A Video Grammar-Based Approach for TV News Localization and Intra-structure Identification in TV Streams

Tarek Zlitni, Walid Mahdi, and Hanène Ben-Abdallah

MIRACL, Multimedia, Information Systems and Advanced Computing Laboratory
University of Sfax, Tunisia
{tarek.zlitni,walid.mahdi}@isimsf.rnu.tn,
hanene.benabdallah@fsegs.rnu.tn

Abstract. The growing number of TV channels led to an expansion of the mass of video documents produced and broadcast on TV channels according to precise rules (e.g. consideration of the graphic charter, recurring of studios…). Thus, the use of a priori knowledge deduced from these rules contributes to the amelioration of the quality of segmentation and indexing of video documents. However, the effectiveness of automatic video segmentation works depends on video type. So, for a better quality of the segmentation, it is necessary to consider a priori knowledge concerning video types. In this context, this paper suggests an approach based on video grammar to identify programs in TV streams and deduce their internal structure. This approach attempts to automatically extract a priori knowledge to conceive the grammar descriptors. The study case of TV news programs is selected to validate the adopted approach since it is one of the most important types of multimedia content.

Keywords: TV programs localization, TV news structuring, video indexing, video grammar, a priori Knowledge.

1 Introduction

Given the increasing number of TV channels, a smart access to their broadcast contents represents a real challenge. The diffusion generates opaque streams whose duration exceeds several hours. An efficient multimedia content structuring approach should first of all proceed by the identification of a program in a large stream, and then detect all various units making up this program. According to production rules, the location indices or separations between the internal entities are recurring (jingles, studio decors). This work aims to use the recurrence of these indices in a process of inter-segmentation and internal structuring of the audiovisual contents.

A priori knowledge is extracted and modeled as descriptors and stored for future uses. These descriptors are used to generate visual grammars which store this knowledge in a structured and relevant way. As a result, each TV channel has its own grammar based on the recurring and discriminating descriptors for the automatic content structuring broadcast. To highlight the concept on which the grammar generation is based, TV news programs are selected as a representative study case for this work.

G. Maino and G.L. Foresti (Eds.): ICIAP 2011, Part II, LNCS 6979, pp. 454–463, 2011.

The remainder of this paper is structured as follows. The second section shows the motivation of this work and explains its general concept. In the third section, the steps of extraction of a priori knowledge and the coding of the relevant descriptors are presented. The fourth section deals with descriptors extraction, and the fifth section explains the structuring and storing of the obtained descriptors. Finally, conclusion and some future research directions are given in section six.

2 Motivation and Grammar Concept

TV channels broadcast several hours of various programs in a continuous way. Although resulting streams are long and heterogeneous, the points of location between different programs and the units within each one are recurring for all program instances. For example, the start jingles are quasi-invariant for at least one year, and thus it would be unimportant to repeat the treatment of their detection to each program occurrence during this period. Based on these issues, this work intends to factorize the recurrence of the delimitation points and to index them in a structured way as a grammar. The extraction, modelling and storing of the descriptors of these points are used as a priori knowledge. They are subsequently exploited in the processes of segmentation as ways for detection and validation of the appearance of their descriptors.

This work provides a grammar which collects and structures a set of a priori knowledge useful for the identification and the structuring of TV programs (Fig. 1). The use of this knowledge has various advantages. On the one hand, it presents the

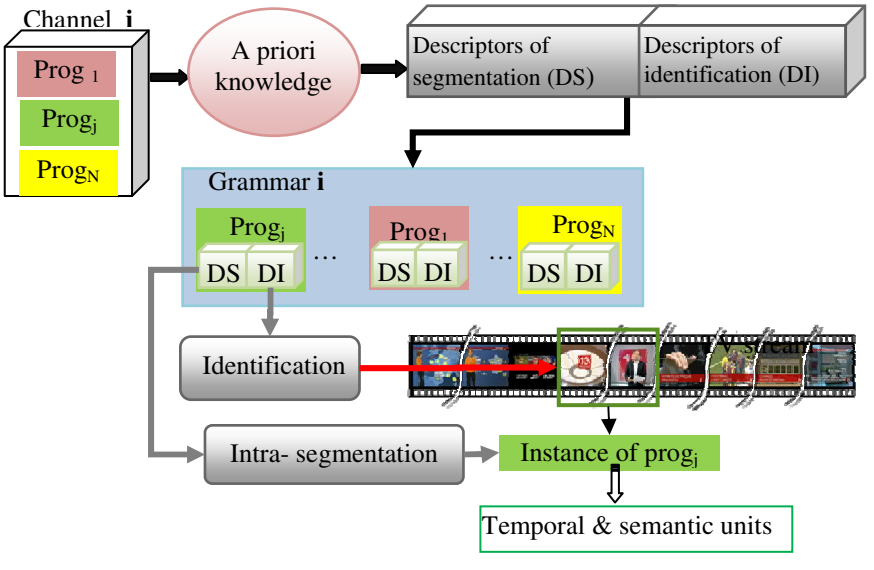

Fig. 1. Generation and exploitation of video grammar for the structuring of TV streams

principle of the descriptors factorization. In fact, video grammar is conceived mainly to model the recurring entities of TV channels. In other words, since the instance descriptors are always the same, they are extracted only once and reused with each process of segmentation (actually for a long duration). Such a step is undertaken in order to avoid tedious processes for the extraction of the primitives, thus the important profit in terms of execution (time calculation, memory capacity…). On the other hand, grammar also represents a complementary source to confirm the detected structural units. Consequently, a temporal or semantic unit is validated at the same time by the descriptors of signal level and the grammatical rules defined for the semantic concept of which this unit has been instanced. For example, a unit of the shot presenter type is detected by the appearance of a person shot and validated by the descriptors of this concept (presenter face and studio decor) defined by the grammar.

This way, each channel has its own grammar consisting of identification indicators of the programs in streams and the descriptors that play the role of indices for the deduction of the internal structure of the channel's programs.

3 A Priori Knowledge Extraction and Modeling

Although several previous works dealt with the segmentation of TV programs and TV news in particular [8], [10], few works start investigation from the inter-segmentation phase (identification) of a program in TV streams. This work highlights the identification phase and the location of the programs in streams since in the majority of the cases, the programs to be structured are incorporated in TV streams of long durations that can reach 24h and even more.

Once the programs are identified, the second part consists of enriching grammar by indices helping afterwards to identify the internal structure of these programs (1).

$$grammar(i) = \bigcup_{Prog_j \in Channel(i)} descriptors(Prog_j) \quad (1)$$

TV channel grammars consist of two types of descriptors (2) respectively for the two segmentation levels of a TV stream: inter-segmentation (identification) and intra-segmentation (internal segmentation) (Fig. 2) for each program $\boldsymbol{prog_j}$.

$$Descriptors(Prog_j) = DI(Prog_j) \cup DS(Prog_j) \quad (2)$$

- *DI* : identification descriptors.
- *DS:* internal segmentation descriptors.

The first type is made up of a priori knowledge used for the identification of the programs in video streams of a channel.

The second type consists of descriptors facilitating the internal structuring of the programs in a subsequent step. These descriptors are generally visual primitives which describe occurrences of the semantic points of anchoring (semantic entities) that delimit the internal units of a program.

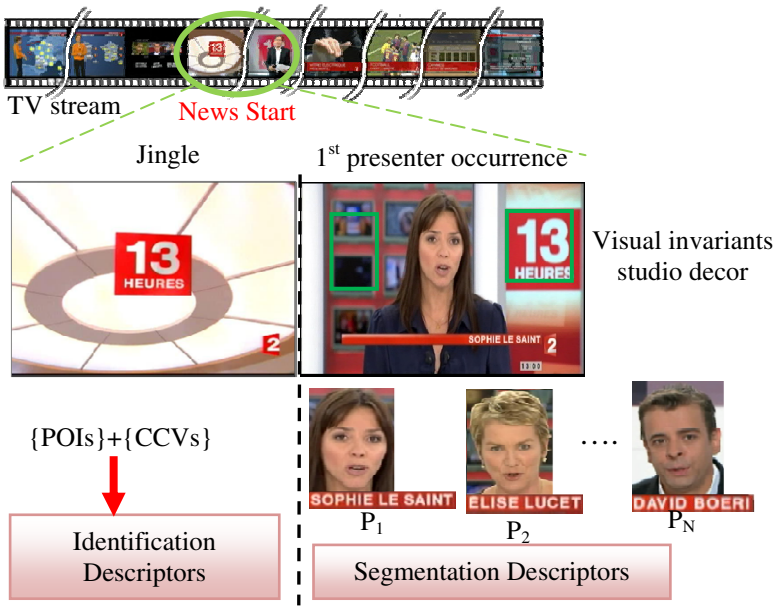

Fig. 2. Sample of the suitable descriptors for the identification and intra-segmentation of TV news

For TV news videos and according to literature [8], the apparition of the presenter(s) in the studio is considered as the anchoring point which delimits the temporal units of news programs, i.e. subjects which make up an instance of TV news program.

So, it would be essential to extract structure and store this knowledge in a relevant way in order to improve the semantic segmentation of the programs. Thus, for this programs type, in addition to the descriptors of trucking, grammar is supplied by visual studio decor descriptors and presenters' descriptors.

4 Descriptors Extraction

4.1 Identification Descriptors

The identification phase is based on Zlitni et al. [9] work's that dealt with the extraction of the adequate descriptors to distinguish the visual jingles from the various programs in a TV stream. Considering their specificity, two discriminative descriptors of video segments representing these jingles were chosen [4].

The first descriptor is a Point Of Interest descriptor (POI) chosen for the following reasons. POI is a point in an image which has particular properties; the peak is located where photometric information is most important within an image. These points are characterized by the robustness face to luminosity variations, blur effects and geometrical transformations.

The second descriptor is a colorimetric descriptor called Color Coherent Vector (*CCV*). Indeed, this descriptor represents each color by two values α and β which represent, respectively the number of coherent pixels and that of the incoherent ones. Moreover, the method used consists of classifying the pixels of the same color, in two categories: coherent and incoherent according to the size of the areas of the image to which they belong. Contrary to histograms, this descriptor presents the distribution of the colors considering their space dispersion.

4.2 Intra-segmentation Descriptors: Case of TV News

The intra-segmentation descriptors are indicators of unit change or the occurrence of an important event in a TV program (a goal in a sport program, new topic in TV news, etc). For the case of TV news, the presenter shot generally serves as an indicator of topic change. So, the descriptors of this entity are extracted and modeled (Fig. 3). Two characteristics specify this entity: presenter face and studio decor features.

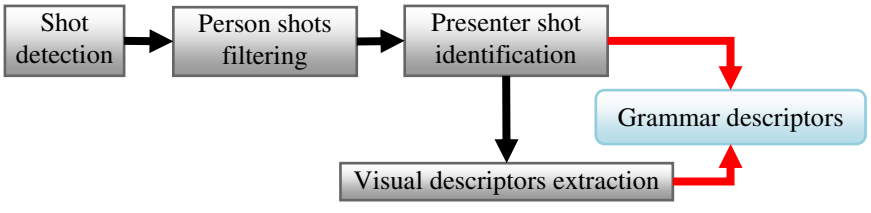

Fig. 3. Alimentation process of the grammar descriptors

4.2.1 Presenter Detection

Similarly to the graphic effects and studio decor of the programs which are invariant (at least for a long duration), the number of presenters per channel and type of program is quite limited. As for the news programs, the same persons are held in turn to present daily TV news programs. Based on this observation, it would be interesting to launch the process of detection and identification of some iteration and store their descriptors rather than call upon this process for hundreds of times during a few weeks.

Following the detection of start jingles the algorithm of identification of the presenter shot is launched. This algorithm is made up of two steps, (i) the detection and filtering of the person shots, (ii) the identification of the presenter's face.

a) Person shots detection and filtering

For the detection of the person shots, two phases are established: the shot detection [7] and the face detection.

To validate the person shots, the face detection technique presented by Viola and Jones [5] and developed by Lienhart and Maydt [6] is adopted, based on descriptors in cascades containing the wavelet of Haar. With this technique, the checking of the face presence in a zone of the image is based on the checking of the existence of a set of classifiers called characteristics of Haar. The application of

these classifiers is achieved in cascade where the order of the classifier depends on its weight.

For the filtering of the presenter shots, a preliminary filtering of the person shots was initially carried out. These are shots containing at least a person. Since a person is identified by his face, the detection of this type of shots is based on the detection of faces. A person shot is then the shot where face(s) appear.

A shot is considered a presenter shot only if it satisfies two essential conditions inspired from the rules of production of TV news defined by Zlitni et al. [11].

- The existence of only one person in the shot (maximum two): there are one or two presenters per news program, generally localized in the shot center (Fig. 5).
- Front view persons: since the presenter addresses the viewers.

b) Presenter identification

In order to recognize presenter face among all the faces detected in TV news, the Eigenfaces approach was adopted: a face recognition approach[2].Eigenfaces consists in measuring the similarity of a requested image with the basic images. Each face image is regarded as a vector in a space having as many dimensions as pixels in the image. The characteristics of the image are extracted by a mathematical method of dimensionality reduction based on the principal components analysis (PCA). An adaptation of this approach consists in computing all the similarities of the faces (V) with the remainder (Fig. 4). It is an operation of clustering of the similar faces. Each face will have a set of similarities (3).

$$Sim(v) = Card(x|\, sim(v,x) \leq TH_{SIM}, \forall x \in V) \tag{3}$$

Since the presenter is the person who appears more in TV news, the face having the maximum of similarity will be considered the presenter face (4).

$$v(P_i) = \max(\{\text{sim}(v)\}) \tag{4}$$

Fig. 4. Distances similarity of face (a) with a set of faces appearing on TV news

An experimental study on different TV news programs was established to evaluate the method suggested for the presenter identification. Very interesting rates were raised (recall ≈ 95) (precision ≈ 92). With the detection phase, the rate of recall is taken into account. Indeed, even if there are false detections (not faces) they have similarity scores equal to zero and are isolated afterwards in the phase of filtering.

Each presenter occurrence is indicated in the grammar by two properties (5): descriptors of his face and his name.

$$P_k = P(f_k, n_k) \tag{5}$$

4.2.2 Decor Descriptors

The second descriptor to characterize the presenter shot is the decor consisting in the decoration of the plateau and the possible graphic effects encrusted in this shot.To represent the graphic recurrence of the decors, a set of visual invariants are selected to graphically symbolize the decor.

As already mentioned, the presenter appears in a studio with a invariant decor. Thus, to represent this invariance, specific zones are selected to extract the descriptors automatically. These zones are in the areas surrounding the presenter. Indeed, following the localization of the presenter's face, we deduce the invariants of the decor close to the face. To reduce description, we were limit to the two areas on the left and on the right of the face (Fig. 5).

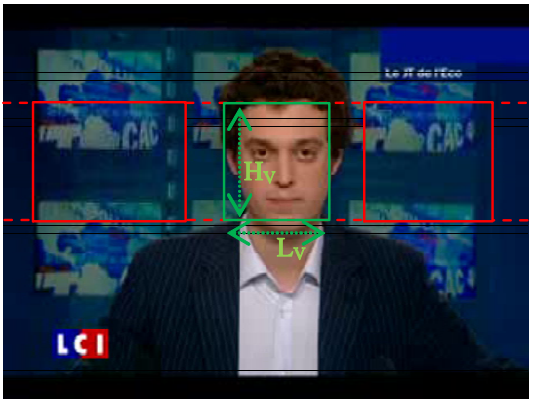

Fig. 5. Automatic localization and extraction of the invariant zones

The decor is usually characterized by particular textures visually easy to identify. The descriptors of texture have the property of identifying various areas in an image. They contain information about the space or statistical distribution of the colors. To compute these descriptors, the co-occurrence matrices are used [3]. These matrices measure the probability of appearance of the pixel pairs located at a distance δ in the image. They are based on the probability compute $M_{\delta,\theta}[i,j]$ which represents the number of times where a pixel of color level appears at a relative distance **δ** of a pixel of level of color and according to a given orientation θ . For reasons of simplification of the descriptors complexity, the direction parameter was

eliminated by fixing it at 0. Several second order metrics can be deduced from these matrices to characterize texture [1]. According to their discriminative effects for the various instances of the invariants, four metrics are retained: contrast (6), local homogeneity (7), homogeneity (8), and entropy (9).

$$C(k,n) = \sum_i \sum_j (i-j)^k \times M_\delta[i,j]^n \tag{6}$$

$$LHo = \sum_i \sum_j \frac{M_\delta[i,j]}{1+|i-j|} \tag{7}$$

$$Ho = \sum_i \sum_j M_\delta[i,j]^2 \tag{8}$$

$$E = \sum_i \sum_j M_\delta[i,j] \times ln(M_\delta[i,j]) \tag{9}$$

The similarity measured between areas which belong to the grammar and which are detected from the video is logically based on a function with a distance ***D*** calculated from a set of vectors f_i. Each vector includes the descriptors f_i corresponding to a specific area. Generally, this function is susceptible to noise and depends on the relevance of descriptors compared to a zone. To avoid this problem, the solution is to define a class for each zone covering a number of samples (i.e. several V_{fi} vectors) from the same spatial area but located on different images of the presenter shot as a first. Each class is considered as training data for a particular area to capture any possible changes that may occur within each zone in terms of noise, change of brightness or color, etc. Then the descriptors of each class are analyzed to associate to each region class (j) the weight of intra-class (w_{fj}) respective to each descriptor (f_i).

To determine the weight (w_{fj}), the standard deviation (σ_{fj}) of each descriptor (f_i) is calculated with the class of zone (j). (σ_{fj}) measures the variability of data in a descriptor region class. Thus, the larger the value of standard deviation (σ_{fj}); the more unpredictable the descriptor (f_i) is for the region class (j). Then, we calculate the weight (w_{fj}) of the descriptor (f) in the region class (j) according to equation (10).

$$w_{f_j} = \frac{(1 - \frac{\sigma_{f_j}}{\sum_{f=1}^{F} \sigma_{f_j}})}{\sum_{f=1}^{F} (1 - \frac{\sigma_{f_j}}{\sum_{f=1}^{F} \sigma_{f_j}})} \tag{10}$$

Obviously, the weights (w_{fj}) defined by this way for a class of zone (j) must satisfy the following constraint:

$$\sum_{f=1}^{F} w_{f_j} = 1$$

For some samples of TV news selected from different channels, the following results are obtained:

Table 1. Values of weight features of different TV news

	w1	w2	w3	w4
P1	0,331	0,333	0,091	0,244
P2	0,333	0,333	0,016	0,316
P3	0,332	0,332	0,225	0,110
P4	0,332	0,333	0,144	0,190
P5	0,333	0,333	0,008	0,325
P6	0,333	0,333	0,016	0,316
P7	0,333	0,333	0,004	0,333
P8	0,331	0,333	0,200	0,135
P9	0,332	0,333	0,047	0,287

The global zone descriptor for each visual invariant is defined by a weighed summation of these metrics form the texture descriptor (11).

$$\text{desc(Decor)} = w_1 E \cup w_2 LHo \cup w_3 C \cup w_4 Ho \tag{11}$$

5 Grammar of TV News Programs

After having detected and identified the presenter, his descriptors are coupled with the decor descriptors, a single description of a shot presenter is obtained (12).

$$PD_k = < P_k, \{desc(Decor)\} > \tag{12}$$

With each appearance of a new presenter, the process of identification is started in order to validate it. Each new presenter is added afterwards to the grammar's list of presenters already identified in the former iterations. So, there will be X iterations for identified X presenters that appear tens even hundreds of times ($N*X$), for only one year of streaming of a channel, resulting in an important reduction of structuring algorithm complexity.

Considering their redundancy, and in addition to the same reasons of complexity reduction, the decor descriptors are extracted just after identifying the first occurrence of the presenters.

Thus, each TV news program is described in the grammar by the descriptors of its jingle for the identification in TV streams and the descriptors of the studio decor in addition to the descriptors of all the presenters who can present this program to be able to deduce his internal structure.

A generalization of the step suggested for the majority of the programs of a channel leads to a grammar/channel composed by all the descriptors of location and the internal structuring of these programs.

The identification and segmentation of the programs of a TV channel are realized through the integration of research and comparing techniques of the descriptors of grammar with those of TV streams.

6 Conclusion

In this paper, the generation of video grammars is approached. The role of these grammars is to model and store a priori knowledge, useful for the structuring of

video streams, in the form of descriptors. Indeed, the recurring descriptors are classified into two levels: descriptors for the program identification in streams, and others for the internal structuring of the located programs. For the second type, TV news programs are selected as a case of study. In this type of media, we focus on the presence of the presenters' shots as the internal unit of structuring.

The perspectives are to extend the number of descriptors of intra-segmentation to have almost complete grammars for the various channels. We also think of suggesting and integrating extraction modules of the relevant and common descriptors for the majority of television broadcasts.

References

1. Haralick, R.M., Shanmugam, K., Dinstein, I.: Textural Features for Image Classification. IEEE Transactions on Systems, Man, and Cybernetics 3, 610–621 (1973)
2. Turk, M., Pentland, A.: Eigenfaces for Recognition. Journal of Cognitive Neuroscience 3, 71–86 (1991)
3. Arvis, V., Debain, C., Berducat, M., Benassi, A.: Generalization of the Co-occurrence Matrix for Color Images: Application to Color Texture Classification. Image Anal. Stereol., 63–72 (2004)
4. Zlitni, T., Mahdi, W.: A visual grammar approach for TV program identification. International Journal of Computer and Network Security 2(9), 97–104 (2010)
5. Viola, P., Jones, M.J.: Rapid Object Detection using a Boosted Cascade of Simple Features. In: Conference on Computer Vision and Pattern Recognition, USA, vol. 1, pp. I-511–I-518 (2001)
6. Lienhart, R., Maydt, J.: An Extended Set of Haar-Like Features for Rapid object Detection. In: IEEE International Conference on Image Processing, USA, vol. 1, pp. 900–903 (2002)
7. Jacobs, A., Miene, A., Ioannidis, G.T., Herzog, O.: Automatic shot boundary detection combining color, edge, and motion features of adjacent frames. In: TRECVID Workshop Notebook Papers, pp. 197–206 (2004)
8. Haller, M., Kim, H., Sikora, T.: Audiovisual Anchorperson Detection for Topic-Oriented Navigation in Broadcast News. In: IEEE International Conference on Multimedia and Expo., Canada, pp. 1817–1820 (2006)
9. Zlitni, T., Mahdi, W., Ben-Abdallah, H.: A new approach for TV programs identification based on video grammar. In: 7th International Conference on Advances in Mobile Computing and Multimedia, Malaysia, pp. 316–320 (2009)
10. Misra, H., Hopfgartner, F., Goyal, A., Punitha, P., Jose, J.M.: TV News Story Segmentation Based on Semantic Coherence and Content Similarity. In: Boll, S., Tian, Q., Zhang, L., Zhang, Z., Chen, Y.-P.P. (eds.) MMM 2010. LNCS, vol. 5916, pp. 347–357. Springer, Heidelberg (2010)
11. Zlitni, T., Mahdi, W., Ben-Abdallah, H.: Towards a modeling of video grammar based on a priori knowledge for the optimization of the audiovisual documents structuring. In: 2nd International Conference on Computer Technology and Development, Egypt, pp. 517–521 (2010)