Journal of Computer Sciences and Applications, 2019, Vol. 7, No. 1, 1-9
Available online at http://pubs.sciepub.com/jcsa/7/1/1
Published by Science and Education Publishing
DOI:10.12691/jcsa-7-1-1

TV Stream Table of Content: A New Level in the Hierarchical Video Representation

Zein Al Abidin Ibrahim*

Lebanese University, Faculty of Sciences, Section I, Beirut, Lebanon
*Corresponding author: zein.ibrahim@ul.edu.lb

Received November 12, 2018; Revised December 13, 2018; Accepted December 28, 2018

Abstract With the rapid development of nowadays technologies, TV could keep its position as one of the most important entertainment and sometimes educative utilities in our daily life. However, keeping this position required a lot of major changes to take place in order for the TV to follow up with the digital revolution, such as, digital broadcasting, High Definition TV, TV on demand. TV-REPLAY, WebTV, etc. This evolution accompanied with many other factors such as the vast spread of communication means and the low prices of storing media have all resulted in many other indispensable technologies for video content storing, structuring, searching and retrieval. Video content can be of various types: a sequence of frames, a sequence of shots, a sequence of scenes, or a sequence of programs which is what the TV stream is usually composed of. Video content structuring would be of a great benefit to help indexing searching and retrieving information from the content efficiently. For example, structuring a soccer game into Play/Break phases facilitates later the detection of goals or summarizing the soccer video. Another example is to structure a news program into stories where each story is composed of an anchorperson segment followed by a report, which facilitates later the search of a specific story or an intelligent navigation inside the news program. However, all the existing analysis methods are dedicated for one type of video content. Such methods generate very poor results if it is applied on a TV stream that is composed of several video programs. So, it is important to detect a priori the boundaries of each program and then identify the type of each program in order to run the dedicated analysis method based on the type. For a TV viewer, a TV stream is a sequence of programs (P) and breaks (B). Programs may be separated by breaks and may include also breaks. For analysis purpose, the stream can be considered as a sequence of audio and video frames with no markers of the start and end points of the included programs or breaks. Most of TV channels that produce TV streams provide a program guide about the broadcasted programs. However, such guides usually lack precision, especially with the existence of live programs which makes the prediction of their start and end very hard. Moreover, program guides do not include any information about the breaks (i.e. commercials). Hence, one of the important steps to structure TV video content is to segment it into different programs and then choose the appropriate method to segment each program separately based on its type. The TV stream structuring consists in detecting the start and end of all the programs and breaks in the stream and later trying to annotate automatically each program by some metadata that summarizes its content or identifies its type. This step can be performed by analyzing the metadata provided with the stream (EPG or EIT), or analyzing the audio-visual stream itself. In this article, we define what we call TvToC (TV stream table of content) that adds a new level in the hierarchical video decomposition (traditional video ToC). Then, we provide a comparative study of all the methods and techniques in the domain of TV stream segmentation. Besides, a comparison of the different approaches is done to highlight the advantages and the weaknesses of each of them.

Keywords: *TV stream structuring, video structuring, near duplicate detection, classification*

Cite This Article: Zein Al Abidin Ibrahim, "TV Stream Table of Content: A New Level in the Hierarchical Video Representation." *Journal of Computer Sciences and Applications*, vol. 7, no. 1 (2019): 1-9. doi: 10.12691/jcsa-7-1-1.

1. Introduction

With the rapid development of digital capturing, storing and communication devices, the capturing, production and sharing of multimedia content has become very easy and very common. With a simple click on a mobile phone, or on a computer key board using a recording and video production software, you can produce, share or even broadcast TV easily. Moreover, the social media networks have facilitated more the spread of multimedia content, e.g. sharing a video with thousands of people or watching a TV stream on a computer or a smartphone. However, to get the most benefit from this huge number of stored video streams, they need to be easily accessed, retrieved and browsed which is still considered a problematic issue to be addressed.

The traditional provided way to access video content is to use the fast rewind and fast forward with different speeds in order to navigate to the part of the video that interests a user manually. Such navigation is usually considered inefficient since it is time-consuming especially when the video is long and it has no-relation with the video content. That is why, providing intelligent video content access methods is of big interest. For example, a story in a news program can be skipped with a simple click on a remote control if it does not interest the viewer. The structure of the video is the key of such intelligent access. A lot of exiting work in the literature has addressed the video content structuring.

An important method proposed to access video content is inspired from textual-book access methods [1]. In a book, the table of content (ToC) is one of the efficient mechanisms to access the content without reading the whole document. The ToC helps the reader to find the chapters or sections of interests and to navigate directly to the part of interest in the document. Moreover, a document contains index words that are considered as relevant keywords to the readers and their locations in the document. Such index can be used to reply the query of users. So, the ToC helps the readers to navigate within the document intelligently while the index helps them to retrieve information from the document. ToC helps to give a summary at the beginning of the document that helps to overview the entire content.

The video content structuring methods have mapped the idea of ToC to the video content. With the help of video ToC, we can browse and retrieve information much easier. However, to construct such ToC for video content, several challenges would be in question. Contrarily to a book, videos are not always of apparent and common structure. Some of them could be well structured such as news programs (an introduction, a presentation about a topic, a report, a presentation about the next topic, a report and so on) or tennis game (points, games, sets) while others are very difficult to be structured such as a soccer game for example (hardly structured in play/break phases). On the other hand, each type of video content will need its own method of structuring, e.g. a method that structures a news program cannot structure a movie program. Consequently, TV stream that normally contains more than one program (several video segments belonging to different programs) from different types and natures should be separated into programs, and then, each program type should be identified in order to run the relevant structuring method accordingly. As a result, additional information related to the boundaries of each program needs hence to be included in the ToC when it would be aligned with video content. The process of detecting the boundaries of programs in a TV stream in the objective of segmenting it into separated programs is nowadays called TV stream macro-segmentation. This name was given to differentiate the process of detecting boundaries from the usual segmentation that is done in a single program to segment it into many smaller parts (scenes, shots, etc.).

Before start presenting the TV stream structuring methods, we may ask ourselves an important question, why we need to structure TV streams if we know that TV channels produce the streams before broadcasting it and thus they should have precise metadata about the broadcasted streams (start, end and description of programs). In practice, broadcasted TV streams have no metadata except the electronic program guide (EPG) or the event information table (EIT) which lack precision especially if you have live programs that you cannot predict their start and/or end times a priori. Moreover, TV channels do not provide precise data about their content to prevent third-parties to archive and build novel TV services (TVoD, Catch-up Tv …) without returning back toward the channels. We should not also forget that the process of production of streams is very complex and many persons are involved in the process which makes the preparation of metadata not trivial task. Furthermore, delivering precise metadata to viewers would open them the possibility to skip commercials which are the first financial source of TV channels (in recorded streams or catch-up TV service) [2].

The aim of this article is to present an overview of the TV stream structuring methods in the literature and discussing the approaches and results obtained. The article is organized as follows: Section 2 defines what we call the TV stream table of content (TvToC). We present a state of the art of the existing method for TV stream segmentation in section 3. The dataset used, the evaluation measures calculated and the results obtained by each approach is provided in section 4 in addition to a discussion of the efficiency of each of them. We conclude the article in section 5.

2. TV Stream ToC

Before the manipulation of what we have called TV streams, video content segmentation or structuring has considered the video content of one type except for some of them that contains commercials. In [3], the video content structuring was defined as the task of decomposing the video into units and constructing the relationships between them. In text documents we find chapters, paragraphs, sentences and words. Similarly, in a video, we find the video itself, group of scenes or stories, scenes, shots, sub-shots, keyframes. Others consider the video content segmentation as a classification problem in which shots are clustered into groups in order to obtain video scenes which are clustered in order to obtain stories and so on.

The six-level video units are defined as follows:

1. **Video**: Flow of video and audio frames presented at a fixed rate.
2. **Story or Group of scenes**: Several scenes that capture continuous action or series of events. This element is relevant for some video genres such as news reports and movies.
3. **Scene**: A series of shots that is semantically related and temporally adjacent. It is usually composed of a series of shots recorded in the same location.
4. **Shot**: A sequence of frames that are recorded continuously with the same camera.
5. **Sub-shot or micro-shot**: A segment in a shot that corresponds to the same camera motion. Each shot may be composed of one or more consecutive sub-shots depending on the camera motion.
6. **Key-frame**: The frame that represents a shot or a sub-shot. Each shot and sub-shot may be represented by one or more key-frames.

In Figure 1, we present the six-level hierarchy. Each unit in a level can be produced by aggregating several units in the lower level (clustering-based techniques) or segmenting units in the upper level (segmentation-based techniques). For example, a scene can be identified by aggregating several shots or by segmenting a story. The literature is very rich in techniques that address one or several levels of this hierarchy (segmentation-based or classification-based approaches). You refer to [3-9] for more information about the segmentation-based techniques and to [10,11] for a review of the classification-based ones.

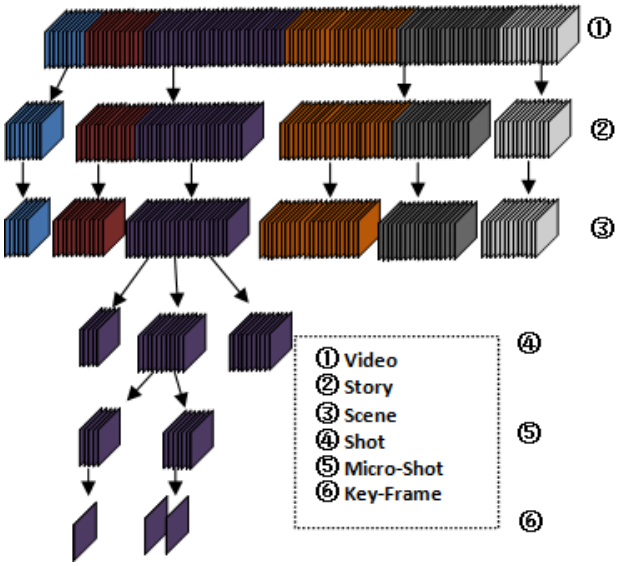

Figure 1. ToC: The six-level video content hierarchy

Unfortunately, the six-level hierarchy cannot be constructed for all types of videos. Some of them do not have a clear structure. In the literature, we can identify two main types of videos: Structured videos (News, Tennis game …) and unstructured or semi-structured videos (i.e. soccer game, video surveillance …). The structured video is the one that is produced according to a script or plan and can be edited later [7]. For unstructured and semi-structured content, instead of decomposing the video into the six-level hierarchy, it is decomposed into logical units. For example, we cannot decompose a soccer game into scenes and stories. However, most of the techniques in the literature decompose a soccer game into Play/Break sequences. The Play unit represents the sequence of shots in which the ball is inside the field and the game is going on while the Break unit represents the case when the ball is outside the field (Read [12,13,14] for more information). For a video surveillance, the units are not clear. Techniques of the literature consider the activity in the game as Play and the non-activity as Break (Read [15] for more information about video surveillance).

For a TV stream, the six-level hierarchy is not sufficient. For a user, he may be interested to browse a video by scenes or stories which are not the case for a TV stream viewer. A TV stream may be composed of a large number of scenes and stories [16]. It contains several heterogeneous programs which are usually separated and interrupted by breaks (commercials) and each has its ToC. For TV stream browsing and retrieval, it is more practical to append some levels to the hierarchy that facilitate the navigation by programs and then we have for each program its ToC that allows us to navigate deeply within it.

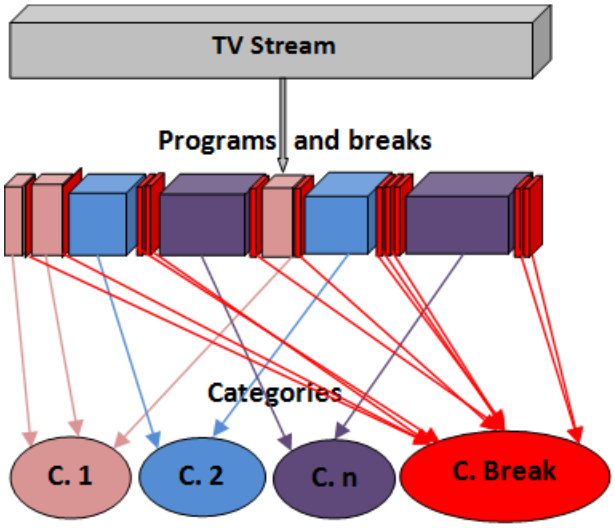

Figure 2. TvToC: The extension of traditional ToC

Figure 2 shows the levels that may be added to the hierarchy. The new ToC will be named TvToC. In such hierarchy, the user may skip programs that do not interest him and go deeply in others. A level that links programs of the same type is inserted. This level may be done by categorizing the programs of the TV streams (i.e. [10,20]).

The new units are defined as follows:

- **TV stream:** defined as contiguous sequence of video and audio frames produced by TV channels. It is composed of a series of heterogeneous programs (P) and breaks (B) without markers at the signal level of the boundaries of the programs and the breaks. Two consecutive programs are usually, but not always, separated by breaks. Each program may be also interrupted by breaks.
- ***Break (B)***: Every sequence with commercial aim such as commercials, interludes, trailers, jingles, bumpers and self-promotions. In some references [2,17], breaks are also called inter-programs or non-programs.
- ***Program (P)***: Every sequence that is not of break type (movies, TV games, weather forecasts, news…). Programs have culture, informative or entertainment aim. Sometimes, a program may be composed of several parts separated by break sequences.
- ***TV stream structuring:*** Known also as TV stream macro-segmentation is the process of precisely detecting the first and the last frames of all the programs and breaks of the stream and in annotating all these segments with some metadata. As a consequence, TV stream structuring allows user to recover the original programs that construct the continuous stream.

3. State of the Art

Most of the structuring methods proposed in the literature focused on structuring a single program and they

didn't handle streams containing several heterogenous programs. In our review, we have focused on two complementary tasks: The first task is how the stream is segmented into sequences of Program/Break while in the second, we present, if proposed, the method to label the segmented programs with some metadata and what is the source of these metadata.

In order to segment TV streams, several types of approaches were proposed in the literature:

1. The first type of approaches focuses on segmenting the stream into logical units and then classify each unit as being a part of a program or a part of break such as proposed by [2,18]. The logical units to be classified may be of different granularities (Key-frame, Shot, Scene, Stories …). After the classification step, consecutive units of the same type are merged together.
2. The second type of approaches focus on the detection of discontinuities in the homogeneity of some features [19], the modeling of the boundaries between program and breaks [21], or the detection of the repetition of opening and closing credits [16].
3. The third type of approaches is based on the fact that breaks have repeated behavior. Some of the techniques recognize breaks in a reference database [17] or by searching the repeated logical units [2,18,23]. Some program may have repeated parts such as the opening and closing credits of news programs, the latter should be followed by a classification step in order to separate repeated program segments from repeated break ones.

After the stream is segmented, the labeling of programs by metadata is done using:

1. The metadata provided by the TV channels such as the EPG or EIT (e. g. [2,17,18]).
2. The metadata extracted from the signal itself such as the speech transcripts (e. g. [24]), teletext or the recognition of opening and closing credits of some specified programs.

In this article, the techniques of the literature are categorized into two main categories:

1. The first category contains methods based only on the analysis of metadata available with the stream. They will be noted metadata-based. The only method found in the literature is the one proposed by Poli et al. in [25]. In this method, the audiovisual stream may be partially processed to enhance the prediction.
2. The second category represents methods based on the analysis of the audiovisual stream. They will be noted content-based and can be categorized into two sub-classes:
   - The class of methods that search the boundaries of the programs themselves noted as program-based methods [16,19,21,26,27].
   - The class of methods that detect breaks that may separate consecutive programs noted as break-based methods [2,18,28].

In the following section, we will present the different methods of the literature. Then, we will provide a summary of their advantages and disadvantages. Finally, we will conclude the section with the results obtained by each method and discuss its efficiency.

3.1. Metadata-based Methods

As we have stated, this category of methods uses only metadata to segment TV streams. It contains the method proposed by Poli et al. in [25]. The idea is to rely on the fact that TV channels tries as much as possible to respect some regularity in the program plan to preserve and increase their audience.

Poli et al. propose an extension of the traditional HMM named Contextual HMM (CHMM) and uses a regression tree to predicts the start time, the duration and genre of programs and breaks during a week. In the CHMM, each node represents the genre of the program and the transition models the transition from one program genre to another one. The genre of a program does not depend on the genre of the precedent one but on the time of the day and the day of the week of the broadcast which is called the context of the broadcasted program. That's why Poli et al. propose an extension of the HMM named CHMM. Based on the context, a regression tree is used to predict the minimum, the maximum and the average duration of the broadcast. They use a one year of corrected EPGs to train the model and one week to test the system.

The idea of the Poli's work comes from the fact that the stream structure of a day in a week is very similar to the stream structure of the same day in the previous week. In addition to that, some part of the day is very similar to the same part in the previous day. Moreover, the start time, the duration and genre of programs are almost similar. For example, a news program starts always at the same time, has almost same duration and cannot be replaced by another program (except in some situation). However, the proposed method has several drawbacks: (1) It requires a huge amount of ground truth dataset to train the model; (2) It relies on the fact that TV channels have stable stream structure which is not always the case; (3) The efficiency of the prediction is 95% using a model learned on a one-year stream which requires additional step to improve the efficiency.

Other type of methods was proposed in the literature for program personalization and recommendation purposes (not structuring purpose) [29,30,31,32], for summaries program stream creation [33], or TV program indexing [34].

3.2. Content-based Methods

In this category, we can highlight two type of methods: program-based methods that focus on the detection of program boundaries and break-based methods that detect break segments.

3.2.1. Program-based Methods

One of the assumptions that some of the techniques of the literature based on is the fact that some programs start and end each day at the same time with the same opening and closing credits. That is why, Liang et al. proposes in [16] a method to construct a boundary model to detect repeated shots in different days. The model is then used to segment the stream into programs. Liang et al. test the proposed method on a 10 non-continuous TV streams recorded from the CCTV-4 channel from 17h00 to 21h00. Among the 10 streams, 4 are used to train the boundary

model and 6 to test it. The results obtained in terms of precision and recall are approximately 100%. However, based on the following drawbacks, we think that this method is efficient if applied on a very special case of TV sub-streams but cannot be generalized on any TV stream. First of all, not all the programs have opening and closing credits. Secondly, authors have only considered the most structured parts of the day (from 17h00 till 21h00) while the other parts are less structured and probably contain programs without opening and closing credits. It would be interesting to consider the whole day instead of only this part. Thirdly, the model cannot detect commercials that may interrupt programs which makes it an incomplete macro-segmentation approach. Finally, the method does not propose any way to update the model in order to consider any possible change in the TV schedule.

Similarly, [21] based in his work on the same weak assumption considering that programs start and end with opening and closing credits. They consider also that such opening and closing credits and commercials contain frames with logos and with monochrome background and big text characters. They call these frames *Program Oriented Informative iMages* (POIM). The idea of this work is to detect these POIMs. In order to reduce false alarms, authors use auditory and textual information. An SVM classifier is used to find inter-program transitions and reject all other type of transitions such as commercials. The method is validated on the TRECVID 2005 corpus. Even though the method shows high efficiency, we should highlight the following: Inserting frames with logos and with monochrome background and big text is not a standard way to separate programs. In the absence of opening or closing credits, POIMs, or the miss-detection of POIMs, the consecutive programs will be combined. Secondly, if any POIMs are detected during a program, the program will be over-segmented. Finally, the approach is validated on TRECVID 2005 corpus which is not really TV streams. They are videos of same type with such specific assumptions.

In [16] and [21], the approaches proposed are supervised ones since the first create a supervised model to detect program boundaries and an SVM classifier in the second to retain inter-program boundaries. Since supervised models tend to lost precision with time and need updates and because such methods based on weak assumptions about TV production rules, El-Khoury et al. proposed an unsupervised method to detect boundaries between programs. They based on a stronger assumption which is a same program has homogeneous properties [19]. Their idea is trying to detect the discontinuities of some audiovisual features. During the same program, these features are homogenous and can be modeled by a gaussian law. In a next program, the gaussian law is different than the previous gaussian law of the previous program. In order to detect the changement from one gaussian law to another one, authors use a GLR-BIC (*Generalized Likelihood Ratio – Bayesian Information Criterion*) audio segmentation method that was designed for speaker diarization [22]. The method uses first visual features in order to detect possible transitions from one program to another one. Then, audio features are used and afterwards the two segmentations are merged together. However, the method shows that small segments such as break segments cannot be detected. Authors test their method on a real TV stream composed of 120 hours of French TV stream recorded continuously during 5 days. The results obtained are promising. The originality of the method is that it is unsupervised and can be used for several types of video analysis tasks such as speaker diarization, shot detection, program segmentation, etc. Moreover, the assumption used in the work is very strong. However, the method has two main drawbacks: The first is that short programs may not be detected and secondly that over-segmented programs are not later merged together.

The homogeneity property of features was also used by Haidar et al. in order to segment audiovisual documents using similarity matrices [26]. The idea of the work is to measure the similarity between documents based on some styles [35] and has not as main aim to segment TV streams. The similarity measure can be applied in order to detect near-duplicate videos, to measure how much two videos are similar or to detect similar segments between two videos. In their work, a similarity matrix is generated per feature used and then all the similarity matrices are merged. As an application, the authors compare a long day stream with itself (auto-similarity) in order to structure it. The similarity matrix shows clearly the structure of the stream. The method has some main advantages: (1) The method is independent from any video type, the used features or the duration of the video document; (2) It is generic since it can be applied for TV stream macro-segmentation, video copy detection, video segmentation and other applications; (3) It is unsupervised that method that does not need any training step and the assumption they base on is very strong and can hardly change. The main drawback of the approach is that the authors do not provide any method to extract the structure from the auto-similarity matrix which is not trivial.

Recently, deep learning techniques were used by Hmayda et al. [27] in order to identify tv programs based on features learned by the auto-encoder algorithm. The idea is to recognize TV programs by learning their jingles. The idea here is to construct a training database of visual jingles for several types of TV programs. Then, the features of the various jingles are learned using the stacked sparse auto-encoder network. A 1490-images of four TV program types (News, Meteo, Sport, Documentary) were used in the training phase. The approach is tested on a total of 376 images and the efficiency of program identification reached 95%. Even though authors do not address the problem of TV segmentation, but the approach can be used to classify video frames into program frame or break frame.

3.2.2. Break-based Methods

The techniques of the literature showed that detecting the boundaries of programs is a hard task. That is why other techniques focused on detecting the breaks that may separate programs instead. They have based on the fact that most of TV channels usually separate consecutive programs by breaks or special type of audiovisual frames. The problem is that lot of TV channels interrupt their programs also by breaks. In such case, break-based techniques will segment also the same program into several parts and a way to merge them should be proposed.

As stated before, breaks can be of several types: commercials, trailers, station identification, bumpers.

Most of the techniques proposed to detect breaks are extensions of commercial detection approaches. Lot of commercial detection approaches have been proposed in the literature but not for TV stream segmentation issues. We can categorize them into: (1) multimodal features-based [36-42], (2) recognition-based [40,43], and (3) repetition-based [44,45,46].

The last two categories have proven their efficiency to detect breaks. However, the step of break detection is not sufficient. It should be followed by the three following steps:

- Classification step: Its aim is to differentiate the program segments from break segments. This step is needed since the detected breaks may be parts of programs such as opening and closing credits (i.e. News opening and closing credits), programs broadcasted twice, etc.
- Merging step: Segment the stream into P/B sequences. Then, consecutive P sequences that belong to the same program should be merged.
- Labeling step: Label each program segment with metadata.

The first complete and real-time recognition-based approach in the literature is the one proposed in [17]. The idea is to use a reference video dataset (RVD) containing manually annotated shots which are a priori classified as P or B. To structure a stream, each shot of it is searched in the RVD in order to detect its repetitive behavior and know if it is a program-repeated shot or break-repeated one. This step covers the detection and classification steps mentioned above. In order to make a real-time recognition, authors propose to use a hash function built on a signature calculated from DCT coefficients [47].

Once the breaks are detected, all recognized break shots are considered as breaks that interrupt or separate programs and thus are retrieved from the stream. All remaining segments that have a duration more than one minute are considered as program segments. Authors consider that segments that are shorter than one minute are too short to be a program and they may be a break that is not in the RVD. In order to annotate programs, an alignment of the segmented stream with the electronic program guide (EPG) is performed using a dynamic time wrapping (DTW) algorithm.

The proposed method is tested using a twenty-two days long TV stream recorded from the France2 TV channel from 9/5/2005 to 30/5/2005. Authors use the first day as the RVD and the remaining days to test the system. The first day should be annotated manually which is a time-consuming step, the RVD should contain enough break segments to structure the stream precisely and this RVD will degrade over time.

Several repetition-based approaches have been proposed in the literature [2,18,23]. [18,23] use hashing tables with audio or video signatures while [2] bases on a clustering approach.

In [23], the structuring approach is performed in three steps. During the first step, repeated audio segments are detected based on silence detection and audio hashing. Step 2 is dedicated to the classification of repeated segments into advertisements if the length is between five seconds and two minutes while others are discarded. The remaining non-repeated segments of the stream are considered as program ones if they are longer than a fixed threshold. In order to merge over-segmented programs, authors consider that two consecutive program segments belong to the same program if they are of short duration (Step 3). To test the efficiency of the proposed method, it is applied on three TV streams (twelve, nine, and eleven hours). In this approach, four main drawbacks can be highlighted. The first is that authors consider the use of audio signature may be more efficient that visual one even though the detection of audio segment boundaries is a very hard task. Moreover, the thresholds used to classify repeated segments are not always true and cannot be applied on all type of channels. A third drawback is that the dataset used to test the efficiency is not very long and seems to contain the part of the days that is structured and is not very hard to structure it. Finally, no annotation of the segmented stream is performed.

One of the limitations in the approach proposed in [17] is the use of the RVD. This RVD may not contain all the repeated shots and may degrade over time even if the authors propose a way to update the it. The work proposed in [18] overcome this limitation. The idea is to use the same visual signature and hashing function in order to detect all the repeated shots in the processed stream and in real time. These repeated shots are then classified in order to separate the program repeated shots from break ones. Then consecutive shots of same type (P or B) are merged. The same steps proposed by [17] to segment the stream in P/B sequences and annotate them are applied. The results obtained by [18] are very close to the ones obtained by [17] and the method shows more stability over time. Even though, we can state two main drawbacks. The first is the need of some annotated segments to train the classifier and the second is no real-time TV segmentation method is proposed.

To detect repeated content, Manson et al. proposes in [2], a micro-clustering technique is proposed. The idea is to cluster key-frames in such way similar ones are put in the same clusters. To separate P segments from B ones, inductive logic programming (ILP) based on local, contextual and relational features is applied. Consecutive B segments are then merged and the remaining segments are aligned with the EPG using DTW algorithm in order to annotate them. Authors show the efficiency of the approach by applying it on two weeks of French TV stream. In this approach, authors use 7 days of manually annotated stream to train the ILP which is considered as time-consuming step. Moreover, the contextual features used for each segment to classify it will prevent the structuring to be in real-time but shifted by a time depending on the contextual window used.

More effective repetition detection approach was proposed by Yuan et al. in [48] that bridges the gap between clustering-based and hashing-based techniques. The idea is to detect repeated content with a little prior knowledge. The idea is to produce hash keys with the help of product quantization hashing in order to take advantage of the efficiency of hashing techniques and the power of clustering methods that fits well data distribution. In this work, video frames are firstly described by a 64-vector representation obtained by applying PCA on a 96-vector of GIST features. Then, product quantization step is applied in order to assign each vector to a unique cluster that will be used later to derive compact hash code for the vector. In such a case, similar frames will be assigned to the same cluster. A temporal consistency check is applied

to retain the meaningful repetitions in TV streams. The technique is applied on a 22-day TV stream. The first day is used to train the product quantization codes and the remaining to detect the repetitions. The method outperforms the traditional repetition detection methods but unfortunately it was not extended to segment the TV stream.

4. Comparison of Approaches: Results and Discussion

In order to be fair in the comparison of the proposed approaches, they should use the same dataset in their experimentations and provide the same evaluation measure. This was not the case except for the methods [18] and [17]. The datasets used in the literature are variable (i.e. one day in [16] to 22 days in [18]). The evaluation measures are almost different (i.e. Precision, recall, F-measure, etc.). Even though, there are some characteristics that help us to compare approaches. We can list the following:

- The type of the approach (metadata-based, program-based, or break-based).
- The size and continuity of the dataset used to train and test the approach. Logically, a several-days dataset is better than several-hours one. Moreover, the continuity of the stream composing the dataset is among the important features since, from our knowledge in the domain, some parts of the day are more structured than others. For example, the period [18h00-22h00] is more structured than other parts because of the large number of audiences following the TV in this period. Structuring a stream built as the concatenation of several well-structured chunks of days is easier than taking several continuous days as they are broadcasted by the TV channel. For example, a 24-hours dataset composed of the chunk from 00h00 till 24h00 is continuous while the one composed of the concatenation of the chunks [18h00-22h00] of 6 days is not continuous even if the days are consecutive since the chunks [22h00-18h00] of the 6 days are missing.
- The completeness of the approach. We mean by completeness that the approach handles all the steps of TV stream structuring or some of them. For example, several approaches in the literature do not annotate the segmented stream at all.
- Learning-based approach or no. A learning-based approach is the one that needs during its process to build a model in order to structure the stream. Moreover, there is an important question which is if the built model degrades over time and need to be updated or no.

4.1. Datasets, Evaluation Measures, and Results

In this section, we will provide the reader the different datasets used in the literature to evaluate the structuring approaches, the evaluation measures used and the results obtained.

In Table 1, we list for each approach, the size of the dataset used for training and testing.

Before comparing the obtained results, it is important to list the different evaluation measures that was adopted by these approaches and how they are calculated.

- **Precision, Recall, F-measure:** Three types of these measures are adopted. Some have calculated them at program level, some have used the frame-level, while others have focused on the detection of boundaries. At program level, the precision is equal to the number of programs correctly found in the stream over the total number of programs found. The recall is equal to the number of programs correctly found over the total number of programs that should be found. The F-measure is equal to: 2*Precision*Recall / (Precision + Recall). By the same way, such measures can be calculated for breaks. At frame level, the calculation is done by the same way except that we replace the program by the frame. For example, the precision is equal to the number of frames correctly classified as program frames over the total number of frames classified as program frames. Similarly to what was said for program-level measures, we can calculate the same measures for breaks. At boundary level, the measures focus on the number of boundaries detected. For example, the precision measures the ratio of correctly detected boundaries over the total number of detected boundaries.
- **ARGOS F-measure**: In the evaluation campaign of ARGOS project [49], another way to calculate the F-measure was defined. The measure is based on matching the segments of the ground truth with ones in the results. Contrarily to the frame-level F-measure defined above, each program in the ground truth is matched only once with the one of the results having the longest intersection with it. Using this assumption, F-measure is equal to: 2 * matched intersection / (Number of programs in ground truth + number of programs in the results).
- **Temporal accuracy (TA):** The TA measures the average temporal shifts between the found programs and the real broadcasted ones. Nearer the value is to zero, more accurate is the segmentation. Such measure has no sense for the breaks.

Table 1. Summary of the paper's experimentations: Corpus

References	dataset	
	Training	Testing
[25]	One year	One week
[16]	4 TV streams (17h00 to 21h00)	6 TV streams (17h00 to 21h00)
[21]	3000 POIM images	5 TV streams from TRECVID 2005 (15h each)
[19]	One hour to tune GLR-BIC parameters	5 days
[17]	One day as RVD	20 days
[2]	One week to train the ILP rules	7 days
[18]	About 30% of annotated repeated sequences in a three weeks stream to train the classifiers	21 days

Table 2. Summary of the papers experimentations: measures and results

Papers	Measures used	Results	
		P/B Segmentation	Labeling
[25]	Program-level precision & TA	P= 97%, TA=17 sec	P= 97%
[16]	Program-level precision and recall & TA	P=95.8%, R=100%, TA=28 sec	No labeling step is used
[21]	Boundary-level precision, recall and F-measure	P=88%, R=91.5, F=89.2%	No labeling step is used
[19]	ARGOS F-measure	F=90.5%	No labeling step is used
[17]	Frame-level precision of programs and breaks	$F_{program}$≈99%, F_{break}≈90%	$F_{program}$>88% and <96%
[2]	TA	No segmentation step is used	TA ≈ 3m35s
[18]	Frame-level F-measure of programs and breaks	$F_{program}$≈98%, F_{break}≈90%	$F_{program}$ >90% and <96%

Table 2 shows for each approach in the literature, which measure was used and the results obtained. For some approaches, the corpus was composed of several streams and some of them have calculated the measure for each of them. In such case, we have averaged the measures on the whole dataset.

4.2. Discussions

Since all the approaches do not use the same datasets and the same evaluation measures, the comparison task is not easy to do. However, we can highlight here some of the keys that will help the reader to make his own opinion about each of the proposed approached.

To do so, we have categorized the approaches of the literature into four categories:

- **Category 1:** Contains the approaches that have no TV stream segmentation aims or their assumptions are not evident or very specific for some TV channels [16,21,26,27,48].
- **Category 2:** Contains the approaches that base on an annotated video dataset [17] or needs a big annotated dataset to train the model [25].
- **Category 3:** Contains the unsupervised TV stream segmentation approaches [19,23].
- **Category 4:** Contains the approaches that substitute the pre-annotated video datasets with a stage that will learn the model from the raw data [2,18].

One of the major drawbacks of the approaches of the first category is that some of them has no TV stream structuring aims such as [26,27,48]. Moreover, the approaches [16,21] base on non-evident assumptions such as each program has an opening and closing credits that repeats from one day no another day which make them not applicable on any TV streams.

The approaches of the second category base on some big pre-annotated dataset. However, with time, the dataset becomes old and the accuracy of the system starts to decrease. Thus, the dataset should be always updated to allow the system maintaining its accuracy which is a tedious task.

In contrast to the above two categories, category 3 and category 4 gather, from our point of view, the efficient approaches. First of all, they are not very constrained with a priori information such as pre-annotated datasets or weak assumptions. Even though the approaches of the forth category learn some information from the raw data, this step is done once and its validity is much longer. The approaches (except [23]) are validated on a real continuous TV stream which make their results more realistic than the ones using several chunks of streams extracted from the most structured parts of the days.

5. Conclusion

In this article, we aim to introduce the reader the new level in the video content hierarchy which is the program level. Nowadays, the content of a video stream most probably does not belong to the same program. In order to apply any analysis step on streams, we should recover the structure of the stream into its composed programs. It is an obligatory step since most of the analysis tools available work on videos having the same content. In this article, we presented the reader a new level in the hierarchical video representation. This level is a result of a segmentation step aiming to recover the original structure of stream. We present here an up to date survey of the stream structuring approaches of the literature. For each approach, we have listed its advantages and its drawbacks. Then, we have presented the datasets used in these approaches, the evaluation measures and the results obtained. At the end, we opened a discussion about these approaches, highlighted some clues that may help the reader to conclude which are the most efficient ones.

References

[1] Rui, Y., Huang, T. S. and Mehrotra, S., "Constructing Table of Content for Videos," *Journal of Multimedia Systems*, 7(5), 359-368, September 1999.

[2] Manson, G. and Berrani, S. A., "Automatic TV Broadcast Structuring," *International journal of Digital Multimedia Broadcasting*, vol. 2010, January 2010.

[3] Wang, M. and Zhang, H., "Video Content Structuring," *Scholarpedia Journal*, 4(8): 9431, 2009.

[4] Hanjalic, A., "Shot-boundary detection: unraveled and resolved?," *IEEE Transactions on Circuits and Systems for Video Technology*, 12(2), 90-105, February 2002.

[5] Lienhart, R., "Reliable Transition Detection in Videos: a Survey and Practitioner's Guide," *International Journal on Image Graphics*, 1(3), 469-486, July 2001.

[6] Koprinska, I. and Carrato, S., "Temporal Video Segmentation: a Survey," *Signal processing: Image Communication Journal*, 16(5), 477-500, January 2001.

[7] Rui, Y., Xiong, Z., Radhakrishnan, R., Divakaran, A. and Huang, T. S., "A Unified Framework for Video Summarization, Browsing and Retrieval," *Technical report, MERL Technical Report*, September 2004.

[8] Snoek, C. G. M. and Worring, M., "Multimodal Video Indexing: A Review of the State-of-the-art," *Multimedia Tools and Applications*, 25(1), 5-35, January 2005.

[9] Wilson, K. W. and Divakaran, A., *Broadcast Video Content Segmentation by Supervised Learning*, Divakaran A. (eds) Multimedia Content Analysis. Signals and Communication Technology. Springer, Boston, USA 1-17.

[10] Brezeale, D. and Cook, D. G., "Automatic video classification: A survey of the literature," *IEEE Transactions on Systems, Man and Cybernetics*, 38(3), 416–430, May 2008.

[11] Roach, M., Mason, J., Xu, L. and Stentiford, F., "Recent Trends in Video Analysis: a Taxonomy of Video Classification Problems," in *Proceedings of the International Conference on Internet and Multimedia Systems and Applications, IASTED*, Hawaii, USA, August 2002.

[12] D'Orazio, T. and Leo, M., "A review of vision-based systems for soccer video analysis," *Pattern Recognition*, 48(8), 2911-2926, August 2010.

[13] Tjondronegoro, D., Chen, Y.-P. P. and Pham, B., "The Power of Play-Break for Automatic Detection and Browsing of Self-Consumable Sport Video Highlights," in *6th International ACM Multimedia Information Retrieval Workshop (MIR'04)*, New York, USA, October 2004.

[14] Xie, L., Xu, P., Chang, S.-F., Divakaran, A. and Sun, H., "Structure Analysis of Soccer Video with Domain Knowledge and Hidden Markov Models," *Pattern Recognition Letters*, 25(7), 767-775, May 2004.

[15] Ali, M. H., Shafie, A. A., Fadhlan, H. and Roslizar, M. A., "Advance Video Analysis System and its Applications," *European Journal of Scientific Research*, 41(1), 72-83, 2010.

[16] Liang, L., Lu, H., X. Xue, and Y. P. Tan. "Program Segmentation for TV Videos," in *Proceedings of the IEEE International Symposium on Circuits and Systems*, 1549-1552, Kobe, Japan, May 2005.

[17] Naturel, X., Gravier, G. and Gros, P., "Fast Structuring of Large Television Streams using Program Guides," in *Proceedings of the 4th International Workshop on Adaptive Multimedia Retrieval*, 223-232, Geneva, Switzerland, March 2006.

[18] Ibrahim, Z. A. A., Gros, P. and Campion, S., "AVSST: an Automatic Video Stream Structuring Tool," in *Third Networked and Electronic Media Summit*, Barcelona, Spain, October 2010.

[19] El-Khoury, E., Senac, C. and Joly, P., "Unsupervised Segmentation Methods of TV Contents," *International Journal of Digital Multimedia Broadcasting, Hindawi Publishing Corporation*, vol. 2010, March 2010.

[20] Ibrahim, Z. A. A., Ferrane, I. and Joly. P., "A Similarity-Based Approach for Audiovisual Document Classification Using Temporal Relation Analysis," *EURASIP Journal on Image and Video Processing*, vol. 2011, March 2011.

[21] Wang, J., Duan, L., Liu, Q., Lu, H. and Jin, J. S., "A Multimodal Scheme for program Segmentation and Representation in Broadcast Video Streams," *IEEE Transactions on Multimedia*, 10(3), 223-232, Geneva, Switzerland, March 2006.

[22] El-Khoury, E., Senac, C. and Joly, P., "Speaker Diarization: Towards a more Robust and Portable System," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 489-492, Hawaii, USA, June 2007.

[23] Zeng, Z., Zhang, S., Zheng, H. and Yang, W., "Program Segmentation in a Television Stream using Acoustic Cues," in *Proceedings of the International Conference on Audio, Language and Image Processing*, 748-752, Shangai, China, July 2008.

[24] Guinaudeau, C., Gravier, G. and Sébillot, P., "Improving ASR-based Topic Segmentation of TV Programs with Confidence Measures and Semantic Relations," in *11th Annual Conference of the International Speech Communication Association, Interspeech' 10*, 1365-1368, Makuhari, Japan, September 2010.

[25] Poli, J. P., "An Automatic Television Stream Structuring System for Television Archives Holders," *Journal of Multimedia Systems*, 14(5), 255-275, September 2008.

[26] Haidar, S., "Comparaison des Document Audiovisuels par Matrice de Similarité, " *PHD Thesis*, University of Toulouse 3 – Paul Sabatier, September 2005.

[27] Hmayda, M., Ejbali, R. and Zaied, M., "Program Classification in a Stream TV using Deep Learning," in *18th Conference on PDCAT*, 123-126, Taipei, Taiwan, December 2017.

[28] Nat06

[29] Ardissono, L., Gena, C. Torasso, P. Bellifemine, F. Chiarotto, A. Difino, A. and Negro, B., "Personalized Recommendation of TV Programs," in *8th Congress of the Italian Association for Artificial Intelligence*, Springer-Verlag Berlin Heidelberg, Pisa, Italy, 2003.

[30] Lee, H., Kim, J. G., Yang, S. J. and Hong, J., "Personalized TV Services based on TV-anytime for personal Digital Recorder," *IEEE Transactions on Consumer Electronics*, 51(3), 885-892, August 2005.

[31] Nickum, L. A., "Personal Preferred Viewing using Electronic Program Guide," *US Patent, Numb. 7617512, url: "http:www.freepatentonline.com/7617512.html"*, November 2009.

[32] Rovira, M., Gonzalez, J., Lopez, A., Mas, J., Puig, A., Fabregat, J. and Fernandez, G., "Indextv: a MPEG7 based Personilized Recommendation System for Digital TV," in *IEE International Conference on Multimedia and Expo*, 823-826, Taipei, Taiwan, June 2004.

[33] Kawai, Y., Sumiyoshi, H. and Yagi, N., "Automated Production of TV Program Trailer using Electronic Program Guide," in *Proceedings of the ACM International Conference on Image and Video Retrieval*, 49-56, Amsterdam, Netherland, July 2007.

[34] Liu, Z., Gibbon, D. C. and Shahraray, B., "Multimedia Content Acquisition and Processing in the Miracle System," in *IEEE Consumer Communications and Networking Conference*, 272-276, Las Vegas, USA, January 2006.

[35] Haidar, S., Joly, P. and Chebaro, B., "Mining for Video Production Invariants to Measure Style Similarity," *International Journal of Intelligent Systems, Wiley*, 21(7), 747-763, July 2006.

[36] Albiol, A., Fulla, M. J., Albiol, A. and Torres, L., "Detection of TV Commercials," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 541-544, Québec, Canada, May 2004.

[37] Dimitrova, N., Jeannin, S., Nesvadba, J., McGee, T., Agnihotri, L. and Mekenkamp, G., "Real Time Commercial Detection using MPEG Features," in *Proceedings of the International Conference on Information Processing and Management of Uncertainty in Knowledge-based Systems*, 1-6, Annecy, France, July 2002.

[38] Duan, L. Y., Wang, J., Zheng, Y. Jin, J. S., Lu, H. and Xu, C., "Segmentation, Categorization, and Identification of Commercial Clips from TV Streams using Multimodal Analysis," in *Proceedings of the 14th annual ACM international conference on Multimedia*, Santa Barbara, USA, October 2006.

[39] Hua, X. S., Lu, L. and Zhang, H. J., "Robust Learning-Based TV Commercial Detection," in *Proceedings of the IEEE International Conference on Multimedia and Expo*, 149-152, Amsterdam, Netherlands, July 2005.

[40] Lienhart, R., Kuhmunch, C. and Effelsberg, W., "On the Detection and Recognition of Television Commercials," in *Proceedings of the IEEE international Conference on Multimedia Computing and Systems*, 509-516, Ontario, Canada, June 1997.

[41] McGee, T. and Dimitrova, N., "Parsing TV Program Structures for Identification and Removal of Non-story Segments," in *Proceedings of the SPIE Conference on Storage and Retrieval for Image and Video Databases*, 243-251, California, USA, January 1999.

[42] Sadlier, S. A., Marlow, S., O'Connor, N. and Murphy, N., "Automatic TV Advertisement Detection from Mpeg Bitstream," *Journal of Pattern Recognition Society*, 35(12), 2719-2726, January 2002.

[43] Sanchez, J. M., Binefa, X. and Vitria, J., "Shot Partitioning based Recognition of TV Commercials," *Journal of Multimedia Tools and Applications*, 8(3), 233-247, December 2002.

[44] Duygulu, P., Chen, M. Y. and Hauptmann, A., "Comparison and Combination of two Novel Commercial Detection Methods," in *Proceedings of the IEEE International Conference on Multimedia and Expo*, 1267-1270, Taipei, Taiwan, June 2004.

[45] Gauch, G. M., and Shivadas, A., "Finding and Identifying Unknown Commercials using Repeated Video Sequence Detection," *Journal of Computer Vision and Image Understanding*, 103(1), 80-88, July 2006.

[46] Covell, M., Baluja, S. and Fink, M., "Advertisement Detection and Replacement using Acoustic and Visual Repetition," in *Proceedings of the IEEE International Workshop on Multimedia Signal Processing*, 461-466, Victoria, Canada, October 2006.

[47] Naturel, X. and Gros, P., "Detecting Repeats for Video Structuring," *Multimedia Tools and Application*, 38(2), 233-252, May 2008.

[48] Yuan, J., Gravier, G., Campion, S., Liu, X. and Jegou, H., "Efficient Mining of Repetitions in Large-Scale TV Streams with Product Quantization Hashing," in *Workshop on Web-Scale Vision and Social Media, in conjunction with ECCV*, 271-280, Firenze, Italy, October 2012.

[49] Joly, P., Benois-Pineau, J., Kijak, E. and Quenot, G., "The ARGOS Campaign: Evaluation of Video Analysis and Indexing Tools," *Signal Processing: Image Communication, Special Issue on Content-based Multimedia Indexing and Retrieval*, 22(7-8), 705-717, September 2007.

© The Author(s) 2019. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).