A Probabilistic Framework for TV-News Stories Detection and Classification

Francesco Colace[1], Pasquale Foggia[2], Gennaro Percannella[1]
[1] *DIIIE – Università di Salerno, Via Ponte don Melillo, 1 I-84084, Fisciano (SA), Italy*
[2] *DIS – Università di Napoli "Federico II", Via Claudio, 21 I-80125 Napoli, Italy*
fcolace@unisa.it, pasquale.foggia@unina.it, pergen@unisa.it

Abstract

In this paper we face the problem of partitioning the news videos into stories, and of their classification according to a predefined set of categories. In particular, we propose to employ a multi-level probabilistic framework based on the Hidden Markov Models and the Bayesian Networks paradigms for the segmentation and the classification phases, respectively. The whole analysis is carried out exploiting information extracted from the video and the audio tracks using techniques of superimposed text recognition, speaker identification, speech transcription, anchor detection. The system was tested on a database of Italian news videos and the results are very promising.

1. Introduction

Among all the different sources of video material nowadays available, news videos received great attention by the scientific community. This is mainly due to the fact that broadcasters are interested in building large digital databases of their resources, so to allow reuse, after a suitable indexing procedure, of the archived material for other TV programs.

An important step towards an effective indexing is the partitioning of a news video into stories and the classification of the detected stories within a certain set of categories (world news, national news, politics, sports, weather, advertising, etc…). Up to now, research efforts have been concentrated only on the problem of news-story detection. The typical approach [1] implies the partition of the video into sequences of frames, called shots, obtained by detecting transitions that are typically associated to camera changes. Once the shots have been individuated, they are classified on the basis of their content. Two different classes are typically considered: an anchor shot and a news-report shot class. Successively, a news video can be segmented into stories; each story is obtained by linking each anchor shot together with all successive news report shots, until another anchor shot, or the end of the news video, occurs.

This approach is based on simple model of the news story. According to this model an anchor shot represents always the beginning of new story. Consequently, this approach can be profitably applied only for the analysis of the TV-news using the above news story model. However, in the last few years due to the convergence of the technologies of the television, computer and network worlds, more and more broadcasters are proposing brand new and quite complex story models. In these cases, more sophisticated techniques have to be considered in order to provide a reliable news-story detection.

In the recent past, some authors have proposed new approaches for TV-news analysis which try to cope with the limitations deriving from the adoption of a simple model of news story. In particular in [2], the authors propose a two-level multi-modal framework to solve the problem of segmentation and classification of news video into single-story semantic units. At the shot level they propose a decision tree approach to classify the shot according to a wide set of categories (intro, single anchor, two anchors, advertising, weather, sport, … shot). After this first step the detection and classification of the stories is obtained through an HMM approach according to a simple taxonomy (generic news, sport, ...). This approach seems to be very interesting and reliable, but it can only segment an input news video into story units.

So starting from this general framework in this paper we describe our system for TV news stories detection and classification. The two phases are realized in the probabilistic framework provided by Hidden Markov Model and Bayesian Networks, respectively. One of the major contributions of this paper is in the use of Bayesian network for scene classification. In fact with the HMM approach we are able only to segment and not classify the video in scene. Another important aspect is in the generality of our approach. We can use it on various TV News because Bayesian Network can easily model the various TV news structures. In our experimentation, for example, we used news from two different TV networks.

0-7803-9332-5/05/$20.00 ©2005 IEEE

The paper is organized as follows: in Section 2, we provide the motivations and the details of the proposed system architecture, while in section 3 the experimental results are reported. Finally, in section 4 we draw conclusions and indicate future directions of our research.

2. System architecture

The architecture of the proposed system for news video analysis is depicted in Figure 1. The main components of the system are the *news-story detection* and *news-story classification* modules. These modules constitute also the main contribution of this paper. However, for the sake of completeness in the figure also the shot boundary detection module has been included. In fact, this module is necessary as it provides the segmentation of the news video into shots. Once detected, the shots are classified and grouped in story units by the news-story detection module. Finally, the news-stories are classified by the news-story classification module.

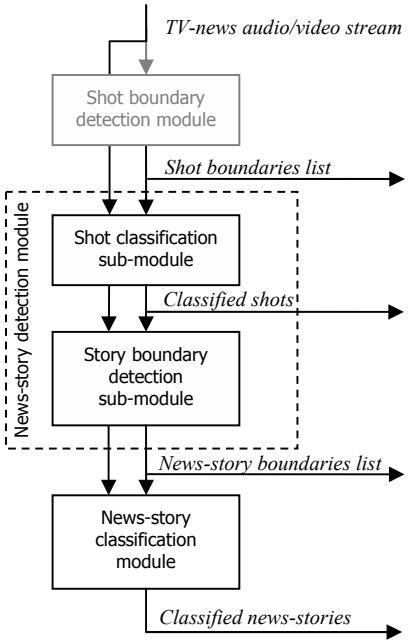

Figure 1. System architecture of the proposed system for news stories detection and classification.

2.1. News Story Detection Module

The news-story detection module receives as input the shot boundaries list and the audio/video stream and provides as output the list of news-story boundaries. This is accomplished through a first *shot classification* phase followed by a *story boundary detection*.

The classification is realized by a specific sub-module that classifies the shot according to the categories proposed in [2] and reported in Table 1.

Table 1. The categories shot proposed in [2] and used in this paper.

Shot Categories		
1) Intro	2) Anchor	3) Double Anchor
4) Meeting	5) Speech/Interview	6) Live-Report
7) Still Image	8) Sports	9) Text-Scene
10) Special	11) Finance	12) Weather
13) Commercials		

2.1.1. Shot Classification Sub-Module. The shot classification sub-module implemented in this paper is inspired by the approach proposed in [3], based on the use of a probabilistic object called *multiject*. The latter map low-level features to high-level semantics. Example of multijects include detectors for explosions, mountains, beaches, generic outdoor environments, music, etc. These semantic concepts interact each other in videos. In order to model this interaction, the authors in [3] propose the integration of multiject in a *Multinet*. Here, we propose to use the formalism of Bayesian Networks, that are capable to generalize the behavior of multijets and are based on a sounder theoretical ground. A Bayesian Network provides a graphical representation of uncertain knowledge that can be easily built and interpreted, with formal probabilistic semantics (see [4] for details). As a consequence Bayesian Networks are very suitable for statistical manipulations.

A Bayesian network encodes the joint probability distribution of a set of random variables $\boldsymbol{X}=\{X_1,..., X_n\}$ using a direct acyclic graph S. Each node of S is associated to a random variable X_i, while each arc between two nodes represents the conditional dependence between the relative random variables. Finally, the description is completed by the set P of the a priori probability distributions of the random variables. In our system the Bayesian Network is structured in two levels. The nodes of the first level of the network are associated to a set of sensors which compute low level features from the audio/video stream of each shot. In Table 2, we have reported the complete list of the used sensors and the states that can be assumed by the relative nodes of the network. At the second level of the network there are 13 nodes corresponding to the categories reported in Table 1. The output of each node of the second level is the probability that the shot under analysis belongs to the class associated to that node. Practically speaking, given the evidences on the sensors, each shot is classified according to the node of the second level that provides the highest probability. The rationale for this structure of the network relies on the fact that the

presence of specific low level features can drive the classification towards a specific category. For instance, the fact that the Face Detector sensor reveals no faces in a shot decreases the probability that the shot belongs to the Anchor category, while it increases the probability that the shot belongs to the Text Scene category.

It is worth noting that, differently from the approach based on multijects, Bayesian Networks not only classify the shots, but they also provide a ranking of the classification. This could allow us to define more sophisticated techniques for shot classification. However, a more important advantage of the Bayesian Networks with respect to multijects relies on the fact that the whole model can be automatically trained for a specific TV network.

Table 2 The sensors used in this paper and the states that can be assumed by the relative nodes of the Bayesian Network.

Sensor	States
Audio	{Silence, Speech, Music, Noise, Speech+Noise, Speech+Music}
Face Detector	{No faces, One face, More than one face }
Text #1 (position)	{Text in the central part of the frame, Text in the lower part of the frame, No text}
Text #2 (words)	{Known words, No known words}
Motion	{No/slow, average, fast}
Color #1 (background)	{Known background, Unknown background}
Color #2 (settings)	{outdoor , indoor}

2.1.2. Story Boundary Detection Sub-Module. The story boundaries are detected by a specific sub-module based on an HMM, using the shot classification and information extracted directly from the audio/video, according to the approach in [2]. In order to identify story boundaries, each shot is modeled by a triple of values: its category (intro, anchor, …), scene/location change and speaker change. For example, an input shot represented by **(Icc)** means that this is an Intro shot with changes in background/location and speaker from the previous shot. Then the sequence of so modeled shots is passed to a HMM that provides the final story segmentation. As in [2], we considered a left to right HMM with 4 internal states.

2.2 News Story Classification Module

This module provides the classification of the news stories detected by the previous stage of the system. Classification is performed through a two levels Bayesian Network, as in the shot classification sub-module. The nodes of the first level are related to the outputs of a set of sensors that process the transcription of the audio track. These sensors extract semantic information on the news story on the basis of the presence of some specific keywords [7]. In particular, we use the same sensor already used in the shot classification sub-module (called Text #2 in Table 2), but trained on a different vocabulary. The nodes of the second level correspond to the news stories classes reported in Table 3.

The presence of this module in our system is a significant improvement with respect to the approach in [2]. In fact, even if the authors proposed a classification of the news stories after the detection, it was carried out with respect to a very restricted set of categories related in a fixed way to the shot categories. On the other side our approach is able to deal with a wider, modifiable set of categories.

Table 3. List of the news story categories.

News story categories	
Introduction	Finance
Local Politics	World Politics
Local News	World News
Sports	Weather Forecast

3. Experimental results

In the recent past some efforts have been spent by other researchers in building video databases for benchmarking purposes [5, 6]. Unfortunately, these data can not be used for our experimentations since in most cases they are not publicly available [5], while in other cases they are not adequate for our aims [6]. Hence, in order to assess the performance of the proposed system we had to build a new test database.

When building a database, it is important to reproduce as much as possible the variability of the phenomenon under study. In our case, the variability is due both to different news video editions of a same TV-network and to different TV-networks. For this reason, the database used in this paper (about four hours) is composed by four news videos from the main Italian public network (namely, RAI 1) and four videos from the main Italian private network (namely, CANALE 5). Then, the tests were carried out on the two TV-networks, separately. Particular care was taken in order to include in the database the main news editions from these TV-networks. In Table 4 it is reported the composition of the database used in this paper.

Table 4. Composition of the database used.

TV-network	Edition	Recording date	Length (mm:ss)
RAI 1	13:30	14-07-2003	27:50
		29-07-2003	26:38
	20:00	10-07-2003	32:54
		11-07-2003	32:02
CANALE 5	13:00	17-07-2004	34:34
		19-07-2004	35:12
	20:00	03-07-2004	30:20
		10-07-2004	31:08

It is worth nothing that here we are interested in assessing the performance of the proposed system without considering the errors due to the sensors and the module of shot boundaries detection. Hence, as in [2] we used no-error (ideal) sensors. It is clear that the reported performance represents an upper bound with respect to the performance that could be obtained using a real implementation of the whole system.

In order to obtain a more realistic estimate of the performance of the proposed system, a four-fold cross validation was performed. Therefore, we divided the database in four subsets for each TV-network. In this case, each subset is composed by a single news video. Then, we performed four tests: in each one, three videos were used as training set and the remaining for testing. Finally, the overall performance for each TV-network is obtained as the average performance on the four folds.

As a first step of our tests, we evaluated the performance of the news-story detection module. Table 5 reports the global performance obtained on the RAI 1 and CANALE 5 data sets by the shot classification and the story boundary detection sub-modules. The performance of the shot classification sub-module is expressed in terms of percentage of the shots which were correctly classified (*CCS*). Differently *Recall* and *Precision* are used to report the performance of the story boundary detection sub-module. Note that the performance of the story boundary detection sub-module is also the performance of the whole news-story detection module.

Table 5. Performance of the *shot classification* and *story boundary detection* sub-modules.

Dataset	*Shot class.*	*Story boundary detection*	
	CCS	Recall	Precision
RAI 1	97.6%	0.935	0.967
CANALE 5	98.0%	0.930	0.952

As a final step of our tests, we evaluated the performance of the news-story classification module. It is worth noting that this module has been tested only on the news stories correctly segmented by the news-story detection module. Hence, the results were obtained only on a subset of the dataset (about 90.3%/90.7% of the whole RAI 1/CANALE 5 dataset). The percentage of news stories correctly classified was 97.6% and 97.4% on RAI 1 and CANALE 5 subsets, respectively. Also in this case it is possible to note that the performance on the two different datasets are very similar.

The experimental results highlight the good overall performance achieved by both the news story detection and classification modules. It is also more interesting to notice that the performance is almost the same on the two different datasets. This result seems to confirm out initial hypothesis that the probabilistic model employed here allows the proposed system to adapt itself to the specific TV network through the training phase.

4. Conclusions and future work

In this paper a system for the automatic detection and the classification of stories in news videos has been presented. The proposed system employs a probabilistic framework based on the Hidden Markov Models and the Bayesian Networks paradigms for the partitioning and classifications phases, respectively. The system has been tested on a database of news videos of two different Italian TV networks. First experimental results are encouraging.

As a future step of our research we intend to evaluate the performance of the proposed system when state of the art sensors are used.

5. References

[1] X. Gao, and X. Tang, "Unsupervised Video-Shot Segmentation and Model-Free Anchorperson Detection for News Video Story Parsing", *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 12, No. 9, pp. 765 776, 2002.

[2] L.Chaisorn, T-S. Chua, C-H. Lee, "A multi-modal approach to story segmentation for news video", *World Wide Web 6(2)*, pp. 187-208, 2003

[3] M.R. Naphade, T.S. Huang, "A Probabilistic Framework for Semantic Video Indexing, Filtering and Retrieval", *IEEE Transactions on Multimedia*, Vol. 3, No. 1, pp 141-151, 2001

[4] F. Jensen, "An Introduction to Bayesian Networks", Springer – Verlag, New York, 1993.

[5] http://www-nlpir.nist.gov/projects/trecvid/trecvid.data.html

[6] ISO/IEC JTC1/SC29/WG11/N2467, Description of MPEG-7 Content Set.

[7] W. Cohen, Y. Singer, "Context Sensitive Learning Methods for Text Categorization", *ACM Trans. Inform. Syst. 17*, 2, 141–173, 1999.