Parsing News Video Using Integrated Audio-Video Features

S. Kalyan Krishna, Raghav Subbarao, Santanu Chaudhury[1],
and Arun Kumar[2]

[1] Department of Electrical Engg., I.I.T,
New Delhi-110016, India
santanuc@ee.iitd.ac.in

[2] Centre for Applied Research in Electronics, I.I.T,
New Delhi-110016, India
arunkm@care.iitd.ac.in

Abstract. In this paper we have proposed a scheme for parsing News video sequences into their semantic components using integrated aural and visual features. We have explored use of the Token Passing Algorithm with HMM for simultaneous segmentation and characterization of the components. Experimentation with about 100 sequences have shown impressive results.

1 Introduction

Content of video sequences like News can be effectively analysed by considering both audio and video cues. For example, field reports in a News Video can not be recognised as an indivisible semantic component without the audio cue as, in general, it consists of multiple shots and scenes. In this paper, we propose a scheme that combines aural and visual cues for parsing a News video sequence into semantic components. This scheme will be useful for development of applications such as News Archival Systems, News-on-Demand Systems, etc.

Some work has been done in the past for interpretation of video sequences by integrating audio and video features. The approach proposed in [1] uses some audio properties along with color and motion information to detect scene and shot breaks. Adams et al. [2] have focused on semantic classification through the identification of meaningful intermediate-level semantic components using both audio and video features. In [3], a method is presented for analysis of News videos by sequentially identifying and segmenting semantic components by exploiting both aural and visual features. A scheme for classification of video scene breaks using both audio and visual features has been suggested in [4]. Dagtas and Abdel-Mottaleb [5] have presented a technique for detecting highlights in a multimedia content using multi-modal features. Another approach for using audio and video features has been suggested in [7].

The method we propose here is different from the previous approaches in many ways. In this approach we have considered a novel way of segmenting

S.K. Pal et al. (Eds.): PReMI 2005, LNCS 3776, pp. 538–543, 2005.

the sequence via recognition of its components through deferred decisions for exploiting contextual information. Use of token passing algorithm with HMM model for this purpose is a novel and useful contribution of this work. Further, it combines audio and video features in an integrated fashion instead of using independent decisions based on individual features. Another advantage of our scheme is its ability to handle large intra-class variations due to the use of a suitable set of stable features and their representation as mixture of Gaussians.

2 News Video Parsing Problem

We address the problem of parsing a News Video sequence into a given set of semantic components with no prior knowledge about the shots in the video sequence.The semantic classes that we have identified and are attempting to detect in a News Video sequence are:

1. Newsroom clips which consist of a News reader presenting News in a studio.
2. Field Report clips which consist of media personnel reporting from the field.
3. Field Interviews / Analysis clips which consist of reporters conducting an interview in the field, persons in the field being interviewed from the Newsroom, or speeches being made in the field.
4. Studio Room Interview / Analysis clips which consist of persons being interviewed by the News reader in the studio, panel discussions etc.
5. Headlines clips n which the News reader deals with multiple news item in brief.

On completion of the parsing process, the complete news sequence is segmented into these semantic components.

3 Feature Description and Extraction

3.1 Visual Features

Field Reports in News Videos have multiple and quick shot breaks whereas News Reader sequences are fairly static and have none or very few shot breaks. Features capturing information about shot breaks in the video sequence provide useful cues for parsing. We have used colour Histogram intersection between consecutive frames for capturing global change in colour distribution. For capturing local changes we have used L1 norm of the image difference between two consecutive frames

Relative motion in the scene is an important visual feature. In News reader and interview components, for example, the motion is less and is localized to small regions whereas in Field Reports, motion is significant and it is not localised. We have used bin-count of histogram of optic flow values as features. In our implementation we used the Lucas-Kanade algorithm for computing optic flow. We have also considered spatial distribution of motion by counting the number of of pixels with high optic flow value in each cell of a 7x7 grid overlaid on the frame.

Another visual cue that characterizes video sequences is the layout of the scene. The layout of News Reader scenes is stable and different from that of different Newsroom scenes or Interview scenes. In Field Reports and Headlines on the other hand the visual layout does not remain constant.We have used wavelet based layout features proposed in [6]. This scheme uses Daubechies Wavelets. We have used variance of wavelet coefficients at different levels as elements of feature vector. Use of layout based features for interpretation of News video has not been explored in the past.

3.2 Audio Features

The common audio scenarios found in news videos are silence, speech, music, environmental sounds and combinations of the latter three. For example, News reader and interview sequences in most News videos consist of pure speech. Field reports on the other hand contain a fair amount of noise due to environmental sounds. Headlines are often accompanied by music. Further, change in the speaker in videos provide a very powerful evidence for transition points and context of the content. Interview sequences, for example, consist of multiple speaker transitions as against News Reader clips which are monologues. Motivated by the above observations, we use statistics of Short Time Energy, Short Time Zero Crossing Rate and Short Time Fundamental Frequency to differentiate between different scenarios. In addition we capture speaker transitions by extracting Short Time Cepstral Coefficients.

3.3 Feature Vector Construction

Following the extraction of audio and visual features we use feature fusion to integrate the two to form one combined audio visual feature vector. The audio features (with the exception of short time fundamental frequency) are produced at the rate of 1 sample / sec. whereas video features are obtained at 25 samples / sec. In order to bring them both to the same temporal scale we repeat our audio features 25 times in every second and then combine them with the visual features. This gives us a feature vector that is generated at the rate of 25 samples / sec. Such a feature vector combines both the audio as well as the visual cues in the video sequence while taking into account their interdependence.

4 Parsing Scheme

The problem of parsing a News Video sequence into a given set of semantic components with no prior knowledge of the location of the breaks between them is very similar to a string parsing problem in continuous speech recognition. Taking into account the success of the Hidden Markov Model paradigm for this problem we decided to use an HMM based classifier for the semantic characterization of News videos.

4.1 HMM Model

A continuous HMM is determined by three groups of parameters: the state transition probability, the observation symbol probability $B = \{b_j(k)\}$; and the initial state distribution. In our case the observations are the feature vectors. We have modelled B as a mixture of Gaussian distribution with a diagonal covariance matrix. Therefore, we assume that the different elements of the feature vector are uncorrelated. We have used EM based approach for estimating Gaussians. The topology of a general HMM used for the semantic characterization is shown in figure 1. As can be seen the first and the last states are assumed to be *non-emitting* to facilitate embedded re-estimation of the HMMs. HMM's corresponding to each class of semantic components are first trained using Baum-Welch algorithm using example sequences. After training, we perform embedded re-estimation. For each training sequence, based purely on the sequence of labels (i.e. ignoring all boundary information) a composite HMM spanning the entire sequence is built. This HMM is retrained using sample sequences.

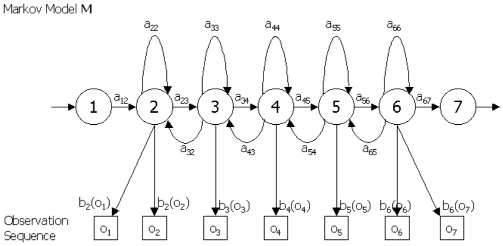

Fig. 1. HMM Generation Model

4.2 Description of the Parsing Algorithm

For a sequence with N frames every path which passes through N emitting states forms a potential transcription of the video sequence. Each of these paths has an associated log-probability which is obtained by adding the log probabilities of the transitions between the consecutive states and the log probability of each emitting state generating the corresponding observation vector. The decoding can now be viewed as the process of finding the path with the highest log probability. We use a variation of the standard Viterbi Algorithm known as the *Token Passing Algorithm* to achieve this.

5 Experimentation and Results

Experiments were conducted on a database consisting of a 100 news video clips (of duration one to three minutes) from 4 News channels broadcasting News in different languages (including BBC and CNN). We have experimentally determined the optimal HMM model for identification of each semantic component.

Using leave-one-out strategy we have obtained ROC curves for each semantic class by changing the number of nodes in the HMM model and obtained optimal configuration for HMMs. We have found that for optimal performance Newsroom HMM require 9 states, Field Report and Field Interview / Analysis HMMs require 10 states and the Headline HMM require 8 states.

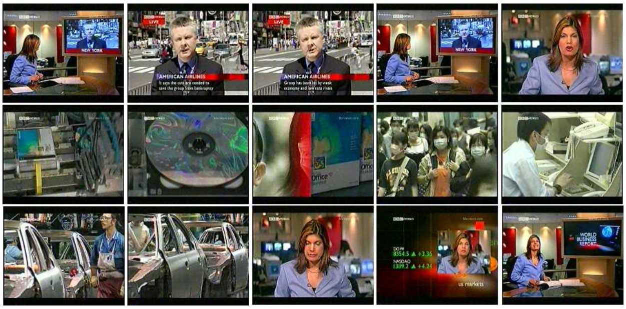

Fig. 2. An example News Video sequence. The frames have been extracted from different semantic components to show the overall structure of the clip.

The figure 2 shows an example News Video sequence. Manual annotation for this News Video sequence (in terms of frame numbers) was as follows: 0 - 135: NEWSROOM; 136 - 1256: FIELD INTERVIEW; 1257 - 1690: NEWSROOM; 1691 - 3159: HEADLINE; and 3160 - 4010: NEWSROOM. The machine annotation produced by our system was: 0 - 132: NEWSROOM - 4501.437500; 133 - 1176 FIELD INTERVIEW - 35901.585938; 1177 - 1406 FIELD REPORT-3241.435791; 1407 - 1690 NEWSROOM - 12143.261719; 1691 - 3251 HEADLINE - 49149.359375;and 3252 - 4010 NEWSROOM - 22067.953125. (The numbers on the side of each segment transcription represent the log probability of the transcription). As can be seen, that using our approach majority of the semantic components have been correctly detected.

We have validated our approach by carrying out leave-one-out test with the complete database. The results have been tabulated in the form of a Confusion Matrix (see table 1). We have found the approach to work well.

Table 1. Confusion Matrix (indicates frame based statistics); N.R: News reader, F.I: field interview, Int.: interview, Hdln: Headline

	N.R	Field	F.I	Int.	Hdln
N.R	87.90%	3.44%	3.14%	3.70%	1.83%
Field	5.83%	83.08%	7.40%	0.86%	2.83%
F.I	7.34%	8.00%	80.88%	2.62%	1.17%
Int.	0.69%	0.22%	2.46%	96.31%	0.32%
Hdline	4.74%	2.55%	4.32%	0%	88.38%

6 Conclusion

In this paper, we have presented a scheme for parsing News video using a combination of audio and video features. Since we have used mixture of Gaussians to handle variations in feature values and token passing algorithm for transcription enabling segmentation of the sequence by taking into account complete temporal context, we have obtained good classification results. Similar approach can be used for parsing other types of video sequences.

Achnowledgements

This work has been supported by the grant for DST project "E-Video: Video Information Processing with Enhanced Functionality."

References

1. Jincheng Huang, Zhu Liu and Yao Wang, "Integration of Audio and Visual Information for Content-based Video Segmentation," *Proc. of IEEE International Conference on Image Processing (ICIP'98),* Invited Paper on "Content-based Video Search and Retrieval", Vol. 3, pp. 526 - 530, Chicago, IL, Oct. 4-7, 1998.
2. W.H. Adams, G. Iyengar, C.Y. Lin, M. R. Naphade, C. Neti,H. J. Nock, and J.R. Smith, "Semantic Indexing of Multimedia using Audio, Text and Visual Cues," EURASIP J. Appl. Signal Processing, pp. 170-185, 2003.
3. Qian Huang, Zhu Liu, Aaron Rosenberg, David Gibbon and Behzad Shahraray "Automated Generation of News Content Hierarchy By Integrating Audio, Video, and Text Information," *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing,* 1999.
4. J. Huang, Z. Liu, and Y. Wang, "Joint Video Scene Segmentation and Classification based on Hidden Markov Model," *IEEE International Conference on Multimedia and Expo (ICME2000),* August 2000, New York.
5. S. Dagtas, Mohamed Abdel-Mottaleb,"Multimodal detection of Highlights for Multimedia Content", Multimedia Systems, Vol. 9, pp. 586-593, 2004.
6. Wang J.Z., Wiederhold G., Firschein O., Wei Xin S., 'Content Based Image Indexing and Searching using Daubechies Wavelets', Digital Library (1997), Pg. 311-328.
7. Sofia Tsekeridou, Ioannis Pitas, "Content-based video parsing and indexing based on Audio-visual interaction", IEEE transactions on circuits and systems for video technology, vol. 11, no. 4, April 2001.