A METHOD AND APPARATUS FOR SEGMENTING A CONTENT ITEM

- silence - (1)
- music - (2)
- speech - (3)
- noise - (4)

video frames

Abstract: A content item is segmented as follows: for each of a plurality of frames of the content item (e.g. for each of a plurality of audio frames of a content item which comprises audio content and video content), class probabilities are calculated, the class probabilities being the probabilities that the frame belongs to a plurality of pre-defined classes or subclasses; the location of a transition in the (e.g. audio) content is determined upon detection of a change in at least two of the calculated class probabilities occurring at the same time; and the content item is segmented into a plurality of portions based on the location of the transition in the (audio) content.
A method and apparatus for segmenting a content item

FIELD OF THE INVENTION

The present invention relates to segmenting a content item. In particular, it relates to segmentation of the video content using audio classification.

BACKGROUND OF THE INVENTION

With the proliferation of digital multimedia content creation and delivery, innovative solutions and systems enabling efficient access and retrieval of information of interest are becoming indispensable. This realisation has given rise to substantial research in the field of multimedia content analysis, referring to computerised understanding of semantic meaning of a multimedia document (for example, a movie). The techniques of content analysis comprise algorithms from signal processing, pattern recognition and artificial intelligence, generating metadata that enables (semi-) automatic annotation of video material. This metadata describes the content at different semantic levels, varying from low-level signal-related properties to higher-level information, for example recognition of faces in images. The results of content analysis can be used in various applications, for example video indexing and retrieval, video-genre detection, video summarisation, etc.

A particular application of content analysis is automated, semantic (i.e. scene-based) segmentation (chaptering or editing) of the audiovisual content of the content item (or document). The central idea, basically, is to detect video scenes and chapter the video accordingly. However, definition of a video scene is quite subjective. Where a video-shot is relatively easy to detect, being a sequence of contiguous video frames taken from a single camera act, shots can be grouped into a scene regarding different criteria. Even so, objective and thus computable criteria for extracting scenes, also involving filmmaking rules, are feasible.

While multimedia content analysis is concerned with metadata generation, innovative methods and systems for metadata handling (transport, storage, presentation) have also been developed. A well-known example of this is Electronic Programming Guide (EPG). EPGs are constructed based on the EPG data, which typically includes the most characteristic video program information that is easy to annotate manually. For example,
EPG data that is commonly included in digital files supplied by service providers and broadcasters (for example, Digital Video Broadcast-Service Information DVB-SI) comprises information such as timing, title, genre, abstract or keywords (for example, actors, director) associated with a video program.

Presently, most methodologies of content-based audiovisual segmentation either entail substantial human involvement, as in professional (DVD) authoring in film studios, or they are automatic but also vastly disregarding of the semantics, as in typical consumer DVD-recorders that simply insert chapter-markers at regular time intervals. Some software packages offer tools for automated content-based editing, but these are limited and still require significant user interaction. Therefore, such tools are not suitable for use by a typical home user.

For these reasons, development of algorithms that could enable (semi-) automatic chaptering of video, while at the same time minding its semantics, has become an objective of substantial research. A multitude of algorithms have been proposed. However, most suffer from imperfection of either being restrained to partial solutions, or else being "strayed" in the attempt to encompass all complex aspects of human audiovisual perception and conventions of filmmaking. The latter, even is successful, entails considerable computational complexity and often-objectionable processing delay.

One such known method effectively joins detected camera shots into semantic scenes regarding heuristic, higher-level cues, such as face and speech detection/recognition and speaker tracking. While this work wells with for example, dialogues and narrative content, it is less effective with analysing other types of scenes.

Alternatively, some methods try to detect, semantically, meaningful translations in video based on statistical observation/recognition of patterns of (semantically) lower-level audiovisual features. The problem here is that low-level information becomes meaningful, i.e. discriminative only when observed over relatively longer periods of time. Therefore, such approaches are typically found to be more adequate for longer-term classification (genre detection), rather than for shorter-term (scene-based) chaptering.

As a compromise between the aforementioned two approaches, some methods seek mid-level cues, that is, cues that have sufficient discriminative power but are also generic enough to be extracted from arbitrary content. One known method designates a scene-change as a conjunction of a camera shot-cut and an audio silence. However, this invariably leads to over-segmentation - "discovery" of a larger number of scenes than would
be found by a human, and susceptibility to classification errors, since only one of the two
cues needs to be erroneous to misclassify the entire event (scene change).

Another approach is the reliance on audio classification. Such systems, e.g. the
one described in US2004/0041831, segment video by solely regarding the corresponding
audio component. Particularly significant have been found to be cues obtained from audio
classification, referring to identification of an audio signal as being speech, music, silence,
noise, etc. The significance of audio classification is explained by the fact that much scenery
in video is inherently driven by audio; obviously dialogues contain speech, dramatic scenes
are often accompanied by background music, etc. Also, film directors are known to
explicitly use audio transitions to establish or support semantic links between (visually
dissimilar) camera shots. However, known systems that use audio classification provide a
relatively inaccurate segmentation.

SUMMARY OF THE INVENTION

To mitigate the problems of existing systems, the present invention proposes a
segmentation method that takes advantage of mid-level (audio) classification techniques.

According to an aspect of the present invention, there is provided a method for
segmenting a content item, the method comprising the steps of: for each of a plurality of
frames of said content item, calculating class probabilities, the class probabilities including
the probabilities that said frame belongs to a plurality of pre-defined classes or subclasses;
determining the location of a transition in said content item upon detection of a change in at
least two of the calculated class probabilities occurring at the same time; and segmenting said
content item into a plurality of portions based on the location of the transition in said content
item.

According to another aspect of the present invention, there is provided
apparatus for segmenting a content item, the apparatus comprising: an audio classifier for
calculating class probabilities, the class probabilities including the probabilities that each of a
plurality of frames of said content item belongs to a plurality of pre-defined classes; an
analyser for determining the location of a transition in said content item upon detection of a
change in at least two of the calculated class probabilities occurring at the same time; and
segmenting means for segmenting said content item into a plurality of portions based on the
location of the transition in said content item.

Preferably, said content item comprises video content and audio content and
said audio content comprises said plurality of frames.
In this way the method and apparatus of the present invention gives scene-change detection that have clear semantic interpretation which would not have, necessarily, been generated by the existing systems. The present invention benefits from the fact that change in a probability of an audio frame belonging to a certain (sub)class can provide meaningful information even if the probability remains relatively low. The audio frames for which the probabilities are calculated do not necessarily have the same size as the stored audio frames, e.g. the audio frames for which the probabilities are calculated could each have the same size as a video frame (e.g. 40 ms for PAL video). A change in at least two of the calculated class probabilities is preferably detected when it occurs in the same frame, but could also be detected when it occurs in neighbouring frames, for example.

The segmentation method of the present invention may also rely on video classification in addition to audio classification.

In a preferred embodiment, the method starts by computing, for each segment, or frame, of the audio content of a content item, likeliness of that segment belonging to each of pre-defined general audio classes, or possible sub-classes thereof. The general classes include speech, music, silence, background noise, environmental sounds, etc. Music, for instance, can be further classified according to tempo, beat or genre (Jazz, Rock Vocal, etc) and different types of background noise or environmental sounds can be distinguished as well (for example. Crowd, Applause, Shots, Explosions, etc).

Adopting audio (classification) as the sole criterion for the video segmentation, reduces the processing delay significantly, enabling even real-time operation (after an initial delay of some minutes needed to fill the analysis buffer). This all without necessarily sacrificing the generality or performance, since much video scenery is inherently driven by audio, and since directors use audio to support or establish semantic links between camera shots. For example, dialogues are dominated by speech, unvoiced scenes are normally accompanied by some audio; rarely is a scene completely silent. Some audio may consist of background music, or environment sounds such as explosions, gunshots, footsteps etc. Therefore, the accentuation of a particular audio source will typically depend on the character and semantic changes in the scene. For instance, in movies, the background music typically dominate static (dramatic) parts, and the environmental sounds are more prominent during the action parts. Furthermore, adopting audio (classification) as the sole criterion provides a useful output for the existing systems delivering audio classification for other, audio-only applications (for example, music retrieval or play-list generation based on audio similarity).
Preferably, the location of the transition in the audio content is determined when a change in the at least two of the calculated class probabilities exceeds a predetermined threshold.

In this way, multiple classes are taken into account reducing classification errors which helps to avoid over segmentation of the video content. Dealing with probability of occurrence of multiple (audio) classes at a same time, rather than going for fewer dominant ones, mitigates susceptibility to classification errors and could also overcome the over-segmentation (discovery of a larger number of scenes than would be found by a human).

For each audio frame, all the audio-class probabilities may be grouped to form a feature vector. This is different than just selecting a winning class (ones having the highest probability) for each frame. Not all of the available classes-probabilities will necessarily be included with the same weight or, alternatively, computation of probability of some classes may be skipped in the first place. The filtering may effectively be dependent on the video genre, since the discriminative power of each of the audio classes may vary accordingly. For example, discriminating music of a feature movie according to genre will in most cases not be very meaningful. Here, the video-genre information (for example movie) could be extracted from the EPG data, if available, or else obtained as a result of content analysis.

Involving video genre information as a criterion for controlling the audio classification could not only enhance the performance, but also ensure applicability of the algorithm to any type of video scenery.

Next, the feature vectors comprising selected audio-class probabilities may be interpolated, to obtain the same audio-class probabilities but at the level of video frames. (Typically, audio classification will be delivering at a higher rate than the video frame rate). Such obtained feature vectors may then be analysed inside a sliding window. At each instance, the data contained inside the window may be processed such to generate a measure of likeliness of a semantic (audio) scene-change inside the window. The location of each potential transition will be output and optionally refined with the aid of visual features, for example, camera shot-cut detections.

The method allows natural extension with any other features and classes (for example, from the visual domain), which could further improve the overall performance or even facilitate different applications of the algorithm.
BRIEF DESCRIPTION OF DRAWINGS

For a more complete understanding of the present invention, reference is made to the following description taken in conjunction with the accompanying drawings, in which:

Figure 1 illustrates a schematic block diagram of the apparatus according to a preferred embodiment of the present invention;

Figure 2 illustrates a flow chart of the computation of the audio-scene transitions;

Figures 3a, 3b, 3c illustrate graphical representation of resulting class probabilities of a sample file, the corresponding cluster indices and assignments of the clusters indices;

Figures 4a and 4b show graphical representation of audio classification, clustering and computation of transition probability according to an embodiment of the present invention; and

Figure 5 illustrates precision figures for the audio-classification based segmentation according to a preferred embodiment of the present invention.

DETAILED DESCRIPTION OF PREFERRED EMBODIMENTS

With reference to Figure 1, a preferred embodiment of the present invention will be described. The apparatus 100 comprises an input buffer 101. The input of the input buffer 101 is connected to the input terminal 103 of the apparatus 100. The output of the input buffer 101 is connected to the input of a de-multiplexed 105. The multiple outputs of the de-multiplexer 105 are connected to a camera shot extractor 107, a video genre detector 109 and the input terminal 110, of an audio classifier 111. The output of the camera shot extractor 107 is connected to a segmenting means 113. The output of the segmenting means 113 is connected to the output terminal 115 of the apparatus 100.

The output of the video genre detector 109 is connected to a control unit 117. The first output of the control unit 117 is connected to a first input of a first multiplier 123. The second input of the first multiplier 123 is connected to a first output terminal 125 of the audio classifier 111. The second output of the control unit 117 is connected to a first input of a second multiplier 127. The second input of the second multiplier 127 is connected to the second output terminal 129 of the audio classifier 111.

The respective outputs of the first and second multipliers 123, 127 are connected to respective inputs of an adder 131. The output of the adder 131 is connected to an analyser 133. The output of the analyser 133 is connected to segmenting means 113.
The audio classifier 111 comprises dividing means 135. The input of the dividing means 135 is connected to the input terminal 110 of the audio classifier 111. The output of the segmenting means 135 is connected to a classifier 137 and a sub-classifier 139. The output of the classifier is connected to the sub-classifier 139 and the first output terminal 125 of the audio classifier 111. The output of the sub-classifier 139 is connected to the second output terminal 129 of the audio classifier 111. The audio classifier 111 may comprise first and second control terminals 119 and 121 connected to the control of the classifier 137 and the sub-classifier 139, respectively, from the control unit 117.

In operation, a content item is fed to the apparatus 100 via the input terminal 103 and input buffer 101. The multimedia data within the file is then de-multiplexed into its respective video data, audio data and, if available, EPG data. Alternatively, the EPG data may be provided in a separate digital file. The EPG data will normally incorporate the genre information, which is of interest for later processing. In the absence of EPG data, the genre can be detected automatically by means of the video genre detector 109, as indicated by the dashed lines in Figure 1.

Once separated from the visual data, the audio is input on the input terminal 110 of the audio classifier 111. The audio is divided into a plurality of frames by the dividing means 135. Each audio frame is assigned a set of probability figures, each indicating the likelihood of that frame belonging to one of a plurality of pre-defined classes by the classifier 137 (for example, speech, music, etc). If a certain class is prevailing, further differentiation into sub-classes follows, for example, music could be further classified according to mood or genre by means of the sub-classifier 139. This multi-stage processing may be realised by cascading independent classifiers 137, 139 each dedicated to a specific classification task, or alternatively by (re-)training the same generic system.

Although only two classifications are shown here, it can be appreciated that additional classifiers may be cascaded to deal with complex classifications.

The system further allows discarding or reducing the contribution of some classes, based on the expectancy about their discriminative power inside the given video genre. For example, sub-class "applause" or "crowd" of the class "noise" is expected to be more relevant for, say, talk shows, and less relevant for, say, feature movies. This is effectively achieved by weighting, that is, assigning an element from pre-defined vectors \( \alpha = [\alpha_0, ..., \alpha_m] \) and \( \beta = [\beta_0, ..., \beta_M] \) output from the control unit 117. One weight vector \( \beta \) is applied to the sub-class probabilities output from the sub-classifier 139 on the second output terminal 129 by the second multiplier 127 to generate a first set of feature vectors of the
audio-class probabilities and the other $\alpha_i$ is applied to the class probabilities output from the classifier 137 on the output terminal 125 by the first multiplier 123 to generate a second set of feature vectors of the audio-class probabilities. In setting $\alpha_i = 0$, the contribution of class "i" is discarded. Alternatively, in the case of cascading classifiers, the corresponding classifiers could simply be switched off, by the control signals on the control terminals 119, 121 switching either the classifier 137 or sub-classifier 139 off. The weights and rules for assigning them may be stored in the form of a look-up table (not shown here).

The obtained sets of feature vectors of audio-class probabilities output from the first and second multipliers 123, 127 are added by the adder 131 and fed to the analyser 133 where the vectors are interpolated and analysed inside a sliding window, as described below with reference to Figure 2. The output of the analyser 133 is fed to the segmenting means 113. The output of the analyser 133 is a sequence of video frames indices, each designating positions of a semantic transition or scene change. As a final refinement, each of the indices may be aligned with the nearest camera shot-cut detection extracted by the camera shot extractor 107 which is also fed to the segmenting means 113. Other refinement strategies and supplemental features are clearly also conceivable, but are not discussed here.

The computation of the audio-scene transitions is done in two stages in the analyser 133. The first stage is used to compute the probability of a transition at each time instance on the basis of the class probabilities. The time instance corresponds to one video frame. This gives a time-dependent signal, in which the local maxima correspond to the points where (locally) the transition is most probable. The second stage extracts the local maxima in this signal (conforming with specific constrains) and defines them as the semantic transition points.

This two stage process will be described in more detail with reference to Figure 2. Basically, the first stage comprises the steps of: buffering, step 201, input feature vectors output from the adder 131 of Figure 1. The feature vectors are then clustered, step 203, inside the buffer by means of a k-means algorithm. The feature vectors are substituted by an index of the cluster to which they have been assigned. The transition probability at the temporal point corresponding to the centre of a sliding analysis window is computed, step 205. First, all the feature vectors inside the window are clustered. Second, the distribution of the cluster indices inside the window is evaluated to see whether they are distributed uniformly along the entire window, or the first and second half of the window contain essentially different cluster indices.
In more detail, in step 203, the feature vectors contained in a window are converted into cluster indices. This is done firstly to reduce the complexity of the problem, as the multiple dimensions of the feature vectors became just a single number from a finite set. But it also has the important effect of grouping together the parts of the signal that resemble each other, which will be necessary to detect the probability of transition.

A well-known k-means clustering algorithm is used. It is not critical to obtain a perfect clustering, so any initialisation of the algorithm will do. Moreover, it is not necessary to run the algorithm for a long time. After a few iterations the clusters will be consistent enough to go to the next stage. After running the k-means, each feature vector is substituted by the index of the cluster in which it belongs.

From an original signal feature vectors are extracted as follows. Let \( N \) be the number of features and \( \chi_{n[i]} \) the value of the \( n \)-th feature of the \( i \)-th sample can now be written as:

\[
x[i] = (X_1[i], X_2[i], \ldots, X_N[i])
\]

Let \( w \) be the width (number of samples) of a window, starting at the \( j \)-th sample, and let \( M \) be a desired number of clusters. The feature vectors enclosed by the window can be arranged as columns of a matrix

\[
X[J,W] = \{x[j], x[J + 1], \ldots, x[J + W]\} \quad j = S - (i - l) + 1, \quad i = 1, 2, \ldots,
\]

and the results of clustering as elements of a vector

\[
y[j,W] = (y[j], y[j+W], \ldots, y[j+W]) \quad y[w]e C_y = \{c_1, c_2, \ldots, c_M\} \quad j \leq w \leq j + W
\]

since each feature vector will be assigned a single cluster-index. In both equations, \( S \) designates the shift of the window, for example, round \((0.1 \cdot W)\).

This is illustrated in Figures 3a, 3b, and 3c. Figure 3a is a graphical representation of the class probabilities calculated for a sample content item. This was carried out over 30 frames using four predefined audio classes, namely, silence, music, speech and noise. Figure 3b is a graphical representation of the cluster indices of the
probabilities shown in Figure 3a derived from the method described above. Figure 3c is a graphical representation of the assignment of the cluster indices of Figure 3b.

The next step is to determine the frequency with which each of the cluster indices appear inside the window, and use this to compute a measure of dissimilarity between the first and the second half of the window. This is computed by computing mutual information between the two vectors of cluster indices, each covering one half of the window, step 205. This is equivalent to computing the relative entropy (the so-called Kullback-Leibler distance) between the given window and another one with different indices in each half. It is also equivalent to subtracting the entropy of the window from the sum of the entropies of the halves, which gives us a hint that a comparison between the halves is being held.

The probability of the occurrence of class \( C_i \) is now computed as

\[
P_a = \frac{\text{number of samples that } y[w] = C_i}{W} \tag{4}
\]

The entropy of \( y = y[w] \) is defined as:

\[
H(y) = -\sum_{c \in C} P_{C_i} \log(P_{C_i}) \tag{5}
\]

The probability of a (scene) transition is now computed as the mutual information between \( y \) and \( z \)

\[
P = H(y; z) = H(y) + H(z) - H(y, z) = H(y) + 1 - H(y, z) \tag{6}
\]

with \( z \) defined as:

\[
z = z[w] = \begin{cases} 
0 & 1 \leq w \leq \frac{W}{2} \\
1 \frac{W}{2} + 1 \leq w \leq W
\end{cases} \tag{7}
\]
and $H(y, z)$ being the "joint entropy", i.e. entropy computed by taking the joint probability distribution of $y$ and $z$ (probability distribution of a signal obtained by concatenation of vectors $y$ and $z$).

Mutual information and the relative entropy (Kullback-Leibler distance) are commonly defined as given below.

$$H(y, z) = D(p(y, z) || p(y)p(z)) = \sum_{y \in C_y} \sum_{z \in C_z} p(y, z) \log \frac{p(y, z)}{p(y)p(z)}$$  \hspace{1cm} (8)

$$D(p \parallel q) = \sum_{x \in X} p(x) \log \frac{P(x)}{q(x)}$$  \hspace{1cm} (9)

Here, $P(x)$ and $q(x)$ are two probability density functions, describing the probability of occurrence of each symbol $x$ from $X$.

It should be realised that, by sliding the window, step 213, after each such computation, one obtains a time-dependent signal reflecting the probability of transition in the entire content. This is illustrated in Figures 4a and 4b.

Figure 4a is a graphical representation of the audio class probabilities of a sample content item. This was carried out over 500 frames using four predefined audio classes, namely, silence, music, speech and noise. The class probabilities are clustered and the transition probabilities are computed as described above. The transition probabilities of the class probabilities of Figure 4a are graphical shown in Figure 4b.

The computed transition probabilities are buffered, step 207 and shown in Figure 4b, significant transitions are indicated by local maxima, step 209. Therefore a point will be considered an audio-scene transition if it is the absolute maximum of the transition-probability signal inside a window centred at that point and having a given size, step 211. This size will be a parameter that can be adjusted to recall more or less of the candidate peaks, leading this way to a coarser or narrower segmentation - or, the signal is benchmarked against some ground truth (for example, obtained by manual annotation), to increase either the precision or the recall of the expected transitions.

The formal definition of $H(y, z)$ is given by (8) and (9), and the proof equation (6) can be found. The equality $H(z) = 1$ in (6) follows straightforward from definitions (6) and (7) above. (Note that for $z$ defined as in (6), the set of possible classes is given as $C_z = \{0, 1\}$.)
Finally, given a chosen peak-distance $V$, it is decided that a transition happened at time instance $i$ when

$$p[i] \geq p[j], i \leq j \leq i + v$$

Consequently, the choice of $V$ effectively determines the minimal distance between any two consecutive transitions (i.e. scene-change detections).

The method of the embodiment of the present invention has the ability to produce segmentation that approaches the ground truth when it comes to the total number of scenes. Current solutions, e.g. combining audio-silence and shot-cut detection, show a typical over-/under-segmentation. The method of the present invention is much less susceptible to any ambiguity of definition and detection of a semantic audio-silence, as it considers probabilities of multiple audio classes at a same time.

Although the preferred embodiment discloses segmentation on the basis of audio classification, it can be understood that the segmentation may also rely on video classification in addition to audio classification.

The present invention provides the user with improved capabilities for editing and browsing video databases and libraries (including commercial skip), for example those constructed from personal TV recordings. The same feature can also be employed to aid other content analysis systems that can benefit therefrom, such as those aimed at video summarisation. The feature in particular applies to consumer systems such as Personal Video Recorders, for example based on DVD and HDD.

Although a preferred embodiment of the present invention has been illustrated in the accompanying drawings and described in the foregoing detailed description, it will be understood that the invention is not limited to the embodiment disclosed, but is capable of numerous modifications without departing from the scope of the invention as set out in the following claims.
CLAIMS:

1. A method for segmenting a content item, the method comprising the steps of:
   for each of a plurality of frames of said content item, calculating class
   probabilities, the class probabilities including the probabilities that said frame belongs to a
   plurality of pre-defined classes or subclasses;
   determining the location of a transition in said content item upon detection of a
   change in at least two of the calculated class probabilities occurring at the same time; and
   segmenting said content item into a plurality of portions based on the location
   of the transition in said content item.

2. A method according to claim 1, wherein said content item comprises video
   content and audio content and said audio content comprises said plurality of frames.

3. A method according to claim 1 or 2, wherein the location of the transition in
   the content item is determined when a change in the at least two of the calculated class
   probabilities exceeds a predetermined threshold.

4. A method according to claim 1, 2 or 3, wherein the calculated class
   probabilities of each frame are grouped to form a feature vector.

5. A method according to claim 4, wherein calculated audio class probabilities
   are weighted and filtered according to the genre of the video content before being grouped.

6. A method according to any one of the preceding claims, wherein the step of
   determining the location of the transition in the audio content comprises the step of:
   computing the location of the transition from derived local maxima of the
   calculated class probabilities within a predetermined time interval.

7. A method according to claim 6, wherein the predetermined time interval
   corresponds to a frame and the location of the transition is computed for each frame.
8. A method according to claim 6 or 7, wherein the step of computing the location of the transition comprises the steps of:
   (a) estimating a potential location of a transition within each predetermined time interval;
   (b) making the potential location of a transition the temporal centre of a sliding analysis window;
   (c) clustering all the features vectors within said sliding analysis window;
   (d) determining the distribution of the clusters within said sliding analysis window;
   (e) changing the potential location of the transition;
   (f) repeating steps (b) to (e) until a uniform distribution of the clusters is achieved; and
   determining the location of the transition as the temporal centre of the sliding analysis window having the uniform distribution.

9. Apparatus for segmenting a content item, the apparatus comprising:
   an audio classifier for calculating class probabilities, the class probabilities including the probabilities that each of a plurality of frames of said content item belongs to a plurality of pre-defined classes;
   an analyser for determining the location of a transition in said content item upon detection of a change in at least two of the calculated class probabilities occurring at the same time; and
   segmenting means for segmenting said content item into a plurality of portions based on the location of the transition in said content item.

10. Apparatus according to claim 9, wherein said content item comprises video content and audio content and said audio content comprises said plurality of frames.

11. A computer program product comprising a plurality of program code portions for carrying out the method according to any one of claims 1 to 8.
FIG. 1
Feature data

201 Buffer

203 Cluster

205 Compute "mutual information" and probability of transition

207 Buffer a number of transition probabilities, equal to window length

209 Determine local maxima and transition points

Audio-scene transition

211 End of sequence?

213 Slide window

No

Yes Exit

FIG. 2
FIG. 4a

FIG. 4b
FIG. 5

Total = 139
Over-Segmented (OS) = 31