Using grammar induction to discover the structure of recurrent TV programs

Bingqing Qu, Félicien Vallet, Jean Carrive, Guillaume Gravier

Using Grammar Induction to Discover the Structure of Recurrent TV Programs

Bingqing Qu
University of Rennes 1
French National Audiovisual Institute
bqu@ina.fr

Félicien Vallet, Jean Carrive
French National Audiovisual Institute
fvallet@ina.fr, jcarrive@ina.fr

Guillaume Gravier
CNRS
IRISA & INRIA Rennes
guillaume.gravier@irisa.fr

***Abstract*—Video structuring, in particular applied to TV programs which have strong editing structures, mostly relies on supervised approaches either to retrieve a known structure for which a model has been obtained or to detect key elements from which a known structure is inferred. In this paper, we propose an unsupervised approach to recurrent TV program structuring, exploiting the repetitiveness of key structural elements across episodes of the same show. We cast the problem of structure discovery as a grammatical inference problem and show that a suited symbolic representation can be obtained by filtering generic events based on their reoccurring property. The method follows three steps: *i*) generic event detection, *ii*) selection of events relevant to the structure and *iii*) grammatical inference from a symbolic representation. Experimental evaluation is performed on three types of shows, viz., game shows, news and magazines, demonstrating that grammatical inference can be used to discover the structure of recurrent programs with very limited supervision.**

***Keywords*—TV program structuring, symbolic representation, structural grammar induction, unsupervised approach**

I. Introduction

Large scale audiovisual archives offer an extremely abundant digital TV program library for users and content providers. For instance, the French National Audiovisual Institute, a repository of French radio and television audiovisual archives, has more than five million hours of radio and television programs stored. However, in order to be useful for later usage such as Internet diffusion, browsing and sharing, such large-scale archives need to be appropriately indexed. In particular, structuring programs, i.e., obtaining a temporal segmentation of programs into their basic constituents, is a crucial step for high-quality indexing, enabling better description as well as direct access to meaningful excerpts.

TV program structuring consists in automatically recovering the structure of a program from the video material, where structure refers to the way in which a program is organized by editors. In the video, the underlying program structure is often reflected in editing rules. Also, the structure of a program is consistent across the different episodes. For example, TV news programs usually start with a brief outline of the reports announced by the anchorperson. Headlines are followed by an alternation of anchorperson's announcement of the upcoming topic and news reports. Most news programs end with interview segments, weather forecasts or program trailers. Program structuring aims at detecting the existence and the temporal boundaries, i.e., the start and end frames, of such constituting elements designated as the structural elements, or events, of the program. In the framework of recurrent programs, i.e., of programs for which several episodes are available, a structural element refers to a video segment with a particular syntactic meaning and which can be regularly found in different episodes.

In this paper, we report ongoing work investigating grammatical inference to discover the basic structural elements as well as their temporal ordering, i.e., the temporal structure, by analyzing a collection of episodes from the program with minimal prior knowledge about the program genre and about the type of structural elements which might be present. In particular, we make very limited assumptions on what structural elements should be, as opposed to supervised approaches which seek to retrieve structural elements previously deemed as relevant for a type of program. To skirt the issue of not knowing which structural elements to look for, we exploit the repetitive nature of recurrent TV programs. A recurrent TV program is a program with multiple episodes which are periodically broadcasted, e.g., daily, weekly. News, entertainments, games and magazines are typical recurrent programs. Most episodes, if not all, follow the same editorial structure: structural elements appear in almost the same order with very similar durations, separated by sequences which repeat across episodes at more or less the same time instants. This last property is successfully used in [1] to detect separators. We adopt a similar idea, further exploiting separators to yield a symbolic representation of episodes suited for grammatical inference. By searching for recurrent elements throughout the episodes and selecting the ones which are relevant to the structure, one can infer the grammar of the show, i.e., the time ordered sequence of structural elements that each episode follows. Assuming such a grammar can be found, a model of the structure of the show can then be established to process additional episodes.

As a proof of concept, we focus here on grammar inference, implementing a three steps approach based on the sole visual modality. Firstly, a batch of broad scope event detection tools are used to find various types of events in all episodes. Secondly, events detected are analyzed across episodes to select the ones relevant to the structure of the program. Finally, a symbolic representation is derived from the segmentation given by structural element s and grammatical inference is applied to yield a grammar of the program by analyzing the symbolic time-ordered representation of each episode.

The rest of the paper is organized as follows. Section II

reviews the existing techniques for TV program structuring. Section III explains the overall method and details each step towards grammar induction for unsupervised TV program structuring. Experimental evaluations are reported in Section IV, followed by conclusions and perspectives in Section V.

II. Related work

TV program structuring has been extensively studied, almost exclusively relying on supervised approaches, which can be classified in two categories, depending on whether a particular program is targeted or not.

Previous work on TV program structuring mainly focused on the case where information on the structure is available as prior knowledge, thus enabling the use of supervised classification techniques. This is typically the case when targeting news or sports, two domains which have received tremendous attention (see, e.g., [2], [3], [4], [5]). Assuming the entire structure of the program is relevant, models of the structure can be learned from annotated data. Hidden Markov models, in multiple variants, have been widely used to this end [6], [7]. Event detection has also been used as an alternative to structure modeling. In this case, models are designed for the events of interest, e.g., goal or penalty in soccer, violent scenes in movies, and parameters are estimated on annotated training data.

Resorting to prior knowledge on the structure offers relatively accurate structuring algorithms but is limited by nature to specific types of programs, requiring training data for each new type. As an alternative, research has targeted the detection of typical structural elements, i.e., of events related to the structure independently of the different program types. Such elements are very diverse, such as anchorperson [8], [9], typical scenes or separators [1], [10] as defined by their repetitiveness. Separators, which are short sequences separating the different parts of a program, are of particular interest to unsupervised structuring. However, program-independent structural elements are insufficient to yield a complete program structure.

Work reported in this paper attempts to make the best of both worlds, i.e., being independent of the program type and obtaining a complete model of the structure.

III. Structure discovery with grammatical inference

The global objective we are targeting consists in inducing a structural grammar for a recurrent TV program, taking advantage of the existence of a collection of episodes of the same program. To avoid any confusion, the term *program* refers to the name of the show, thus being disconnected from video material, while *episode* refers to an exemplar of the program. A structural grammar is comprised of the set of structural elements, representing the different elements which compose each episode, and their temporal order. For example, report and anchorperson's announcement are structural elements for a news program, on top of which a temporal model with occurrence probabilities can be built to form a structural grammar.

A natural idea to discover the temporal model is to use grammatical inference on a symbolic representation of the video material representing each episode. The main difficulty

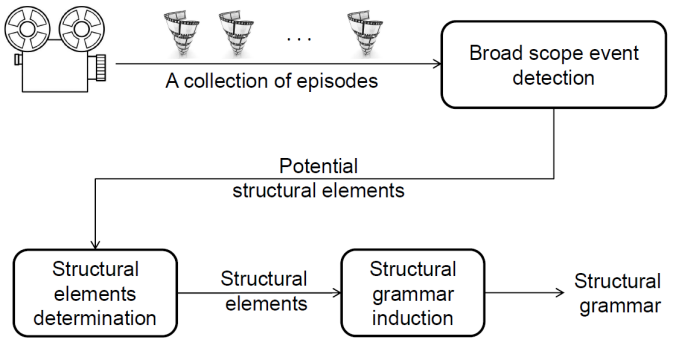


Fig. 1. General architecture of the three steps approach for the grammatical inference of a program structure.

therefore lies in obtaining a symbolic representation of the episodes in the absence of prior knowledge on what structural elements are to be considered. To solve this problem, we take advantage of the redundancy of information across all episodes to obtain symbols in an almost unsupervised manner. Redundancy appears in the structural elements, as well as in the so-called separators which recurrently occur between structural elements. The key idea of our method is therefore to analyze recurrent separators to in turn find recurrent elements which are deemed structural elements and which can be used for grammatical inference.

Targeting this general idea, we propose a three steps approach as illustrated in Figure 1. In the first step, a set of broad scope event detectors are used to find events within each episode, which might be of interest as a potential structural element or as a potential separator. In the second step, we assess the set of events detected along two lines. Density estimation is used to assess repeatability, i.e., to find events which recurrently occur at about the same instant in each episode. Role recognition is further used to assign properties to structural elements so as to help in deriving a symbolic representation. Finally, we induce the structural grammar of the program by leveraging multiple sequence alignment techniques.

One of the benefits of this approach is the very limited supervision that is required, thus making it possible to virtually apply the method on any collection quite straightforwardly. In particular, no data annotation is required at any step of the process. Apart from the selection of episodes, minimal prior knowledge on the program genre is required to select relevant event detectors in the first step and, in the second step, to deduct a meaningful symbolic representation from the set of structural elements found.

We detail in turn each of the three steps.

A. Detection of broad scope events

Ideally, a large number of event detectors should be used, which are generic enough to apply to a large number of shows. This is however not very practical because of implementation and run time issues. A number of key event detectors must therefore be selected based on a trade-off between the type of programs to process, the complexity at run time and, to a lesser extent, the implementation complexity. In this work, five event detectors were used as described below. Note that only minimal knowledge on the type of programs to process was

considered, resulting in general purpose detectors, which were applied to all types of programs in evaluation.

Shot detector. Shots are basic units for program structure which are required for structuring purposes. Shot boundaries were detected using the implementation of J. Mathe et al. [11] which seeks for hard cuts.

Dissolve detector. Gradual transitions are also common in TV programs, which usually signal the start and end of a scene. We focused on dissolve transitions, the most common gradual transition. Dissolve were detected using an extension of [12] in which the dissolve feature description was improved using double chromatic difference.

Monochrome image detector. Monochrome images—mostly white or black—are usually added to the TV streams at edition time to separate the different parts of a program. Detecting monochrome images is therefore an obvious choice for structure discovery in TV programs, in particular due to the simplicity of implementation. In this work, monochrome images are detected by simply verifying the histogram variance of the images.

Person clustering. Persons are essential features for almost any type of TV program. Especially, in many programs, a few number of key persons appear and are strong structure cues, such as the anchorperson in news shows or the host in game shows. TV show conductors usually appear as the most dominant face in a program, i.e., the one which appears most. Dominant person is usually implemented using person clustering based on faces and clothing. Taking news as an example, the anchorpersons clothes are usually carefully chosen so as to be easily distinguishable from guests (and one from another in case of multiple anchors) and obviously do not change within an episode. Person clustering was implemented using Viola and Jones face detection [13] and dress bounding box determination [14]. A key frame of each shot with a face is firstly otained. For each key frame, face features [15] and clothing histograms are then used in a two-step clustering algorithm, shown in Figure 2, based on K-means and hierarchical clustering to obtain person clusters.

Shot reverse shot detector. Shot reverse shot is a classical video technique depicting shots alternating between two characters facing one another, usually engaged in some face to face interaction. In the case of TV programs, we assume that a segment of shot reverse shot represents a dialog and that such interaction between two characters are relevant to the structure. Based on the results of person clustering, we detect shot reverse shot segments by a straightforward analysis of the cluster interlacing.

B. Determination of structural elements

Events detected in the first step are obviously not all relevant to the structure of a program and cannot be used as is to obtain a symbolic representation of episodes. For instance, monochrome images appear between two parts of a game show as a separator, i.e., a structural element, but can also be found at other places such as in a night scene. The second step therefore consists in selecting valid structural elements from which the symbolic representation is obtained. Selection of structural events implements two complementary strategies. On the one hand, role recognition is used to further

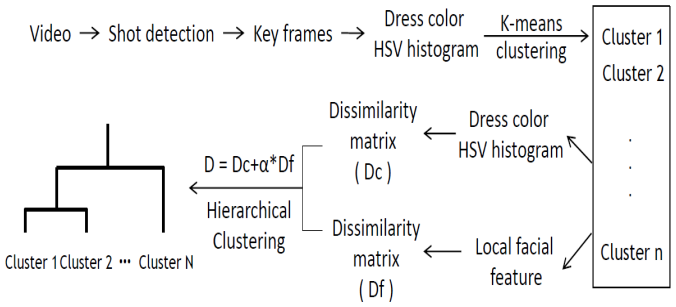


Fig. 2. Flowchart for person clustering

characterize the outcome of person clustering and identify important persons. On the other hand, the temporal distribution of the events across episodes is analyzed to find elements which repeat with relative temporal stability.

Role recognition. Role recognition enables determining the role of each person cluster resulting from face and clothing clustering. We mostly focus here on the conductor, or anchor, role which is clearly a strong cue with respect to structure. The first characteristic of the anchor is that he/she appears frequently, at more or less regular intervals and at key places, e.g., start and end of the each episode.

We used similar features as those defined in [16] to characterize a cluster, viz.: total duration of appearance (TFL); total number of distinct appearances, i.e., number of non consecutive clusters (TFT); duration of the longest segment in which the person appears (LFL); time range between the first and last occurrence (FR); duration in which the speaker is engaged in a dialog (DPT). Given such features, a dominant person is assumed to ideally have the following properties: he/she is the one that appears most (highest TFL); he/she is filmed the most frequently (highest TFT); he/she participates the most frequently in dialogs (highest DPT); his/her range of appearance (FR) and longest time of appearance (LFL) should not be the lowest. To account for varying episodes and programs lengths, all five features are scaled in $[0, 1]$. Decision on the dominant person is taken based on the sum of the five normalized features, the cluster for which the sum is maximal being identified as the dominant person.

Density filtering. In addition to role recognition, we exploit the property of repeatability across episodes to filter out events which occasionally appear in some episodes and which are therefore not relevant for the global structure of the program. Since different episodes of a recurrent program share a common temporal structure and have almost identical structural elements, the latter should appear in the vast majority of episodes and at about the same time.

Identifying elements which appear frequently at roughly the same time in all episodes is performed using temporal density analysis with a kernel function. For each type of event, we project onto the temporal axis the occurrences across all episodes of the event considered, counting the number of occurrences. To smooth out limited time variations, a kernel density estimation is used based on the following function

$$f(x;h) = \frac{1}{nh^d} \sum_{i=1}^{n} K(\frac{x - x_i}{h}) \ , \qquad (1)$$

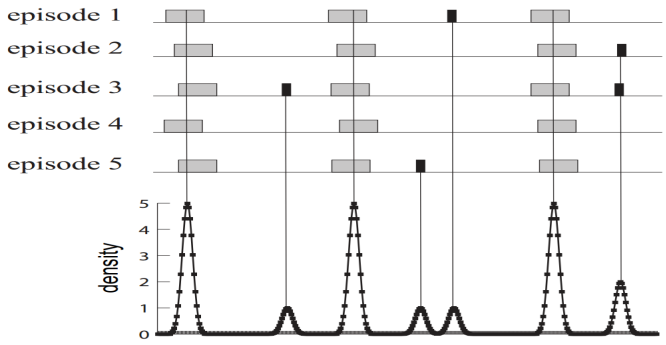


Fig. 3. Example of temporal density filtering to select structural events corresponding to structure separators (from [1])

where $f(x;h)$ is the estimated density function, h the bandwidth and $K()$ is the kernel. A Gaussian kernel was used with optimal bandwidth automatically chosen [17]. Peaks in the density estimation identify segments with high structural power due to their repetitiveness across episodes. Structural elements are selected based on thresholding of the density function to retain only relevant separators. This process is illustrated in Figure 3, where events in black are removed while the ones in gray are retained because of their high temporal concentration

C. Structural Grammar Induction

As a result of structural element determination, each episode can be represented as a time-ordered sequence of symbols with one symbol per structural element. Selecting and identifying valid structural elements for a program requires semantic interpretation of the structural elements detected via role recognition and temporal density filtering. This identification in turn requires minimal prior knowledge. For instance, a structural element corresponding to a sequence of white frames is a separator, while a long duration shot containing the dominant person at the beginning of the program is the conductor's opening.

Based on simple rules to identify valid structural elements, each episode is represented as a symbolic sequence depicting the succession of valid structural elements. The symbolic sequences corresponding to the different episodes of a program are usually similar due to the temporal stability of recurrent TV program structure. However, slight differences still exist between different episodes. Inferring a grammar from such sequences can be done by discovering the common patterns across symbolic sequences, a problem which is straightforwardly handled via grammatical inference. We adopt multiple sequence alignment techniques which can align the symbolic sequences in the way that alphabet symbols, i.e., valid structural elements, in a given position are homologous, superposable or play a common functional role. Many multiple sequence alignment tools have been developed, e.g., T-Coffee, MAFFT and ClustalW. We adopt in this work the ClustalW algorithm [18], which is one of the most widely used sequential tools for multiple sequence alignment due to its high accuracy, effectiveness and free availability [19]. While more complex grammatical inference techniques exist, based on regular expressions or context free grammars [20], we limited ourselves to multiple sequence alignment to study the meaningfulness of symbolic representations derived in an unsupervised manner.

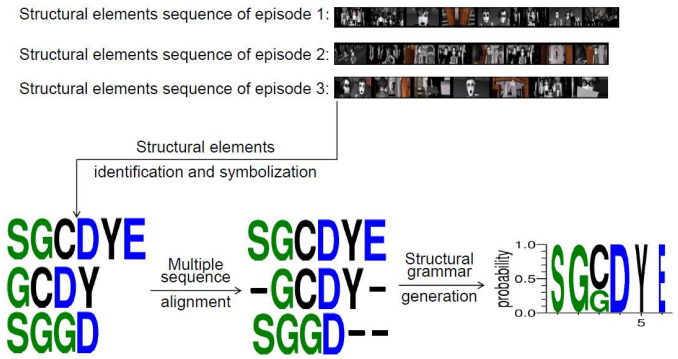


Fig. 4. Illustration of structural grammar induction from three episodes with symbols SGCDYE.

TABLE I. DESCRIPTION OF THE DATA USED FOR EVALUATION

Dataset	Date	Episodes	Type	Average duration
GAME	Sep. - Dec. 1991	12	Game	37.4 m
NEWS	Jan. - Dec. 2007	12	TV news	36.9 m
MAG	Jan. - Jul. 1997	12	Magazine	56.2 m

The process of grammar induction from a symbolic representation is illustrated in Figure 4 with three episodes. Symbols, i.e., SGCDYE, represent the valid structural elements that were identified. A graphical representation of the resulting grammar was obtained using WebLogo [21]. A stack of symbols is used to illustrate each position in the grammar: The height of objects within the stack indicates the relative frequency of each symbol while the stack width is proportional to the fraction of valid symbols in that position.

IV. EXPERIMENTAL RESULTS

Experiments are conducted on three recurrent programs from different types, viz., game, news and magazine. Evaluation considers in turn the three steps of our method for structural grammar inference. We first measure performance of the event detectors considered. Second, we evaluate structural element determination. Finally, we discuss the grammar inferred for each of the three programs to assess their relevance.

A. Data set description

Three different programs, with 12 episodes each, are used for inference and evaluation, as given in Table I. Two programs, *Que le meilleur gagne* (GAME) and *20h News* (NEWS), were taken with episodes selected over a large time period spanning 1991 to 2007. *Que le meilleur gagne* is a game show with four parts divided by separators. The program, leaded by a conductor, mainly contains interview scenes and questions/answers scenes with full text segments. The daily news show follows a standard pattern for such shows. The third program *Thalassa* (MAG) is a magazine about sea stories, where episodes were taken over a single year (1997). A conductor leads the show which is composed of reports and discussions. While the same conductor appears in all the episodes of GAME, two distinct conductors can be found both for NEWS and MAG.

B. Performance of broad scope detectors

Initial general purpose detectors are a key to subsequent steps and must therefore exhibit an acceptable level of accu-

TABLE II. PERFORMANCE OF DISSOLVE TRANSITION AND PERSON CLUSTERING

Dataset	Dissolve			Person
	Prec.	Rec.	F	Purity
GAME	0.62	0.79	0.70	72.7%
NEWS	0.64	0.81	0.71	49.5%
MAGS	0.52	0.95	0.67	69.6%

racy. We report here evaluation of dissolve transition detection and person clustering. Shot detection and monochrome image detection are based on standard techniques which yield very high accuracy. Performance for shot reverse shot detection is directly linked to performance of person clustering.

Results are presented in Table II. Dissolve detection is evaluated in terms of recall and precision. Detectors perform similarly for the three programs with correct recall and precision rates. Person clustering is evaluated by means of cluster purity. While correct purity values are obtained for GAME and MAGS, purity is rather low on NEWS. This is likely due to the fact that a fairly large number of persons appear in news programs and that scenes of reports in news are very variable with non discriminative clothing. Nevertheless, clustering results were found reliable enough for structure inference.

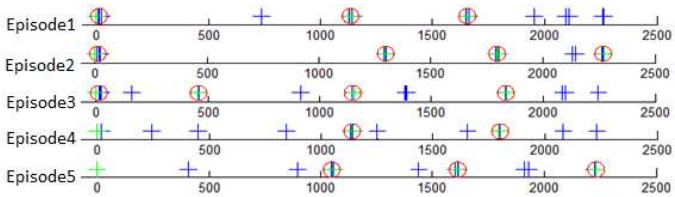
C. Structural elements detection

Selection of structural elements from the output of event detectors is evaluated both qualitatively and quantitatively.

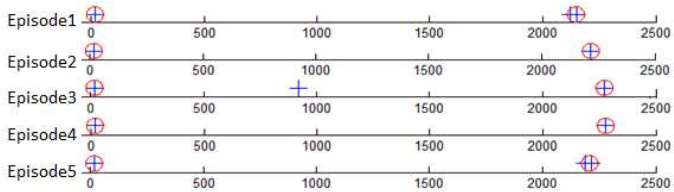
Regarding temporal density filtering to select recurrent events, we observed for GAME that peaks in the density coincide for monochrome frames and for dissolve transitions, with a relatively consistent number of such regions, i.e., 3 or 4 per episode. For NEWS, monochrome images were found to be often around the same temporal positions. Additionally, two short sequences of monochrome images are found in each episode, resp. at the beginning and end of the episode. Positions of events and separators are illustrated in Figure 5. Based on their characteristics, such elements are considered as separators on both programs. For evaluation purposes, a separator is considered as correctly detected when overlapping with a reference separator as annotated in the data. Precision and recall in GAME are resp. 94 % and 67 %. In NEWS, where we have a stronger structure, a recall of 96 % is achieved with similar precision as in GAME.

Dominant person detection succeeded in 83 % of the episodes for GAME, 92 % for NEWS and 75 % for MAG. NEWS are obviously easier with clear features identifying the anchorperson as the dominant speaker: fewer gestures, large area showing clothing, neutral and still facial expression, etc. In spite of relatively low cluster purity, dominant person detection is accurate. At the other end of the spectrum, MAG is a difficult content with long interviews for which occurrence time of the interviewee is almost as long as that of the anchor. Results in grammatical inference below indicate that limited accuracy in MAG did not hurt structure inference, even in spite of the limited number of episodes considered.

Moreover, identifying monologues of the dominant person turned out to be fairly easy based on the duration of the shots containing the dominant person. Dialog segments were also accurately determined combining person clustering and shot reverse shot detection, with performance directly proportional to the purity of the person clusters.



(a) GAME



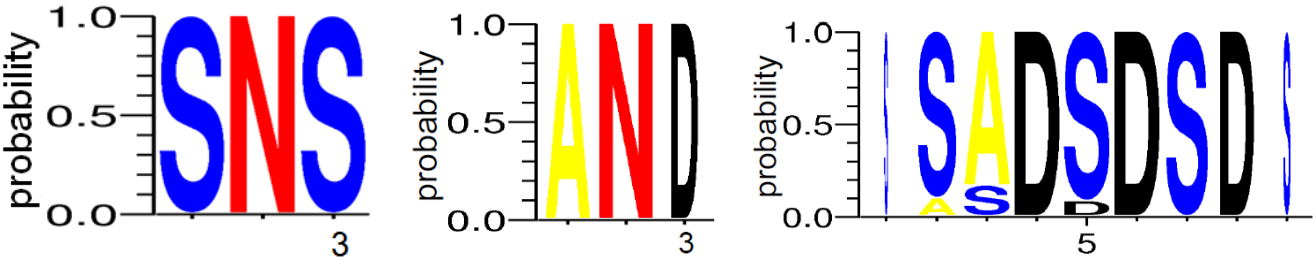
(b) NEWS

Fig. 5. Examples of separators for five different episodes of GAME and NEWS, where "+" in green (resp. blue) represents monochrome images (resp. dissolve) and "o" represents separators

D. Structural grammar inference

Results for grammar induction are illustrated in Figure 6 for the three programs. For NEWS, separators, denoted as S, are first determined. The segment between two separators, accounting for most time of the program, is considered as news content, denoted as N. The grammar of NEWS is therefore that of an introduction, followed by news reports, followed by a conclusion, a grammar shared by all shows. For MAG, selected symbols are based on dominant person's monologue and dialogs segments, yielding a simple deterministic grammar. A continuous segment with long duration, denoted N, is considered as a report while A and D represent anchorperson's monologue and dialogs respectively. For GAME, separators, dominant person (i.e., monologues of the conductor) and dialogs are the valid structural elements, resp. denoted S, D and A. The grammar inferred is more complex than for NEWS and MAG, reflecting the greater variability across shows. The main syntax is as follows: The game starts with an introduction (separator) followed by a dialog (between the anchor and the participants). We then have an alternating pattern of anchor (dominant person) and game phases (appearing as separator).

With very limited supervision, i.e., basic prior knowledge about TV program structure, possible structure is yielded for each program. All the identified structural elements in this ongoing work are the most common ones, so little bias caused by prior knowledge influences the final structural grammars. Obviously, all three grammars are concise because of the simple symbolic representation that was adopted. Yet, each one represents the structure of the corresponding program, thus demonstrating that grammar inference can efficiently handle structure inference in videos.Results on the game program interestingly prove that non deterministic video structures can be discovered. This last result opens the door to inferring structures at a finer grain. Considering symbolic description at a finer granularity will increase grammar diversity and complexity, which we believe can be handled via probabilistic grammar inference. Taking NEWS as an example, the news



(a) Grammar of NEWS (b) Grammar of MAG (c) Grammar of GAME

Fig. 6. Grammars induced for resp. NEWS, MAG and GAME. See text for details.

content, coarsely denoted N, can be divided into a repeating alternating an anchor's introduction and a report. This can be reflected in a grammar, either using multiple sequence alignment as considered here (assuming the number of reports is comparable across news episodes) or relying on more complex grammatical inference techniques with better factoring and generalization capabilities.

V. Conclusion

Preliminary work described in this paper shows how a symbolic representation suited for grammatical inference can be obtained from a collection of episodes of the same program, with almost no supervision and no specific training data. Starting from a set of general purpose event detectors, structural elements are derived with minimal prior knowledge by exploiting role recognition and recurrence across episodes. Grammatical inference finally brings a final layer of abstraction by evidentiating the overall structure of the program from the joint analysis of multiple episodes. Experimental evaluation on three types of programs shows that coarse yet relevant structures can be discovered from examples, even for non deterministic programs structures.

Results reported here mostly hint that unsupervised video structuring in recurrent collections using grammatical inference is viable and deserves further attention. The framework proposed in this paper as a proof of concept remains general and can be extended in a number of directions. Obviously, obtaining a symbolic description at a finer grain with limited supervision has yet to be achieved. Increasing the number of general purpose detectors and targeting multiple modalities seems like the most natural path to follow. But adding detectors will challenge the determination of structural elements and the grammar induction step, requiring more elaborate grammars to be considered.

Acknowledgment

The authors wish to acknowledge the help of François Coste and Vincent Claveau, IRISA, for their advise. They in particular have been very useful in guiding the choices made regarding grammar induction.

References

[1] A. E. Abduraman, S.-A. Berrani, and B. Merialdo, "An unsupervised approach for recurrent tv program structuring," in *of the European Interactive TV Conference*, 2011.

[2] H. Zhang, Y. Gong, S. W. Smoliar, and S. Y. Tan, "Automatic parsing of news video," in *Multimedia Computing and Systems, 1994., Proceedings of the International Conference on*. IEEE, 1994.

[3] M. Bertini, A. Del Bimbo, and P. Pala, "Content-based indexing and retrieval of tv news," *Pattern Recognition Letters*, 2001.

[4] H. Li, J. Tang, S. Wu, Y. Zhang, and S. Lin, "Automatic detection and analysis of player action in moving background sports video sequences," *Circuits and Systems for Video Technology, IEEE Transactions on*, 2010.

[5] L.-Y. Duan, M. Xu, Q. Tian, C.-S. Xu, and J. S. Jin, "A unified framework for semantic shot classification in sports video," *Multimedia, IEEE Transactions on*, 2005.

[6] L. Xie, P. Xu, S.-F. Chang, A. Divakaran, and H. Sun, "Structure analysis of soccer video with domain knowledge and hidden markov models," *Pattern Recognition Letters*, 2004.

[7] E. Kijak, G. Gravier, L. Oisel, and P. Gros, "Audiovisual integration for tennis broadcast structuring," *Multimedia Tools and Applications*, 2006.

[8] A. Hanjalic, R. Lagensijk, and J. Biemond, "Template-based detection of anchorperson shots in news programs," in *Image Processing, 1998. ICIP 98. Proceedings. 1998 International Conference on*. IEEE, 1998.

[9] X. Gao and X. Tang, "Unsupervised video-shot segmentation and model-free anchorperson detection for news video story parsing," *Circuits and Systems for Video Technology, IEEE Transactions on*, 2002.

[10] A. E. Abduraman, S.-A. Berrani, and B. Merialdo, "Audio/visual recurrences and decision trees for unsupervised tv program structuring," in *Proceedings of the 8th International Conference on Computer Vision Theory and Applications (VISAPP)*, 2013.

[11] *http://johmathe.name/shotdetect.html*.

[12] A. Mittal, L.-F. Cheong, and L. T. Sing, "Robust identification of gradual shot-transition types," in *Image Processing. 2002. Proceedings. 2002 International Conference on*. IEEE, 2002.

[13] P. Viola and M. Jones, "Robust real-time object detection," *International Journal of Computer Vision*, vol. 4, 2001.

[14] G. Jaffré, P. Joly *et al.*, "Costume: A new feature for automatic video content indexing," in *Proceedings of RIAO*. Citeseer, 2004.

[15] J. Sivic, M. Everingham, and A. Zisserman, "who are you?-learning person specific classifiers from video," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009.

[16] D. B. Jayagopi, S. Ba, J.-M. Odobez, and D. Gatica-Perez, "Predicting two facets of social verticality in meetings from five-minute time slices and nonverbal cues," in *Proceedings of the 10th international conference on Multimodal interfaces*. ACM, 2008.

[17] Z. Botev, J. Grotowski, and D. Kroese, "Kernel density estimation via diffusion," *The Annals of Statistics*, 2010.

[18] J. D. Thompson, D. G. Higgins, and T. J. Gibson, "Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice," *Nucleic acids research*, 1994.

[19] M. Larkin, G. Blackshields, N. Brown, R. Chenna, P. A. McGettigan, H. McWilliam, F. Valentin, I. M. Wallace, A. Wilm, R. Lopez *et al.*, "Clustal w and clustal x version 2.0," *Bioinformatics*, 2007.

[20] Y. Sakakibara, "Efficient learning of context-free grammars from positive structural examples," *Information and Computation*, 1992.

[21] G. E. Crooks, G. Hon, J.-M. Chandonia, and S. E. Brenner, "Weblogo: a sequence logo generator," *Genome research*, 2004.