Chapter 4

Audio-Video Segmentation

The previous chapter described feature extraction and modeling techniques used in this work. In this chapter, a method is proposed for segmenting the audio and video data to find the category change point using support vector machine and autoassociative neural network models. The proposed methods for audio and video segmentation are presented in Section 4.2. Section 4.3 describes a method of combining audio and video segmentation results.

4.1 Introduction

Systems that are designed for audio and video classification and indexing for audio and video retrieval usually take segmented audio’s and video’s rather than raw audio/video data as input. Changes in audio/video signal characteristics help in detecting the category change point between different categories of broadcast audio and video. In general, considerable work has been done in audio/music segmentation [47], [49] and [3].

Speech is further classified into pure-speech and non-pure speech [16]. An audio stream is segmented into speech, music, environment sound, and silence [15]. Various acoustic features are short-term energy, zero crossing rates, band periodicity, and noise-frame ratio [26]. BICC based video classification and video shot detection are described in [2]. Various segmentation techniques are described in [3] and [4].
There are several reasons why segmentation is performed for audio-video. It is easier to build analysis and classification systems for segmented objects than for raw data. As the volume of the available material becomes very huge, manual segmentation is impossible. Most of the approaches presented in literature use only the average zero-crossing rate and the energy features, and applies a simple threshold technique, while some of them use a number of features in the time, frequency, and cepstrum domains, as well as model based classification methods to achieve a automatic performance.

Audio and video segmentation, in general, is the task of segmenting a continuous stream of audio and video in terms of homogeneous regions, where the rule of homogeneity depends on the task. A human listener can easily distinguish audio and video signals into these different audio and video types by just listening to a short segment of an sample signal. However, solving this problem using computers has proven to be very difficult [22]. Category change points in audio and video signal such as news to advertisement and advertisement to serial are some examples of segmentation boundaries. Systems that are designed for classifying audio and video signals usually take segmented audio’s and video’s rather than raw audio and video data as input. However, the task in practice is a little more complicated as these transitions are not so obvious all the times both audio and video sequences. For example, the environmental/circumferences change while a news report is broadcast. Thus, many times it is not obvious even to a human abilities, whether a category change point should occur or not.

Changes in audio and video signal characteristics help in detecting the category change point between different categories of broadcast audio. This chapter proposes a method to detect the category change point in broadcast audio and video data consisting of different categories namely advertisement, news, songs, serial, and movie.
The category change point detection is made using mel frequency cepstral coefficients (MFCC) features and color histogram features extracted from the broadcast audio and video data based on support vector machine and autoassociative neural network models (AANN).

4.2 Audio-Video Segmentation

Audio and video segmentation, in general, is the task of segmenting a continuous audio and video stream in terms of acoustically homogeneous regions, where the rule of homogeneity depends on the task. A human identification can easily distinguish audio and video signals into these different audio and video types by just listening to a short segment of an signal. If has more difficulties to identify the remix audio-video datasets which are very popular day today life. Category change points in audio and video signal such as news to advertisement, advertisement to song are some examples of segmentation boundaries. Systems that are designed for classifying audio and video signals usually take segmented audio and video data sets rather than raw audio and video data as input.

Changes in audio and video signal characteristics helps in detecting the category change point between different categories of broadcast audio and video. Most of the approaches presented in literature uses only the audio and the video features, and applied a simple threshold technique to segment the audio and video.

4.2.1 Audio and Video Segmentation using SVM

In this work, the audio and video specific information from the mel frequency cepstral coefficient and color histogram are captured using support vector machine (SVM). The algorithm used for category change detection is described below:
The entire audio/video stream $S$ is divided into $N$ number of analysis frames such that $S = \{s_l : l = 1, \ldots, N\}$ and the feature vector is obtained for each frame.

A window with $L_f$ frames has been selected and a frame $2k+1$ within the window is selected in such a way to separate the window into two portions.

It is assumed that the category change is at this $k + 1^{th}$ frame. Then every frame in the window that is located left of the $k + 1^{th}$ frame is set to the (+1) class and located right of the $k + 1^{th}$ frame is set to the (-1) class.

Now the window is divided into two different sets of frames namely $w^-$ and $w^+$, where $w^- = \{s_l : l = 1, \ldots, k + 1\} \in (+1)$ class and $w^+ = \{s_l : l = k + 1, \ldots, L_f\} \in (-1)$ class.

The SVM is trained using these two classes and the hyperplane is obtained between these two classes.

Fig. 4.1: SVM based segmentation algorithm.
• By using this hyperplane the frames of these two classes are classified. If the window of audio/video stream is from a single category, the hyperplane could not classify the data into two distinct classes. So the rate of misclassification will be very high. On the other hand if the window of audio/video stream is from two different categories, the hyperplane is capable of classifying the data into two distinct classes. Hence, the rate of misclassification will be very low.

• From this we know that the SVM training misclassification rate can be used to decide whether the true category change occurs at the $k + 1^{th}$ frame.

• Two types of misclassification rate are computed namely $SVM^-(k+1)$ and $SVM^+(k+1)$, where $SVM^-(k+1)$ is the rate of the (-1) class misclassified as (+1) and $SVM^+(k+1)$ is the rate of the (+1) class misclassified as (-1).

• If the misclassification rates are smaller than the threshold $t_{mc}$, we can conclude that true category change occurs at $k + 1^{th}$ frame; otherwise it is concluded that there is no category change at $k + 1^{th}$ frame.

• Then the window is shifted one frame to the right of current position and the procedure is repeated. Likewise the entire speech stream must be examined.

The proposed method is unsupervised because it can work without the knowledge of the identity of category’s and there is no need for training category models beforehand. The proposed audio (video) segmentation uses a sliding window of about 2 seconds assuming the category change point occurs in the middle of the window. The sliding window is initially placed at the left end of the audio (video) signal. The SVM is trained to classify the feature vectors in the left half of the window, and the feature vectors in the right half of the window as shown in Fig. 4.1.
The SVM is tested with all these feature vectors. A low misclassification or a high correct classification indicates a category change point such as news to advertisement because the SVM is able to discriminate the two classes. SVM training and testing are repeated by moving the window with a shift one frame until it reaches the right end of audio (video) signal. The proposed method is unsupervised because it can work without the knowledge of the identity of category (audio and video) and there is no need for training category models beforehand.

4.2.2 Experimental Results

The experiments are conducted using the TV tuner card TV PROB II to record television broadcast program. A total dataset of 50 dual categories are used in our studies. This includes 10 categories of each class with another category combinations. A two seconds window size is used in our experiments. SVM classifier with Gaussian kernel function is used. The misclassification threshold \( t_{mc} \) of 0.075 has been used for reliable category change detection.

4.2.3 Audio and Video Segmentation using AANN

In this work, a new approach for audio and video segmentation is proposed using autoassociative neural network (AANN). A five layer autoassociative neural network (AANN) model is used to capture the distribution of the audio/video feature vectors. Autoassociative neural network models are feed-forward neural networks performing an identity mapping of the input space, and are used to capture the distribution of the input data [18]. The second and fourth layers of the network have more units than the input layer. The third layer has fewer units than the first or fifth. The activation functions at the second, third and fourth layer are non-linear.

The structure of the AANN model used in our study are \( 39L \ 78N \ 14N \ 78N \)
39L, and 64L 128N 18N 128N 64L, for audio and video segmentation respectively, where \( L \) denotes a linear unit and \( N \) denotes a non-linear unit. The integer value indicates the number of units used in that layer. The non-linear units use \( \tanh(s) \) as the activation function, where \( s \) is the activation value of the unit. Backpropagation learning algorithm is used to adjust the weights of the network to minimize the mean square error for each feature vector. The algorithm used for category change detection is described below: Given the audio and video features \( S = \{s_l : l = 1, \ldots, N\} \), where \( l \) is the frame index and \( N \) is the total number of frames in the audio/video signal. The proposed algorithm for detecting category change is summarized as follows:

1. A analysis window of \( 2k + 1 \) frames are selected in such a way to separate the window into the portions of \( k \) frames.

2. We consider all the frames in the analysis window that are located left of \( k + 1^{th} \) as left half window \( w_l \) At the same time, we consider all the frames that are
located right of $k + 1^{th}$ as right half window $w_r$.

3. AANN is trained using the frames in $w_l$ and the model captures the distribution of this block of feature vectors. Then feature vectors in $w_r$ are given as input to the AANN model and the output of the model is compared with the input to compute the normalized squared error $e_k$. The normalized squared error ($e_k$) for the feature vector $x$ is given by

$$
e_k = \frac{\|x - o\|^2}{\|x\|^2} \quad (4.1)$$

where $o$ is the output vector given by the model. The error $e_k$ is transformed into a confidence score $s$ using

$$s = \exp(-e_k) \quad (4.2)$$

The average confidence score is calculated by summing the confidence score of the individual frames and the result is divided by the number of frames in the block ($k$). We tried with the weighted summation of the frame scores within the block and there is no improvement in the performance. If true category change occurs at $k + 1$, then $w_l$ and $w_r$ will be from different categories and the average confidence score for this $k + 1^{th}$ will be very low. Likewise, if $k + 1^{th}$ frame is not the true category change point and both $w_l$ and $w_r$ are for the same category then the average confidence score will be very high. The next possibility is that either $w_l$ or $w_r$ may have the audio and video features from both the categories. If this is the case, the average confidence score will be in between the above two values.

4. The value of $k+1$ is incremented by one and the steps from 1 to 3 are repeated until $2k+1$ reaches $N$.  

65
5. Detect the category change points from the confidence score by applying a threshold. The threshold \( t_s \) is calculated from the confidence score as follows:

\[
t_s = s_{\text{min}}(1 + a), \quad 0 < a < 1
\] (4.3)

where \( s_{\text{min}} \) is the global minimum confidence score and \( a \) is the adjustable parameter.

The threshold step is performed as in any other detection algorithm: the threshold is tuned in accordance with some trade off between false alarms and missed detections. The research community tends to treat false alarms as less cumbersome than missed detections. Because over segmentation caused by a high number of false alarms is easier to remedy than under-segmentation caused by high number of miss detections. In our algorithm when \( a \) is nearer to 0, the number of miss detections will be more; if it is nearer to 1, the number of false alarms will be more. So the parameter \( a \) is selected to achieve over segmentation.

The AANN model is trained to capture the distribution of the feature vectors in the left half of the window, and the feature vectors in the right half of the window are used for testing as shown in Fig. 4.2. The structure of AANN model plays an important role in capturing the distribution of the feature vectors.

Average confidence score is obtained for the right half of the window. A low confidence score indicates that the characteristics of the audio (video) signal in the right half of the window are different from the signal in the left half of the window, and hence, the middle of the window is a category change point. The above process is repeated by moving the window with a shift of one frame until it reaches the right end of the audio (video) signal.
4.2.4 Experimental Results

The proposed method uses a sliding window of size 2 seconds assuming the category change point occurs in the middle of the window. The sliding window is initially placed at the left end of the signal. The AANN model is trained to capture the distribution of the feature vectors in the left half of the window, and the feature vectors in the right half of the window are used for testing. The output of the model $o$ is compared with the input to compute the normalized squared error $e_k$. The normalized squared error ($e_k$) for the test feature vector $x$ is given by

$$e_k = \frac{\|x - o\|^2}{\|x\|^2} \quad (4.4)$$

where $o$ is the output vector given by the model. The error $e_k$ is transformed into a confidence score ($s$) using

$$s = \exp(-e_k) \quad (4.5)$$

Average confidence score is obtained for the feature vectors in the right half of the window. The above process is repeated by moving the window with a shift of one frame until it reaches the right end of the signal. The category change points are detected from the average confidence scores by applying a threshold. The threshold ($t_s$) is calculated from the confidence score using

$$t_s = s_{\min}(1 + a), \quad 0 < a < 1 \quad (4.6)$$

where $s_{\min}$ is the global minimum confidence score and $a$ is the adjustable parameter.

A low average confidence score indicates that the characteristics of the signal in the right half of the window are different from the signal in the left half of the window, and hence, the middle of the window is a category change point.
4.3 Combining Audio-Video Segmentation

4.3.1 Performance Measures

The performance of audio segmentation is assessed in terms of two types of error related to category change point detections namely false alarms and missed detections. A false alarm of category change point detection occurs when a detected category change point is not a true one. A missed detection occurs when a true category change cannot be detected. The false alarm rate ($\alpha_r$) and missed detection rate ($\beta_r$) are defined as

$$\alpha_r = \frac{\text{Number of false alarms}}{\text{Number of actual category change points} + \text{Number of false alarms}} \quad (4.7)$$

$$\beta_r = \frac{\text{Number of missed detections}}{\text{Number of actual category change points}} \quad (4.8)$$

A category change detection system has two possible types of errors. Type-I errors occur if a true change is not spotted within a certain window. Type-II errors occur when a detected change does not correspond to a true change in the reference (false alarm). Type-I and Type-II errors also correspond to recall ($RCL$) and precision ($PRC$) respectively, which are defined as

$$PRC = \frac{\text{Number of correctly found category change points}}{\text{Total number of change points found}} \quad (4.9)$$

$$RCL = \frac{\text{Number of correctly found category change points}}{\text{Number of actual category change points}} \quad (4.10)$$

In order to compare the performance of different systems, the F-measure is often used and is defined as

$$F = \frac{2 \times PRC \times RCL}{PRC + RCL} \quad (4.11)$$

The F-measure varies from 0 to 1, with a higher F-measure indicating better performance.

The evidence from audio and video segmentation are combined using weighted sum rule are described below.
The audio and video based misclassification rate obtained from SVM are combined using:

\[ m_{av} = wm_a + (1 - w)m_v \] (4.12)

where

\( m_a, m_v, m_{av} \) are miss classification rate for audio, video, and combined audio-video respectively.

Similarly, the audio and video based confidence scores obtained from AANN are combined using:

\[ s_{av} = ws_a + (1 - w)s_v \] (4.13)

where \( s_a, s_v, s_{av} \) are confidence scores for audio, video, and combined audio-video respectively.

4.3.2 Experimental Results

- **The database**: Performance of the proposed audio and video segmentation system is evaluated using the TV broadcast audio and video data collected from various channels in our regional language, comprising different durations of audio and video namely news, advertisement, sport, serial and movie ranging from 6 seconds to 1 hour. The audio and video consists of varying durations of the categories, i.e. advertisement followed by news, advertisement followed by movie etc., and the audio is sampled at 8 kHz and encoded by 16-bit. Similarly the video data are recorded at 25 frames/seconds, 240 * 320 pixel size of various television channels at different timings to ensure variety of data.

- **Acoustic feature extraction**: The acoustic feature, MFCC, is extracted as de-
scribed in Section 3.3.1. A frame size of 20 ms and a frame shift of 10 ms is used. Hence, an audio signal of 2 seconds to 6 seconds duration results in $200 \times 39$ to $600 \times 39$ feature vectors. The silent frames are not considered for processing.

- **Visual feature extraction**: The 64 dimensional color histogram, is extracted as described in Section 3.3.2. A frame size of 20 ms and frame shift of 10 ms is used. Hence, an video signal of 2 seconds to 6 seconds duration results in $50 \times 64$ to $150 \times 64$ feature vectors. All the frames are considered for processing.

- **Category change point detection**: The sliding window of 2 second is initially placed at the left end of the signal. The misclassification rate and confidence score for the middle frame of the window is computed as described in Section 4.2. The window is shifted by one frame towards right and the same procedure is repeated for the entire signal.

Fig. 4.3 and Fig. 4.4 shows the results for audio, video and audio-video segmentation using SVM and AANN, respectively. The category change points are manually marked. The manual segmentation results are used as the reference for evaluation of the proposed audio segmentation algorithm. The vertical lines show the category change points and the corresponding confidence scores show that these points are correctly detected by our algorithm. In AANN training the number of epochs is varied and the performance is analyzed. There is no significant change in the performance even though the number of epochs is increased to 1000. Hence, the AANN models are trained for only 200 epochs. Fig. 4.5 and Fig.4.6 shows the performance audio, video and audio-video segmentation using SVM and AANN, respectively.
Our experiments show that the category change points are correctly detected most of the times when one of the category is news. This is because the amplitude of the news signal varies a lot from the other categories. This is because MFCC, color histogram features are proven to be one of the best features in literature to distinguish audio and video signals from other categories. However, segments of less than 1 second duration are not detected correctly. But most of the significant true change points are detected precisely by the proposed audio and video segmentation framework.

4.4 Summary

This chapter addresses the issue of detecting category change points in a continuous audio and video stream collected from TV broadcast audio/video data. MFCC and
color histogram features are extracted from the audio and video signal. Using the acoustic and visual features based on SVM and AANN models, a category change point detection algorithm is proposed. This algorithm uses an automatic threshold technique which detects the category change points to a greater extent possible. The performance of the system is evaluated on a large dataset collected from television broadcast audio/video data of various channels. Most of the category change points are correctly detected and the system gives a performance of 99.67% and 98.48% for using SVM and AANN.

Fig. 4.4: Audio-video segmentation using AANN.
Fig. 4.5: Performance of audio-video segmentation using SVM.
Fig. 4.6: Performance of audio-video segmentation using AANN.