Available online at www.sciencedirect.com
SCIENCE DIRECT®
Computer Vision and Image Understanding 96 (2004) 216–236
Computer Vision and Image Understanding
www.elsevier.com/locate/cviu

A multi-modal system for the retrieval of semantic video events

Arnon Amir[a,*], Sankar Basu[b,1], Giridharan Iyengar[c], Ching-Yung Lin[b], Milind Naphade[b], John R. Smith[b], Savitha Srinivasan[a], Belle Tseng[b]

[a] *IBM Almaden Research Center, 650 Harry Road, San Jose, CA 95120, USA*
[b] *IBM T.J. Watson Research Center, 19 Skyline Drive, Hawthorne, NY 10532, USA*
[c] *IBM T.J. Watson Research Center, 1101 Kitchawan Road, Yorktown Heights, NY 10598, USA*

Received 5 April 2002; accepted 2 February 2004
Available online 5 August 2004

Abstract

A framework for event detection is proposed where events, objects, and other semantic concepts are detected from video using trained classifiers. These classifiers are used to automatically annotate video with semantic labels, which in turn are used to search for new, untrained types of events and semantic concepts. The novelty of the approach lies in the: (1) semi-automatic construction of models of events from feature descriptors and (2) integration of content-based and concept-based querying in the search process. Speech retrieval is independently applied and combined results are produced. Results of applying these to the Search benchmark of the NIST TREC Video track 2001 are reported, and the gained experience and future work are discussed.

Keywords: Multimedia indexing; Event detection; Semantic video annotation; Content-based video retrieval

* Corresponding author.
E-mail address: arnon@almaden.ibm.com (A. Amir).
[1] Current address: The National Science Foundation, 4201 Wilson Blvd., Arlington, VA 22230, USA.

1077-3142/$ - see front matter © 2004 Published by Elsevier Inc.
doi:10.1016/j.cviu.2004.02.006

1. Introduction

Content-based retrieval of video events is an emerging field of research. People envision video retrieval systems that allow retrieval of relevant clips, scenes, and events based on multi-modal queries which could include textual description, image, audio and/or video samples. Examples of such queries are "John Deere tractor," "Clips of people who are water skiing," "rocket launch," "Shots of a lunar rover," "Find Ronald Reagan speaking," and "Clips that deal with floods."[2] These are semantic-level queries which should match the visual and/or the audio content of the video. Such systems involve multi-modal video and audio indexing, automatic or semi-automatic learning of semantic concepts and their multi-modal representation, advanced query interpretation and matching algorithms, imposing many new challenges to research.

Content-based retrieval of video events is a multi-disciplinary field, built on computer vision, content-based image retrieval (CBIR, see, e.g. [4,33]), spoken document retrieval (SDR, see, e.g. [10]), video analysis, and machine learning. Previous work in event detection includes statistical methods trained for domain-specific detection and classification of events, such as Petkovic and Jonker's work [26] on classifying tennis strokes to six categories, Gong et al. [14] on event detection in soccer and human motion classification such as walk and hand gestures [3,5,13]. Anjum and Aggarwal [1] present a system for automatic segmentation and tracking of continuous body motion. Their methodology uses tracking of several different body parts and computation of the temporal angular relationships between those parts.

Other work use object recognition for event indexing, such as Qian et al. [28], who use multiple modalities for indexing and retrieval. These methods require domain-specific knowledge and make use of special algorithms tailored for the event detection task in that domain, and often depend on camera view angle. A more generic approach to gait detection and human motion tracking and analysis is proposed in [22]. Spatio-temporal features are classified using HMM that is trained on multiple examples of labeled motions. By using machine learning to classify events, this method is more general and might apply to other types of events. However, generic approach to event detection in heterogeneous collections of videos is still at its very early stages.

A generic approach to event detection requires semantic annotation of not only events but also objects, locations, people, and other semantic concepts. For example, the event of "people spending leisure time on the beach" requires the detection of people, beach (sand, water, and sky), and leisure activity. We denote all of those semantic bits of information as *concepts*. By generic we mean that those concepts are not taken from any specific domain, and the methods applied to detect them should generalize to multiple concepts, rather than doing specific detailed analysis such as in the case of classifying human actions. In this context, events are one type of concept. Like with the above example, the relationships and dependencies between

[2] All the examples in this section are topics from the 2001 NIST TREC Video track benchmark.

concepts provide a way to detect their presence or absence. Often the presence of one concept may suppress the probability of other concepts. For example, the presence of ''rocket launch'' in a shot would greatly reduce the probability of finding ''people spending leisure time on the beach'' in the same shot. The use of context for improving multi-modal concept detection in video was demonstrated in [21]. Detection of generic events in heterogeneous video collections is complementary to work on detection of specific events in domain-specific video.

We refer to events at the shot level. It is merely a choice of representation rather than of a statement about the actual duration of a specific event. A shot is said to contain an event if the event occurs at least once within the shot period of time. A shot is a continuous, and often homogeneous, sequence of frames. Hence for many visual events, a shot would be a very natural choice of the event's temporal support. The same applies to many instances of other types of concepts, such as scene location, presence of objects, of faces, and more.

The Text REtrieval Conference (TREC [32]), cosponsored by NIST and DARPA, has started in 1991. Since then, TREC has become the main international forum for benchmarking of various information retrieval (IR) tasks, such as document retrieval, web search, and spoken document retrieval (SDR). Those IR fields have made substantial progress through the years of being promoted by TREC (see, e.g. [12]). The Video track, started in 2001, supports research in video indexing and retrieval by providing video data, defining tasks, running a thorough evaluation of the results, and putting the stage to exchange information, ideas and opinions between the participating groups. It focuses mainly on visual content-based search and leaves only a secondary, optional role for use of speech retrieval in its tasks. Since year 2002, the video track has adopted MPEG-7 (ISO/IEC [16,30]) as the format to exchange metadata between participating groups. While MPEG-7 provides a unified framework for metadata representation, it does not specify the methods for extracting metadata descriptions from video, nor how to match and search in the stored metadata.

In this work we touch both of these research challenges. We describe a system that analyzes video by segmenting it into shots, extracting key-frames, and extracting audio-visual descriptors from each of the shots. These descriptors include not only low-level features, but also semantic events and concepts. The retrieval integrates those features, concepts, and semantics to find thought for need of information, including non-indexed events and other concepts.

The system was tested through participation in the TREC 2001 video track. This video retrieval benchmark consists of 74 topics, searched in a corpus containing approximately 11 hours of video. The topics are designed to retrieve video based on semantic visual contents. We report these results and post-TREC improved results, and discuss some of the lessons learned which are implemented towards the TREC 2002 video track.

The rest of the paper is organized as follows. The video analysis and indexing is described in Section 2, including the concept classifier and its training process. Section 2.5.5 provides a quick overview of the speech retrieval process. Section 3 describes the experimentation with the TREC benchmark and their results. These results are then discussed in Section 4, with some suggestions for future work.

2. Video analysis

The video content is analyzed through shot detection, feature extraction, and semantic content classification, as shown in Fig. 1. The video is first segmented into shots and descriptors are extracted for each shot. The descriptors are stored and used as input into the concepts classification system which assigns semantic labels to each shot. The system also ingests any meta-data related to the content such as title (if available), format, and source.

2.1. Shot boundary detection (SBD)

The video content is pre-processed by splitting it into shots using the *IBM CueVideo Toolkit* [9]. As the shot boundaries are detected, key-frames are extracted and saved as JPEG files. In addition, all MPEG I-frames are extracted, stored, indexed, and used for accessing the shots.

CueVideo uses a 512 bins (3 bits/channel) three-dimensional color histograms in RGB color space to represent frames and to compare pairs of frames, up to seven frames apart (see Fig. 2). Histograms of 60 frames around the current frame are stored in a moving window buffer. Statistics of frame differences in the moving window are used to compute adaptive thresholds. Hence no manual tuning is required.

A state machine is used to detect and classify the different shot boundaries. The 13 states are listed at the bottom of Fig. 2, along with the graph of the actual machine states. At each frame a state transition is made from the current state to the next state, and any required operation is taken (e.g., report a shot, save a key-frame to file). The algorithm reports the boundary location and duration in frames and in SMPTE, and classifies the boundary type as Cut, Fade-in, Fade-out, Dissolve or Other. It operates in a single pass on the MPEG file and processes fully decoded frames, thus supporting also other video formats such as Quicktime. It is robust to possibly incompatible MPEG streams, and processes video at about twice faster than its real-time rate on a 800 MHz P-III.

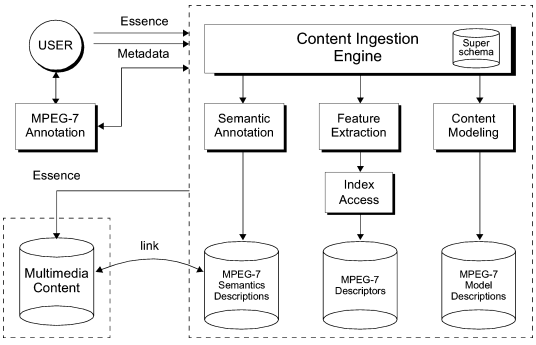

Fig. 1. The video ingestion process. Models of concepts are built during training phase using annotated training video. The models are applied to new input video during the ingestion process.

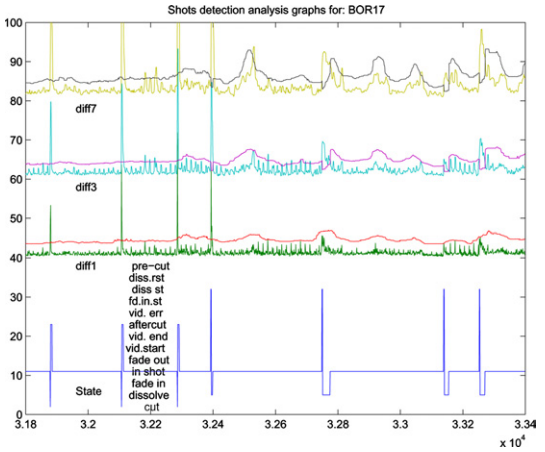

Fig. 2. This example represents a 53 second video sequence with four cuts, followed by three dissolves, in relatively high encoding noise (from TREC 2001, *bor17.mpg*, frames 31,800–33,400).

Table 1
Shot boundary detection results on the TREC 2001 shot boundary detection test set

	Ins. rate	Del. rate	Precision	Recall
Cuts	0.039	0.020	0.961	0.980
Gradual	0.589	0.284	0.626	0.715
All	0.223	0.106	0.831	0.893

The results of the shot boundary detection on the TREC 2001 SBD test corpus are shown in Table 1. The data set consists of 42 videos, total of 5.8 h, about 3000 shots of which third are gradual and two-thirds are cuts.

While the system performed very well on detecting cuts (best amongst all the participating systems), the detection of gradual changes could be greatly improved.[3] An analysis of the errors in detecting gradual changes shows that in many cases which were reported as errors, there was a detection of a boundary but the reported duration of the gradual effect was too short compared to the actual duration.

2.2. Feature extraction

The feature detectors take as input a set of feature vectors extracted from video shots and frames. The system extracts the following set of features for each of the key-frames and I-frames.

[3] For improvements made in later years see [2].

- *Color*. A normalized, linearized 3-channel *HSV* histogram is used, with 6 bins for hue, 6 bin for saturation, and 12 bins for intensity.[4]
- *Texture.* A two-dimensional dependence matrix, which captures the spatial dependence of gray-level values contributing to the perception of texture is called a gray-level co-occurrence matrix (GLCM). A GLCM is a statistical measure extensively used in texture analysis. In general we denote

$$p(i,j,d,\theta) = \frac{P(i,j,d,\theta)}{N(d,\theta)}, \tag{1}$$

  where $P(\cdot)$ is the GLCM for the displacement vector d and orientation θ and $N(\cdot)$ is the normalizing factor to make the left-hand side of Eq. (1) a probability distribution. We compute four GLCMs of the quantized (32 gray-levels) V channel, where $d = 1$ and $\theta \in \{0°, 45°, 90°, 135°\}$. For each of the four matrices, six statistical features of the GLCM are computed. The features are contrast, energy, entropy, homogeneity, correlation, and inverse difference moment [17].
- *Structure*. To capture the structure within each image/region, a Sobel operator with a 3×3 window is applied to each region and the edge map is obtained. Using this edge map, a 24-bin histogram of edge directions is obtained as in [18].

These features are used as the input for semantic concept detectors, as well as for traditional content-based retrieval (CBIR).

2.3. Semi-automatic concept annotation of training data

A basic concept detector is a classifier which takes feature vectors from a shot as input and decides if the concept applies to the shot. It may further provide a concept score to the shot. The classifier is also denoted as the *model* of the concept.

Each model is trained using a training video set, annotated with labels of the corresponding concept. To allow a small number of concepts to have maximal impact, we have primarily investigated concepts that apply broadly to video content, such as indoor vs. outdoor, nature vs. man-made, face detection, sky, land, water, and greenery. We have also investigated several specific concepts including airplanes, rocket launches, fire, and boats. While concept classifiers allow video content to be annotated automatically using this small vocabulary, the integration of the different search methods together (content-, concept-based) allows more effective retrieval of other types of events as well.

Annotating multiple concepts over a sizable training corpus requires efficient video annotation tools. A video annotation tool was developed to support the annotation of training data (Fig. 3). It allows users to annotate each shot in a video sequence. The user assigns Events, Scenes, and Objects labels to the shot. The user

[4] A linearized histogram of multiple channels is obtained by concatenating the histogram of each of the channels.

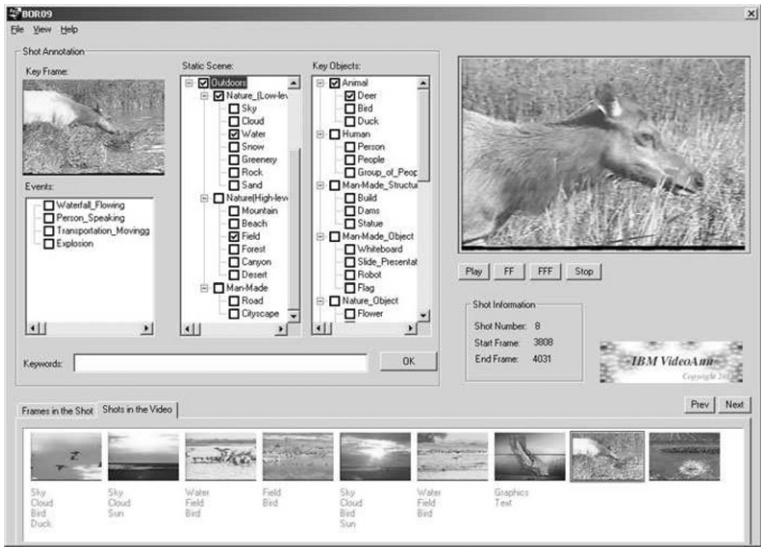

Fig. 3. The video annotation tool allows users to label the events, scenes, and objects in the video shots. Labels are selected from a lexicon, hierarchically organized as a small taxonomy in three categories: Events, Static Scenes, and Key Objects.

may further specify regions in the shot's key-frame and associate the Objects labels with specific regions.

A small lexicon was created to describe different types of events, scenes, and objects; the following excerpt gives some of the annotation terms:

- *Events*: water skiing, boat sailing, person speaking, landing, take-off/launch, and explosion.
- *Scenes*: outer space (moon, mars), indoors (classroom, meeting room, laboratory, factory), outdoors (nature, sky, clouds, water, snow, greenery, rocks, land, mountain, beach, field, forest, canyon, desert, waterfall), and man-made (road, cityscape).
- *Objects*: non-rigid objects (animal, deer, bird, duck, human), rigid objects (man-made structure, building, dam, statue, tree, flower), and transportation (rocket, space shuttle, vehicle, car, truck, rover, tractor).

The user assigns labels to one shot at a time. The tool is semi-automatic in that it automatically learns and propagates designated labels to ''similar'' unlabeled shots as described in [24]. The user may confirm or reject the propagated labels. The final annotation is saved in an MPEG-7 file.

2.4. Concept modeling

Several different types of statistical learning models were examined, including Bayesian Networks, Multinets [23], and Gaussian Mixture Models (GMM). Statis-

tical models were developed using a collection of labeled training video. We modeled the following concepts (a selected subset of the annotated elements):

- *Events*: fire, smoke, rocket launch.
- *Scenes*: greenery, land, outdoors, rock, sand, sky, water.
- *Objects*: airplane, boat, rocket, vehicle.

2.4.1. Statistical concept modeling

In the statistical modeling approach, the features extracted from the video content are modeled by a multi-dimensional random variable X. The features are assumed to be independent identically distributed random variables drawn from known probability distributions with unknown deterministic parameters. For the purpose of classification, it is assumed that the unknown parameters are distinct under different hypotheses and can be estimated. In particular, each semantic concept is represented by a binary random variable. The two hypotheses associated with each such variable are denoted by H_i, $i \in \{0,1\}$, where H_0 denotes absence and H_1 denotes presence of the concept. Under each hypothesis, we assume that the descriptor values are generated by the conditional probability density function $P_i(X)$, $i \in \{0,1\}$.

In case of Scene concepts, we use static descriptors that represent the features of each key-frame. In case of Events, which have specific temporal characteristics, we can construct temporal descriptors using time series of static descriptors over multiple video I-frames. In our previous work on multi-frame-based temporal event modeling we had used Hidden Markov Models (HMM) with Gaussian Mixtures to represent observations within each state of the HMM [23].

Due to the small number of positive examples for events such as Rocket Launch, and the difficulty in object segmentation and motion extraction, we have made a simplifying assumption that events occur within shots and that key-frames represent the dominant observation states of these events. By modeling the key-frames with Gaussian Mixtures we learn the dominant static representations of these events.

We use a *one-zero* loss function [27] to penalize incorrect detection. This is shown in the following equation:

$$\lambda(\alpha_i|\omega_j) = \begin{cases} 0 & \text{if } i = j, \\ 1 & \text{otherwise.} \end{cases} \tag{2}$$

The risk corresponding to this loss function is equal to the average probability of error and the conditional risk with action α_i is $1 - P(\omega_i|x)$. To minimize the average probability of error, class ω_i must be chosen, which corresponds to the maximum a posteriori probability $P(\omega_i|x)$. This corresponds to the minimum probability of error (MPE) rule.

In the special case of binary classification, the MPE rule can be expressed as deciding in favor of ω_1 if

$$\frac{p(x|\omega_1)}{p(x|\omega_2)} > \frac{(\lambda_{12} - \lambda_{22})P(\omega_2)}{(\lambda_{21} - \lambda_{11})P(\omega_1)}. \tag{3}$$

The term $p(x|\omega_j)$ is the *likelihood* of ω_j and the test based on the ratio in Eq. (3) is called the *likelihood ratio test* (LRT) [7,27]. More details about the statistical modeling can be found in [19].

2.4.2. Model's parameter estimation

For modeling the TREC video content, we assume that the conditional distributions over the descriptors X under the two hypotheses—concept present (H_1) and concept absent (H_0)—have been generated by distinct mixtures of diagonal Gaussians. The modeling of these semantic concepts involves the estimation of the unknown (but deterministic) parameters of these GMMs using the set of annotated positive examples in the training set. For this purpose the features associated with training data corresponding to each label are modeled by a mixture of five Gaussians. The parameters (mean, covariance, and mixture weights) are estimated by using the Expectation Maximization (EM) algorithm [8].

The rest of the training data is used to build a negative model for each label in a similar way. The LRT is used in each test case to determine which of the two hypotheses is more likely to account for the descriptor values. The likelihood ratio can also be looked upon as a measure of the *confidence* of classifying a test image to the labeled class under consideration. A ranked list of confidence measures for each of the labels is produced by repeating this procedure for all the labels under consideration.

The performance of statistical models such as the GMM depends to a large extent on the amount of training data. Due to the relatively small amount of labeled training video data in the TREC video corpus, we adopted a ''leave one video out'' strategy. This means that we trained a model for each concept as many number of times as the number of videos. During each such training, one video clip was left out from the training set. The trained models for the two hypotheses were used to detect the semantic concept in the clip that was left out.

2.4.3. Region-based concepts

The manual annotation process is involved with marking of bounding boxes encompassing relevant regions for each Object label. When the concept model is applied to new data, it is also applied at a regional level. Those regions are generated by an automatic rough segmentation process, which assign bounding boxes to a few salient regions. To fuse the decisions from those several regions, we use the following hypothesis: if a concept is to be declared absent in a frame, it must be absent in each and every region tested. Hence the maximal score of the concept in any of the regions is used as the score of the concept occurrence in the frame. For concepts which are global in terms of feature support, this region-based step is not needed. Localized or regional concepts include *rocket*, *face*, *sky*, and so forth.

2.4.4. Concept fusion

The objective of concept fusion is to combine multiple statistical models for the different video concepts. Separate GMM models are trained for each of the different feature vectors (color, texture, and structure). As a result, we get several classifications and associated confidence levels for a single concept in a test image. While

the different classifiers may be combined in many ways, we explored straightforward methods such as taking sum, maximum, and product of the individual confidences in computing an overall classification confidence for the concept.

While this strategy of "late fusion" is fairly simple, one can envision other "early fusion" methods such as concatenating different feature vectors into a single vector and then building a single GMM. We did not pursue this strategy due to the high dimensionality of the descriptors, especially in view of the paucity of training video content depicting the concepts of interest. However, it may be possible to consider discrimination in reduced dimensionality subspaces of the combined feature space by using techniques such as principal component analysis (PCA).

2.5. Using search and retrieval to detect new events and concepts

Training a concept detector for every possible type of event or concept would take an infinite amount of resources—manual annotation of a training corpus and system training to build the models. On the other hand, one could assume that once we build a large enough set of basic concept detectors, it would be possible to define many new events and concepts by mapping those to existing ones. This bottom-up approach is discussed in [15]. If annotation is available for the new concept, a classifier might be trained in concepts space. In the absent of a large training set for the new concept, its mapping to existing concepts may be performed via a search and retrieval process.

Before searching for any new events and concepts, the feature extraction process and the trained concept detectors are applied to the (test) video corpus. Feature vectors and detected concepts are stored in MPEG-7 files. Next, we turn to search for the new topics—events and concepts for which the system was not trained. The search is initiated by the few examples provided as part of the topic definition. In a sense, the retrieval process is a way of learning to detect a new concept from a few examples rather than using a large set of pre-labeled training data. The model in this case is the final form of the query, derived from the topic definition and fed into the retrieval system.

Different retrieval procedures may be applied. Those include low-level content-based retrieval (CBR), semantic model-based retrieval (MBR), speech retrieval (SDR), and combined methods. In addition, the query formulation might be done automatically from the given topic, or with the aid of a human, in a manual or an iterative process. Several of those approaches are discussed in this section, and their performance is then compared in Section 3.

For TREC video retrieval, each of 74 provided topics includes textual description and sample content, which includes anywhere from a single image to several images, audio, and video clips. Although it is commonly agreed that users are unlikely to arrive with media samples in hand, this query model is appropriate for focused browsing and query refinement "find me more like this." This is a common practice in content-based image retrieval systems [11]. Here, however, a combination of CBR, MBR, and speech is used for the matching task.

2.5.1. Search task I: automatic content-based retrieval

Content-based retrieval is the most amenable in the case that the query provides example content. Automatic CBR is the case where the given topic definition is passed as-is to the retrieval system. For the case of CBR-only (no use of text/speech), the following approach was adopted: the query image and video content were analyzed using shot detection, key-frame selection, and feature extraction to produce a set of features of the query content, similar to those of the indexed videos. Then, the query features were matched against the target features.

We considered two approaches for automatic content-based matching: (1) matching of single key-frame descriptors between the query and shots in the corpus and (2) matching of multiple I-frames descriptors between the query and shots in the corpus. The latter is based on multiple pair-wise comparisons of frames and aggregation of the results into a single decision using standard "all" or "any" semantics, corresponding to logical AND and OR operations, respectively. The multi-to-multi frame matching is illustrated in Fig. 4.

2.5.2. Search task II: interactive content-based retrieval

In interactive retrieval, a user is allowed to conduct multiple rounds of searching operations in which each successive round refines or builds on the results of previous rounds. The video query pipeline process is shown in Fig. 5. Queries are processed in a multi-stage search. In the interactive mode, at each iteration, the user selects related concepts and marks clusters or examples of related shots. At each stage of the search, a query Q_i produces a result list R_i. The result list R_i is then used as input into a subsequent query Q_{i+1}, and through various selectable operations for combining and scoring R_i with the matches for Q_{i+1}, the result list R_{i+1} is produced Fig. 6.

Each round consists of the following: (1) a similarity search in which target shots are scored against query content (using single frame or multi-frame search) and (2) a combining of these search results with the previous results list. This way, each successive round combines potentially new results with the current list. We have used a choice of the following aggregation functions for combining the scores:

$$D_i(n) = D_{i-1}(n) + D_q(n) \tag{4}$$

and

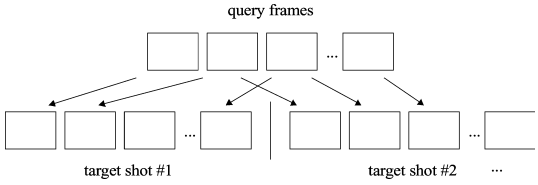

Fig. 4. Content-based retrieval matches multiple query frames against multiple frames in the target shots.

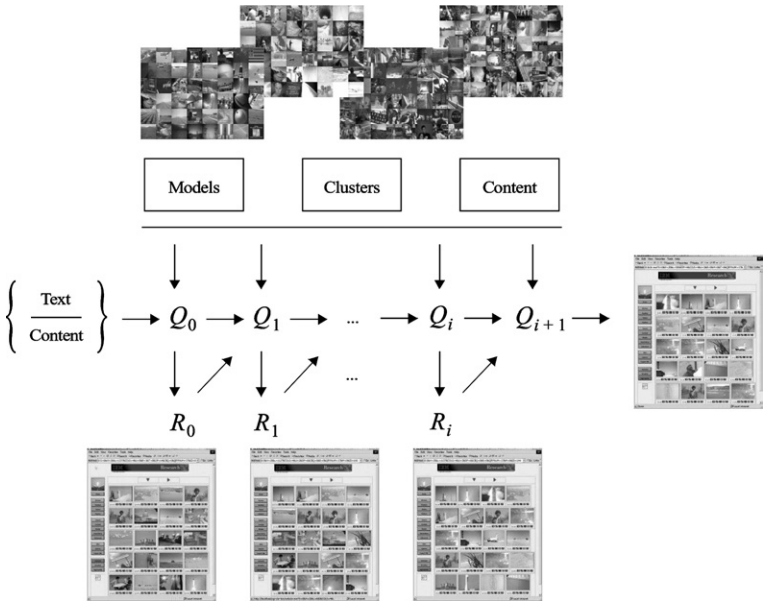

Fig. 5. The video content retrieval engine integrates methods for searching in an iterative process in which the user successively applies content-based and model-based searches.

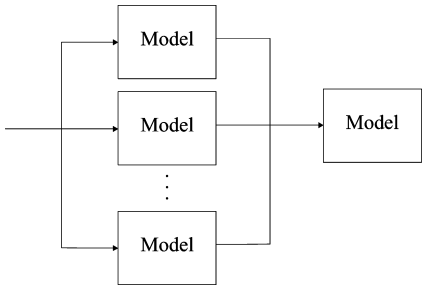

Fig. 6. Ranking of database entries based on the detection of semantic concepts can be performed with various configurations of the concept models. Here we show how multiple concept models can be applied in parallel and/or sequence to affect the final ranking.

$$D_i(n) = \min(D_{i-1}(n), D_q(n)), \tag{5}$$

where $D_q(n)$ denotes the score of video shot n with the present query and $D_{i-1}(n)$ and $D_i(n)$ denote the combined score of video shot n for the previous query and the current round, respectively. Eq. (4) has the effect of weighting the most recent query equally with the previous queries, while Eq. (5) has the effect of ranking most highly the target shots that best match any one of the query images. Other combining functions that use different join predicates are possible [20]. The user can continue this iterative search process until the desired video content is retrieved.

2.5.3. Model-based retrieval

The model-based (or concept-based) retrieval (MBR) allows a user to retrieve shots based on the semantic labels, automatically produced by the concept models. Each semantic label has an associated confidence score. In the automatic MBR, the concepts detected in the query are matched with concepts extracted from corpus shots. In the interactive MBR, the user can retrieve results for a selected concept by issuing a query for the particular semantic label. The target video shots are then ranked by confidence score (higher score gives lower rank).

A new concept is made by specifying a combination of existing concepts. The user may specify a weighted sum of several concepts as a single query. For example, the user may express "nature" as nature = 0.5 * outdoors + 0.25 * water + 0.25 * sky. The user my further apply additional concepts, one at a time, to build more complicated expressions of concepts, while using the interim results to decide which concept to apply next and which operator to use.

2.5.4. Interactive content + model-based retrieval

The interactive integrated search is carried out by the user successively applying the content-based and the model-based search methods. For example, a user looking for video shots showing a beach scene can issue the following sequence of queries in the case that beach scenes have not been explicitly labeled:

1. Search for concept "outdoors."
2. Aggregate with concept "sky."
3. Aggregate with CBR of a query image (or possibly an image selected in the interim result set) resembling the desired scene.
4. Aggregate with concept "water."
5. Aggregate with selected relevant image, video shot.

Different operators can be used for rescoring the results at each stage of the query and combining with the previous results. The user may choose any of the following options:

1. *Inclusive*: each successive search operation issues new query against target database:

$$D_0(n) = D_q(n), \tag{6}$$

2. *Iterative*: each successive search operation issues query against current results list and scores by new query:

$$D_i(n) = D_q(n), \tag{7}$$

3. *Aggregative*: each successive search operation issues query against current results list and aggregates scores from current results and new query results:

$$D_i(n) = f(D_{i-1}(n), D_q(n)), \tag{8}$$

where $f(\cdot)$ corresponds to max or avg. The distance scores $D_i(n)$ are based on feature similarity (for CBR) and concept confidence (for models). For the models, $D_q(n) = 1 - C_q(n)$, where $C_q(n)$ gives the confidence of the query label for video shot n, and $D_{i-1}(n)$, and $D_i(n)$ are defined as above. The lossy filtering is accounted for in that some target shots n^* might have confidence score $C_q(n^*) = -\infty$. Eq. (8) combines the label score of each target video shot for the current query plus the cumulative label score of the previous queries, whereas Eq. (7) takes only the latest score.

2.5.5. Using speech retrieval to search for new concepts

In addition to automatic analysis and modeling of the visual video content, we also investigate the use of speech indexing for video retrieval [6,31]. Since this paper focuses on the visual modelling, we provide here only a brief overview of the speech indexing part of the system, and count the main differences of this task from traditional SDR.

A speech index is prepared using speech recognition, using the traditional spoken document retrieval (SDR) approach [10,12]. We used the speech indexing and retrieval system of *CueVideo* [9]. The indexing is performed in several steps. First, a speech recognition system (*IBM ViaVoice* real-time engine, with the broadcast news language model of 60,000 words) is used to transcribe the audio and generate a continuous stream of timed words. The words and times from the speech recognizer output are segmented into document-units of 100 words, with partial overlap between consecutive units. This step is followed by tokenization to and part-of-speech tagging such as noun phrase, plural noun, etc. Morphological analysis uses the part-of-speech tag and a morph dictionary to reduce each word to its morph. For example, the verbs "lands," "landing," and "land" are all reduced to "land." Then, stop words are removed using a standard stop words list. For each of the remaining words, the number of document-units that it belongs to (the inverse document frequency) is computed and is used to weight the word using the common term frequency times inverse document frequency (tf*idf) criteria [29].

The search for visual events and concepts using speech retrieval might be more complicated than it first appears. Given a topic, or a statement of information need, a text to speech query has to be constructed. In many cases, however, the words used to write the topic statement are not the words one would expect to hear at or near relevant shots. For example, events related to the topic "people spending leisure time at the beach" are more likely to have words like "vacation," "family," and "sun" rather than the word leisure. "Spending" is unlikely to help for this topic neither.

Generating a good query in an automatic way was beyond the scope of the work presented here. We used text query terms which were manually composed by a person who did not see the video corpus and was given only the textual part of each topic definition. No refinement was done to those queries. Once the query text was composed, it was sent to the speech retrieval server.

When a query is processed, each query word is tokenized, tagged, and morphed. Video segments are then retrieved in a ranked list. The relevancy score is calculated by the traditional OKAPI formula [29]. This score gives higher weight to rare query

word over a more frequent one, and a higher score to document units which contain more instances of the query word.

The two main differences between traditional SDR and video retrieval are the small size of document units (down to single shots at TRECVID) compared to complete documents in SDR, and the higher word error rate (WER) that results from lower audio quality and from videos which contain speech and music mixed together. Another difference arises from the misalignment between document units and shots—the former may typically span over 5–10 shots. To allow query integration with CBR and MDR results, which are shot-based, all the shots within the retrieved video segment receive the same speech-based score.

3. Concept detection evaluation using the TRECVID 2001 benchmark

The TRECVID 2001 benchmark involves two search tasks, denoted known-item search and general search [25]. They total 74 topics (queries), covering events, objects, locations specific people, etc. Each topic is composed of a textual definition and example media content. The search test set consists of approximately 11 h of video. The evaluation was carriedout using standard quantitative metrics for evaluating retrieval effectiveness, based on ground truth assessment of the retrieved shots.

Topics were in general composed such that they cannot be retrieved by speech alone. For example, the instances of topics like "scenes with person doing water ski" and "scenes with views of canyons" were not the subject of the video clips containing them. They appeared at isolated, short shots during the discussion of other matters (e.g., hydroelectric power plants, showing a dam and lake landscapes). As they were not mentioned in the speech, they could not benefit from speech retrieval at all. A few topics like "scenes showing astronaut driving the lunar rover" and "rocket launch" had matches in NASA videos in which the lunar rover or the rocket launch are mentioned several times and made speech retrieval extremely useful.

The topics were typically defined at a highly semantic-level, e.g., "retrieve video clips of Ronald Reagan speaking," as opposed to "retrieve video clips of a red flower," which might be addressed to some extent with CBR without dealing with a semantic visual representation at all. Adding the fact that speech was of no help for many of the topics, retrieving at a semantic level became a much more challenging task.

The evaluation metric was based on precision and recall. These are commonly used for the evaluation of information retrieval tasks (see, e.g. [32]). Given a ranked list of matching shots, precision is the percentage of correct matches within the retrieved set and recall is the percentage of correct matches retrieved from all the (ground truth) instances occur in the corpus,

$$\text{precision} = \frac{\text{correct retrieved matches}}{\text{total retrieved matches}}, \tag{9}$$

$$\text{recall} = \frac{\text{correct retrieved matches}}{\text{total ground truth occurrences}}. \tag{10}$$

Precision is usually plotted at different recall points (by considering the top K ranked matches for different Ks). This however requires more data and more ground truth than was available this year. Thus only one precision–recall point was computed for the entire set of results of each query.

The two search tasks, known-item (KI) and general search (GS), had slightly different evaluation processes. Topics in the KI task are more specific (such as a specific place as oppose to outdoor) and are expected to have a smaller number of instances in the data. All the correct matches for KI topics were listed in a complete ground truth set, allowing computation of both precision and recall of top 100 matches. In the general search, the topics are of a more general notion and the number of instances in the corpus is higher and unknown, e.g., "video shots showing nature scenes." The top 20 matches were submitted for each GS topic. Each match was evaluated by two NIST assessors who decided if this is a correct or an incorrect match, and precision at 100 was computed for each topic. In both search tasks an average precision was computed over all the corresponding topics.

3.1. Retrieval results for general search

The results of the general search are shown in Table 2. The average number of hits over all 36 GS queries is displayed. Up to 20 matches were submitted for each topic. The interactive visual content-based and model-based retrieval, denoted below as visual-based retrieval (VBR) method is compared to the automatic speech recognition (SDR) results and the combined search.

The results show a significant increase in retrieval quality using the interactive query formulation over automatic one. In both automatic and interactive runs the best results were achieved for the combined VBR + SDR retrieval.

Specific examples comparing retrieval performance for interactive VBR and SDR approaches are given in Table 3. In some cases, such as topics VT66 and VT47, the SDR approach gave better retrieval results. In these topics, the relevant information was not easily captured by the visual concepts. For other topics, such as VT55, VT49, VT43, and VT42, the interactive VBR approach gave better performance than the SDR approach.

Table 2
Video retrieval results (average hits/query in 20 top matches over 36 GS topics)

Approach	Hits/query	Average precision (%)
Automatic VBR	4.1	21
Automatic SDR	1.9	19
Automatic VBR + SDR	5.0	25
Interactive VBR	4.3	23
Interactive SDR[a]	1.9	19
Interactive VBR + SDR	6.1	31

[a] The same SDR results of the automatic run where used.

Table 3
Video retrieval results (hits/query) comparing interactive VBR and SDR methods for specific queries

Topic #	Description	SDR	VBR
VT66	Clips about water project	9	3
VT47	Clips that deal with floods	8	1
VT55	Pictures of Hoover Dam	3	8
VT49	Lecture showing graphic	4	20
VT43	Shots showing grasslands	0	8
VT42	Shots of specific person	1	9

We also compared the interactive VBR approach to non-interactive (or automatic) VBR in which only a single iteration of searching was allowed. The results for two of the topics given in Table 4 show a significant increase in retrieval performance using the interactive VBR approach.

3.2. Retrieval results for known item search

The retrieval results for 38 known item topics is summarized in Table 5. Up to 100 best matches were submitted for each topic. As was said earlier, these topics specified much higher semantic level than of the general search, like identification of specific persons. Such specific concepts impose a greater challenge on the CBR approaches.

Overall, the KI topics had much fewer items than the general search has. Each topic has, on average, only 5.5 correct matches in the corpus (the "rocket launch" topic has 60 items—the KI topic with the largest number of items). Since the retrieval unit is defined to be a shot, a single event that spans over several shots would correspond to several known items.

Table 4
Video retrieval results (hits/query) comparing automatic and interactive CBR methods for specific queries

Topic #	Description	Automatic CBR	Interactive CBR
VT54	Glen Canyon Dam	3	12
VT15	Shots of corn fields	1	5

Table 5
Video retrieval results (avg. precision/recall over 38 known-item searches)

Approach	Average precision (%)	Average recall (%)
Automatic VBR	0.9	21.0
Automatic SDR	2.9	0.9
Automatic VBR + SDR	0.6	24.6
Interactive VBR	3.2	34.8
Interactive SDR[a]	2.9	0.9
Interactive VBR + SDR	3.2	34.8

See text for details.
[a] The same SDR results of the automatic run were used.

The SDR results were in particular very low, even in compare to those of the general search. We found that the main reason for that was a glitch in the submitted results. While the SDR retrieves video segments, 5-15 shots long, only the first shot of each segment was submitted. A naive fix would be to submit all the shots which correspond to each segment. We repeated the evaluation for this case, using the NIST evaluation tools and ground truth. The results are shown in Table 6.

These results are a bit closer to those of the general search, yet the overall results are not as good as for GS.

Fig. 7 shows the precision–recall for topic "rocket launch," shown here due to its large number of correct items which allows us to compute such a graph. This topic was one of the few where speech was much better than CBR in both precision and recall, despite of our efforts in modelling launch and rocket concepts. Apparently, this topic is always discussed in the speech when it shows up in the video.

One may argue that these results are still low. At their current level, these results are still not useful for practical use. We consider this as the first step in a long journey towards a very ambitious goal. The task of searching for high-level visual con-

Table 6
Post-TREC Video retrieval results (average precision/recall over 38 known-item searches)

Approach	Average precision (%)	Average recall (%)
Automatic CBR	2.5	20.0
Automatic SDR	6.0	10.0
Automatic CBR + SDR	4.0	22.0
Random retrieval (reference)	0.18	3.3

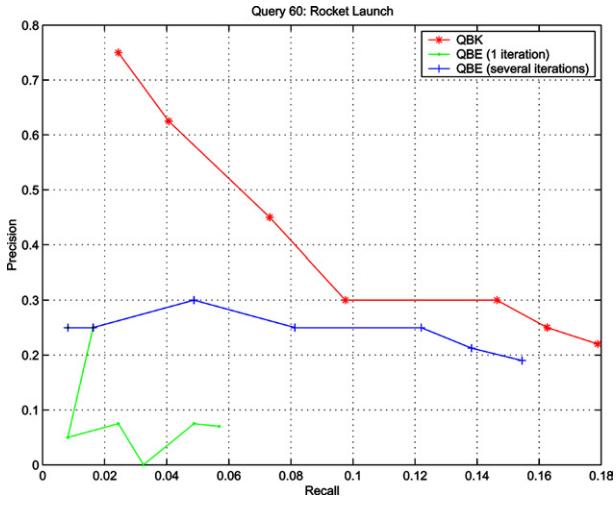

Fig. 7. Precision–recall graph for topic 60 "rocket launch." This shows a CBR precision level similar to the precision level obtained in the general search task.

cepts using automatic indexing tools of audio-visual content is still proved to be very challenging and a lot of research need to be done.

4. Summary

In this paper, we describe a system for automatic and interactive content-based and model-based detection and retrieval of events and other concepts. Models of semantic concepts are built by training classifiers using labeled training video. The system then extracts feature descriptors from shots, classifies the shots using those semantic models for different events, scenes, and objects. The retrieval system allows the integration of content-based and model-based retrieval, along with speech retrieval, to detect new types of events and concepts for which the system was not trained. The search and query refinement may be done either in an automatic or iterative search modes.

The results clearly demonstrate that low-level visual features can be used to classify shots by semantic-level concepts. However, this requires a lot of labeled data which are not handy and are very time consuming to generate. Let alone the problem of negotiating copyright issues to get access to useful video data. We hope that with NIST help and with a collaborative effort of the participating groups there will be more data and in particular labeled data available in the future for those purposes.

The results also demonstrate that combined CBR + SDR retrieval is superior to each of the methods alone. The main advantage of speech-based video retrieval, whenever speech is available, is the direct access to semantic information conveyed by spoken words and sentences. However, queries which directly correspond to the visual content, such as "pink flower" has nearly no chance to be found by speech retrieval. In such cases, speech retrieval only adds irrelevant matches. More elaborate methods for combining SDR and CBR results need to be developed. Automatic analysis of the query and decision of whether to use speech retrieval on the query or not is an open problem.

Interactive search has a clear advantage over automatic search. This is expected due to the difficulty in automatic query formulation and modality weights. The interactive search could leverage from relevance feedback, or active learning, however we did not use those techniques in this work.

The problem of mapping between low-level features and concepts is of major importance. Many special detectors need to be developed, either by using machine learning and large amount of annotated data for training, or by developing tailored algorithms for detecting faces, people, text etc. The combination of multiple such filters will allow to improve the retrieval of semantic queries. Since no single group can effort to develop all different kinds of visual detectors, we would like to have detectors and detection results being shared among the TREC participants and being used by all groups for the search tasks. Last year the group from CMU has made the first step in this direction by making their speech transcripts available to the lowlands group, following the tradition that was carried on at the SDR track. NIST will have a major role in promoting and supporting this collaboration. A joint effort in collect-

ing and labeling large amounts of training video data using standard description scheme such as MPEG-7 is also required to enable the research in this field.[5]

Acknowledgments

We are very grateful to Paul Over and Ramazan Taban, NIST, for organizing the video track. Many thanks to the NIST assessors who devoted much of their time in watching each and every shot while assessing all the submitted results.

References

[1] A. Anjum, J.K. Aggarwal, Segmentation and recognition of continuous human activity, in: Proc. IEEE Workshop on Detection and Recognition of Events in Video (EVENT-01), Vancouver, Canada, 2001.
[2] A. Amir, M. Berg, G. Iyengar, C-Y. Lin, C. Dorai, M. Naphade, A. Natsev, C. Neti, H. Nock, I. Sachdev, J.R. Smith, Y. Wu, B. Tseng, D. Zhang, Ibm research trecvid-2003 video retrieval system, in: Proc. NIST TRECVID 2003, Gaithersburg, MD, 2003.
[3] J.K. Aggarwal, Q. Cai, Human motion analysis: a review, Comput. Vision Image Understanding 73 (3) (1999) 428–440.
[4] A. Del Bimbo, Visual Information Retrieval, Morgan Kaufmann, San Francisco, CA, 1999.
[5] M. Brand, N. Oliver, A. Pentland, Coupled hidden Markov models for complex action recognition, Proc. Comput. Vision Pattern Recogn. (1997) 994–999.
[6] E. Brown, S. Srinivasan, A. Coden, D. Ponceleon, J. Cooper, A. Amir, Towards speech as a knowledge resource, IBM Syst. J. 40 (4) (2001) 985–1001.
[7] R.O. Duda, P.E. Hart, Pattern Classification and Scene Analysis, Wiley Eastern, New York, 1973.
[8] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, Proc. R. Stat. Soc. B (39) (1977) 1–38.
[9] S. Srinivasan, A. Amir, G. Ashur, D. Petkovic, D. Ponceleon, The IBM CueVideo Toolkit Version 2.1. Available from: <http://www.almaden.ibm.com/projects/cuevideo.shtml/>.
[10] J. Makhoul, et al., Speech and language technologies for audio indexing and retrieval, Proc. IEEE 88 (8) (2000) 1338–1353.
[11] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, P. Yanker, Query by image and video content: the QBIC system, IEEE Comput. 28 (9) (1995) 23–32.
[12] J. Garofolo, G. Auzanne, E. Voorhees, The trec spoken document retrieval track: a success story, November 16–19, 1999.
[13] D.M. Gavrila, The visual analysis of human movement: a survey, Comput. Vision Image Understanding 73 (1) (1999) 82–98.
[14] Y. Gong, L.T. Sin, C.H. Chuan, H.-J. Zhang, M. Sakauchi, Automatic parsing of tv soccer programs, in: Proc. IEEE Internat. Conf. on Multimedia Computing and Systems, Washington, DC, 1995, pp. 165–174.
[15] Niels Haering, Niels da Vitoria Lobo, Visual Event Detection, the Kluwer Int. Series in Video Computing, Kluwer Academic Publishers, New York, NY, 1992.

[5] Concepts donation and collaborative annotations were implemented at TRECVID 2002 and 2003, after the time of writing this paper.

[16] ISO/IEC JTC 1/SC 29/WG 11/N3966. Text of 15938-5 FCD Information Technology—Multimedia Content Description Interface—Part 5 Multimedia Description Schemes, Final Committee Draft (FCD) edition, March 2001.
[17] R. Jain, R. Kasturi, B. Schunck, Machine Vision, MIT Press and McGraw-Hill, New York, 1995.
[18] Anil K. Jain, Aditya Vailaya, Shape-based retrieval: a case study with trademark image databases, Pattern Recogn. 31 (9) (1998) 1369–1390.
[19] M. Naphade, S. Basu, C. Lin, J. Smith, B. Tseng, A statistical modeling approach to content-based video retrieval, in: Proc. 16th Internat. Conference on Pattern Recognition, ICPR 2002, Quebec City, Canada, vol. 2, 2002, pp. 953–956.
[20] A. Natsev, Y.-C. Chang, J.R. Smith, C.-S. Li, J.S. Vitter, Supporting incremental join queries on ranked inputs, in: Proc. Conference on Very Large Databases (VLDB), Rome, Italy, September 2001.
[21] Milind R. Naphade, Thomas S. Huang, Detecting semantic concepts using context and audiovisual features, in: Proc. IEEE Workshop on Detection and Recognition of Events in Video (EVENT-01), Vancouver, Canada, 2001.
[22] R. Sharma, N. Krahnstover, M. Yeasin, Towards a unified framework for tracking and analysis of human motion, in: Proc. IEEE Workshop on Detection and Recognition of Events in Video (EVENT-01), Vancouver, Canada, 2001.
[23] M. Naphade, I. Kozintsev, T.S. Huang, K. Ramchandran, A factor graph framework for semantic indexing and retrieval in video, in: Proc. IEEE Workshop on Content-based Access to Image and Video Libraries (CBAIVL), Hilton Head, SC, June 12, 2000.
[24] M.R. Naphade, C.Y. Lin, J.R. Smith, B.L. Tseng, S. Basu, Learning to annotate video databases, in: IS&T/SPIE Symposium on Electronic Imaging: Science and Technology—Storage & Retrieval for Image and Video Databases 2002, vol. 4676, San Jose, CA, January, 2002.
[25] P. Over, R. Taban, The trec-2001 video track framework, in: 10th Text Retrieval Conference (TREC-10), Gaithersburg, Maryland, USA, November 2001.
[26] M. Petkovic, W. Jonker, Content-based video retrieval by integrating spatio-temporal and stochastic recognition of events, in: Proc. of IEEE Workshop on Detection and Recognition of Events in Video (Event 2001), Vancouver, BC, Canada, July 8, 2001.
[27] H.V. Poor, An Introduction to Signal Detection and Estimation, second ed., Springer-Verlag, New York, 1999.
[28] R. Qian, N. Hearing, I. Sezan, A computational approach to semantic event detection, in: Proceedings of Computer Vision and Pattern Recognition, Fort Collins, CO, vol. 1, June 1999, pp. 200–206.
[29] S.E. Robertson, A. Walker, K. Sparck-Jones, M.M. Hancock-Beaulieu, M. Gatford, OKAPI at TREC-3, in: Proc. Third Text Retrieval Conference, 1995.
[30] J.R. Smith, Mpeg-7 standard for multimedia databases, in: ACM Intl. Conference on Data Management (ACM SIGMOD), Santa Barbara, California, May 2001.
[31] S. Srinivasan, D. Petkovic, Phonetic confusion matrix based spoken document retrieval, in: Proc. of Twenty-Third Annual Int. ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR-2000), Athens, Greece, July 2000.
[32] E. Voorhees, D. Harman, Overview of the ninth text retrieval conference, in: Proceedings of the Ninth Text Retrieval Conference (TREC-9), Gaithersburg, Maryland, USA, November 16–19, 1999.
[33] A. Yoshitaka, T. Ichikawa, A survey on content-based retrieval for multimedia databases, IEEE Trans. Knowledge Data Eng. 11 (1) (1999) 81–93.