MULTIMODAL SPACE FOR RUSHES REPRESENTATION AND RETRIEVAL

Sergio Benini, Luca Canini, Pierangelo Migliorati and Riccardo Leonardi

DEA-SCL, University of Brescia, Via Branze 38, 25123, Brescia, Italy
Email: {*firstname.lastname*}*@ing.unibs.it*

ABSTRACT

In the field of video content analysis, growing research effort aims at characterising a specific type of unedited content, called *rushes*. This raw material, used by broadcasters and film studios for editing video programmes, usually lies unannotated in a huge database. In this work we aim at retrieving a desired type of rush by representing the whole database content in a multimodal space. Each rush content is mapped into a trajectory whose coordinates are connected to multimodal features and filming techniques used by cameramen while shooting. The trajectory evolution over time provides a strong characterization of the video, so that different types of rushes are located into different regions of the multimodal space. The ability of this tool has been tested by retrieving similar rushes from a large database provided by EiTB, the Basque Country main broadcaster.

Index Terms— Retrieval, Multimodal Analysis, Rushes

1. INTRODUCTION

In the broadcasting and film-making industries, *rushes* is a term for indicating raw footage used to generate the final productions such as TV programmes and movies. Rushes are potentially very reusable video content but are largely unexploited because only few people in the production team know what rushes contain and metadata with annotations are generally very limited. Therefore a growing research effort is aiming at developing techniques for structuring, indexing and retrieving rushes. For example, in the context of European funded research, the FP6 project RUSHES [1] is focusing on automatic semantic annotation, indexing and retrieval for the reuse of such raw and unedited audio-visual content in a media professional environment.

Since only a small portion of the rushes is actually used in the final productions, 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 [2] mainly deal with rushes summarisation, believing that this might also help other tasks, such as search and retrieval. However, we 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 [3], efficient retrieving from large video archives depends on the availability of indexes, and effective indexing requires a multimodal approach in which different modalities (auditory, visual, etc.) are used in collaborative fashion.

In contrast to edited videos, rushes are often characterised by unorganised structure, limited editing, the presence of redundant content, and are mainly accompanied with natural sounds and few or no on-screen texts. For this reason, retrieval techniques presented so far are mainly based on visual information as other modalities are sometimes absent or difficult to obtain. In [4] the authors index the rushes shots by "visual words" which are related to colour, texture and the combination of the two. In [5] the same features extracted from each keyframe are considered, as well as the color, texture and shape of the semi-automatically segmented objects. The approach in [6] instead takes into account motion features only. By analysing motion sequential patterns, the proposed two-level hierarchical HMM is capable of mapping low-level motion features into high-level semantic concepts.

Even if the limited presence or the absence of some of the traditional information channels could discourage a multimodal approach, this limited availability may be considered by itself as an useful information for retrieving similar rushes.

Therefore, in this paper we propose a novel approach for characterising the "multimodal identity" of a single rush and for retrieving similar footage from professional archives. To do this, we represent rushes into a space which is similar to those used for defining the identity of design objects [7]. In order to build this space we investigate the multimodal low-level features of the rush content and the filming techniques adopted by the cameramen while shooting. A single rush is then mapped into a geometrical trajectory, whose evolution over time provides a strong characterization of the investigated material. As a result, different types of rushes occupy different regions of the "multimodal space". Since it has been observed by many authors [8] that a temporal continuity of low-level features related to chromatic composition, audio and motion usually implies a persistent semantics, in the experimental part this analysis space is used for retrieving similar rushes from a professional broadcaster database.

This document is organized as follows. Section 2 describes the investigated type of data. In Section 3 the multimodal space is presented, while the characterization of its axes is described in Section 4. Section 5 discusses how a rush can be represented in the given space by a trajectory or by a geometric solid which summarises the trajectory characteristics. In Section 6 our framework is tested for retrieving similar rushes. Conclusions are finally drawn in Section 7.

2. RUSHES DATA

Different types of rushes footage are used by broadcasters to build documentaries or news programs, or by production companies to edit movies. By analysing the material provided by the RUSHES partner EiTB [9], we identified three main different categories of rushes: news footage, rushes for documentaries and raw material for comedies or sit-coms.

News footage can contain any type of audio-visual content, ranging for example from interviews to different kind of sports (Figure 1).

Fig. 1: Two frames extracted from news footage: a journalist (left) and a football match (right).

Rushes for documentaries instead contain footage showing natural environments such as mountains, coastlines, countryside life, etc. (Figure 2). They are mainly characterised by the presence of natural and background sounds and by a distinguished use of camera shooting techniques, such as those employed during helicopter views of natural environments.

Fig. 2: Frames extracted from aerial views, usually employed for producing documentaries.

The third type is script-content rushes, that is, footage shot to produce movies, dramas or situation comedies. The characteristics of this type of material are a limited editing, the dominant presence of human speech, a distinctive use of camera techniques and a high level of redundancy. In fact this footage usually presents many takes of the same scene, mainly due to actor errors. Script-content material may also contain some segments not really related to the storytelling, such as scene preparation by assistants, clap boards, talks between actors and director, scenes with fixed camera, undesirable content such as colour bars and frames whose colour is uniform or blurred, usually referred to as *junk frames* (Figure 3).

Fig. 3: Frames from sit-com shooting: a dialogue scene (left) and a blurred junk frame (right).

In the following we try to characterise these types of rushes by exploiting their multimodal features for retrieval purposes.

3. RUSHES MULTIMODAL SPACE

In [7] the author presents a tool to describe the identity of a design object, by placing the product in a 3D space according to its shape, efficiency and social context. In this space the three axes refer to the so called *natural, temporal* and *energetic* dimension, respectively.

In a similar way, we characterise the identity of a rush by positioning it in a multimodal space whose dimensions are related to the physical properties of the filmed video. In order to make explicit the existing bonds between the low-level features of a rush and its semantics, at first we associate each axis to a couple of adjectives in a dichotomic semantic relationship. To the natural axis we link the couple *warm/cold*. The temporal one is described in terms of *dynamic/slow*, while the dichotomy *energetic/minimal* is associated to the third axis. Then, we look for low-level features and filming techniques used by cameramen while shooting rushes and we associate them to the selected dichotomies (see Section 4).

The association between the semantic axes and the extracted multimodal features aims at closing the semantic gap between the physical video properties and the shown high-level concepts. In the defined space, a rush is represented by a trajectory that describes the temporal evolution of its multimodal low-level features (as shown in Figure 4). Observing this trajectory moving along the semantic dimensions, a chance for a high-level interpretation of the investigated material is provided.

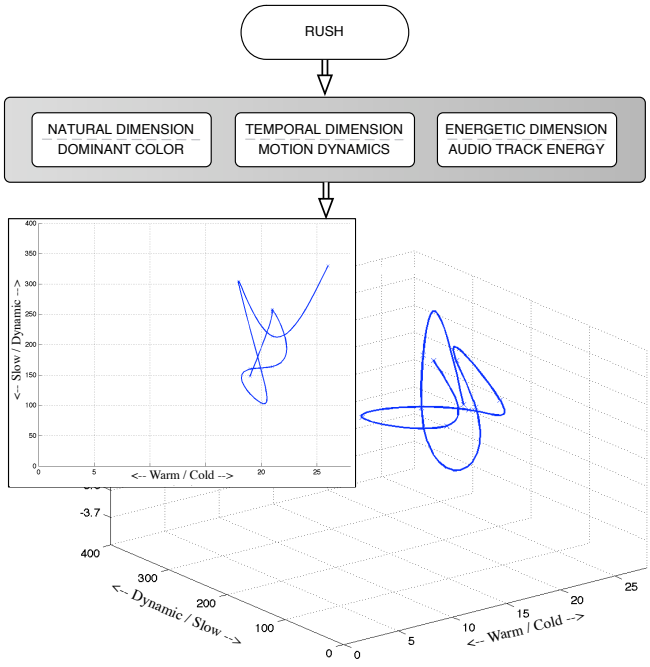

Fig. 4: Multimodal space: general framework (top) and trajectory from an excerpt of a football rush (bottom).

4. SPACE DIMENSIONS

A video can be considered as the transmission of a concept in an audio-visual appearance. This concept is mainly conveyed by the subject of the scene, by its shape, by its movements and by its general behaviour. In addition, there are many other factors that give an important contribution, such as the shooting techniques, the shot type (*e.g.*, long, medium, close-up), the use of colour (of the objects and of the scene illumination), camera movements, the aesthetic organization of the scene, the accompanying audio, etc.

To define the axes of our space, we link the associated dichotomies with specific physical properties of the video and with filming techniques employed by cameramen. For the natural axis we consider the value of the dominant colour of the scene, for the temporal axis we select the motion dynamics due to camera and object movements, and finally we associate the energy of the audio track to the energetic axis.

4.1. Dominant colour

Colours present in a scene are an important visual attribute for its characterization. Humans perceive and understand what they see thanks also to colours, to their spatial distribution, to the presence or absence of particular hues, etc. Consequently, it is crucial to extract from an image those features which are able to give, in a compact way, information about colours, like the *Colour Descriptors* defined by the MPEG-7 standard [10]. Therefore for the natural axis we consider the dichotomy warm/cald and we associate it to the *dominant colour*, *i.e.*, the most representative colour of an image. For example, in Figure 5 the frame on the left has a green dominant colour, while that on the right is gray-blueish.

Fig. 5: The frame on the left has a green dominant colour, the frame on the right white-grayish.

Given a rush, for each shot one key-frame is extracted as specified in [11]. Then, in order to compute the dominant colour in the CIE-Luv space, the following procedure is adopted. At first, the average colour for all the pixels of the key-frame determines the value of the first cluster centroid. Then, a recursive procedure of cluster subdivision is applied by adding a perturbation to all centroids until the percentage reduction of the distortion from one step to the next is lower then a fixed threshold. The distortion δ is given by:

$$\delta = \sum_{i=1}^{Q} \sum_{l=1}^{P_i} \|\vec{q}_i - \vec{p}_{i,l}\|^2 \qquad (1)$$

where Q is the number of clusters, P_i the number of pixels $\vec{p}_{i,l}$ of the image that belong to the i-th cluster, that is, they are at lower euclidean distance from the centroid $\vec{q}_i$ than from other centroids. Finally, the dominant colour is given by the centroid of the most populous cluster.

A novel procedure to map the dominant colour components on a one-dimensional warm/cold scale (natural axis) is then proposed. The Black Body radiation, whose spectral composition depends only on temperature (Figure 6, top), provides a suitable starting point. However, this radiation has not green hues and the chromatic distance between its points is not linear with temperature. To solve these problems, we first build a dummy radiation by switching the position of the green and the blue channels. Then, we linearise both the original and the dummy radiation. Finally, combining these results with an appropriate non-uniform quantization law, we build the natural axis (Figure 6, bottom) and we map the dominant colour on the $\mathcal{N}$-th interval of the axis at lowest euclidean distance.

4.2. Motion dynamics

Motion dynamics are very important in the characterization of a video sequence. The analysis of motion fields and shot

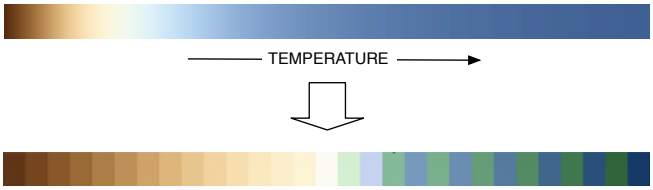

Fig. 6: From the Black Body radiation (top) to the built natural axis (bottom).

pace are two of the most common features used to extract information about the video tempo. Since for the temporal axis we consider the dichotomy dynamic/slow, we bind the axis to a compact and effective descriptor of the motion dynamics. Consequently, a shot is mapped on this axis using an index $\mathcal{T}$ given by:

$$\mathcal{T} = \left[2 - min\left(2, \frac{l_{shot}}{l_{avg}}\right)\right] + k \cdot \mathcal{M} \qquad (2)$$

The first term of $\mathcal{T}$ is related to the shot length l_{shot} and uses the average shot length l_{avg} computed on a large rushes database provided within the EU project RUSHES [1]. A short shot presents a big value of $\mathcal{T}$, since short shots convey high pace. Moreover, if l_{shot} is at least double than l_{avg}, this contribution becomes null. The second term is connected to the motion activity. It captures the intuitive notion of intensity of action, not distinguishing between camera and object motion, and it is given by the standard deviation of motion vector modules. This term is then averaged over the entire shot, obtaining $\mathcal{M}$, and normalized to the same scale as the first term by a coefficient k.

4.3. Audio track log-energy

Audio usually plays a key role in understanding the concepts conveyed by a multimedia content. Even if some rushes are characterised only by natural sounds or by no sounds at all, we exploit also this information for cataloguing rushes. Therefore we decide to take into account a feature which roughly describes the audio characteristics instead of using, for example, an accurate frequency analysis. To the energetic axis we link the dichotomy energetic/minimal and we associate it to the energy of the audio track.

Log-energy $\mathcal{E}$ is computed for each shot by using a 8 kHz single-channel audio signal. To highlight the presence of brief and intense events (like thunders, football supporters cheering for a goal, etc.), only audio samples above an adaptive threshold are taken into account.

5. DRAWING RUSHES TRAJECTORIES

In the defined multimodal space, a rush is drawn as a cloud of points, where each point, defined by a triplet $\{\mathcal{N}, \mathcal{T}, \mathcal{E}\}$, represents a shot. During the video playback, these points are connected in temporal order by a cubic spline, creating a trajectory which describes the evolution of the rushes multimodal identity, as shown in the bottom of Figure 4.

5.1. Solid summaries

Drawing the trajectory of an entire rush may result in a too complex description of its multimodal identity. A condensed representation is provided by a 3D-solid that summarises the fundamental characteristics of the trajectory (Figure 7).

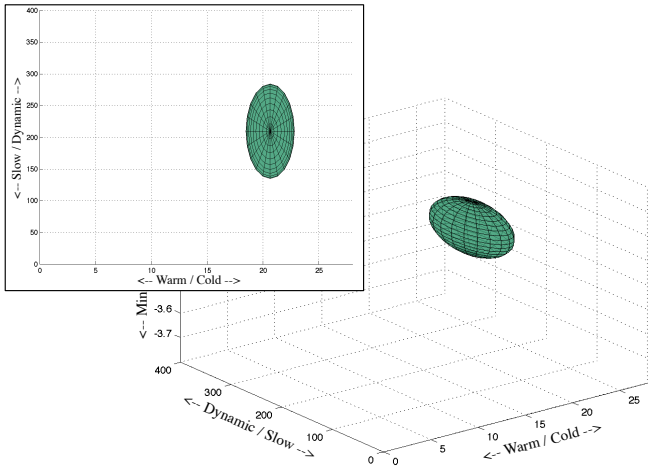

Fig. 7: A 3D-solid built on the whole football rush used in Figure 4.

The solid *colour* is the average dominant colour, while its *geometric shape* is defined by the smoothness of the trajectory: the smoother the trajectory, the smoother the solid surface. The solid *centroid* c is obtained by averaging the shot positions, while solid *dimensions* σ_j are computed as the standard deviations of the shot triplets $\{\mathcal{N}, \mathcal{T}, \mathcal{E}\}$ over the three axes. Based on these parameters, it is possible to define a "multimodal distance" D between the solids $\mathcal{S}_A$ and $\mathcal{S}_B$ representing two different rushes:

$$D(\mathcal{S}_A, \mathcal{S}_B) = \sum_{j=1}^{3} \alpha_j |c_{A,j} - c_{B,j}| + \frac{1}{\beta} \sum_{j=1}^{3} \alpha_j |\sigma_{A,j} - \sigma_{B,j}|$$

where coefficients α_j are used to normalize the axes to a common scale, while β adequately weighs the two terms.

6. RETRIEVAL BY MULTIMODAL IDENTITY

Our framework has been tested on a database provided by EiTB [9], the main Basque Country broadcaster, within the EU project RUSHES [1]. This corpus mainly comprehends material for news, documentaries and for producing situation comedies. The database used for tests contains 77 videos

of rushes material that are manually annotated with four different semantic labels: *aerial*, *football*, *interview* and *script-content*, which specialise even more the categories described in Section 2. These four semantic labels constitute the ground-truth for the following performance evaluation. To investigate the database structure and how different rushes types are related in terms of multimodal distance, we built the similarity matrix shown in Figure 8, where rushes are ordered according to their semantic labels.

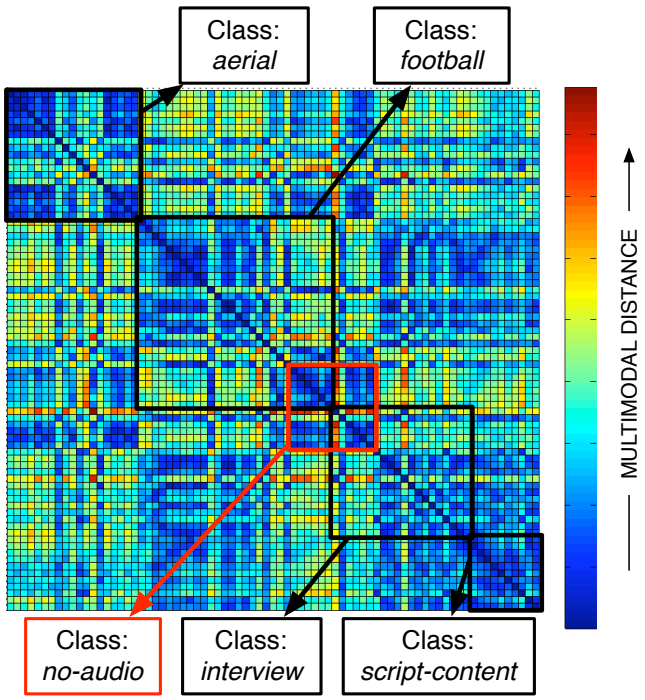

Fig. 8: Similarity matrix based on multimodal distances.

Observing the matrix, it is evident that rushes of the same type are mainly clustered together at small multimodal distance (*i.e.,* in blue colour clusters). Moreover we can observe the presence of another cluster, called *no-audio*, shared between the labels *football* and *interview*, which contains material with no sound. Further minor correlations can be observed in the matrix, for example between the *interview* and the *script-content* categories, since both are partially post-edited, and between the class *aerial* (most with natural sounds only) and the *no-audio* one.

Starting from these considerations, we verified the efficiency of our approach by building an application of rushes retrieval based on their multimodal identity. Given a query rush video, the application is able to retrieve from the database all those rushes whose 3D-solids are at low multimodal distance from the query one. The system performance is evaluated in terms of Precision-Recall (P-R) measured with respect to the four semantic labels of the retrieved rushes.

In Figure 9 we present the comparison of retrieval performance obtained by using all the three dimensions of the multimodal space with those achieved by using single low-level features, *i.e., dominant colour* (natural axis), *motion dynamics* (temporal axis) and *audio energy* (energetic axis). Each curve in Figure 9 is averaged on the results obtained considering all single rushes as queries.

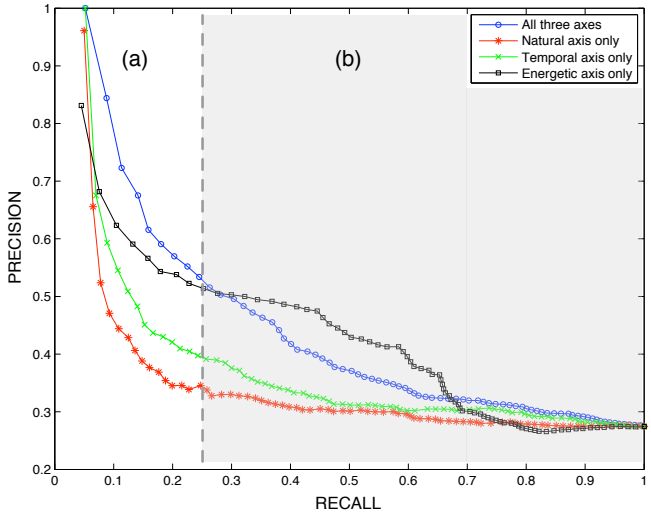

Fig. 9: P-R curves of retrieved results. The (a) "retrieval functional area" identifies the relevant portion of the P-R curves, by considering only the first positions of the ranked retrievals.

Considering the dimensions of a rushes database, we have to limit the number of the retrieved results presented to the user. Since that, and considering the fact that a professional user is interested in high precision only on the very first positions of the ranked list of retrieved videos, we identify a "retrieval functional area" for comparing system performances. This area, marked with (a) in Figure 9, here considers the first 25% of the desired type of rushes present in the database, but it is expected to be further reduced when dealing with a real application scenario. In this region, the combination of the three multimodal axes is better performing, in terms of P-R curves, than the systems employing single axes only.

Outside the area of interest, the energetic axis (audio only) proves to be highly effective in rushes retrieval. This is due to the fact that each class of rushes material in our database has a well distinguished audio (or no audio at all): natural sounds with some commentary for *aerial*, tv report and crowd cheering for *football*, people speech for *interview*, dialogues and surrounding sounds for *script-content*.
In Figure 10, P-R curves of single semantic categories are displayed. In the "retrieval functional area", identified with (a), the best performance is achieved on script-content material. This is motivated by the fact that all rushes belonging to the *script-content* class are from the same sit-com, so that they share a strong common multimodal identity. Less performing results are instead obtained on the *interview* class, since it

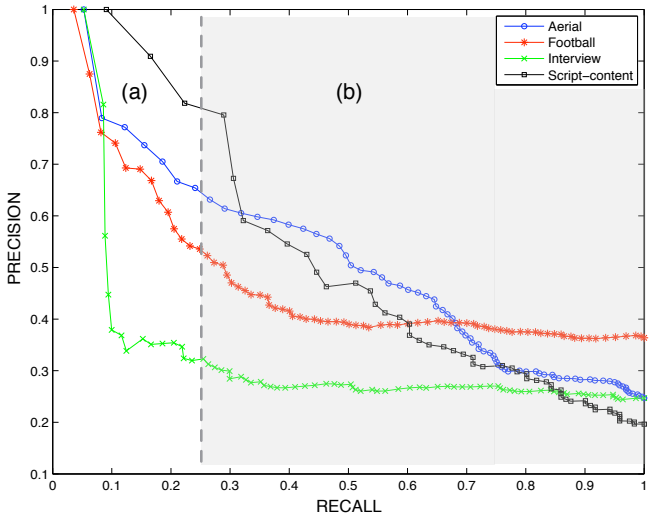

Fig. 10: P-R curves of single semantic classes. The (a) "retrieval functional area" identifies the relevant portion of the P-R curves.

contains more heterogeneous material than other categories, as shown in the similarity matrix of Figure 8.

7. CONCLUSIONS

In this paper we proposed a novel approach for characterising the multimodal identity of rushes and for retrieving similar footage from professional archives. To do this we built a multimodal space in which a rush is represented by a trajectory whose behavior is determined by low-level features related to the chromatic composition of the scene, objects and camera movements, audio and filming techniques used by cameramen. The given characterisation provides a chance for a high-level interpretation of the rushes, since we linked the axes of the multimodal space to specific semantic concepts.

The ability of our framework has been tested for retrieving similar rushes from a large database. Obtained results suggest that the proposed multimodal approach for retrieving rushes generally outperforms systems working with a single modality only. Future work aims at integrating our system with current EiTB search-engine which is only based on text.

8. ACKNOWLEDGEMENTS

This research work has been partially supported by EU project *RUSHES (FP6-045189)*. We would also like to thank EiTB for the provision of the rushes database.

9. REFERENCES

[1] "RUSHES: Retrieval of multimedia semantic units for enhanced reusability," http://www.rushes-project.eu.

[2] "TRECVID: TREC video retrieval evaluation," http://www-nlpir.nist.gov/projects/trecvid.

[3] C.G.M. Snoek and M. Worring, "Multimodal video indexing: A review of the state-of-the-art," *Multimedia Tools and Applications*, vol. 25, no. 1, pp. 5–35, 2005.

[4] B. P. Allen and V. A. Petrushin, "Searching for relevent video shots in BBC rushes using semantic web techniques," in *TRECVID Workshops*, 2005.

[5] C. Foley et al., "TRECVid 2005 Experiments at Dublin City University," in *TRECVID Workshops*, 2005.

[6] C.-W. Ngo, Z. Pan, and X. Wei, "Hierarchical hidden markov model for rushes structuring and indexing," in *International Conference on Image and Video Retrieval*, Tempe, Arizona, USA, July 2006, pp. 241–250.

[7] C. T. Castelli, "Trini diagram: imaging emotional identity 3d positioning tool," *The International Society for Optical Engineering (SPIE)*, vol. 3964, pp. 224–233, December 1999.

[8] H. Sundaram and S.-F. Chang, "Computable scenes and structures in films," *IEEE Transactions on Multimedia*, vol. 4, no. 4, pp. 482–491, December 2002.

[9] "EiTB: Euskal Irrati Telebista," http://www.eitb.com.

[10] B. S. Manjunath, P. Salembier, and T. Sikora, *Introduction to MPEG-7: Multimedia Content Description Interface*, pp. 187–201, Wiley, 2002.

[11] RUSHES FP6-045189, "D13: Report on preliminary development of low level av media processing and knowledge," http://www.rushes-project.eu, 2008.