TRECVID 2003 Experiments at MediaTeam Oulu and VTT

Mika Rautiainen[†], **Jani Penttilä**[‡], **Paavo Pietarila**[‡], **Kai Noponen**[†], **Matti Hosio**[†], **Timo Koskela**[†]
Satu-Marja Mäkelä[‡], **Johannes Peltola**[‡], **Jialin Liu**[†], **Timo Ojala**[†] **and Tapio Seppänen**[†]

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

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

Abstract

MediaTeam Oulu and VTT Technical Research Centre of Finland participated jointly in semantic feature extraction, manual search and interactive search tasks of TRECVID 2003. We participated to the semantic feature extraction by submitting results to 15 out of the 17 defined semantic categories. Our approach utilized spatio-temporal visual features based on correlations of quantized gradient edges and color values together with several physical features from the audio signal. Most recent version of our Video Browsing and Retrieval System (VIRE) contains an interactive cluster-temporal browser of video shots exploiting three semantic levels of similarity: visual, conceptual and lexical. The informativeness of the browser was enhanced by incorporating automatic speech transcription texts into the visual views based on shot key frames. The experimental results for interactive search task were obtained by conducting a user experiment of eight people with two system configurations: browsing by (I) visual features only (visual and conceptual browsing was allowed, no browsing with ASR text) or (II) visual features and ASR text (all semantic browsing levels were available and ASR-text content was visible). The interactive results using ASR-based features were better than the results using only visual features. This indicates the importance of successful integration of both visual and textual features for video browsing. In contrast to previous version of VIRE which performed early feature fusion by training unsupervised self-organizing maps, newest version capitalises on late fusion of features queries, which was evaluated in manual search task. This paper gives an overview of the developed system and summarises the results.

1 Introduction

This paper first describes our experiments in the task of detecting semantic features. Then it introduces our recent version of prototype Video Browsing and Retrieval System VIRE and the experiments in manual and interactive search tasks. In the end of the paper concluding remarks are given.

2 Semantic Feature/Concept Detection

To avoid confusion with low-level features, we denote semantic features as semantic concepts.

2.1 Visual Features

The efficiency of visual low-level features in the detection of semantic concepts was tried out in our TREC 2002 Video Track experiments [15]. The obtained results attracted towards further testing with a larger set of concepts. In brief, the propagation of small example sets uses the local neighbourhoods and geometrical distances between low-level feature points to define high-level concept confidences for every shot in a feature space. We have observed in [16] that different semantic concepts (e.g. high-level features) may co-exist in a video shot. Our visual concept detectors do not cluster space into separable classes, but instead provides continuous confidence measures for all items in the space. Due to this, it has been possible to reduce amount of training into small positive example sets for each concept. The method uses following three temporal low-level features taken from the visual content of a video shot.

Motion Activity

Motion is a prominent property in a video frame sequence. We have used features describing motion activity, based on definitions of MPEG-7 Visual standard [4] MPEG-7 Motion Activity descriptor computes four features from the motion vectors within a video sequence: discrete intensity, direction, spatial distribution and temporal distribution of motion activity. Our system uses the following four values from these 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 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 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.

Temporal Color Correlogram (TCC)

Color properties of a video shot were obtained into a Temporal Color Correlogram (TCC) feature. Its efficiency against traditional static color descriptors has been indicated in [2][3]. TCC captures the correlation of HSV color pixel values in spatio-temporal neighborhoods. The feature approximates temporal dispersion or congregation of uniform color regions within a video sequence. Traditionally static color features do not consider temporal dynamics as they are often computed from a single keyframe. TCC is computed by sampling 20 video frames evenly over bounded video sequence. The details about the computational parameters (such as color space quantization and spatial constraints) are described in [3].

Temporal Gradient Correlogram (TGC)

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 bounded video sequence. Due to temporal sampling, autocorrelogram is able to capture also temporal changes in spatial edge orientations. From each sample frame, the edge orientations (obtained with Prewitt kernels) are quantized into four segments depending on their orientation being horizontal, vertical or either of the diagonal directions. The details of the feature and the used parameters can be found from the study utilizing it in the detection of city/landscape scenes and people from video shots [16].

Feature Fusion and Combined Result Sets

A simple method of propagating small positive example sets was used in the experiments to provide final confidence values for the high-level concepts. In brief, positive examples are used to create ranked result sets that are later combined with a variant of Borda count voting principle [17].

Geometric distances (dissimilarities) to the example shot in low-level feature space result in initial rank ordered lists $D_l^f(k)$, where semantic concept *f* of an example shot *k* is propagated by 'confidence ranks' to its nearest neighbours in a low-level feature space *l*. The ranked result list for each example was computed with L$_1$-distance in the feature space *l*. Then, ranked lists $D_l^f(k)$ created from *L* feature spaces were combined with variants of Borda count. In this work, two fusion approaches based on minimum and sum operations were experimented:

$$R_n^f(k) = \Theta(\frac{D_1^f(k)}{D_{1\max}^f(k)},...,\frac{D_L^f(k)}{D_{L\max}^f(k)}) \qquad \textbf{(1)}$$

where $R_n^f(k)$ = fused rank of a result shot *n* to the feature result sets 1…*L*

$D_l^f(k)$ = rank to the query example *k* representing semantic concept *f* in the feature space *l*

$D_{l\,\max}^{f}(k)$ = maximum rank to the query example k in its result set. Equals to the size of data

Θ = fusion operator, minimum or sum

After fusion with low-level feature results is achieved, feature *f* result set $R^{f}(k)$ contains the items with ranks describing the 'voted' confidence to the example shot *k*. Next, the sets of results $(R^{f}(1),..., R^{f}(K))$ are further combined with a method that considers the ranks as votes. The method selects the best (minimum) ranks for each item to form the final confidence rank. This operation, which is congruent with a Boolean *OR* operation in fuzzy systems, further sorts and selects the 2000 most confident result shots organized by the confidence rank S_n^f . The following formulae describe the combination procedure:

$$S_n^f = min(\frac{R_n^f(1)}{R_{\max}^f(1)},...,\frac{R_n^f(K)}{R_{\max}^f(K)}) \qquad \textbf{(2)}$$

$$C_f = \left[sort\{S_1,...,S_N\}\right]_{2000} \qquad \textbf{(3)}$$

where S_n^f = minimum rank of a result shot n to the examples $1…K$

$R_n^f(k)$ = rank to the query example k with a feature f

$R_{\max}^f(k)$ = maximum rank to the query example k in its result set $R^{f}(k)$. Equals to N

C_f = Final ranked set of results for the semantic concept *f*

$[\ ]_X$ = X top-ranked items in a sorted list

N = Size of the items in the feature space, equals to the size of the TRECVID 2003 test set

Face Detection

Previously described low-level features are insufficient for some high-level feature tasks in TRECVID 2003 as some of the features were based on presence of a person in a shot. Features 'news subject face' and 'news subject monologue' required the detection of at least one person (a news subject) to be contained in a shot. In order to detect these semantic concepts, face regions were located from the shot's key frame images, which were provided by NIST. Prior the actual face detection, key frames containing skin were filtered to reduce computational effort of a face classifier. Skin regions were detected from 10x10 image sub-blocks according a method introduced in [16]. The connected skin-block regions were further scrutinized using face detector that extracted facial features with overlapping 19x19 windows in multiple resolutions. The facial features were classified with trained support vector machine [6]. Finally, heuristic rules were used to reduce false and overlapping detections. The face detector was empirically found to give apt results for the 'news subject face' when one or two face regions were detected per key frame. Empirical findings were obtained using TRECVID 2003 development set.

2.2 Audio Features

MPEG7 Audio

spreaddev, harmdev, flengthdev: Two of the basic algorithms were selected from the MPEG-7 Part 4 Audio Standard [12]. Low-level features that were selected from MPEG-7 were spectral spread and harmonicity ratio. 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. The standard deviation of the harmonicity ratio over one shot was used as a feature. 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. Using this feature it was possible to classify brown noise-like signals such as car motor sound, which in general have the smallest possible comb filter length. The feature for one shot, was the standard deviation of the comb filter length values. 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. Used values were standard deviation values over a shot.

F0dev, F0med: The F0 was calculated as presented in [13]. F0 was used for two purposes: F0 detection was used to make a voiced/unvoiced decision and its value as a basic feature. Features derived from F0 were: standard deviation and median over a shot. They were naturally used for detecting speech related semantics: female voice detection and finding monologue parts of the audio signal.

clustmean, clustmed, clustdev: In order to be able to separate monologue and conversation parts we used features derived from K-means clustering of mel-cepstrum coefficients. Mel-cepstrum models the human perception and it is widely used in speech and speaker recognition [14]. The method used calculates mel-cepstrum coefficients for voiced frames and clusters the coefficients to 10 clusters with K-means algorithm. It returns the central vectors of 10 clusters for each frame in a shot. The distances between clusters in current and previous frames are calculated and the cluster distance mean, median and standard deviation were used as features to distinguish between monologue and conversation frames.

SNR: Signal to noise ratio, SNR, for each shot was calculated. It was calculated through the power between voiced and unvoiced frames in a shot. This was thought to give an estimate of the outdoor shots as a higher background noise level of the outside recordings.

Additional Features

We computed the amplitude envelope functions, relative difference functions and onsets, based on methods described in [9] and [10], and derived various low-level features based on this data. From this feature set we selected the following features for semantic concept detection:

spectral_kurtosis: This is the averaged kurtosis calculated across the frequency domain on each of the sample points on the subband amplitude envelopes. This measure has its highest values when the subband amplitude envelopes have a sharp peak in their frequency distribution and low values when the subband energy is evenly distributed across the frequencies. Thus, this measure is directly related to the spectral flatness of the shot.

onset_kurtosis: This measure gives the average spectral kurtosis of onsets. It is calculated from the relative amplitude envelopes at time points respective to onset times and gives a measure of how evenly the onset subband energies are distributed across the frequency domain.

hi_onset_ratio: This is the proportion of onsets whose frequency distribution is slanted towards high frequencies.

spectral_variance: The averaged variance taken across the frequency domain on each of the sample points on the subband amplitude envelopes. This gives a measure related to the spectral volatility of a shot.

ioi_dev: The standard deviation of the inter-onset intervals within a shot. This feature has its lowest values when the onset trail is either periodical and steady or very fast.

onset_loudness_variance: The variance of onset loudness within a shot.

onset_skewness: Mean onset skewness is a measure of the asymmetry of the frequency distribution of onsets. It is calculated from the relative amplitude envelopes at time points respective to onset times. We use this measure as a supplement to f0-based features in female speech detection since the onsets in female speech are dominantly distributed towards the higher frequencies.

*zero-crossing ratio: Z*ero-crossing ratios were calculated as described in [1].

Feature Selection and Classification

The classifier used was a standard Quadratic Gaussian Classifier. The training data features were normalized to a unit variance and zero mean. The features were selected for each class by first selecting a larger feature base by heuristic methods. Then the features obtaining the best classification performance were validated for the intermediate level classification.

Initially the classification was used to produce intermediate semantic concepts from opposing semantic pairs collected from the training data. Because the annotations of the test data were found to be unreliable, all training data was manually verified. Table 1 shows the intermediate semantic concepts and the features selected for their classification. The training data was selected from the collaborative annotation data including all siblings for a given concept.

Semantic Concept Pairs	Used Features
‘Outdoors’ *vs.* ‘Indoors’	*spectral_kurtosis, onset_kurtosis, SNR, zero-crossing ratio*
‘Monologue’ *vs.* ‘Conversation’	*hi_onset_ratio, F0dev, clustmean, clustmed, clustdev*
‘Male Speech & Monologue’ *vs.* ‘Female Speech & Monologue’	*onset_skewness, F0med*
‘Sport’ *vs.* ‘Monologue & Conversation’	*spectral_variance, ioi_dev, onset_loudness_variance, SNR*
‘Vehicle Noise’ *vs.* ‘Monologue & Conversation’	*spreaddev, harmdev, flengthdev*

Table 1. Intermediate semantic concepts

The ‘Conversation’ concept was created from the shots annotated as ‘Speech’, but not as ‘Monologue’. The output of the classifier was considered as the confidence for a concept.
Table 2 shows the final semantic concepts and the intermediate semantic classifiers that produce confidence value for the first class. Speech/music discriminator [11] confidence was combined with speech related concepts to improve the detection. For the combination of intermediate classifier outputs, we have used exactly the same procedure than in fusion of multiple modalities (see Eq. 4 in Chapter 2.3).

Final Semantic Concept	Intermediate Semantic Classifiers and Combinations
‘Outdoors’	Confidence of ‘Outdoors’ *vs.* ‘Indoors’
‘News Subject Monologue’	Confidence of ‘Monologue*’ vs.* ‘Conversation’, masked with results from ‘Speech/music’.
‘Sport’	Confidence of ‘Sport’ *vs.* ‘Monologue & Conversation’
‘Female Speech’	Confidence of ‘Male Speech & Monologue’ *vs.* ‘Female Speech & Monologue’, masked with results from ‘Speech/music’
‘Car, Truck, Bus’	Confidence of ‘Vehicle Noise’ *vs.* ‘Monologue & Conversation’

Table 2. Final semantic concepts for run MT6. The outputs of intermediate classifiers
are used to create confidence values (for the underlined name) to generate the final concept.

2.3 Fusion of Visual and Audio Features

In order to find the effect of combining multiple modalities, late fusion of independent feature detector outputs (auditory and visual) was experimented with four semantic concept categories: ‘Outdoors’, ’News subject monologue’, ‘Sporting event’ and ‘Physical violence’. The fusion was achieved according the Borda count variant method explained in Chapter 2.1. In order to emphasize results that have correlations between detector outputs, equation 1 was modified to sum the normalized confidence ranks of separate feature detectors instead of selecting the minimum (strongest confidence rank) from the feature detector outputs:

$$S_n^f = sum(\frac{D_{audio}^{fx}}{D_{audio\max}^{fx}}, \frac{D_{visual}^{fy}}{D_{visual\max}^{fy}}) \qquad \textbf{(4)}$$

, where S_n^f = sum of the normalized confidence ranks of item *n* in the visual and audio result sets

D_k^f = confidence rank of item *n* using a detector *k* in a feature *f.*

$fx = fy$ (fusing same feature detector outputs) or $fx \neq fy$ (fusing different feature detectors)

$D_{k\max}^f$ = maximum rank to the detector *k* in its result set R_k . Equals to test set size in TRECVID 2003

2.4 Semantic Feature Task Experiments in TRECVID 2003

Seven runs were submitted for TRECVID 2003 semantic concept task. Table 3 shows the definitions of the submitted runs and their identifiers. Some of the runs used a fixed visual feature configuration whereas others used pre-validated feature configurations for each semantic concept. In this chapter, a group of semantic concepts were tested with varying visual feature configurations. The group will be denoted with symbol Ψ , consisting of the topics 1,3,4,5,6,7,9,10,12,13,14,16.

Run ID	Used Features	Feature fusion method	Semantic Concepts
FI_OU_MT1	*TGC, TCC, MA*, face, f0	*MIN, SUM*	*Ψ*, 2, 8
FI_OU_MT2	TGC & TCC	MIN	Ψ
FI_OU_MT3	MA & TGC &TCC	SUM	Ψ
FI_OU_MT4	MA & TGC & TCC	MIN	Ψ
FI_OU_MT5	TGC	-	Ψ
FI_OU_MT6	*MPEG-7 Fea, Additional Fea*	Gaussian classifier & SUM	*1,8,9,11,13*
FI_OU_MT7	*Visual, Audio Concepts*	SUM	*1,11,13,16*

Table 3. Submitted semantic concept runs for TRECVID 2003. Cells with cursive text reflect that the run used a validation process to obtain the best configuration for each semantic concept, normal text indicates that the property is fixed for every feature in the run. Ψ ={1,3,4,5,6,7,9,10,12,13,14,16}

Results of Visual Runs

Runs from MT1 to MT5 tested sets of visual feature configurations and fusion methods for a fixed group of semantic concepts Ψ. Run MT1 consists of pre-validated (best) configurations for each concept in Ψ. Additionally, it also contains results for concepts 'news subject face' and 'female speech' that had specifically designed detectors (see respectively 2.1 and 2.2 for more details). Validation of the configurations in MT1 was obtained in TRECVID 2003 development set using collaborative annotation data [18].

The effect of various feature configurations and fusion methods is visible in Table 4. Run-wise means of the average precisions (MAP) show that pre-validated configurations of MT1 obtained the best overall detection result. Best non-validated result with fixed features was obtained using TGC and TCC with minimum rank feature fusion. Surprisingly the use of motion didn't lead to any improvement from this. Using only one feature, TGC, provided the weakest MAP. Semantic concepts with the highest individual results were 'weather news' (MT1: 0.501), 'news subject face' (MT1: 0.107), 'sporting event' (MT1: 0.106) and 'people' (MT1: 0.096). For the fusion methods, minimum rank performed better than sum of ranks.

About domain specific features, our 'news subject face' detector obtained the average precision of 0.107, whereas the median and maximum results were 0.0835 and 0.182 (26 runs submitted results for 'news subject face').

Run ID	MAP of Group Ψ
FI_OU_MT1	0.090
FI_OU_MT2	0.075
FI_OU_MT3	0.037
FI_OU_MT4	0.063
FI_OU_MT5	0.043
Median	0.084
Max	0.314

Table 4. Mean of the average precisions (MAP) for Ψ. Table also shows median and maximum of MAP for 26 runs that submitted results for these concepts.

Results of Audio Runs

The results obtained using only audio features were submitted in run MT6. Table 5 shows the average precisions for the individual concepts. Generally speaking, the results using audio were rather low. However, by fusing the detectors with visual features, encouraging results were obtained for some concepts (see multi-modal results for details).

Feature ID	Audio Run (MT6)
Outdoors (1)	0.013
News subject monologue (11)	0.003
Sporting event (13)	0.070
Female speech (8)	0.042
Car/Truck/Bus (9)	0.005

Table 5. Average precisions for the five semantic concepts. Best performing concept is 'sporting event' whereas 'news subject monologue' is the worst

Result of the Multi-modal Run

Run MT7 contains four multi-modal results obtained from combinations of two detectors (visual and audio-based). The selection of feature detector combinations is based on empirical findings with TRECVID 2003 development set and collaborative annotation results. The used fusion method is explained in Chapter 2.3. Table 6 shows run concepts and their detector combinations.

	Combinations of Semantic Concepts	
Semantic Concept	**Visual Semantic Concept Detector**	**Audio Semantic Concept Detector**
'Outdoors'	'Outdoors' from MT1	'Outdoors' from MT6
'News Subject Monologue'	'News Subject Face' from MT1	'News Subject Monologue' from MT6
'Sporting Event'	'Sporting Event' from MT1	'Sporting Event' from MT6
'Physical Violence'	'Physical Violence' from MT1	'Music' from 'Speech/music' detector (see 2.2)

Table 6. Final semantic concepts

Table 7 shows the average precisions for multi-modal feature detectors in four semantic concept categories. By comparing the results to the individual visual and audio-based detectors, one can see that in two concepts the multi-modal features provide the best overall result.

Feature ID	Combined(MT7)	Visual(MT1)	Audio(MT6)
Outdoors (1)	0.035	**0.052**	0.013
News subject monologue (11)	**0.013**	-	0.003
Sporting event (13)	**0.245**	0.106	0.07
Physical violence (16)	0.024	**0.031**	-

Table 7. Average precision for the four semantic concepts. Table also shows corresponding result from a single modality detector, if available.

3 Manual and Interactive Search

3.1 Video Browsing and Retrieval System

Our video browsing and retrieval system VIRE was used in manual and interactive search experiments. System uses J2SE, QuickTime for Java and MySQL JDBC components. The system consists of a server and client(s); the server provides query services for various feature indexes and formulates the final result set by fusing the sub-results of independent features. Client software offers two views, one for constructing automatic video queries and another for browsing. References to media data and additional meta-information are stored in MySQL-database. The browsing view of the client offers cluster-temporal [19] navigation to the database videos.

VIRE system provides search features in three semantic levels. Lowest level provides visual similarity between two video objects and is also known as a content-based example search. Second level draws out the semantics from low-level features by pre-trained semantic concept detectors. Third level is highly semantic textual search from automatic speech recognition transcripts.

Visual Similarity with Generic Features

Generic features are measured from the physical properties of visual video data. Generic feature vectors of query and database shots are compared using a geometric distance metric that measures dissimilarities. Different features can be used simultaneously in searching. The results of individual feature queries are combined with late fusion techniques.

The visual similarity is constructed from three physical properties of a shot: (I) *Color* is the most widely used content property in CBVR research. Similarity by color gives initially very good perceptual correspondence between two color images that are small or short of details. After images are reviewed in detail, other properties for similarity emerge. (II) *Structure* of the edges in a visual imagery is a strong cue for many computer vision applications, such as classification city and landscape images or segregating natural and non-natural objects. This can also provide invaluable in queries where statistical color information is insufficient in describing the main properties of an image. (III) *Motion* is a property that is intrinsic for multimedia video. It describes properties that cannot be perceived from a static image. In some occasions it can be the only property that can reliably describe the content of a sequence, a fast-paced action would be an example of such. Following features from these categories have been used in our experiments: Temporal Color Correlogram (TCC), Temporal Gradient Correlogram (TGC) and Motion Activity (MA). These features are already described in Chapter 2.1. To compute the geometric dissimilarities of the described feature vectors, we have used L1 norm. Each feature vector is normalized prior computing the dissimilarity.

Conceptual Similarity with Semantic Concepts

Semantic features are lexically defined concepts that exist in a shot with a certain confidence value. Trained semantic concept detectors are used to create a series of confidences for each video shot as described in Chapter 2. Our search system utilizes concept confidences as features that are being compared against other shots' confidence values. The problem with floating confidence values is that they are usually generated with an algorithm that attempts to find out the degree in which possibility a concept in question truly exists in the shot. Therefore the confidence value usually starts to fail in some threshold value and generates non-relevant results for a concept-query. Our system uses 15 semantic concepts to describe a shot conceptually. These features are ordered based on their confidence values and the ordered list is utilized in computation of conceptual similarities between shots. The ordered list of concepts is used differently in manual and interactive searching.

Manual concept query is based on user-defined concept list that describes for each of the 15 features whether it is required (1) or not required (0) property of the result. In order to compute the overall similarity, each of the list item *l* is used to compute dissimilarity for that feature in a database shot *n.* Dissimilarity is computed based on the difference between the detected confidence and binarized confidence value defined by the user. After computing dissimilarities for every concept, overall dissimilarity between the concept list and a shot is computed as the sum of individual dissimilarity values. All shots are ranked by the smallest dissimilarity value and ordered list of shots is outputted as the conceptual query result.

Example based concept query is used when browsing by conceptual features. Here the concept list of required/not required concepts is constructed from the example shot confidences. Each confidence value is measured against the upper and lower limits of the scale resulting a difference between maximum and minimum confidence (confidence distance to the values 1 and 0 respectively). The confidences are ranked based on the shortest distances, resulting a list of the most discriminative concepts that describe shot in conceptual terms. For example, a shot containing an airplane flying on the sky could have top three descriptive concepts as *airplane*, *outdoors* and *no face*. This list is further used similar manner to a manual concept query, where a list of binary concepts is used as a query parameter. Our system configuration was set to use top three concepts for the example based concept queries.

Following 15 semantic concepts were incorporated in our conceptual search: outdoors, news subject face, people, building, road, vegetation, animal, female speech, car/truck/bus, aircraft, news subject monologue, non studio setting, sporting event, weather news and physical violence. The confidence values were obtained as a result of our participation to the semantic feature detection task. See Chapter 2 for details. Semantic detectors were used from the run MT1, except the features 11,13,16,8 that were from the runs MT6 and MT7. Selected concepts were outcome of an educated guess about the performance of individual feature detectors.

Lexical Similarity with Term-Frequency Inverse Document Frequency

Text features are derived from the automatic speech recognition (ASR) data that was made available for all participants by LIMSI-CNRS [8]. The ASR transcripts were indexed into a database treating each white-space delimited token as a word. A stop word list was used to exclude grammatical and otherwise undiscriminating words that would have led to poor resolution. Remaining words were then stemmed using the Porter stemming algorithm [7]. Available speaker segmentation helped to create better contextual organization for the index and whip up the semantic capabilities. This was achieved by expanding neighbouring shots to every index word until the boundaries of speaker segments were reached. Due to this, every shot became topically connected to its neighbours, presuming that the speaker would not change his/her semantic context during speech.

To evaluate queries, we used relevance metrics to measure which shots were suitable given a set of topic related query words. The ranking of the shots was computed using Term-Frequency Inverse Document Frequency (TFIDF) [5] based classification method that pinpoints relevant words occurring in the ASR transcripts. First, the given query words were automatically stemmed to remove any suffixes. Second, the TFIDF relevance metric was computed for every shot in the database. Finally, the shots were sorted according the computed TFIDF metric.

Since lexical similarity was also a search feature for browsing, we constructed an example-based text search that used example shot ASR-text as a source for query. A list of lexically similar shots was then created with TFIDF as was described earlier.

Late Feature Fusion in Manual and Interactive Queries

In order to construct a final ranked list of most similar shots, system has to formulate sub-queries with previously described similarities. User can select whether he/she wants to retrieve shots based on all or some of the semantic levels. For example, user can define a following complex search (manually):

- Lexical: 'sphinx pyramid'
- Conceptual: concepts 'outdoors', 'buildings' and 'road' are required
- Visual: example image (about Sphinx) search using 'color' and 'structure'

The process of creating visual similarity search follows the same process that is used to generate the semantic concept results (in Chapter 2.1), except the clipping of top ranked results. The fusion operator Θ has been set to sum for low-level visual feature search and the features available in manual search are TCC, TGC and MA. For browsing with visual similarity, we have constrained the system to use only one search configuration: TGC and TCC search combined with sum operator.

To provide final result lists for previously described example, VIRE system commences sub-queries for each of the defined semantic level. VIRE combines the sub-query result lists that contain rankings based on brute-force computations of geometric L_1 dissimilarities for the entire database. The sub-results can be considered as votes of individual feature 'experts' *e*. To make a fusion of these feature lists, we use a variant of Borda count voting method like in Chapter 2.1. Following formulae describe the procedure

$$S_n^t = sum(\frac{R_n^t(1)}{R_{\max}^t},...,\frac{R_n^t(E)}{R_{\max}^t}) \qquad \mathbf{(4)}$$

$$C_f = \left[sort\{S_1,...,S_N\}\right]_{1000} \qquad \mathbf{(5)}$$

where S_n^t = minimum rank of a result shot *n* to the search topic *t* by the query 'experts' 1…*E* (*E*∈ *{1,2,3}*)

$R_n^t(e)$ = rank to the query topic *t* given by an 'expert' *e* (e.g. a visual similarity query service).

$R_{\max}^t$ = maximum rank for topic *t* . Equals to the size of TRECVID 2003 test set

C_t = Final ranked set of results for the search topic *t*

$[\ ]_X$ = *X* top-ranked items in a sorted list

Manual Query Interface

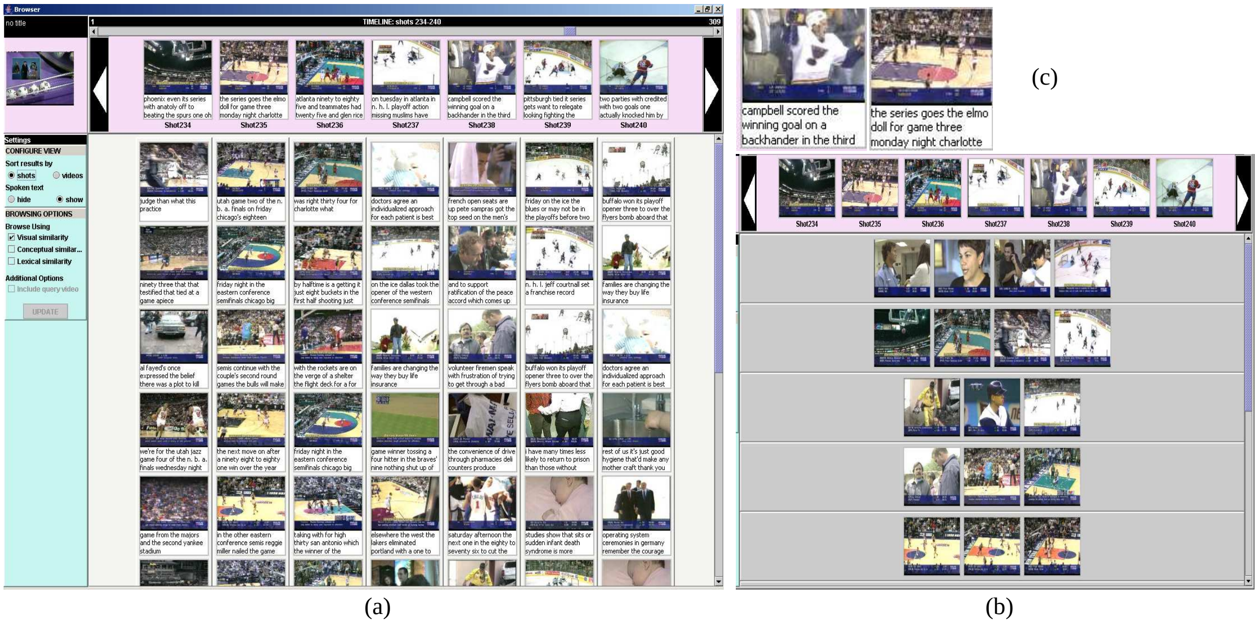
Manual query interface provides search features in three semantic levels. Changing a topic from the menu launches a timer and loads new example shots or images provided by the topic description. User can select any combination of the three visual features (TGC, TCC or MA) for any example video shot, but for example image only TCC and TGC are supported. User can set any configuration of '1s' and '0s' for 15 semantic features. Additionally, a lexical word search can be constructed by selecting a set of words from the topic description. After results for the search have arrived, user can select any interesting shot as a start point for navigation in the browsing interface.

Interactive Browsing Interface

The novelty of our approach in the interactive search task relies on cluster-temporal browsing [19] and ways of viewing the result shots. The motivation is to reduce the effect caused by ambiguous results 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 *cluster-temporal browsing* that combines timeline presentation of videos with content-based retrieval. Cluster-temporal browsing implies that the video content is not utilized alone, but in conjunction with temporal video structure.

Figures 1 (a) and (b) 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 a fast overview of the video content. 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 view that provides linkage to other database videos. The similarity is measured based on the features selected in the lower left panel. Selectable features are named 'visual', 'conceptual' and 'lexical'. User can select a single feature or any combination depending on what properties he/she wants to browse with.



phoenix even its series with anatoly off to beating the spurs one oh

the series goes the elmo doll for game three monday night charlotte

atlanta ninety to eighty five and teammates had twenty five and glen rice

on tuesday in atlanta in n. h. l. playoff action missing muslims have

campbell scored the winning goal on a backhander in the third

pittsburgh tied it series gets want to relegate looking fighting the

two parties with credited with two goals one actually knocked him by

Figure 1 (a) Cluster-temporal browsing interface. The similarity view organizes result shots column-wise (b) Another organization of the similarity view: similar shots grouped by videos. Images depict a news search for ice hockey and basketball sports content. (c) ASR text is visible under the result items

When user locates interesting shots in the similarity view, he can open the related video shot in the topmost row so that the interesting shot is located in the middle column. After the video shots are changed in the top row, system re-computes the similarity view. At any time, user can update the current view by changing to other feature combinations. User can also select two ways to organize similar shots in a similarity view. The traditional view puts the results of a single search into columns under the top-row shots. Another view computes the amount of result shots originating from a single video to create video ranks. Ranks are then used to organize the shots into rows by their ranked parent video. Another novelty feature in the browser is the added visualization of the textual shot content obtained from the ASR transcripts. The requirements to update the similarity view are heavy, since the browsing speed should be close to real-time. To update a view, system must perform parallel query processing for several example-based queries. Multi-threaded index queries with efficient query cache provide reasonable access times even for the most complex feature parameters.

3.2 Experimental Setup

MediaTeam submitted 10 result runs, six for manual and four for interactive search tasks. NIST provided 25 search topics that were used in the experiments. Topic nr. 120 was left out from the interactive experiments leaving 24 search topics. A topic contained one or more example clips of video or images and textual topic description to aid the search process. From the visual similarity example search, only TCC and TGC features were used. Textual search was based on given words in the topic description.

NIST provided segmentation for all videos, from which more than 32000 video shots belonged to the feature search test collection. IBM organized a collaborative annotation for TRECVID 2003 participants, which resulted in annotation of feature search development collection, creating over 60 hours of annotated training material. TRECVID 2003 data consisted mainly of ABC/CNN news from 1998 and about 13 hours of C-SPAN programming.

In manual queries a test user processed all 25 topics. He used VIRE system’s manual search interface to generate six different result sets with following configurations (run ID):

- ASR-based lexical search (M5),
- single example based visual search (M6),
- all examples based visual search (M7),
- all examples based visual + ASR search (M8),
- All examples based visual + conceptual search (M9).

Run M6 was based on a list of single examples that was provided by Mr. Thijs Westerveld from the University of Twente.

The interactive search task was carried out by a group of eight new users. Test users, two of them females, were mainly information engineering undergraduate students, having good skills in using computers, but little experience in searching video databases. They all are used to web-searching. 8 users, 24 topics and two variants of VIRE system were divided into following configurations

- I1V: Variant A: S1[T1-T6],S7[T7-T12],S2[T13-T18],S8[T19-T24]
- I2VT: Variant B: S2[T1-T6],S8[T7-T12],S1[T13-T18],S7[T19-T24]
- I3V : Variant A: S3[T1-T6],S5[T7-T12],S6[T13-T18],S4[T19-T24]
- I4VT: Variant B: S4[T1-T6],S6[T7-T12],S5[T13-T18],S3[T19-T24]

System variant A disabled the support for lexical searching and ASR text visualizations so that searching was based entirely on visual content. System variant B enabled lexical searching and textual visualizations. By setting the experiment like above, the effect of learning was reduced between the system variants and the effect of fatigue was minimized with break and refreshments. The effect of learning within the topic sets was not controlled, most of the users processed the topics in numerical order. All users were given half an hour introduction to the system, with emphasis on the search and browsing interface functions demonstrated with a couple of example searches. Users were told to use approximately twelve minutes for each search, during which they navigated in the shot database and selected shots that seemed to fit to the topic description. Total time for the experiment was about three hours. 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. During the change of system configurations (halfway of the experiment), users were given refreshments and a break.

3.3 Search Results from TRECVID 2003

Average precisions for the 10 different search configurations are shown in Table 8. The descriptions of the runs are found from Chapter 3.2. MAP shows the mean value of the average precisions in 25 (manual) and 24 (interactive) topics. As can be seen, the performance of interactive search is much better than the manual search results. The average time for interactive search was 10.68 with system variant A (visual) and 11.63 with variant B (visual+textual). For the system variant A, the average number of hits at depth 30 was 8.19 whereas for the variant B respective value was 10.88. The same values for hits at depth 10 were 6.33 and 4.81 respectively. This means that the people spent about 9% more time browsing with visual and textual cues resulting in 31-33% increase in found matches within the first 10-30 results.

Manual runs elapsed average time was 1.5 minutes. With best manual configuration (M5) the average hits per topic at depths 10 and 30 were 2 and 4.44

Search Run ID	MAP	Elapsed Time Average (min)
OUMT_I1V (interactive)	0.172	11.2
OUMT_I2VT (interactive)	0.241	12.1
OUMT_I3V (interactive)	0.156	10.2
OUMT_I4VT (interactive)	0.207	11.2
Median (interactive)	0.184	9.5
Max (interactive)	0.476	15
OUMT_M5 (manual)	0.098	1.5
OUMT_M6 (manual)	0.005	1.5
OUMT_M7 (manual)	0.023	1.5
OUMT_M8 (manual)	0.024	1.5
OUMT_M9 (manual)	0.004	1.5
OUMT_M10 (manual)	0.006	1.5 (estimated)
Median (manual)	0.072	3.12
Max (manual)	0.218	15

Table 8. Results for search runs over 25 and 24 search topics

Some of the most successful topics for the interactive test users were 116, 118, 114, 110, 106 and 102. According to the answers in questionnaire, the VIRE system was easy to learn, and somewhat easy to use. This was an improvement from the commentary we received last year about VIRE [15] and was partially caused by the added functionality that utilised more semantic ASR textual transcripts. From the different browser configurations people preferred browsing with visual and lexical features enabled and the best system variant was type B (ASR transcripts visible) with the result shots organised in columns, not by video rows.

4 Conclusions

Large experiments were conducted with TRECVID 2003. Our experiments showed that our remarkably simple method based on training with small example sets using color and structural features generated slightly above median results. Our multimodal fusion was not behaving as we expected, improvements could be achieved for example with multi-modal classifiers.

The most successful component in this work was cluster-temporal browser that provided many levels of semantics for the user to pursue more meaningful results. Our system provided the user multiple parallel paths from which he/she could choose the direction for the navigation. We have combined the temporal adjacency of shots and content-based similarity searches into one interface. This gives the user better understanding about the various inter-relations of video shots during browsing.

The manual search methods based on text were most successful; there was no improvement in results when any combinations of visual, conceptual and lexical features were used together. One reason for the failure of combined features may be in the implementation issues we have found after the experiments were finished, however it is clear that the fusion techniques should be further improved. Nevertheless, the second best manual configuration was obtained with all visual examples together with a lexical search. The use of one search example was performing worse than using all of them.

Acknowledgments

The financial support of the National Technology Agency of Finland and the Academy of Finland is gratefully acknowledged.

References

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

[2] 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.

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

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

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

[6] Heisele B, Poggio T & Pontil M (2000) Face Detection in Still Gray Images, Tech. Report 1687, Center for Biological and Computational Learning, MIT.

[7] Porter M (1980) An Algorithm for Suffix Stripping Program. Program, 14(3):130—137.

[8] Gauvain JL, Lamel L & Adda G (2002) The LIMSI Broadcast News Transcription System. Speech Communication, 37(1-2):89-108.

[9] Scheirer, ED (1998) Tempo and Beat Analysis of Acoustic Musical Signals J. Acoust. Soc. Am. 103:1, pp 588-601.

[10] Klapuri, A. (1999) Sound Onset Detection by Applying Psychoacoustic Knowledge. IEEE International Conference on Acoustics, Speech and Signal Processing.

[11] Penttilä J, Peltola J & Seppänen T (2001) A speech/music discriminator-based audio browser with a degree of certainty measure. Proc. Infotech Oulu International Workshop on Information Retrieval, Oulu, Finland, 125-131.

[12] MPEG7 standard: ISO/IEC FDIS 15938-4: Information technology -- Multimedia content description interface -- Part 4: Audio

[13] Dubnowski JJ, Schafer RW & Rabiner LR (1976) Real-Time Digital Hardware Pitch Detector, IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. 24, No. 1, s. 2-8.

[14] Deller JR, Hansen JHL & Proakis JG (2000) Discrete-Time Processing of Speech Signals, IEEE Press, 908 s.

[15] Rautiainen M, Penttilä J, Vorobiev D, Noponen K, Väyrynen P, Hosio M, Matinmikko E, Mäkelä SM, Peltola J, Ojala T & Seppänen T (2002) TREC 2002 Video Track experiments at MediaTeam Oulu and VTT. Text Retrieval Conference TREC 2002 Video Track, Gaithersburg, MD.

[16] Rautiainen M, Seppänen T, Penttilä J & Peltola J (2003) Detecting semantic concepts from video using temporal gradients and audio classification. International Conference on Image and Video Retrieval, Urbana, IL.

[17] Ho T, Hull J & Srihari S (1994) Decision combination in multiple classifier systems. IEEE Transactions on Pattern Analysis and Machine Intelligence 16(1): 66–75.

[18] IBM VideoAnnEx Server Page MPEG-7 (31.10.2003) http://mp7.watson.ibm.com/VideoAnnEx/

[19] Rautiainen M, Ojala T & Seppänen T (2003) Cluster-temporal video browsing with semantic filtering. Advanced Concepts for Intelligent Vision Systems, Ghent, Belgium.

[20] Prosody-based classification of emotions in spoken Finnish Seppänen T, Väyrynen E &Toivanen J (2003) 8th European Conference on Speech Communication and Technology, Geneva, Switzerland, p.26.