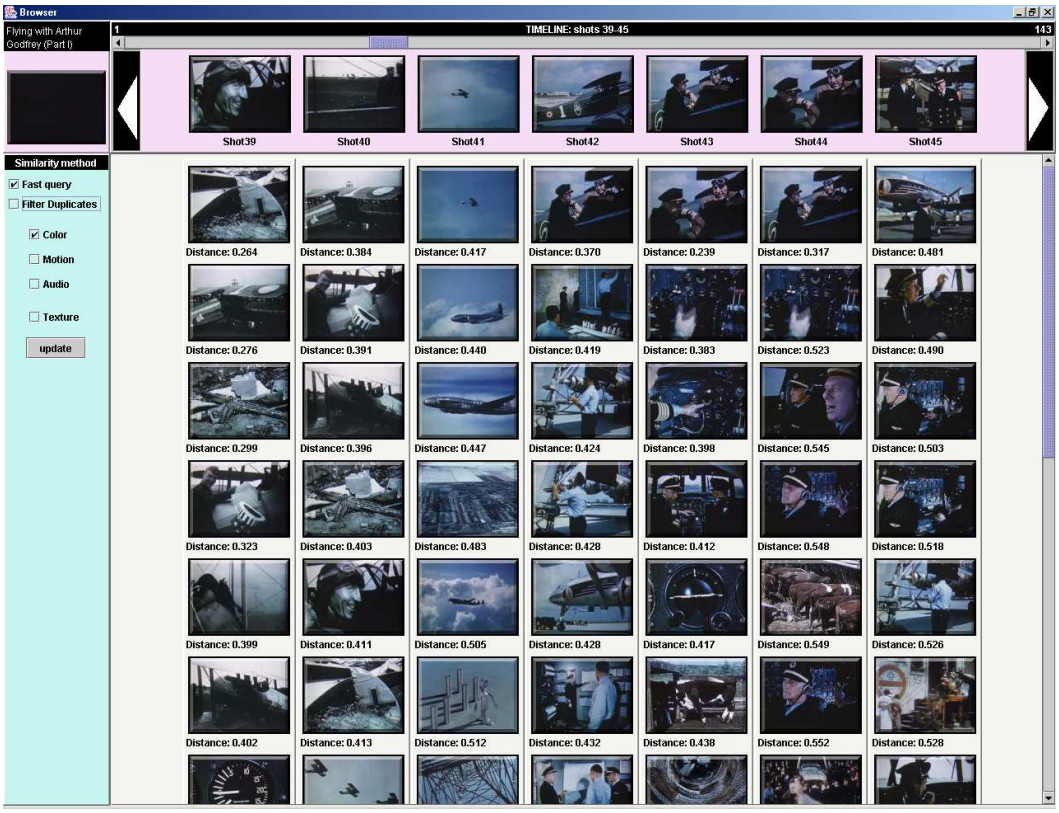

Figure 3. Content-oriented browsing interface. Lower right panel shows the similarity lattice.

The requirements to update the similarity lattice are heavy, since the browsing speed should be close to real-time. To update a single lattice, system must perform parallel query processing in several individual example-based queries. Multi-threaded index queries to self-organizing maps provide efficient access mechanism to proximity search for every feature combination.

3.2 Experimental Setup

We tested our system in both manual and interactive search. NIST provided 25 search topics that contained varied from very specific ('Find shots with Eddie Rickenbacker') to more generic concepts ('Find overhead views of cities'). Topic included one or more example clips of video or images to aid the search process. From the image examples, only a color correlogram could be used as a generic search feature whereas the video examples offer color, motion and audio or any of their combinations. All topics were processed both manually and interactively in search test video collection that consisted 40 hours of video material. NIST also provided segmentation for all videos with more than 24000 video shots, from which over 14000 belonged to the search test collection.

In manual query the user initiated the query by selecting appropriate generic feature cues from the topic examples, sets weight to indicate relevant and non-relevant examples and decides appropriate filters. Three different search configurations were tried for each search topic: search using generic features only, search using generic features with semantic feature filtering, and search using generic features with text feature filtering. Self-organized index was used with 70 best matching nodes.

A group of eight new users carried out the interactive search. Test users, most of them males, were information engineering undergraduate students, having good skills in using computers, but less experience in searching video databases (obviously at least somewhat experienced in www-search). Every user reported to be somewhat familiar with the search topics. 25 topics were divided into four sets that were randomly given to the test users so that two users carried out the same set of topics. All users were given half an hour introduction to the system, with emphasis on the search and browsing interface functions demonstrated with couple of example search. Users were told to use approximately ten minutes for each search, during which they navigated in the shot database and selected shots that seemed to fit to the topic description. Users were also told to fill a questionnaire about their experiences. The machines that the system was running on were 400-800 MHz PCs with Windows 2000 operating system. After the tests were finished, the two result sets for one topic were joined by removing duplicate matches and averaging their rank values.

3.3 Results

Average precisions for the four different search configurations are shown in Table 2. Manual configurations used either only Generic Features (color, motion or audio), Generic Features together with semantic feature filtering (outdoors, landscape) or Generic Features together with text filter ("red Chevrolet"). Number of hits at depth 10 describes the total amount of found matches when considering the 10 best ranked matches for a topic. Overall average shows the mean value of the average precisions in 25 topics. As can be seen, the performance of interactive search overcomes greatly the manual search results. This indicates clearly the importance of the human factor in search. The average time in making an interactive browsing was 10.08 minutes. During this time, the average of 12.7 matches were found from the database. This means that average of 1.26 matches were found during each minute of searching. Another interesting observation is the counter-effectiveness of semantic and text filters to the results.

Table 2. Results for different configurations over 25 search topics

Type of search configuration	Nr. of hits at depth 10	Overall Average
Interactive	151	0.26
Manual (Generic Features)	38	0.03
Manual (Gen. Features + Sem. Filters)	22	0.02
Manual (Gen. Features + Text Filter)	12	0.01

The six most successful topics are listed in Table 3, best resulting topics being topmost. Finding George Washington, Price Tower and James H. Chandler overall seemed to be the most successful topics in manual and interactive searching across the participating system runs.

Table 3. Six most successful topics (topic number in parenthesis)

Interactive Search	Manual Search (Generic Features)
James H. Chandler (76)	James H. Chandler (76)
George Washington (77)	Microscopic living cells (97)
Parrots (91)	Eddie Rickenbacker (75)
Price Tower in Oklahoma (84)	Musicians (80)
Microscopic living cells (97)	Snow covered mountains (90)
Nuclear explosion (95)	Overhead view of cities (86)

According to the answers in questionnaire, the VIRE system was easy to learn, but somewhat harder to use. This was due to the ambiguous results that fuzzy-based similarity measurements return in many instances. The browsing interface was appreciated, although the near-real time responses in updating the browsing view would have been preferred to be completely real-time.

4 Conclusions

We approached content-based video retrieval from many different perspectives in TREC 2002 Video Track. Our experiments showed that the extraction of semantic concepts is a challenging task. However, the results in detecting instrumental sound and speech were promising. Visual semantic features have still a lot to improve: the low visual quality of the videos definitely affected the results, though.

The most successful component in our work was the content-oriented browsing that gave the computer merely the role of an assistant in the cognitive process of semantic searching. Our system was subordinated to provide the user multiple parallel paths from which he could choose the direction for his navigation independently. We combined the temporal connectivity of temporally adjacent shots and fuzzy shot similarities into one view to provide the user comprehensive information about the inter-relations between the video shots.

The manual search methods must still overcome big challenges to reach a satisfactory level of performance in the high-level semantic search problems. However, there are search problems that are more suitable to automatic search than others, for example locating cityscape views or retrieving shots containing speech. It seems that while efficient features can be computed from different modalities, they are not alone appropriate for automated semantic retrieval. Does the missing link lie within a single feature quality, or rather in a way to combine multiple modalities? This still remains an intriguing research problem.

5 References

[1] Carey M, Parris E & Lloyd-Thomas H (1999) A comparison of features for speech, music discrimination. Proc. ICASSP.

[2] Hoyt J & Wechsler H (1994) Detection of human speech in structured noise. Proc. ICASSP.

[3] Kedem B (1986) Spectral analysis and discrimination by zero-crossings. Proc. IEEE. Vol. 74, No. 11.

[4] Saunders J (1996) Real-time discrimination of broadcast speech/music. Proc. ICASSP.

[5] Scheirer E & Slaney M (1997) Construction and evaluation of a robust multifeature speech/music discriminator. Proc. ICASSP.

[6] Ojala T, Rautiainen M, Matinmikko E & Aittola M (2001) Semantic image retrieval with HSV correlograms. Proc. 12th Scandinavian Conference on Image Analysis, Bergen, Norway, pp.621-627.

[7] Rautiainen M & Doermann D (2002) Temporal color correlograms for video retrieval. Proc. 16th International Conference on Pattern Recognition, Quebec, Canada.

[8] MPEG-7 standard: ISO/IEC FDIS 15938-3 Information Technology - Multimedia Content Description Interface - Part 3: Visual.

[9] Gerhard D (2000) Audio signal classification, School of Computer Science, Simon Fraser University, Burnaby, Canada.

[10] Rabiner L & Juang B-H (1993) Fundamentals of speech recognition, Prentice Hall.

[11] Gray R & Davisson L (1985) Random processes: a mathematical approach for engineers, Prentice-Hall Information and System Sciences Series, Prentice Hall.

[12] MPEG-7 standard: ISO/IEC FDIS 15938-4: Information technology -- Multimedia content description interface -- Part 4: Audio

[13] Wang Y, Liu Z & Huang JC (2000) Multimedia content analysis, IEEE Signal Processing Magazine, November 2000, pp.12—36.

[14] Scheier E & Slaney M (1997) Construction and evaluation of a robust multifeature speech/music discriminator, Proc. ICASSP-97, April 21-24, Munich, Germany.

[15] Fellbaum C (1998) WordNet: An electronic lexical database. The MIT Press.

[16] Jurafsky D & Martin JH (2000) Speech and language processing: an introduction to natural language processing. Computational Linguistics, and Speech Recognition. Prentice-Hall.

[17] Salton G & Yang C (1973) On the specification of term values in automatic indexing. Journal of Documentation, Vol. 29, 351–372.

[18] SOM Toolbox (2.11.2002) http://www.cis.hut.fi/projects/somtoolbox/documentation/somalg.shtml

[19] Vailaya A, Jain A & Zhang HJ (1998) On image classification: city vs. landscape. IEEE Workshop on Content-Based Access of Image and Video Libraries, pp. 3-8.