COST292 experimental framework for TRECVID 2008

Q. Zhang,[1] G. Tolias,[2] B. Mansencal,[3] A. Saracoglu,[4]
N. Aginako,[5] A. Alatan,[4] L. A. Alexandre,[6] Y. Avrithis,[2]
J. Benois-Pineau,[3] K. Chandramouli,[1] M. Corvaglia,[7] U. Damnjanovic,[1]
A. Dimou,[10] E. Esen,[4] N. Fatemi,[8] J. Goya,[5] F. Guerrini,[7] A. Hanjalic,[11]
R. Jarina,[9] P. Kapsalas,[2] P. King,[10] I. Kompatsiaris,[10] L. Makris,[10]
V. Mezaris,[10] P. Migliorati,[7] A. Moumtzidou,[10] Ph. Mylonas,[2] U. Naci,[11]
S. Nikolopoulos,[10] M. Paralic,[9] T. Piatrik,[1] F. Poulin,[8]
A. M. G. Pinheiro,[6] L. Raileanu,[8] E. Spyrou,[2] S. Vrochidis[10]

Abstract

In this paper, we give an overview of the four tasks submitted to TRECVID 2008 by COST292. The high-level feature extraction framework comprises four systems. The first system transforms a set of low-level descriptors into the semantic space using Latent Semantic Analysis and utilises neural networks for feature detection. The second system uses a multi-modal classifier based on SVMs and several descriptors. The third system uses three image classifiers based on ant colony optimisation, particle swarm optimisation and a multi-objective learning algorithm. The fourth system uses a Gaussian model for singing detection and a person detection algorithm. The search task is based on an interactive retrieval application combining retrieval functionalities in various modalities with a user interface supporting automatic and interactive search over all queries submitted. The rushes task submission is based on a spectral clustering approach for removing similar scenes based on eigenvalues of frame similarity matrix and and a redundancy removal strategy which depends on semantic features extraction such as camera motion and faces. Finally, the submission to the copy detection task is conducted by two different systems. The first system consists of a video module and an audio module. The second system is based on mid-level features that are related to the temporal structure of videos.

[1]Q. Zhang, K. Chandramouli, U. Damnjanovic, T. Piatrik and E. Izquierdo are with Dept. of Electronic Engineering, Queen Mary, University of London (QMUL), Mile End Road, London E1 4NS, UK, {qianni.zhang,uros.damnjanovic,tomas.piatrik,krishna.chandramouli,ebroul.izquierdo}@elec.qmul.ac.uk

[2]G. Tolias, E. Spyrou, P. Kapsalas, Ph. Mylonas and Y. Avrithis are with Image, Video and Multimedia Laboratory, National Technical University of Athens (NTUA), 9 Iroon Polytechniou Str., 157 80, Athens, Greece, {gtolias,espyrou,pkaps,fmylonas,iavr}@image.ntua.gr

[3]B. Mansencal and J. Benois-Pineau are with LaBRI, University Bordeaux (LaBRI), 351, cours de la Liberation 33405, Talence, France, jenny.benois,boris.mansencal@labri.fr

[4]A. Saracoglu, E. Esen and A. Alatan are with Middle East Technical University (METU), 06531, Ankara, Turkey, ahmet.saracoglu@uzay.tubitak.gov.tr, ersin.esen@uzay.tubitak.gov.tr, alatan@eee.metu.edu.tr

[5]N. Aginako, J. Goya are with VICOMTech, Mikeletegi Pasealekua, 57 Parque Tecnolgico 20009 Donostia / San Sebastin, Spain,{naginako,jgoya}@vicomtech.es

[6]A. M. G. Pinheiro and L. A. Alexandre are with Universidade da Beira Interior (UBI), Covilha, Portugal, pinheiro@ubi.pt, lfbaa@di.ubi.pt

[7]M. Corvaglia, F. Guerrini and P. Migliorati are with University of Brescia (U. Brescia), Via Branze 38 25123 Brescia, ITALY, {marzia.corvaglia,fabrizio.guerrini@ing.unibs.it,pierangelo.migliorati}@ing.unibs.it

[8]N. Fatemi, F. Poulin and L. Raileanu are with IICT, University of Applied Sciences of Western Switzerland (UAS), St-Roch, Avenue des Sports 20, CH-1401, Yverdon-les-Bains {nastaran.fatemi,florian.poulin,laura.raileanu}@heig-vd.ch

[9]R. Jarina and M. Paralic are with Department of Telecommunications, University of Zilina (U. Zilina), Univerzitna 1, 010 26 Zilina, Slovakia, jarina@fel.uniza.sk

[10]S. Vrochidis, A. Moumtzidou, P. King, S. Nikolopoulos, A. Dimou, V. Mezaris, L. Makris and I. Kompatsiaris are with Informatics and Telematics Institute/Centre for Research and Technology Hellas (ITI-CERTH), 6th Km. Charilaou-Thermi Road, P.O. Box 361, 57001 Thermi-Thessaloniki, Greece, {stefanos, moumtzid, king, nikolopo, dimou, bmezaris, lmak, ikom}@iti.gr

[11]U. Naci, A. Hanjalic are with Delft University of Technology (TU. Delft), Mekelweg 4, 2628CD, Delft, The Netherlands, {s.u.naci,A.Hanjalic}@tudelft.nl

1 Introduction

This paper describes collaborative work of several European institutions in the area of video retrieval under a research network COST292. COST is an intergovernmental network which is scientifically completely self-sufficient with nine scientific COST Domain Committees formed by some of the most outstanding scientists of the European scientific community. Our specific action COST292 on semantic multi-modal analysis of digital media falls under the domain of Information and Communication Technologies. Being one of the major evaluation activities in the area, TRECVID has always been a target initiative for all COST292 participants [1]. Therefore, this year our group has submitted results to four tasks: high-level feature extraction, search, rushes and copy detection. Based on our submissions to TRECVID 2006 and 2007, we have tried to improve and enrich our algorithms and systems according the previous experience [2] [3]. The following sections bring details of applied algorithms and their evaluations.

2 High-level feature extraction

COST292 participated to the high-level feature extraction task with four separate systems. The results of all these systems were integrated to a single run. The first system, developed by the National Technical University of Athens (NTUA), transforms a set of low-level descriptors into the semantic space using Latent Semantic Analysis and utilises neural networks for feature detection and is described in Section 2.1. The second system, developed by the University of Beira Interior (UBI), uses a multi-modal classifier based on SVMs and several descriptors and is described in Section 2.2. The third system, developed by Queen Mary University of London (QMUL), uses three image classifiers based on ant colony optimisation, on particle swarm optimisation and on a multi-objective learning algorithm and is described in 2.3. Finally, the fourth system is developed by the University of Zilina (U. Zilina) in co-operation with LaBRI, uses a Gaussian model for singing detection and a person detection algorithm and is described in Section 2.4.

2.1 Feature extraction from NTUA

NTUA detected the following high-level concepts: *Classroom*, *Airplane flying*, *Two People*, *Cityscape*, *Mountain* ,*Nighttime*.

For the detection of *Classroom*, *Cityscape* and *Mountain*, we use the following approach [4]: First, we extract color and texture MPEG-7 descriptors from the NRKFs and more specifically *Dominant Color*, *Scalable Color*, *Color Layout*, *Homogeneous Texture* and *Edge Histogram*. These low-level descriptions are extracted from image regions that resulted from a coarse color segmentation. Then, a clustering algorithm is applied on a subset of the training set, in order to select a small set of regions that will be used to represent the images. From each cluster we select the closest region to the centroid. This region will be referred to as "region type". Then, for each keyframe we form a model vector description. Let: $d_i^1, d_i^2, ..., d_i^j, i = 1, 2, \ldots, N_R$ and $j = N_C$, where N_C denotes the number of region types, N_R the number of the regions within the image and d_i^j is the distance of the i-th region of the image to the j-th region type. The model vector D_m is formed as in Eq.1.

$$D_m = \left[min\{d_i^1\}, min\{d_i^2\}, ..., min\{d_i^{N_C}\} \right], i = 1, 2, \ldots, N_R \tag{1}$$

Then we apply the Latent Semantic Analysis algorithm. We construct the co-occurrence matrix of region types in given keyframes of the training set. After of the construction of the co-occurrence matrix, we solve the SVD problem and transform all the model vectors to the semantic space. For each semantic concept, a separate neural network (NN) is trained. Its input is the model vector in the semantic space and its output represents the confidence that the concept exists within the keyframe.

The main idea of our *Two People* and *Airplane flying* detection algorithm is based on extracting regions of interest, grouping them according to some similarity and spatial proximity predicates and subsequently defining whether the area obtained, represents the object in question. Thus, the method initially involves detection of salient points and extraction of a number of features representing local the color and texture. At the next step, the points of interest are grouped with the aid of an unsupervised

clustering algorithm (DBSCAN) that considers the density of the feature points to form clusters. In the classification stage, there is a need for a robust feature set allowing the object's form to be discriminated even in a cluttered background. Histogram of Oriented Gradients (HoG) descriptor [5] is used to encode information associated to the human body boundary. The method is based on evaluating well-normalized local histograms of image gradient orientations in a dense grid. The basic idea is that local object appearance and shape can often be characterized rather well by the distribution of local intensity gradients or edge directions, even without precise knowledge of the corresponding gradient or edge positions. In practice, this is implemented by dividing the image window into small spatial regions ("cells"), for each cell accumulating a local 1-D histogram of gradient directions or edge orientations over the pixels of the cell. For better invariance to illumination, shadowing, etc., it is also helpful to perform contrast normalization to the local responses before using them. Finally our object detection chain involves tiling the detection window with a dense grid of HoG descriptors and using the feature vector in a conventional SVM based window classifier.

2.2 Feature extraction from UBI

UBI approach to detect several high-level features is based on the detection of several descriptors that are used for a multimodal classification after a suitable training.

The main decriptor is the histograms of oriented gradients (HOGs) such as the ones proposed in [5]. The number of directions considered depended on the type of object that was to be detected: either 9 or 18 directions. This description information was complemented with color information from the RGB color histograms, texture information from 9-7 bi-orthogonal filters, color correlograms [6] for 1 pixel distance with a color quantization to 16 colors, dominant color descriptor, a combination of shape and texture information using a Scale-Space Edge Pixel Directions Histogram.

The images were subdivided into rectangular sub-images of varied size, depending on the descriptor. These sub-images were processed individually and a classification was obtained from a SVM with RBF kernel or with KNN, depending of the feature. The SVM classification confidence interval for an image was obtained by averaging the classification scores of each of its sub-images. The KNN classification was obtained measuring a weighted distance of each subimage Manhattan distance. The final confidence interval is given by the rate of the class training samples number selected in K.

2.3 Feature extraction from QMUL

The system developed by QMUL uses three image classifiers: a classifier based on ant colony optimisation (ACO), a classifier based on particle swarm optimisation (PSO) and a multi-feature based classifier using a multi-objective learning (MOL) algorithm.

The idea underpinning the ACO model is loosely inspired by the behavior of real ants. The real power of ants resides in their colony brain and pheromone-driven communication within the colony. An important and interesting behavior of ant colonies is, in particular, how ants can find the shortest paths between food sources and their nest. For image classification task, the ACO algorithm is implemented and it is integrated with the semi-supervised COP-K-means approach.

In our proposal, the ACO plays its part in assigning each image to a cluster and each ant is giving its own classification solution. Images are classified based on the probability influenced by heuristic information and pheromone value. The main idea of finding optimal solution resides in marking classification solutions by pheromone as follows:

$$\tau_{(X_i,C_j)}(t) = \rho\ \tau_{(X_i,C_j)}(t-1) + \sum_{a=1}^{m} \Delta\tau^{a}_{(X_i,C_j)}(t) \tag{2}$$

where ρ is the pheromone trail evaporation coefficient ($0 \leq \rho \leq 1$) which causes vanishing of the pheromones over the iterations. $\tau_{(X_i,C_j)}(t-1)$ represents the pheromone value from previous iteration. $\Delta\tau^{a}_{(X_i,C_j)}(t)$ is a new amount of pheromones calculated from all m ants that assign image X_i to the j'th cluster. Definition of $\Delta\tau^{a}_{(X_i,C_j)}(t)$ ensure that the pheromone increases when clusters get more apart and when each cluster has more similar images. The ACO makes the COP-K-means algorithm less dependent on the initial parameters and distribution of the data; hence it makes it more stable.

Furthermore the ACO based multi-modal feature mapping improves inferring semantic information from low-level feature.

PSO technique is one of the meta-heuristic algorithms inspired by Biological systems. The image classification is performed using the self organising feature map (SOFM) and optimising the weight of the neurons by PSO [7]. The algorithm is applied to SOFM for optimising the weights of the neurons. The objective of SOFM is to represent high-dimensional input patterns with prototype vectors that can be visualised in a usually two-dimensional lattice structure [8]. Input patterns are fully connected to all neurons via adaptable weights, and during the training process, neighbouring input patterns are projected into the lattice, corresponding to adjacent neurons. SOFM enjoys the merit of input space density approximation and independence of the order of input patterns. Each neuron represents an image with dimension equal to the feature vector. Two different SOFM networks were used in detecting features. The first network configuration is a dual layer SOFM (DL-SOFM) structure which enables training of only positive models while the negative training models are implicitly generated by the network property. This model provides a high degree of recall, while the second configuration is a single layer rectangular mesh (R-SOFM), enabling explicit training of both positive and negative models. Thus enabling to achieve high precision.

The multi-feature classifier features a multi-feature combination algorithm based on optimisation. To handle the problem of merging multiple features in a more automatic and semantically meaningful way, the this classifier utilises a multi-objective learning (MOL) algorithm based on the multi-objective optimisation (MOO) strategy. This classifier requires only a small amount of initial training from a professional user. The goal is to estimate a suitable multi-feature merging metric for a concept that can improve the visual-to-semantic matching for multimedia classification.

From the training set of a concept, a virtual centroid which lies across all the considered feature spaces is firstly calculated. Then each training sample is used to construct an objective function. Thus the concept is represented by a set of objective functions and MOO strategy is used to find the general optimum of all of these functions. The obtained metric is considered to be the optimal combination metric for the concept and will be used for classification of relevant/irrelevant multimedia samples. For more details of the MOL approach for feature combination and its other applications, please refer to [9].

Six visual features were utilised in this experiment: the colour layout, colour structure, dominant colour, edge histogram and homogeneous texture descriptors from MEPG-7, and grey-level cooccurence matrix feature.

2.4 Feature extraction from U. Zilina

Singing feature extraction procedure consisted of two separate steps: 1) singing detection from audio track, and face detection performed on the keyframes. Following [10], a common MFCC/GMM based classification on one-second audio segments with 50% overlap was performed for audio analysis. The GMM for singing sound was compared with the universal background GMM (UniGMM) that contained all other sounds. Initialy the singing sound data for the model training were manually selected from the "music" class of our own audio database [10]. The classifier was first run on audiotrack of a part of TRECVID 2008 development data, which were manually labeled for singing sounds. The initial recognizer was consequently tuned-up by extending the training data for UniGMM by the audiosegments that were recognized uncorectly as singing, and retraining the UniGMM. Then each video shot in the TRECVID 2008 test collection was assigned as relevant for the singing features if at least 1 second of audio, within the shot, was recognized as singing sound. Finaly the audio analysis was combine with image processing and face detection of the keyframes. The face detection was performed by LaBRI using OpenCV [11], [12]. Only shots with both positive face and singing sound detection were submitted for evaluation.

2.5 Results

All the results of the aforementioned techniques were submitted in a single run. The numerical results are shown in Table 1. Among the 20 features evaluated by NIST, our detectors had relatively better success for concepts *Two People*, *Nighttime* and *Street*.

concept id	1	5	6	7	10	12	13	14	15	16	17	18	19	20
relevant returned	9	15	4	124	44	11	72	11	30	26	114	12	9	37
inferred AP $\cdot 10^{-2}$	0.08	0.21	0.04	3.31	1.02	0.10	2.28	1.28	0.49	0.63	5.44	0.11	0.09	0.63

Table 1: Number of relevant documents returned and inferred Average Precision for the submitted concepts.

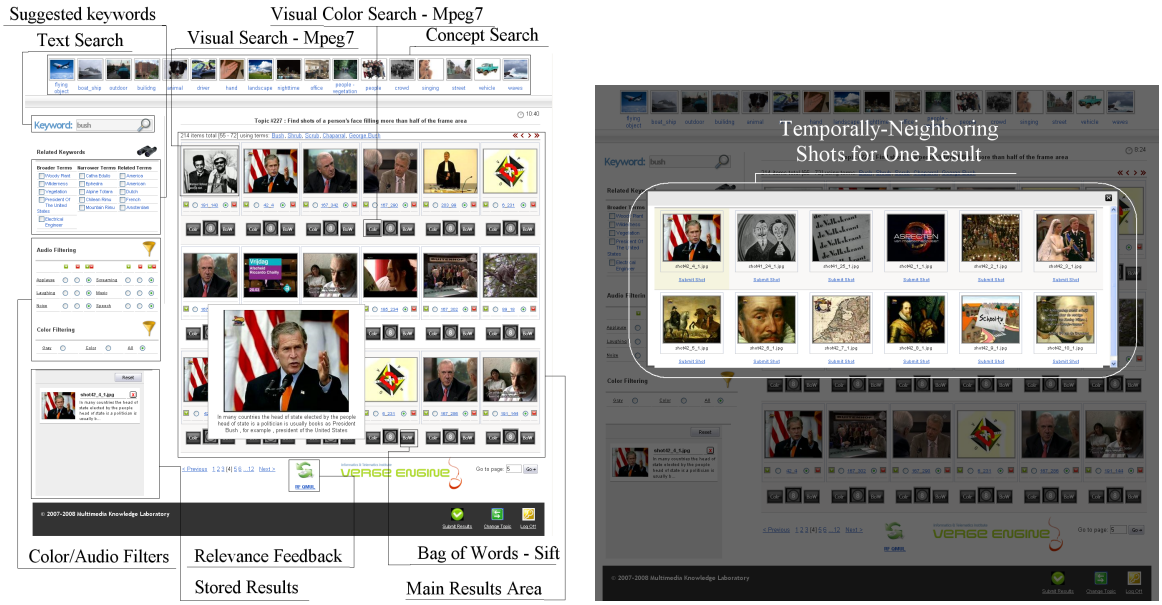

Figure 1: User interface of the interactive search platform

3 Interactive Search

The system submitted to the search task is an interactive retrieval application developed jointly by the Informatics and Telematics Institute (ITI-CERTH), QMUL, U. Zilina and University of Applied Science (UAS). It combines basic retrieval functionalities in various modalities (i.e. visual, audio, textual) and a user friendly graphical interface, as shown in Figure 1, that supports the submission of queries using any combination of the available retrieval tools and the accumulation of relevant retrieval results over all queries submitted by a single user during a specified time interval. The following basic retrieval modules are integrated in the developed search application:

- Visual Similarity Search Module;
- Audio Filtering Module;
- Textual Information Processing Module;
- Term recommendation module;
- Relevance Feedback Module;
- High Level Concept Retrieval Module;

The search system, combining the aforementioned modules, is built on web technologies, and more specifically php, JavaScript and a mySQL database, providing a GUI for performing retrieval tasks over the internet. Using this GUI, the user is allowed to employ any of the supported retrieval functionalities and subsequently filter the derived results using audio and colour constraints. The retrieval results (representative keyframes of the corresponding shots) are presented ordered by rank in descending order, providing also links to the temporally neighbouring shots of each one. The identities of the desirable shots, which are considered as relevant to the query, can be stored by the user (Figure 1). The latter is made possible using a storage structure that mimics the functionality

of the shopping cart found in electronic commerce sites and is always visible through the GUI. In this way, the user is capable of repeating the search using different queries each time (e.g. different combination of the retrieval functionalities, different keywords, different images for visual similarity search, etc.), without loosing relevant shots retrieved during previous queries submitted by the same user during the allowed time interval. A detailed description of each retrieval module employed by the system is presented in the following section.

3.1 Retrieval Module Description

3.2 Visual similarity search

In the developed application, the user has two different search options to retrieve visually similar images. More specifically, content based similarity search is realized using either global information such as the MPEG-7 visual descriptors capturing different aspects of human perception (i.e., colour and texture), or local information as formulated by vector quantizing the local descriptors obtained by applying the Lowes SIFT transform [13].

In the first case, five MPEG-7 descriptors namely Color Layout, Color Structure, Scalable Color, Edge Histogram, Homogeneous Texture are extracted from each image of the collection [14] and stored in a relational database. By concatenating these descriptors a feature vector is formulated to compactly represent each image in the multidimensional space. An empirical evaluation of the systems performance using different combinations of the aforementioned descriptors advocated the choice of two MPEG-7 based schemes. The first one relies on color and texture (i.e., ColorLayout and EdgeHistogram are concatenated), while the second scheme relies solely on color (i.e., only ColorLayout is used).

In the second case, where local information is exploited, the method adopted is an implementation of the bag-of- visual words approach as described in [15]. Specifically, a large amount of local descriptors (training data) is used for learning the visual words which are extracted by applying clustering using a fixed number of clusters. Subsequently, each image is described by a single feature vector that corresponds to a histogram generated by assigning each key-point of the image to the cluster with the closest center. The number of selected clusters (100 in our case) determines the number of bins and as a consequence the dimensionality of the resulting feature vectors. For both cases retrieval is performed efficiently using a multi-dimensional indexing structure.

An r-tree structure is constructed off-line by using the feature vectors of all images and the corresponding image identifiers. R-tree(s) [16] are structures suitable for indexing multidimensional objects and known to facilitate fast and efficient retrieval on large scale. Principal Component Analysis (PCA) was also employed to reduce the dimensionality of the initial space. In the query phase, a feature vector is extracted from the query image and submitted to the index structure. The set of resulting numbers correspond to the identifiers of the images that are found to resemble the query one. Since the order of these identifiers is not ranked according to their level of similarity with the query example, an additional step for ranking these images using custom distance metrics between their feature vectors is further applied to yield the final retrieval outcome.

3.3 Textual information processing module

The textual query module attempts to exploit the shot audio information in the best way. This audio information is processed off-line with the application of Automatic Speech Recognition and Machine Translation to the initial video, so that specific sets of keywords can be assigned to each shot. Indexing and query functions were implemented using the KinoSearch full-text search engine library [17], which creates a binary index optimized for speed. The advantages of a post-coordinated retrieval system [18] were maintained by using a full-text engine on the keywords without any intervening attempts at Natural Language Processing to identify meaningful keyword combinations. We were thus able to maintain a full multidimensionality of relationship among the terms [19].

Stopwords were first removed from keyword files followed by the application of the Porter stemmer for English to all remaining keywords. Term weights for each keyword were then computed using the BM25 text algorithm, incorporating a collection frequency weight in the form of inverse document frequency (defined as the number of image/key-frames found to contain the keyword) and term

frequency (the number of times the term appears as a keyword for a given image/key-frame) [20].

The textual query module contains various devices to achieve post-coordination of the full-text index such as logical operators (Boolean operators OR, AND, and NOT), term grouping (using parentheses), and phrases (delimited by quotes), which all serve to improve precision. On the other hand, recall is boosted by using query expansion which is implemented by generating a list of synonyms for each query term from a local WordNet [21] database and then adding the most common synonyms to the original set of query terms until there are five terms in total. These are then used for subsequent matching against the keyword index.

Pre-coordinated term frequency weights are updated at query time with feedback from the post-coordinated query expansion device. This is first accomplished by boosting the weights of images/key-frames containing keywords that match an original query term. Secondly, weights are summed each time a match occurs between a keyword and an original or expanded query term. This enables ranking of the most relevant resources at the top of the results, thereby removing much of the noise generated by the increased recall achieved through query expansion.

To assist the user in subsequent query iteration tasks, a hierarchical navigation menu of suggested keywords is generated from each query submission (accomplished in parallel with the production of the expanded query described above). Traditional thesaurus term relationships [22] are generated from the WordNet database with broader terms and narrower terms derived from hypernyms and hyponyms, respectively. Related terms are provided by synonyms not used during query expansion.

Although performance of the module is satisfactory in terms of time-efficiency and navigation, the quality of the results greatly depends on the reliability of the speech transcripts.

3.4 Term recommendation module

For a given query, the term recommendation module proposes a set of related terms to the user in order to facilitate the query formulation process. The suggested terms are selected by taking into account their distribution in the automatic speech recognition data.

The selection of the recommended terms is done by taking into account the most frequent terms in the automatic speech recognition data. Out of this set of frequent terms, we semi-automatically selected a subset of proper names including persons, cities, countries, and organisations. For example, we considered terms such as *Milosevic*, *Amsterdam*, *Netherlands*, *Council*. Stemming algorithms were applied to this list to increase the performance of the recommendation module. For example, the terms *German*, *Germans* and *Germany* were all assigned to a single stem *German*.

Then we generated a co-occurence matrix containing the support of each possible pair of stems. Two stems are considered co-occurent if they appear together in the same transaction. A transaction corresponds to the automatic speech recognition data related to one given shot. The support of a pair of stems is the number of times they appear together in the whole set of transactions. For a given user query, the recommendation module proposes the most correlated terms.

3.5 Audio filtering

A user has an option to use additional filtering of the search results by applying audio content based filtering on the shots retrieved. Should a certain sound occurs in the video shots, a user has an option either to take or omit such shots from the list of shots retrieved. The following six sound classes are defined applause, laugh, screaming/crying, music, loud noise, and speech. First, common MFCC based low-level audio features were extracted from audio waveform. Then the Maximum Likelihood GMM based classifier was used for classification of audio patterns. Sliding one-second analysis window with a half-second shift was applied for classification. Finally, a video shot is assigned as relevant for the certain audio class if at least 1 second of audio, within the shot, is assigned to the given sound class. The system was trained on about 4 hours of various audio data that were manually extracted from various audio and video content (TV and radio broadcast programs, free Internet sources, TRECVID 2006 and 2007, and CD and mp3 audio recordings, the speech files consist of speech sounds in various languages, namely English, Slovak, Czech, Dutch, Chinese, and Arabic). More details about the audio classification system can be found in [10].

3.6 Relevance feedback module from QMUL

Relevance feedback (RF) scheme was initially developed for information retrieval systems in which it performs an online learning process aiming at improving effectiveness of search engines. It has been widely applied in image retrieval techniques since the 1990s. RF is able to train the system to adapt its behaviour to users' preferences by involving human into the retrieval process. An image retrieval framework with RF analyse relevant or irrelevant feedback from the user and uses it to predict and learn user's preferences. At the mean time, more relevant image can be successively retrieved.

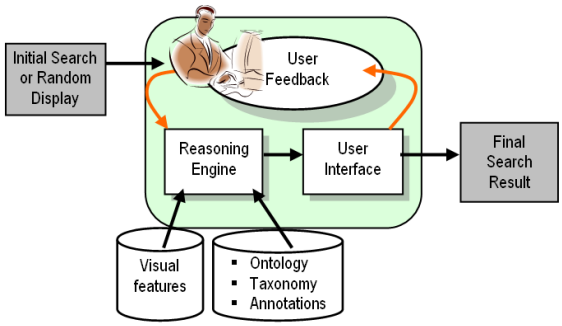

Figure 2: Generalised hybrid content-based image retrieval systems with relevance feedback.

A system contains RF process is illustrated in Figure 2. It needs to satisfy several conditions:

- Images are presented to the user for his/her feedback, but same images should not be repeated in different iterations.
- The input to the module is relevant/irrelevant information provided by the user on iterative bases.
- The module should automatically learn user's preferences by adapting the system behaviour using the knowledge feedback from the user.

A general image retrieval system with RF such as the one displayed in Figure 2 can use any kind of descriptors from low-level information of available content itself to prior knowledge incorporated into ontology or taxonomy.

When a learning approach is considered, many kind of reasoning engine can be used to determine relevant information. There are several common classes of RF modules such as: Descriptive models (e.g. Gaussians, GMMs), Discriminative models (e.g. SVMs, Biased Discriminative Analyses) and Neural networks (e.g. SOMs, Perceptrons).

In our framework, one of the RF modules is implemented by QMUL based on SVM. It combines several MPEG7 or non-MPEG7 descriptors as a cue for learning and classification. SVM is one of the developed supervised learning algorithms. It empirically models a system that predicts accurate responses of unseen dataset based on limited training sets [23].

In submitted runs with QMUL RF, all experiments were conducted using linear SVM for the sake of efficiency. Given the initial search result using visual similarity search or text-based search, users were asked to select at least one positive and one negative examples on screen as feedback. Usually two to five iterations were performed depending on users' preferences, within the time limitation. Four MPEG7 descriptors: Colour Layout, Colour Structure, Edge Histogram and Homogeneous Texture and one non-MPEG7 descriptor: Grey Level Co-occurrence Matrix were used and combined to conduct visual feature based RF [24].

3.7 High Level Concept Retrieval Module

This module provides high level concept (e.g. animal, landscape, outdoor, etc.) selection for the user. After an off line preprocessing the images are sorted based on similarity co-efficients for each concept.

The procedure required for the extraction of high level concept information is based on a combination of MPEG-7 and SIFT [13] based global image features.

A set of MPEG-7-based features are concatenated to form a single MPEG-7 feature, namely, color structure, color layout, edge histogram, homogeneous texture and scalable color. The SIFT-based feature is created in a 2-stage procedure. A set of 500 keypoints, on average, is extracted from each image. SIFT descriptor vectors (128 elements) are computed for those keypoints, and a set of 100 visual words is created by clustering the multidimensional space that is created. Using the Bag of Words (BoW) methodology [15], a new feature is created from the histogram of distances from each visual word. The MPEG-7 feature is exploited independently from the BoW feature. A set of SVM classifiers (LIBSVM [25] was employed in this case) is used to create classification models for the MPEG-7 and BoW features, using the first half of the development set for the training procedure. In order to keep the training set balanced between positive and negative input samples, a negative to positive ratio of 5 was kept for all concepts. The SVM parameters were set by an unsupervised optimization procedure, embedded in LIBSVM. The output of the classification is the Degree of Confidence (DoC), by which the query may be classified in a particular concept. After the classification process for both MPEG-7 and BoW features is completed, the respective DoCs are used as input to a second SVM. This stage-2 classification uses for training the second half of the development set and self-optimized parameters to create a classification model. The results from this 2-stage SVM on the testing set are sorted by DoC, and the 2000 higher in the rank are indexed in the database to support the concept retrieval.

3.8 Results

We have submitted five runs to the Search task. These runs employed different combinations of the existing modules as described below:

Run1. Text search + visual search

Run2. Visual search + High level concepts

Run3. Visual + text search + High level concepts

Run4. Visual search + RF + High level concepts

Run5. Visual + text search + RF + High level concepts

Here it has to be mentioned that the audio filtering was available in all runs while the text search included the also the results from the term recommendation module. The results achieved using these runs are illustrated in Table 2 below.

Table 2: Evaluation of search task results.

Run IDs	run1	run2	run3	run4	run5
Precision out of 100 retrieved shots	0.0708	0.0838	0.0633	0.0562	0.0562
Average precision	0.0441	0.0599	0.0559	0.0468	0.0312

Comparing the runs 1, 2 it is obvious that the high level concepts module improved the results of the system by a factor of 16% proving that it performs better than text search. In run 3 it can be observed that the text search failed to improve the results of run 2 as the performance slightly drops. In that case it seems that the text search module was incapable of providing good results as the annotations that were produced, after automatic machine recording and translation, were not descriptive of the video shot content.

The visual search in general performs better than text search, while their combination together with the high level concept retrieval module works the best out of all five runs. It seems that the RF module were capable of retrieving more relevant shots. However, the achieved scores using the RF module were not improved due to the limited time and the fact that the users did the experiments were relatively inexperienced with the interactive approaches. In all five runs, the limitation of time was considered to be 15 minutes rather than 10 minutes as defined in this year guidlines.

4 Rushes Task

After a thorough study of available test data, the typical rushes content structure has been identified. The main property of rushes content is repetitiveness of the same takes of camera with slight variations. Hence the first problem to address in the summary construction was to find the most effective way to remove the repetition from the content. The second objective, based on the assumption that all the repetitions have been removed, consisted in selecting such video segments which could have attracted the viewer. Thus mid level semantic features had to be combined in order to select such video segments. Finally, as rushes content contains a lot of unscripted parts, the third objective was to filter this content, such as junk frames (developed by VICOMTech and LaBRI).

Compared to our previous participations in the rushes task, the breakthrough in TRECVID 2008, as results show, consisted in redundancy removal. More details on our approach can be find in [26].

In [26], we make some highlights of our methods. *Redundancy removal*, the core of the approach, consisted in application of a spectral clustering approach such as normalized cuts [27] (developed by QMUL). First all video frames were clustered together. Then clusters close in time were grouped into scenes. Similar scenes are removed based on eigenvalues of frame similarity matrix. The second step in redundancy removal (developed by TU. Delft) consists of selecting only segments which contain semantic features such as detected camera motions [28] or faces [12] (developed by LaBRI).

The experiment showed that the new redundancy removal approach is very much efficient. On the *RE* criterion, the COST292 system had a very good score. On the contrary, and what could have been expected, on the *IN* criterion, results are not so good. It is partially due to the fact that the feature set for clustering was limited, as only MPEG7 Color Layout Descriptor was used. Besides high level features semantic set has to be enriched in the future.

5 Copy Detection Task

The two submitted runs to this task are respectively contributed by two partners of COST292: the Middle East Technical University (METU) and University of Brescia (U. Brescia).

5.1 Contribution of METU

5.1.1 Video Module

Video indexing maps the huge amount of content to a lower dimensional space for effective representation of the information conveyed. In this respect generally index values can be characterized as short and descriptive. However in our approach we utilize a representation that can be characterized as long and coarse index values inspired by the structure of gene databases as in [29]. The discriminative power resides in the length and robustness in the coarse structure of the index values, whereas fast queries are possible with the utilization of special indexing that involves multi-resolution and hierarchic structures. In a nutshell our approach works as a feature matching algorithm between a query and reference videos in which features are extracted from spatial and temporal units of the video. These units are constructed from a uniform grid, in which a spatial overlap is attained. And multi-resolution property is introduced by utilizing down-sampled versions of the frames into the equation. Furthermore, temporality is achieved by extending each grid element in time yielding a rectangular prism. For each prism, a feature vector is computed from a set of feature extractors that spans three low-level video contents that are color, texture and motion. Our set of feature extraction methods is constructed based on MPEG-7 visual descriptors. It should be noted that the modifications are introduced in order to decrease the computational complexity and introduce the ability of coarse representation. Note that, these pseudo-MPEG-7 features [30] are extracted from each grid element and concatenated to form a single feature vector for a single prism that extends through time and space on the video. Coarse representation is further accentuated by a quantization of the feature vectors. Also, with this step computational complexity of a query/feature search is decreased considerably. Quantization models for the feature groups, including quantization levels and reconstruction values, are computed from a 100 hour of the reference video database by employing Lloyd-Max Quantization Algorithm. After building the feature database of reference videos, queries are searched on this database by localizing the matching segment(s). For the matching, dot-product of each consecutive

feature vector is averaged over the duration of the query and the segment with the maximum average is localized.

5.1.2 Audio Module

Audio copy detection system should be robust against several attacks i.e. compression, bandwidth limitation and mixing with uncorrelated audio data. Our audio analysis method is based on a similar audio hashing scheme defined by Haitsma et al. [31]. However, in our approach instead of 32 bit hash representation of audio content, 15 bit hash values are used. Each audio frame is Hamming windowed before STFT. The spectrum is divided into 16 non-overlapping bands in the range of 300Hz-3000Hz according to Bark frequencies. Using the energy values of these 16 bands, a 15 bit hash value is computed for the frame. The hash function used to generate the audio hash is as follows;

$$\begin{aligned} H(n,m) &= 1 \quad if \quad EB(n,m) - EB(n,m+1) > 0 \\ &= 0 \quad if \quad EB(n,m) - EB(n,m+1) \leq 0 \end{aligned}$$

where $EB(n,m)$ denotes the energy in frame n and band m. Search operation is based on the extracted hash values which allow for efficient database queries. Firstly matching hash positions are determined from the hash table for each value of the query. For each exact match position 500 frame window on both query and reference data is taken. After calculating BER, window is reported as a true match if BER is lower than 30%. A post-processing is applied in order to determine start and end positions of intersecting result intervals for a query. In this data set the attacks applied to original content are considerably harsh so 15 bit representation gives a coarse representation of the audio signal which increases the robustness against strong signal degradations as compared to a 32 bit representation. While 15 bit representation increases the number of match positions which increases the number BER computation, it decreases the miss probability. And 15 bit representation is also advantageous with low memory constraints.

5.1.3 Video-only Runs and Results

COST292.v.metuq1: In this run the aforementioned video module is utilized. Dominant color is represented as a single color component for all color channels and all grids in all levels. This is backed up by a single structured dominant color for all channels in all grids of first two levels. Texture is represented by a mean brightness value and three primary 3D DCT coefficients for the luminance channels of the grids in first two levels. Also an edge energy value is calculated in luminance channel for all grids in these two levels. Finally, motion is represented by a motion activity value for all grids in first two levels. These features are then quantized to two bit representations and then queries are performed by the linear search method explained previously. Mean F1 performance for all transformations except PIP attack is above the median for this run.
COST292.v.metuq2: This run differs from the first run only in the representation of the color structure. Dominant color and structured dominant color features are converted to color cluster percentages. These clusters are obtained by first segmenting each grid channel with their histograms at predefined values. Cluster limits are obtained by an analysis on the dominant and structured dominant colors obtained from a sample set. These percentages are sorted for direct comparison between features from grids in different videos. As a result, the color structure in each grid channel is represented by the ratios of several distinct color clusters rather. This modification increases the performance of the previous run along with a slight increase in computational complexity. With this modification overall F1 performance of this run is slightly improved especially for the first transformation.

5.1.4 Audio + Video Runs and Results

Our approach to utilizing both the audio and video contents for content copy detection is combining query results from audio only and video only results. Query results from video content are obtained from the algorithm of our second video only run and query results from audio content are obtained by an independent algorithm explained in the previous sub-section. Combining the two query results is a process in which for each query clip, a list of result matches is found by interleaving the matches from

audio only and video only results. Each three results for audio + video runs use different interleaving schemes, which are explained below.

COST292.m.A0metuq1: This run is generated by the audio-only match results only for benchmarking purpose. Results coming from video matches are discarded in order to evaluate our audio only content copy detection algorithm. For all seven transformations applied to the queries the mean F1 score is between 0.8 and 0.9. The mean F1 score for the first transformations is better than the last three transformations, which have mixing with speech attack.

COST292.m.A0metuq2: Confidence values for audio and video match results are not generated by same procedures. An essential step in combining these two set of results is to modify confidence values in a way that they are normalized in comparable ranges. The audio matching algorithm is observed to result in less false alarm and miss rates compared to the video matching algorithm. Thus, video match confidences are modified in a way that ranks at least one audio match above any video match for all queries. This is achieved by multiplying the confidence values of video matches with the maximum confidence found in audio matches for each query given that audio confidence value is ¡ 1. This scales down video match confidences enough to rank the best audio match the highest for any query. Further, if a database match is given by both audio and video match algorithms, this result is combined into one and given a confidence of the sum of two confidences. Temporal location of the combined match is copied from the combining audio match since the audio algorithm works on a higher temporal precision. The query results obtained from this method had slightly better performance compared to the previous benchmark run.

COST292.m.A0metuq3: This run is generated by an altered form of the previous combining algorithm. This time all confidence values from both audio and video matches are left as is to favor video results more. Although results of this run performed better than audio only matches, it performed worse than those of previous run.

5.2 Contribution of U. Brescia

The most of the existing methods for CBCD are mainly based on the use of low-level features and in particular a subset of low-level features: motion, color (intensity, histogram, etc.) and audio. These features are utilized in a broad set of techniques for CBCD, like for instance the analysis of spatio-temporal variations, the reuse of traditional hash algorithms, etc.

In this run we propose and test a new method which is characterized by extracting a mid-level feature, as opposite to more traditional low-level feature based approaches. The mid-level features we used is related to the temporal structure of videos: given the shot boundaries of a certain video, we consider as a mid-level feature the shot length sequence. The idea is indeed to extract the shot video segmentations from videos belonging to both dataset and queryset, and then to compare them.

For each video of the dataset, its signature in terms of shot lengths sequence is extracted: $V = \{v_i\}|_{i \in N}$ where v_i is the length in frames of the $i-th$ shot and N the number of shot of the considered video. The signature is extracted for each query video of the queryset, as well: $Q = \{q_j\}|_{j \in M}$. For each video query, first and last component of the extracted length sequence are discarded since the extrapolation is random and hence in the most of the cases first and last frame of each query video will do not be a boundary frame between shots.

Each query video is then compared with all videos of the dataset. Let's consider the shot length sequence of the query video as a sliding window. This window scans a given video of dataset and stops at each video cut. At this point, a distance d is evaluated between all the M valid shots of the query video and the same number of shots of the dataset video counted from the stop point (Figure 3). As distance measure, the standard Manhattan norm is selected:

$$d(\mathbf{Q}, \mathbf{V}) = \sum_{k=p}^{p+M-1} |q_k - v_k|.$$

where p is the index of the stop point. The distance d is evaluated at each stop of the sliding window. For a whole video, the minimum of these distances is found and its value and position in the dataset video are stored; so for every dataset-query pair we have a distance and its associated position. This operation is performed for all videos of dataset. At last, for every query the best three candidates in terms of distance in the dataset are selected as the retrieved dataset videos. Since it can happen that

there are queries which are not meant to be retrieved, only answers in which the distance is smaller than a chosen threshold are considered. Also, signatures composed by a single valid value must be discarded and a retrieval answer cannot be given.

Because of the degradation of the query video, it is inappropriate to evaluate the simple distance between the shot lengths sequence; instead, the cumulative sequence (which represents the relative distance to the video fragment start) is used to compare two signatures. Besides, signature performance can be strongly compromised if false cut detections occurs during the preliminary shot boundary extraction. In this sense we had to introduce some constraints in the proposed method. First, shots of duration under 100 frames are not considered because a false shot of short duration between adjacent shots can be detected if some kind of fast motion occurred in the video. Second, a shot merging operation has been considered when a false cut can be obtained when a long shot is mistakenly divided in two because of a dissolve, of a strong motion activity, etc.

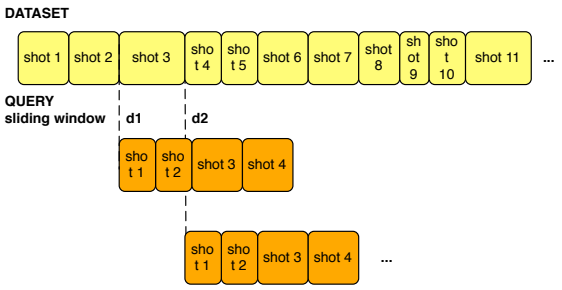

Figure 3: CBCD method using shot boundaries mid-level feature.

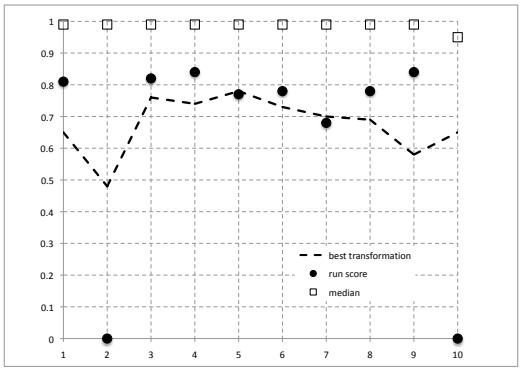

Figure 4: TRECVID experimental results for CBCD (video-only).

5.2.1 Video-only Run and Results

Videos of both TRECVID queryset and dataset have been segmented in shots using [32]. In Figure 4 TRECVID results are shown. First we observed that two queries failed (dot with mean equal to zero) for bad setting during the running process. The others transformations are near the median (dotted line) and in some cases even closed to the best.

6 Conclusion

In this paper, we reported the COST292 experimental framework for TRECVID 2008 evaluation. COST292 participated in four tasks: high-level feature extraction, search, rushes and copy detection. We enhanced the developed 2006 and 2007 systems from various aspects. In addition, contributions have been made to the new copy detection task. The submissions to all these tasks were successful and many used applications have shown good performance.

Acknowledgement

The research and experimental platform leading to COST292 submission to TRECVID 2008 was partially supported by the European Science Foundation under the COST292 Action.

References

[1] Alan F. Smeaton, Paul Over, and Wessel Kraaij. Evaluation campaigns and trecvid. In *MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, pages 321–330, New York, NY, USA, 2006. ACM Press.

[2] J. Calic and all. Cost292 experimental framework for trecvid 2006. November 2006.

[3] Q. Zhang and all. Cost292 experimental framework for trecvid 2007. November 2007.

[4] E. Spyrou, G. Tolias, Ph. Mylonas, and Y. Avrithis. A semantic multimedia analysis approach utilizing a region thesaurus and LSA. 9th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS 2008), 2008.

[5] N. Dalal and B. Triggs. Histogram of oriented gradients for human detection. In *In Proc. Of the 2005 IEEE Computer Society Conference On Computer Vision and Pattern Recognition (CVPR '05), USA, San Diego, June 20-25*, 2005.

[6] J. Huang, R. Kumar, M. Mitra ans W.-J. Zhu, and R. Zabih. Image indexing using color correlograms. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1997.

[7] Krishna Chandramouli and Ebroul Izquierdo. Image classification using self organising feature map and particle swarm optimisation. In *Proceedings of 3rd International Conference on Visual Information Engineering*, pages 313–316, 2006.

[8] T. Kohonen. The self organizing map. *Proceedings of IEEE*, 78(4):1464–1480, September 1990.

[9] Q. Zhang and E. Izquierdo. Combining low-level features for semantic extraction in image retrieval. *Eurasip Journal on Advances in Signal Processing*, 2007:61423 –, 2007. Object signature;Optimal linear combination;Image primitives;.

[10] R. Jarina, M. Paralic, and M. Kuba et al. Development of a reference platform for generic audio classification. 9th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS 2008), 2008.

[11] OpenCV. http://opencvlibrary.sourceforge.net, 2007.

[12] A. Don, L. Carminati, and J. Benois-Pineau. Detection of visual dialog scenes in video content based on structural and semantic features. In *International Workshop on Content-based Multimedia Indexing (CBMI) 2005*, Létonie (Tampere), 2005.

[13] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[14] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–239, Mar 1998.

[15] J. Sivic and A. Zisserman. Video google: a text retrieval approach to object matching in videos. volume 2, pages 1470–1477, 2003.

[16] V. Mezaris, H. Doulaverakis, S. Herrmann, B. Lehane, N. O'Connor, I. Kompatsiaris, and M. Strintzis. Combining textual and visual information processing for interactive video retrieval: Schema's participation in trecvid 2004. In *TRECVID 2004 - Text REtrieval Conference TRECVID Workshop*, MD, USA, 2004. National Institute of Standards and Technology.

[17] Kinosearch search engine library. `http://www.rectangular.com/kinosearch/`.

[18] J. Aitchison, A. Gilchrist, and D. Bawden. Europa Publications Ltd, 4th edition edition, 2000.

[19] Frederick Wilfrid Lancaster. *Vocabulary control for information retrieval*. Information Resources Press, Arlington, Virginia, 2nd edition edition, 1986.

[20] S.E. Robertson and K. Spärck Jones. Simple, proven approaches to text retrieval. Technical Report UCAM-CL-TR-356, University of Cambridge, Computer Laboratory, dec 1994.

[21] Christiane Fellbaum, editor. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press, May 1998.

[22] Guidelines for the construction, format, and management of monolingual controlled vocabularies, 2005. `http://www.niso.org/standards`.

[23] Christopher J.C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, 1998.

[24] D. Djordjevic and E. Izquierdo. Kernel in structured multi-feature spaces for image retrieval. *Electronics Letters*, 42(15):856–857, 2006.

[25] Chih C. Chang and Chih J. Lin. *LIBSVM: a library for support vector machines*, 2001.

[26] S. U. Naci, U Damnjanovic, B Mansencal, J. Benois-Pineau, C Kaes, M Corvaglia, E Rossi, and N Aginako. The cost292 experimental framework for rushes task in trecvid 2008. In *TVS'08*, Vancouver, British Columbia, Canada, 2008.

[27] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8), 2000.

[28] P. Kraemer, J. Benois-Pineau, and M. Gràcia Pla. Indexing camera motion integrating knowledge of quality of the encoded video. In *Proc. 1st International Conference on Semantic and Digital Media Technologies (SAMT)*, December 2006.

[29] Tamer Kahveci and Ambuj K. Singh. An efficient index structure for string databases. In *In VLDB*, pages 351–360. Morgan Kaufmann, 2001.

[30] B. Manjunath, P. Salembier, and T. Sikora. *Introduction to MPEG 7: Multimedia Content Description Language*. Ed. Wiley, 2002.

[31] Jaap Haitsma and Ton Kalker. Robust audio hashing for content identification. In *In Content-Based Multimedia Indexing (CBMI)*, 2001.

[32] N. Adami and R. Leonardi. Identification of editing effect in image sequences by statistical modeling. pages 0–4, Portland, Oregon, U.S.A., Apri 1999. Picture Coding Symposium.