Automatic Summarization for Generic Audiovisual Content

Nuno Matos, Fernando Pereira

Instituto Superior Técnico, Av. Rovisco Pais, 1049-001 Lisboa, Portugal
{nuno.matos, fernando.pereira}@lx.it.pt

Abstract. *Nowadays, with the explosion of multimedia content availability, the selectiveness of its consumption increases in importance. Audiovisual content is no longer brought to us only by the television, being available through many other systems, like Personal Video Recorders and Video on Demand systems, on each one's Personal Computer or, obviously, in the Internet. The exponential growth of websites like YouTube shows that people assign great relevance to audiovisual content in these days. Moreover, common people can easily produce, store, distribute and view those contents as it is not required any specialized skills to do so. To browse for a specific content in any of these systems can be a long, painful task. To manually summarize videos for the user's own purposes, as showing a summary of his/her vacations to friends and family also takes much time and it is a complex task. As people's time is getting more precious and scarce every day, an application capable of saving the time spent in these tasks by automatically summarizing audiovisual generic content arises as useful. Motivated by these factors, this report describes the developed solution for automatic audiovisual summarization for generic content, its motivations, the architecture as well as the process for summarization designed and implemented in the course of this work. To evaluate the quality of the created summaries, a user evaluation study was conducted with encouraging results, showing that the developed application is able, with relative success, to summarize audiovisual generic content.*

Keywords: *Automatic audiovisual summarization; Generic content; Arousal; Motion intensity; Shot cut density; Sound energy.*

1. Introduction

With the recent explosion of multimedia content availability, the selective consumption of audiovisual content has been increasing in importance. Audiovisual content is no longer brought to us only by the television, as happened for many decades, but it is also available to the world from an endless number of systems, notably the personal computer, the Internet, Personal Video Recorders (PVR), Vidoe on Demand (VoD) systems, mobile networks, among others. All these systems store or stream audiovisual content in a multimedia format, often rather big in size (this means in number of bits) and duration, and usually there are massive collections of contents available to the users located in any part of the world. One of the distinctive features of audiovisual content usage is that, in these days, it is not required having any specialized skills at all to produce, process, store, distribute or view this type of content. Common people can make videos from their vacations or special moments with their personal camcorders and store them in their personal computers, or use them on a daily basis on VoD and PVR systems in their homes, and frequently use the Internet to make their own videos available to the world or browse for videos of their interest. One screaming example of this later case is the fantastic boom of YouTube, in recent years, with millions of new video additions per day, from people all over the world. YouTube represents today almost 3% of all daily page views across the Internet, arising from near 0% at the beginning of 2006. This illustrates that the time spent by people consuming audiovisual contents is increasing at a very fast pace.

However, the huge amount of available data is also a problem since everybody has a limited viewing time and thus it is not only important to find quickly the audiovisual material one is looking for but it is also sometimes also important to filter from that material the more relevant or exciting parts, especially if the content is long and has many less relevant parts. In this context, imagine, for example, an application capable of allowing someone to produce a video summary of his/her vacations to family and friends, or to automatically produce trailers or teasers of movies for a VoD system, or even to extract the highlights of a football match or a Formula 1 race to include them as reports in a TV news show. These examples present some critical usages of automatic audiovisual summarization, justifying the development of tools capable of successfully automatically identifying and filtering the most exciting moments from a content asset and include them in a summary. Also manual browsing in looking for a specific audiovisual content is, in the majority of these systems, a common task performed by its users. Searching for a specific birthday video of someone in special, from a personal collection, stored in the personal computer, or choosing a movie in a VoD system, are not rare tasks performed by people in its every day life. The main problem of this kind of tasks is the great amount of time and effort that has to be spent by the user to be successful in his/her search. Audiovisual contents take time to be available to and consumed by the user and, in the majority of the cases, the desired video is found after browsing/viewing many contents which proved to be of low interest. Automatically summarizing audiovisual content may allow the user to spend much less time in browsing tasks, as summaries are smaller files, in size and duration, and therefore take less time to become available and to be consumed by the user who can infer the relevance of the entire content, deciding afterwards if it suits his/her wishes. All this motivates automatic audiovisual summarization. The utility of automatic summarization in the above mentioned systems and situations proves to be rather wide and powerful, motivating the work designed, implemented and evaluated in this paper which objectives are explained next.

Currently, many approaches to the audiovisual summarization problem have been studied and proposed, mainly divided in summarization of specific content versus generic content. The applicability of solutions addressing only specific content, for example football or basketball matches, is obviously limited. Therefore, this paper targets the development of an audiovisual summarization solution for generic audiovisual content. The main objectives of this paper are to design, implement and evaluate an application capable of based on some input audiovisual content, of any kind, producing summaries with the most interesting events occurred in that content in a simple and intuitive way to the user. To do so, the design of the summarization application was based on modeling the excitement, in this paper named as *arousal,* felt by the viewer along the content, including in the summary the segments which provoke more excitement in the viewer. This arousal modeling approach allows any content to be summarized regardless of its kind, origin, etc. assuring, in this manner, the genericity of the application and thus a wide range of applicability.

This paper is organized in seven sections, including this first one and the seventh referring to the conclusions and future work. Section 2 reviews the literature on automatic audiovisual summarization problem through a brief description of some relevant systems and proposes a classification for the audiovisual summarization solutions depending on the adopted technical approach. Section 3 introduces the solution developed in this paper, by presenting its architecture and a functional description of each of its modules. Section 4 presents a description of the processing algorithms in order to allow the reader to get a complete understanding of the entire process proposed for the summarization. Section 5 is dedicated to the description of the application’s Graphic User Interface and intends to provide sufficient information for a proper and easy usage of the summarization application. Section 6 presents and analyses the results of the subjective tests carried out to evaluate the proposed solution. Finally, Section 7 is dedicated to the conclusions and future work.

2. Background and Classification

To get aware of the state-of-the-art on automatic audiovisual summarization, a review of the relevant literature has been made. As a consequence, many systems were found, and four systems were considered more representative of a wide range of approaches to the audiovisual summarization problem and will, therefore, be briefly described next. The first system, developed by James, Echigo, Teraguchi and Satoh [1], proposes a framework for generating personal video summaries based on metadata, this means based on extracted features. This approach intends to summarize only football matches and is based on the extraction of high-level features from available event metadata; these features are after used in a supervised learning context. After a training phase where a user selects his/her personal highlights from a set of training videos, features are extracted from the training set. Those features are after used by a classifier that chooses, using the event metadata from a new video, which segments will be included in the digest according to the user’s preferences. The second system reviewed selected is also only intended to summarize football content, and has been developed by Ekin, Tekalp, Mehrotra [2]. This system is based on both low-level and high-level features: low-level features are used in the low-level analysis for cinematic features extraction algorithms while high-level features are used in the detection of goals, the referee and penalty-box events. The third system has been developed by Hanjalic and Xu [3][4][5] and focuses on the semantic summarization of multimedia content, mainly based on the extraction of moods from video and sounds. This system addresses generic content which makes it a more powerful approach to the problem regarding the systems previously presented as it produces summaries for any kind of audiovisual input. In the scope of this paper, it reveals another very relevant dimension which is the exploitation of affect in audiovisual summarization. The generic and affective aspects of this solution come through the concept of “arousal”: the search for highlights is done by tracing the audiovisual segments where the arousal experienced by the viewer is expected to be high, instead of modeling each potential highlight event individually. For this set of reasons, this system has become a major reference for the work developed io this paper. The last system here reviewed addressing the summarization problem was developed by Ma, Hua, Lu and Zheng [6] and provides a generic framework for user attention modeling and its application to the problem at hand. This system is an example of an interesting alternative to the third system described since both are generic content, affective-based solutions. It aggregates both low-level and high-level features in the construction of its user attention model and presents a complete solution to address the audiovisual summarization problem as it includes an audiovisual summarization model able to produce static or dynamic summaries, i.e. summaries based only on key-frames or also audiovisual segments. The solution differs from the one previously described in its use of high-level features to provide a different approach to affective summarization.

As for the majority of technical problems, the various ways to address the audiovisual summarization problem can be clustered and classified depending on the approach, concepts and tools used. Based on the literature reviewing made for the purpose of structuring the problem at hand, automatic audiovisual summarization, a classification tree for solutions addressing the audiovisual summarization problem is proposed and shown in Figure 1.

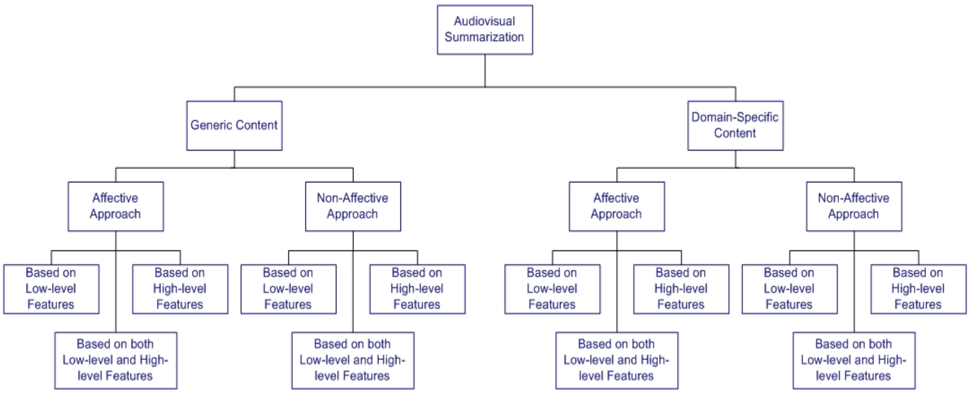

Audiovisual Summarization
Generic Content
Domain-Specific Content
Affective Approach
Non-Affective Approach
Affective Approach
Non-Affective Approach
Based on Low-level Features
Based on High-level Features
Based on Low-level Features
Based on High-level Features
Based on Low-level Features
Based on High-level Features
Based on Low-level Features
Based on High-level Features
Based on both Low-level and High-level Features
Based on both Low-level and High-level Features
Based on both Low-level and High-level Features
Based on both Low-level and High-level Features

Figure 1 – Classification of automatic audiovisual summarization solutions.

As shown in Figure 1, the main dimensions adopted to organize and classify the technologies and solutions for audiovisual summarization are:

1. **Generic versus specific content solutions** – The main difference between these two main families of solutions, generic and specific, relies on the target type of audiovisual content. The generic approach is designed to have the ability to produce summaries for any type of audiovisual content while the specific content approach addresses some specific type of content, e.g. football broadcasts or news programs. For both the generic and specific content solutions, the approach to the summarization problem can be affective or non-affective.

2. **Affective based versus non-affective based solutions** – Both in the context of generic and specific content solutions, it is important to distinguish between affective and non-affective solutions. Audiovisual summarization is one of many possible applications of affective video content analysis. These solutions are characterized by the usage, as filtering criteria for the presence of a video segment in the summary, of a certain amount or type of feelings or emotions provoked on the viewer or a certain amount of attention that the viewer is expected to dedicate to that part of the video. Therefore, affective audiovisual summarization is performed after a process of affect or attention modeling. One of the main appeals of affective summarization is its potential ability to create summaries for any type of input audiovisual content. Both affective and non-affective solutions are based on low-level or high-level audiovisual features or even on a hybrid solution, which is the most common approach as both low-level and high-level information are important to model human reactions to video content.

3. **Low-level features based versus high-level features based solutions** – Low-level features based summarization solutions are those which are based on low-level information automatically extracted from the audiovisual segments. The main difference among the various low-level features based models is what is done with the low-level information extracted. In affective summarization solutions, low-level information is typically used to model affection or attention while in non-affective solutions what is typically done is to include in the summary all segments that have, for example, a sound energy value or a density of cuts higher than a pre-determined threshold value; the other segments are left behind. The main difference for high-level features is that high-level features are mainly used to describe events, and event modeling is mainly done in a domain-specific content context. Therefore, in a generic content context, it is difficult to build an effective audiovisual summarization solution based only on high-level features since the relevant events cannot be known in advance. High-level features based solutions are far more common in domain-specific solutions, based on event modeling, where the list of relevant is predetermined. Combined low-level and high-level solutions also exist, gathering the best of the two worlds, both in generic and domain-specific and affective and non-affective solutions.

There are other ways to classify and organize the same technologies, very likely as good as the one proposed here; but what is most important here is to get one good broad view of this technical field with this view presented in a structured, organized way and not just as a simple list of solutions.

3. System Architecture

As referred before, the reference system for the technical solution adopted for this work was developed by A. Hanjalic [3][4][5]. Therefore, there are many similarities between the architecture proposed by Hanjalic and the system architecture developed and implemented in the context of this paper. In the summarization system presented here, and as in Hanjalic's work, the excitement or arousal felt by the viewer of a certain audiovisual content is modeled in order to present him/her with a summary of the selected audiovisual content. This arousal modeling relies on low-level features extracted from the audiovisual content. While Hanjalic uses only the arousal curve resulting from arousal modeling to decide which segments of the original audiovisual content should be included in the final summary, here an MPEG-7 compliant hierarchical summary description is created; this allows all the segments of the audiovisual content to be labeled and adequately included in the final summary to be created later, based on that description and the user needs. From the hierarchical summary description, MPEG-1 summaries can be made by the user at any time, even much after the description was created, and with various filtering criteria. The system architecture is presented in Figure 2: it shows three core modules, its sub-modules, as well as the inputs and outputs for each sub-module.

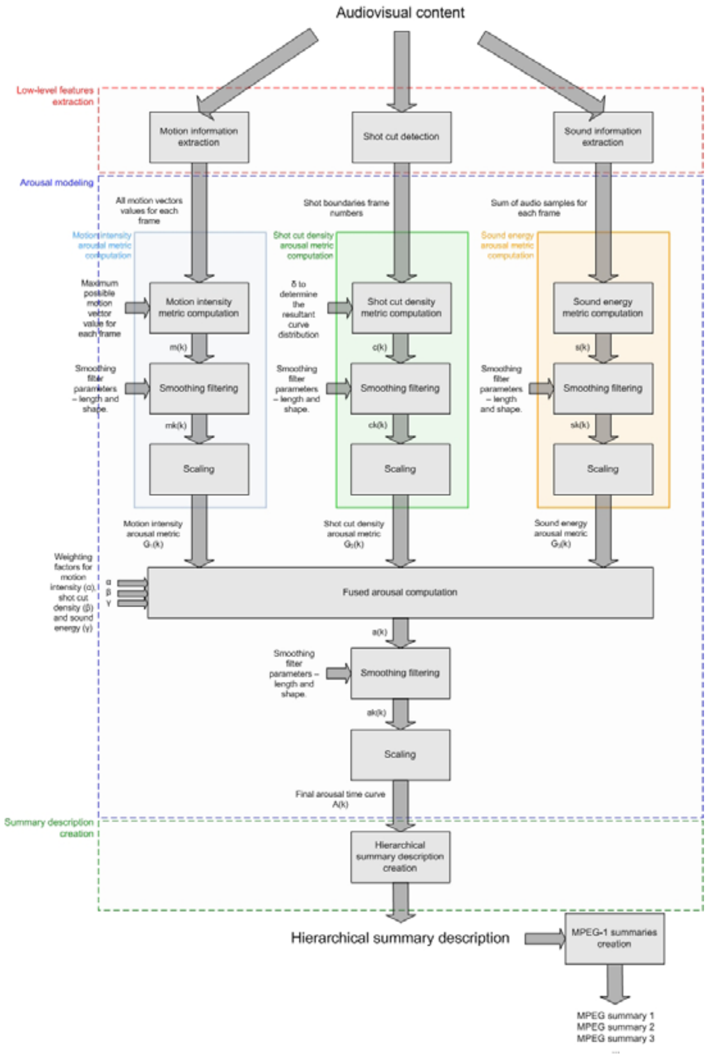

Audiovisual content
Low-level features extraction
Motion information extraction
Shot cut detection
Sound information extraction
Arousal modeling
All motion vectors values for each frame
Shot boundaries frame numbers
Sum of audio samples for each frame
Motion intensity arousal metric computation
Shot cut density arousal metric computation
Sound energy arousal metric computation
Maximum possible motion vector value for each frame
Motion intensity metric computation
δ to determine the resultant curve distribution
Shot cut density metric computation
Sound energy metric computation
m(k)
c(k)
s(k)
Smoothing filter parameters – length and shape.
Smoothing filtering
Smoothing filter parameters – length and shape.
Smoothing filtering
Smoothing filter parameters – length and shape.
Smoothing filtering
mk(k)
ck(k)
sk(k)
Scaling
Scaling
Scaling
Motion intensity arousal metric G1(k)
Shot cut density arousal metric G2(k)
Sound energy arousal metric G3(k)
Weighting factors for motion intensity (α), shot cut density (β) and sound energy (γ)
α
β
γ
Fused arousal computation
a(k)
Smoothing filter parameters – length and shape.
Smoothing filtering
ak(k)
Scaling
Final arousal time curve A(k)
Summary description creation
Hierarchical summary description creation
Hierarchical summary description
MPEG-1 summaries creation
MPEG summary 1
MPEG summary 2
MPEG summary 3
...

Figure 2 – Architecture of the developed automatic audiovisual summarization solution.

The three core modules in the system architecture are:

- **Low-level features extraction** - The first step in the summarization process is the extraction of low-level features, necessary to model the arousal for the input audiovisual content, in MPEG-1 (coded) format, to be summarized. This module has a fundamental role as it provides the necessary information about the audiovisual content to effectively model the arousal. Each of the feature extraction processes – one per feature – is done independently, and it is possible to produce summaries based only on one or more of the three selected features. The three low-level features chosen were motion intensity, density of shot cuts and sound energy for their relevance in influencing the viewer's reactions when watching a video. An increase of object motion, as well as camera motion, typically implies an increase in the arousal. Shot lengths, or its patterning, are used many times by movie directors to inflict a desired pace of action. Normally, a higher density of shot cuts, or consequently shorter shot lengths, means action and stressful segments, while longer shot lengths are useful to provoke more relaxing and calmer moments for the viewer. A change in shot lengths during a video is likely to cause significant changes in the viewer's arousal, similarly to motion intensity. The third feature chosen for arousal modeling was sound energy. As motion intensity, the loudness or energy of the audio signal has a direct influence on the emotions that viewers may experience while watching a video. An increase of the sound energy in determined segments of the audiovisual content leads to a boost of the audience's arousal. In a football match, for example, when a goal event occurs or when a rough tackle takes place, normally the commentator shouts or starts to talk louder and the audience cheers or boos. In an action movie, gunfire or explosions are related to action sequences and, therefore, segments where the arousal experienced significantly increases. All this, justifies the choice for these three low-level features. The outputs of this module serve as input for the arousal modeling module which comes next.

- **Arousal modeling** - The information obtained by extracting the low-level features from the audiovisual content is used to model its arousal. As in the extraction module, arousal is modeled independently for each feature, producing an arousal curve for each of them. A smoothing filter is applied after in the process, for each of the features, with the objective to transform the (sometimes) abrupt arousal changes resulting from the feature extraction in smoother arousal changes more likely to express the viewer's feeling when watching a video. After smoothing filtering, scaling is applied to scale the resulting curve to percentage curve, which is fundamental to permit the comparison between all arousal curves. When all features' arousals are modeled, a fusion function is applied to integrate them into one single final arousal curve. The arousal curve illustrates the arousal evolution along the audiovisual content duration. In the fusion process, different weights can be assigned to the various features, producing a different arousal final curve and, therefore, different final summaries.

- **Hierarchical summary description creation** - After arousal modeling, a hierarchical summary description is created, resulting in the final output of the system. According to the frame arousal, frames will be grouped together in segments and those arousal segments will be labeled as "Top Highlights", "Key Points", "Extended Summary" or "Remaining Content" and included in the summary description under those labels. The arousal labels go from the most exciting - "Top Highlights" - to the less exciting - "Remaining Content". The user can view the summary for an audiovisual content by choosing through the first three labels (as the "Remaining Content" represents all audiovisual content) or by entering the length of the desired summary. If the user decides to input the summary length, the summary will include segments from "Top Highlights" to "Remaining Content", respectively, until the desired length is reached. At the end of the process, if the user wishes he/she can produce an MPEG-1 file with the created summary. The MPEG Summaries Creation module is not considered a core module as it may be applied or not immediately after the summarization process. The main output of this system is the hierarchical summary description from which, if it possible to create many types of summaries following the user needs.

4. Processing Algorithms and Metrics

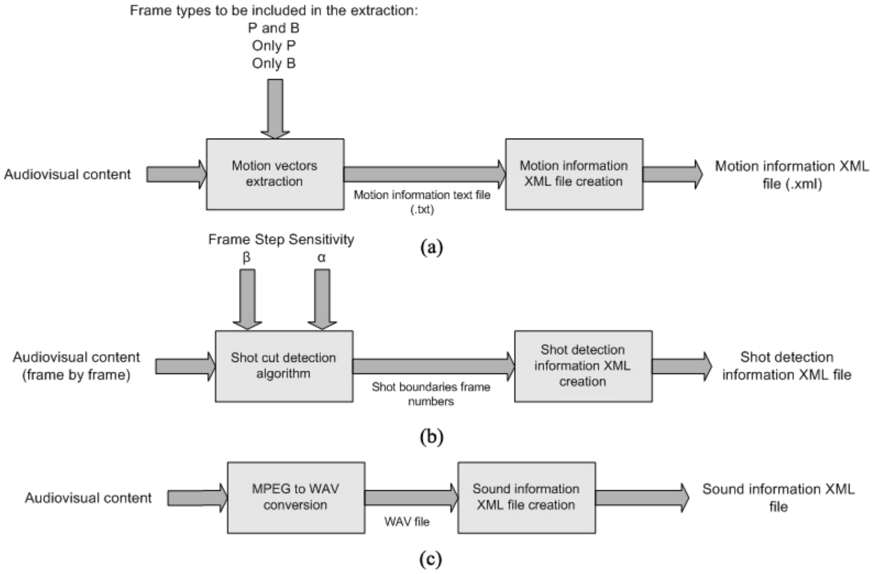
This section intends to provide an overview of the processing algorithms developed to fill the modules for the architecture presented in Figure 2.

4.1 Low-level features extraction

Three feature extraction tracks were introduced in the architecture, one for each low-level feature, with the capability of delivering the information needed for arousal modeling. To model arousal from motion intensity, all motion vectors for each of the video frames (with the exclusion of I frames which do not have motion vectors) are required, to model arousal from density of shot cuts the frame indexes which represent shot boundaries are needed, and to model sound energy all audio samples from the audiovisual content are necessary. The low-level features extraction sub-modules architectures are presented in Figure 3. The feature extraction sub-modules are:

- **Motion intensity extraction** – Motion intensity extraction sub-module directly extracts from the MPEG-1 coded bit stream the necessary motion vectors.
- **Shot cut detection** – To detect the shot boundaries, an algorithm based on luminance and saturation histograms differences is used.

- **Sound energy extraction** – The solution used to obtain all audio samples from the audiovisual content is based on a conversion from MPEG to WAV format, followed by the reading of the samples from the WAV format.



Frame types to be included in the extraction:
P and B
Only P
Only B
Audiovisual content
Motion vectors extraction
Motion information text file (.txt)
Motion information XML file creation
Motion information XML file (.xml)
(a)
Frame Step Sensitivity
β
α
Audiovisual content (frame by frame)
Shot cut detection algorithm
Shot boundaries frame numbers
Shot detection information XML creation
Shot detection information XML file
(b)
Audiovisual content
MPEG to WAV conversion
WAV file
Sound information XML file creation
Sound information XML file
(c)

Figure 3 – Architecture of (a) motion information extraction, (b) shot cut detection and (c) sound information extraction sub-modules.

All these sub-modules will produce as output, a XML file with the information collected in the extraction process. This output will serve as input to the corresponding arousal modeling sub-module described next.

4.2 Arousal modeling

After extracting the selected low-level information and storing it in XML files, the next stage is arousal modeling. Arousal modeling is done in two main steps: Arousal metrics computation and Fused arousal computation. The final arousal curve A(k) can be seen as the fusion of the $G_i(k)$ functions which represent the arousal resulting for feature i information along the audiovisual content. The computation of the $G_i(k)$ functions will be made in the Arousal metrics computation module while A(k) will be generated in the Fused arousal computation module. These processes will be explained in the following sections.

4.2.1 Arousal metrics computation

As referred before, Arousal metrics computation is the immediate stage after low-level information extraction. Each feature's arousal metric is computed independently, originating an individual arousal function, $G_i(k)$, as output. After the individual arousal metrics are computed for each feature, a final fused arousal curve will be generated. In the process of determining an arousal curve for each feature, to be given to the fused arousal process, three main steps have to be taken: **computing the associated metric**, **smoothing that same metric**, and **scaling the smoothed curve**.

1. ***Motion intensity:*** First of all, and following the steps described before, a metric aiming to compute motion intensity at each video frame *k* is proposed. Having all motion vectors values for each frame stored in the motion information XML file, a possible way to represent the motion intensity for each video frame *k* is to compute the average motion vector magnitude for each frame and divide it by the maximum possible motion vector magnitude for that frame, obtaining therefore, for each video frame *k*, the motion intensity in relation to its maximum, possible in percentage. In this context, it is proposed here to adopt as motion intensity metric m(k), the function [5]:

$$m(k) = \frac{100}{\left|\overrightarrow{v_{k\,\max}}\right|} * \frac{\sum_{i=1}^{TotalMV} \vec{v}_i(k)}{TotalMV}\,\% \qquad (1)$$

Where m(k) corresponds to the average magnitude of all motion vectors values extracted for each frame *k*, regardless of its direction, horizontal or vertical – the main factor is magnitude – normalized by the maximum possible magnitude of a motion vector for that frame, $|v_k\max|$. In order to smooth this proposed metric, for the reasons explained in Section 3, a mathematical convolution with a Kaiser window [7], K(N, α), is performed, originating an mk(k) curve. In this work, the values of N are, typically, the total duration of the video divided by 15 and α=5 as these were the values that proved, after intensive testing, to be more suitable for the desired purposes. Note that the mathematical convolution of any metric with the Kaiser window will produce a curve in a different scale range and, therefore, scaling is needed, precisely to scale back the curve for percentage values. To complete the motion intensity arousal metric computation, the curve has to be scaled back to percentage values, as will be also done for the other metrics, in order to allow the comparison between them. The scaling results in a final motion intensity arousal metric, G_1(k) [5]:

$$G_1(k) = \frac{\max(m(k))}{\max(mk(k))} * mk(k)\% \qquad \text{with} \qquad mk(k) = m(k) * K(N,\alpha) \qquad (2)$$

2. ***Shot cut density:*** The same type of reasoning made for motion information has to be made now for shot cut detection. Here, the goal is to relate the shots duration with the arousal experienced by the audience. Shorter consecutive shots are normally related with moments of fast action while long shots often mean calmer and more relaxing segments in the audiovisual content. With this in mind, the objective of the defined shot cut density arousal metric to be proposed is to compute coherent values with each shot's duration. In this way, the metric should result in higher values for shorter shots and lower values for longer shots. The metric adopted to fulfill these requirements, c(k), is [5]:

$$c(k) = 100 * e^{\left(\frac{1-(n(k)-p(k))}{\delta}\right)}\% \qquad (3)$$

For the same reasons, and in the same way as motion, smoothing and scaling have to be performed originating the final arousal curve for shot cut density, G_2(k) [5]:

$$G_2(k) = \frac{\max(c(k))}{\max(ck(k))} * ck(k)\% \qquad \text{with} \qquad ck(k) = c(k) * K(N,\alpha) \qquad (4)$$

3. ***Sound energy:*** As for the other low-level features metric definitions, the goal here is to relate the feature information with the degree of excitement experienced by the audience. For the sound, this relation may be quite simple; the louder the sound of an audiovisual segment, the higher is the arousal experienced by the viewer. Explosions or gunfire in action films, cheers of the audience in sport broadcasts, screams in horror films are all segments with high values of sound energy and all these segments are capable of provoking high arousal experiences in the viewer. On the other hand, silent segments are usually associated to calming moments for the viewer. Therefore, the aim is to use a metric capable of providing higher values for higher sound energy values and lower values for lower sound energy values. To do so, s(k), representing for each frame *k* the sum of the squares of the audio samples is computed [5]:

$$s(k) = \sum_{i=1}^{TotalSamples} \left(AudioSample_i(k)\right)^2 \qquad (5)$$

Smoothing is applied in the same way, and for the same reasons, as previously described for the other metrics, convoluting s(k) with the Kaiser window, with N and α having the same values as before, originating sk(k):

$$sk(k) = s(k) * K(N,\alpha) \qquad (6)$$

Scaling in sound energy is done using two scaling functions with different purposes. The first has the same function as the scaling functions used for the other low-level features arousal metrics, i.e. to scale back the curve resulting from the smoothing filtering back to percentage values. As sound energy does not have a maximum value by default because it is not computed in percentage, the smoothed sound energy curve must be scaled according to its own peak. In this manner, sound energy curves from different contents can be compared. Therefore, the first scaling function is performed by (7). If no other scaling function was applied, a sound energy arousal curve with a 100% value on the highest sound energy value would result, which is not desirable because, as for motion intensity and shot cut density, the purpose of the sound energy arousal curve is to model the arousal experienced by the viewer while watching an audiovisual content and not to produce a curve relating the sound energy values to its maximum value. Therefore, a second scaling function (9) is necessary to transform the first scaled function into a curve capable of represent the level or arousal related to sound energy felt by the viewer along the content. The solution found was to compute the mean of sk_n(k), i.e. the mean of the sound energy values related to its peak (8). This will give more precise information about the arousal related to the sound energy experienced by the viewer along the content. A high mean value shows that the sound energy of most of the audiovisual content is close to the peak value and, therefore, the arousal curve should be flatten and scaled down, as the audiovisual content is very constant in terms of sound energy. If the mean value is low, then the sound energy peak is considerably higher than the rest of the content meaning that the variations are more significant and, therefore, the curve should be able to highlight its peaks as they represent important and relevant changes in terms of arousal. The formulas below represent the scaling functions for sound energy [5]:

$$sk_n(k) = \frac{sk(k)}{\max(sk(k))} \quad (7) \quad \text{and} \quad \overline{sk_n} = \frac{1}{K}\sum_{k=1}^{K} sk_n(k) \quad (8)$$

Finally, to create the final sound energy arousal curve satisfying the requirements above and using the scaling functions described above, $G_3(k)$ [5] is computed as:

$$G_3(k) = 100 * sk_n(k) * (1 - \overline{sk_n})\% \quad (9)$$

4.2.2 Fused arousal computation

The last sub-module before creating the summary description is Fused arousal computation. Fused arousal computation aims to integrate all arousal curves resulting from the low-level features arousal metrics computation processes. This constitutes the main objective of fused arousal metric computation process and, therefore, the main challenge when defining a metric to achieve this goal. As the maximum of each $G_i(k)$ can be located on different frame indexes, a weighted average of the various feature metrics seems to be an appropriate fusion function as it proved to be faithful enough to the variations of each individual $G_i(k)$ function. As the default solution, all three features will be considered with the same weight, i.e. 1/3 of the weight in the creation of the final arousal curve. However, the user can change the weights assigned to each feature if he/she wishes to see how the final summary evolves (differently); this may be more relevant for specific type of content. Than the Fused arousal metric, a(k), is computed as [5]:

$$a(k) = \sum_{i=1}^{3} w_i G_i(k) \quad (10)$$

In (10), w_i refers to the weight assigned to each feature and its sum must be 1. Smoothing is done to merge the neighboring maxima of each $G_i(k)$ function and is applied in the same way as before. After smoothing, scaling has to be performed to scale back the values to percentage values. The fused arousal curve, A(k) is the final result, computed after scaling and is computed as [5]:

$$A(k) = \frac{\max(a(k))}{\max(ak(k))} * ak(k) \quad \text{with} \quad ak(k) = a(k) * K(N, \alpha) \quad (11)$$

4.3 Hierarchical summary description creation

The main output of the entire summarization process, derived from the fused arousal curve resulting from previous modules, is a MPEG-7 compliant XML file, which contains a hierarchical summary description of the audiovisual content. A hierarchical summary description provides the means to represent the audiovisual content in segments labeled according to their importance. The most important level contains the top of the hierarchy and, as one goes down, less important segments will be included in the summary.

The creation of the MPEG-7 hierarchical summary description has two main motivations: To allow the user to view and create many different summaries fulfilling different needs, e.g. different types or with different lengths without having to repeat the entire process for summarization, and to create a summarization output capable of allowing interoperability with other systems. To achieve the first requirement, the obvious choice was to create a XML description capable of hierarchically representing the audiovisual content in terms of its summarization relevance. To fulfill the second requirement, the solution was to create the XML file with a standard format; in this case, the MPEG-7 standard was chosen since it defines precisely a description tool, *HierarchicalSummary*, with this purpose. From one MPEG-7 compliant hierarchical summary description, it can be created an infinite number of MPEG-1 files representing different summaries according to the user needs in terms of length and type.

4.4 MPEG-1 summaries creation

The MPEG-1 summaries creation process was implemented integrating the MPGTX project [8] in the developed system. MPGTX is a command line MPEG audio/video/system toolbox with the ability to slice and join MPEG-1 files. This is precisely what is needed to create MPEG-1 summaries as, from the hierarchical summary description and off the user's choice of parameters, a summary is generated, formed by a group of segments from the audiovisual content. In this way, the application will create the MPEG-1 files with the desired summary by slicing from the audiovisual content the relevant segments and joining them together. From one MPEG-7 compliant hierarchical summary description, an infinite number of MPEG-1 files representing summaries can be created as the user can choose to create summaries by length or relevance.

5. Graphics User Interface

This section aims to provide an overview of the application's Graphic User Interface (GUI). The application in composed by a single Windows Form, which is divided in 5 main areas, as shown in Figure 4. Figure 4 represents the state of the application after extracting motion information. In the charts and tab colors, light blue is related to motion information, green to shot cut information and orange to sound information.

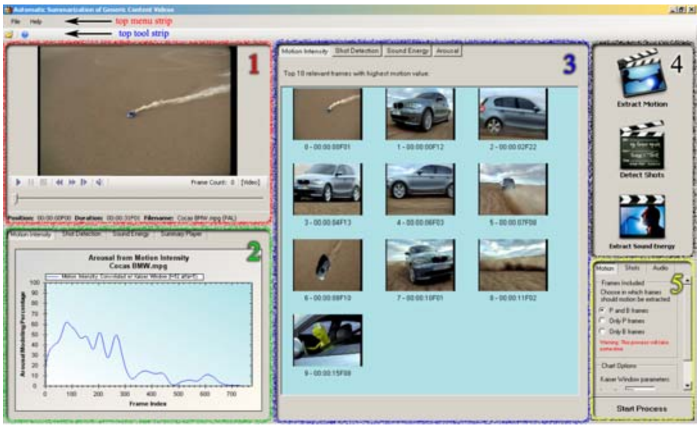


Figure 4 – Application's GUI.

The five areas highlighted in Figure 4 have the following main functions:

1. **Player** – Intended to play the audiovisual content.
2. **Charts/Summary player tab control** – Area destined to present the arousal charts from each feature to the user and also to play the final summary. Has 4 tabs, three related to the arousal charts of each of the low-level features and the fourth destined to play the final created summary.
3. **Main tab control** – Area meant for presenting information resulting from the low-level features extraction to the user and also to parameterize the fused arousal computation process as well as the summary creation and viewing processes.
4. **Side menu** – Has three buttons which serve as start for the respective low-level features extraction processes.
5. **Side menu options tab** – Area reserved for the option related to each low-level features extraction process.

The next section presents the user evaluation study conducted to evaluate the system's performance.

6. Performance Evaluation

Since the development of any multimedia application is not finished without a serious and meaningful evaluation of its performance regarding the user objectives, this section aims to present and analyze the results obtained by a subjective evaluation study which was carried out to evaluate the developed application's performance. Considering the type of system developed, it was considered that the adequate performance evaluation methodology should follow a subjective and not objective approach. The reviewing of the relevant literature confirmed that there are no objective evaluation methodologies available for the problem at hand. Thus, it was decided to design an adequate evaluation methodology and conduct a user evaluation study to assess how the developed application performs in view of the initially defined objectives, notably:

1. Evaluate how good is the experience provided by the created summaries, according to each summary type, i.e. if the "Top Highlights" summary captures only the indispensable segments of the content; if "Key Points" summary is able to provide some context to a "Top Highlights" summary, capturing only interesting segments without being too extensive; and if an "Extended Summary" is able to exclude the more boring and least interesting segments of the content.
2. Evaluate if any important segments are excluded in any of the created summaries.

The next section presents the test methodology designed to achieve these two objectives.

6.1 Test methodology

In order the test performed is credible, it had to be well defined enough to be reproducible by other experts in order the resulting results and conclusions are statistically similar.

1. **Test Questions** – The test questions defined for this test to address the objectives above are:
   - **Question 1** – Does the summary viewed satisfy its type definition, i.e. contains the top most relevant/exciting 10%, 25% and 50% of the original content?

a) Not at all b) Badly c) Reasonably d) Mostly e) Totally

- **Question 2** – Were any relevant segments ruled out of the viewed summary for each summary type, i.e. Top Highlights, Key Points and Extended Summary?

a) All b) Many c) Some d) Few e) None

2. Sequence of Testing – A group of 13 volunteers were asked to view the original audiovisual content and to give their subjective assessment to the questions above, for each of the summaries created, following the sequence of steps defined next:

a. Open and visualize the original content, starting from "Basketball" content.

b. For the original content at hand, visualize one single time its three possible summaries, i.e. "Top Highlights", "Key Points" and "Extended Summary".

c. Answer to questions 1 and 2, marking with a cross (X), in the evaluation tables, the desired classification mark, for each one of the three summaries just visualized, i.e. "Top Highlights", "Key Points" e "Extended Summary".

d. Go back to point 1. for all lasting contents until tables 1 and 2 are completely filled.

3. Test Material – The test set was constituted by 6 audiovisual pieces. Based on these 6 pieces, 18 summaries were produced and exported for MPEG-1 files, using the developed summarization application. For each audiovisual piece, three summaries were produced: i) "Top Highlights" summary; ii) "Key Points" summary; and iii) "Extended Summary". The set of 6 pieces was divided into two classes, with 3 pieces per class:

- **Sport content** – Contents containing sport broadcasts, with two clips from football matches and one clip from a basketball match.
- **Entertainment content** – Contents containing clips from TV series containing action events; one from "Lost", other from "Prison Break" and other from "Heroes".

As it is intended the user to see the original content as well as three types of summary created for that content, the original contents were clipped in order to reduce the test's duration in a way that the results can be considered meaningful but without exhausting the tester. The contents duration as well as their spatial resolution are presented in Table 1.

	Content duration	**Content resolution**
Sport content		
BASKETBALL.mpg	10:02	320×240
FOOTBALL1.mpg	13:27	418×288
FOOTBALL2.mpg	10:10	418×288
Entertainment content		
ACTION1.mpg	14:14	624×352
ACTION2.mpg	12:48	624×352
ACTION3.mpg	14:38	624×352

Table 1 – Test material characteristics.

6.2 Results and analysis

Table 2 shows all results obtained in this user evaluation study for both questions 1 and 2. In relation to question 1, the average results showed that 41,45% and 42,31% of the inquired considered that the viewed summary mostly and totally satisfies its type definition. None of the inquired considered that the summary did not satisfy at all its type definition and only 3,85% considered that it satisfied badly. Analyzing the results by summary type, all summaries had positive results, with the results of d) and e) combined always superior to 65%, which was the poorer result, revealed in "TopHighlights" type summary. "KeyPoints" and "ExtendedSummary" type summaries presented very good results, with d) and e) added in the order of 90%. In terms of content the results were more satisfactory in sport than in entertainment content. We believe that this is deeply related with the fact that exciting moments in sport broadcasts are easier to identify that in TV series and do not depend so much on the viewer's interpretation.

Regarding question 2, in average, 40,17% and 32,91% of the inquired considered that only few and none of relevant segments were excluded of the viewed summaries, results that when added, represent a good result of near 73%. Performing an analysis by type, "TopHighlights" type summary is the one that, understandable, shows the poorer results, as it is the smaller one in duration and, therefore, the one with more probability to exclude any relevant segment. Even so, it presents an added results of d) and e) of near 50% with the majority of the inquired considering that some segments were excluded from the summary and only 14,10% considering than many relevant segments were excluded. “KeyPoints” and “ExtendedSummary” performed quite well in question 2, as in question 1, with combined results of d) and e) of near 82% and 90% respectively.

Question 2 presented a greater discrepancy of results between sport content and entertainment content, with the summaries of sport content achieving better results than the one’s regarding entertainment content. Once more, we believe that this is related with the fact of entertainment content having a storyline and, therefore, the “quality” of the summary is more permeable to each user’s interpretation than in sport content where the main events, as goals in football, are normally identified by every user.

	Question 1 *The summary viewed satisfies its type definition, i.e. contains the top most relevant/exciting 10%, 25% and 50% of the original content?*					**Question 2** *Any relevant segments were ruled out of the viewed summary, for each summary type, i.e. Top Highlights, Key Points and Extended Summary?*				
	a) Not at All	b) Badly	c) Reason ably	d) Mostly	e) Totally	a) All	b) Many	c) Some	d) Few	e) None
Sport content										
TopHighlights	0,00%	0,00%	23,08%	51,28%	25,64%	0,00%	10,26%	30,77%	41,03%	17,95%
KeyPoints	0,00%	0,00%	7,69%	48,72%	43,59%	0,00%	0,00%	15,38%	56,41%	28,21%
ExtendedSummary	0,00%	0,00%	7,69%	38,46%	53,85%	0,00%	5,13%	7,69%	41,03%	46,15%
Entertainment content										
TopHighlights	0,00%	17,95%	28,21%	28,21%	25,64%	0,00%	17,95%	43,59%	23,08%	15,38%
KeyPoints	0,00%	2,56%	7,69%	58,97%	30,77%	0,00%	2,56%	17,95%	46,15%	33,33%
ExtendedSummary	0,00%	2,56%	0,00%	23,08%	74,36%	0,00%	2,56%	7,69%	33,33%	56,41%
TopHighlights Average	0,00%	8,97%	25,64%	39,74%	25,64%	0,00%	14,10%	37,18%	32,05%	16,67%
KeyPoints Average	0,00%	1,28%	7,69%	53,85%	37,18%	0,00%	1,28%	16,67%	51,28%	30,77%
ExtendedSummary	0,00%	1,28%	3,85%	30,77%	64,10%	0,00%	3,85%	7,69%	33,33%	56,41%
Sport content Average	0,00%	0,00%	12,82%	46,15%	41,03%	0,00%	5,13%	17,95%	46,15%	30,77%
Entertainment content	0.00%	7,69%	11,97%	36,75%	43,59%	0,00%	7,69%	23,08%	34,19%	35,04%
Total Average	0,00%	3,85%	12,39%	41,45%	42,31%	0,00%	6,41%	20,51%	40,17%	32,91%

Table 2 – Evaluation results for both questions 1 and 2.

7. Conclusions and Future work

Despite the promising results obtained with the user evaluation study, the solution developed still leaves room for improvement, mainly regarding the issues highlighted next. The first of them regards a revision of the shot detection algorithm which presents some false positives and false negatives, although not in a worrying number, for the majority of contents; a revision of the shot detection algorithm implemented could, therefore, be performed targeting the improvement or substitution of this algorithm. Regarding the low-level information extraction processes, other motion and sound extraction algorithms can also be studied in the future. Moreover, other low-level features may be introduce in the system to make the fused arousal metric more representative and robust. In terms of the developed application, some additional features can be also added in the future, namely a summarization wizard able to guide the user step by step through the entire process for summarization and, possibly, a more complete summary player, capable of giving the user the ability to leap from segment to segment inside the summary, perhaps by clicking in the respective segment on the chart. Despite presenting good results and fulfilling its function in a rather satisfactory manner, the fused arousal metric can also be subject of revision or even other fusion metrics can be studied in order to enhance the system’s performance. Finally, regarding performance evaluation more complete tests can be done in the future, mainly with more types and longer pieces of content to evaluate how the system’s performs under those different and more complete conditions. A more complete set of questions can also be placed to the users, in order to evaluate, in more detail the meaningfulness of each of the created summaries. The test can also be conducted in a larger scale, with a higher number of subjects participating, collecting, in this manner, many more scores and, therefore, more statistically accurate results.

In conclusion, there is still a lot of research to do in the field of automatic audiovisual summarization, specially if the semantic value of the content is to be taken into account … automatically modeling semantics still is, and very likely will be for a long time, a difficult task …

References

1. A. Jaimes, T. Echigo, M. Teraguchin and F. Satoh, “Learning personalized video highlights from detailed MPEG-7 metadata”, in Proc. IEEE ICIP, vol. 1, pp. 133-136, Rochester, New York, USA, Sep. 2002.
2. A. Ekin, A. Murat Tekalp and R. Mehrotra, “Automatic soccer video analysis and summarization”, IEEE Transactions on Image Processing, vol. 12, nº 7, pp. 796-807, Jul. 2003.
3. A. Hanjalic and L. Q. Xu, “User-oriented video content analysis”, IEEE Workshop on Content-based Access to Image and Video Libraries (CBVAIL ’01), pp. 50-57, Kauai, HW, USA, Dec. 2001.
4. A. Hanjalic, “Extracting moods from pictures and sounds”, IEEE Signal Processing Magazine, vol. 23, nº 2, pp. 90-100, Mar. 2006.
5. A. Hanjalic and L. Q. Xu, “Affective video content representation and modeling”, IEEE Transactions on Multimedia, vol. 7, nº 1, pp 143-154, Feb. 2005.
6. Y. F. Ma, X. S. Hua, L. Lu and H. J. Zhang, “A generic framework of user attention model and its application in video summarization”, IEEE Transactions on Multimedia, vol. 7, nº 5, pp. 907-918, Oct. 2005.
7. Kaiser window: http://en.wikipedia.org/wiki/Kaiser_window
8. MPGTX project: http://mpgtx.sourceforge.net/