Multimed Tools Appl (2010) 48:23–49
DOI 10.1007/s11042-009-0375-8

RUSHES—an annotation and retrieval engine for multimedia semantic units

Oliver Schreer · Ingo Feldmann · Isabel Alonso Mediavilla · Pedro Concejero · Abdul H. Sadka · Mohammad Rafiq Swash · Sergio Benini · Riccardo Leonardi · Tijana Janjusevic · Ebroul Izquierdo

Published online: 2 October 2009

Abstract Multimedia analysis and reuse of raw un-edited audio visual content known as rushes is gaining acceptance by a large number of research labs and companies. A set of research projects are considering multimedia indexing, annotation, search and retrieval in the context of European funded research, but only the FP6 project RUSHES is focusing on automatic semantic annotation, indexing and retrieval of raw and un-edited audio-visual content. Even professional content creators and providers as well as home-users are dealing with this type of content and therefore novel technologies for semantic search and retrieval are required. In this paper, we present a summary of the most relevant achievements of the RUSHES

O. Schreer · I. Feldmann
Fraunhofer Institute for Telecommunications/Heinrich-Hertz-Institut, Berlin, Germany

O. Schreer
e-mail: oliver.schreer@hhi.fraunhofer.de

I. Feldmann
e-mail: ingo.feldmann@hhi.fraunhofer.de

I. A. Mediavilla · P. Concejero
Telefónica I+D, Madrid, Spain

I. A. Mediavilla
e-mail: iam@tid.es

P. Concejero
e-mail: pedroc@tid.es

A. H. Sadka · M. R. Swash
Brunel University, London, UK

A. H. Sadka
e-mail: abdul.sadka@brunel.ac.uk

M. R. Swash
e-mail: rafiq.swash@brunel.ac.uk

project, focusing on specific approaches for automatic annotation as well as the main features of the final RUSHES search engine.

Keywords Rushes · Video retrieval · Annotation · Visualisation

1 Introduction

Due to the explosive growth of audio-visual data and the widespread use of multimedia content in the Web, increasing demands exist to handle this huge amount of data. Therefore, novel ways of meaningful description on higher level are required and the search itself turns to be the main paradigm for fast and efficient data access. It appears that the origin of multimedia content is moving from professionally produced content to user-generated content. This can be recognized by a number of public facing search sites and Internet portals such as YouTube, Google Video and Yahoo!Video. Furthermore, in the professional domain, novel search and retrieval techniques become of high relevance. Broadcasters are the producers of vast quantities of video footage. Some of the material will be used for productions, but there is still plenty of footage that has been shot but not necessarily ever used. Since the amount of available material is very large the requirements for storage, although decreasing, are still significant. Broadcasters usually have a strict media management policy that keeps unedited media content (including outtakes) for a short period (e.g. 1 year) and material with higher re-use expectations (stock footage) for longer time (e.g. 5 years).

Since only a small portion of the rushes is actually used in the final productions at broadcasters, it is generally believed that the ability to summarize such rushes might contribute significantly to an overall rushes management and exploitation solution. For this reason, a number of research groups participating to the “rushes exploitation” task in the TRECVID 2008 campaign [22] mainly dealt with rushes summarisation, believing that this might also help other tasks, such as search and retrieval.

However, it can be observed that rushes material usually has well-defined and distinctive multimodal properties which, if correctly exploited, might enable the retrieval task without the need of a preliminary summarisation stage. In fact, as stated in [30], efficient retrieving from large video archives depends on the availability of

S. Benini (✉) · R. Leonardi
University of Brescia, Brescia, Italy
e-mail: sergio.benini@ing.unibs.it

R. Leonardi
e-mail: riccardo.leonardi@ing.unibs.it

T. Janjusevic · E. Izquierdo
Queen Mary, University of London, London, UK

T. Janjusevic
e-mail: tijana.janjusevic@elec.qmul.ac.uk

E. Izquierdo
e-mail: ebroul.izquierdo@elec.qmul.ac.uk

indexes, and effective indexing requires a multimodal approach in which different modalities (auditory, visual, etc.) are used in collaborative fashion.

The European FP6 funded research project RUSHES designed, implemented, and validated through trial a system for indexing, accessing and delivering raw, unedited audio-visual footage, known as rushes. The reuse of such content in the production of new multimedia assets is offered by semantic media search capabilities [26].

After 2 years work, significant results have been achieved whereby the focus is on the development of novel algorithms and techniques for annotation, indexing and search of large video repositories of un-edited audiovisual content. In the next section, the overall design of the developed RUSHES system is presented. In Section 3, novel techniques for multi-modal analysis of un-edited audiovisual content are described. Finally in Section 4, new components for visualization and browsing of large video repositories are explained. The paper ends with a conclusion and outlook.

2 The RUSHES system

In Fig. 1, the overall RUSHES workflow is depicted showing data storage, the automatic and manual annotation modules and the user interface. The developed user interface will offer novel search capabilities including relevance feedback mechanisms for two individual user scenarios: the home user and the professional user in the broadcaster domain.

New content will be ingested in the content database (DB) and then processed by a large set of low- and mid-level classifiers. The aim of the project was to develop novel classifiers on a high semantic level in order to allow users an easy access to the desired content in the database. Furthermore, the focus of development was also

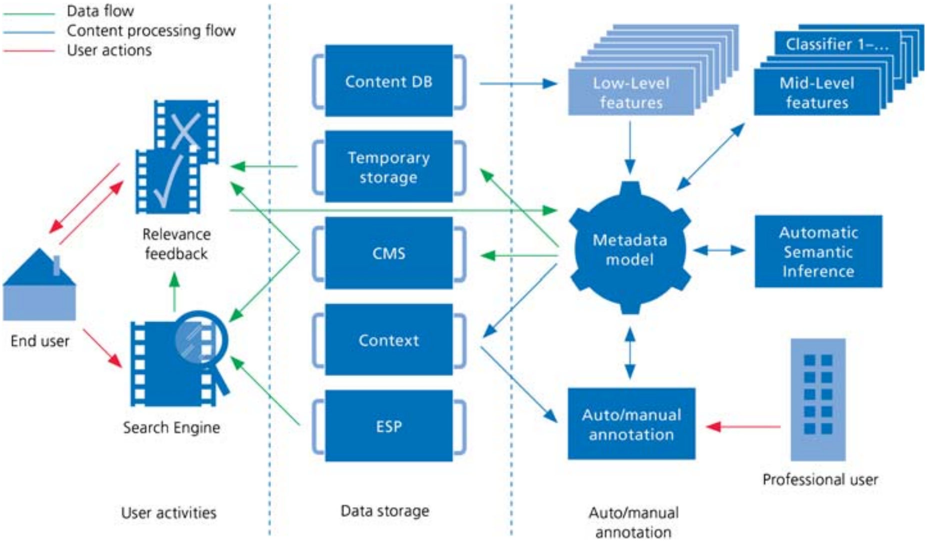

Fig. 1 The RUSHES system for annotation, indexing, search and retrieval

on the dynamic properties of the video content. The Basque broadcaster EiTB [11], involved in the project, provided the consortium at the beginning of the project with requirements definitions and user scenarios based on professional user evaluation. The technical work packages then derived a set of classifiers, which seems to be realistic being developed during the course of the project. The following classifiers have been developed and integrated in the RUSHES search engine:

- Vegetation classifier
- Classifier based on water segmentation
- Classifier based on regular shape detection
- Classifier based on text detection
- Classifier based on recognition of faces
- Classifier based on interview detection
- Music/speech classification
- Analysis of 3D camera motion in order to classify rotation, linear movement, zoom in/zoom out
- Analysis of 3D properties of the scene in order to classify flatness of a scene

In order to give an insight view to some key modules of the RUSHES system, we are presenting here the new approach on 3D scene structure analysis for automatic annotation and indexing. A complete overview of the full set of classifiers can be found in a detailed project deliverable [6] reporting the development of low level AV media processing and knowledge discovery.

The complete audio-visual analysis is implemented in a so-called CCR graph (CCR = content capture and refinement). Due to close relationship with the FP6 Integrated Project PHAROS and involvement of FAST in both projects, the RUSHES project is able to benefit from this collaboration regarding our integration activities. PHAROS is developing horizontal framework technologies for audio-visual search [23] and can provide a complete CCR framework. Hence, RUSHES can be considered as the first user of the technology developed in another FP6 project. This bilateral cooperation demonstrates the benefits of research and development on European level.

After automatic analysis of the video, the metadata model (MDM) will be generated as the fundamental database. This MDM is stored as a MEX file in the content management system (CMS) for the search later on. MEX is the schema of the metadata (annotations) used for the exchange of the metadata information among the different components of the RUSHES system. In addition to the automatic annotation, the professional user has the opportunity to add manual annotations to the content as well. The search is performed by the enterprise search platform (ESP) developed by FAST. This search engine provides the required functionalities in order to search quickly through the MEX file, which is the textual description of the complete video database by means of semantic key words and tags.

A novel user interface has been developed which allows the navigation through the large video repository in many different ways. At first, the videos themselves can be explored by advanced visualization of videos classified in a hierarchy which is automatically generated. Novel timeline zoom capabilities have been developed in order to access quickly the desired part of the video. Furthermore, key frames are available as well as static and dynamic video summaries for display of the video repository. The search is supported by relevance feedback capability in order to allow the user a refinement and re-ranking of the search results.

3 Novel techniques for multi-modal analysis of un-edited audio-visual content

In the following sections, two examples of novel audio-visual classifiers and techniques will be presented in more detail.

3.1 Multi-modal synchronisation

Multimedia synchronisation is performed widely using Synchronised Multimedia Integration Language (SMIL) which integrates streaming audio and video (images, text or any other type) [27]. SMIL allows the authors to use a text editor to write script codes for multimedia synchronisation and presentation e.g. <par> command synchronises audio and video by playing them at the common time line [32]. In addition, time-alignment method is used for synchronisation of multimedia files. In [9], time alignment method is used to synchronise the closed-caption with voice in a video clip. However, our interest is to synchronise outputs of classifiers in common timelines in a XML file as this will improve retrievability of multimedia that corresponds with user's demand. In this section, we propose new methods to synchronise outputs of classifiers in common time-lines.

Figure 2 shows the metadata synchronisation scheme that is designed to synchronise MEX file(s) produced by low/mid-level classifiers in order to improve searchability and search accuracy. In addition, during synchronisation, the system calculates and generates statistical reports (availability of faces, vegetation, etc.) in percentage format as well as the number of shots in the video clip. This report is used for a video content visualisation, search result categorisation and presentation of video content in a very comprehensive way by grouping the availability of items.

As seen in Fig. 2, classifiers analyse a video clip independently. Metadata synchronisation module retrieves classifiers results and synchronises the results by generating

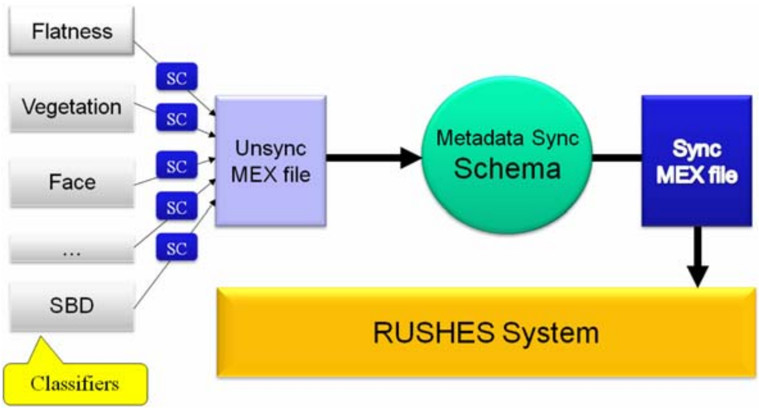

Fig. 2 Metadata synchronisation architecture

Table 1 General relationships of annotated and temporal segment

(1)	TSI	<=	ASI	&	TSE	>=	ASE				
(2)	TSI	<=	ASI	&	TSE	>	ASI	&	TSE	<	ASE
(3)	TSI	>	ASI	&	TSI	<	ASE	&	TSE	>=	ASE
(4)	TSI	>	ASI	&	TSE	<	ASE				

TSI temporal segment initial timestamp, *TSE* temporal segment end timestamp, *ASI* annotated segment initial timestamp, *ASE* annotated segment end timestamp

segment(s) based on (Section 3.1.1) static temporal segment, (Section 3.1.2) textual keyword and (Section 3.1.3) shot boundary detection.

Table 1 shows relationships of annotated segment and temporal segment (which is created during the synchronisation process). As seen, there are four possible conditions so all annotated segments will be synchronised that are true under these four conditions.

The temporal concept was applied for reasoning about actions. In [2] a formalism based on a temporal logic is proposed for reasoning about actions because it enabled to describe a much wider range of events/actions than other methods. The formalism was used to characterise events, processes, actions, and properties which can be described in English sentences. The difference is between the approach presented in [2] and the proposed method as this [2] applied temporal logic for reasoning actions e.g. past, present and future where the proposed method applies temporal logic to synchronise annotated segment(s) on common time-lines.

3.1.1 Metadata synchronisation based on static temporal segment

Metadata synchronisation based on static temporal segment synchronises a MEX file produced by classifiers by generating temporal segment(s) and grouping all operators output which is within the temporal segment time line. The temporal segment size needs to be specified when inputting unsynchronised document(s). In the example shown in Fig. 3, the temporal segment size fit to 30 s and as video length is 120 s, it

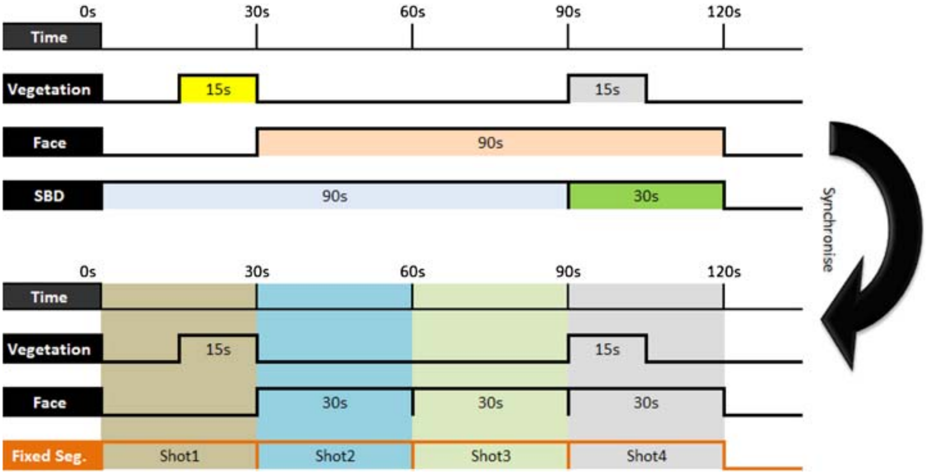

Fig. 3 Metadata synchronisation timeline based on fixed temporal segment

generates four fixed segments/shots even though shot boundary detection shows two segments/shots in this clip. This scheme is good when there is no need to consider any classifiers including shot boundary detection while MEX file(s) needs to be synchronised based on their common time lines.

3.1.2 Metadata synchronisation based on textual keyword

Metadata synchronisation based on textual keyword synchronises a MEX file by prioritising one classifier at a time. The architecture of the metadata synchronisation based on textual keyword is depicted in Fig. 4. This prioritisation is defined by the keyword which is provided at once, when a query is submitted. The prioritised classifier time line(s) (StartTimeStamp and EndTimeStamp) is used to generate temporal segment(s) for the synchronisation. Minimum temporal segment threshold is introduced to avoid very small video shots that has to be defined when a query is submitted e.g. 30 s. The prioritised classifier time line is used, if it is greater than the threshold time line otherwise threshold time line is used to generate temporal segment(s).

An example timeline, presented in Fig. 5, is showing the behaviour of this synchronisation scheme. This method plays a key role when search is carried out on a particular object e.g. face, vegetation, etc in which case this method can produce an optimum synchronisation. This method uses supervised keyword library and uses WordNET to expand the query as shown in Fig. 4.

3.1.3 Metadata synchronisation based on shot boundary detection

Metadata synchronisation based on shot boundary detection synchronises a MEX file produced by classifiers by generating dynamically temporal segment(s) using the shot boundary detection time line information. Therefore, this scheme is dependent on shot boundary detection operator. In the case of the classifier failing to detect a shot or generating a very lengthy shot, maximum threshold is introduced to handle this shot detection error and to avoid lengthy shots in order to improve on the search performance. The maximum threshold value has to be defined on a query submission.

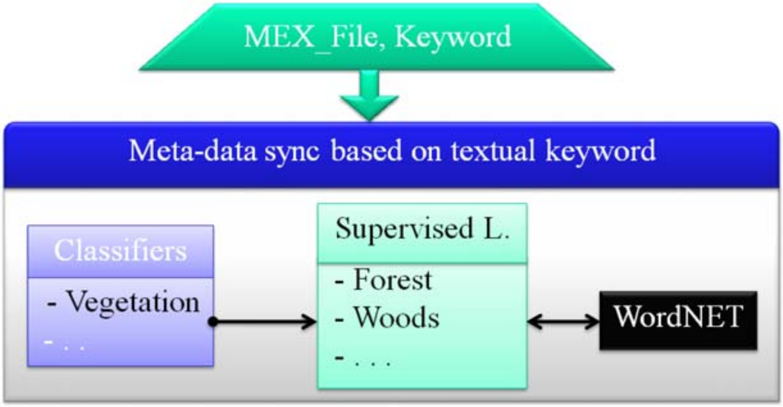

Fig. 4 Architecture of metadata synchronisation based on textual keyword

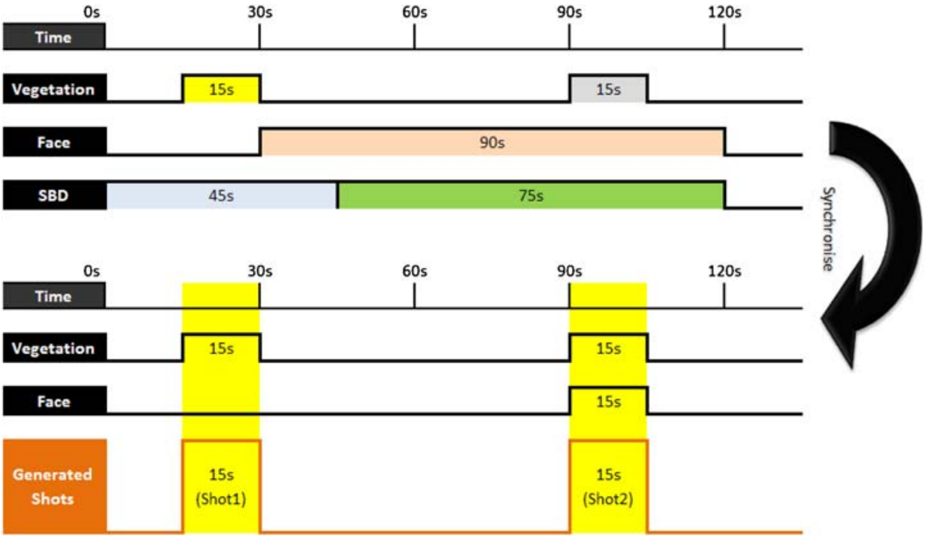

Fig. 5 Metadata synchronisation timeline based on a textual keyword

Figure 6 shows a graphical representation of synchronising the MEX file of a video clip which contains annotations (vegetation, face and shot boundary detection) and the synchronised version of the MEX file that has two shots generated from shot boundary detection time-lines and the annotations are fitted into these two shots.

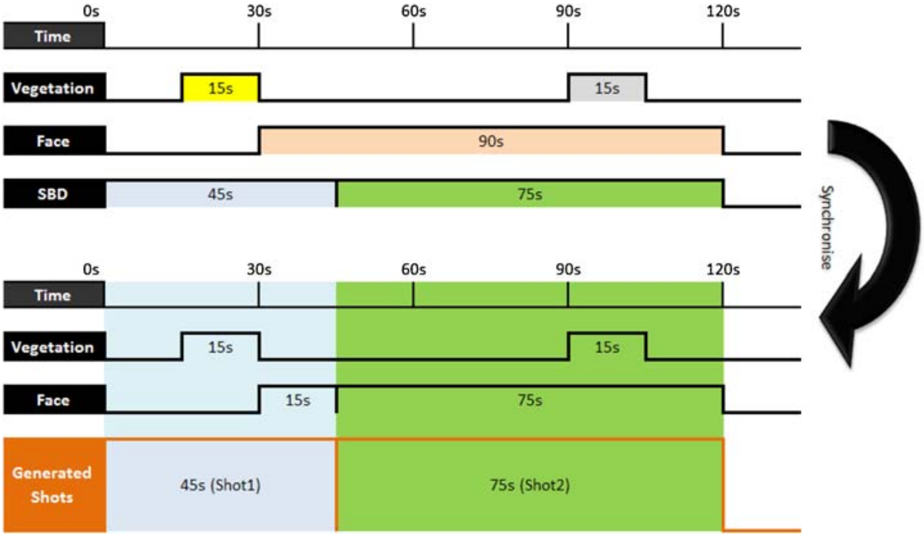

Fig. 6 Metadata synchronisation timeline based on shot boundary detection

3.1.4 Results

Metadata synchronisation based on shot boundary detection (proposed system) is tested on real data (unsynchronised MEX files) which was generated from the RUSHES video database using the RUSHES classifiers e.g. shot boundary detection, vegetation detection, etc. The system performance is stable on any number of complex annotations and video shots. Figure 7 shows the results indicating that as number of annotations and video length increases, the frame rate decreases. This is due to the lengthy shots with annotations; the system processes all shots and annotations for synchronisation purposes and generates statistical reports. As seen, this module does not affect the overall system performance. Standard video frame rate is 25 fps where this proposed system synchronises MEX files at a much higher frame rate as shown in Table 2 because it manipulates textual data (MEX file) not visual data (e.g. video image).

3.2 Analysis of 3D scene structure

One main objective of RUSHES was to develop novel semantic classifiers describing the spatio-temporal properties in an image sequence. Even from the professional users, interest has been received to provide classifiers, which allow distinguishing between different types of helicopters flights or global properties of landscapes such as hills, valleys, flat regions.

In the past, various motion descriptors were defined in the well known MPEG-7 standard [17]. Nevertheless, the exploitation of camera motion information for video search and retrieval applications is still very limited in the literature. In contrast, the estimation of scene structure and depth information based on a moving camera has received much attention [3, 28].

The aim of this section is to demonstrate the potential of scene structure based video annotation and retrieval. The general analysis chain for high-level 3D scene structure analysis and annotation is illustrated in Fig. 8.

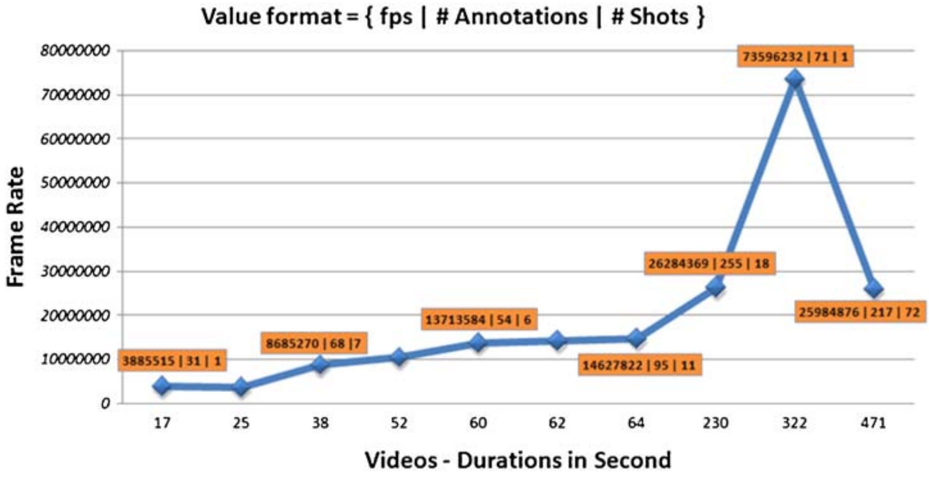

Fig. 7 Experiment results, generated from Table 2

Table 2 The experiment is conducted on MEX files generated from RUSHES video database

ID	Video MEX file	Sync-time (s)	Duration(s)	No. of shots	No. of annotations	Frame per second
1	MEX_Foot_318.xml	0.00011	17	1	31	3,885,515
2	MEX_PR_259_ManSailing.xml	0.00017	25	2	45	3,636,177
3	MEX_Foot_317.xml	0.00011	38	7	68	8,685,270
4	MEX_Row_743.xml	0.00013	52	7	104	10,399,468
5	MEX_Foot_321.xml	0.00011	60	6	54	13,713,584
6	MEX_AR_101_Getaria.xml	0.00011	62	1	6	14,170,781
7	MEX_Foot_308.xml	0.00011	64	11	95	14,627,822
8	MEX_AR_904_ SanSebastianFromHighBuilding.xml	0.00022	230	18	255	26,284,369
9	MEX_AR_102_Factory.xml	0.00011	322	1	71	73,596,232
10	MEX_AR_BBC.xml	0.00045	471	72	217	25,984,876
11	MEX_PR_610_ImagesOnBeach.xml	1.00028	697	37	653	17,420
12	MEX_PR_619_ FlyingBilbaoAndFootballStadium.xml	1.00069	1880	24	1,034	46,968
13	MEX_PR_623_ CrowdedSquareAndInterview.xml	45.51796	2838	77	2,951	1,559

We distinguish between three major component parts, the low level, medium level and high level scene structure extraction modules. In order to extract low level scene descriptors we apply a state of the art 2D feature tracker based on the well known KLT tracker [29]. Based on the properties of perspective projection of the features to the images it is possible to estimate the 3D camera path as well as the camera parameters for each image frame. We use the outcome of this module, i.e. focal length, 3D camera orientation, 3D camera position etc., to estimate the 3D correspondences of the 2D feature points. Further, the camera parameter information can be used to validate the tracked features in order to remove outliers and false detections. The

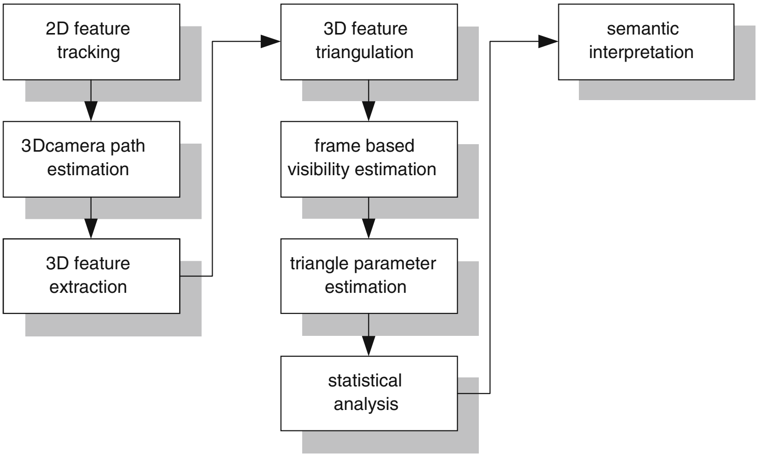

Fig. 8 Work flow of high-level scene description

result of the low level analysis is a sparse set of robustly estimated 3D features points as well as the 3D camera path and the camera parameter information (see Fig. 9).

The medium level analysis works on the sparse set of 3D feature points only. Its goal is to simplify the large set of 3D feature points in order to extract typical scene structure information. Note that the 3D feature points were obtained by the analysis of multiple frames. In this way, they are not restricted to single image frames. Rather, they reflect properties of the overall scene. In contrast, for video annotation of long unedited sequences, we are interested in local image based information rather than having a global overall set of scene parameters. Therefore, we need to find a way to model the properties of the given scene structure based on single image frames. To solve this problem, we propose a frame based visibility validation algorithm which relies on the triangulation of the 3D point parameter set. Triangles which are fully or partly occluded are removed (see Fig. 9). In a final step, the remaining triangle set will be analyzed. In order to describe the scene structure, we perform statistical analysis of the properties of the modelled triangles. In detail, we exploit the orientation of the normals of the triangles, the triangle area and their relative distance to the camera. To be more specific, for every triangle we observe the three angles enclosed by the coordinate planes and its normal. By denoting the resulting sets of angles with α for the set of angles enclosed by the normals and the YZ-plane, β for the XZ-plane and γ for the XY -plane we can perform a first statistical analysis of the 3D model. A final step is the high level semantic interpretation of the resulting data set. The extracted medium level data contain rich high level semantic information. For example, a 2D histogram analysis of all visible triangle normals can be applied. The result is a parameter for the type of scene structure which is visible in each of the frames. We will illustrate this for an example of the extraction of the *flatness* of a scene.

The high level analysis is illustrated in Fig. 10. A 3D histogram analysis of the angles α and β of the visible triangle normals, shown in Fig. 10, right was applied to the set of data. It can be seen from the sample image that there are three main scene orientations in the given image, i.e. the left-hand and right-hand hill as well as the main plain. In the 2D histogram in Fig. 10, these regions $b1$, $b2$, and $b3$ appear as significant clusters of normal orientations. Despite of the outliers and the sparse nature of the data set a robust high level scene description can be made based on this 2D histogram, i.e. the scene contains a valley. Please note that a simple scene interpretation model is used, i.e. it is distinguished between flat or non-flat scenes.

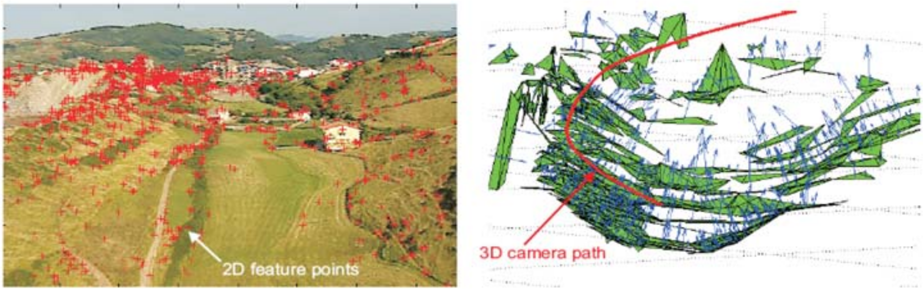

Fig. 9 Original image and visible 2D feature points (*left*), triangulated 3D feature points (*right*)

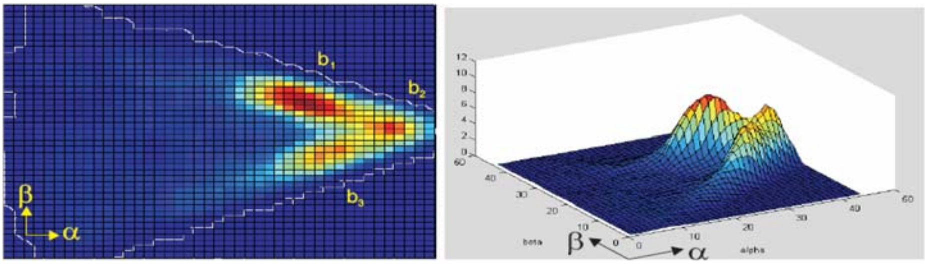

Fig. 10 2D histogram analysis of the angles α and β of the visible triangle normals, shown in Fig. 9

The high potential of our proposed method lies in more sophisticated analysis of the medium level features. One can easily see from the example in Fig. 10 that much more meaningful information can be extracted by including information about the distance of the triangles to the camera, or by grouping the triangles according to their histogram clusters and analyzing their corresponding 2D image position etc. Without loss of generality, we restricted our analysis to the detection of ‘flat-non flat’ scene parts.

The purpose was to develop a high-level semantic classifier which provides automatic annotation of a given scene by its ‘flatness’. In order to validate the efficiency of this approach, the weighted variance distribution of the visible scene triangle normals for a number of $N = 21$ bins has been calculated. In Fig. 11, the weighted variance of normal orientations, i.e. angles enclosed by normals and the main coordinate planes XY, XZ, YZ, as well as the mean value of the variance of all angles is shown along the complete sequence. The weighted variance is considered to be a very good indicator for the flatness of the scene.

It has to be noted that the ‘flatness’ is plotted as an inverse value, i.e. the scene is annotated to be very flat if the flatness value is low. The x-axis of the figure marks the frames of the sequence. It can be seen that the analyzed image sequence has two major flat parts at the beginning (frames 200–350) and the end (frames 800–1,000) of the scene. In order to classify and annotate finally individual segments of the video

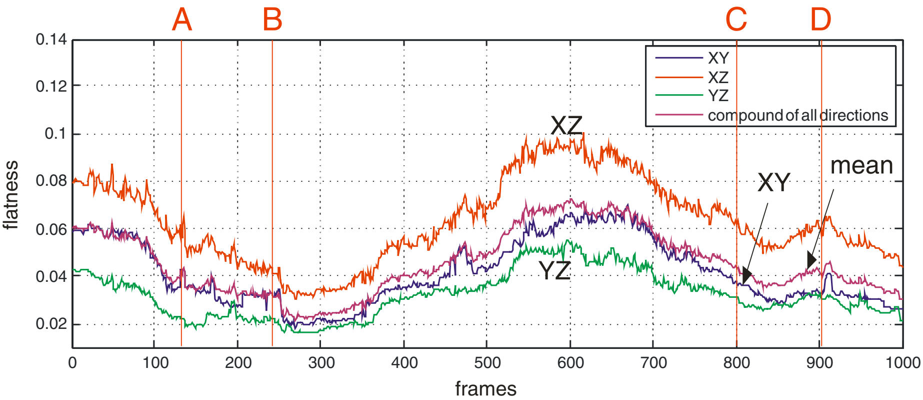

Fig. 11 Weighted variance of triangle normals

Table 3 Performance evaluation of the flatness classifier

	Macro average		Micro average	
	Precision	Recall	Precision	Recall
Flatness	66.41	56.78	77.61	73.46

into flat and non-flat scene parts, a simple threshold can be applied. The threshold has been defined heuristically by observation of flatness value and the scene itself. In the final performance evaluation of this approach, the automatic annotation has been compared to manually annotated ground truth data. In Table 3, the precision and recall is presented, whereas, we provide micro and macro averages for both performance values. Micro average denotes precision and recall for all frames at ones and macro average is the average of the outcomes for each video sequence. More details of the concept presented above can be found in [12].

4 New tools for visualization and browsing

The content of a large video database can only be browsed by means of the key frames, thus the problem of displaying and navigating through a large video database is a problem of visualisation of huge image/key frame collections. The challenge is the trade-off between the image size, so that the user can understand what is contained therein, and the amount of images that can be displayed simultaneously, so that a user needs the minimum necessary actions to understand the content and its organization, and find the desired items. In recent years there has been a boom of visualization mechanisms for displaying large collections of images, mainly exploiting hierarchical organisation of the investigated material. In information visualisation there is a number of techniques for visualising hierarchical structures, such as data mountains [25], hyperbolic tree [19] and 3D hyperbolic visualisation [21], treemaps [4, 18] and cone trees [14, 24].

However, these visualisation solutions do not always serve as efficient aids for the user, since the excess complexity of the user interface sometimes induces an additional obstacle for performing the browsing task. We have tried these approaches and found that they do not fulfil all our requirements.

Therefore, we have decided to develop two ad-hoc solutions within the RUSHES project. Since in the case of rushes, most material lies un-annotated in huge unorganised databases, and a semantic clustering is not always feasible, the first solution provides a tool for visually browsing the content. The second solution instead, deals with the case of semantic clustering and navigation of semantically annotated content.

4.1 Visual browsing tool

In the case of large video databases, it is helpful for browsing to structure the given material into a hierarchical structure, where each layer contains a complete partition of the database content and where each node contains a quick preview of key-frames highly representative of the visited content. The hierarchical summaries are obtained by a visual clustering of key-frames extracted from shots, where visual content is represented through a dictionary of visual words, as described in [5]. Even if the grouping of similar content is based on visual similarity rather than semantics, the proposed arrangement assists the browsing process by reducing the semantic gap between low-level features and high-level semantic concepts familiar to the user.

4.1.1 Navigation interface

The interface provides a visual navigation tool, forming a thread between the concealed structure of the content and the user's needs. A screenshot of the whole user interface (developed using Prefuse [15] and Java Swing) is shown in Fig. 12.

It consists of two windows which enable (1) the exploration of the content by interactive search and (2) the vision of rushes previews. In the upper window the hierarchical preview organisation is presented using a tree view, as better shown in Fig. 13.

The tree view is chosen since it is a common way of representing a hierarchical structure and its biggest advantage is that the majority of users are familiar with it. The presented tree contains the hierarchical preview and in case of large amount of content it might consist of a huge number of nodes and branches, where nodes are clusters of visually similar key-frames.

Due to the limited display dimensions, visualising the entire tree at any time is space-consuming and unnecessary. An adequate solution is to dynamically change

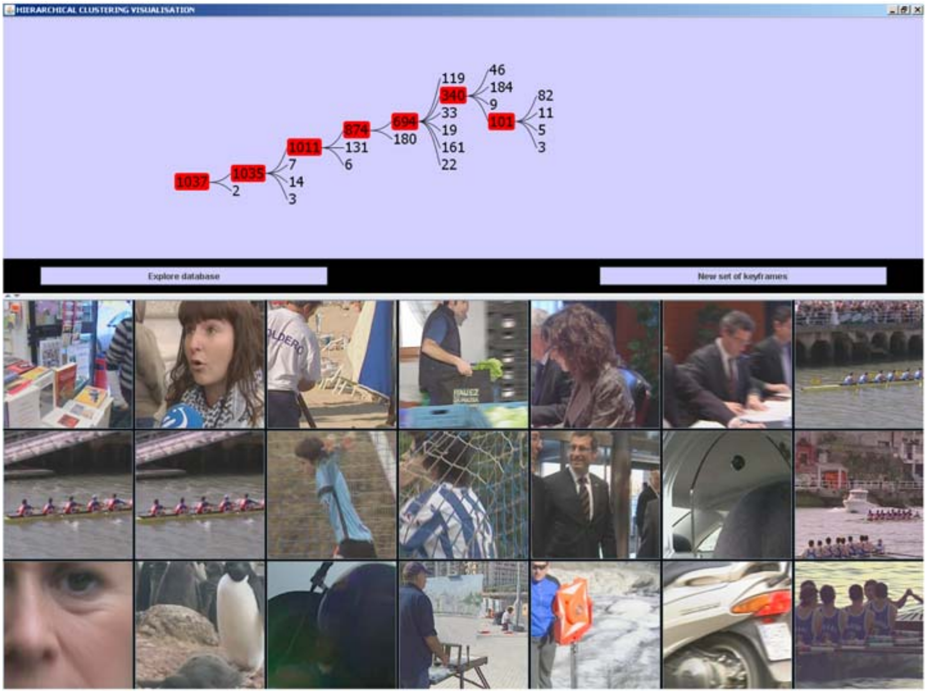

Fig. 12 Screenshot of the interface. In the *upper window* the tree view of the database structure and the two buttons, which allow the interactive exploration of the content. In the *bottom window*, the grid with node preview key-frames is displayed

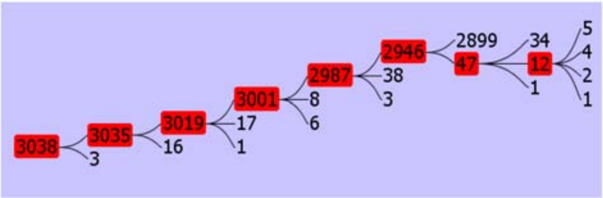

Fig. 13 Tree view of the database structure (*upper window*)

the number of displayed levels: a node can be expanded to reveal its child-elements, if any exist, or collapsed to hide departing branches and child-elements. In a similar way, the user does not see the entire tree while browsing, but only the set of nodes actually involved in his exploration. To highlight the user's search path, all nodes starting from the root to the currently explored node, are colour-encoded.

Considering that there is no meta-data on the observed rushes, and that the information we have is purely visual, the only property we visualise for each node is the number of key-frames contained in the node (see Fig. 13). For giving an insight into the node content, instead of putting the representative key-frame, we display the preview of nodes in the bottom window, since the content in a node can often be too diverse to be represented by a single key-frame (this is specifically true for nodes in the higher levels of the hierarchy).

When clicking on a certain node, the selected item is coloured in red, the next level child-elements are shown, and the related preview is accessed and shown in the lower window, as shown in Fig. 14.

The preview key-frames belonging to the currently visited node are displayed on a grid, where temporally close key-frames are placed sequentially in order to assist content understanding. Positioning the mouse over a key-frame gives the information about the name of the video it belongs to. In this way the user can distinguish if similar key-frames placed close together belong to the same data item.

In short, the tree metaphor in the upper window of the interface fulfils the following tasks:

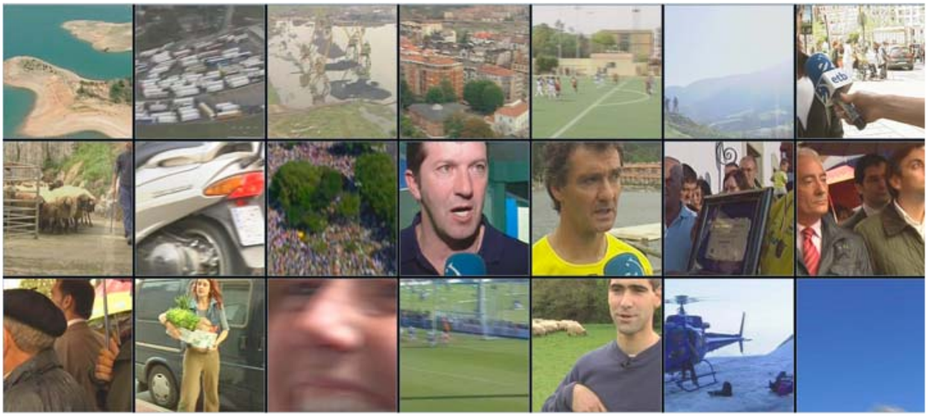

Fig. 14 Preview key-frames of the currently selected node (*lower window*)

- provides the visualisation overview at all times;
- describes the parent-child relations between visual content;
- provides the information on the number of key-frames in each node;
- facilitates the comprehension of the current position within the database by colour encoding;
- supports the user moving forward, backward and making progressive search refinements.

General guidelines that led the development of our solution are based on one main principle of information visualisation, known as “focus and context”: the user preserves the global perspective by seeing the database structure and his current position at each step of the search, while getting more information by observing at the same time the visual summary of the selected node.

4.1.2 Access methods

To perform an efficient exploration the user can click on any node of the hierarchy and visualise its content preview. Then by *sequential access* the user can move backward and/or forward through the tree to refine his search, while constantly being aware of the current position inside the database.

However, during the initial stage of exploration, even a professional user might be completely unaware of which direction to take to locate relevant content. For example, when using traditional navigation tools, in case an interesting key-frame is not presented at first, we observed that users often perform some “random” attempts of exploration, in order to look for something that might better suit their queries.

The specificity of the random exploration mainly lies in the fact that other state-of-the-art exploratory tools do not deal with such a repository in the early unstructured part of the browsing task, i.e., when the rushes content is still unknown to the user (see for example [1, 31] and [7]). The proposed novel random access schema aims at reducing the time for browsing initialisation and content grasping by statistically modelling the probability to access collections of hierarchically arranged previews. This system functionality, called *random exploration*, imitates in a way the user random behaviour in the situation when displayed key-frames are of no interest for the user and he wants to move on. In this case the application randomly selects another node and visualises its summary, thus opening a new search direction.

The random browsing strategy is modelled by a statistical law, whose density function represents the probability to access one node of the hierarchy and display its summary. When the user selects this navigation modality, the algorithm randomly selects one level of the hierarchy with uniform probability; then, inside the chosen level, the probability of accessing one specific node is shaped by the distribution of data in that level of the database. In particular, in the current implementation, the probability of selecting a node inside one level is proportional to the colour entropy of the node itself (computed through a vector quantization process at the node level, as in [5]), so that more informative nodes are more likely to be chosen with respect to less informative ones. In future implementations, when shaping the access probability, we aim to integrate some user profile data, for example by including information related to user’s profile and browsing history.

By modelling the node access by such a statistical method, we expect to reduce the expected time of browsing needed by the user to find his/her search goals. However, if the shown visual preview is of no interest for the user, another node can be randomly selected, and a different content set at a different level of the hierarchy is shown.

This random walk can continue until the user is able to find an interesting key-frame and decides to follow the new information scent, for example using a sequential access to nodes in order to visualise content previews.

Both random and sequential explorations are assisted by the display of the visual content previews in the bottom application window (as in Fig. 12). The aim of the visual summary is to show a set of representative key-frames for each node in the hierarchy. However since there are no semantic labels that could assist us in defining the most appropriate representative set for the selected cluster, we adopt a similar approach to the one presented in [8], and we randomly extract from each node the set of key-frames to be displayed. In case the user wants to perform further exploration on the same node, he can request for additional content, and a new random set is then extracted from the node and displayed.

4.2 Semantic browsing tool

Regarding the solution based on semantic browsing, firstly the user can navigate through the hierarchy that is the result of the designed automatic organization process. Secondly, the visualization solution allows for quick understanding by the user of the content represented by the displayed images. This means a minimum size and quality of images and appropriate mechanisms navigating through the structure. The previous mentioned approaches only allow the display of "flat" structures, as opposite to hierarchical solutions, and very often impose difficulties for the second objective. We have decided to use Adobe Flex as development technology, mainly because of the advantage of creating engaging web applications offered by this popular platform.

First of all, RUSHES browsing interface addresses the challenge of meaningfully grouping the images by proposing an automatic classification algorithm, based on a hybrid cluster analysis solution, that effectively classifies the key frames based on the co-occurrence of semantic concepts or annotations. Cluster analysis is a well-defined field that is applied for the organization of a collection of patterns (feature vectors) into groups (or clusters) based on some similarity metric. Cluster algorithms are divided into partitional algorithms, which provide a single division of the pattern space into groups (the best known of these being the k-means algorithm) and hierarchical algorithms, which provide a sequence of nested partitions.

The standard hierarchical algorithms produce a binary tree, in which each parent holds exactly two children. This disposition is impractical for browsing huge media repositories since the number of levels in the hierarchy would be too high in big databases. We need to create more populated clusters. This target is easily reached by means of partitional clustering, but the well-known partitional clustering algorithms [10] produce only flat partitions (although extensions have been proposed) and have the additional problem of determining the right number of clusters.

Our proposed solution [20] takes advantage of both hierarchical and partitional clustering algorithms, forcing them to work together for better results. The proposed clustering algorithm begins with a hierarchical processing of the whole set of elements, which is used to obtain a target number of clusters k by means of a parameter that represents the magnitude of the gap between two successive steps in the hierarchical clustering process. As this value grows, the difference between the closest nodes, chosen in each step, becomes higher. Once this value is computed, the same group of data is structured according to a partitional clustering, and therefore divided in *k* clusters. This process ends when every node is clustered into one single parent node.

For testing and demonstration purposes the set of 3,064 key frames have been extracted and manually annotated with 21 concepts. They were input to the clustering algorithm previous to the display in the navigation tool.

The browsing UI proposes an easy-to-use mechanism to navigate through the hierarchy where every cluster is represented by its centroid image and its descriptive concepts. The screenshots depicted in Figs. 15 and 16 show this browsing tool.

The user is enabled to explore the database using any of the three windows:

- In the top window the selected cluster (in bigger size) together with the sibling clusters and the parent one are shown. The user is enabled to navigate the repository by clicking on the corresponding image so he/she is able to get back to the previous branch (parent cluster) or to sibling branches. When the tool starts, the top level of the hierarchy is shown.
- Central window shows the content of the selected cluster and its descriptive concepts. As mentioned before, every cluster is represented by a key-frame, with a "+" icon whenever the cluster contains other child branches or collections of images. The tool also provides zooming capabilities (in and out) of a final key-frame and gives access to other RUSHES tools enabling the user to annotate the video the key-frame belongs to or to play the video or a summary of it. In the lower part of the window clusters already visited by the user are represented in order to help the user with the browsing history.
- Finally, the left window shows the tree representation of the repository where information of every cluster is depicted. User can browse the database using this tree which can be shown and hidden under user's request. The tree is intended to show the user the level in the hierarchy he/she is browsing.

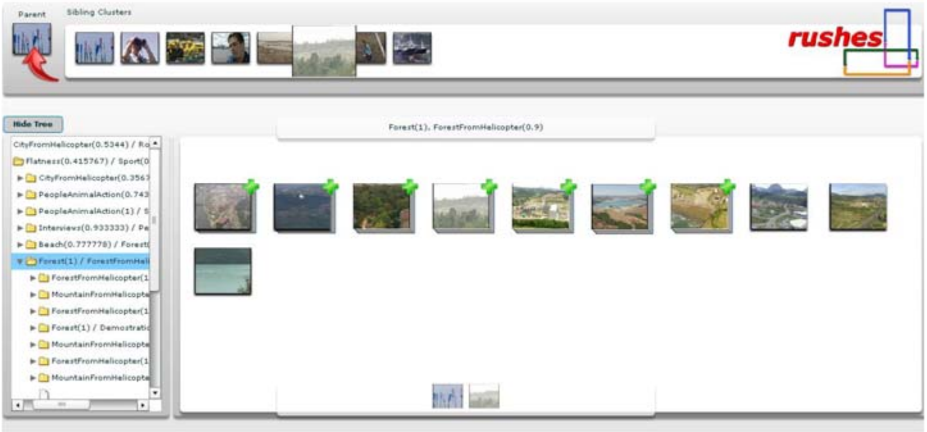
Parent
Sibling Clusters
rushes
Hide Tree
Forest(1), ForestFromHelicopter(0.9)
CityFromHelicopter(0.5344) / Ro
Flatness(0.415767) / Sport(0
CityFromHelicopter(0.3567
PeopleAnimalAction(0.743
PeopleAnimalAction(1) / S
Interviews(0.933333) / Pe
Beach(0.777778) / Forest(
Forest(1) / ForestFromHeli
ForestFromHelicopter(1
MountainFromHelicopte
ForestFromHelicopter(1
Forest(1) / Demostratio
MountainFromHelicopte
ForestFromHelicopter(1
MountainFromHelicopte

Fig. 15 Navigation tool—showing clusters within the branch *Forest—Forest from helicopter*

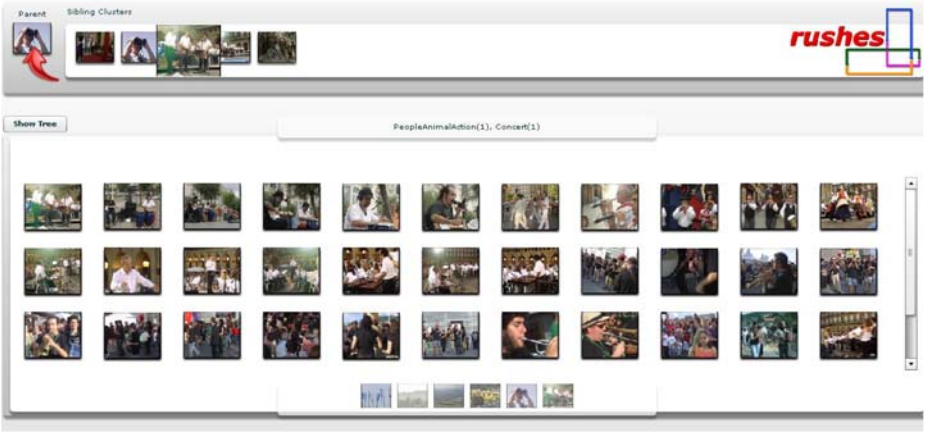

Fig. 16 Browsing interface showing key-frames annotated with *Concert*

4.3 User evaluation of both visual and semantic browsing tools

4.3.1 Initial evaluation of the random exploration functionality in the visual browsing tool

For testing and demonstration purposes of the novel random exploration method, a set of 134 raw videos of about 14 h length in total have been provided by EiTB. The videos belong to several domains (e.g., interview, football, aerial views, rowing, etc.) and from these videos a total set of 3,064 key frames has been extracted.

Evaluation was based on two professional use-cases defined in [13] and the initial usability tests have been performed by five journalists from the main Basque broadcaster.

After performing the specified tasks, EiTB journalists were asked to fill a questionnaire for rating their satisfaction with the most important aspects of the proposed solution, to state positive and negative aspects of the application, and to give personal comments on potential improvements. As a main outcome of the evaluation process (see [16] for further details), the navigation tool with random access exploration has been highly appreciated by the journalists as a useful tool for browsing, especially when they did not know where to find the desired content. During the evaluation process we also analysed different behavioural patterns among different users that will be taken into account for further improvements of the visual browsing tool.

Figure 17 shows the results of the user evaluation where scale from 1 to 5 is used for stating the level of agreement (maximum 5 and minimum1) with the statements given in the questionnaire. We can see that the questions Q10—"The key-frames displayed in the bottom window provide a good overview of the node content" and Q11—"Random exploration is helpful for browsing when I do not know where to find the desired content" demonstrate a positive initial evaluation regarding the proposed random exploration and content preview display method. The colour encoding of the nodes was also highly marked as a useful feature (Q13) and the visual browsing application was pleasant to use (Q14). The low average mark was

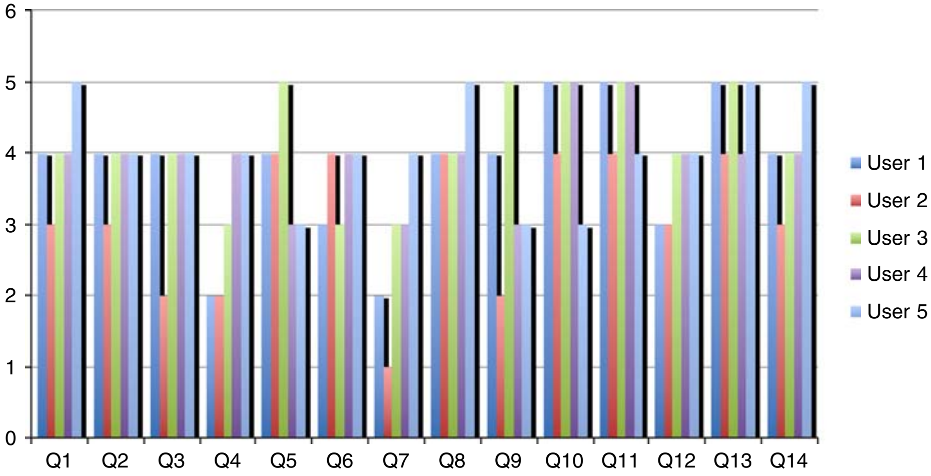

Fig. 17 Ratings per user per question for evaluating user satisfaction with the visual browsing tool application

received for the intuitiveness of the interface (Q7) and further improvements will be performed in order to make the users interaction with the application easier.

Users also provided positive and negative comments regarding the application. They were comfortable with using the tool, liked the idea of random browsing for the content exploration and considered that the key-frames displayed in the grid is a good content representation method. As additional requirements, users suggested to add more information about the clip (name, duration, date) and to add ToolTips to buttons in order to explain their functionalities.

4.3.2 Final usability trial of both browsing tools

As part of the final usability trials, 49 users were recruited in seven of the project partners' sites. These were a posteriori divided in two groups, according to their expertise with multimedia analysis systems. The lower expertise group was composed of 30 persons, and the higher expertise group was composed of 19 persons. The participants responded to a questionnaire with a rating response scale of nine points (1 to 9).

The variance analysis of the results using as a between-subjects factor in our definition gives significant differences between both groups in most of the questionnaire items for both semantic and visual browsing tools. In average, the higher expertise group rated the tools 1 point higher than the lower expertise group. Average rating for semantic browsing tool was 6 points in the lower expertise versus 7 points for the higher expertise. And for visual browsing average for the lower expertise group was 6.12, versus 7.6 for the higher expertise group. We have concluded that higher expertise rated slightly higher the visual browsing tool in comparison with the other RUSHES module.

One important result we obtained is that both groups considered positively the capabilities of automatic annotation provided by RUSHES, while the lower expertise group rated much lower the need to manually annotate or correct the annotations, in comparison with the high expertise group.

5 Conclusion

After two years research and development within the FP6 project RUSHES, a set of results have been presented showing the scientific and technological strength of the consortium on audio-visual analysis, indexing search and retrieval of un-edited raw multimedia assets. During the development of the RUSHES search engine and the integration of all the components, a new level of bilateral cooperation between the integrated project PHAROS and the STREP project RUSHES has been achieved. The open and distributed architecture of PHAROS could be successfully used in the RUSHES system by integrating the CCR framework. Furthermore, two novel approaches for multi-modal analysis of raw un-edited audio-visual data have been presented. These approaches have to be considered as examples showing the high scientific level within the consortium. Finally, a set of tools for navigation and browsing of large video repositories conclude the paper. The RUSHES search engine has been demonstrated successfully to the public at the CEBIT (largest industrial computer and telecommunication fair of the world) in Hannover in March 2009.

Acknowledgements This work is a result of the FP6 project “RUSHES” Proposal no.: FP6-045189, which is funded by the European Commission. We would also like to thank Leticia Fuentes Ardeo, Mikel Frutos Hernandez and journalists at EiTB for priceless help during the experimental evaluation.

References

1. Adcock J, Cooper M, Pickens J (2008) Experiments in interactive video search by addition and subtraction. In: CIVR’08: proceedings of the 2008 international conference on content-based image and video retrieval. ACM, New York, NY, USA, pp 465–474
2. Allen JF (1984) Towards a general theory of action and time. Artif Intell 23(2):123–154
3. Beardsley PA, Torr PHS, Zisserman A (1996) 3D model acquisition from extended image sequences. In: ECCV’96: proceedings of the 4th European conference on computer vision-volume II. Springer, London, UK, pp 683–695
4. Bederson B (2001) Photomesa: a zoomable image browser using quantum treemaps and bubble maps. In: Proceedings of the 14th annual ACM symposium on user interface software and technology, pp 71–80
5. Benini S, Bianchetti A, Leonardi R, Migliorati P (2006) Extraction of significant video summaries by dendrogram analysis. In: Proceedings of the international conference on image processing, ICIP’06. Atlanta, GA, USA, 8–11 October
6. Benini S et al (2009) D21 report on final development of low level AV media processing and knowledge discovery, 2009. RUSHES Project, FP6-045189, Deliverable D21, WP2
7. Benmokhtar R, Dumont E, Merialdo B, Huet B (2006) Eurecom in trecvid 2006: high level features extractions and rushes study. In: TrecVid 2006, 10th international workshop on video retrieval evaluation, November 2006, Gaithersburg, USA
8. Borth D, Schulze C, Ulges A, Breuel TM (2008) Navidgator—similarity based browsing for image and video databases. In: KI’08: proceedings of the 31st annual german conference on advances in artificial intelligence. Springer, Berlin, pp 22–29
9. Cho J, Jeong S, Choi BU (2004) Automatic classification and skimming of articles in a news video using Korean closed-caption. In: Gelbukh AF (ed) Computational linguistics and intelligent text processing. Lecture notes in computer science, vol 2945. Springer, Berlin, pp 498–501
10. Duda RO, Hart PE, Stork DG (2001) Pattern classification, 2nd edn. Wiley, New York

11. EiTB. Euskal irrati telebista. http://www.eitb.com/
12. Feldmann I, Waizenegger W, Schreer O (2008) Extraction of 3D scene structure for semantic annotation and retrieval of unedited video. In: IEEE 10th workshop on multimedia signal processing, pp 82–87
13. Fuentes Ardeo L et al (2008) Requirement analysis and use-cases definition for professional content creators or providers and home-users. RUSHES Project, FP6-045189, Deliverable D5, WP1
14. Hearst MA, Karadi C (1997) Cat-a-cone: an interactive interface for specifying searches and viewing retrieval results using a large category hierarchy. In: Proceedings of the 20th annual international ACM SIGIR conference on research and development in information retrieval, pp 246–255
15. Heer J, Card SK, Landay JA (2005) Prefuse: a toolkit for interactive information visualization. In: CHI'05: Proceeding of the SIGCHI conference on human factors in computing systems. ACM, New York, NY, USA, pp 421–430
16. Janjusevic T, Benini S, Izquierdo E, Leonardi R (2009) Random assisted browsing of Rushes archives. J Multimedia (in press)
17. Jeannin S, Divakaran A (2001) Mpeg-7 visual motion descriptors. IEEE Trans Circuits Syst Video Technol 11(6):720–724
18. Johnson B, Shneiderman B (1991) Tree-maps: a space-filling approach to the visualization of hierarchical information structures. In: Proceedings of the IEEE conference on visualization. IEEE Computer Society Press, pp 284–291
19. Lamping J, Rao R (1994) Laying out and visualizing large trees using a hyperbolic space. In: Proceedings of the 7th ACM symposium on user interface software and technology. ACM, pp 13–14
20. Lozano A, Villegas P (2007) Recursive partitional hierarchical clustering for navigation in large media databases. In: Eighth international workshop on image analysis for multimedia interactive services, WIAMIS 2007. Santorini, Greece, 6–8 June
21. Munzner T (1998) Exploring large graphs in 3D hyperbolic space. IEEE Comput Graph Appl 18(4):18–23
22. Over P, Smeaton AF, Awad G (2008) The trecvid 2008 bbc rushes summarization evaluation. In: TVS'08: Proceedings of the 2nd ACM TRECVid video summarization workshop. ACM, New York, NY, USA, pp 1–20
23. PHAROS IST-45035. Platform for searching of audiovisual resources across online spaces. http://www.pharos-audiovisual-search.eu
24. Robertson GG, Mackinlay JD, Card SK (1991) Cone trees: animated 3D visualizations of hierarchical information. In: Proceedings of the SIGCHI conference on human factors in computing systems: reaching through technology, pp 189–194
25. Robertson GG, Czerwinski M, Larson K, Robbins DC, Thiel D, Dantzich MV (1998) Data mountain: using spatial memory for document management. In: Proceedings of the 11th annual ACM symposium on user interface software and technology, pp 153–162
26. RUSHES FP6-045189. Retrieval of multimedia semantic units for enhanced reusability. http://www.rushes-project.eu
27. Rutledge L, Hardman L, van Ossenbruggen J (1999) The use of SMIL: multimedia research currently applied on a global scale. In: Modeling multimedia information and systems conference, pp 1–17
28. Shade J, Gortler S, He L-W, Szeliski R (1998) Layered depth images. In: SIGGRAPH'98: proceedings of the 25th annual conference on computer graphics and interactive techniques. ACM, New York, NY, USA, pp 231–242
29. Shi J, Tomasi C (1994) Good features to track. In: 1994 IEEE conference on computer vision and pattern recognition (CVPR'94), pp 593–600
30. Snoek CGM, Worring M (2005) Multimodal video indexing: a review of the state-of-the-art. Multimedia Tools and Applications 25(1):5–35
31. Villa R, Gildea N, Jose JM (2008) Facetbrowser: a user interface for complex search tasks. In: El-Saddik A, Vuong S, Griwodz C, Del Bimbo A, Candan KS, Jaimes A (eds) Proceedings of the international conference on multimedia. ACM, pp 489–498
32. W3C. Synchronized multimedia integration language. World wide web consortium—web standards. http://www.w3.org/TR/REC-smil/

Oliver Schreer graduated in Electronics and Electrical Engineering and received his Dr. -Ing. degree in electrical engineering at the Technical University of Berlin in 1993 and 1999, respectively. Since August 1998, he is working as project leader of the Immersive Media & 3D Video Group in the Image Processing Department of Heinrich-Hertz-Institute. In this context he is engaged in research for 3D analysis, novel view synthesis, real-time video conferencing systems and immersive TV applications. From 2000 to 2003, he was the responsible person for the European IST-project VIRTUE at HHI. Since 2001, he is Adjunct Professor at the Faculty of Electrical Engineering and Computer Science, Technical University Berlin. Since November 2006, he is Assistant Professor (Privatdozent) at Institute of Computer Engineering and Microelectronics in the Computer Vision and Remote Sensing Group. Since 2007, he is project manager of the European FP6 project RUSHES on "Retrieval of multimedia Semantic units for enhanced reusability".

Ingo Feldmann is working as project leader of the Immersive Media & 3D Video-Group in the Image Processing Department. He received his Dipl. -Ing. degree in Electrical Engineering from the Technical University of Berlin in 2000 respectively. Since September 2000 he is with the IP department, where he is engaged in several research activities in the field of 2D image processing, 3D scene reconstruction and modelling, digital cinema, multi-view projection systems, real-time 3D video conferencing systems and immersive TV applications. He was involved in different German and European projects which were related to these topics, like ATTEST, VIRTUE, ITI, Tsdk, Prime, Rushes, and 3DPresence. He was involved into several contributions for the MPEG 3DAV ad hoc group.

Isabel Alonso Mediavilla received her degree in Telecommunication Engineering from the Universidad de Valladolid in 2004. She started her career in Telefonica R&D in 2002, initially with a scholarship and eventually in 2004 as a Researcher. She has been working in the area of communications in real-time and metainformation and involved in some research projects in the IST and VI Framework programmes, such as AKOGRIMO, NM2 and, in particular, RUSHES.

Pedro Concejero is Doctor (PhD) in Psychology, his dissertation deals with application of ROC curve methods for detection in marketing research systems. After a period as research fellow and associate professor in Universidad Complutense, he joins Telefonica R&D. He has worked for a long time in usability and human factors research. He is currently focused on research projects in the IST and VI Framework programmes, in particular MESH and RUSHES.

Abdul H. Sadka is the Head of Electronic and Computer Engineering and the director of the centre for Media Communications Research at Brunel with 15-years experience in academic leadership and excellence. He is an internationally renowned expert in visual media processing and communications with an extensive track record of scientific achievements and peer recognised research excellence. He has managed so far to attract over 2M GBP worth of research grants and contracts in his capacity as principal investigator. He has been the coordinator and chair of executive board of a large EC funded Network of Excellence "VISNET" on Networked Audio-visual Media Technologies. He has published widely in international journals and conferences and is the author of a highly regarded book on Compressed Video Communications published by Wiley in 2002. He holds three patents in the video transport and compression area. He acts as scientific advisor and consultant to several key companies in the international Telecommunications sector and is the founder and managing director of VIDCOM Ltd.

Mohammad Rafiq Swash graduated with a first class honours degree in Computer System Engineering from Brunel University in 2008. Mr. Swash is a recipient of a UG University Prize for the Best Final Year Project, Granham Hawkes Prize and Brunel University Modal in 2008. Mr. Swash worked for Global Betbrokers as a Software engineer and software project development leader for 3 years. Also Mr. Swash is a senior committee member of ASA UK and a member of IEEE and IET. Mr. Swash is currently pursuing his PhD programme in the Centre for Media Communications Research (CMCR) at Brunel University under Professor A. H. Sadkas supervision and his current research interests include automatic video image annotation and retrieval.

Sergio Benini was born in Verona, Italy. He's received his MS degree in Electronic Engineering (cum laude) at the University of Brescia in 2000 with a thesis which won a prize granted by Italian Academy of Science. Between May 2001 and May 2003 he's been working in Siemens Mobile Communication R&D, on mobile network management projects. He received his PhD degree in Information Engineering from the University of Brescia in 2006, working on video content analysis topics. During his Ph.D. studies, between September 2003 and September 2004 he has conducted a placement in British Telecom Research, Ipswich, U.K. working in the "Content & Coding Lab". He is currently an Assistant Professor in the Telecommunications group of DEA at the University of Brescia, Italy.

Riccardo Leonardi has obtained his diploma (1984) and PhD (1987) degrees in Electrical Engineering from the Swiss Federal Institute of Technology in Lausanne. He spent 1 year (1987–88) as a post-doctoral fellow with the Information Research Laboratory at the University of California, Santa Barbara (USA). From 1988 to 1991, he was a Member of Technical Staff at AT&T Bell Laboratories, performing research activities on image communication systems. In 1991, he returned briefly to the Swiss Federal Institute of Technology in Lausanne to coordinate the research activities of the Signal Processing Laboratory. Since February 1992, he has been appointed at the University of Brescia to lead research and teaching in the field of telecommunication. His main research interests cover the field of digital signal processing applications, with a specific expertise on visual communications, and content-based analysis of audio-visual information. He has published more than 100 papers on these topics. Since 1997, he acts also as an evaluator and auditor for the European Union IST and COST programmes.

Tijana Janjusevic received her MS degree in Department for Telecommunications, University of Belgrade, Serbia in 2005. She is currently a PhD candidate at Multimedia and Vision Group (MMV), Queen Mary, University of London, UK. Her research interests include information visualisation and user interfaces for visual data mining.

Ebroul Izquierdo is Chair of Multimedia and Computer Vision and head of the Multimedia and Vision Group at Queen Mary, University of London. For his thesis on the numerical approximation of algebraic-differential equations, he received the Dr. Rerun Naturalium (PhD) from the Humboldt University, Berlin, Germany, in 1993. From 1990 to 1992 he was a teaching assistant at the department of applied mathematics, Technical University Berlin. From 1993 to 1997 he was with the Heinrich-Hertz Institute for Communication Technology (HHI), Berlin, Germany, as associated and senior researcher. From 1998 to 1999 Dr. Izquierdo was with the Department of Electronic Systems Engineering of the University of Essex, as a senior research officer. Since 2000 he has been with the Electronic Engineering department, Queen Mary, University of London. He is a Chartered Engineer, a Fellow member of The Institution of Engineering and Technology (IET), a senior member of the IEEE, a member of the British Machine Vision Association and was acting chairman of the Visual Information Engineering professional network of the IET. He is member of the programme committee of several international conferences. He is an associate editor of the IEEE Transactions on Circuits and Systems for Video Technology (TCSVT). He has served as guest editor of three special issues of the IEEE TCSVT, three special issues of the journal Signal Processing: Image Communication and three special issue of the EURASIP Journal on Applied Signal Processing. He has published over 300 technical papers including book chapters.