Unsupervised scene detection and commentator building using multi-modal chains

Gert-Jan Poulisse · Yorgos Patsis ·
Marie-Francine Moens

© Springer Science+Business Media, LLC 2012

Abstract This paper presents a novel unsupervised method for identifying the semantic structure in long semi-structured video streams. We identify *chains*, i.e., local clusters of repeated features from both the video stream and audio transcripts. Each chain serves as an indicator that the temporal interval it demarcates is part of the same semantic event. By layering all the chains over each other, dense regions emerge from the overlapping chains, from which we can identify the semantic structure of the video. We present two clustering strategies that accomplish this task, and compare them against a baseline Scene Transition Graph approach. We then develop a commentator that provides a semantic labeling of the resultant video segmentation.

Keywords Semantic event detection · Feature extraction · Multi-modal scene segmentation · Video summarization

1 Introduction

Over a period of two weeks, we recorded large swathes of the BBC, Beijing 2008 Olympic coverage in order to develop a methodology to automatically segment long video streams into its constituent semantic components, which we refer to as the scenes or semantic events.

G.-J. Poulisse (✉) · M.-F. Moens
Department of Computer Science, Katholieke Universiteit Leuven, Leuven, Belgium
e-mail: Gert-Jan.Poulisse@cs.kuleuven.be

M.-F. Moens
e-mail: Marie-Francine.Moens@cs.kuleuven.be

Y. Patsis
IBBT, Ghent-Ledeberg, Belgium

Y. Patsis
e-mail: gpatsis@etro.vub.ac.be

Y. Patsis
ETRO—DSSP, Vrije Universiteit Brussel, Brussels, Belgium

An example of a scene would be an Olympic event such as the 100 meters butterfly, from the moment the athletes emerge from their changing rooms, dive into the water, touch home at the race conclusion, and any post-race follow up. The scene concludes when the next semantic event begins, often the next race. This paper focuses on identifying such semantic events, with the discovered semantic scenes then serving as input to an automatic event summarizer and commentator.

Olympic video is a challenging subject matter; it is not as expository in nature as news video, which we examined previously in [17], and as such the information content contained within the audio transcripts is generally poorer and less informative. A single Olympic scene may be five minutes or an hour long, and due to the more ad-hoc nature of the continuous live coverage, there are less discernible production rules in comparison to news broadcasts. One difficulty faced was the need to distinguish between several semantically distinct, but visually similar events that followed one after the other (i.e. a sequence of swimming races) as a result of the continuous coverage. In short, we found that our methodology for successfully segmenting news video [17], with its reliance on a number of textual features, was not appropriate to the current task, and that a different, multi-modal approach was necessary, presented in section 3.

The segmentation of expository text into subtopics, work done by [7], introduced the notion of text-chains to measure lexical cohesion. Repeated terms formed textual chains, and changes in chain similarity within an interval reflect the change in discourse structure and can be used to split the text. We adapt and extend this basic idea to video scene identification. Each repeated feature demarcates an interval, and indicates that the interval under consideration is part of the same semantic event, section 4.

The idea is to extract features from several modalities and to unify them in a universal system of multi-modal chains, each representing a particular type of similar feature. Once in this representation, the contribution from each multi-modal feature is equivalent. When viewed on the time axis, dense regions emerge from the overlapping chains, thus outlining the semantic structure of the video. A preliminary description of this methodology has been presented by the authors in [16].

We use two clustering strategies to segment the video, and analyze their resultant performance in section 5, and when augmented with audio in section 6. Up till this point, our work has been unsupervised. With the addition of a small amount of extra knowledge automatically acquired from Wikipedia articles, we develop a commentator, section 7, to add some semantic meaning to our recently discovered Olympic scenes. We identify the sport, some relevant keywords, and the participating athletes.

The contribution of this work is the segmentation of video from a challenging domain into its constituent semantic units-scenes using multi-modal chains of repeated features, and then to provide a semantic annotation and indexation for these scenes.

2 Related work

Video analysis is an important task in the frame of video summarization, indexing and retrieval. Many video programs have story structures that can be recognized through the clustering of video contents based on low-level visual features and the analysis of the temporal arrangement of the composing elements, possibly leading to a hierarchical decomposition of the video [3, 13, 22, 27]. Video analysis often relies on a multi-modal approach, blending evidence from different sources including the transcribed audio text and the imagery [23]. For instance, event and speaker identification can be

used for movie indexing [9]. Rarely, specific cinematic rules have been relied on for content analysis [10].

In discourse theory, [21] several discourse structures are described for text, such as sequential, ringed, monolithic, and piecewise. Video of expository nature typically only has a sequential structure of semantic events. Nonetheless, we posit that occasionally there is some hierarchy. Many of the 5–10 minute semantic events that occur in the Olympic footage occur in a flat sequence. However, longer events (half an hour or more), such as the gymnastics coverage, usually possess some shallow hierarchy. In such a scenario, gymnastics would naturally be the encapsulating semantic event, but there would be subordinate segments, such as the performance of one athlete on the parallel bars, followed by a different segment of a different athlete, such as the balance beam. We feel that multi-modal-chains are well suited towards exploiting the sequential nature of the semantic events in the Olympic footage. The challenge, however, is in discerning the overarching semantic scene hidden behind the subordinate sequences.

Our work is most closely related to research done in the TRECVID 2004 news story segmentation task, illustrated by such works as [1, 18] in the sense that both seek to segment long video streams. Our task is harder, as news stories have a more formalized structure, a richer set of underlying features, and are typically more focused in their description of a semantic event (five to ten minutes for a news story while certain Olympic events can be up to an hour long).

In terms of subject matter, our work is of a more generic nature and inspires comparison with sports video analysis. This topic includes sports tactics summarization (i.e., classifying a play sequence into an existing tactic pattern and recognizing unknown patterns) and semantic event recognition within a sports game [2, 4]. The latter type of analysis is relevant for this paper. Sometimes video and transcribed audio information is augmented with closed-caption [14] or webcasted text, in order to improve the event recognition and the recognition of the players and their team [24]. Webcast text is a text broadcast channel for sports games, which is co-produced with the broadcast video and is easily obtained from the Web. For instance, [25] analyzes Webcast text to cluster and detect text events in a supervised way using probabilistic latent semantic analysis.

It should be noted that semantic event detection in sports video typically happens in a supervised fashion within a sports game [2]. In contrast, we segment video containing different sports. Occasionally multiple, distinct instances of the same sport occurs in sequence, which also have to be differentiated between. To do this, we use the multi-modal channels of the sports video by aligning chains of named entities in the transcribed speech with chains of visually related shots in a completely unsupervised way.

With regard to automated annotation of the sports video segments, several approaches relying on supervised classification of video features, among which are audio features such as typical game sounds [11, 19, 26]. Examples of this include crowd cheering, referee whistling, commentator excitement and the sound of the ball being struck in a tennis game or hitting the rim in basketball [11, 26]. In [19], important sporting occurrences are recognized when the accompanying energy signal is in excess of the average broadcast signal. [20] performs speaker identification when segmenting documentary films, with the changing distribution of speakers influencing the final partitioning.

In the sports video segmentation approach of [11], segments are annotated based on the presence of sports keywords in subtitles or transcribed speech, which are defined in a manually constructed ontology. We take the approach of learning the keywords that are typical for a particular sport from external documents, specifically Wikipedia articles. This follows the recent trend of using Wikipedia information for automated knowledge

acquisition, such as done in [12], where Wikipedia category names and structure are used to induce knowledge of concepts, their relations, and attributes. In conjunction with the extraction of salient named entities, this allows us to summarize and implicitly categorize our segmented video scenes.

3 Features

We have developed two types of multi-modal chains, one based on text and the other based on visual features. Chains are an abstraction; they represent a temporal grouping of features. This abstraction makes it possible to later fuse many different types of features in an ensemble to identify scenes in the video.

Our textual source consists of Teletext subtitles that are enclosed with the broadcast signal. As one can see in the transcript provided below, some of the sentences have lexical inconsistencies or are incomplete. We are unsure whether this is due to errors at the signal origin or due to errors in recording and reconstituting the signal at our end. Also, the quality of the transcript is entirely in the control of the broadcaster. In our experience, scripted shows, such as news broadcasts, have much more complete transcripts; the text that is read out is often very close to the one provided by the Teletext. A broadcast such as the Olympics is different because it is broadcast live, with a corresponding drop in transcription accuracy. The segment below corresponds to an entire five minute race, and much of the commentators' dialogue, which is audible in the video, is omitted in the subtitles (Fig. 1).

The most prominent features for determining scene boundaries in this excerpt are the athlete names. The remaining non-trivial terms, e.g. 'swim', are considerably less discriminative as this particular swimming event is preceded and followed by additional swimming races.

As such, we focus our efforts on identifying named entities (typically person names, names of organizations, and locations). We used the state of the art, Stanford Named Entity

```
Four people make a massive Andrew Bree of violence when well to
make this semi-final.
Right next to Kosuke Kitajima of Japan, the world record holder
and defending Olympic champion.
I thought Andrew Bree swam really well.
He has a good chance of making the final.
He had a lifetime best to reach the semi-.
All the guys who make it to the final are probably looking at
lifetime marks did he not understand.
Look how powerful kipper Gema looks.
Kosuke Kitajima.
Andrew Bree is in a pretty good condition - position.
while it is going well.
If he can get a good touch here, he has a good chance of making
the final.
The time, a new Olympic Andrew Bree finishing in 5th.
Another good swim for him.
```

Fig. 1 Teletext transcript of the Men's 200 m breaststroke 1st semi-final

Recognizer, which is based on conditional random fields with Gibbs sampling, to label all the named entities in the text [6].

Within the context of a sliding window, T_1 , named-entity chains are made linking sentences containing near-identical named entities based on their Levenshtein edit distance. Thus a particular named entity has to occur within one window's length of the previous instance of that named-entity for it to be added to the chain. The resultant named-entity chains are in essence local clusters. The reason for this constraint is to avoid the not too inconceivable case where named-entities occur in distinct scenes far apart in the video. In such a scenario, the current setup would produce two separate chains for the same named entity, each representing a different temporal position within the video. Based on our experience in [17], where we found that a suitable window size was typically proportional to the expected size of the semantic event, we chose a window size of twenty-five sentences, roughly equivalent to five minutes of video. Typically named-entity repetition will have occurred by this point (even in semantic scenes far longer than five minutes), and a bigger window only introduces errors, usually because named-entities are matched across several smaller semantic events. We create named-entity chains as follows:

- A Levenshtein edit distance threshold lt has been empirically established on a held-out validation set. Based on the size of the sliding window controlling the temporal interval T_1 , chains $c_i = \{ne_a, \dots, ne_b\}$ are constructed for all m named entities NE in the video, $NE = \{ne_1, \dots, ne_m\}$ where $Lev(ne_a, ne_b) < lt$ and $|a-b| < T_1$. Indexes a and b correspond to temporal positions in the video. Timing information provided by the Teletext transcript allows for the time alignment of the text with the associated video shot.

To exploit the visual information present, we make an equivalent feature, based on the spatial-pyramid feature developed by [8]. In essence, this procedure extracts SIFT features for an image which are then transformed to a set of visual words and then sampled over increasingly fine sub-regions. This feature has been shown to work well in recognizing scene categories. Our choice was motivated by its ability to capture the gist of a scene; the ability to identify visually related images such as stills with near-similar backgrounds, but different objects or persons in the foreground. The spatial-pyramid kernel is given by:

$$I(H_X^l, H_Y^l) = \sum_{i=1}^D \min(H_X^l(i), H_Y^l(i)) \quad (1)$$

$I(H_X^l, H_Y^l)$ is the histogram intersection function, which determines the number of matching points at resolution l within the spatial-pyramid grid position i of the image. In a d -dimensional feature space, the number of sampling locations is $D=2^{dl}$. When L represents the total number of spatial-pyramid resolution levels, the kernel $k^L(X, Y)$ calculates the summation of every intersection of each sampled histogram at each grid resolution; the histogram intersection is given a weight proportional to each resolution level. We use $L=4$ as suggested by [8].

$$k^L(X, Y) = \frac{1}{2^L} I(H_X^0, H_Y^0) + \sum_{l=1}^L \frac{1}{2^{L-l+1}} I(H_X^l, H_Y^l) \quad (2)$$

We perform automatic shot cut detection on our Olympic video to partition the underlying frames $\{f_1 \dots f_n\}$, into shots $S = \{s_j\}_{j=1}^J$. The frames are grouped such that each shot is given by a sequence of frames $s_j = \{f_{c_j} \dots f_{d_j}\}$, $1 \leq c < d \leq n$ using the visual similarity kernel given by equations 1 and 2 [8]. The resultant shot-segmentation is convincing, as upon

evaluation the output is always within a few frames from a reference boundary. We note that the shot-cut detector is robust against smooth transitions such as wipes, and that fades and dissolves are infrequent in the data-set. Fundamentally, we trust the resultant shot-cuts, and we believe that it minimally influences the semantic scene detection error.

From each shot, we select representative frames. These frames are then matched using the previously described spatial pyramid kernel thus forming chains of visually related shots. The procedure was as follows:

- A kernel threshold kt is tuned on a held-out validation set so that visually related shots such as the ones in Fig. 2 are identified. Like with the named-entity chains, we constrain the matching of shots to within a certain interval T_2 (we found that an interval of 40 shots was more than sufficient) to avoid temporally disparate matches.
- Chains $c_i = \{s_a, \dots, s_b\}$ are then constructed for all n shots S in the video, $S = \{s_1 \dots s_n\}$ where $k^L = (s_a, s_b) < kt$ and $|a - b| < T_2$. Indexes a and b correspond to temporal positions in the video.

Multi-modal chains are indicators of repeated temporal occurrences within a portion of a video segment. Once chains have been generated, the contribution of constituent features becomes indistinguishable for the eventual scene determination.

4 Scene clustering

We have developed two clustering strategies which use multi-modal chains in order to discern the semantic structure of the video, and these are compared against a baseline Scene Transition Graph approach [27] from the literature.

Fig. 2 Two pairs of visually related images

The first is essentially a greedy bottom-up approach. All multi-modal chains are initially ordered sequentially in time, and then one by one all temporally overlapping chains are merged together. Since chains consist of at least two or more shots, it is reasonable to suppose that a number of shots in the video are not members of a chain, nor are they bounded by one (that is, the shot does not lie between the start and end of any chain). We assume most characteristic descriptors for a semantic event are introduced in the beginning and middle sections, and that they taper off as the scene progresses. If this holds for video production, most chains will have been established early on in the scene, and non-repeating shots may occur at the end. These shots, unbounded by any chains, are merged with the closest preceding chain, thus extending the bounds of the returned scene.

The resultant merged chains, and any remaining unmerged chains, describe semantic events with boundaries corresponding to the first and last occurrence in time of the entities that make up each chain. We refer to this approach as Greedy-Clustering in our reporting in section 5, because of the way scenes are formed by ‘greedily’ merging overlapping chains.

- Step 1: Using chains c_i from section 3, initialize the set of scenes Sc where $sc_i = \{c_i\}$. Each chain c_i is assigned to a single scene. Each $c_i = \{c_i^{head}, c_i^{tail}\}$ where c_i^{head} represents the first occurrence of the underlying feature in the video and c_i^{tail} the last occurrence.
- Step 2: Scenes are merged till no further action is possible. Scenes are merged when chains temporally overlap, i.e.

$$(c_1^{head} < c_2^{head} < c_1^{tail} < c_2^{tail}) || (c_1^{head} < c_2^{head} \&\& c_1^{tail} > c_2^{tail})$$

The new scene boundaries for sc^* are given by

$$\{\min(c_1^{head}, c_2^{head}), \max(c_1^{tail}, c_2^{tail})\}$$

- Step 3: Any shot $S = \{s_i^n\}$ not part of an overlapping chain is assigned to the closest temporally preceding scene sc .

Figure 3 is an illustrative example from the Women’s 200 m freestyle final. It contains one named-entity chain, and two shot-chains. Each entity’s position is relative to its temporal occurrence in the video. In Fig. 3, all three chains overlap and so can be merged; in this example the resultant scene would run from the first utterance of ‘Commonwealth’ till the conclusion of the second shot containing the swimmer facing the water.

Our second strategy, referred to as Density-Clustering in our reporting in section 5, involved sampling the density of overlapping chains. We count the number of multi-modal chains that occur at every frame. This can then be plotted, and after smoothing, the resultant local minima identify the scene boundaries.

- Step 1: Given chains c_i from section 3, where $c_i = \{c_i^{head}, c_i^{tail}\}$, c_i^{head} represents the first occurrence of the underlying feature in the video and c_i^{tail} the last occurrence. Each chain is sampled over the length of the video V , such that at position j in V , given by $V[j]$, $V[j]$ is incremented between $head < j < tail$.

Fig. 3 Greedy-clustering

- Step 2: A smoothing, least-squares filter¹ is applied to $V[j]$, such that

$$V[j]_{smooth} = smooth(V[j])$$

- Step 3: Scene boundaries are placed at local minima in $V[j]_{smooth}$. A local minimum function places scene boundaries in $V[j]_{smooth}$ where local minima exceed a threshold δ . δ is an empirically defined threshold, and we performed experiments using two thresholds; δ_1 and δ_2 .

This process is best illustrated graphically, in Fig. 4 using the same multi-modal chains of Fig. 3. Each black horizontal bar represents a multi-modal chain that runs for a certain time interval, represented on the x-axis in terms of frames of the video. The grey bars then represent the cumulative number of multi-modal chains present at each frame position, from which we can then infer a curve, whose local minima (the red circles) represent scene boundaries.

We compare these methods against a segmentation using a Scene Transition Graph (STG) [27] segmentation process. In this methodology, clustered shots, similar to shot-chains, form vertexes in a graph. We also include our clustered named entity feature, as additional vertexes. A directed edge is drawn from one vertex to another representing the video progression, one shot transitioning to another. Edges which, if removed, divide the graph into two disconnected graphs are known as “cut-edges”. After removing all cut-edges from the STG, each disconnected sub-graph represents a scene, with boundaries at the cut-edges. A scene consisting of multiple shots presents as a cycle in a STG.

The chain paradigm can be seen as a temporal graph, with one node per chain, and edges representing the temporal transitions between occurrences of chain members. A single chain then, forms a cycle; multiple co-occurring chains simply extend the length of the cycle in the sub-graph representing the underlying scene. As with STG segmentation, scenes are found by splitting the temporal graph into distinct sub-graphs. The difference with STG segmentation is in the treatment of single shot-clusters with only one incoming and outgoing cut-edge; in the Greedy approach these are merged into the immediately preceding scene formed

¹ Matlab: Savitzky-Golay filter, which is a moving average with filter coefficients determined by an un-weighted linear least-squares regression and a polynomial model of specified degree (degree 7 used here)

Fig. 4 Density-clustering

by a chain, while in the STG approach each forms a single scene. Density-Clustering, is different from the Scene Transition Graph approach since a scene is identified by the relative degree of repetition over a timespan.

5 Results and analysis

We evaluate our scene detection experiments on 4 h of BBC coverage of the Beijing 2008 Olympics, corresponding to a day's worth of sporting events. This particular day's events consist mostly of various swimming races, interspersed with other sporting highlights and post race interviews along with coverage of the women's gymnastics finals and the women's cycling time-trial. We take the officially listed Olympic events as the basis for our ground-truth semantic events. For example, the Men's 200 m breaststroke 1st semi-final would constitute a semantic event in our ground-truth. We augment this with additional labeling, such as when the BBC produces highlights or breaks away from the coverage for a general wrap-up of the day's performances in the studio.

There is some discussion possible about established evaluation metrics for semantic event detection in video. The TRECVID 2004² story segmentation task is evaluated using a variant of precision and recall; boundaries are counted as corrected matches if they are placed within five seconds of the ground truth reference boundaries.

$$\begin{aligned} \text{Precision} &= \frac{\# \text{scene boundaries correct}}{\# \text{scene boundaries found}} \\ \text{Recall} &= \frac{\# \text{scene boundaries correct}}{\# \text{ground truth scene boundaries}} \\ F - \text{measure} &= 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \end{aligned}$$

The problem with this approach is that it leaves little room for 'partial credit', i.e. boundaries that are placed more than five seconds from the reference but still more than adequately demarcate a scene. A user performing an information retrieval task would only mildly be affected if they were presented with a thirty-minute scene which is one minute shorter or longer due to a misplaced boundary. Another point worth raising is that not every semantic event ought to be treated as equal. Short, i.e. less than one-minute segments such as typical program highlights, rarely contain much information, are sparse in features, and are thus difficult to detect. In some cases they are simply highlights, precursors to longer segments appearing later in the program. The failure to detect these short segments distracts from the ability to successfully recognize longer, more relevant scenes.

² <http://www-nlpir.nist.gov/projects/tv2004/tv2004.html>

We evaluate our two clustering strategies with the TRECVID style precision and recall. We present our Greedy-clustering results, as well as two variants of our Density clustering method using differing local minima thresholds δ_1 and δ_2 . The different thresholds trade precision for recall. Both segmentation approaches are highly dependent on repetition. While they are quite able to capture the center of mass of a story, data at the boundaries do not contain sufficient repetition for exact boundary placement. Consequently, we use a boundary evaluation window of one minute. We also provide the results of our implementation of the Scene Transition Graph methodology [27] applied on our dataset. Both shot and named-entity features are used, with all other parameters unchanged.

The results of Table 1 suggest that, at best, around half (recall=0.48) of the proposed scene boundaries are placed within one minute of the reference boundaries. However this does not give a comprehensive insight into the quality of the identified scenes. A single scene is only truly identified when both its boundaries are correctly placed.

In the case of under-segmentation, the returned video segment contains a number of semantically distinct scenes. We found that Greedy-Clustering tended towards under-segmentation, while Density-Clustering and STG tended to over-segment. Part of the STG over-segmentation is attributed to singleton shot clusters which get partitioned into distinct sub-graphs. We feel that a smaller, over-segmented scene is of greater utility for browsing or indexing purposes than an under-segmented one. In this respect, Density-Clustering- δ_1 is of greater utility than Greedy-Clustering, as it is less prone to under-segmentation.

If the boundaries assigned by a particular method are off, we can still assess whether the segmentation has largely circumscribed a particular semantic event. This can be quantified in terms of a metric which we call the event-coverage, i.e. the proportion of unique scenes found in a particular segmentation in relation to the total number events in the ground truth.

$$\text{Coverage} = \frac{\# \text{unique events in hypothesized segmentation}}{\# \text{events in groundtruth segmentation}}$$

An example which illustrates how this metric is computed is given in Table 2. A hypothesized segmentation (the last two columns) is compared against the actual ground truth segmentation (the first two columns). In the hypothesized segmentation, each boundary is aligned to the closest preceding boundary in the ground truth (based on frame number). The corresponding scene label from the ground truth is then used to label the associated scene in the hypothesized segmentation. Each non-duplicate label can then be counted for the event-coverage computation.

In the hypothesized segmentation of Table 2, over-segmentation, e.g. “Men’s 100 m freestyle 1st semi-final”, only contributes one unique event to the coverage count. Under-segmentation, e.g. “Women’s 200 m butterfly 1st semi-final” is not correctly identified in the hypothesized segmentation and so is not counted. The resultant coverage is then 4/7.

Table 1 Olympic event detection results

27 ground truth boundaries	Greedy-Clustering	Density-Clustering δ_1	Density-Clustering δ_2	STG
Boundaries found	18	31	21	27
Boundaries correct	10	13	10	9
Precision	0.56	0.42	0.48	0.33
Recall	0.37	0.48	0.37	0.33
F-measure	0.44	0.45	0.42	0.33

Table 2 Alignment of a reference segmentation and hypothesized segmentation for the purpose of event-coverage calculation. Each cell corresponds to a scene in a segmentation

Ground truth boundaries in terms of frame number	Ground truth scene label	Hypothesized boundaries in terms of frame number	Scene label in hypothesized segmentation
86024	Men's 100 m freestyle 1st semi-final	87104	Men's 100 m freestyle 1st semi-final
		91702	Men's 100 m freestyle 1st semi-final
98149	Men's 100 m freestyle 2nd semi-final	105051	Men's 100 m freestyle 2nd semi-final
106300	Women's 200 m freestylefinal		
116675	Men's 200 m butterfly final	124807	Men's 200 m butterfly final
127523	Caitlin McClatchey interview		
128698	Women's 200 m butterfly 1st semi-final		
137724	Women's 200 m butterfly 2nd semi- final	146141	Women's 200 m freestyle 2nd semi-final

The results in terms of coverage are presented in Table 3. We also calculated the coverage when discarding short scenes, which typically consist of highlights or previews. Such scenes typically consist of non-repeating shots and named entities, and are difficult to recognize as a result. By only allowing scenes longer than one minute (the typical length of a highlight segment), 8 semantic events were discarded, leaving 19 segments longer scenes.

It turns out that both segmentation approaches presented had difficulties identifying segments that were under one minute, hence the considerable improvement when discounting the segments smaller than one minute from the ground truth for event-coverage calculations. To put this into perspective, 8 min of discarded events constitute less than 2 % of the complete video. Upon doing this however, it also becomes clear that Density-Clustering- δ_1 is able to identify 80 % of all longer semantic events in the video material which is more relevant to the user. The primary reason these, overall, less relevant, small events are not identified is because there is little repetition that can be captured within the chain paradigm.

6 An audio extension

We have incorporated an audio classifier into our system in an effort to improve the effectiveness of our Density-Clustering approach. Based on work in [15], the classifier is

Table 3 Alignment event-coverage

	Greedy- Clustering	Density-Clustering δ_1	Density-Clustering δ_2	STG
Coverage	0.48	0.56	0.48	0.35
Coverage discounting segments <1 min long	0.63	0.79	0.63	0.50

used to distinguish speech, music, silence and noise in an audio fragment. The classifier is trained on BBC Olympic broadcasts recorded on different days, distinct from the dataset used for our scene segmentation analysis.

User defined duration of the spotted silence segments is added in order to provide cues for thematic changes in the audio clip. An implementation in Weka³ of the C4.5 algorithm (Decision Trees J48) is used as a classifier and the duration of a segment for decision making is set to 640 msec. The intuition is that the changes in audio condition at shot boundaries indicate a genuine scene change, while uninterrupted sound over a shot transition merely indicates a change in visual perspective.

The audio feature does not form a chain-type feature, as it is simply a cue for a possible scene change, and cannot therefore be readily integrated in our clustering strategies. The resultant output of the analysis of our audio feature then is all positions in the video that lie on shot boundaries and have changing audio content to either side. We combine this with our Density-Clustering strategy, giving the positions suggested by the audio feature a negative weight. These are interpolated into our density plot curve (see Fig. 4), after which we identify the local minima which represent the scene boundaries for this method. We improve on our initial Density-Clustering- δ_1 run by interpolating in the audio feature using differing local-minima thresholds μ_1 and μ_2 . The results are presented in Table 4.

The inclusion of the audio feature shows an overall decrease in F-measure due to a tendency to over-segment. Nonetheless, the higher recall and increased coverage in Audio- μ_1 illustrates the ability to discern more semantic events, which is of potential value for browsing applications for example.

7 A sports commentator

Once scene cuts have been produced by the segmentation engine, a natural progression then is to give some semantic connotation to these scenes. We do this by building a scene commentator. Our above procedures are unsupervised, and therefore additional knowledge is needed to bridge the semantic gap. We chose to build a commentator that has learned key words and phrases from an external corpus, i.e. Wikipedia, for various Olympic events using the Chi-square criterion. The commentator would then identify these words and phrases in the segmented scenes. Examining the “2008 Summer Olympics” Wikipedia⁴ page, 31 Olympic sports are listed, and they form the principle categories in our commentator. We crawled the respective Wikipedia entries for each sport to build up a corpus of potential keywords. A typical category structure starts with the page for the individual sport at the 2008 Olympic Games, e.g. “Gymnastics at the 2008 Summer Olympics.” Each listed sporting event in that category is then indexed, e.g. “Gymnastics at the 2008 Summer Olympics Women's artistic team all-around,” as well as every page that describes the sport in general, such as “Artistic Gymnastics.” Every category followed this general pattern and we thus assembled a collection of web-pages containing potential keywords.

We use the Chi-square test (at significance level $p=0.10$) to identify word n-grams (n ranging from one to four) in each of our 31 categories which occur significantly more frequently than random in our collection. As pre-processing we remove punctuation, and perform stemming and stop word removal. We also omit athlete names that are mentioned in certain Wikipedia articles (typically the medal winners of that

³ <http://www.cs.waikato.ac.nz/ml/weka/>

⁴ http://en.wikipedia.org/wiki/2008_Summer_Olympics

Table 4 The effect of audio on the segmentation process

	Density-Clustering- δ_1 (no audio)	Density- Clustering- (audio) μ_1	Density-Clustering- (audio) μ_2
Precision	0.42	0.22	0.30
Recall	0.48	0.74	0.52
F-measure	0.45	0.33	0.37
Coverage	0.56	0.66	0.48
Coverage discounting segments <1 min long	0.79	0.84	0.58

event) in an attempt to remain as generic as possible. For the gymnastics category, a representative sampling of n-grams found were: “apparatus”, “balance beam”, “women’s gymnastics”, “tumbling”, “uneven bars”, etc.

When identifying sports categories in our scene segments, we do not preclude the assignment of multiple categories to a single scene, which might be the case when under-segmentation occurs. Each segment was then processed for the presence of n-grams identified by the Chi-square test for each sporting category. We have evaluated the commentator on our best segmentation run, Density Clustering- δ_1 , by examining the annotations produced by the commentator for every segment. We count how many times our commentator correctly noted a sport in each segment, and the converse, how many times it correctly noted when a sport was not present in a segment. The segmentation under evaluation contained 5 of the 31 sporting categories, and the results are presented in Table 5.

We had considered the possibility of using the commentator results as a possible feedback for the segmentation process. However, the established sporting categories are too broad for the obtained scene segments. For example, we have a “swimming” category, but to be useful for our segmentation process we would also have to be able to distinguish between individual swimming races such as the “100 meter Men’s freestyle” and the “100 meter Men’s breaststroke.” To satisfy the chi-square criterion to identify such fine-grained topics in the scene segments would require a significantly larger volume of sports texts on these subcategories as training data. This data requirement is also the reason why the ‘Cycling’ category performed poorly. While other categories have sufficiently discriminate keywords, this is not the case here.

Our current commentator reports the sporting activities present in a segment, and the names of any athletes mentioned. Typical words for the sports category learned from the Wikipedia data and also found in the transcript are also reported. For instance, the commentator results describing the event shown in Fig. 5 are:

Table 5 Commentator performance in classifying Olympic-scenes

Density Clustering- δ_1	Archery	Fencing	Cycling	Swimming	Gymnastics
True Positive	3	1	9	16	7
False Positive	4	1	16	2	4
True Negative	24	28	6	13	20
False Negative	0	1	0	0	0
Actual number of segments	3	3	10	18	7
Accuracy	87 %	94 %	48 %	94 %	87 %

Fig. 5 Several automatically identified multi-modal chains that make up part of a gymnastics scene. Their horizontal position reflects their temporal position in the video

- Classification: Gymnastics.
- Keywords: apparatus, beam, deduction, difficulty, dismount, execution, floor exercise, gymnasts, height, landing, performed, routine, beam, floor, gymnast, vault.
- Athletes: Shawn Johnson, Liukin, Steliana Nistor, Alicia Sacramone.

8 Conclusion

We presented a novel, unsupervised approach for segmenting a long video stream. We extracted two multi-modal features to aid us in this task; one based on local clusters of similar named entities, the other based on local clusters of similar shots. These multi-modal chains served as an input for two methods, Greedy-Clustering, and Density-Clustering. Both were compared against a baseline segmentation approach using Scene Transition Graphs, which they exceeded in terms of F-measure and the coverage metric.

While both our methods have comparable F-measures, Density-Clustering is the better option for determining the semantic structure in the video stream, both in terms of identifying boundaries and in the overall number of unique scenes found.

Inclusion of an audio feature, in the form of a silence/music/speech detector, further improved recall and scene coverage, but at the expense of some over-segmentation. We suspect that our system will do well on broadcasts of an expository nature, where there are clear semantic boundaries between each scene. In our current dataset, we illustrate this by distinguishing between events of different sports (i.e. cycling vs. gymnastics) and by distinguishing between events of the same sport but with different semantic content (i.e. in swimming, the Men's 100 m freestyle 2nd semi-final followed by the Women's 200 m freestyle final)

Further improvements in scene recognition might be realized by performing speaker diarization, the identification of every speaker, and this would fit well into the chain paradigm. Our commentator could be improved by using latent topic models such as Latent Dirichlet Allocation[5] to infer better word to sport associations.

Summarization and segmentation are two sides of the same coin. More refined segmentation of a source allows for better analysis and summarization later on. Our commentator is able to describe a semantic event that is occurring and its participants given a good initial segmentation. We emphasize that the core work of segmenting the video stream into semantic events is primarily unsupervised. The parameters for the similarity criterion for each multi-modal chain and the local minima for the Density-Clustering can readily be determined; only the audio component requires additional supervision. With minimal extra knowledge automatically learned from Wikipedia articles, a semantic meaning can be assigned to the resultant scene segments in the form of important keywords and any participating persons. The resultant labels provide a semantic connotation, and can be used for further indexing purposes or other cross-media fusion tasks. Once segmented, the resultant video scenes are open to further analysis. Possibilities include object identification, action recognition, and the alignment of names and faces.

Acknowledgments The work reported is supported by IWT-SBO project AMASS++ (Advanced Multimedia Alignment and Structured Summarization, IWT 060051) and TOSCA-MP (Task-oriented search and content annotation for media production, FP7-ICT 287532).

References

1. Amir A, Argillander J, Berg M, Chang S-F et al (2004) IBM Research TRECVID-2004 Video Retrieval System. In Proceedings of TRECVID
2. Babaguchi N, Nitta N (2003) Intermodal collaboration: a strategy for semantic content analysis for broadcasted sports video. In Proceeding of the International Conference on Video Processing
3. Benini S, Bianchetti A, Leonardi R, Migliorati P (2006) Extraction of significant video summaries by dendrogram analysis. In Proceedings of the International Conference on Image Processing
4. Bertini M, Del Bimbo A, Nunziati W (2005) Common visual cues for sports highlights modeling. *Multimed Tool Appl* 27:215–218
5. Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. *J Mach Learn Res* 3:993–1022
6. Finkel JR, Grenager T, Manning CD (2005) Incorporating non-local information into information extraction systems by Gibbs sampling. In Proceedings ACL
7. Hearst MA (1997) TextTiling: segmenting text into multi-paragraph subtopic passages. *Comput Ling* 23 (1):33–64
8. Lazebnik S, Schmid C, Ponce J (2006) Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In Proceedings CVPR
9. Li Y, Narayanan S, Kuo CCJ (2004) Content-based movie analysis and indexing based on audiovisual cues. *IEEE Trans Circ Syst Video Tech* 14(8):1073–1085
10. Li Y, Lee S-H, Yeh C-H, Kuo C-CJ (2006) Techniques for movie content analysis and skimming: tutorial and overview on video abstraction techniques. *IEEE Signal Process Mag* 23(2):79–89
11. Liu S, Xu M, Li H, Chia L-T, Rajan D (2006) Multimodal semantic analysis and annotation for basketball video. *EURASIP J Adv Sig Process* 1–13
12. Nastase V, Strube M (2008) Decoding wikipedia categories for knowledge acquisition. *Proceedings AAAI* 1219–1224
13. Ngo C-W, Ma Y-F, Zhang HJ (2005) Video summarization and scene detection by graph modeling. *IEEE Trans Circ Syst Video Tech* 15(2):296–305
14. Nitta N, Babaguchi N (2002) Automatic story segmentation of closed-caption text for semantic content analysis of broadcasted sports video. In Proceedings of International Workshop on MM Information Systems, 110–116
15. Patsis Y, Verhelst W (2008) A speech/music/silence/garbage/ classifier for searching and indexing broadcast news material. In Proceedings of Database and Expert Systems Applications, 585–589

16. Poulisse GJ, Moens M-F (2010) Unsupervised scene detection in olympic video using multi-modal chains. In *Proceedings of CBMI*, 103–108
17. Poulisse GJ, Moens M-F, Dekens T, Deschacht K (2010) News story segmentation in multiple modalities. *Multimed Tool Appl* 48:3–22
18. Quenot G, Moraru D, Ayache S, Charhad M, Guironnet M, Carminati L, Mulhem P, Gensel J, Pellerin D, Besacier L (2004) CLIPS-LIS-LSR-LABRI experiments at TRECVID 2004. In *Proceedings of TRECVID*
19. Sadler, DA, O'Connor N (2005) Event detection in field sports video using audio-visual features and a support vector machine. *IEEE Trans Circ Syst Video Tech* 1225–1233
20. Sidiropoulos P, Mezaris V, Kompatsiaris I, Meinedo H, Trancoso I (2009) Multi-modal scene segmentation using scene transition graphs. In *Proceedings of ACM Multimedia* 665–668
21. Skorochod'ko EF (1972) Adaptive method of automatic abstracting and indexing. *Inf Process* 71:1179–1182
22. Vasconcelos N, Lippman A (2000) Statistical models of video structure for content analysis and characterization. *IEEE Trans Image Process* 9(1):3–19
23. Wang Y, Liu Z, Huang J-C (2002) Multimedia content analysis-using both audio and visual clues. *IEEE Signal Process Mag* 17(6):12–36
24. Xu C, Wang J, Wan K, Li Y, Duan L (2006) Live sports event detection based on broadcast video and web-casting text. In *Proceedings ACM Multimedia*
25. Xu C, Zhang Y-F, Zhu G, Rui Y, Lu H, Huang Q (2008) Using webcast text for semantic event detection in broadcast sports video. *IEEE Trans Multimed* 10(7):1342–1355
26. Xu M, Xu C, Duan L (2008) Audio keywords generation for sports video analysis. *ACM Trans Multimed Comput Comm Appl* 4(2):article 11
27. Yeung M, Yeo B-L, Liu B (1998) Segmentation of video by clustering and graph analysis. *J Comput Vis Image Understand* 7(1):94–109

Gert-Jan Poulisse has a B.Sc. from Stanford University and a M.Sc. from the University of Twente. He is currently pursuing his Ph.D at the Katholieke Universiteit Leuven. His research is on multimodal analysis and semantic content recognition in multimedia.

Yorgos Patsis was born in Creta (Greece) in 1972. After his graduation from the Department of Physics of Aristotle University of Thessaloniki he continued his studies in London (UK) obtaining an MSc in Digital Signal Processing with an emphasis in audio processing. He then worked as a design engineer in Soundcraft (UK) in the development of digital live consoles. In 2003 he continued in university research for speech and audio applications at AIIA lab of Aristotle University of Thessaloniki. At the same time he was working as an engineer in music for live and studio production. In 2006 he started working in speech and audio research for department ETRO of Vrije Universiteit Brussels in Belgium and since 2009 he is responsible for an audiovisual lab at ETRO used for research and industrial application.

Marie-Francine Moens is a research professor (BOF-ZAP) at the Katholieke Universiteit Leuven, Belgium. She received a Ph.D in Computer Science in 1999 from this university. She leads a research team specialized in information retrieval and text analysis. She is author of two monographs published by Springer and numerous articles in proceedings of international conferences and journals. She is involved in the organization or program committee of major conferences on computational linguistics and information retrieval (ACL, SIGIR, CIKM, EACL, ECIR). She is the current chair of the European Chapter of the Association for Computational Linguistics.