vdFP – Video Fingerprinting Technologies for Media and Security Applications

D1: Report on Existing Technologies

DATE 2010-02-09

ABSTRACT In this document we report on existing technologies for video fingerprinting, audio fingerprinting, near duplicate detection and video linking inside and outside our consortium.

AUTHOR, COMPANY John Schavemaker (TNO), Peter Jan Doets (TNO), Werner Bailer (JRS), Harald Stiegler (JRS), Felix Lee (JRS), Helmut Neuschmied (JRS), Wessel Kraaij (TNO), Paul Brandt (TNO), Pieter Eendebak (TNO), Elena Ranguelova (formerly TNO), Andy Thean (formerly TNO)

KEYWORDS VIDEO FINGERPRINTING, AUDIO FINGERPRINTING, NEAR DUPLICATE DETECTION, VIDEO LINKING

RELATED ITEMS BROADCAST MONITORING SYSTEMS, COPY DETECTION, MEDIA MONITORING, TRECVID, LOGO RECOGNITION, OBJECT REDETECTION, SETTING DETECTION

DOCUMENT HISTORY

Release	Date	Reason of change	Status	Distribution
0.10	2009-11-13	Copy from internal document	Living	Confidential
0.20	2009-11-19	Changes to JRS sections and intro	Living	Confidential
0.30	2009-11-24	Changes to TNO sections, inclusion of Creative Commons License	Living	Creative Commons Attribution Non-Commercial Share Alike
1.00	2009-11-26	Changed font of license section and made document final	Living	Creative Commons Attribution Non-Commercial Share Alike
1.11	2010-02-09	Minor change concerning technology of Margaux Matrix, merged with version 1.10	Living	cc BY NC SA

Table of Contents

1 Executive Summary

Video fingerprinting is a proven and commercially available technique that can be used for content-based copy detection. The task of a video-fingerprinting system is to detect whether a particular segment of video is (partly) based on the same original video as video footage in a database of reference videos. Typical applications are in the media domain (detection of copyright infringement, counting broadcasts of advertisements, detection of changes in advertisements) but new application domains are in development (e.g. detection of illegal digital video material such as child-abuse material on hard-disks).

In this document we report on the existing technologies for video fingerprinting inside and outside our consortium. Besides video fingerprinting, we also report on adjacent technologies like audio fingerprinting, near duplicate detection (video fingerprinting for videos that are not exactly the same) and video linking (finding copies of visual objects in different videos). For each technology item, we present the state-of-the-art and give a list of commercially available systems. Furthermore, for each item, we discuss its availability in the consortium.

2 Introduction

TNO and **JOANNEUM RESEARCH** have developed a cooperation-model to increase their competitive position and to make better use of internally available technologies and know-how. This project aims to develop a joint research program and marketing plan in the field of video fingerprinting together with two local industrial partners and one end user organization.

The JOANNEUM RESEARCH Institute of Information Systems **(IIS**) and TNO Information and Communication Technology have found common interest in the field of video fingerprinting. Video fingerprinting (vdFP) is a technique that can be used for content based copy detection. The task of a vdFP system is to detect whether a particular segment of video is (partly) based on the same original video as video footage in a database of reference videos. Typical applications are in the media domain (detection of copyright infringement, counting broadcasts of advertisements, detection of changes in advertisements) but new application domains are in development (e.g. detection of illegal digital video material such as child-abuse material on hard-disks).

Two industrial partners and one end-user organization will participate in this project: the Dutch SME **ZiuZ** with experience and interest in bringing video fingerprinting solutions into the security/police market, the Austrian SME **HS-ART Digital** with background and experience in marketing such solutions in the media market. The Dutch National Audiovisual archive of **Sound and Vision (S&V)** will also participate in this project since they have particular interest in video fingerprinting technologies as part of the *Images for the future* project, a large 7-year national effort focused on digitizing Dutch video and film content.

3 Video Fingerprinting

In this chapter we describe some typical applications of video fingerprinting in Section 3.1. In the next section, 3.2, we discuss the state-of-the-art in video fingerprinting. Sections 3.4 and 3.5 introduce video-fingerprinting systems available in our consortium. Section 3.6 lists other commercially available systems for video fingerprinting. Finally, in Section 3.7 we show evaluation results of video fingerprinting for TRECVID 2008, TRECVID 2009 and MediaCampaign.

3.1 Applications

The following application scenarios for video fingerprinting can be identified [LTC07]:

- Monitoring of TV advertisements (counting)

- Detection of copyright infringement (e.g. camcorded feature films, illegal re-encoded movies on the internet, or movie extracts)

- Re-use of archive material (post production artifacts: logo insertion, picture in picture)

- Detection of known CA material on seized hard disks

3.1.1 Copy Detection

One of the most prominent applications for copy detection is finding illegal copies of video content (cf. [HB01,HHB02]). A related problem is the identification of known unwanted content in public access video databases [CBF06].

In [SEI09] the requirements for fingerprinting of to protect movies shown in cinema are describes and preliminary test results are presented.

The Institute National pour l'Audiovisual [INA] is developing a fingerprinting system for monitoring broadcast channels for archive content. The solution is also marketed as a commercial product.

3.1.2 Media Monitoring

The objective is to identify known or new creatives (a creative is the occurrence of the same spot, in one language, one media), to cluster them into campaigns (sets of semantically related creatives) and finally to estimate expenditure.

The requirements are differ from other fingerprinting applications. The problem requires near real-time matching, as creatives information shall be available 1-2 hours after broadcast and videos need to be matched against creatives from the last 1-2 months. The fingerprinting algorithm needs to deal with different video sources (analog/digital, IPTV, different bandwidth), spots of 1-240 seconds duration and must operate on several channels 24/7. The target performance is a spot identification accuracy >98.5%, false negatives < 0.5% and false positives < 1%.

3.1.3 Archive Documentation

In archive documentation, video fingerprinting can be used to identify reuse of video clips and thus to apply documentation from the complete programme to source clips or vice versa. This use case is described in more detail in Section 6.

3.2 State-of-the-art in Scientific Research

Video fingerprinting techniques borrow heavily from robust image duplicate search and the general framework as has been developed for audio fingerprinting. A brief overview of video fingerprinting techniques will be presented based amongst other on a short survey by Law-To et al. [LTC07]. Video fingerprinting systems do essentially have the same base structure as audio fingerprinting systems. The most prominent differences are: feature extraction is not carried out in the frequency domain but exploits temporal, spatial, local and global features (such as luminance). In addition, produced video is composed of shots, which can help to decompose a video clip in relatively coherent segments.

3.2.1 Approaches

Amongst the multitude of approaches for similarity matching of video clips, many apply still image features for clustering key frames of video sequences. Some of these approaches use more video oriented features such as sequences of key frames and camera motion [ZEX05] or motion trajectories [CCM97]. However, while the feature extraction and matching approaches proposed in these works are relevant for our problem, similarity matching in these approaches is quite coarse. The approaches are optimized towards compact feature descriptions and scalability rather than precise matching of the content of a sequence. If large sets of material need to be processed, similarity matching can be used as a preprocessing step to find candidate sequences that are then matched with a more precise measure.

A number of approaches have been presented in the context of the TRECVID 2008 copy detection task.

The method described in [KBG08] uses the MPEG-7 visual features ScalableColor, ColorLayout, Color-structure, Homogeneous Texture and EdgeHistogram as signatures (extracted by MPEG-7 XM software). The signatures of every first frame per 2 seconds in the query video are compared to the center frames of every shot in the data set. Custom similarity measures are used to compare MPEG-7 descriptors: Meehl index, pattern differences and city block distance. The ratio of most similar match to the 5th one is used for classifying between similar and dissimilar.

The authors of [CJ08] propose to use only intensity information. The method is based on classifying key frames of a segment (max. 500) into (two) classes with different template labels. As templates the bounding boxes of Harris points are used. The partition method is based on the overlaps of the templates. This approach allows to deal with different editing patterns, e.g. the detection of temporal redundancy (insertions of caption or pattern, black margins). The key frame signature generated by scaling the image down to 9x11 blocks, results in a 99-dimensional vector.

In [GZL08] an approach for segmenting the video temporally and extracting features for the segments is presented. The features used for segmentation are global intensity histogram, intensity ordinal measurements and enhanced local color histogram, the features used as descriptors of the segments are local edge histogram descriptor, Canny edge, SIFT and Gabor color moment. A segment boundary is found when difference of frame-differences exceeds a threshold for at least two of the features. There are different strategies for search for long and short query video clips, but no details about matching are given in the paper.

The authors of [Zha08] present two approaches for video fingerprinting and a simple combination for video and audio fingerprints (multiplying confidence values of the two approaches). One video fingerprinting method uses MPEG-7 visual descriptors extracted from spatial-temporal elements (grid-time prisms). A coarse representation is created using Lloyd-Max quantization. For matching, the maximum average of the dot-product of each consecutive feature vector over the duration of the query is located. The other video fingerprinting approach uses shot lengths as signature of each video. The cumulative shot length sequence (distance to the video fragment start) is used for matching. To reduce problems with false positives in the shot detection, shots under 100 frames are merged with adjacent ones.

A "bag of visual terms" approach is used in [HFG08] for representing key frames built by k-NN between SIFT descriptors. Probabilistic latent space modeling is used to find the local matches between query video and reference video key frames. In a post processing step, RANSAC is applied in the time domain to achieve temporal consistency, i.e. ensure coherent segment matches in time up to translation and scaling factors.

The approach in [DGJ08] uses local features extracted from key frames. Key frame extraction is done either on a regular basis or based on motion activity; experiments indicate that regular basis is better. The Hessian - affine region extractor is used for interest point detection and SIFT descriptors are extracted for the regions. In an off-line step partitioning of the SIFT descriptor space of the database is performed, yielding a set of "centroids" (similar to the "Video Google" approach). The SIFT descriptor is assigned to centroid and only a list index to the appropriate centroid is stored. The SIFT descriptor is further assigned a binary signature. In the matching step the centroid must be identical and the Hamming distance is used to calculate the distance of the signatures. If Hamming distance is below a threshold, the distance is weighted by a weighting function not described in the paper. The frame matching score is determined from the sum of all matched descriptors of their weighted distance between query and source descriptor. Frame grouping and geometrical verification are used to build "global" overview out of local approach.

In [JLB08] two fingerprinting approaches as well as way of applying them sequentially to improve the results are presented. The first use trajectories of interest points and classifies them into three classes. Matching is done by comparing the sets of class labels of the trajectories. The second is an extensions of a still image approach. It extracts "dissociated dipoles", i.e. non-local differential operators extracted around Harris points are extracted, yielding 20-dimensional normalized features.

The authors of [GG08] apply Nonnegative Matrix Factorization to each video frame and use dimensionality reduction to get matrices of rank 2. The resulting matrices are matched.

In [LWS08] a local feature based approach is used, extracting patches along the trajectory of tracked feature points. Classification of the motion behavior in a patch according to an "inconsistency" criterion is done. If the motion in the patch is uniform, "inconsistency" is close to 0, if not uniform (e.g. motion in different directions in a patch) close to 1. The motion behavior (classification) in a patch generates a so called "inconsistency sequence". Sequences are matched and have to registered temporally. The approach is rather slow, and has problems with short shots (i.e. short trajectories) and static sequences.

In [KSV08], a so called best matching unit (BMU) signature is calculated for each frame. The signature is based on ColorLayout and EdgeHistogram and self organizing maps (SOMs) trained on the extracted features. The approach compute all possible alignments of query and clip from the database and determine the distance based on low pass filtered BMU signatures (considering 4 subsequent map units).

The authors of [Lia08] use key frames sampled every 20 frames and extract up to 512 SURF descriptors from each key frame. A key frame descriptor is created as a histogram of SURF descriptor visual words (i.e. descriptors clustered into bins). The approach iteratively refines match position and range on individual frame descriptor matches to discard outliers.

The approach described in [OHP08] is based on normalized Hu moment invariants (NHMI), extracted globally from every frame. The 2nd to 6th moment are used, normalized by the power of the 1st in order to accommodate for gamma and quality changes. A signature of 6 integers per frame is created from the moments. The descriptors are matched within a sliding window, going over all videos.

An image signature based on the trace transform has been proposed in [BB09]. The image signature is extracted globally as well as locally around points that are robust in scale space. The proposed approach will be the MPEG-7 image signature, the work on the video signature is ongoing.

The authors of [DL09] propose a system for mining broadcasts for recurring video sequences based on gradient histograms. The approach assumes that no transformations have been applied to the clips. The system searches for short matching clips and connects them to longer sequences using domain specific filter rules. The approach is capable of yielding frame precise information about repeated sequences at a speed several times faster than real time on a state of the art computer.

Other recent works are [LKF09,LS09,RB09].

3.2.2 Fingerprint Generation

In correspondence to audio fingerprinting, a necessary first step for video fingerprint generation is normalization of data. This involves decompression, re-sampling if necessary. Some techniques require shot segmentation information. Law-To et al. [LTC07] mention the following different techniques for feature extraction: temporal patterns of shot boundaries [IIS99], a technique which will not work well for short video segments; spatiotemporal signatures based on the differential luminance in a grid based frame partitioning, looking at spatial and temporal differences [OKH02]; global descriptors based on motion, color, and spatiotemporal distribution of intensities [HB01]. The latter approach has been derived from a technique developed for robust image duplicate search [BBN96]. Li et al [LJZ05] propose a method for the monitoring of TV advertisements, based on global color histograms. This is a relatively easy problem, since it is not necessary to offer robustness against moderate distortions or production artifacts. To address the latter type of transformations, local descriptors such as salient points, Harris points, space time interest points, sift etc. seem more robust [IKY06, JFB04]. This approach is known to be robust against occlusion, variations in luminance, which is good, but is also invariant to viewpoint changes, which is not necessarily good for copy detection. In addition, candidate local descriptors need to be pruned, in order to create a compact signature for each frame. Another possibility to achieve compact fingerprints is to limit the number of frames that are fingerprinted, e.g. by sampling or by looking at just the key frames that correspond to large motion activity.

Finally, signatures are computed based on concatenation of feature values (and in the case of local features, spatial position) leading to trajectories. Typical values of signature sizes are 60,000 features for 1 hour of video [JBF07].

3.2.3 Fingerprint Matching

A typical architecture for fingerprint matching is based on two steps. In a first step, copy candidates are selected using a fast but coarse algorithm. In the second step, these matches are ranked according to the more precise fingerprint distance computations. One of the approaches to implement scoring is voting for a video-id for each local feature extracted from the test video.

3.3 Broadcast Monitoring System “Genifer”

Joanneum Research developed the video broadcast monitoring system “Genifer”. It enables the observation of up to 70 broadcast channels. The fingerprint database can manage the fingerprint of up to 200.000 films and of 10.000 short movies (e.g. advertising films). A short movie has a length between 30 seconds and 10 minutes.

The redetection of films is mainly based on the detection of the shot structure. For that the relative length of consecutive shots are extracted. Using the relative lengths gives some robustness against speed changes. The fingerprint of a film contains a series of these numbers according its shot structure.

Before a film can be redetected the fingerprint data of the film has to be extracted and has to be saved in a database. To enable a fast search database index values are calculated from shot groups. For the calculation of the index value the absolute shot length values are determined with a tolerance range

The film detection process is divided in three parts:

1. **The film begin time detection**: For detecting a film the first time the database index values for shot groups are used. A film is detected if 14 of 20 consecutive shots are detected.
2. **The continuous film detection**: Once a film is detected the expected relative length of the actual shots can be continuously fetched from the database. Through the saved matching history it is possible to recognize missing or false detected shot borders and to use this information for the matching of the relative shot length.
3. **The film end time detection**: The end of a film is observed if in the last two minutes no matching shot pairs are detected or if several of the last 20 shot length values do not match.

In difference to a film the duration of a short movie is less or equal 15 minutes. It is possible that the short movie has only one shot. For this reason absolute shot length values are used for the fingerprint. Additionally for each shot a key-frame is saved with resolution of 64x48 pixels. The database index values are calculated from one, two, and three consecutive shots. The index value is composed of the absolute shot length values and of four image parameter values from each key-frame.

The image parameter values have to be robust against different image qualities and image modifications like cropping and image format changes (e.g. letter box) and the needed calculation time should be as low as possible. This is the case for the calculation of image symmetry values. An image parameter value indicates the relative brightness difference of according pixels in two different image areas. To reduce the influence of image noise and overexposure or underexposure only pixels values of a defined value range are used for the calculation of an image parameter value. If the number of relevant pixels is below a relative threshold than the image parameter value is set as undefined.

Once a short movie is redetected by the database search index an additional check is done by matching the observed key-frames with the according key-frames saved in the database.

3.4 ZiuZ Video-Fingerprinting System

ZiuZ is in the front line of development high-tech applications in the field of digital image and video processing. Ziuz is predominantly focused on the development of applications to support police investigations involving image or video but is also expanding its market segments.

ZiuZ has developed a dedicated video fingerprinting system for the Dutch police to aid vice officers in investigations of child abuse cases. The system can automatically classify confiscated material into legal and not-legal based on a national fingerprint database of known child-abuse material. The system also tells users whether videos have been seen by other investigators.

Material in the child-abuse domain usually consists of short, low-quality, low-resolution videos. The ZiuZ system is tuned to that specific application but it may also be of interest to parties that have old, low-quality archive material like Sound & Vision.

ZiuZ has filed a patent on the underlying fingerprinting technology (http://www.ziuz.com/).

3.5 Existing Commercial Systems

The following commercial systems have been identified after searching the internet with ”video fingerprinting” , ”video copy detection” and looking at several directories, such as the Wikipedia entry on video fingerprinting:

- Civolution (www.civolution.com)
- Ipharro (www.ipharro.com)
- Audible Magic (www.audiblemagic.com)
- Advestigo (www.advestigo.com)
- Ziuz (www.ziuz.com)
- Auditude (www.auditude.com)
- Vobile (www.vobilein.com)
- Yuvsoft (www.yuvsoft.com)
- Zeitera (www.zeitera.com)
- Vidyatel (www.vidyatel.com)

3.6 Evaluation

3.6.1 TRECVID 2008 Results

TNO has initiated a dedicated content based copy detection task at TRECVID [SOK06] in close co-operation with INRIA-IMEDIA. Similar to the copy detection showcase at CIVR 2007, the TRECVID 2008 task was based on a scenario where video copies were synthetically created, by inserting reference clips in unrelated material and applying distortions and transformations in order to simulate real copied clips. The list of transformations included quality decreasing transformations such as: blur, frame drops, compression, aspect ratio change, gamma change, cam cording and post production artifacts such as logo insertion, picture in picture and random combinations of transformations.

The test collection consisted of 201 test clips: 67 clips consisted only of reference material (educational TV programs from the Netherlands Institute for Sound & Vision), 67 clips were cut from unrelated material (rushes from BBC dramatic series), 67 clips were composed of a segment of reference video padded with unrelated material. All shot lengths were determined at random in a range between 3 seconds and 3 minutes. All transformations were parameterized, and parameters were chosen at random within a pre-specified range for each test clip.

Each test clip was transformed by 10 different transformations. As a result, the performance of systems can be evaluated for different individual transformations (201 ’random’ data points per transformation). The test collection was split in a set consisting of the video material only, and a set of audio files. Twenty two teams participated in the TRECVID CBCD task, some of them new to the problem, some of them very experienced teams. Most of the teams participated submitted runs for the video only task, only one team (TNO) submitted teams for the audio only task. Two teams submitted runs for the combined audio+video task.

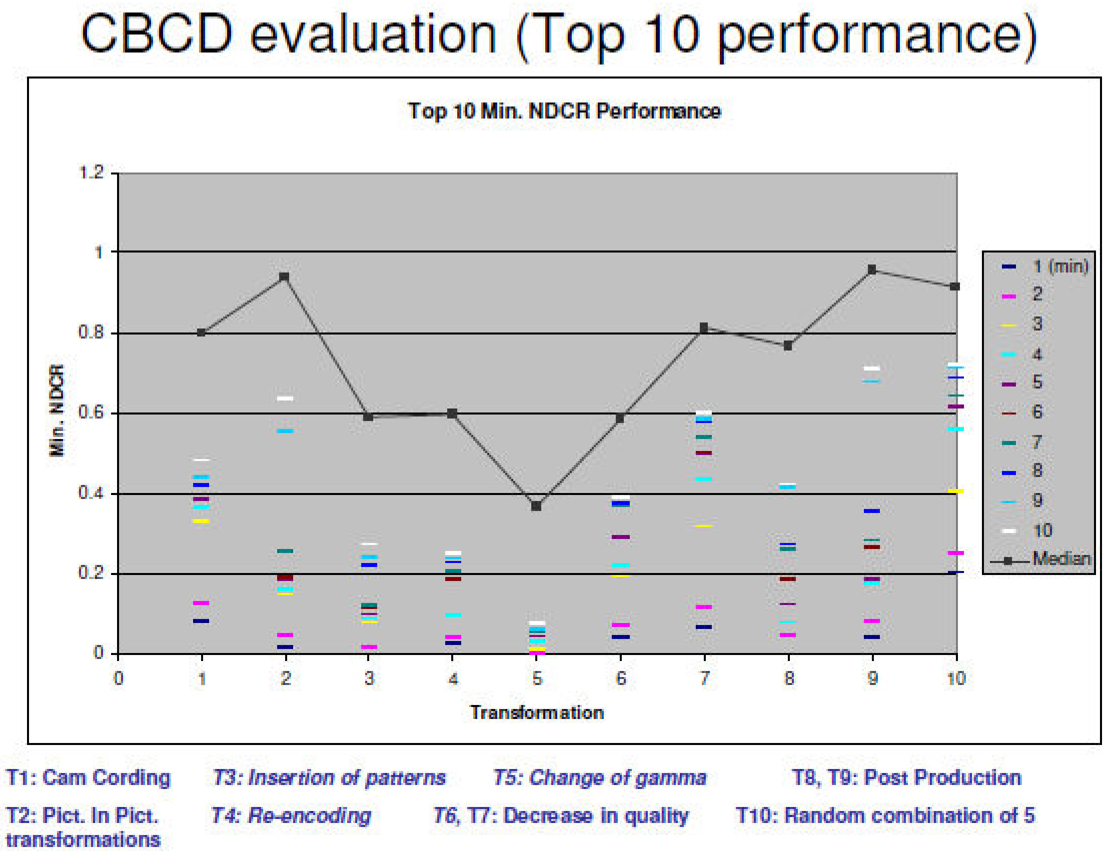
System performance was measured on three aspects [OAR09]:

detection quality has been measured by the Normalized Detection Cost Rate, which is a weighted average between the false alarm rate and the probability that the system misses a copy.

localization accuracy The asserted and actual extents of the copy in the reference data are compared using precision and recall and these two numbers are combined using the F1 measure, recall and precision are measured at the optimal operating point where the normalized detection cost is minimal.

processing speed mean processing time per test clip.

The results of the TRECVID CBCD evaluation for these three measures are summarized in figures 6, 7, and 8.



CBCD evaluation (Top 10 performance)

T1: Cam Cording

T2: Pict. In Pict. transformations

T3: Insertion of patterns

T4: Re-encoding

T5: Change of gamma

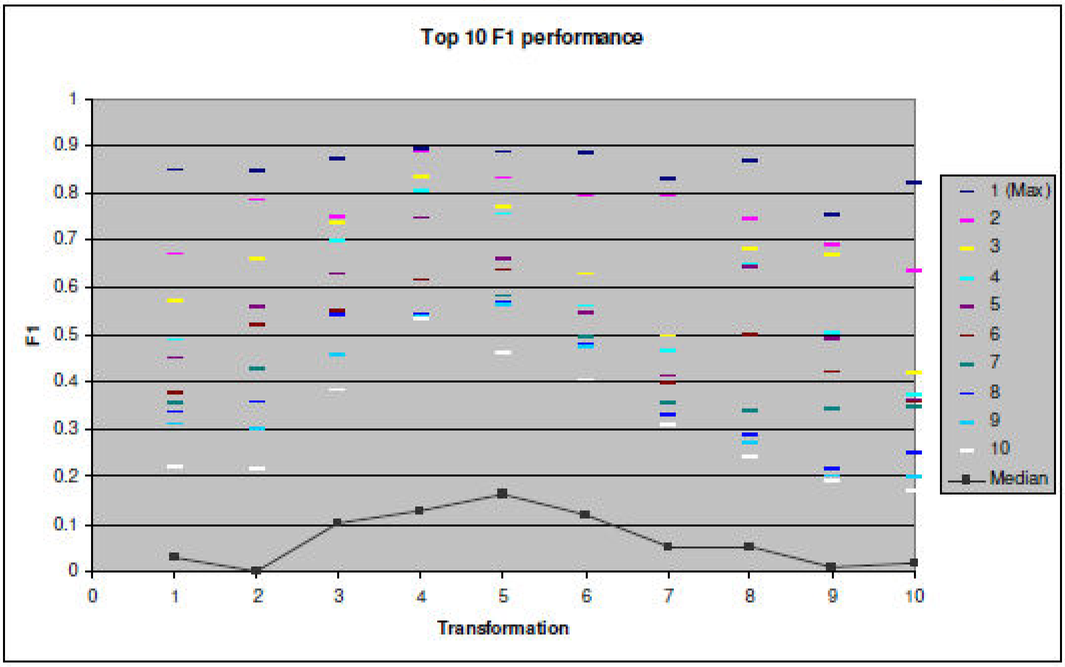
T6, T7: Decrease in quality

T8, T9: Post Production

T10: Random combination of 5

Figure 1: TRECVID 2008 CBCD: Overview of the normalized detection cost rate (NDCR) scores for the top 10 systems, per transformation and the median score of all submitted runs.

CBCD evaluation (Top 10 performance)

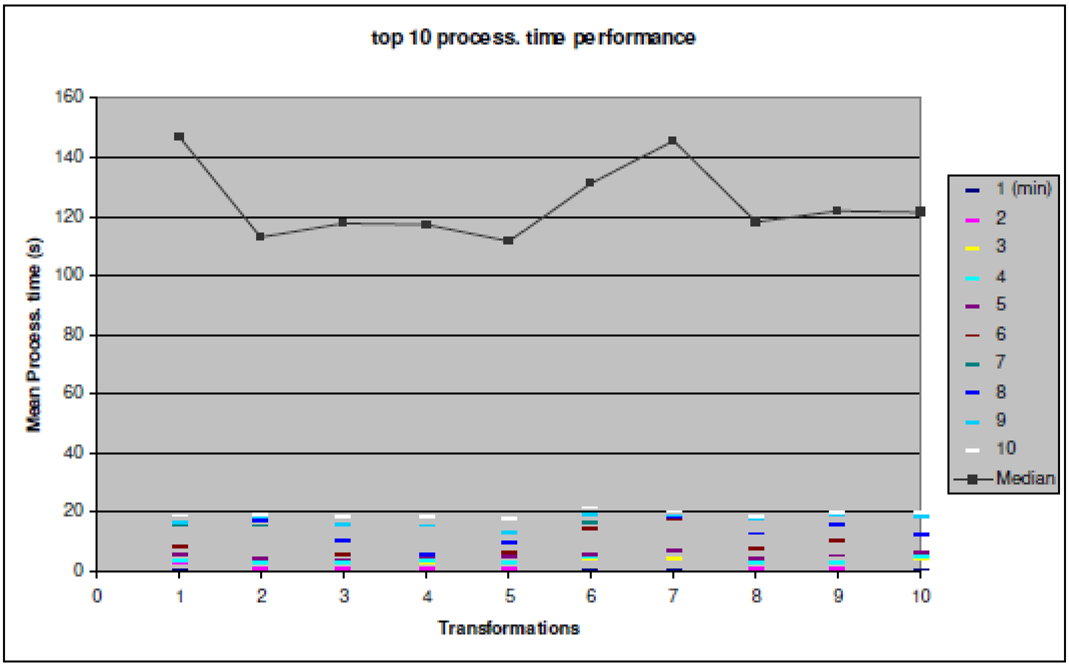


T1: Cam Cording T3: Insertion of patterns T5: Change of gamma T8, T9: Post Production

T2: Pict. In Pict. transformations T4: Re-encoding T6, T7: Decrease in quality T10: Random combination of 5

Figure 2: TRECVID 2008 CBCD: Overview of the F1 score, quantifying the localization performance, for the top 10 systems.

CBCD evaluation (Top 10 performance)



T1: Cam Cording T3: Insertion of patterns T5: Change of gamma T8, T9: Post Production

T2: Pict. In Pict. transformations T4: Re-encoding T6, T7: Decrease in quality T10: Random combination of 5

Figure 3: TRECVID 2008 CBCD: Overview of the average processing time per test clip, for the

top 10 systems.

The general impression from the discussion at the TRECVID CBCD workshop is that top systems perform accurate enough for practical use, and that the main focus for future work should be on a more realistic estimate of false negatives and measuring scalability. Best performance in the TRECVID 2008 CBCD benchmark was achieved by two groups from INRIA (LEAR and IMEDIA) and a team from France Telecom. These systems have an impressive performance. The INRIA LEAR system [DGJ09] was developed as a modified image search system, inspired by the bag-of-features work by Sivic and Zisserman [SZ03]. The best INRIA LEAR run used 2 million frames to fingerprint 200 hours of video, yielding 874 million descriptors. Scalability will be an issue for large datasets. The system from Orange [GB09] has a similar lay-out as it is also based on a frame representation based on local visual descriptors. Temporal consistency of frame matches is scored using a Markovian framework. The Orange system seems to display a favorable speed - detection quality trade off with respect to the other top systems. Best run of the INRIA Imedia team [JLB09] was based on a hybrid system using specific oriented dissociated dipoles around multi-resolution color Harris points. This representation is claimed to be more efficient and more compact.

3.6.2 TRECVID 2009 Results

The TRECVID content based copy detection task uses the normalized detection cost rate (NDCR) as evaluation measure, which is calculated from the probability of a miss, the false positive and the costs associated with misses and false positives.

In 2009, two profiles have been evaluated. The *balanced* profile assigns a cost of 1 to both the misses and false positives, the *no false alarms* profile assigns costs of 1 to misses and 1000 to false positives. Seven different video transformations have been applied to the copies to be detected. Table 1 and Table 2 show the results of the video only runs of the TRECVID 2009 content based copy detection task. The actual NDCR is calculated from a submitted threshold, the minimum NDCR is the optimal calculated value.

Table 1: Results for the balanced profile. The columns contain the min/medium NDCR over the submissions, the rows the min/median/max NDCR over the transformations.

actual NDCR	submissions		min. NDCR	submissions	
transformation	min	median	transformation	min	median
min	0,14	2,11	min	0,14	0,86
median	0,41	2,31	median	0,33	0,96
max	0,94	5,75	max	0,73	1,00

Table 2: Results for the no false alarm profile. The columns contain the min/medium NDCR over the submissions, the rows the min/median/max NDCR over the transformations.

actual NDCR	submissions		min. NDCR	submissions	
transformation	min	median	transformation	min	median
min	0,24	631,06	min	0,22	0,89
median	0,38	1.014,62	median	0,28	0,99
max	0,69	2.248,37	max	0,67	1,00

3.6.3 MediaCampaign Image/Video Fingerprinting Evaluation

In MediaCampaign we perform Fingerprinting to find out if the investigated advertisement is identical or similar to an existing advertisement for TV, Press and Internet. The analysis process works in two steps called similarity matching and exact matching. In the first step, all existing creatives are compared with the investigated spot in a fast way to find similar creatives. In the second step, the results of the similarity matching are tested against the input spot to find out if they are identical.

For the evaluation of Image and Video Fingerprinting the whole working data set is tested against the back data. In the press/internet case these makes a total amount of 2371 spots and for the TV case

641 spots. All evaluations were made automatically. Recall and precision values are computed separately for *exact matches* (advertisements which are recognized with a match value of 1) and for *similar matches* (match value < 1). In the exact matching evaluation only identical advertisements are counted as *true positives* in contrast to similarity matching, where advertisements which belong to the same company or product are used as *true positives*. For exact matches NMR has generated ground truth for the whole working data set. In false positives cases the wrongly matched images are very similar with a slightly different text or color.

Table 3: Results of the exact matches for Image and Video Fingerprinting

	# EM in ground truth	**# true positives**	**# false positives**	**Precision**
Image	1632	1597	35	0.98
Video	317	160	79	0.67

To generate the recall and precision for similar images, the first 10 result images images/videos were investigated. When no similar image/video was included in these 10 result images, it is assumed that no similar images/videos are included in the historical data set.

Table 4: Results of the similarity matches for Image and Video Fingerprinting

	Similar on pos 1	**Similar on pos > 1**	**Recall**
Image	448	51	0,89
Video	89	56	0.61

Stability and performance tests were made by analysing the whole working data multiple times. Both Fingerprinting modules have passed these tests without any crashes and with constant memory usage. Another test was made, where the same advertisement was analyzed 1000 times. In this test an identical result was produced in every iteration for Image and Video Fingerprinting.

Performance tests were made on an Intel Pentium 4, 3.00 GHz with 2 GB of RAM using Windows XP. The average analysis time needed from Image Fingerprinting is 11 seconds. The analysis time depends heavily on the fact, if an exact match is found or not because the analysis terminates when an exact match is found. In cases where no exact match is found (new creatives) the analysis can take more than 20 seconds. The average time needed to learn a new creative is about 0.5 seconds. The Video Fingerprinting needs on average 2 times real time for analysis and 5 to 10 times real time to learn advertisement into the database. This means that a 30 second video can be analyzed in 60 seconds but it takes more than 150 seconds to learn this video.

Table 5: Fingerprint evaluation summary

	Description	**Results**
Test material	All spots from the official test set were used for stability and quality evaluations.	
Stability and Performance	Stability: Stress test using working set multiple times Performance: Average analysis time on an Intel Pentium 4, 3 GHz with 2 GB of RAM	Stability: No issues for Image Fingerprinting The Video Fingerprinting crashes a few times but recovers autonomous Performance: Image fingerprint, average analysis time: 11 seconds learn time: 0.5 seconds Video fingerprint, average analysis: 2 * real time learn: 5-10 * real time

	Description	Results
Consistency	Analysis of the same ad 1000 times.	The identical results have been produced in every iteration
Recognition criteria	Precision and recall Exact matches: Results with an match value = 1 where evaluated against the ground truth Similar matches: Results with match value < 1 were evaluated to find out if similar ads are listed on top of the result list.	Image Fingerprinting Exact matches: Precision: 0.98 Similar matches: Recall: 0.89 Video Fingerprinting Exact matches: Precision: 0.65 Similar matches: Recall: 0.61

4 Audio Fingerprinting

In this chapter we discuss the state-of-the-art in audio fingerprinting in Section 4.1. In the next section, 4.2, we introduce the TNO Audio Mining Toolkit, available in our consortium. Section 4.3 lists other commercially available systems for audio fingerprinting.

4.1 State-of-the-art in Scientific Research

Audio fingerprinting techniques are more mature than video fingerprinting. Digital audio and digital audio piracy have been the main driving factors for this research, which started before the DVD became a commodity medium and before the uptake of high bandwidth internet for the masses. We will give a brief overview of techniques, following the short survey paper by Cano et al. [CBK05].

4.1.1 Fingerprint Generation

The first step in audio fingerprint generation is normalization of the signal, involving (if necessary) digitization, resampling, conversion to mono, channel simulation etc. Subsequently, the signal is divided into overlapping frames, where the frame size corresponds to the typical rate of changes in the underlying acoustic events (range of tens to hundreds of milliseconds). Some algorithms extract time domain features [LB07,LBP07], but usually the next step is transformation to the frequency domain. Various transformations have been proposed, such as discrete cosine transformation, Haar transformation, modulated complex lapped transform or Walsh-Hadamard. The most commonly applied transformation is fast Fourier (FFT).

The second step is feature extraction. A large variety of feature extraction methods have been reported. The main objective is to reduce the dimensionality and invariance to distortions. Models of the human auditory system have been an important reference for this step. Example of features for audio fingerprinting are: Mel-Frequency Cepstrum Coefficients (MFCC) [BKW99, CBM02], energy levels in different bands [KKK01], energy band differences (in time and frequency axis) [HK02], Spectral Flatness Measure (SFM) [AHH01MPL04]. Others propose an information-theoretic approach to find optimal features [BPJ03, KHS05, SB04, SJL06]. Recently, two approaches from computer-vision have found their way into the audio fingerprinting world. Some algorithms consider the frequency domain representation of a number of consecutive frames (spectrograms) as images [BC07, KHS05].

In line with this approach, pairwise boosting techniques like AdaBoost that were successful in face recognition are now used to optimize feature extraction and representation [KHS05, SJL06]. Since most of the features are just snapshots of the signal at a certain time, it is common to add higher-order features (such as the derivative and acceleration) that somehow capture the temporal variations in the feature values. In some cases, normalization and compacting operations are applied. A very coarse quantization step (e.g. binary or ternary) is usually applied in order to improve resilience against distortions. This quantization also helps to keep memory requirements low and speed up the matching process.

The last step consists of the computation of an aggregated fingerprint for an audio file (e.g. a song). There are systems that summarize all feature vectors for all frames in a single vector, however most systems generate fingerprints that scale linearly in the number of frames. Sometimes redundancies are exploited to compress the signature size, e.g. exploiting the repetitive structure of songs, or use an intermediate level representation (e.g. audio classes like phonemes used in Hidden Markov Models (HMMs) for Automatic Speech Recognition) to re-encode the audio information in a lower dimensional form. Finally, there are systems that use the time-series of features to train a model, e.g. a Gaussian Mixture Model (GMM) or a Vector Quantization (VQ) code-book; this model-based representation is then stored in the database. In summary, each step in the fingerprint generation aims at one or more of the following goals:

- **Dimensionality reduction and compact representation** Examples include feature extraction, sample rate conversions and spectral representations, e.g. PCA, OPCA and SVD.
- **Increase robustness to distortion** Examples include the use of (invariant) features, coarse quantization.

- **Emphasize unique characteristics of the signal** Examples include the use of derivatives of feature time series.
- **Capture perceptual characteristics** There are two main reasons for a fingerprinting system to consider using the perceptual characteristics and match the Human Auditory System (HAS). First, many deliberately introduced signal distortions preserve the most important perceptual characteristics. Second, some applications explicitly aim at 'perceptual similarity', or fingerprints as a perceptual digest.

Table 2 summarizes the audio fingerprinting systems reported in literature by universities, research institutes and corporate research labs. In the column 11 type the table distinguishes between systems that extract features at a constant rate (CR), summarize the feature time-series in a fixed size fingerprint (FS), and fingerprints that are based on acoustic events (VR). Most papers elaborate only on the fingerprint generation and the distance metric, and do not discuss appropriate indexing structure and matching procedures.

Reference	Feature	Representation	Distance Metric
Fraunhofer [AHH01, KAH02]	SFM (MPEG 7 scalable)	VQ codebook	Euclidean
Microsoft [BPJ03]	OPCA on spectral energy	Time Series	Euclidean
Philips[1] [HK02]	Haar on spectrogram	Time Series	Hamming
Cantmetrix [VCD04]	Statistics of spectral energy	Vector	Euclidean
Shazam [Wan03]	Location of spectral peaks	Time Series	# co-located peaks
MusicIP[2] [HH03]	SVD on spectrogram	Vector	Euclidean
Audible Magic [WBK00]	Time-averaged MFCCs	Time Series	Euclidean
Dolby Labs [RBC07]	Spectrogram projections	Time Series	Hamming
Google [BC06,BC07]	Hashed sign of Haar wavelets	Time Series	# matching bytes
Orange [LB07,LBP07]	Intervals between temporal maxima	Time Series	Dedicated score function
Relatable[3] [WR01]	Mean time and freq. features	Vector	Manhattan
Tuneprint [SB04]	OPCA on Bark energies	Time Series	Euclidean
Bogazici Univ. [ÖSM05]	SVD on MFCCs	Time Series	Euclidean
CEFRIEL [LMP04,MPL04]	Spectral energy, SFM, SCF	Time Series	Euclidean
KAIST Univ. [SJL06]	Spectral Moments	Time Series	Hamming
UPF [CBM02]	PCA of MFCC	HMM	Max Likelihood
Carnegie [KHS05]	Optimized Haar on spectrogram	Time Series	Max Likelihood
Budapest Univ. [RVK01]	Quantized Bark energies	Vector	Euclidean
Washington Univ. [SAP04]	Centroid of modulation scale	Time Series	Euclidean
Ryerson Univ. [RK06]	MFCC	GMM	Max Likelihood

Table 1: Main characteristics of several audio fingerprinting algorithms

[1] the audio fingerprinting portfolio was sold to Gracenote in 2005

[2] used by MusicBrainz since March 2006

[3] used by MusicBrainz till March 2006

4.1.2 Fingerprint Matching

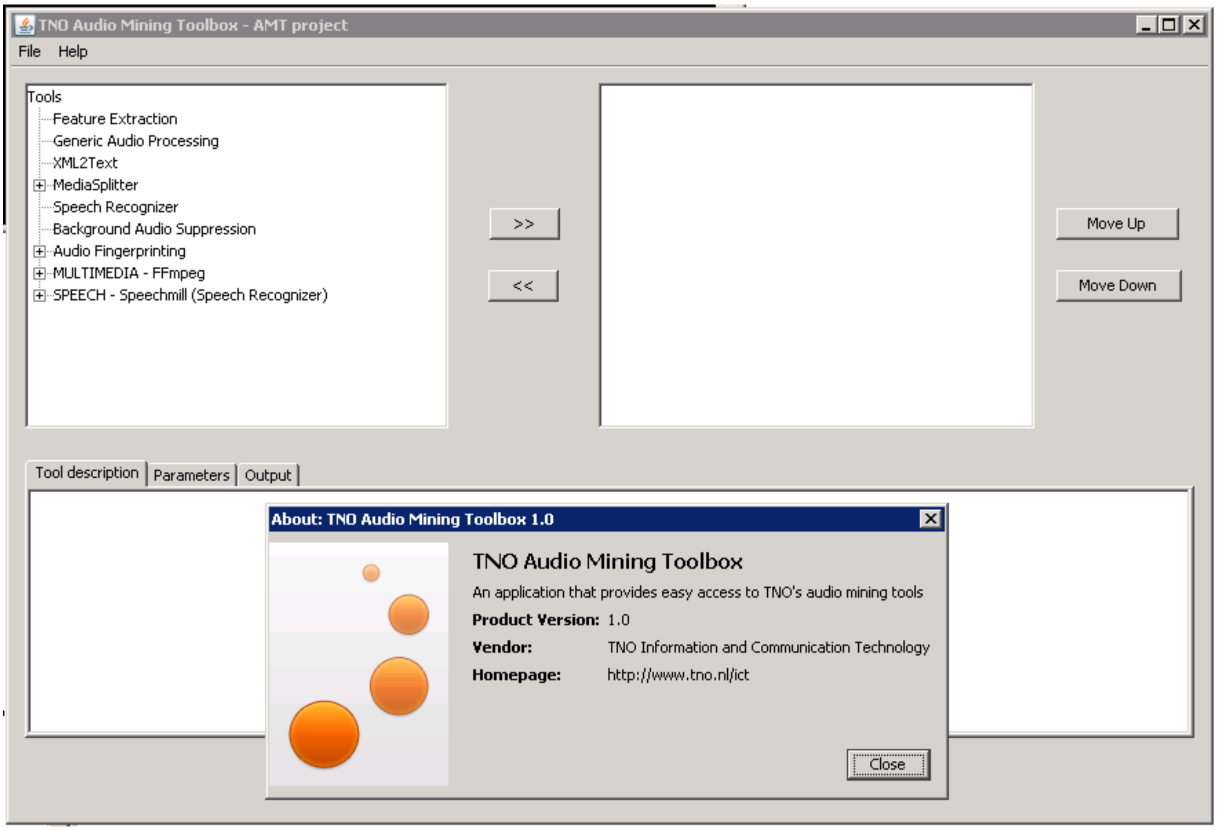
The type of distance measure used for matching fingerprints is usually determined to a large extent by the chosen fingerprint model. Common methods are Euclidean, Cross-Entropy, Manhattan or Hamming distance. In addition one can compute the distance of feature sequences to fingerprints (omitting the fingerprint model generation step for the test video). This usually applies to systems that represent the feature time-series as a model. An important issue for deployment of fingerprint matching is speed. Several methods exist to make matching faster than template matching by exhaustive search. Just like in text retrieval, indexes help to reduce the number of comparisons that have to be made to the potentially similar fingerprints. However the design of an index structure requires great care, since it is easy to increase the false negative rate. Also, the index should be designed in such a way that they are easily updatable when new fingerprints are inserted into the database.

Finally, a decision threshold has to be set which optimizes the false alarm rate and miss rate.

4.2 TNO MultimediaN Audio Mining Toolkit

Within the MultimediaN project TNO has developed an audio mining toolkit that acts like a software framework for integration and prototyping of audio mining techniques. The following audio-mining algorithms have been included:

1. Feature extraction for analyzing audio and music information retrieval
2. Audio fingerprinting for identification and search purposes (fingerprinting patented by Philips)
3. Speech recognition
4. Background audio suppression
5. Audio codec and format conversion
6. Audio segmentation



TNO Audio Mining Toolbox

An application that provides easy access to TNO's audio mining tools

Product Version: 1.0

Vendor: TNO Information and Communication Technology

Homepage: http://www.tno.nl/ict

Figure 4: TNO Audio Mining Toolkit.

The TNO Audio Mining Toolkit makes the following possible:

- to visualize the technology with a simple demo;
- to apply an algorithm on a single content item as well as for a content batch;
- to combine an algorithm with other technology in a flexible manner;
- to provide a scalable approach for use on multiple servers.

TNO Audio Mining toolkit allows one to offer rapid and concrete media mining functionality to customers in the content and media domain.

4.3 Existing commercial systems

Solutions Targeting Audio Since most video also contains audio, and audio fingerprinting technology is considered to be more mature than video fingerprinting, we devoted some attention to

- BMAT (www.bmat.com)
- AudioID (business.mufin.com)
- Audible Magic (www.audiblemagic.com)
- Melodyguard (www.melodyguard.com)
- Gracenote (www.gracenote.com)

4.3.1 Evaluation

An RIAA/IFPI initiative that spawned a lot of research and IP on audio copy detection is documented at http://www.ifpi.org/content/section_news/20010615.html.

5 Near Duplicate Detection

In this chapter we discuss the state-of-the-art in near-duplicate detection in Section 3.1. In the next section, 3.2, we describe some typical applications of near-duplicate detection.

5.1 State-of-the-art in Scientific Research

Most copy detection approaches are based on the following assumptions: (i) the actual content of the videos to be matched is identical, (ii) partial matches need to be identified and (iii) the algorithm needs to be robust against a number of distortions, such as changes of sampling parameters, noise, encoding artifacts, cropping, change of aspect ratio etc. The first assumption does not hold for near duplicates, while robustness to distortions is only necessary to a very limited degree in this application, as the content to be matched is often captured and processed under similar conditions.

The problem of matching near duplicate video segments can be transformed into a problem of matching sequences of feature vectors extracted from the video segments. The feature vectors can contain arbitrary features and can be sampled with different rate from the videos. The task is then to find suitable distance measures between sequences of these feature vectors. Two classes of approaches have been proposed for this problem.

One is based on the Dynamic Time Warping (DTW) paradigm [MR81], which tries to align the samples of the sequences so that the temporal order is kept but the distance is globally minimized. The approach has been applied to detecting repeated takes in rushes video [KS07]. The authors of [TKR99] propose a method that is conceptually very similar to DTW but includes further strict constraints, e.g. it is assumed that start and end of the two video segments are temporally aligned and only the content in between may vary in timing. The distance measure Nearest Feature Line (NFL) [ZQL00] is also conceptually related. It does not align samples of the two sequences but calculates the nearest point as the intersection of a line that is orthogonal to the line between two samples in feature space and passes through a sample of the other sequence. The distances in feature space between the intersection points and the corresponding points in the other feature sequence are summed to yield the total distance of the sequences.

The other class of distance measures is based on the idea of the edit distance between strings, i.e. the cost of inserting, deleting or replacing samples in the sequence. The authors of [ALK99] propose such a measure called vString edit distance. The values of vectors in the feature sequence are mapped to a set of discrete symbols and three new edit operations are introduced: fusion/fission of symbols (in order to deal with speed changes), swapping of symbols or blocks of symbols and insertion/deletion of shot boundaries. The distance is defined as a weighted linear combination of the traditional edit distance (using only equality or inequality of the symbols) and a modified one taking also the difference between the symbol values into account. The drawbacks are that the sequence of feature vectors needs to be mapped to a discrete set of symbols and that operations such as fission/fusion and handling of shot boundaries need to be modeled separately. The Longest Common Subsequence (LCSS) model is a variant of the edit distance, supporting gaps in the match. It has been applied to measuring the distance between trajectories in 2D space [VKG02] and it has been shown that it performs better than other methods (including DTW) for this problem [ZHT06].

An approach based on the string edit distance that avoids quantization is described in [BBN06]. The MPEG-7 ScalableColor, ColorLayout and EdgeHistogram descriptors are sued as features for the frames. The edit distance between two strings of descriptors is calculated from the descriptor similarities. When applied to news video the approach yields high precision values up to a recall rate of 0.75, for higher recall rates precision drops significantly. A similar approach is used in [DM08] and [CLG08].

An approach based on the LCSS model using quantization is proposed in [KC05]. The features used are DCT coefficients.

In [BLT09] a LCSS based approach for detection and clustering of repeated takes is proposed that does not need quantization. Sequences of feature vectors are extracted from the video and compared using appropriate distance measures (which can be different for parts of the feature vectors). The algorithm can also identify the start and end of a repeated clip and is robust to deletions and insertions. A modified single linkage hierarchical clustering algorithm is used to form clusters of takes of one scene. In the experiments MPEG-7 ColorLayout and EdgeHistogram of key frames as well as the average motion activity of short video segments are used as features. The experiments show that the algorithm reaches precision of around 85% and recall of around 70%, given a good segmentation of the parts to be matched (e.g. no take separated, not several takes merged). Under automatically produced segmentations that either under- or oversegment the precision remains unchanged, while the recall drops to 50-60%.

Another approach based on applying the string edit distance to video in presented in [YC09]. The approach is inspired by the FASTA algorithm for DNA sequence matching. Matching of frame descriptors is based on a discrete set of visual words and a vocabulary tree. The algorithm first builds matrix of matching sub-sequences (i.e. without insertions or deletions) and then joins sufficiently long matching sub-sequences. The algorithm reaches a score of 0.76 on the finding extracts (ST2) task of the MUSCLE VCD benchmark.

The approach described in [RBL09] extracts key frames at certain relative or absolute positions from each shot. The MPEG-7 ColorLayout and EdgeHistogram descriptors are used to describe the key frames globally. Agglomerative hierarchical clustering is performed on the key frames. The clustering algorithm is constrained by temporal consistency requirements. The ration of the inter- and intra distances is used as stopping criterion. Normalized Mutual Information (NMI) is used for evaluating the results, the authors report NMI around 0.85.

In [RJE08] the authors propose an approach for clustering repeated takes only based on visual features of the key frames. The temporal order is not considered in key frame matching, however, in a second step the sequences of activity values of candidate repeated takes are matched.

The authors of [BBL08] propose a footprint for video shots based on a binary 2-dimensional 30 bin histogram. The histogram is calculated from the projection of the extracted visual features of the shot into 2d space using PCA or LDA. Matching can be done with Boolean operations on the footprints.

In [GPP08] an approach for clustering repeated takes based on locality sensitive hashing is proposed. For the features extracted from each frame (e.g. HSV histograms) the feature distances to all other frames are calculated and an adaptive threshold is determined. Thresholding yields candidate matching segments. To manage the scalability problem, a variant of locality sensitive hashing is used to efficiently retrieve matching segments.

In [SRT08] the authors determine the similarity of key frames from the chi-square test between their HSV histograms. The mutual distances between the key frames are used to build a graph and the minimum spanning tree of the graph is determined. Edges exceeding a threshold are removed, the remaining connected subgraphs are interpreted as clusters containing similar shots.

5.2 Applications

An important application for near duplicate detection is content organization and summarization in audiovisual post production. In film and video production usually large amounts of raw material (“rushes”) are shot and only a small fraction of this material is used in the final edited content. The reason for shooting that amount of material is that the same scene is often taken from different camera positions and several alternative takes for each of them are recorded, partly because of mistakes of the actors or technical failures, partly to experiment with different artistic options. The action performed in each of these takes is similar, but not identical, e.g. has omissions and insertions, or object and actor positions and trajectories are slightly different.

The result of this practice in production is that users dealing with rushes have to handle large amounts of audiovisual material which makes viewing and navigation difficult. In post-production of audiovisual content, where editors need to view and organize the material in order to select the best takes to be used (the ratio between the playtime of the rushes and that of the edited content is often 30:1 or more). In general the different takes of one scene can be shot from different camera positions. Even for a very similar camera position the content between the different takes may vary, as actors and objects move differently or there are insertions and omissions in the action being performed. In addition there are takes that stop earlier (mostly due to mistakes) or start in the middle of the scene (“pickups”). The algorithm for detecting and grouping retakes have to deal with this variability.

Another application of near-duplicate detection is topic tracking of news stories as they develop over time. Some approaches work on a story level and use features such as speech transcripts [HC06] that are often not available in other content such as rushes. Other approaches work on matching sequences of key frames, tolerating gaps and insertions [DPF04]. Also the approach for tracking news stories described in [ZS05] uses among others matching of key frame sequences.

6 Video Linking

Video linking is technology that provides a means for searching video segments which contain a certain object or shot at a certain location. An application for video linking is the documentation of audiovisual archives. While edited content is well documented (at least in larger broadcast archives) and thus reusable, raw material is in most cases not documented due to the amount of material and its redundancy [DH04].

Video linking has strong historical connections with the general computer vision challenge of object recognition. In the following, we present after a short state-of-the-art section, some example applications of JRS and TNO:

- Logo recognition by JRS (Section 6.2)
- Object redetection by JRS (Section 6.3)
- Setting detection by JRS (Section 6.4)
- Object recognition by TNO (Section 6.5)

Section 6.6 lists some commercially available solutions for video linking and object recognition.

6.1 State-of-the-art in scientific research

The three main issues of video linking are:

- Finding compact, discriminative feature descriptors for objects that are invariant to changes in scene, scale, illumination, occlusion, background clutter, noise, etc.
- Rapid object matching;
- Scalable to large amounts of video data.

Recently, most scientific research was targeted at using interest points and local descriptors in large scale video retrieval systems. Most local descriptors have the desired compact and discriminative properties combined with invariance to illumination, viewpoint, occlusion, etc. Furthermore, good recognition results for objects can be achieved on a subset of detected interest points (bag-of-features) or using visual words (clusters of interest points). Fast matching and scalability is achieved by using the appropriate data-structures (hashing, KD-trees) and search strategies (approximate nearest neighbor matching).

6.2 JRS Logo Recognition

Logo recognition may be understood in different ways depending on the application: Logo recognition in the document domain, in order to decide whether the scanned document has to be further investigated [e.g. CLM00], recognition of vehicle logos [HZ07, MQW07] or detection of TV broadcaster logos in order to remove them e.g. in recordings. In this context “logo recognition” is understood as an object recognition task specialized on recognition of brands in general videos. This task is not dedicated to special types of videos (neither sports in general nor to specific types of sports). Being just a special subtask any general object recognition methods may be used, e.g. local appearance based methods like SIFT [LOW99, XQZ08], SURF [BET08] or shape matching methods [DRB09, YOU07]. But there also exist specialized algorithms [HH03, GOR03] attacking the logo recognition challenge directly by exploiting the logos’ properties, e.g. they consist mostly only of a few colors, have a high contrast and optionally a short writing. Evaluations on logo recognition level of these approaches still lack. Research in the last years has tended to use rather general object recognition algorithms instead of specialized ones. Commercial logo recognition applications known by the author are currently BrandDetector [HSA], MargauxMatrix [MMA] and RepuCom [RC]. Whereas BrandDetector is based on SIFT, MargauxMatrix is based on Magellan (http://www.omniperception.com/), and RepuCom is based on its core on SpikeNet [SN], a commercial pattern recognition library based on neural networks.

Evaluation of JRS custom SIFT implementation in terms of logo recognition has shown that precision and recall values depend on many criterions: types of distortion, logo template, video quality... Evaluation values published in scientific papers tend to better precision and recall values, because distortions are synthetic and are measured most of the time only with one type, whereas practice has shown that different types of distortions, e.g. noise and perspective distortion occur at the same time and in higher severity. A rough overview of JRS SIFT evaluation values depending on the type of distortion in the domain of logo recognition is given in the following table:

Distortion	Logo	Precision	Recall
Planar	Sky	75,87%	91,39%
	Tamoil	75,23%	92,10%
	Tamoil	24,24%	56,69%
Perspective	Sky	11,18%	17,63%
	Tamoil	11,54%	17,00%
Motion Blur	Sky	0,00%	0,00%
	Tamoil	14,06%	5,92%
	Tamoil	26,83%	4,49%
Rigid surface	Sky	51,10%	92,82%
Non-rigid surface	Sky	78,09%	76,95%
Strong Illumination changes	Sky	0,00%	0,00%

Table 2: Evaluation of JRS' SIFT implementation

6.3 JRS Object Redetection

Objects can be recognized at different levels of specificity and we distinguish between object classification and identification. Over the last decades work on automatic object recognition of both types has been done from numerous directions with varying success. Object classification is only feasible if the objects of interest and/or the specific application domain are known a priori and when a limited number of object classes is used. In spite of these restrictions state of the art object classification techniques perform poor for real-world problems. For object identification the opposite is true, as a number of solutions for object identification problems lead to satisfying results nowadays. One object identification problem is object re-detection, which aims at finding occurrences of specific objects in a single video or a collection of still images and videos. In contrast to general object identification, the samples for learning an object are taken directly from the video on which the object re-detection is performed.

Many different approaches to object recognition exist, including neural network based approaches, graph matching, genetic algorithms and fuzzy systems. The major difference between these approaches is their different representation of objects which discriminates them into the so-called model-based and appearance based approaches [SM97]. Model-based approaches use 3-dimensional models of the object shape to represent an object with geometric features such as lines, vertices and ellipses, while global or local photometric features are used for appearance based approaches. In recent years methods using local features [LOW04, MS05] have become most popular because they have solved a number of object recognition problems. With these features robustness to small perspective view changes as well as to partial occlusion is achievable and objects can be recognized anywhere in an image, with arbitrary size and rotated, without using a previous object segmentation step. Local features are usually extracted from numerous image regions around interest points [SM97] and store visual information (color or texture) of these regions in local descriptors. A typical object recognition system that works with local features performs the recognition task in the following steps: (a) First some objects of interest are learned. Therefore local descriptors are extracted from images of these objects and stored in an object database. (b) To recognize objects in a test image, again local descriptors are extracted from this image and (c) matched against the descriptors in the object database. After the best matching descriptor pairs are found, (d) an optional verification step can be performed to decide whether an object appears in the test image or not.

JRS has implemented an object recognition system, which uses a combination of SIFT and MPEG-7 color layout descriptors. The performance of this object recognition system has been evaluated on two test sets: a car and a person test set. In the car test set, a model of a car occurrence is learned from a single reference image. The model is then used to detect occurrences of the same car in other images. This is demonstrated by searching for a car in a set of 50 images. The data set contains 13 images which show the same car in different views, environments and under different lighting conditions, and 37 images showing other cars and random scenes. The person test data set contains video scenes of a person walking inside a building and random shots taken from the TRECVID test data set, which includes different persons in various scenes. The person data set contains 742 images.

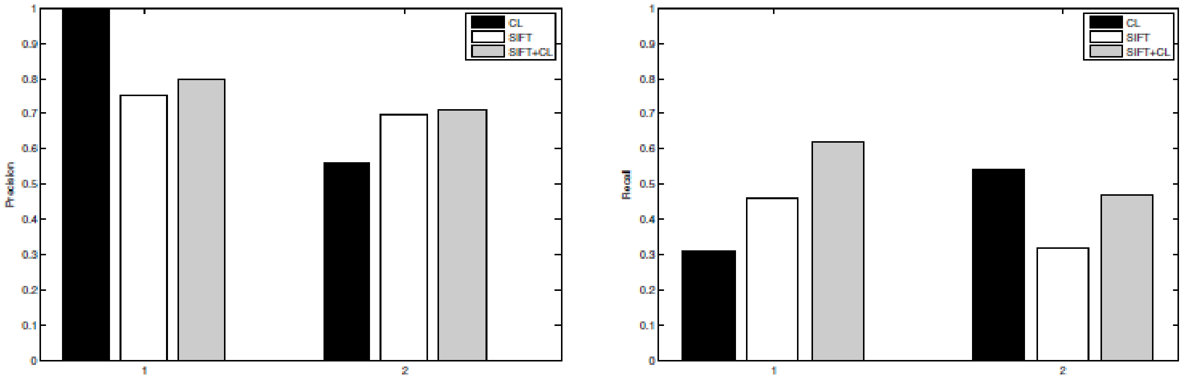


Figure 1: Precision (left) and Recall (right) of car (1) and person (2) dataset

6.4 JRS Setting Detection

While the detection of foreground objects (e.g. persons, vehicles, animals etc.) is the focus of many approaches, JRS has developed the setting detection algorithm which aims to summarize video sequences with similar camera settings by identify the similar scene background additionally.

The basic idea of setting detection proposed in [LB08] is finding spatial temporal similar regions shown in the video sequences. The setting detection algorithm extracts key frames from each video and compares the key frames, described by the region covariance descriptor [TPM06], to each other. The comparison delivers a distance matrix which describes the similarity of each pair of key frames. According to the distance matrix, the key frames are partitioned into clusters which reflect their linkage, i.e. common background scene, common foreground. We used the k-means clustering algorithm for the partition process. Another key component in our setting detection approach is the region covariance descriptor which will be described more in detail in the following.

The region covariance descriptor is used as an appearance model for an image region or image object. In contrast to other feature descriptors (e.g. SIFT, Histogram of oriented Gradients etc.), region covariance descriptor allows more freedom for the type of features. For instance the pedestrian tracking system in [TPM07] used RGB color values, pixel intensities, derivatives of the pixel intensity and edge orientations simultaneously as features. Moreover, covariance region descriptor has some very useful properties:

- It provides a natural fusion method for multiple features and modalities.
- It is invariant to mean changes, therefore robust against global illumination changes.
- It is low-dimensional descriptor, e.g. in comparison to histogram-based descriptors.
- The dimension of covariance matrix is independent on the size of region.

Based on the integral image technique, a fast estimation of the covariance matrix for any rectangular image region is available. However, a drawback of the region covariance descriptor is that the covariance matrices do not conform to the Euclidean geometry. For instance, the average of two covariance matrices cannot be computed by averaging the elements with the same index. The solution of this problem is the application of the Riemannian geometry [TPM07]. Further drawback of region covariance descriptor can be found in the distance measure between two covariance matrices. The distance function proposed in [TPM06] is based on the solution of the generalized eigenvalues problem which is rather time consuming.

The setting detection algorithm developed by JRS has been tested on 6 randomly selected TRECVID 2007 BBC-Rushes videos. The result of this test shows that the exact matching between key frame clusters and the ground truth of the video segmentation according their setting is hard to achieve. In particular the determining number of clusters is a challenge task. However, reasonable und useful clustering results can be provided and help to get an overview of the raw video.

The result of the test is shown in the following table:

Video	Length [mm:ss]	Key Frames	No. Settings Ground Truth	No. Settings Estimated	Accuracy
MRS07063	33:52	1056	7	7	92%
MRS25913	25:42	1019	7	10	66%
MRS044731	35:07	591	6	5	83%
MRS144760	27:10	739	7	8	84%
MRS157475	25:57	1137	10	5	71%
MS216210	24:13	870	7	10	71%

Table 2: Evaluation of JRS' Setting Detection algorithm

6.5 TNO MultimediaN Object Detection

One topic of ongoing research is matching of specific objects and locations in large (usually unlabelled) image collections. A tool that allows large image databases to be visually 'googled' would help the police to link crimes by matching crime scenes. For example, a photo taken inside a suspect's home could be compared to child pornography databases. The key idea behind such a tool is to automatically detect salient (prominent, distinguishing) patches in an image and describe each patch independently to the viewing angle, distance to the camera and the acquisition conditions. In this way, finding the correspondences between two photos of the same scene is facilitated regardless of the transformations relating them. A framework for generic object detection was pioneered by Schmid and Mohr and can be summarized as: detect, describe, match. The three steps will be described in more detail below.

6.5.1 Detection: Keypoints and Salient Regions

In many images there are regions which possess some distinguishing, invariant and stable property which can be detected independently with high repeatability. This property makes them a good choice for the representative image patches whose correspondence is sought. The detected salient regions should change invariantly with the transformation relating the two images. The salient patches can be detected either as groups of image pixels in the vicinity of a keypoint or directly as salient regions.

Keypoints in images that can be detected repeatedly are mostly related to corner like structures. The Harris corner detector (and variations) is a basic building block for many detectors. The Harris corner detector itself however is not scale or affine invariant. To introduce scale invariance often scale-spaces are introduced. Inspired by this, Lowe [LOW99] proposed a method for extracting keypoints which are invariant to image scaling and rotation and partly invariant to change in illumination and camera viewpoint. This approach is known as the Scale Invariant Feature Transform (SIFT) as it transforms the image data into scale-invariant coordinates relative to local features. The SIFT features are the scalespace extrema, subject to a stability criterion (for details the reader is referred to [LOW99]).

Although the SIFT algorithm performs very well it is not invariant to affine transformations. Since affine transformations appear with changes of the camera viewpoint several people have developed affine invariant detectors. A comparison of six state-of-the-art affine covariant region detectors is presented in [MS05]. For structured scenes, containing homogeneous regions with distinctive boundaries (as usually are the indoor scenes), the MSER (Maximally Stable Extremal Region) and IBR (intensity-based region) detectors perform best as they analyze the image isocontours directly.

Similarly to MSER, we proposed to analyze image isocontours by decomposing the image into binary cross-sections and computing two main types of saliency maps for each. The first type are the regions darker/brighter than their surroundings (similarly to MSER), and we propose a new type of salient regions manifested as significant irregularities on strong contrast borders. They are combined into a final map based on the stability of their support over the cross-sections.

Our detector uses morphological operators (for details the reader is referred to [RP06]), hence the name Morphology-based Stable Salient Regions (MSSR) detector. We have shown [RB06] that while the MSSR achieved comparable repeatability and matching performance to MSER and IBR, it is best in identifying perceptually salient regions. In Figure 1, bottom we have plotted all matched regions from the scene and the ones satisfying the spatial consistency constraints are shown connected with lines.

The MSSR detector has been successfully patented by TNO.

6.5.2 Region Descriptors

In a second step of a generic matching application, the detected regions are encoded using a robust (invariant to geometric and photometric modifications) descriptor, and matching between the descriptors is performed. For the case of keypoints the descriptor is computed over the neighborhood of the point, while in the case of the salient regions, the image values within the region (after normalization) is used. A popular choice is the SIFT descriptor (usually of dimension 128) computed over the normalized regions- a 3D histogram of gradient location and orientations [LOW99]. SIFT descriptors produce the best performance for different scene types, geometric and photometric transformations [MS05].

6.5.3 Matching

Descriptors are usually matched by using distance metrics (e.g. Euclidean or Mahalanobis distance) and selecting pairs with the shortest distance (nearest neighbor method). Since nearest neighbor queries in high dimensional spaces always have a worst-case quadratic running time, various approximations have been developed. Several geometric constraints can be added to further improve the matching. In Figure 9 we have plotted detected region, matches using a descriptor and the final matches after adding a spatial consistency constraint.

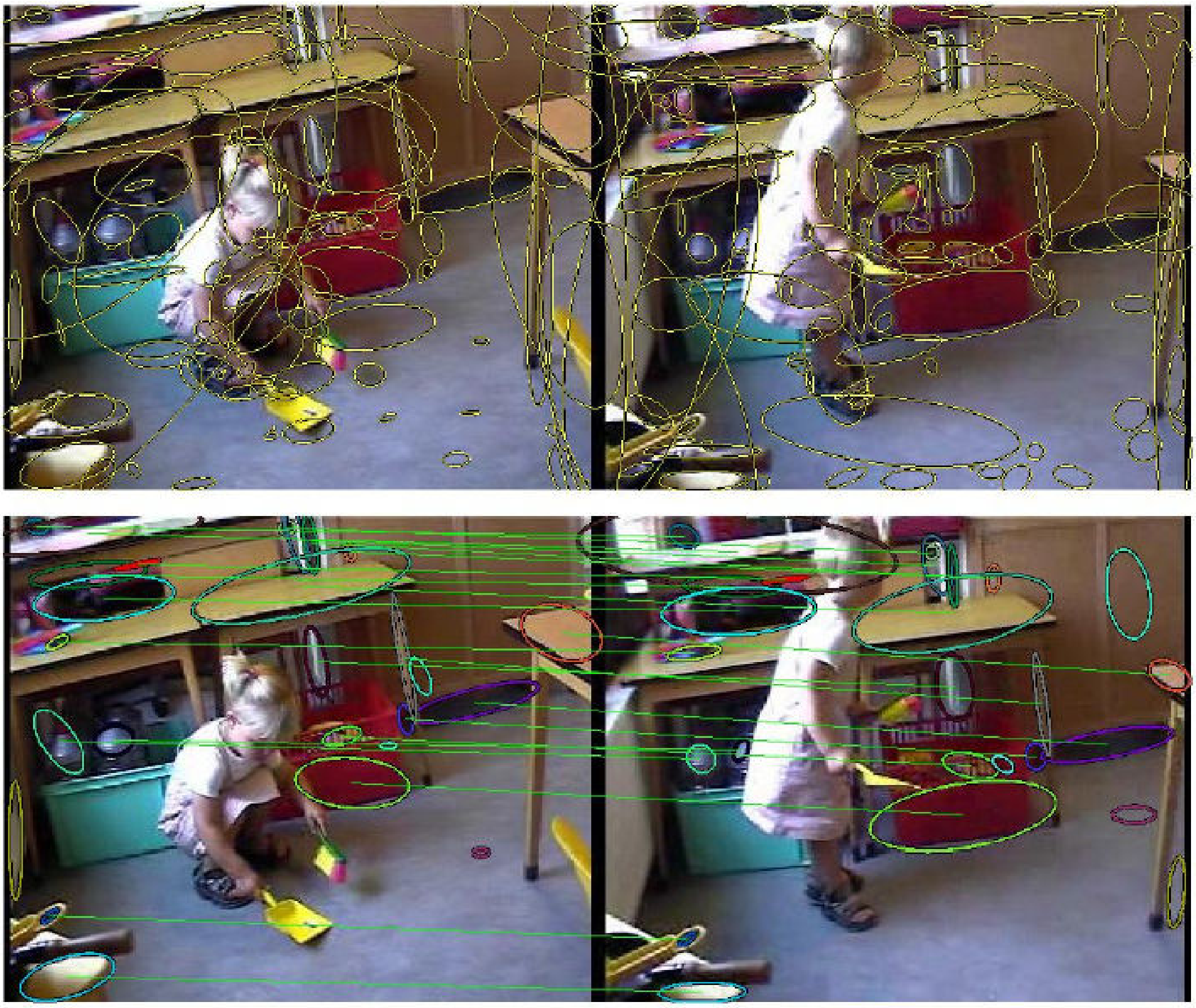


Figure 5: MSSR region detection. On top all detected regions, in the bottom corresponding regions and spatially consist matches.

6.5.4 Visual Words

Breaking down an image into an invariant set of image patches allows for applying insights from text retrieval. The image patches can be thought of as 'visual words' and a set of images can be treated as a set of 'documents' in which sets of visual words can be searched. The detected regions together with their descriptor are called visual words. Just as a normal text consists of words at specific locations, an image consists of visual words spread throughout this document. Using the analogy we can transfer search methods from text-retrieval into the visual domain.

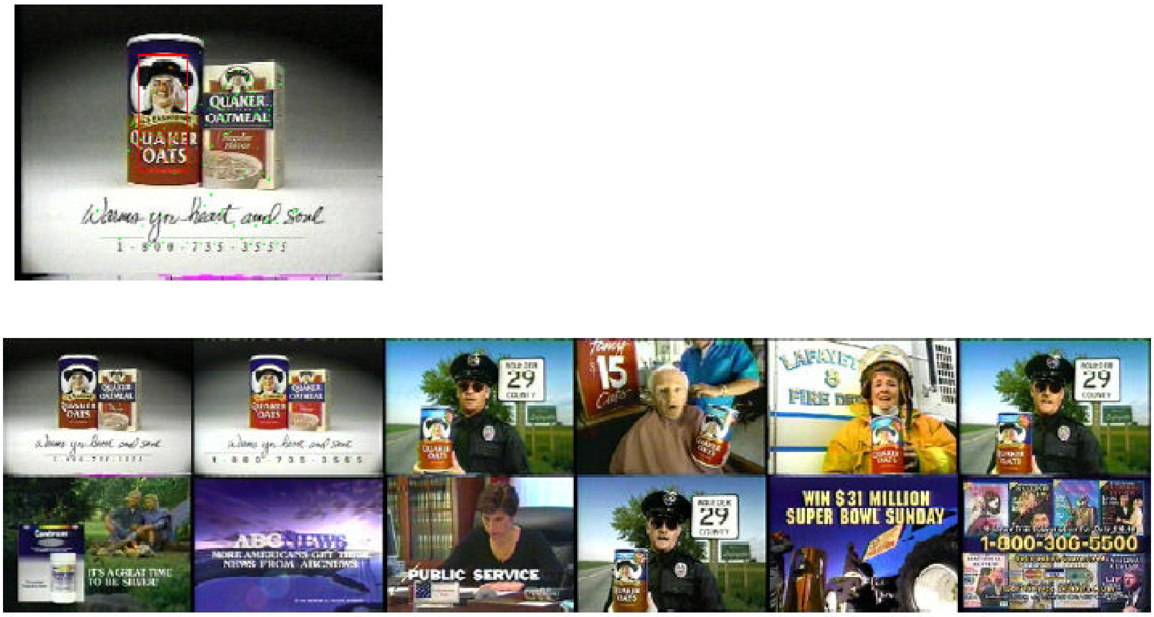

QUAKER OATS
QUAKER OATMEAL
Warms your heart and soul
1-800-735-3555
29
15
WIN $31 MILLION SUPER BOWL SUNDAY
1-800-306-5500
IT'S A GREAT TIME TO BE SILVER
ABC NEWS
PUBLIC SERVICE

Figure 6: Query by example.

In text searches several words are combined into groups. For example ‘color’, ‘color’ and ‘colors’ are all combined into the same concept. In the visual domain this corresponds to grouping salient regions with similar descriptors into clusters. Some of these clusters are not very useful for searching. For example when comparing English text the words `the’ and `and’ are usually omitted from the search. The same can be done by creating stop lists for visual words that are not very discriminative.

Using these techniques it is possible to search effectively through large image sets to locate an object of interest or a particular scene. In Figure 10 the result is shown of a query using an example image of a food box. The results contain not only the original image, but also many variations.

6.6 Existing Commercial Systems

There are several commercial and a number of academic systems that are able to search for a particular object in a large image/video databases:

- OmniPerception Magellan (http://www.omniperception.com/) is an analysis tool that provides a solution for the automatic identification, monitoring and reporting of brand and logo exposure in TV, film and other image media.
- SpikeNet (http://www.spikenet-technology.com/) has products to detect online copyright infringement, detecting logos on broadcast TV images for media analysis and do content analysis and indexing of video flows.
- Evolution Robotics, Inc. (http://www.evolution.com/core/ViPR/) develops intelligent products and solutions based on ViPR (TM), which is implementation of scale-invariant feature transformation (SIFT) algorithms.
- MILPIX (http://www.milpix.com/en/) develops state-of-art algorithms for image processing, indexation and retrieval, thanks to deep collaboration with research institute INRIA LEAR.
- Kooaba (http://www.kooaba.com/) uses the SURF algorithm on mobile phones to recognize objects in images and link the user directly to web content or other digital services.

7 Conclusions

In this document we reported on existing technologies for video fingerprinting inside and outside our consortium. Besides video fingerprinting, we also reported on adjacent technologies like audio fingerprinting, near duplicate detection and video linking. For each technology we present the state-of-the-art and gave a list of commercially available systems. Furthermore, for each item, we discussed its availability within the consortium.

For video fingerprinting we have found that within the consortium there are two systems available:

- JRS Genifer system
- ZiuZ video fingerprinting system

Both systems are going to be evaluated on Sound & Vision data in the project pilot.

For audio fingerprinting we have only one system available: a TNO implementation of an algorithm patented by Philips in the Audio Mining Toolbox. This implementation of audio fingerprinting can only be used for evaluation purposes. The toolbox contains other processing functionality that may be of interest.

For video linking the consortium members JRS and TNO have a multitude of algorithms and applications available. Most are based on SIFT-style detection and matching with some extensions and alterations depending on the application.

8 References

[AHH01] E. Allamanche, J. Herre, O. Hellmuth, B. Fr¨oba, T. Kastner, and M. Cremer. Content-based identification of audio material using mpeg-7 low level description. In 2nd International Symposium on Music Information Retrieval (ISMIR), October 2001.

[ALK99] Donald A. Adjeroh, M. C. Lee and Irwin King, “A distance measure for video sequences”, Comput. Vis. Image Underst., vol. 75, nr1-2, pp. 25-45, 1999.

[BB09] Miroslaw Bober and Paul Brasnett, “MPEG-7 Visual Signature Tools”, Proc. ICME, New York, USA, 2009, pp. 1540-1543.

[BBL08] Bredin, H., Byrne, D., Lee, H., O'Connor, N. E., and Jones, G. J., “Dublin City University at the TRECVid 2008 BBC rushes summarisation task,” In Proceedings of the 2nd ACM Trecvid Video Summarization Workshop, Vancouver, British Columbia, Canada, October, 2008.

[BBN06] M. Bertini, A. Del Bimbo, and W. Nunziati, "Video Clip Matching Using MPEG-7 Descriptors and Edit Distance", proceedings of the International Conference on Image and Video Retrieval (CIVR), Tempe, July 2006.

[BBN96] D. N. Bhat, D. N. Bhat, S. K. Nayar, and S. K. Nayar. Ordinal measures for visual correspondence. In In IEEE Conference on Computer Vision and Pattern Recognition, pages 351–357, 1996.

[BC06] S. Baluja and M. Covell. Content fingerprinting using wavelets. In 3rd European Conf. on Visual Media Production (CVMP), pp 198 - 207, November 2006.

[BC07] S. Baluja and M. Covell. Audio fingerprinting: Combining computer vision & data stream processing. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), volume 2, pages 213 – 216, April 2007.

[BET08] Herbert Bay, Andreas Ess, Tinne Tuytelaars, Luc Van Gool, "SURF: Speeded Up Robust Features", Computer Vision and Image Understanding (CVIU), Vol. 110, No. 3, pp. 346--359, 2008

[BKW99] T. Blum, D. Keislar, J. Wheaton, and E. Wold. Method and article of manufacture for content-based analysis, storage, retrieval and segmentation of audio information. U.S. Patent 5,918,223, june 1999.

[BLT09] Werner Bailer, Felix Lee and Georg Thallinger, “A Distance Measure for Repeated Takes of One Scene,” The Visual Computer, vol. 25, no. 1, pp. 53-68, Jan. 2009.

[BPJ03] C. J. C. Burges, J. C. Platt, and S. Jana. Distortion discriminant analysis for audio fingerprinting. IEEE Transactions on Speech and Audio Processing, 11(3):165 – 174, May 2003.

[BR99] R. Baeza-Yates and B. Ribeiro-Neto. Modern Information Retrieval. Addison Wesley, Harlow, 1st edition, 1999.

[CBF06] Michele Covell, Shumeet Baluja and Michael Fink, “Advertisement Detection and Replacement using Acoustic and Visual Repetition”, Proc. IEEE Workshop on Multimedia Signal Processing, pp. 461-466, Oct. 2006.

[CBK05] P. Cano, E. Batlle, T. Kalker, and J. Haitsma. A review of audio fingerprinting. Journal of VLSI Signal Processing, vol 41, iss 3, 271 - 284, November 2005.

[CBM02] P. Cano, E. Batlle, H. Mayer, and H. Neuschmied. Robust sound modeling for song detection in broadcast audio. In in Proc. AES 112th Int. Conv, 2002.

[CCM97] Shih-Fu Chang, William Chen, Horace J. Meng, Hari Sundaram and Di Zhong, “VideoQ: an automated content based video search system using visual cues,” Proceedings of the fifth ACM international conference on Multimedia, Seattle, WA, USA, pp. 313-324, 1997.

[CJ08] J. Chen and J. Jiang, “University of Bradford at TRECVID 2008: Content Based Copy Detection Task”, Proceedings of TRECVID workshop, Gaithersburg, MD, 2008.

[CLG08] Chasanis, V., Likas, A., and Galatsanos, N., “Video rushes summarization using spectral clustering and sequence alignment,” In Proceedings of the 2nd ACM Trecvid Video Summarization Workshop, Vancouver, British Columbia, Canada, October 2008.

[CLM00] Chen, Jingying; Leung, Maylor K.; Gao, Yongsheng, “New approach for logo recognition”, Proc. SPIE Vol. 4043, p. 272-279, 2000, Optical Pattern Recognition XI, David P. Casasent; Tien-Hsin Chao; Eds.

[DH04] Beth Delaney and Brigit Hoomans, “Preservation and Digitisation Plans: Overview and Analysis”, PrestoSpace Deliverable 2.1, User Requirements Final Report”, http://www.prestospace.org/project/deliverables/D2-1_User_Requirements_Final_Report.pdf, 2004.

[DGJ08] M. Douze, A. Gaidon, H. Jegou, M. Marszałek, C. Schmid, “INRIA-LEAR’s Video Copy Detection System“,Proceedings of TRECVID workshop, Gaithersburg, MD, 2008.

[DGJ09] M. Douze, A. Gaidon, H. Jefgou, M. Marszalek, and C. Schmid. INRIA-LEAR’s video copy detection system. In Proceedings of TRECVID 2008, paper to appear, 2009.

[DL09] Ina Döhring and Rainer Lienhart, “Mining TV Broadcasts for Recurring Video Sequences,” ACM International Conference on Image and Video Retrieval (CIVR 2009), July 2009.

[DM08] Dumont, E. and Mérialdo, B., ”Sequence alignment for redundancy removal in video rushes summarization,” In Proceedings of the 2nd ACM Trecvid Video Summarization Workshop, Vancouver, British Columbia, Canada, October, 2008.

[DPF04] Pinar Duygulu, Jia-Yu Pan and David A. Forsyth, “Towards auto-documentary: tracking the evolution of news stories”, Proc. 12th annual ACM international conference on Multimedia, New York, NY, USA, pp. 820-827, 2004.

[DRB09] Donoser Michael, Riemenschneider Hayko, Bischof Horst, “Efficient Partial Shape Matching of Outer Contours”, Proceedings of Asian Conference on Computer Vision (ACCV), 2009.

[GB09] N. Gengembre and S.-A. Berrani. The orange labs real time video copy detecion system - trecvid 2008 results. In Proceedings of TRECVID 2008, page to appear, 2009.

[GG08] O. Gursoy, B. Gunsel, “Istanbul Technical University at TRECVID2008”, Proceedings of TRECVID workshop, Gaithersburg, MD, 2008.

[GOR03] M. Gori et. al., “Edge-backpropagation for noisy logo recognition”, ScienceDirect, Pattern Recognition, Volume 36, Issue 1, January 2003, Pages 103-110.

[GPP08] Gorisse, D., Precioso, F., Philipp-Foliguet, S., and Cord, M., “Summarization scheme based on near-duplicate analysis,” In Proceedings of the 2nd ACM Trecvid Video Summarization Workshop, Vancouver, British Columbia, Canada, October 2008.

[GZL08] Z. Gao, Z. Zhao, T. Liu, X. Nan, M. Mei, B. Zhang, X. Liu, X. Peng, H. Zheng, Y. Zhao and A Cai, “BUPT at TRECVID 2008”, Proceedings of TRECVID workshop, Gaithersburg, MD, 2008.

[HB01] A. Hampapur and R.M. Bolle, “Comparison of distance measures for video copy detection”, Proc. IEEE International Conference on Multimedia and Expo, pp. 737-740, Aug. 2001.

[HFG08] M. Héritier, S. Foucher and L. Gagnon, “CRIM Notebook Paper - TRECVID 2008 Video Copy Detection Using Latent Aspect Modeling Over SIFT Matches”, Proceedings of TRECVID workshop, Gaithersburg, MD, 2008.

[HC06] Winston Hsu and Shih-Fu Chang, “Topic Tracking across Broadcast News Videos with Visual Duplicates and Semantic Concepts”, Proc. International Conference on Image Processing (ICIP), Atlanta, GA, USA, Oct. 2006.

[HH03a] F. Holm and W. T. Hicken. Audio fingerprinting system and method, September 2003.

[HK02] J. Haitsma and T. Kalker. A highly robust audio fingerprinting system. In 3rd International Conference on Music Information Retrieval (ISMIR), October 2002.

[HH03b] Richard J.M. den Hollander and Alan Hanjalic, “Logo Recognition in Video stills by String Matching”, ICIP 2003

[HHB02] A. Hampapur and K. Hyun and R.M. Bolle, “Comparison of sequence matching techniques for video copy detection", Proc. Storage and Retrieval for Media Databases 2002, pp. 194-201, Dec. 2002.

[HSA] http://www.hs-art.com/html/products.html

[HZ07] Humayun Karim Sulehria, Ye Zhang, “Vehicle logo recognition using mathematical morphology”, Source Proceedings of the 6th WSEAS Int. Conference on Telecommunications and Informatics, 2007.

[IIS99] P. Indyk, G. Iyengar, and N. Shivakumar. Finding pirated video sequences on the internet. Technical report, Stanford University, 1999.

[IKY06] K. Iwamoto, E. Kasutani, and A. Yamada. Image signature robust to caption superimposition for video sequence identification. In ICIP, pages 3185–3188, 2006.

[INA] Institute National pour l’Audiovisual (INA), http://www.ina.fr

[JBF07] A. Joly, O. Buisson, and C. Frelicot. Content based copy detection using distortion-based probabilistic similarity search. In IEEE Transactions on Multimedia, 2007.

[JFB04] A. Joly, C. Frelicot, and O. Buisson. Feature statistical retrieval applied to content-based copy detection. In International Conference on Image Processing, 2004.

[JLB08] Alexis Joly, Julien Law-to and Nozha Boujemaa, “INRIA-IMEDIA TRECVID 2008: Video Copy Detection”, Proceedings of TRECVID workshop, Gaithersburg, MD, 2008.

[JLB09] A. Joly, J. Law-to, and N. Boujemaa. Inria-imedia trecvid 2008: Video copy detection. In Proceedings of TRECVID 2008, paper to appear, 2009.

[KAH02] T. Kastner, E. Allamanche, J. Herre, O. Hellmuth, M. Cremer, and H. Grossmann. Mpeg-7 scalable robust audio fingerprinting. In 112th AES Convention, May 2002.

[KBG08] O. Kucuktunc, M. Bastan, U. Gudukbay, O. Ulusoy, “Bilkent University Multimedia Database Group at TRECVID 2008”, Proceedings of TRECVID workshop, Gaithersburg, MD, 2008.

[KC05] Kim, Y. and Chua, T., “Retrieval of News Video Using Video Sequence Matching,”. In Proceedings of the 11th international Multimedia Modelling Conference, 2005.

[KHS05] Y. Ke, D. Hoiem, and R. Sukthankar. Computer vision for music identification. In Computer Vision and Pattern Recognition (CVPR), pages 597– 604, June 2005.

[KKK01] A. Kimura, K. Kashino, T. Kurozumi, and H. Murase. Very quick audio searching: introducing global pruning to the time-series active search. In in Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, pages 1429–1432, 2001.

[KS07] Jim Kleban, Anindya Sarkar, Emily Moxley, Stephen Mangiat, Swapna Joshi, Thomas Kuo and B. S. Manjunath, “Feature fusion and redundancy pruning for rush video summarization,” Proc. International workshop on TRECVID video summarization, Augsburg, DE, pp. 84-88, 2007.

[KSV08] M. Koskela, M. Sjöberg, V. Viitaniemi and J. Laaksonen, “PicSOM Experiments in TRECVID 2008”, Proceedings of TRECVID workshop, Gaithersburg, MD, 2008.

[LB07] J. Lebossé and L. Brun. Audio fingerprint identification by approximate string matching. In 8th International Conference on Music Information Retrieval (ISMIR), October 2007.

[LB08] F. Lee and W. Bailer, “Organizing Rushes Video by Visually Similar Setting”, Proc. ACM International Conference on Image and Video Retrieval, p279-287, Niagara Falls, CA, 2008

[LBP07] J. Lebossé, L. Brun, and J. C. Pailles. A robust audio fingerprint’s based identification method. In 3rd Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA), volume 4477 of Lecture Notes in Computer Science (LNCS), pages 185–192, June 2007.

[LJZ05] Y. Li, J. Jin, and X. Zhou. Video matching using binary signature. In Proceedings of the 2005 International Symposium on Intelligent Signal Processing and Communications Systems (ISPAC 2005), pages 317–320, 2005.

[LKF09] Gustavo Leo, Hari Kalva and Borko Furht, "Video Identification using Video Tomography", Proc. ICME, New York, USA, 2009, pp. 1030-1033.

[LMP04] R. Lancini, F. Mapelli, and R. Pezzano. Audio content identification by using perceptual hashing. In IEEE Int. Conf. on Multimedia and Expo (ICME), pp 739 - 742, June 2004.

[LVB04] M. M. Lazarescu, S. Venkatesh, and H. H. Bui. Using multiple windows to track concept drift. Journal of Intelligent Data Analysis, vol 8, iss 1, pp 29 - 59, 2004.

[LWS08] D. Le, X. Wu, S. Satoh, S. Rajgure and J. Gemert, "National Institute of Informatics, Japan at TRECVID 2008", Proceedings of TRECVID workshop, Gaithersburg, MD, 2008.

[Lia08] Y. L. Liang et al., " THU and ICRC at TRECVID 2008", Proceedings of TRECVID workshop, Gaithersburg, MD, 2008.

[LOW04] Lowe, David G., "Distinctive image features from scale-invariant keypoints". International, Journal of Computer Vision 60(2), 91–110 (2004)

[LOW99] Lowe, David G., "Object recognition from local scale-invariant features". (1999) Proceedings of the International Conference on Computer Vision. 2. pp. 1150–1157

[LS09] Seungjae Lee and Young Ho Suh, "Video Fingerprinting Based on Orientation of Luminance Centroid", Proc. ICME, New York, USA, 2009, pp. 1386-1389.

[LTC07] J. Law-To, L. Chen, A. Joly, I. Laptev, O. Buisson, V. Gouet-Brunet, N. Boujemaa, and F. Stentiford. Video copy detection: a comparative study. In CIVR '07: Proceedings of the 6th ACM international conference on Image and video retrieval, pages 371–378, New York, NY, USA, 2007. ACM.

[MMA] http://www.margaux-matrix.com/index.php?page=whatwedo

[MPL04] F. Mapelli, R. Pezzano, and R. Lancini. Robust audio fingerprinting for song identification. In 12th European Signal Processing Conference (EUSIPCO), pages 2095 – 2098, September 2004.

[MQW07] Mei WANG, Libo QIU, Guohong WANG, "A Vehicle-logo Recognition Method Based on Wavelet Transform and Invariant Moment", Proceedings of Information Computing and Automation, 20-22 Dec 2007, China

[MR81] C. S. Myers and L. R. Rabiner, "A comparative study of several dynamic time-warping algorithms for connected word recognition", The Bell System Technical Journal, vol. 60, nr. 7, pp. 1389-1409, Sept. 1981.

[MS05] Mikolajczyk, K., Schmid, C., "A performance evaluation of local descriptors." IEEE, Transactions on Pattern Analysis and Machine Intelligence 27(10), 1615–1630, (2005).

[OAR09] P. Over, G. Awad, T. Rose, J. Fiscus, W. Kraaij, and A. F. Smeaton. Trecvid 2008 - goals, tasks, data, evaluation mechanisms and metrics. In Proceedings of TRECVID 2008, 2009. forthcoming.

[OHP08] O. Orhan, J. Hochreiter, J. Poock, Q. Chen, A. Chabra and M. Shah, " University of Central Florida at TRECVID 2008 Content Based Copy Detection and Surveillance Event Detection",Proceedings of TRECVID workshop, Gaithersburg, MD, 2008.

[OKH02] J. Oostveen, T. Kalker, and J. Haitsma. Feature extraction and a database strategy for video fingerprinting. In VISUAL, pages 117–128, 2002.

[ÖSM05] H. Özer, B. Sankur, N. Memon, and E. Anar. Perceptual audio hashing functions. EURASIP Journal on Applied Signal Processing, 1780 - 1793, 2005.

[RB09] Regunathan Radhakrishnan and Claus Bauer, "Video Fingerprinting Based on Moment Invariants Capturing Appearance and Motion", Proc. ICME, New York, USA, 2009, pp. 1532-1535.

[RBC07] R. Radhakrishnan, C. Bauer, C. Cheng, and K. Terry. Audio signature extraction based on projections of spectrograms. In IEEE Int. Conf. on Multimedia and Expo (ICME), July 2007.

[RBL09] E. Rossi, S. Benini, R. Leonardi, B. Mansencal and J. Benois-Pineau, "Clustering of scene repeats for essential rushes preview," Image Analysis for Multimedia Interactive Services, International Workshop on, pp. 234-237, 10th Workshop on Image Analysis for Multimedia Interactive Services, 2009.

[RC] http://www.repucom.net/

[RJE08] Ren, J.; Jiang, J.; Eckes, C.,"Hierarchical modeling and adaptive clustering for realtime summarization of rush videos in TRECVID'08", Proceedings 16th ACM International Conference on Multimedia, Vancouver, British Columbia, Canada, October 2008.

[RK06] A. Ramalingam and S. Krishnan. Gaussian mixture modeling of short-time fourier transform features for audio fingerprinting. IEEE Transactions on Information Forensics and Security, vol 1, iss 4, pp 457 - 463, December 2006.

[RP06] Ranguelova, E., Pauwels, E. J. "Morphology-based Stable Salient Regions Detector", IVCNZ, pp. 97-102, 2006

[RVK01] G. Richly, L. Varga, F. Kovas, and G. Hosszu. Optimised soundprint selection for identification in audio streams. IEE Proceedings - Communications, vol 148, iss 5, pp 287 - 289, October 2001.

[SAP04] S. Sukittanon, L. E. Atlas, and J. W. Pitton. Modulation-scale analysis for content identification. IEEE Transactions on Signal Processing, vol 52, iss 10, pp 3023 - 3035, October 2004.

[SB04] G. R. Schmidt and M. K. Belmonte. Scalable, content-based audio identification by multiple independent psychoacoustic matching. Journal of the Audio Engineering Society, 52(3):366 – 377, March 2004.

[SEI09] Craig Seidel, "Content Fingerprinting from an Industry Perspective", Proc. ICME, New York, USA, 2009, pp. 1524-1527.

[SJL06] J. S. Seo, M. Jin, S. Lee, D. Jang, S. Lee, and C. D. Yoo. Audio fingerprinting based on normalized spectral subband moments. IEEE Signal Processing Letters, 13(4):209 – 212, April 2006.

[SM97] Schmid, C., Mohr, R.: Local greyvalue invariants for image retrieval. IEEE Transactions, on Pattern Analysis and Machine Intelligence 19(5), 530–535 (1997)

[SN] http://www.spikenet-technology.com/products.htm

[SOK06] A. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and TRECVid. In Proceedings of the 8th ACM SIGMM International workshop on Multimedia Information Retrieval, 2006.

[SRT08] Sasongko, J., Rohr, C., and Tjondronegoro, D., "Efficient generation of pleasant video summaries," In Proceedings of the 2nd ACM Trecvid Video Summarization Workshop, Vancouver, British Columbia, Canada, October, 2008.

[SZ03] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In Proceedings of ICCV, pages 1470–1477, 2003.

[TKR99] Yap-Peng Tan, Sanjeev R. Kulkarni and Peter J. Ramadge, "A framework for measuring video similarity and its application to video query by example", Proc. International Conference on Image Processing, Kobe, JP, pp. 106-110, 1999.

[TPM06] Oncel Tuzel, Fatih Porikli, and Peter Meer, "Region covariance: A fast descriptor for detection and classification", Proc. European Conference on Computer Vision (ECCV), pp. 589-600, May 2006

[TPM07] Oncel Tuzel, Fatih Porikli, and Peter Meer, "Human Detection via Classification of Riemannian Manifolds", Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1-8, June 2007

[VCD04] V. Venkatachalam, L. Cazzanti, N. Dhillon, and M. Wells. Automatic identification of sound recordings. IEEE Signal Processing Magazine, vol 21, iss 2, pp 92 - 99, March 2004.

[VKG02] Michail Vlachos, George Kollios and Dimitrios Gunopoulos, "Discovering Similar Multidimensional Trajectories", Proc. 18th International Conference on Data Engineering, San Jose, CA, USA, pp. 673-684, 2002.

[Wan03] A. Wang. An industrial strength audio search algorithm. In 4th Int. Conf. on Music Information Retrieval (ISMIR), October 2003.

[WBK00] E. H. Wold, T. L. Blum, D. F. Keislar, and J. A. Wheaton. Method and apparatus for creating a unique audio signature, November 2000.

[WR01] S. Ward and I. Richards. A system and method for acoustic fingerprinting, March 2001.

[XQZ08] Liangfu Xia, Feihu Qi, Qianhao Zhou, “A learning-based logo recognition algorithm using SIFT and efficient correspondence matching”, Information and Automation, 2008. ICIA 2008. International Conference on

[YC09] Mei-Chen Yeh and Tim Cheng, "Video Copy Detection by Fast Sequence Matching," ACM International Conference on Image and Video Retrieval (CIVR 2009), July 2009.

[YOU07] http://www.youtube.com/watch?v=K2auR9QxNP0&feature=related

[ZEX05] Xingquan Zhu, Ahmed K. Elmagarmid, Xiangyang Xue, Lide Wu and Ann Christine Catlin, “InsightVideo: toward hierarchical video content organization for efficient browsing, summarization and retrieval,” IEEE Transactions on Multimedia, vol. 7, nr. 4, pp. 648-666, Aug. 2005.

[ZS05] Zhai, Y. and Shah, M. 2005. Tracking news stories across different sources. In Proceedings of the 13th Annual ACM international Conference on Multimedia, 2005.

[Zha08] Q. Zhang et al., “COST292 experimental framework for TRECVID2008“,Proceedings of TRECVID workshop, Gaithersburg, MD, 2008.

[ZHT06] Zhang Zhang, Kaiqi Huang and Tieniu Tan, “Comparison of Similarity Measures for Trajectory Clustering in Outdoor Surveillance Scenes”, Proc. 18th International Conference on Pattern Recognition, Washington, DC, USA, pp. 1135-1138, 2006.

[ZQL00] Li Zhao, Wei Qi, Stan Z. Li, Shi-Qiang Yang and H. J. Zhang, “Key-frame extraction and shot retrieval using nearest feature line (NFL)”, Proc. ACM workshops on Multimedia, Los Angeles, CA; USA, pp. 217-220, 2000.

9 License

THE WORK (AS DEFINED BELOW) IS PROVIDED UNDER THE TERMS OF THIS CREATIVE COMMONS PUBLIC LICENSE ("CCPL" OR "LICENSE"). THE WORK IS PROTECTED BY COPYRIGHT AND/OR OTHER APPLICABLE LAW. ANY USE OF THE WORK OTHER THAN AS AUTHORIZED UNDER THIS LICENSE OR COPYRIGHT LAW IS PROHIBITED.

BY EXERCISING ANY RIGHTS TO THE WORK PROVIDED HERE, YOU ACCEPT AND AGREE TO BE BOUND BY THE TERMS OF THIS LICENSE. TO THE EXTENT THIS LICENSE MAY BE CONSIDERED TO BE A CONTRACT, THE LICENSOR GRANTS YOU THE RIGHTS CONTAINED HERE IN CONSIDERATION OF YOUR ACCEPTANCE OF SUCH TERMS AND CONDITIONS.

1. Definitions

a. **"Adaptation"** means a work based upon the Work, or upon the Work and other pre-existing works, such as a translation, adaptation, derivative work, arrangement of music or other alterations of a literary or artistic work, or phonogram or performance and includes cinematographic adaptations or any other form in which the Work may be recast, transformed, or adapted including in any form recognizably derived from the original, except that a work that constitutes a Collection will not be considered an Adaptation for the purpose of this License. For the avoidance of doubt, where the Work is a musical work, performance or phonogram, the synchronization of the Work in timed-relation with a moving image ("synching") will be considered an Adaptation for the purpose of this License.

b. **"Collection"** means a collection of literary or artistic works, such as encyclopedias and anthologies, or performances, phonograms or broadcasts, or other works or subject matter other than works listed in Section 1(g) below, which, by reason of the selection and arrangement of their contents, constitute intellectual creations, in which the Work is included in its entirety in unmodified form along with one or more other contributions, each constituting separate and independent works in themselves, which together are assembled into a collective whole. A work that constitutes a Collection will not be considered an Adaptation (as defined above) for the purposes of this License.

c. **"Distribute"** means to make available to the public the original and copies of the Work or Adaptation, as appropriate, through sale or other transfer of ownership.

d. **"License Elements"** means the following high-level license attributes as selected by Licensor and indicated in the title of this License: Attribution, Noncommercial, ShareAlike.

e. **"Licensor"** means the individual, individuals, entity or entities that offer(s) the Work under the terms of this License.

f. **"Original Author"** means, in the case of a literary or artistic work, the individual, individuals, entity or entities who created the Work or if no individual or entity can be identified, the publisher; and in addition (i) in the case of a performance the actors, singers, musicians, dancers, and other persons who act, sing, deliver, declaim, play in, interpret or otherwise perform literary or artistic works or expressions of folklore; (ii) in the case of a phonogram the producer being the person or legal entity who first fixes the sounds of a performance or other sounds; and, (iii) in the case of broadcasts, the organization that transmits the broadcast.

g. **"Work"** means the literary and/or artistic work offered under the terms of this License including without limitation any production in the literary, scientific and artistic domain, whatever may be the mode or form of its expression including digital form, such as a book, pamphlet and other writing; a lecture, address, sermon or other work of the same nature; a dramatic or dramatico-musical work; a choreographic work or entertainment in dumb show; a musical composition with or without words; a cinematographic work to which are assimilated works expressed by a process analogous to cinematography; a work of drawing, painting, architecture, sculpture, engraving or lithography; a photographic work to which are assimilated works expressed by a process analogous to photography; a work of applied art; an illustration, map, plan, sketch or three-dimensional work relative to geography, topography, architecture or science; a performance; a broadcast; a phonogram; a compilation of data to the extent it is protected as a copyrightable work; or a work performed by a variety or circus performer to the extent it is not otherwise considered a literary or artistic work.

h. **"You"** means an individual or entity exercising rights under this License who has not previously violated the terms of this License with respect to the Work, or who has received express permission from the Licensor to exercise rights under this License despite a previous violation.

i. **"Publicly Perform"** means to perform public recitations of the Work and to communicate to the public those public recitations, by any means or process, including by wire or wireless means or public digital performances; to make available to the public Works in such a way that members of the public may access these Works from a place and at a place individually chosen by them; to perform the Work to the public by any means or process and the communication to the public of the performances of the Work, including by public digital performance; to broadcast and rebroadcast the Work by any means including signs, sounds or images.

j. **"Reproduce"** means to make copies of the Work by any means including without limitation by sound or visual recordings and the right of fixation and reproducing fixations of the Work, including storage of a protected performance or phonogram in digital form or other electronic medium.

2. Fair Dealing Rights. Nothing in this License is intended to reduce, limit, or restrict any uses free from copyright or rights arising from limitations or exceptions that are provided for in connection with the copyright protection under copyright law or other applicable laws.

3. License Grant. Subject to the terms and conditions of this License, Licensor hereby grants You a worldwide, royalty-free, non-exclusive, perpetual (for the duration of the applicable copyright) license to exercise the rights in the Work as stated below:

a. to Reproduce the Work, to incorporate the Work into one or more Collections, and to Reproduce the Work as incorporated in the Collections;

b. to create and Reproduce Adaptations provided that any such Adaptation, including any translation in any medium, takes reasonable steps to clearly label, demarcate or otherwise identify that changes were made to the original Work. For example, a translation could be marked "The original work was translated from English to Spanish," or a modification could indicate "The original work has been modified.";

c. to Distribute and Publicly Perform the Work including as incorporated in Collections; and,

d. to Distribute and Publicly Perform Adaptations.

The above rights may be exercised in all media and formats whether now known or hereafter devised. The above rights include the right to make such modifications as are technically necessary to exercise the rights in other media and formats. Subject to Section 8(f), all rights not expressly granted by Licensor are hereby reserved, including but not limited to the rights described in Section 4(e).

4. Restrictions. The license granted in Section 3 above is expressly made subject to and limited by the following restrictions:

a. You may Distribute or Publicly Perform the Work only under the terms of this License. You must include a copy of, or the Uniform Resource Identifier (URI) for, this License with every copy of the Work You Distribute or Publicly Perform. You may not offer or impose any terms on the Work that restrict the terms of this License or the ability of the recipient of the Work to exercise the rights granted to that recipient under the terms of the License. You may not sublicense the Work. You must keep intact all notices that refer to this License and to the disclaimer of warranties with every copy of the Work You Distribute or Publicly Perform. When You Distribute or Publicly Perform the Work, You may not impose any effective technological measures on the Work that restrict the ability of a recipient of the Work from You to exercise the rights granted to that recipient under the terms of the License. This Section 4(a) applies to the Work as incorporated in a Collection, but this does not require the Collection apart from the Work itself to be made subject to the terms of this License. If You create a Collection, upon notice from any Licensor You must, to the extent practicable, remove from the Collection any credit as required by Section 4(d), as requested. If You create an Adaptation, upon notice from any Licensor You must, to the extent practicable, remove from the Adaptation any credit as required by Section 4(d), as requested.

b. You may Distribute or Publicly Perform an Adaptation only under: (i) the terms of this License; (ii) a later version of this License with the same License Elements as this License; (iii) a Creative Commons jurisdiction license (either this or a later license version) that contains the same License Elements as this License (e.g., Attribution-NonCommercial-ShareAlike 3.0 US) ("Applicable License"). You must include a copy of, or the URI, for Applicable License with every copy of each Adaptation You Distribute or Publicly Perform. You may not offer or impose any terms on the Adaptation that restrict the terms of the Applicable License or the ability of the recipient of the Adaptation to exercise the rights granted to that recipient under the terms of the Applicable License. You must keep intact all notices that refer to the Applicable License and to the disclaimer of warranties with every copy of the Work as included in the Adaptation You Distribute or Publicly Perform. When You Distribute or Publicly Perform the Adaptation, You may not impose any effective technological measures on the Adaptation that restrict the ability of a recipient of the Adaptation from You to exercise the rights granted to that recipient under the terms of the Applicable License. This Section 4(b) applies to the Adaptation as incorporated in a Collection, but this does not require the Collection apart from the Adaptation itself to be made subject to the terms of the Applicable License.

c. You may not exercise any of the rights granted to You in Section 3 above in any manner that is primarily intended for or directed toward commercial advantage or private monetary compensation. The exchange of the Work for other copyrighted works by means of digital file-sharing or otherwise shall not be considered to be intended for or directed toward commercial advantage or private monetary compensation, provided there is no payment of any monetary compensation in con-nection with the exchange of copyrighted works.

d. If You Distribute, or Publicly Perform the Work or any Adaptations or Collections, You must, unless a request has been made pursuant to Section 4(a), keep intact all copyright notices for the Work and provide, reasonable to the medium or means You are utilizing: (i) the name of the Original Author (or pseudonym, if applicable) if supplied, and/or if the Original Author and/or Licensor designate another party or parties (e.g., a sponsor institute, publishing entity, journal) for attribution ("Attribution Parties") in Licensor's copyright notice, terms of service or by other reasonable means, the name of such party or parties; (ii) the title of the Work if supplied; (iii) to the extent reasonably practicable, the URI, if any, that Licensor specifies to be associated with the Work, unless such URI does not refer to the copyright notice or licensing information for the Work; and, (iv) consistent with Section 3(b), in the case of an Adaptation, a credit identifying the use of the Work in the Adaptation (e.g., "French translation of the Work by Original Author," or "Screenplay based on original Work by Original Author"). The credit required by this Section 4(d) may be implemented in any reasonable manner; provided, however, that in the case of a Adaptation or Collection, at a minimum such credit will appear, if a credit for all contributing authors of the Adaptation or Collection appears, then as part of these credits and in a manner at least as prominent as the credits for the other contributing authors. For the avoidance of doubt, You may only use the credit required by this Section for the purpose of attribution in the manner set out above and, by exercising Your rights under this License, You may not implicitly or explicitly assert or imply any connection with, sponsorship or endorsement by the Original Author, Licensor and/or Attribution Parties, as appropriate, of You or Your use of the Work, without the separate, express prior written permission of the Original Author, Licensor and/or Attribution Parties.

e. For the avoidance of doubt:

   i. **Non-waivable Compulsory License Schemes**. In those jurisdictions in which the right to collect royalties through any statutory or compulsory licensing scheme cannot be waived, the Licensor reserves the exclusive right to collect such royalties for any exercise by You of the rights granted under this License;

   ii. **Waivable Compulsory License Schemes**. In those jurisdictions in which the right to collect royalties through any statutory or compulsory licensing scheme can be waived, the Licensor reserves the exclusive right to collect such royalties for any exercise by You of the rights granted under this License if Your exercise of such rights is for a purpose or use which is otherwise than noncommercial as permitted under Section 4(c) and otherwise waives the right to collect royalties through any statutory or compulsory licensing scheme; and,

   iii. **Voluntary License Schemes**. The Licensor reserves the right to collect royalties, whether individually or, in the event that the Licensor is a member of a collecting society that administers voluntary licensing schemes, via that society, from any exercise by You of the rights granted under this License that is for a purpose or use which is otherwise than noncommercial as permitted under Section 4(c).

f. Except as otherwise agreed in writing by the Licensor or as may be otherwise permitted by applicable law, if You Reproduce, Distribute or Publicly Perform the Work either by itself or as part of any Adaptations or Collections, You must not distort, mutilate, modify or take other derogatory action in relation to the Work which would be prejudicial to the Original Author's honor or reputation. Licensor agrees that in those jurisdictions (e.g. Japan), in which any exercise of the right granted in Section 3(b) of this License (the right to make Adaptations) would be deemed to be a distortion, mutilation, modification or other derogatory action prejudicial to the Original Author's honor and reputation, the Licensor will waive or not assert, as appropriate, this Section, to the fullest extent permitted by the applicable national law, to enable You to reasonably exercise Your right under Section 3(b) of this License (right to make Adaptations) but not otherwise.

5. Representations, Warranties and Disclaimer

UNLESS OTHERWISE MUTUALLY AGREED TO BY THE PARTIES IN WRITING AND TO THE FULLEST EXTENT PERMITTED BY APPLICABLE LAW, LICENSOR OFFERS THE WORK AS-IS AND MAKES NO REPRESENTATIONS OR WARRANTIES OF ANY KIND CONCERNING THE WORK, EXPRESS, IMPLIED, STATUTORY OR OTHERWISE, INCLUDING, WITHOUT LIMITATION, WARRANTIES OF TITLE, MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE, NONINFRINGEMENT, OR THE ABSENCE OF LATENT OR OTHER DEFECTS, ACCURACY, OR THE PRESENCE OF ABSENCE OF ERRORS, WHETHER OR NOT DISCOVERABLE. SOME JURISDICTIONS DO NOT ALLOW THE EXCLUSION OF IMPLIED WARRANTIES, SO THIS EXCLUSION MAY NOT APPLY TO YOU.

6. Limitation on Liability. EXCEPT TO THE EXTENT REQUIRED BY APPLICABLE LAW, IN NO EVENT WILL LICENSOR BE LIABLE TO YOU ON ANY LEGAL THEORY FOR ANY SPECIAL, INCIDENTAL, CONSEQUENTIAL, PUNITIVE OR EXEMPLARY DAMAGES ARISING OUT OF THIS LICENSE OR THE USE OF THE WORK, EVEN IF LICENSOR HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

7. Termination

a. This License and the rights granted hereunder will terminate automatically upon any breach by You of the terms of this License. Individuals or entities who have received Adaptations or Collections from You under this License, however, will not have their licenses terminated provided such individuals or entities remain in full compliance with those licenses. Sections 1, 2, 5, 6, 7, and 8 will survive any termination of this License.

b. Subject to the above terms and conditions, the license granted here is perpetual (for the duration of the applicable copyright in the Work). Notwithstanding the above, Licensor reserves the right to release the Work under different license terms or to stop distributing the Work at any time; provided, however that any such election will not serve to withdraw this License (or any other license that has been, or is required to be, granted under the terms of this License), and this License will continue in full force and effect unless terminated as stated above.

8. Miscellaneous

a. Each time You Distribute or Publicly Perform the Work or a Collection, the Licensor offers to the recipient a license to the Work on the same terms and conditions as the license granted to You under this License.

b. Each time You Distribute or Publicly Perform an Adaptation, Licensor offers to the recipient a license to the original Work on the same terms and conditions as the license granted to You under this License.

c. If any provision of this License is invalid or unenforceable under applicable law, it shall not affect the validity or enforceability of the remainder of the terms of this License, and without further action by the parties to this agreement, such provision shall be reformed to the minimum extent necessary to make such provision valid and enforceable.

d. No term or provision of this License shall be deemed waived and no breach consented to unless such waiver or consent shall be in writing and signed by the party to be charged with such waiver or consent.

e. This License constitutes the entire agreement between the parties with respect to the Work licensed here. There are no understandings, agreements or representations with respect to the Work not specified here. Licensor shall not be bound by any additional provisions that may appear in any communication from You. This License may not be modified without the mutual written agreement of the Licensor and You.

f. The rights granted under, and the subject matter referenced, in this License were drafted utilizing the terminology of the Berne Convention for the

Protection of Literary and Artistic Works (as amended on September 28, 1979), the Rome Convention of 1961, the WIPO Copyright Treaty of 1996, the WIPO Performances and Phonograms Treaty of 1996 and the Universal Copyright Convention (as revised on July 24, 1971). These rights and subject matter take effect in the relevant jurisdiction in which the License terms are sought to be enforced according to the corresponding provisions of the implementation of those treaty provisions in the applicable national law. If the standard suite of rights granted under applicable copyright law includes additional rights not granted under this License, such additional rights are deemed to be included in the License; this License is not intended to restrict the license of any rights under applicable law.