RECENT ADVANCES AND CHALLENGES IN TV STRUCTURING

Patrick GROS

INRIA
Campus universitaire de Beaulieu - 35042 Rennes - France
Patrick.Gros@inria.fr

ABSTRACT

TV represents a huge source of data. Even if a TV stream exhibits a strong structure to the viewer, in terms of programs and breaks, this structure is completely implicit in the stream, which is a simple sequence of images and audio frames. This paper presents recent works achieved to recover the structure of such a stream. 4 categories of works are presented, as well as their results and respective requirements in term of annotation. The paper ends by outlining the challenges to be solved in this largely opened field of research.

Index Terms— TV stream structuring, program detection, commercial detection

1. INTRODUCTION

TV contents processing represents a big and interesting challenge. From an applicative point of view, it is the source of many innovative services, e.g. catch-up TV, TV on demand, which rely on new ways to access, aggregate and display these TV contents. From a more fundamental point of view, TV data represent huge volumes of multimodal data that challenge most existing processing techniques, like classification or clustering algorithms. TV contents appear as a good playground to study and develop techniques suited to such large-scale multimodal data that can be found in other application domains like biomedical or meteorological data.

Using a portion of TV stream in new applications first requires recovering its structure in order to split it in coherent segments. Such an operation would be trivial is all the information used to create this stream were available. Unfortunately, such information are usually lost in the production process; when they exist they are unavailable since the TV channels are often reluctant to diffuse such information that could help other companies to build services on their contents or even to remove the advertisements from the stream; when available, they can be inadequate for some applications. As an example, TV regulation authorities cannot rely on the data provided by the channels to verify if these channels comply with local regulations.

The goal of the structuring process is to recover the stream structure in terms of programs and breaks and to identify the exact nature of these segments: advertisement, trailer, self-promotion, sponsoring, or news, fiction... The problem is thus twofold: segmenting the stream in coherent segments and classying these segments into various categories.

When shorter segments have been identified, they can be also structured. News reports can be segmented according to the various topics, many live programs like games have various sequences corresponding to different stages of the program. A lot of work has been achieved on this topic that we let out of the scope of the present paper.

The present paper proposes a survey of the main approaches in the field of TV stream structuring. Section 2 presents the main categories of methods while section 3 presents some results. The conclusion presents some challenges of the field for the coming years.

2. TV STREAM STRUCTURING

The TV stream structuring problem has not been extensively addressed in the literature. Most of previous works focus on structuring a single program or a collection of programs without dealing with streams containing several heterogeneous programs. However, the literature is rich with systems that are dedicated to detect commercials, which could be considered as the basis of any TV stream structuring system (i.e., [1, 2, 3, 4, 5, 6]). However, these techniques are not sufficient to structure the streams because commercials are not the only kind of breaks. TV stream structuring can be divided into two complementary tasks. (i) The first task consists in segmenting the stream in Program/Break sequences where the precise start and end of programs and breaks are provided. (ii) In the second task, each segmented program or break is labeled with metadata in order to identify it and to facilitate the retrieval of information from the stream.

The first task of the process can be based on different approaches. (i) Segmenting the stream into logical units and then classifying each segment as a program or a break segment [7]. These segments may be of different granularities

P. Gros acknowledges the support of OSEO - France for its support through the Quaero project. He thanks his students, X. Naturel, G. Manson, Z.A.A. Ibrahim who made the real work on this topic.

(Key-frame, Shot, Scene...). Then, consecutive segments of the same content (same commercial, e.g.,) are concatenated. (ii) Searching the start and the end of program segments based on the detection of discontinuities in the homogeneities of some features [8], modeling the boundary between programs and breaks [9], or detecting the repetition of opening and closing credits [10]. (iii) Searching the start and the end of break segments by recognizing them in a reference database [11] or based on their repeated behavior [7, 12]. The latter should be followed by a classification step in order to separate repeated program segments from break ones.

This segmentation may use two kinds of data. (1) Metadata: some methods almost exclusively use the metadata available with the stream in order to structure it [13]. (2) The stream contents: others use the audiovisual stream itself to structure TV streams. Furthermore, these methods can be classified into two subclasses. (a) Methods that search the boundaries of the programs themselves. This type of methods is noted as program-based methods [8, 9, 10, 14]. (b) Methods that search to detect breaks, which may separate consecutive programs. These methods are called break-based methods [7, 11, 12].

The methods of the literature can be classified in four categories based on the techniques they rely on.

Category 1. A prototype of the first category is the metadata-based method developed by Poli [13]. His main idea is to use a large set of already annotated data to learn a model of the program guide and thus of the stream structure. A hidden Markov model and a decision tree are used to learn this model that predicts the start time and the genre of all programs and breaks appearing during a week. This is the only method that is totally based on television schedules, and it requires a huge amount of annotated data for the learning stage (up to one year for each channel, up to four years used in the work). An additional step is required afterward to analyze the stream since the prediction is not perfect. But the analysis can be restricted around the moments where a beginning or end of a program is predicted to appear. This reduces the computation need by a huge factor. One of the main outcomes of this method is the experimental proof that, on the channels used, the stream structure is very stable over the years.

Category 2. The second category contains the program-based methods that recover the structure of the TV stream by detecting the programs boundaries [8, 9, 10, 14]. In [10], the authors start from the assumption that, when considering two consecutive days, a given program starts approximately at the same time with the same opening and closing credits. As a consequence, their method relies on the repetitive behavior of the open and closing credits of programs in order to detect their start and end time. The assumption used by the authors is not always true. Some programs do not have opening and closing credits. In addition, the TV channel broadcasts change completely in weekends. Likewise Liang et al. and Wang et al. propose in [9] a method based on the opening and closing credits of programs. The idea is to detect special images called Program-Oriented Informative iMages (POIMs). These POIMs are frames containing logos with monochrome backgrounds and big text characters. From the authors' point of view, these POIMs appear in opening and closing credits and at the end of commercial segments. Unfortunately, if opening and closing credits seem to be quite common on Chinese TV, they are not always present in other countries. Moreover, these POIM frames are not always present at the end of commercials and are variable from channel to channel. Contrarily to the methods proposed in [9, 10], El-Khoury et al. propose an original unsupervised method based on the fact that each program has homogeneous properties [8]. Consequently, the programs are extracted by detecting the discontinuities of some audiovisual features. The authors start from the idea that during a program, a selected set of features behaves in a homogeneous manner. In this method, short programs may not be detected and consecutive segments that belong to the same program are not merged. Moreover, detecting the boundaries of the breaks is easier and more precise than detecting the program ones.

Category 3. In the third category fall the recognition-based techniques that detect break segments. Naturel's work [11] is the only complete structuring solution that is based on a reference database containing manually annotated breaks. This database is used to detect the breaks of the database broadcasted again in the following part of the stream. The authors use hashing tables with video signatures in order to detect such repetitions, which are used later to get the stream structure. The manual annotation of the database is the main constraint and drawback of the method. As a matter of fact, the validity of the reference database is rather short since the commercials change very often. On the other hand, an automatic technique is proposed to update the database and thus to face the continuous change of the breaks. Unfortunately, the experimental data set used in this paper is not long enough to validate this updating approach.

Category 4. The techniques of the last category are the break-based methods that are based on the detection of repeated audio-visual sequences in the TV stream. The underlying idea is that breaks and especially commercials have a repetitive behavior. Several methods that use this principle have been proposed. For example, Zeng et al. [12] use hashing tables with audio signatures in order to detect such repetitions. On the other hand, Serrano and Manson in [7, 15] use a clustering-based approach, which groups similar key-frames and visual features and then use inductive logic programming (ILP) to classify them into program and break segments. This last method, based on a supervised symbolic machine learning technique (ILP), shows that it is possible to learn the structure of the stream from raw data, establishing a link between Naturel's and Poli's methods. The drawback of this method is that it needs 7 days of manually annotated data to train the

system. Moreover, ILP restricts the usable information to the local context of each segment. In addition to that, authors have chosen to classify each segment independently from its repetitions. From our point of view, most of the times, a segment and its repetitions are of the same type except in the case of trailers. The trailers segments can be filtered using predefined rules.

Ibrahim's method [16] is a variation of the previous one taking into account the contextual information of all occurrences in the stream of a given piece of contents. This last step adds a noticeable improvement to the results as shown in the experimentations. But the technique uses supervised classifiers (several of them are compared) and also requires at least one or two weeks of manually annotated data to train the classifiers.

Contrary to the other methods in this category, the method proposed in [12] relays on audio signatures. Its authors justify this choice by the fact that audio can overcome the limitation of the time-consuming video decoding. Using audio is a good idea, but it should be noticed that video decoding is not so time-consuming nowadays. Moreover, detecting audio segment boundaries is not so easy. As a consequence, video signatures are used to overcome the latter problem. In addition, video signatures may be more robust than audio ones since the audio signal is very sensitive to noise and this may affect the repetition detection. On the other hand, the audio stream used to evaluate the method is not long enough, and the number of repetitions it contains is not provided. The rules used for the segmentation are very simple and their effectiveness is not clearly evaluated, for example, by a comparison with the ground truth. Finally, the programs segmented by this method have to be annotated manually. The metadata provided with the stream (EPG), which is an interesting source of information, is not used.

As a conclusion of this state-of-the-art survey, it should be noticed that the techniques based on the repetition detection are the more suitable ones as the user searches to segment the stream into programs and breaks. Of course, the breaks are not the only repetitive segments. Some program segments can appear several times in the stream, like opening and closing credits, flashbacks, news reports, and even a whole program can be repeated. Thus, a classification step is required to differentiate repeated segments that are programs from those that correspond to breaks.

3. RESULTS

Comparing the various methods in a fair way is rather difficult. In the various papers, they are tested and evaluated using different data sets that can come from various countries where the regulations are different, making the stream more of less difficult to analyze, and they can contain continuous parts of streams of just parts taken at specific hours (14-24). Here also, it is clear that nightly hours are often more challenging due to the lack of breaks between some programs. Furthermore, the measures are not always exactly the same. Reimplementing everything could be the solution, but as the papers do not describe all the software blocks used in the analysis, the result could be unfair to some of the methods. As a conclusion, comparing methods is not a simple task at all.

In Poli's work [13], the author tries to determine the start and end time of segments corresponding to programs or to break sequences. The results obtained from the automatic analysis are compared to the one obtained manually by a society (Mediametrie). It should be noticed that the algorithms provide the limits up to one image, when the manual data have an accuracy of 1s only. The results are provided for seven days, in terms of number of programs or break sequences correctly retrieved and of accuracy of the temporal limits.

Day	Nb of Programs	% of found programs	Temporal precision
Saturday	70	97,1 % (68/70)	7 s
Sunday	67	94,0 % (63/67)	26 s
Monday	90	97,8 % (88/90)	17 s
Tuesday	81	98,8 % (80/81)	0 s
Wednesday	77	93,5 % (72/77)	46 s
Thursday	93	97,8 % (91/93)	17 s
Friday	95	96,8 % (92/95)	0 s

Table 1. Table coming from [13]. For 7 days, accuracy of the detection of program limits. The number between parentheses provides the number of programs / break sequences found.

The method provides very good results (see Tab. 1), even during nights. The results in terms of temporal precision are also very good, with a accumulated error of 0s for all the programs detected during a full day twice in the week. The price to pay is the amount of data used during the training stage. One year of manually annotated data was used to learn the models. This was possible because the work has been achieved on the French national TV archive were this manual annotation is done daily by many professional annotators. It should be noticed that Poli works by recovering a predicted program grid from the stream. So, only predicted segments are recovered: as these segments are already labeled according to their category, there is no classification problem.

In the other works that use less information, the problem is twofold. First detect segments, and then classify them, since the detection phase cannot provide a label at the same time. These two phases are usually evaluated independently. In Manson's work, the first goal is to detect all the repetition of breaks. These repetitions are then used to delimit the programs afterward. The segmentation is evaluated in term of F_{2s}, the F-measure of the detection with an accuracy of 2s and R_{2s} recall with a precision of 2s, i.e. a segment is correctly detected if its boundaries could be detected with an error smaller than 2s. The results are provided on tab 2.

	Channel 1	Channel 2
Nb of segments	5549	6630
$F_2 s$	71,52 %	77,99 %
R_{2s}	91,45 %	87,69 %

Table 2. Table from Manson's PhD. Segmentation results. The two channels are two French non-specialized channels.

These results are based on unsupervised techniques (clustering), avoiding a long manual annotation. 4 weeks of data from each channel are used for evaluation. For the classification phase, the best results are obtained by separating the long programs (LP) from the rest, than by categorizing the short segments in various categories. The results are shown on tab 3. In this second phase, one week of data is used to learn the classifiers, 4 to evaluate the method.

		Long prgms	Others
Channel 1	Long prgms	1479	152
	Others	241	6775
Channel 2	Long prgms	1070	385
	Others	121	11827

Table 3. Table from Manson's PhD. Confusion matrices for the first phase of classification.

The short segments are then classified in several categories: short programs (SP), commercials (C), trailers (T) and sponsoring messages (S), and other breaks (OB). The global confusion matrix is presented on Tab 4.

	C	T	S	OB	LP	SP
C	1744	4	5	19	38	2
T	131	194	5	17	32	23
S	36	6	0	0	39	10
OB	48	0	2	197	24	11
LP	34	4	7	10	983	5
SP	54	2	0	2	83	63

Table 4. Table from Manson's PhD. Confusion matrix after the classification phase.

Naturel's and Ibrahim's methods were evaluated using the same data set of 3 weeks of TV. The aim of the first stage of these methods is to separate the frames corresponding to programs from those corresponding to breaks. Thus they are evaluated in terms of binary classification, using the precision and recall measures and combining them in the F-measure. Since the relative importance of the two classes is very different, each of them is respectively taken as the reference class that is search for. Fig. 1 and 2 provide the results in terms of F-measure for the program class, and for the break class respectively. Each of these figures shows three curves: the one obtained by Naturel (blue), one obtained by Ibrahim when using only the local context to classify a segment (pink) (a tentative to mimic Manson's method without reimplementing it), and the last one (orange) obtained by Ibrahim when using all occurrences of a repeated piece of contents to classify it.

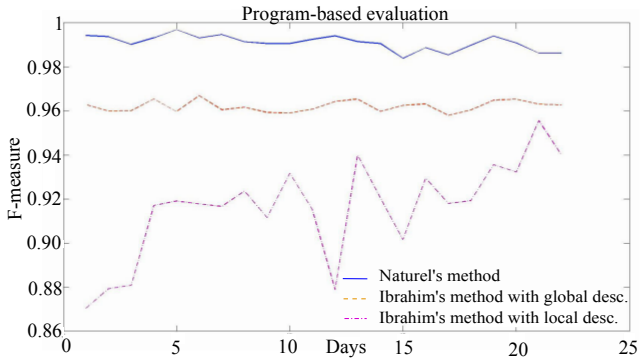

Fig. 1. Results in terms of program F-measure.

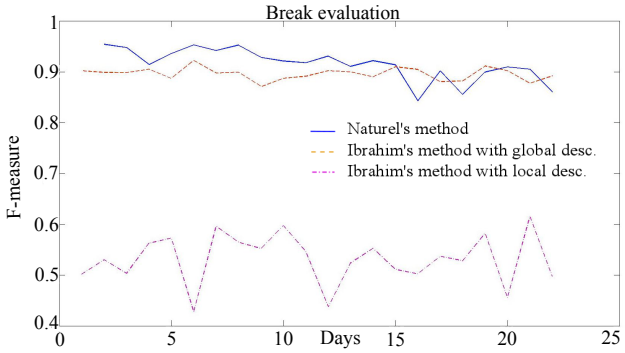

Fig. 2. Results in terms of break F-measure.

Several lessons can be learned from these results. First, when a piece of contents is repeated, classifying all the occurrences together is better. The alternative where each occurrence is classified independently with respect to its local context loses too much information. Second, it appears that learning contents rather than structural information is very powerful. It restricts the need of manual annotation to one day. Manson's or Ibrahim's require a lot more. On the other hand, Manson's system could work properly with some stream taken several months after the learning corpus, something clearly impossible with Naturel's method.

4. DISCUSSION AND CONCLUSION

The methods developed so far provide pretty good results (only very partial results could be provided here because of space limitations, but the reader is invited to read at the original papers). Both in terms of stream segmentation in coherent segments, and in terms of classification of these segments as programs or breaks of several kinds, the results are good.

Of course, there are still problems. During nights, many programs are not separated by breaks, some programs announced in the program guide are replaced without notice; Some days with unusual events (Olympics) have a very disturbed schedule, soccer games that have no fixed duration bring a lot of trouble. Nevertheless, the methods provide some useful information that could already be used in professional systems.

The main limitation comes from the requirement of a huge manual annotation load to get the appropriate data to train the classifiers. The aim of further research should not be to withdraw any manual work, but to reduce it to a more reasonable amount. As a matter of fact, most approaches rely on a brut force approach where the user is supposed to provide data to train a classifier designed to recognize rather sophisticated and high level concepts without any help of the system itself. An alternative would be to start with an unsupervised method, and to ask the user to interpret what this method could extract. this would for example allow to find first many occurrences of similar contents, or to find contents with very similar contexts, and thus to limit the manual annotation efforts. Such a problem cold also benefit from iterative methods where everything has not to be provided at once, but where the system can be trained to recognize simple but frequent situations first, before spending some more time on general concepts that are not easy to distinguish for humans (where is the limit between a short program and a promotional film?)

Another limit is the need to test the methods on much longer streams. Two or three weeks of continuous stream is already a lot of data, but such systems will be used on months of videos. The main problem with such durations is that one has to annotate everything manually in order to provide a ground truth even if the method is completely automatic and requires minimal manual help. Some companies already do such a work, and using it could overcome the problem (at least partly since their segmentation is usually less precise that what we could need).

In conclusion, if first methods have been proposed that show that the problem can be solved, some work has still to be done to develop methods that are easy and not too expensive to use in industrial contexts.

5. REFERENCES

[1] P. Duygulu, M. Chen, and A. Hauptmann, "Comparison and combination of two novel commercial detection methods," in *IEEE ICME*, 2004, vol. 2, pp. 1267–1270.

[2] L.Y. Duan, J. Wang, Y. Zheng, J.S. Jin, H. Lu, and C. Xu, "Segmentation, categorization, and identification of commercial clips from TV streams using multimodal analysis," in *14th ACM Multimedia*, 2006, pp. 201–210.

[3] J.M. Gauch and A. Shivadas, "Finding and identifying unknown commercials using repeated video sequence detection," *CVIU*, 103(1), pp. 80 – 88, 2006.

[4] X.S. Hua, L. Lu, and H.J. Zhang, "Robust learning-based TV commercial detection," in *IEEE ICME*, 2005, pp. 149–152.

[5] R. Lienhart, C. Kuhmunch, and W. Effelsberg, "On the detection and recognition of television commercials," in *IEEE Int. Conf. on Multimedia Computing and Systems*, 1997, pp. 509–516.

[6] N. Dimitrova, S. Jeannin, J. Nesvadba, T. McGee, L. Agnihotri, and G. Mekenkamp, "Real time commercial detection using MPEG features," in *Int. Conf. on IPMU*, 2002, pp. 1–6.

[7] G. Manson and S.A. Berrani, "Automatic TV broadcast structuring," *Int. Journal of Digital Multimedia Broadcasting*, 2010.

[8] E. El-Khoury, C. Sénac, and P. Joly, "Unsupervised segmentation methods of TV contents," *Int. Journal of Digital Multimedia Broadcasting*, vol. 2010, 2010.

[9] J. Wang, L. Duan, Q. Liu, H. Lu, and J.S. Jin, "A multimodal scheme for program segmentation and representation in broadcast video streams," *IEEE Trans. on Multimedia*, 10(3), pp. 393–408, 2008.

[10] L. Liang, H. Lu, X. Xue, and Y.P. Tan, "Program segmentation for TV videos," in *IEEE Int. Symp. on Circuits and Systems*, 2005, vol. 2, pp. 1549–1552.

[11] X. Naturel, G. Gravier, and P. Gros, "Fast structuring of large television streams using program guides," in *4th Int. W. on AMR*, 2006, vol. 4398 of *LNCS*, pp. 223–232.

[12] Z. Zeng, S. Zhang, H. Zheng, and W. Yang, "Program segmentation in a television stream using acoustic cues," in *Int. Conf. on Audio, Language and Image Processing*, 2008, pp. 748 –752.

[13] J.P. Poli, "An automatic television stream structuring system for television archives holders," *Multimedia Systems*, 14(5), pp. 255–275, 2008.

[14] S. Haidar, *Comparaison des documents audiovisuels par matrice de similarité*, Ph.D. thesis, Paul Sabatier University of Toulouse 3, 2005.

[15] S.A. Berrani, G. Manson, and P. Lechat, "A non-supervised approach for repeated sequence detection in TV broadcast streams," *Signal Processing: Image Communication*, 23(7), pp. 525–537, 2008.

[16] Z.A.A. Ibrahim and P. Gros, "TV stream structuring," *ISRN Signal Processing*, 2011(0), 2011.