**Projet ANR- 11-IS02-001**

# MEX-CULTURE / Multimedia libraries indexing for the preservation and dissemination of the Mexican Culture

# Deliverable

# MID-term report on summarization and scalable search

Programme Blanc International II- 2011 Edition

# A IDENTIFICATION

| | |
|---|---|
| Project acronym | MEX-CULTURE |
| Project title | Multimedia libraries indexing for the preservation and dissemination of the Mexican Culture |
| Coordinator of the French part of the project (company/organization) | Centre d'Etude et de Recherche en Informatique et Communications – Conservatoire National des Arts et Métiers |
| Coordinator of the Mexican part of the project (company/organization) | Instituto Politécnico Nacional - Centro de Investigación y Desarrollo de Tecnología Digital |
| Project coordinator (if applicable) | Michel Crucianu |
| Project start date<br>Project end date | 01/01/2012*<br>31/12/2014 |
| Competitiveness cluster labels and contacts (cluster, name and e-mail of contact) | Cap Digital Paris-Région<br>Philippe Roy<br>Philippe.Roy@capdigital.com |
| Project website if applicable | http://mexculture.cnam.fr |

** The Mexican partners are only financed since November 2012.*

| *Coordinator of this report* | |
|---|---|
| *Title, first name, surname* | *Monsieur Henri Nicolas* |
| *Telephone* | *05 40 00 35 20* |
| *E-mail* | *nicolas@labri.fr* |
| *Date of writing* | *29/09/2013* |

| | |
|---|---|
| Rédacteurs : | Gabriel Sargent (LABRI),<br>Henri Nicolas(LABRI)<br>Andrei Stoian (CEDRIC-Cnam),<br>Michel Crucianu(CEDRIC-Cnam)<br>Karina Perez-Daniel (LABRI)<br>Sofian Maabout(LABRI)<br>Jenny Benois-Pineau(LABRI)<br>Mireya García-Vázquez (CITEDI-IPN)<br>Alejandro Ramírez-Acosta (CITEDI-IPN) |

# B INTRODUCTION

## B.1 MOTIVATION

Cultural heritage content has a significant importance in promoting diversity in a globalized world, so making this content easily available to a broad audience is a critical issue. Large volumes of such content must be indexed and users must be provided with means for a fast and easy access to the multimedia information, including browsing according to multiple criteria and visualization of summaries or of detailed content stored in the archives. Given the size of the databases envolved, such operations must rely on *automatic* content-based indexing, as well as on scalable summarization and content-based retrieval.

Summarization of AV (Audio/Video) content is motivated by the need to reduce information overload on the human consumer, to provide him access to important information in the AV content, or to enable the viewer to consume more content in a given time. AV summaries are useful in many applications for entertainment, news access and educational purposes. But summaries are not the only method for providing access to multimedia content. Smilarities according to various "dimensions" can be directly employed for searching large multimedia databases starting from already selected content. Both multimedia summarization and content-based retrieval require both effectiveness and *efficiency* in dealing with very large amounts of content. The strong requirements in terms of scale and response time of interactive multimedia-related applications, together with the high-dimensional and often complex descriptions that characterize multimedia content, make this scalability problem very challenging.

## B.2 AUDIO-VISUAL CORPORA

The methods devised in the project will be applied to the large databases of the FONOTECA NACIONAL (National Sound Archive of Mexico), part of CONACULTA (National Council for Culture and the Arts of Mexico) and the Video library (TVUNAM, more than 100,000 hours of video) of the UNAM (National Autonomous University of Mexico). Since these databases will only become progressively available during the project, a large video database provided by INA (sub-contractor of Cnam) will be employed for the early evaluation of the methods devised in this project.

## B.3 GENERAL FRAMEWORK FOR SUMMARIZATION AND SCALABLE RETRIEVAL

One of our goals in this project is to devise and evaluate new methods for the scalable retrieval of multimedia content following mainly two paradigms: query-by-example (QBE, where the query is a data item and the system has to retrieve the database items that are most similar to the query) and relevance feedback (RF, where the retrieval session is composed of successive round during which the user tells the system whether the returned results are satisfactory or not). These paradigms can only reach their full potential if they accept *partial* queries (regarding specific *temporal segments* and/or specific *areas* of video

frames) and scale to large databases. The scalability challenge is significantly reinforced by the use of partial queries. While video content-based retrieval typically relies on compact, global descriptions of entire video sequences, we intend to consider more detailed, *sequence*-based video matching in order to obtain more effective results. This, however, requires significant developments in *multidimensional* sequence matching. Effective RF solutions typically use kernel methods with active learning, but the efficiency of existing proposals for the scalability of RF is quite limited. Since kernel functions can often be seen as similarity measures, in order to support fast active learning in large databases we intend to devise LSH-inspired solutions for the efficient identification of the most informative unlabeled samples; note that in such cases the "query" is not an item in the database but rather a decision boundary.

# C STATE-OF-THE-ART IN SUMMARIZATION AND SCALABLE SEARCH

This section presents a brief overview of the state of the art in the domain of summarization and scalable search and of cross-media description of audio-visual content.

## C.1 CROSS-MEDIA DESCRIPTION OF AUDIO-VISUAL CONTENT

We consider here the problem of segmentation and summary of documentaries using audio-visual features. As mentioned in [C.1.1] the structural segmentation of documentaries this is a difficult issue, as the structure of such videos doesn't follow fixed production rules which usually vary according to the producers, and are almost never explicitely defined.

We therefore restrict the definition of the structural segments to connex portions of the video which displays a continuity in terms of audio (eg. auditive environment) and video (similar visual environment).

Generic approaches have been proposed for video segmentation using audio and video in the scope of scene detection [C.1.2]. They mainly consist in a two-step process : shot segmentation (from visual analysis) then scene segmentation obtained by the fusion of contiguous shots which share similar audio-visual properties. The first step is performed using visual features only. The second step uses a wide variety of audio and video features (eg [C.1.3]). This step is mainly achieved through similarity criteria analysis [C.1.3], clustering techniques [C.1.4], support vector machines [C.1.5], scene transition graphs [C.1.6]. As the evaluation databases vary from one work to another, it is hard to compare their performances.

The segmentation approach based on audio-visual descriptors proposed in the project is described in Section D.3.

## C.2 SUMMARIZATION

Nowadays a large number of video information is available and the amount of videos has shown an exponential increase, which implies a lot of resources to store them and a lot of time to determine their content. Video summarization gives an overview of the video content and allows us to have a fast understanding of it and it helps the user to navigate and retrieve through a large amount of videos. However, the diversity in length and content make of automatic video summarization a challenging task. Recently have been proposed several video summarization approaches such as [C.2.1-7], where the video summary is given in one dimension, this is, one single video summary per each video.

The main contribution of this work is the idea of using *data cube* approach to present a scalable multidimensional video summary. Such *scalability* is given in terms of the ability of navigate into the data cube, such that, when a keyframe of the video summary is selected by the user, a new summary of that keyframe, in the next level of the data cube, is displayed. This scheme enables the selection of a new keyframe of the new summary to navigate again into the data cube in several levels of detail.

The essential characteristic of data cubes is their ability to deal with very large data bases. Among the techniques to handle very large amount of information, data cube gives a multidimensional representation of a database which allows a quick navigation into large datasets by changing its visual representation in order to present the data at different levels of aggregation according to the user requirements. Nevertheless, despite its proven effectiveness in other fields such as the multidimensional traffic GPS data quality [C.2.8], text cube architecture to organize social media information [C.2.9] and density analysis of geoinformation [C.2.10]; as well as taking into account the important advances in its performance by reducing the size of the data cube [C.2.11], the design of efficient cube computation with identification of interesting cube groups [C.2.12] and by implementing the parallel cube computation with distributive memory [C.2.13] among others, the data cube model had not been extended to visual tasks until 2010, when Jin *et al* [C.2.14] introduce this concept to computer vision field where the dimensions where considered as meta information (date, gps, title, etc) as well as image visual features. However the visual summarization technique based on data cubes is still an open challenge to be tackled. For that reason we propose to develop a system for summarizing video documentaries which supports a quick navigation between the different levels of detail in the cube in order to display a single summary according to the selected query.

The approach which we develop in the context of video summarization based on data cube approach is described in Section D.4.

## C.3 SCALABLE RETRIEVAL

Multimedia content-based retrieval require both effectiveness and efficiency in dealing with very large amounts of content. To make retrieval of multidimensional data sublinear in the size of the database, many index-based access methods were put forward in the database community [C.3.18]; some of them were adapted to multimedia content. The strong requirements in terms of scale and response time of interactive multimedia-related applications, together with the high-dimensional and often complex descriptions that

characterize multimedia content, make this scalability problem very challenging. While most of the early proposals concerned exact retrieval, significant progress was later made on *approximate* retrieval, following results showing that approximations can strongly improve efficiency while retaining high effectiveness. Most of the proposals dealing with the largest amounts of multimedia data rely on approximate retrieval (see [C.3.9], [C.3.15], [C.3.23]). Locality Sensitive Hashing (LSH), introduced in [C.3.7] (see also [C.3.1], [C.3.5]), is a general principle for devising approximate similarity-based retrieval solutions. LSH-inspired methods were successfully developed for multimedia content including audio (e.g. [C.3.21], [C.3.22]), images (e.g. [C.3.8], [C.3.2]) and video (e.g. [C.3.15], [C.3.16]). Later developments of LSH methods for multimedia data focused on improving the balance between effectiveness and efficiency, e.g. multi-probe LSH [C.3.11], *a posteriori* multi-probe LSH [C.3.8], spectral hashing [C.3.20] or multi-level LSH [C.3.21]. Other developments concerned the extension of LSH-based methods to more complex data, like geometric configurations of local features [C.3.2], [C.3.17]. There is comparatively little work on the scalability of relevance feedback (RF, another important paradigm in multimedia retrieval, see e.g. [C.3.3]). We can mention specific solutions like [C.3.12], [C.3.13], [C.3.14], [C.3.4] for SVM-based RF or a solution [C.3.10] employing boosting and *a posteriori* multi-probe LSH. The efficiency of such solutions has to be further improved. Since kernel functions are often similarity measures, it is possible to develop LSH-inspired solutions for scaling up retrieval with RF based on effective kernel methods [C.3.6].

In the current state-of-the-art, scalable retrieval is not connected to the scalability of media description. But we see two connections between these two concepts of scalability. Consider a scalable, hierarchical representation of content such as wavelets or Gaussian pyramids. First, interactive retrieval with relevance feedback can be made more scalable by iteratively increasing, during consecutive relevance feedback loops, the resolution on which content descriptors have been extracted. Second, we believe that a hierarchical representation of content allows to design new multi-level LSH (e.g. [C.3.21]) solutions that also support scalable retrieval. Research in scalable indexing of video data with Daubechies wavelet pyramids has shown that such approaches "scale" in the sense of content description.

While video content-based retrieval typically relies on compact, global descriptions of entire video sequences, one can expect that more detailed, *sequence*-based video matching should allow to obtain more effective results. This is currently avoided because the cost of sequence matching that is usually considered to be prohibitive. But recent work like [C.3.18] shows that (one-dimensional) sequence matching with Dynamic Time Warping (DTW) can be performed efficiently. However, an efficient comparison of video sequences with DTW (or related measures) requires significant developments in *multidimensional* sequence matching and our current work addresses this issue.

# D PROPOSED METHODS FOR SUMMARIZATION AND SCALABLE SEARCH

## D.1 GENERAL ARCHITECTURE OF THE PROJECT

Indexing and retrieval of multimedia information is based on architecture that we called architecture for scalable search. This architecture will allow to store and organize multimedia

information in a scalable manner, allowing a multimedia resource search in large databases scale easily and quickly.
The feature of this architecture is that the indexing and retrieval of information is performed in the compressed domain of multimedia information [D.1.1, D.1.2]. Figure 1 shows the proposed architecture.

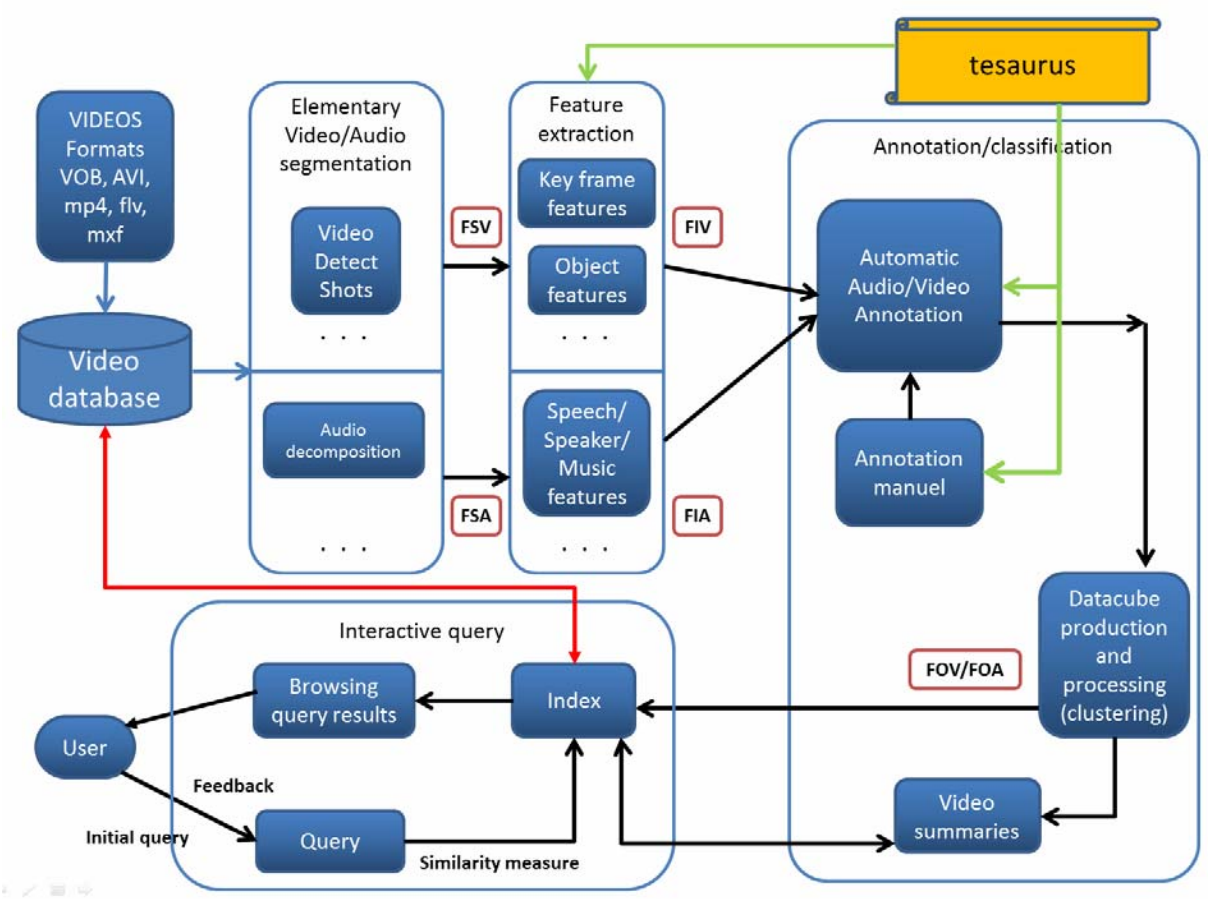


Figure 1: Architecture for content-based audiovisual scalable indexing and retrieval.

In Figure 2 we can see that the decomposition tasks are reflected in this scalable search architecture. It shows then the interrelationship between the different activities of each partner:

**Task 1:** Scalable description of visual encoded content. Extract effective local and global spatio-temporal descriptors from JPEG2000 compressed flow.

**Task 2:** Description of speech/audio content. Speech/audio signal segmentation, description and classification of sound events, Mexican native language speech recognition.

**Task 3:** Audiovisual summaries and scalable retrieval. Develop scalable methods for structuring the audiovisual database and for interactive content-based multimodal retrieval.

**Task 4:** Software development of multi-modal algorithms. Make the databases available for the project and develop the software platform.

Tasks 1, 2 and 3 all develop software components (for extracting content descriptions, for performing content summarization and for scalable search). All software integration activities were grouped in Task 4.

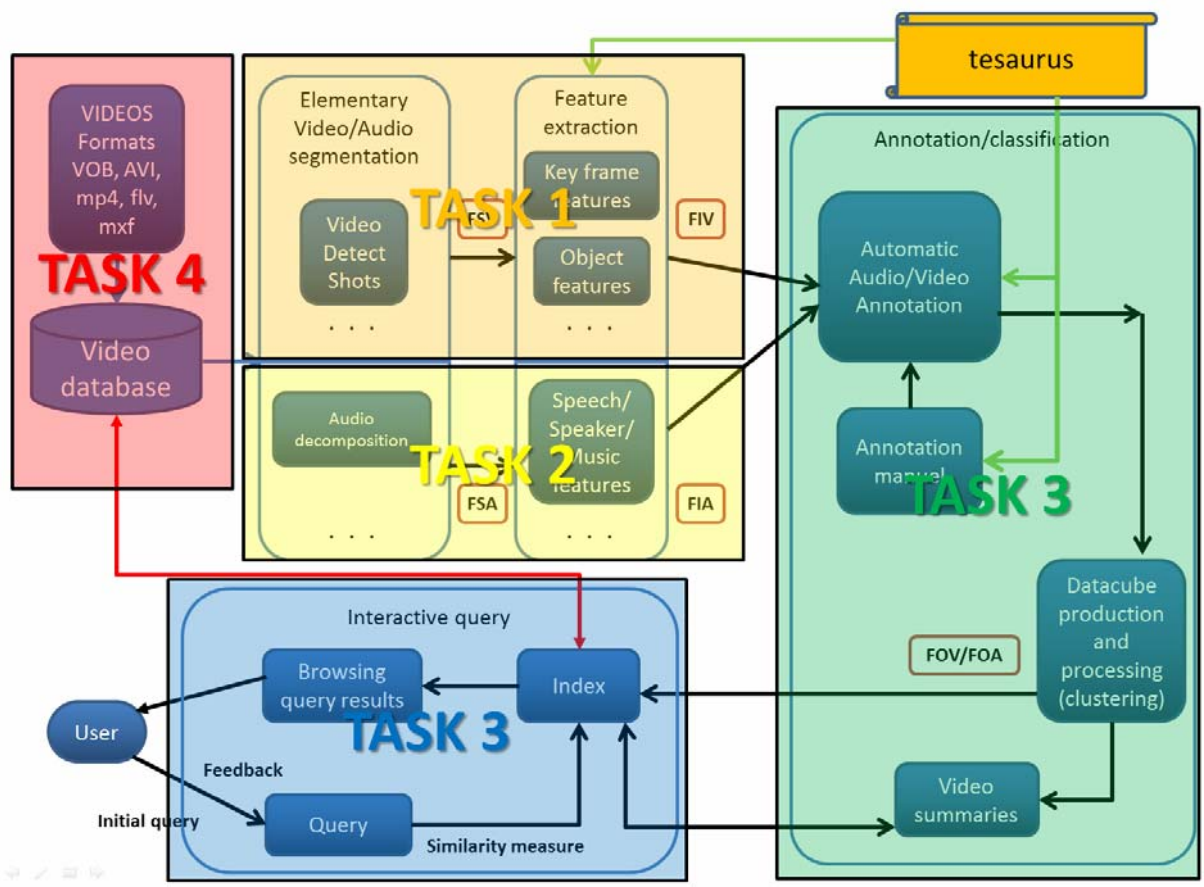



Figure 2: Task in the architecture for content-based audiovisual scalable indexing and retrieval.

The architecture for content-based audiovisual scalable indexing and retrieval includes the following:

a) **Video database**: the information of audiovisual content is in the compressed domain (audio format: MP4; video format: JPEG2000).
b) **Elementary video/audio segmentation**: the audiovisual content structure analysis aims at segmenting an audiovisual content into a number of structural elements that have semantic contents.
c) **Feature extraction**: the extraction of features of the audiovisual content in the structural elements, represents the base for scalable indexing and retrieval.
   - For visual encoded content we consider a scalable method of signal representation such as a spatio-temporal motion compensated wavelet-based solution. The objects to be extracted are sets of connected components that are homogeneous in color and/or in texture with respect to a given criterion and that

have a motion different from the global motion of the scene [D.1.3, D.1.4, D.1.5, D.1.6].

- The features of the audio encoded content are obtained with the automatic extraction of the descriptors. These descriptors are made up by different parameters that represent sound events of the speech/audio signals and by parameters that help the speech recognition performance with identification tasks [D.1.7, D.1.8].

d) **Annotation/classification**: these rely heavily on audiovisual structure analysis and the extracted audiovisual features. The annotation is the basis for the detection of audiovisual semantic concepts and the construction of semantic indices for audiovisual content [D.1.9]. The classification is to find rules using extracted features and then assign the audiovisual content into predefined categories [D.1.10]. The scalable wavelet-based descriptors (Task1) will be combined with audio descriptors (Task2) in order to use them in a global data-cube model which will represent each audiovisual document. To the constitution of visual summaries, we apply data summarization techniques coming from information retrieval domain [D.1.11].

e) **Interactive query**: The scalability challenge is significantly reinforced by the use of partial queries. The signal descriptions issued from tasks 1 and 2 will be employed; these descriptions are well-adapted to partial queries. It is will extend the Locality Sensitive Hashing (LSH) approach in two directions: improve the balance between effectiveness and efficiency for the types of content concerned by this project, and devise LSH-based methods for scalable retrieval with relevance feedback RF [D.1.12].

## D.2 TASK 1: SCALABLE DESCRIPTION OF VISUAL CONTENT

This section presents the work which has been done for Task 1. It consists in the extraction of video descriptors based on motion, color and MFCCs (features encoding roughly the spectral envelope of the audio signal).

### D.2.1 EXTRACTION OF COLOR-BASED DESCRIPTORS

Scalable color descriptors have been extracted from the original video in order to be used in the datacube and for the creation of video summaries. See Section D.4.1.

### D.2.2 RECOGNITION OF ACTIONS FOR SEARCH AND SUMMARIZATION BASED ON MOTION DESCRIPTORS

To summarize video databases and perform content-based retrieval, several issues have to be addressed: video content description and associated similarity (or dissimilarity) measures, index design for storing such descriptions, algorithms for efficient retrieval and summarization. Content description should focus on the aspects that are considered most significant in characterizing the relevant visual content of video scenes. For Mexican cultural videos, we consider that *motion* is the key aspect that should be described. In the recent litterature, several features have been put forward as motion descriptors. They can be global

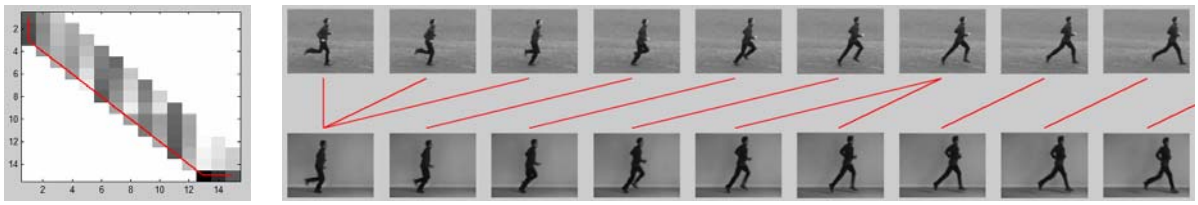
descriptors that take the spatio-temporal volume of the action and describe it as a whole, a time-series of frame by frame descriptors, or aggregates of frame by frame descriptions.

Global approaches usually aim to describe a human performing a certain action captured on video under controlled conditions (single fixed point of view, constant scale, etc.). Either the human body in motion is described using a volumetric descriptor (time being the third dimension) or the human silhouette is extracted and described using a shape descriptor, then parametric model is built.

Local features describing the dynamics of video patches were first defined as extensions of image interest points in the spatio-temporal domain [D.2.8]. An alternative recent solution, proposed in [D.2.1], consists in describing the trajectories of 2D interest points (SURF) tracked over frames. Recently also, trajectory-based features describe patches around high contrast points that are tracked using optical flow. These features [D.2.10] are based on shape (histogram of gradients, HOG), optical flow (histogram of optical flow, HOF) and optical flow gradient (motion boundary histogram, MBH); they were further extended by applying vector field measures to the optical flow field [D.2.6]. Local features of each frame are aggregated into frame features by quantifying them and computing bag of visual words (BoW) histograms.

Once the frame-level description is chosen, an entire video containing one or several actions must be represented. Various ways of aggregating frame signatures have been proposed such as summing up all the frames or creating a tree of atomic action signatures as in [D.2.4]. In [D.2.13] the authors propose to keep frame-level features (in this case SIFT features with motion information) and use the Earth Mover Distance (EMD) to compute the best match between frames, without regard for their order; EMD provides more flexibility than DTW, but some discrimination power could be lost for complex actions.

**Figure 3: Movement description based on trajectories obtained from optical flow.**



**(a)** **(b)**

**Figure 4: Warping path (a) and corresponding alignment (b) between two video sequences.**

We have been experimenting with several of these features on various databases, including KTH [D.2.11], Olympic Sports [D.2.9], Weizmann [D.2.5] or HMDB [D.2.7]. We are currently investigating the potential of multi-scale representations of sequences of such rame features in improving the balance between effectiveness and computation cost.

We show some intermediate results on the KTH and Weizmann databases for the evaluation of query by example. Using trajectory MBH BoW histograms as frame-level features, we show that DTW achieves better results than frame-to-frame matching (L1). The test dataset is obtained by segmenting the KTH videos into 40-frame subsequences and eliminating those that do not contain any movement. The evaluation metric is the Mean Average Precision obtained by querying the entire dataset with the set of all subsequences. The first results below concern the KTH database:

| | Boxing | Handclapping | HandWaving | Jogging | Running | Walking | MEAN |
|---|---|---|---|---|---|---|---|
| L1 | 75 | 63 | 84 | 19 | 9 | 49 | 50 |
| DTW | 75 | 66 | 93 | 19 | 9 | 48 | 52 |

We obtained the same trend on the Weizmann dataset using 30-frame subsequences:

| | Bend | Jack | JumpL | JumpR | pjump | SideL | SideR | wave1 |
|---|---|---|---|---|---|---|---|---|
| L1 | 96 | 98 | 86 | 75 | 95 | 83 | 82 | 79 |
| DTW | 97 | 98 | 86 | 75 | 95 | 83 | 82 | 82 |

We have also found that frame descriptions are sparse (on average, 10% of the bins not null) and we can increase sparseness by using a threshold that helps removing noise.

Most of the features employed for describing motion can be represented as sequences of high-dimensional vectors. To compare such sequences in a scalable way, we have been working on an indexing method inspired by existing indexing solutions for one-dimensional sequences compared with DTW, but better adapted to sequences of sparse high-dimensional vectors. We are also working on associated algorithms for efficient retrieval and similarity self-joins. Indeed, finding compact clusters in a set of data can be based on a similarity self-join that identifies pairs of data points whose similarity is above a threshold and leaves apart isolated data points (see e.g. [C.3.16]).

## D.3 TASK 2: EXTRACTION OF SPEECH / AUDIO CONTENT

A baseline system for video documentary segmentation into perceptually continuous segments has been developed.

It consists in three steps : feature extraction, segmentation and labeling.

1. **Feature extraction:** In step 1, we consider the following features : the 13 first MFCCs coefficients including the 0th order (*Mel-Frequency Cepstral Coefficients* are features encoding roughly the spectral envelope of the audio signal), dominant color and lightness (maximal indexes of Hue and Lightness histograms from the HLS decomposition). The features are first extracted at a fixed period : the analysis window is fixed to 1024 samples, and the hop size to 512 samples for the MFCC, and the visual features are extracted for each image of the video. They are expressed on the same time-scale : for each multiple of the time period T=0.5s, we calculate the mean of the feature vectors contained in the analysis window of length 0.5s and centered on the multiple of T. The features are then combined by

concatenation (early fusion) : at each multiple of T is associated the MFCC coefficients, the dominant color and lightness of the corresponding time instant, as shown in figure 5.

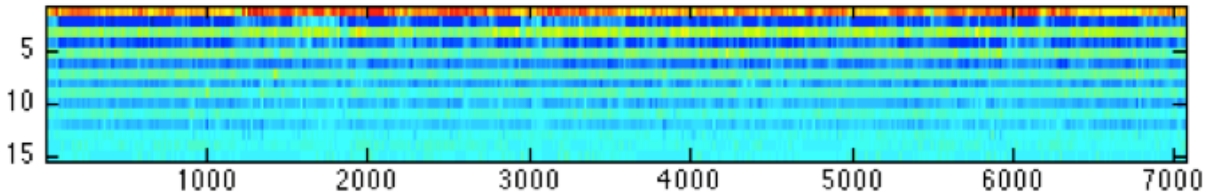



**Figure 5.** The 13 first MFCCs coefficients (indexes 1 to 13 of the y-axis) concatenated with the dominant color and lightness (indexes 14 and 15 respectively) for the INA video « Mexique Magique1 ». The x-axis corresponds to the time axis (with T=0.5s between two feature vectors).

2. **Segmentation** is achieved through the selection of a fixed number of peaks from an homogeneity criterion calculated on the concatenated feature vectors. This criterion is obtained by calculating at each instant *t* the *Generalized Likelihood Ratio* (GLR) considering an analysis window of duration 90s centered on time *t*. The GLR results from the comparison of two multidimensional gaussian models (with dimension = 15), one modeling the set of features contained before *t* within the analysis window (G1), and one modeling the set of features contained after *t* within the analysis window (G2). If G1 and G2 are similar, *t* is likely to belongs to an homogene segment and the GLR has a low value. Homogeneity reputure (G1≠G2) are then characterized by high values of the GLR [D.3.9]. An example is given below (at log scale for better readability).

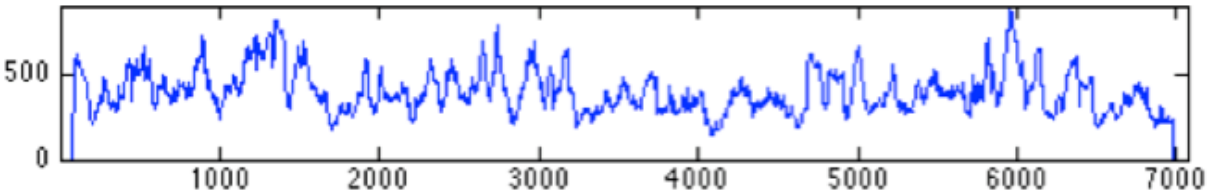



**Figure 6.** Display of the log-GLR calculated using the concatenated features obtained from step 1 for the video « Mexique Magique1 ». The x-axis corresponds to the time axis (with T=0.5s between two feature vectors).

Then we calculate the breakdown criterion by Seck on the log-GLR curve on order to keep its dominant peaks [D.3.9].

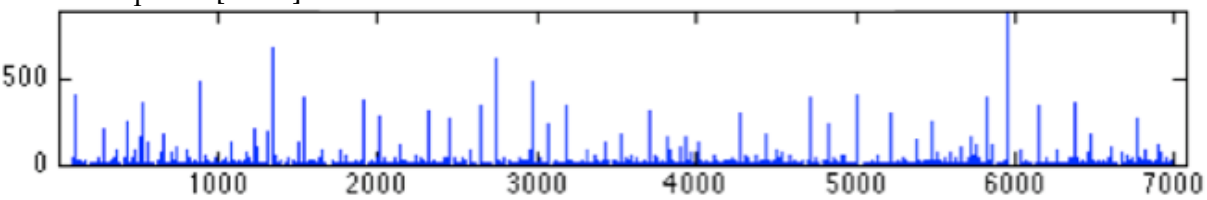



**Figure 7.** Display of Seck's breakdown criterion calculated on the log-GLR for « Mexique Magique1 » The x-axis corresponds to the time axis (with T=0.5s between two feature vectors).

A first segmentation of the video is obtained by selecting the P highest peaks (P is currently fixed to 50, in order to favor oversegmentation). This segmentation will be refined during the labeling process.

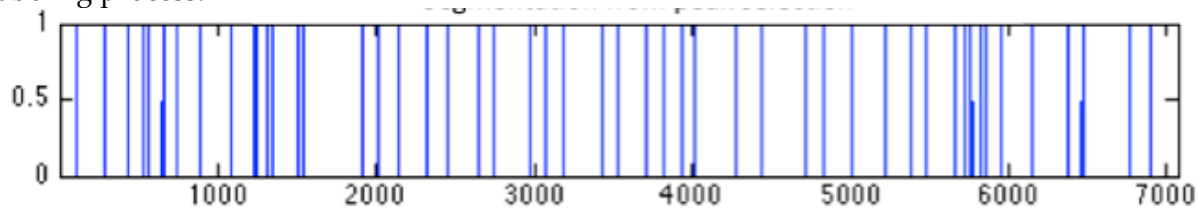



**Figure 8.** Segmentation of « Mexique Magique1 » obtained by selecting the P higest peaks from Seck's criterion. The x-axis corresponds to the time axis (with T=0.5s between two feature vectors).

3. **Labelling**. We calculate for each segment the average of the concatenated feature vectors they contain. The segments are then clustered using a K-means approach, using their average feature vectors and a weighted squared euclidean distance :

$$d_E^2(x,y)=\Sigma w_i(x_i-y_i)$$

with $w_i$=1/13 if i={1...13} (corresponding the the dimentions of the MFCCs), and $w_i$=1/2 otherwise (the 2 visual features). This is a rough way to balance the audio and visual modalities. K is yet chosen arond 10-15 clusters. Consecutive segments gathered within the same cluster are merged. The following figure presents the results of the clustering in the case of « Mexique Magique1 » for K=10 clusters.

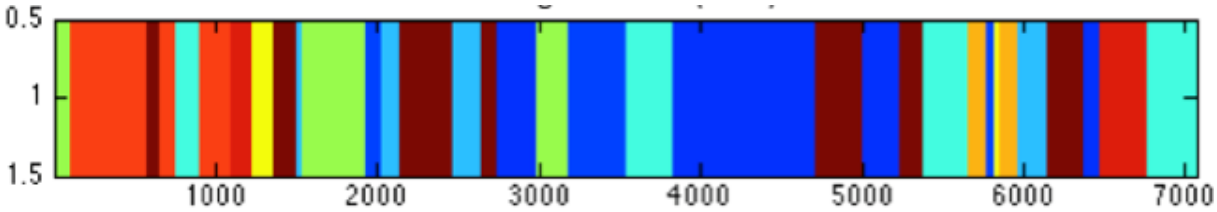



**Figure 9.** Structure of obtained from the K-means clustering the segments obteined for « Mexique Magique1 » with K=10 clusters. The x-axis corresponds to the time axis (with T=0.5s between two feature vectors).

**Application to video summarization :**

An additional step is considered in order to produce a summary of the considered videos. For each of the K clusters, we select the segment whose average feature vector is closest to the cluster's centroïd. An example of the selected segments for « Mexique Magique1 » is shown below.

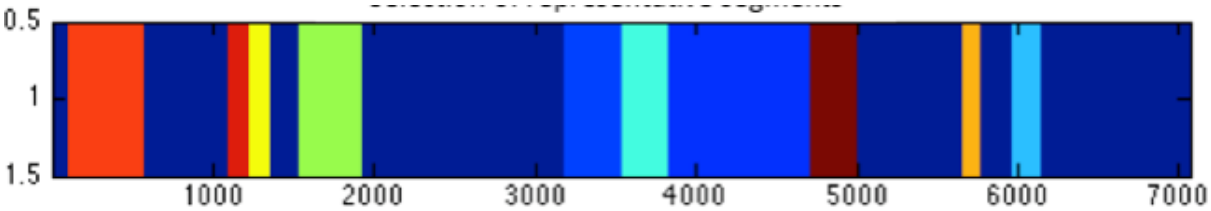



**Figure 10.** Display of the most representative segment for each of the K clusters obtained in the labeling step for « Mexique Magique1 » with K=10 clusters. Dark blue segments are not retained for the summary. The x-axis corresponds to the time axis (with T=0.5s between two feature vectors).

Then we consider the shot estimation obtained from Advene's estimator. For each selected segment, we select the longest shot contained in it.

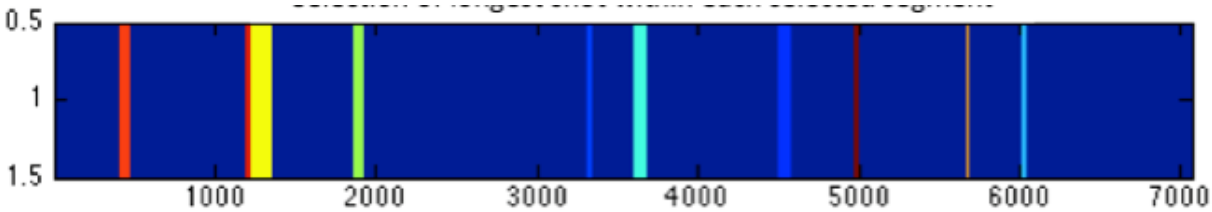



**Figure 11.** Display of the selected shot retained for the resulting summary of « Mexique Magique1 ». Dark blue segments are not retained for the summary. The x-axis corresponds to the time axis (with T=0.5s between two feature vectors).

Their concatenation constitutes the output summary of the video.

## D.4 TASK 3: VIDEO SUMMARIZATION BASED ON DATACUBE APPROACH

This section describes the proposed method for summarizing video content based on *data cube* approach, which allows the multidimensional access to the information stored on it, helping users to navigate and to analyze the video content efficiently by accessing data collections of any dimension subsets, thus providing the scalability property.

In general terms, data cube structure consists of several *dimensions*, where each dimension represents some attribute in the data base, which is represented by a *measure*. Thus, a cell in the data cube represents a measure of interest in the current dimension [D.4.1] as shown in Figure 12.

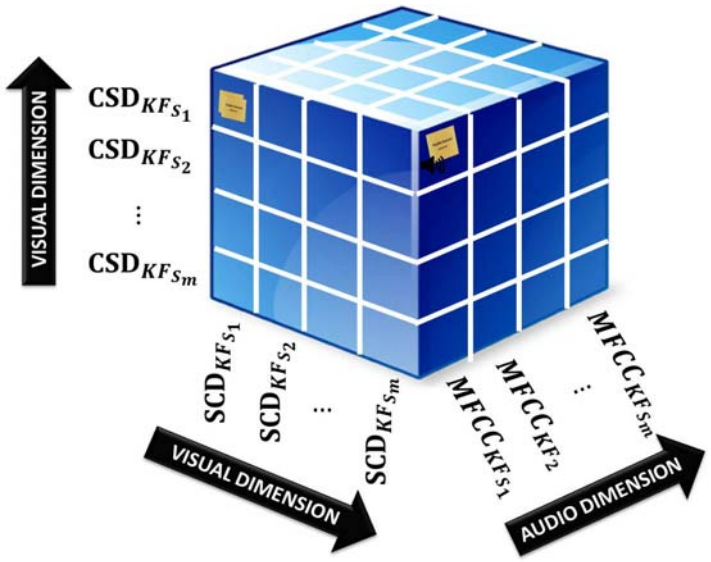



Figure 12: Data Cube approach based on 3 audio-visual dimensions

Figure 12 represents a 3-dimensional data cube structure based on audio-visual features, where each low level feature is considered as one dimension.

Data cube offers a serie of operations to allow flexibility for navigation into the data by displaying a summary at different granularity level. To reach this goal, data cube considers several On Line Analytical Processing (OLAP) operations such as:

- *Rolling up*: This operation implies the data summarization by climbing up hierarchy into the data cube and is also called as consolidation.
- *Drill down*: Inverse operation of rolling-up, where user is able to navigate from higher level summary to lower level summary, this is, to explore into detailed data.
- *Slice:* A slice is a subset of a multi-dimensional array corresponding to a single value for one or more members of the dimensions.
- *Pivot:* This operation is also called rotate operation. It rotates the data in order to provide an alternative presentation of data.

The main challenge of using data cube structure to achieve scalable video summarization is to conceptualize how the data cube is going to be materialized and which kind of dimensions and measures are going to be considered. Since the most relevant information of a frame is given by visual information, the *dimensions* can be considered as a set of image feature descriptors and its related audio information, while the *measure* value is given by the top $k$ representatives obtained from a clustering method. Thus, top $k$ representatives are used to materialize the consecutive levels in the cube. Once that data cube has been materialized the OLAP operations can be computed to display the user the scalable video summary trough a user interface as illustrated in Figure 13.
As shown in Figure 13, developing a video summary based on data cube structure implies two main steps: 1. Feature extraction to conceptualize the dimensions and measures of the data cube to build the base cuboid, and 2. Data cube construction to allow the OLAP operations.

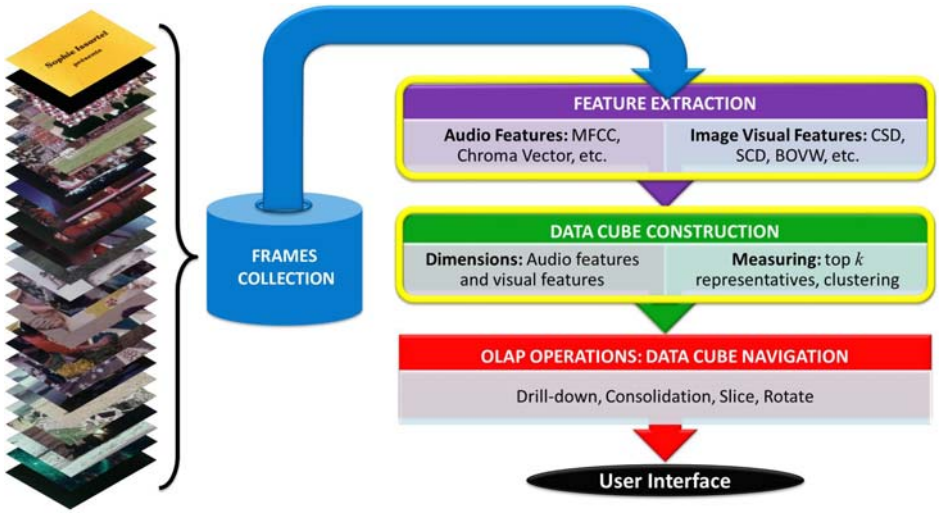



Figure 13. Video Summarization based on data cube approach framework

### D.4.1 Feature Extraction

Typically a data cube consists of a *base cuboid* which is the most top level in the cube. The base cuboid is given by a raw database which contains the multidimensional model of each element considered to build the data cube.

Then to obtain the elements of the raw database, video $V$ is decomposed into $k$ frames per second, such that $V = \{F_1, F_2, \ldots, F_n\}$ where $n$ is equal to the video's duration multiplied by $k$. Assuming that the *dimensions* of the data cube are given by a set of low level descriptors, thus from each video frame, a set of low level audio and video feature descriptors are extracted at different sampling rates, then it is necessary not only to fuse the normalized descriptors obtained to get a full low level description of each frame but also to consider the feature alignment by linear interpolation to overcome the difference between sampling rates. Finally the video is given by $\mathcal{F} = \{\overline{F_1}, \overline{F_2}, \ldots, \overline{F_n}\}$ where $\overline{F_i} = [D_1, D_2, \cdots D_m]\ \forall i \in \{1, 2, \ldots n\}$ while $D_j$ $\forall i \in \{1, 2, \ldots m\}$ is a single normalized feature descriptor which is also considered as one *dimension* in the cube. The process to obtain the data cube is illustrated in Figure 14.

As image feature descriptors, Scalable Color Descriptor (SCD) [D.4.2], Color Structure Descriptor (CSD) [D.4.3], Bag of Visual Words (BOVW) [D.4.4], Pyramid of Histogram of Oriented Gradients (PHOG) [D.4.5], etc. can be used as dimensions. However due to the proven effectiveness in image retrieval based on colors [D.4.6] SCD and CSD are chosen as a color based descriptors to build the raw database, while in the case of audio features, Mel Frequency Cepstral Coefficients (MFCC) and Chroma vectors were considered.

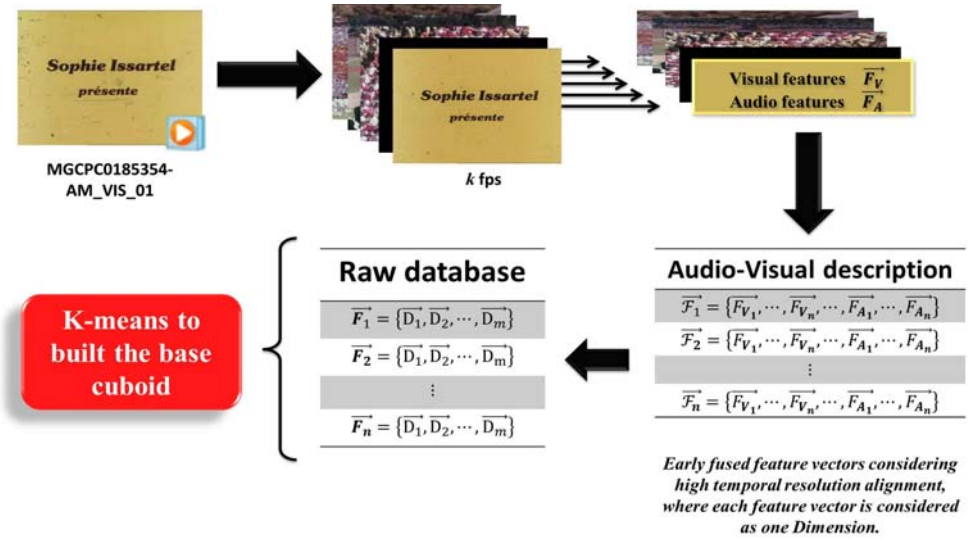



Figure 14: Base cuboid baseline

In order to develop a baseline for our visual summarization approach, the raw database materialization and the development of the main visual summary is formalized as follows

1. Detection and description of low level features
2. Early fusion of several low level (audio and visual) features considering the sampling rate alignment and store them into a raw database, where each raw is given by $\vec{F_i} = \mathcal{D}$
3. Data partitioning using $k$-means by setting $k$ equal to the number of shots manually detected in the video.
4. Post-processing density based clustering stage to separate frames grouped in the same cluster which are similar in features but far in time.
5. Develop the chronological sorting of the obtained clusters, where the number of obtained clusters is bigger than $k$.
6. Selection of one keyframe per each cluster to build the visual summary in the current level.

## *Scalable Color Descriptor*

SCD descriptor [D.4.2] consists of 256-bin color histogram computed on the HSV color space encoded using the Haar transform to make it scalable. This transform develops a sum and a difference operation represent by low and high pass filters respectively. Thus, summing pairs of adjacent bins is equivalent to half size histogram, i. e., 128-bin histogram with 8 levels in Hue, 4 levels in Saturation and 4 levels in Value.

Haar transform claims that it is possible to truncate the high pass coefficients to an integer representation with small number of bits since high pass coefficients depicts the information contained in finer-resolution levels. This information usually expresses high redundancy between the adjacent histogram bins and they have only small values. This descriptor can be compared using the Manhattan distance better known as $L_1$ distance which is given by

$$L_1(x,y) = \sum_{i=1}^{d} |x_i - y_i|$$

( 1 )

## *Color Structure Descriptor*

CSD descriptor [D.4.3] is a modified version of a histogram which captures the distribution of colors in the image as well as the local spatial structure of the colors according to Equation (2)

$$CSD = H_s(m) \qquad m \in 1, \cdots, M \tag{2}$$

where $M \in 256, 128, 64, 32$ is the quantization step, $H_s(m)$ is the number of times a particular color is contained within a squared structuring element as the structuring element scans the image and $s$ is the scale of the associated square structuring element. CSD aims at expressing the local structure of colors in the image through some histogram computation. CSD can be compared using $L_1$ metrics (1).

## *K-means*

K-means is still one of the most widely used algorithms for clustering. Ease of implementation, simplicity, efficiency, speed and empirical success are the main reasons for its popularity.
K-means [D.4.6] is a clustering algorithm which partitions the data set $\mathcal{D} = \{F_1, F_2, \cdots, F_n\}$ into $k < n$ sets $\mathcal{S} = \{S_1, \ldots, S_k\}$ so as to maximize the within intra cluster similarity measure as shown in the equation (3)

$$\arg_{\mathcal{S}} \min \sum_{i=1}^{k} \sum_{x_j \in S_i} \|F_j - \mu_i\| \tag{3}$$

where $x_N$ is a vector which represents the $n$-th element, $\mu_i$ is the center mass (mean) of the cluster $S_i$. K-means consist of the following steps

1. The initial $k$ centers $C = \{c_1, c_2, \ldots, c_k\}$ are randomly selected.
2. for each $i \in \{1, \ldots, k\}$ set $S_i = \left\{F \in \frac{\mathcal{D}}{\|F - c_i\|} \leq \|F - c_j\| \;\; \forall \;\; j \neq 1\right\}$
3. for each $i \in \{1, \ldots, k\}$ set $c_i = \frac{1}{|S_i|} \sum_{F \in S_i} x$
4. repeat 2 and 3 until convergence

To enhance the K-means performance, some other techniques have been proposed to choose the initial clusters, so as to speed up the convergence of the clustering process.

## *Post-processing : Density based clustering*

K-means offers the data partitioning according to feature similarity. However similar frames can be found in different shots of the video, which indeed are chronologically far. To solve this problem, it is possible to use the density based clustering as a post-processing stage by

considering that the density connectivity between the members of the set $S = \{s_1, s_2, \ldots, s_n\}$ is given by the time stamp of each member.

Given the data set $\mathcal{S} = \{S_1, S_2, \ldots, S_k\}$, where $S_i = \{s_1, s_2, \ldots, s_n\}$ $\forall i \in \{1, 2, \ldots k\}$ and considering that $s_j$ exist at the time instant $t_{s_j}$ $\forall j \in \{1, 2, \ldots n\}$, such that $\overline{S}_i = \{t_{s_a}, t_{s_b}, \ldots, t_{s_n}\}$, the member $t_{s_{(j+1)}} \in \overline{S}_i$ iff $t_{s_{(j+1)}} > t_{s_j} + th$, in that case a new subset $\hat{S}$ emerges, where $th$ is a temporal distance threshold (set to 1), otherwise the member $t_{s_{(j+1)}} \in \overline{S}_i$. Thus $\ddot{\mathcal{S}} = \{\ddot{S}_1, \ddot{S}_2, \ldots, \ddot{S}_q\}$ is considered as a set of all new clusters obtained from this procedure, where $\ddot{S}_p$
$\forall p \in \{1, 2, \ldots q\}$ is a new cluster.

Finally $\ddot{\mathcal{S}} = \{\ddot{S}_1, \ddot{S}_2, \ldots, \ddot{S}_q\}$ is chronologically sorted to represent the chronological occurrence of each subset (shot) in the video. Then, the visual summary in the current cuboid, corresponds to the top $k$ representative frame of each subset, thus $Summary = \{kf_1, kf_2, \ldots, kf_q\}$ where $kf_p$ $\forall p \in \{1, 2, \ldots q\}$ is the keyframe of $\ddot{S}_p$.

### D.4.2 Data Cube Construction

Visual data cube approach involves the construction of the top most cuboid, also called the *base cuboid* in the literature, which construction has been detailed in the previous section, but also involves the aggregation of different granularity levels hierarchically, called *cuboids*, where each *cuboid* represents a particular view.

Let us consider a set of dimensions $\mathcal{F} = \{A, B, \ldots, F\}$ where $A, B, \ldots, F$ represent each single dimension, in this example we consider 6 dimensions, i.e., six feature descriptors. This set of dimensions exists along the video $\mathcal{V} = \{F_1, F_2, \ldots, F_n\}$ which is represented by a collection of $F_i$ frames. In order to get the top $k$ representatives from the video, which indeed represent the video summary, the data collection is partitioned using k-means clustering based on audio and visual features.

Assuming there is no partition intersection due k-means, the video summary is given by a set of top $k$ representatives (keyframes) denoted by $\{K_1, K_2, \ldots, K_m\}$, where $m$ is the number of top $k$ representatives found by the proposed approach. Navigation is enabled when user selects *one* keyframe looking for detailed information. Then a summary of the partition related to the selected keyframe along the dimensions $\{\|\mathcal{F}\| - 1\}$ according to the different aggregation possibilities is displayed. Thus, navigation into that subspace is enabled to again select a keyframe and display a new summary and so on, as shown in Figure 15. This figure illustrates the hierarchical scalability through the cube to display video summary at different levels of detail.

The total number of *cuboids* materialized is given by $\mathcal{C} = 2^{\|\mathcal{F}\|}$. It is worth noting that summary in the current level is given by the top $k$ representatives along the dimensions considered for the aggregation of that cuboid. Then, due to the proposed data cube based approach a user is able to navigate into one particular cluster and going down (drill-down operation) or going-up (rolling-up) into the cube.

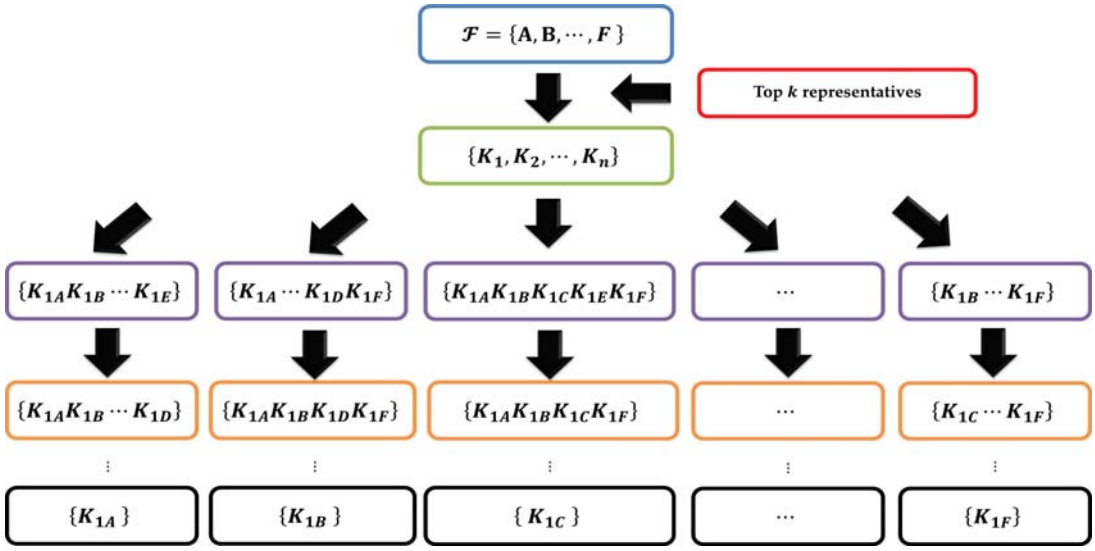



Figure 15 Cuboid aggregation

As shown in Figure 15, each *cuboid* can be seen as conventional database table, which can be processed and handled using a Relational Database Management System (RDBMS). The advantage of this structure is that it requires few assumptions about how data is related or how it will be extracted from the database. Then as a result, the same database can be viewed in different ways, which allow the multidimensional access to video summary.

### RESULTS

The evaluation of the intermediate results is given by the comparison between the ground truth video summary, which was developed under human supervision, and the automatic summary obtained from the proposed method (See D.4.1), which is considered as the base cuboid.

The manual annotation of the video was done using ELAN tool, which is software for video annotation. The annotation file contains the number of shots detected by the user as well as the time instant of beginning and ending of each shot and the keyframe of each shot.

The experiment was developed using the video entitled "Mex Magic 1" from INA Corpus Collection. This video consists of 58 minutes and where manually detected 616 shots from it.

This intermediate result is given in terms CSD and SCD descriptors to get the automatic video annotation (summary), which is also the base cuboid. Since K-means needs the number of clusters, in this first instance, the number of clusters was setting to $k = MDS$, where $MDS$ is the number of shots manually detected during ELAN based annotation. However due to post-processing stage, the number of shots detected by the proposed approach is considerable increased. For that reason, by experimentation, we consider to set $k = 70\%(MDS)$.

Then, according to the intersection matrix between the keyframes manually selected and the keyframes obtained from the proposed approach, the intersection is around 85%, however if the shots instersection is considered instead of keyframe intersection, the accuracy of the proposed video summary achieves 95%, i.e., most of shots obtained from our experiments are very near to the shots manually selected. Figure 16 illustrates a set of keyframes manually selected, compared with keyframes obtained by the proposed approach.

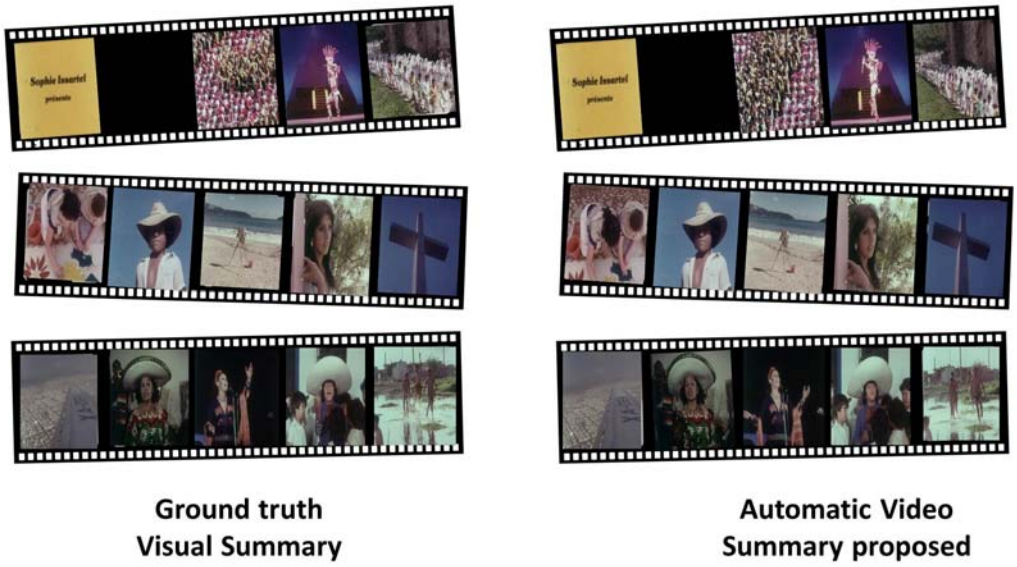



Figure 16: Video summary comparison

From Figure 16 can be seen that keyframes obtained by the automatic video summary represents the same concept of the keyframes obtained by ground truth video summary.

For further experiments, audio features as well as other visual features will be considering to develop the video summary. The proposed approach also considers the use of other clustering methods such as consensus clustering [D.4.7]. Finally, the evaluation methodology will be done using VERT [D.4.8] metric, as described in Section D.5.

## D.5 EVALUATION METRICS

A video summary is defined as short version of the original video which provide at the user an overview of the video content and it can be presented as static keyframes, video skims or multidimensional browser. However the automatic video summarization is still a challenge. To measure the quality of the summary we consider the Video Evaluation by Relevant Threshold (VERT) [D.4.8] as an automatic video summary evaluator.

This method is based on the selection of relevant keyframes, thus in this process is assumed that each shot in the video is represented by one or more keyframes. Then a selection of the

video content to be included in the summary is performed by assigning an importance weight $w_s(f)$ depending on the rank of keyframe $f$ in the selection $S_f$. Finally VERT compares a set of automatically selected keyframes by the proposed summarization method versus the ground truth human-selected keyframes. VERT consists of two measures VERT-P and VERT-R, precision and recall respectively.

### D.5.1 VERT-PRECISION

Let's assume that each keyframe is assigned an importance weight $W_s(x, y)$ according to its position in the selection $(x = 1, \ldots, k)$ and $(y = 1, \ldots, n_i)$, while a non-selected keyframe has a weight of zero, where $k$ is the number of reference summaries (ground truth summaries) and $n$ is the number of keyframes in each one. By the other hand, the proposed summary contains $m$ keyframes, and each keyframe $i$ is assigned a weight$W_C(i)$. VERT-P measures the precision of the position of each candidate keyframe found using our summary approach with respect to the ground truth summaries. Hence, for each keyframe $i$ in the candidate, the maximum weight assigned in the ground truth summary is $T_i = \max_x W_S(x, y_i)$, where $x$ is a ground truth summary, and $y_i$ is the position of keyframe $i$ in $x$.
VERT-P compares the maximum weight $T_i$ with the weight assigned to the current keyframe in the proposed summary, such that

$$VERT \quad P = \frac{\sum_{i=1}^{m} \min[W_C(i), T_i]}{\sum_{i=1}^{m} W_C(i)} \qquad (5)$$

This value used to be between 0 and 1. The maximum is obtained when all keyframes of the proposed summary were selected with a weight that is lower than at least one of the ground truth selections.

### D.5.2 VERT-RECALL

Considering that $C$ is the proposed video summary, $gram_n$ is a group of $n$ keyframes, $W_S(gram_n)$ is the weight of the group $gram_n$ for a ground truth summary $S$, and $W_C(gram_n)$ is the weight of the group $gram_n$ for the proposed summary $C$, the VERT-$R_N$ is given by

$$VERT - R_N = \frac{\sum_{se\{\text{ground truth summaries}\}} \sum_{gram_n \in S} W_C(gram_n)}{\sum_{se\{\text{ground truth summaries}\}} \sum_{gram_n \in S} W_S(gram_n)} \qquad (6)$$

VERT-$R_N$ computes a percentage of $gram_n$ from the ground truth summaries occurring also in the proposed summary. Thus, if a group of $n$ keyframes is considered as a subset of size $n$, then, when $n = 1$ and $n = 2$, the VERT-$R_1(C)$ and VERT-$R_2(C)$ can be defined as

$$VERT - R_1(C) = \frac{\sum_{seR}\sum_{f\in s} W_C(f)}{\sum_{seR}\sum_{f\in s} W_S(f)}$$

$$VERT - R_2(C) = \frac{\sum_{seR}\sum_{(f,g)\in s} W_C(f,g)}{\sum_{seR}\sum_{(f,g)\in s} W_S(f,g)} \quad (7)$$

Such that each $gram_1$ in $VERT - R_1$ contains only 1 keyframe, thus the number of keyframes is equal to number of $gram_1$ and the weight of a group is then the weight of keyframe. While in $VERT - R_2$, there are 2 keyframes in each $gram_2$. Hence, $VERT - R_2$ considers two measures $VERT - R_{S2}$ and $VERT - R_{2D}$. Where the weight of $gram_2$ in $VERT - R_{S2}$ is the average of the weights of the keyframes and in $VERT - R_{2D}$ is the absolute difference between the weights, such that

$$VERT - R_{S2} - W_S(f,g) - \frac{W_S(f) + W_S(g)}{2}$$

$$VERT - R_{2D} = W_S(f,g) = |W_S(f) + W_S(g)| \quad (8)$$

## D.6 GROUNDTRUTHING OF COMPLEX AUDIO-VISUAL CONTENT

### D.6.1 MANUAL ANNOTATION

In order to allow us to evaluate the approaches developed during the project, we need to produce, within the project, reference annotations for the video and audio content (the available content only has some global metadata like filename, show name, channel, author, etc.).

***Annotation types.*** We are preparing a multidimensional annotation of the videos according to the following "dimensions": visual content, actions, speech, music, audio special effects / environmental sound effects. Given the cost of this task, we aim to devise an unambiguous set of labels for every "dimension" in order to produce a corpus of reference content annotations that is coherent and as reproducible as possible.

***Annotation software.*** Three software tools (ANVIL, ADVENE, ELAN) have been evaluated against the following requirements : free and open-source, compatible with MPEG4, easy to use, documented in English, open export format. The standard version of ANVIL is not compatible with the video formats considered. ADVENE proposes an automatic segmentation of the video into shots and a "full zoom-out" on the video. ELAN doesn't have such functionalities, however, it's interface is found to be more suitable for the manual annotation itself (keyboard shortcuts and a faster way to edit the segments labels). Therefore, ELAN will be used for this task, and ADVENE's shot estimation program will be used to produce the segmentations in shots that will be imported on advene through a format conversion process. Figure 17 shows an overview of ELAN's interface.

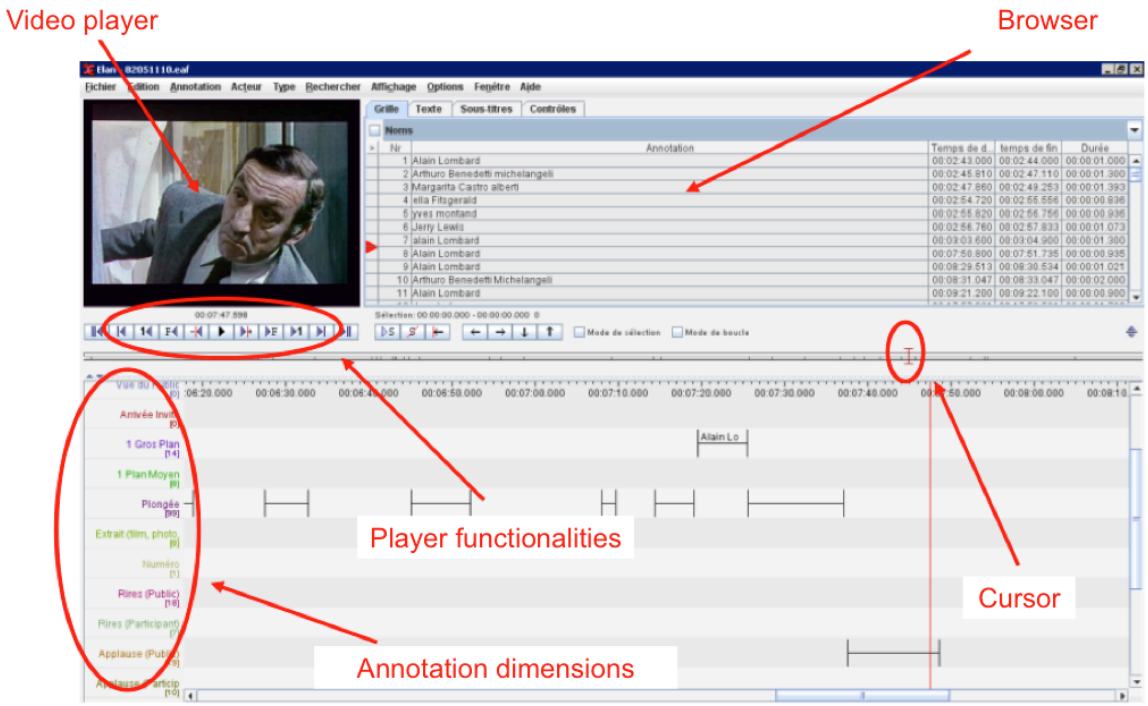



Figure 17: Overview of ELAN's interface

***Format of annotation files.*** A "tabulation-separated values" format following the model of ELAN will be used. An example of such a format is given in Figure 18.

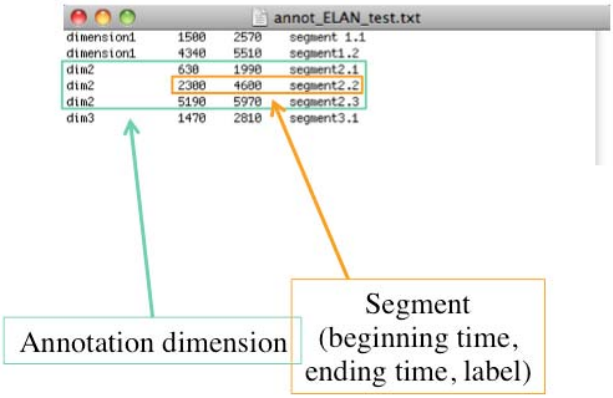



Figure 18: File format.

### D.6.2 FEATURE FILE FORMAT

Figure 1 describes the overall analysis chain of a video. This allows to enlighten three types of feature files. They all derive from the Tabulation Separated Values (tsv) format, in which values or labels are separated by a tabulation.
The feature files are of three types :

- *FSA/FSV (fileIn files)* : low-level features used by processing algorithms
- *FIA/FIV (fileOut files)* : high level features produced by the algorithms
- *FOA/FOV (fileGlobal)* : file gathering *fileIn* and *fileOut* features expressed at the same timescale (highest frequency rate amongst all features) to be used for large-scale research and retrieval.

1. *FSA/FSV* : consist in different types of features which have different sampling rates. We choose to store the informations related to one feature type in a single tsv file, made of a header with the feature name and extraction parameters, and its values

2. *FIA/FIV*: will mainly consist in segmentations in which each segment is associated to a label (symbol, or string without space) or a sequence of values (vector). A tsv file storing the informations related to one high-level feature type is expected to be made of a header containing the feature name, and a list of segments (start time, end time and label). -> such a file may be converted to another tsv file which shares the same format and the same sampling rate than the low-level features analyzed. This intermediate file is created in order to merge easily all the features (high-level and low-level) in one single file (*FOA/FOV*).

3. *FOA/FOV*: All the features used and produced by the algorithms are put altogether through two files, who follow a format derived from the tsv format used in the "low-level feature" case. The first one gathers all the feature parameters and their position within the second file (column indexes). The second one gathers all the feature values, re-sampled at a single sampling frequency and concatenated according to their dimensions.

## E CONCLUSION AND PERSPECTIVES

Since the beginning of the project, the following progress have been done:

- First version of the architecture for content-based audiovisual scalable indexing and retrieval for the projet. It shows then the interrelationship between the different activities of each partner.

- Design of a system to split a video into homogeneous segments using a criterion based on audio/video descriptors. It allows to create video summaries.

- Development of methods to extract audio/video scalable descriptors.

- Definition of a metric to evaluate the quality of the summaries.

- Development of a method based on dynamic time warping (DTW). It has been shown that it can improve the quality of matching (compared to a direct frame-to-frame comparison) for sequences of high-dimensional features describing video sequences.

- We devised an indexing solution for the scalable comparison with DTW of sequences of sparse high-dimensional vectors.

Furthermore, from a methodological/technical point of view, a significant work has been done. In particular, the following points have been specified :

- Specification of file formats to store manual annotations and audio visual features. This is a crutial point to exchange data within the project.

- Choice of a software to perform manual annotation of videos

The main perspective of the project for the coming year will be to design a complete system. It will be based on selected descriptors and will integrate contributions coming from the different partners. To reach this goal, the following points will be adressed:

- Definition of an automatic annotation method based on extracted descriptors and/or manual annotation. For that purpose, the following aspects will have to be developed:

  - Clear definition of a *thesaurus*. It will be chosen according to the available corpus.
  - Manual annotation of the video database (using the thesaurus).
  - Definition of a method to extract automatic annontation from low level audiovisual descriptors.

- Definition of a method to create a video summary directly from the datacube

- Include more audio and visual features in the system, measure their impact on the summary

- Design a way to better adjust the weights of the audio and video modalities

- Compare different fusion methods for segment selection (early vs late...) to evaluate the interest in considering the audio in the summary

# F REFERENCES

[C.1.1] Mitrovic D, Hartlieb S, Zeppelzauer M and Zaharieva M, "Scene Segmentation in Artistic Archive Documentaries", 6th Symposium of the Workgroup Human-Computer Interaction and Usability Engineering, USAB 2010, Klagenfurt

[C.1.2] Kompatsiaris Y, Merialdo B and Lian S, "TV Content Analysis: Techniques and Applications", book chapter in "TV Content Analysis: Techniques and Applications", editors Abduraman A E, Berrani S-A, Mérialdo B, October 2011, ISBN: 9781439855607

[C.1.3] Hari Sundaram, Shih-Fu Chang, "Video scene segmentation using video and audio features", In proc. of ICME, 2000

[C.1.4] Naveen Goela, Kevin Wilson, Feng Niu, and Ajay Divakaran. An svm framework for genre-independent scene change detection. In IEEE International Conference on Multimedia and Expo, pages 532–535, Beijing, China, July 2007.

[C.1.5] Silvia Pfeiffer, Rainer Lienhart, and Wolfgang Efflsberg. Scene determination based on video and audio features. Multimedia Tools and Applications, 15(1):59–81, September 2001.

[C.1.6] Panagiotis Sidiropoulos, Vasileios Mezaris, Ioannis Kompatsiaris, Hugo Meinedo, and Isabel Trancoso. "Multi-modal scene segmentation using scene transition graphs". In 17th ACM International Conference on Multimedia, pages 665–668, Beijing, China, October 2009.

**C.2**

[C.2.1] Y. Gong, X. Liu. Video summarization with minimal visual content redundancies. 2001 International Conference on Image Processing, vol. 3, pages: 362–365, 2001.

[C.2.2] R. Wen, Z. Yuesheng. Intelligent information Hiding and Multimedia Signal Processing. International conference on Digital Object Identifier, pages: 450-453, 2008.

[C.2.3] E. Fendri, H. Ben Abdallah, A novel approach for soccer video summarization. Second international conference on multimedia and information technology, 2:138-141, 2010.

[C.2.4] Z. Xiong, Y. Rui, R. Radhakrishnan, A. Divakaran, T. Huang. A Unified Framework for video summarization Browsing, and Retrieval. Handbook of Image and video processing (Second Edition), pages: 1013-1029, 2005.

[C.2.5] J. Almeida, N. Leite, Vison: video sumaarization for online application .Pattern Recognition Letters, Volume 33, Issue 4, pages: 397-409, 2012

[C.2.6] J. Almeida, N. Leite, R. Torres. Online video summarization on compressed domain. Journal of Visual Communication and Image Representation, 24: 729-738, 2013

[C.2.7] E. Rossi, S. Benini, R. Leonardi, B. Mansencal, J. Benois-Pineau. Clustering of scene repeats for essential rushes preview. Workshop on Image Analysis for Multimedia Interactive Services, pages: 234-237, 2009.

[C.2.8] C. Zhang, S. Yat-sen, Z. Zheng, F. Zhang, and J. Ren. Multidimensional traffic gps data quality analysis using data cube model. International Conference on Transportation, Mechanical, and Electrical Engineering (TMEE), pages: 307-310, 2011.

[C.2.9] X. Liu, K. Tang, and J. Han. Socialcube: A text cube framework for analyzing social media data. International Conference on Social Informatics (SocialInformatics), pages: 252-259, 2012.

[C.2.10] J. Yi and Y. Du. Visual data exploration of space-time characteristics of mesoscale eddies in the south china sea: A cube-based approach. International Conference on Geoinformatics, pages: 1-5, 2012.

[C.2.11] E. Garnaud, S. Maabout, and M. Mosbah. Using functional dependencies for reducing the size of a data cube. In Proceedings of the 7th international conference on Foundations of Information and Knowledge Systems, pages 144-163, 2012.

[C.2.12] A. Nandi, C. Yu, P. Bohannon, and R. Ramakrishnan. Data cube materialization and mining over map reduce. IEEE Transactions on Knowledge and Data Engineering, 24(10):1747-1759, 2012.

[C.2.13] A. Moreira and J. de Castro-Lima. Full and partial data cube computation and representation over commodity pcs. International Conference on Information Reuse and Integration (IRI), pages: 672-679, 2012.

[C.2.14] X. Jin, J. Han, L. Cao, J. Luo, B. Ding, and C. Xide Lin. Visual cube and on-line analytical processing of images. In Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM), pages: 849-858, 2010.

**C.3**

[C.3.1] Moses S. Charikar. Similarity estimation techniques from rounding algorithms. In STOC'02: Proceedings of the 34th annual ACM symposium on Theory of computing, pages 380–388, New York, NY, USA, 2002. ACM.

[C.3.2] Ondrej Chum, Michal Perd'och, and Jiri Matas. Geometric min-hashing: Finding a (thick) needle in a haystack. In CVPR'09: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pages 17–24, June 20–26 2009.

[C.3.3] Michel Crucianu. Encyclopedia of Database Systems, chapter Image retrieval with relevance feedback. Springer Verlag, 2009.

[C.3.4] Michel Crucianu, Daniel Estevez, Vincent Oria, and Jean-Philippe Tarel. Speeding up active relevance feedback with approximate kNN retrieval for hyperplane queries. International Journal of Imaging Systems and Technology, Special issue on Multimedia Information Retrieval, 18:150–159, 2008.

[C.3.5] Mayur Datar, Nicole Immorlica, Piotr Indyk, and Vahab S. Mirrokni. Locality-sensitive hashing scheme based on p-stable distributions. In SCG'04: Proceedings of the 20th annual symposium on Computational geometry, pages 253–262, New York, NY, USA, 2004. ACM.

[C.3.6] D. Gorisse, M. Cord, and F. Precioso. Salsas: Sub-linear active learning strategy with approximate k-nn search. Pattern Recognition, In Press, Corrected Proof, 2011.

[C.3.7] P. Indyk and R. Motwani. Approximate nearest neighbors: towards removing the curse of dimensionality. In STOC'98: Proceedings of the 30th annual ACM symposium on Theory of computing, pages 604–613, New York, NY, USA, 1998. ACM.

[C.3.8] Alexis Joly and Olivier Buisson. A posteriori multi-probe locality sensitive hashing. In MM '08: Proceeding of the 16th ACM international conference on Multimedia, pages 209–218, New York, NY, USA, 2008. ACM.

[C.3.9] Herwig Lejsek, Fridrik Heidar Ásmundsson, Björn THór Jónsson, and Laurent Amsaleg. Nv-tree: An efficient disk-based index for approximate search in very large high-dimensional collections. IEEE Trans. Pattern Anal. Mach. Intell., 31(5):869–883, 2009.

[C.3.10] Saloua Litayem, Alexis Joly, and Nozha Boujemaa. Interactive objects retrieval with efficient boosting. In MM '09: Proceedings of the seventeen ACM international conference on Multimedia, pages 545–548, New York, NY, USA, 2009. ACM.

[C.3.11] Qin Lv, William Josephson, Zhe Wang, Moses Charikar, and Kai Li. Multi-probe LSH: efficient indexing for high-dimensional similarity search. In VLDB'07: Proceedings of the 33rd international conference on Very large data bases, pages 950–961. VLDB Endowment, 2007.

[C.3.12] Navneet Panda and Edward Y. Chang. Exploiting geometry for support vector machine indexing. In SDM, 2005.

[C.3.13] Navneet Panda and Edward Y. Chang. Efficient top-k hyperplane query processing for multimedia information retrieval. In Proceedings of the 14th ACM international conference on Multimedia, pages 317–326, New York, NY, USA, 2006. ACM Press.

[C.3.14] Navneet Panda, King-Shy Goh, and Edward Y. Chang. Active learning in very large databases. Multimedia Tools and Applications, 31(3):249–267, 2006.

[C.3.15] Sébastien Poullot, Olivier Buisson, and Michel Crucianu. Scaling content-based video copy detection to very large databases. Multimedia Tools and Applications, 2009.

[C.3.16] Sébastien Poullot, Michel Crucianu, and Olivier Buisson. Scalable mining of large video databases using copy detection. In MM'08: Proceedings of the 16th ACM international conference on Multimedia, pages 61–70, New York, NY, USA, 2008. ACM.

[C.3.17] Sébastien Poullot, Michel Crucianu, and Shin'ichi Satoh. Indexing local configurations of features for scalable content-based video copy detection. In LS-MMRM: 1st Workshop on Large-Scale Multimedia Retrieval and Mining, in conjunction with 17th ACM international conference on Multimedia, pages 43–50, New York, NY, USA, 2009. ACM.

[C.3.18] Thanawin Rakthanmanon, Bilson Campana, Abdullah Mueen, Gustavo Batista, Brandon Westover, Qiang Zhu, Jesin Zakaria, and Eamonn Keogh. 2012. Searching and mining trillions of time series subsequences under dynamic time warping. In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '12). ACM, New York, NY, USA, 262-270.

[C.3.20] Yair Weiss, Antonio Torralba, and Rob Fergus. Spectral hashing. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, Advances in Neural Information Processing Systems 21, pages 1753–1760. 2009.

[C.3.21] Yi Yu, Michel Crucianu, Vincent Oria, and Lei Chen. Local summarization and multi-level LSH for retrieving multi-variant audio tracks. In MM'09: Proceedings of the 17th ACM international conference on Multimedia, pages 341–350, New York, NY, USA, 2009. ACM.

[C.3.22] Yi Yu, Kazuki Joe, and J. Stephen Downie. Efficient query-by-content audio retrieval by locality sensitive hashing and partial sequence comparison. IEICE Trans. on Information and Systems, E91-D(6):1730–1739, June 2008.

[C.3.23] Crucianu, M. (2012) Scalability issues in visual information retrieval. In: J. Benois-Pineau, F. Precioso, M. Cord (eds) Visual indexing and retrieval. Springer Verlag, 2012, pp. 65-81, ISBN 978-1-4614-3587-7.

**D.1**

[D.1.1] M. S. Lew, N. Sebe, C. Djeraba, and R. Jain, "Content-based multimedia information retrieval: State of the art and challenges," ACM Trans. Multimedia Comput., Commun. Appl., vol. 2, no. 1, pp. 1–19, Feb. 2006.

[D.1.2] C. Morand, J. Benois-Pineau, J.P. Domenger, "Scalable indexing of HD video," Proceeding of International Workshop on Content-based Multimedia Indexing. pp.417–424. CBMI08, 2008.

[D.1.3] J. Benois-Pineau, F. Morier, D. Barba, and H. Sanson, "Hierarchical segmentation of video sequences for content manipulation and adaptive coding," 66(2):181–201, April 1998.

[D.1.4] O. Brouard, F. Delannay, V. Ricordel, and D. Barba, "Spatio-temporal segmentation and regions tracking of high definition video sequences based on a markov random field model," Pages 1552–1555, 2008.

[D.1.5] Francesca Manerba, J. Benois-Pineau, and Riccardo Leonardi, "Extraction of foreground objects from an mpeg2 video stream in rough-indexing framework," volume 5307, pages 50–60. SPIE, 2004.

[D.1.6] P. Salembier and L. Garrido, "Binary partition tree as an efficient representation for image processing, segmentation, and information retrieval," 9(4):561–576, April 2000.

[D.1.7] Pierre Hanna, Pascal Ferraro, and Matthias Robine, "On optimizing the editing algorithms for evaluating similarity between monophonic musical sequences," 2007.

[D.1.8] Hugo Meinedo and Joao Neto, "Automatic speech annotation and transcription in a broadcast news task," In in Proc. ISCA ITRW on Multilingual Spoken Document Retrieval, Hong Kong, 2003.

[D.1.9] Y.Wu,B. L. Tseng, and J.R. Smith, "Ontology-based multi-classification learning for video concept detection," in Proc. IEEE Int. Conf. Multimedia Expo., vol. 2, pp. 1003–1006. Jun. 2004.

[D.1.10] S. Fischer, R. Lienhart, and W. Effelsberg, "Automatic recognition of film genres," in Proc. ACM Int. Conf. Multimedia, pp. 367–368. 1995.

[D.1.11] Nicolas Hanusse, Sofian Maabout, and Radu Tofan, "Matérialisation partielle des cubes de données," In Actes de la conférence Bases de Données Avancées (BDA), pages 21–40, Namur (Belgique), octobre 2009.

[D.1.12] Yi Yu, Michel Crucianu, Vincent Oria, and Lei Chen, "Local summarization and multi-level LSH for retrieving multi-variant audio tracks," In MM'09: Proceedings of the 17th ACM international conference on Multimedia, pages 341–350, New York, NY, USA, 2009.

**D.2**

[D.2.1] Nicolas Ballas, Bertrand Delezoide, Françoise J. Prêteux: Trajectory signature for action recognition in video. ACM Multimedia 2012: 1429-1432.

[D.2.2] Emilie Dumont and Bernard Merialdo. Rushes video parsing using video sequence alignment. In CBMI 2009, 7th International Workshop on Content-Based Multimedia Indexing, June 3-5, 2009, Chania, Crete Island, Greece, Crete Island, GR` ECE, 06 2009.

[D.2.3] Adrien Gaidon, Zaid Harchaoui, and Cordelia Schmid. A time series kernel for action recognition. In British Machine Vision Conference, Dundee, United Kingdom, August 2011.

[D.2.4] Adrien Gaidon, Zaid Harchaoui, and Cordelia Schmid. Recognizing activities with cluster-trees of tracklets. In BMVC, Guildford, Royaume-Uni, September 2012.

[D.2.5] Lena Gorelick, Moshe Blank, Eli Shechtman, Michal Irani, and Ronen Basri. 2007. Actions as Space-Time Shapes. IEEE Trans. Pattern Anal. Mach. Intell. 29, 12 (December 2007), 2247-2253.

[D.2.6] Jain, M., Jégou, H., & Bouthemy, P. (2013). Better exploiting motion for better action recognition. Computer Vision and Pattern Recognition 2013.

[D.2.7] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: A Large Video Database for Human Motion Recognition. ICCV, 2011.

[D.2.8] Ivan Laptev. On space-time interest points. Int. J. Comput. Vision, 64(2-3):107–123, September 2005.

[D.2.9] Juan Carlos Niebles, Chih-Wei Chen and Li Fei-Fei. Modeling Temporal Structure of Decomposable Motion Segments for Activity Classification. 11th European Conference on Computer Vision (ECCV), 2010.

[D.2.10] Michalis Raptis and Stefano Soatto. Tracklet descriptors for action modeling and video analysis. In Proceedings of the 11th European conference on Computer vision: Part I, ECCV'10, pages 577–590, Berlin, Heidelberg, 2010. Springer-Verlag.

[D.2.11] Christian Schuldt, Ivan Laptev and Barbara Caputo. Recognizing Human Actions: A Local SVM Approach, in Proc. ICPR'04, Cambridge, UK.

[D.2.12] Feng Wang, Yu-Gang Jiang, and Chong-Wah Ngo. Video event detection using motion relativity and visual relatedness. In Proceedings of the 16th ACM international conference on Multimedia, MM '08, pages 239–248, New York, NY, USA, 2008. ACM.

[D.2.13] HengWang, Alexander Kläser, Cordelia Schmid, and Liu Cheng-Lin. Action Recognition by Dense Trajectories. In IEEE Conference on Computer Vision & Pattern Recognition, pages 3169–3176, Colorado Springs, United States, June 2011. MSR - INRIA.

**D.3**

[D.3.9] Mouhamadou Seck, Raphaël Blouet, and Frédéric Bimbot. "The IRISA/ELISA Speaker Detection and Tracking Systems for the NIST'99 Evaluation Campaign".Digital Signal Processing, 10(1-3) :154–171, January 2000.

**D.4**

[D.4.1] J. Gray, A. Bosworth, A. Lyaman, H. Pirahesh. Data cube: a relational aggregation operator generalizing group by, cross-tab, and sub-totals. Twelfth International Conference on Digital Object Identifier, pages: 152 – 159, 1996.

[D.4.2] W. Yong-ge and P. Sheng-ze. Research on Image Retrieval based on Scalable Color Descriptor of MPEG7. Advances in Control and Communications, pages: 137:91-98, 2012.

[D.4.3] P. Beek and J.H Errico. The MPEG-7 Color Structure Descriptor: Image description using color and local spatial information. In proceedings of the International Conference on Image processing, pages: 670-673, 2001.

[D.4.4] J. Yu-Gang and N. Chong-Wah. Bag of Visual Words explanation using visual relatedness for video indexing. In proceedings of the 31st annual International Conference on Research and Development in information retrieval, pages: 769-770, 2008.

[D.4.5] A. Bosh, A. Zisserman and X. Munoz. Representing shape with a spatial pyramid kernel. In proceedings of the International Conference on Image and Video Retrieval, pages: 401-408, 2007.

[D.4.6] K. Jain. Data Clustering: 50 years beyond K-means. Pattern Recognition Letters, 31:651-666, 2010.

[D.4.7] S. monti, P. Tamayo, J. Mesirov and T. Golub. Consensus clustering: A resampling based method for class discovery and visualization of gene expression microarray data. Machine learning, functional genomics special issue, 2003.

[D.4.8] L. Yingbo and B. Merialdo. VERT: Automatic Evaluation of Video Summaries. In Proceedings of the International Conference on Multimedia, pages: 851-854, 2010.