In this chapter, a method is proposed for classification of audio-video data. Section 5.2 describes the method for audio, video and combined audio-video classification using SVM. Similarly, Sections 5.3 presents the method for audio, video, and combining audio-video classification using AANN.

5.1 Introduction

Today, digital multimedia applications are part of our day-to-day lives. A typical broadcast audio and video database contains millions of audio and video samples, including news, advertisements, serial, sports, and movie. The need to automatically recognize to which category an audio and video clip belongs, makes audio-video classification an emerging and important research area. Content based classification and retrieval of audio is essentially a pattern recognition problem in which there are two basic issues: feature selection, and classification based on the selected features [3], [9], [143].

An effective representation should be able to capture the most significant properties of audio and video for the task, robust under various circumstances and general enough to describe various audio and video categories. Audio and video feature extraction and classification techniques have been addressed by many existing works over the few years. The general methodology of audio classification involves extracting
discriminatory features from the audio data and feeding them to a pattern classifier. Video classification involves extracting video features from the video data information [253], [256], [96], and [92].

Audio and video classification can provide powerful tools for content based multimedia management. If an audio and video clip can automatically be classified, it can be stored in an organized database, which can improve the management of audio-video dramatically. In this chapter, we analyse effective algorithms to automatically classify audio-video clips into one of five classes: news, advertisement, serial, song, and movie. For these categories, a effective acoustic and visual features that include MFCC and color histogram are extracted to characterize the audio and video content.

SVM is applied to classify audio into their respective classes by learning from training data. SVM finds the optimal class boundary between the different categories. The AANN is used to capture the distribution of the acoustic and visual feature vectors. The AANN model captures the distribution of the acoustic and visual features of a class, and the back-propagation learning algorithm is used to adjust the weights of the network to minimize the mean square error for each feature vector.

5.2 Audio-video Classification using SVM

Support vector machine is trained to distinguish acoustic (visual) features of a category from all other categories. One SVM is created for each category. For testing, acoustic (visual) features are given as input to the SVM models and the distance between each of the feature vectors and the SVM hyperplane is obtained. The average distance is calculated for each model. The category of audio is decided based on the maximum distance. Performance of the proposed audio-video classification system is evaluated using the Television broadcast audio-video database collected from various channels.
5.2.1 Audio and Video Classification using SVM

Audio samples are of different length, ranging from one minute to six minutes, with a sampling rate of 8 kHz, 16-bits per sample, monophonic and 128 kps audio bit rate. The waveform audio format is converted into raw values (conversion from binary into ASCII). Silence segments are removed from the audio sequence for further processing. 39 MFCC coefficients are extracted for each audio sample as described in Section 3.3.1.

Video samples are of different length, ranging from one minute to six minutes, recorded and digitized at a resolution of 320 * 240 pixels and 25 frames per second. From television broadcast video 6 minutes video genres stream are taken for training and testing. Each frame will have 64 dimensional color histogram feature vector. Compared to audio classification, video classification more complicated. A non-linear support vector classifier is used to discriminate the various categories. The N-class classification problem can be solved using N SVMs. Each SVM separates a class from all the remaining classes (one-vs-rest approach).

Support vector machine is trained to distinguish MFCC (color histogram) features of five categories. Support vector machines are created for each category. The training data finds on optimal way to classify audio frames into their respective classes. The derived support vectors are used to classify audio data. For testing MFCC (color histogram) feature vectors are given as input to SVM model and the distance between each of the feature vector and the hyperplane is obtained. The average distance is calculated for each model.

The category of the audio is decided based on the maximum distance. The training data is segmented into fixed-length and overlapping frames (in our experiments we used...
20 ms frames with 10 ms frame shift.) When neighbouring frames are overlapped, the
temporal characteristics of audio and video content can be taken into consideration
in the training process. Since a 8 kHz sampling rate is deployed, 20 ms audio frames
consists of 160 values.

5.2.2 Combining Audio and Video Classification

Combining the modalities has been done at the score level. The methods to com-
bine the two levels of information present in the audio signal and video signal have
been proposed. The evidence from audio and video classifications are combined using
weighted sum rule.

The audio and video classification results obtained by SVM are combined using:

\[ m_j = \frac{w}{n} a_j + \frac{1 - w}{p} v_j, 1 \leq j \leq c, \]  

(5.1)

where

\[ a_j = \sum_{i=1}^{n} x_i^j \]  

(5.2)

\[ v_j = \sum_{i=1}^{p} y_i^j \]  

(5.3)

\[ x_i^j = \begin{cases} 1, & \text{if } c_i^a = j \\ 0, & \text{otherwise} \end{cases} \]  

(5.4)

\[ y_i^j = \begin{cases} 1, & \text{if } c_i^v = j \\ 0, & \text{otherwise} \end{cases} \]  

(5.5)
$c_i^a =$Category label for $i^{th}$ audio frame.

$c_i^v =$Category label for $i^{th}$ video frame.

$v_j =$ video based score for $j^{th}$ category.

$a_j =$ audio based score for $j^{th}$ category.

$m_j =$ Combined audio and video based score for $j^{th}$ category.

$c =$ number of category.

$n =$ number of audio frames.

$p =$ number of video frames.

$w =$ weight.

The category is decided based on the highest $m_j$.

The weight for each of modality is decided by parameter $w$ and it is chosen such that the system gives optimal performance for audio-video based classification. Individual results (audio/video) are combined using weighted sum rule. Audio and video frames are combined based on 4:1 ratio of frame shifts.
5.2.3 Experimental Results

For conducting experiments, audio and video data are recorded using a TV tuner card from various television regional language channels at different timings to ensure quality of database collected and various genres consists of the following data stream: 100 samples of data in all the five predefined categories such as news, advertisement, sports, serial, and movie. The training data set includes 2 to 6 minutes of audio and video stream for each genres. Audio stream is recorded at 8khz with mono channel and 16 bits per sample. Video clips are recorded with a frame resolution of 320*240 pixels and frame rate of 25 frames per second. The audio-video classification results are given in Table 5.1, and Fig 5.1.

![Fig. 5.1: Performance of SVM for audio-video classification.](image-url)
5.3 Audio-video Classification using AANN

Autoassociative neural network is used to capture the distribution of the acoustic (visual) feature vectors of a category. Separate AANN model is trained to capture the distribution of acoustic (visual) feature vectors of each category. For testing, each acoustic (visual) feature vector is given as input to each of the models. The output of the model is compared with the input to compute the normalized squared error. The normalized squared error is transformed into a confidence score as described in Sections 4.2.3. The average confidence score is calculated for each model. The category is decided based on the highest confidence score.

5.3.1 Audio and Video Classification using AANN

The acoustic feature vectors are given as input to the AANN model and the network is trained of 100 epochs. One epoch of training is a single presentation of all training vector. The training takes about 2 minutes on a PC with dual core 2.2 GHz CPU. Similarly the experiments are conducted using histogram as features for video classification. Each frame will have 64 dimensional vector.

The structure of AANN model plays an important role in capturing the distribution of the feature vectors. The number of units in the third layer (compression layer) determines the number of components captured by the network. The AANN model projects the input vectors onto the subspace spanned by the number of units ($n_c$) in the compression layer. If there are $n_c$ units in the compression layer, then the feature vectors are projected onto the subspace spanned by $n_c$ components to realize them at the output layer.
The distribution of the acoustic and visual features are captured using AANN. The class to which the audio (video) sample belongs is decided based on the confidence score.

### 5.3.2 Combining Audio and Video Classification

Proposed audio-video classification algorithm using AANN:

\[
s = \frac{w}{n} \sum_{i=1}^{n} s_i^a + \frac{(1-w)}{p} \sum_{i=1}^{p} s_i^v
\]  

(5.6)

\( n \) is the number of frames in audio signal.

\( p \) is the number of frames in video signal.

\( s_i^a \) is the confidence score of the \( i^{th} \) audio frame.

\( s_i^v \) is the confidence score of the \( i^{th} \) video frame.

\( s \) is the combined audio and video confidence score.

\( w \) is weight.

The category is decided based on the highest confidence score obtained from the models. Audio and video frames are combined based on 4:1 ratio of frame shifts. The weight for each of modality is decided by the parameter \( w \) is chosen such that the system gives optimal performance for audio-video based classification.
5.3.3 Experimental Results

The performance of the proposed audio-video classification system is evaluated using the TV broadcast audio-video data collected from various channels, comprising different durations of audio-video ranging from five seconds to one hour. The acoustic feature MFCC is extracted as described in Section 3.3.1. The color histogram features are extracted as described in Section 3.3.2. A frame size of 20 ms and a frame shift of 10 ms is used. The audio-video classification results are reported in Table 5.2, and Fig. 5.2.

Fig. 5.2: Performance of AANN for audio-video classification.

5.4 Summary

In this chapter, support vector machine (SVM) and autoassociative neural network (AANN) models are used for modelling the features. MFCC and color histogram features are extracted to characterize audio and video content. A non-linear support
vector machine learning algorithm is applied to obtain the optimal class boundary between the various classes namely news, advertisement, serial, sports and movie, by learning from training data. The average performance considering different duration of training data for audio-video classification using SVM and AANN is shown in Fig 5.3. The evidence from acoustic and visual features are combined using weighted sum rule.

![Fig. 5.3: Average performance of audio-video classification using SVM and AANN.](image)

Experimental results show that proposed audio-video classification gives an accuracy of 98.72% and 96.21% using SVM and AANN, respectively.
Table 5.1: Performance of audio, video and audio-video classification using SVM (in %).

<table>
<thead>
<tr>
<th>Category</th>
<th>Modality</th>
<th>Advt.</th>
<th>News</th>
<th>Sports</th>
<th>Serial</th>
<th>Movie</th>
</tr>
</thead>
<tbody>
<tr>
<td>Advt.</td>
<td>Audio</td>
<td>97.4</td>
<td>0.78</td>
<td>0.0</td>
<td>0.0</td>
<td>1.82</td>
</tr>
<tr>
<td></td>
<td>Video</td>
<td>97.96</td>
<td>1.38</td>
<td>0.44</td>
<td>0.0</td>
<td>0.22</td>
</tr>
<tr>
<td></td>
<td>Audio-Video</td>
<td>99.04</td>
<td>0.64</td>
<td>0.0</td>
<td>0.0</td>
<td>0.32</td>
</tr>
<tr>
<td>News</td>
<td>Audio</td>
<td>1.42</td>
<td>96.09</td>
<td>0.50</td>
<td>0.50</td>
<td>1.49</td>
</tr>
<tr>
<td></td>
<td>Video</td>
<td>0.78</td>
<td>96.22</td>
<td>0.0</td>
<td>1.15</td>
<td>1.85</td>
</tr>
<tr>
<td></td>
<td>Audio-Video</td>
<td>0.59</td>
<td>98.18</td>
<td>0.0</td>
<td>0.57</td>
<td>0.66</td>
</tr>
<tr>
<td>Sports</td>
<td>Audio</td>
<td>2.08</td>
<td>0.96</td>
<td>95.48</td>
<td>0.0</td>
<td>1.48</td>
</tr>
<tr>
<td></td>
<td>Video</td>
<td>1.98</td>
<td>1.09</td>
<td>96.46</td>
<td>0.47</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>Audio-Video</td>
<td>0.58</td>
<td>0.38</td>
<td>99.04</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Serial</td>
<td>Audio</td>
<td>1.73</td>
<td>2.87</td>
<td>0.0</td>
<td>94.9</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td>Video</td>
<td>0.86</td>
<td>1.47</td>
<td>0.0</td>
<td>96.28</td>
<td>1.39</td>
</tr>
<tr>
<td></td>
<td>Audio-Video</td>
<td>0.54</td>
<td>0.87</td>
<td>0.0</td>
<td>97.99</td>
<td>0.60</td>
</tr>
<tr>
<td>Movie</td>
<td>Audio</td>
<td>1.17</td>
<td>1.02</td>
<td>0.0</td>
<td>1.03</td>
<td>96.78</td>
</tr>
<tr>
<td></td>
<td>Video</td>
<td>0.92</td>
<td>0.78</td>
<td>0.86</td>
<td>0.66</td>
<td>96.78</td>
</tr>
<tr>
<td></td>
<td>Audio-Video</td>
<td>0.30</td>
<td>0.60</td>
<td>0.0</td>
<td>0.58</td>
<td>98.52</td>
</tr>
</tbody>
</table>
Table 5.2: Performance of audio, video and audio-video classification using AANN (in %).

<table>
<thead>
<tr>
<th>Category</th>
<th>Modality</th>
<th>News</th>
<th>Advt.</th>
<th>Sports</th>
<th>Serial</th>
<th>Movie</th>
</tr>
</thead>
<tbody>
<tr>
<td>News</td>
<td>Audio</td>
<td>94.75</td>
<td>0.96</td>
<td>1.83</td>
<td>1.61</td>
<td>0.85</td>
</tr>
<tr>
<td></td>
<td>Video</td>
<td>94.09</td>
<td>1.73</td>
<td>1.13</td>
<td>1.48</td>
<td>1.57</td>
</tr>
<tr>
<td></td>
<td>Audio-Video</td>
<td>96.07</td>
<td>1.64</td>
<td>0.56</td>
<td>0.48</td>
<td>1.25</td>
</tr>
<tr>
<td>Advt.</td>
<td>Audio</td>
<td>0.99</td>
<td>94.14</td>
<td>1.30</td>
<td>1.62</td>
<td>1.95</td>
</tr>
<tr>
<td></td>
<td>Video</td>
<td>1.14</td>
<td>94.19</td>
<td>1.51</td>
<td>1.64</td>
<td>1.52</td>
</tr>
<tr>
<td></td>
<td>Audio-Video</td>
<td>0.32</td>
<td>96.78</td>
<td>1.43</td>
<td>1.16</td>
<td>0.31</td>
</tr>
<tr>
<td>Sports</td>
<td>Audio</td>
<td>2.21</td>
<td>1.76</td>
<td>93.23</td>
<td>1.57</td>
<td>1.23</td>
</tr>
<tr>
<td></td>
<td>Video</td>
<td>1.27</td>
<td>1.20</td>
<td>94.44</td>
<td>1.57</td>
<td>1.52</td>
</tr>
<tr>
<td></td>
<td>Audio-Video</td>
<td>0.98</td>
<td>0.51</td>
<td>96.17</td>
<td>1.22</td>
<td>1.12</td>
</tr>
<tr>
<td>Serial</td>
<td>Audio</td>
<td>1.48</td>
<td>1.46</td>
<td>1.54</td>
<td>94.08</td>
<td>1.44</td>
</tr>
<tr>
<td></td>
<td>Video</td>
<td>1.15</td>
<td>1.73</td>
<td>1.24</td>
<td>94.41</td>
<td>1.47</td>
</tr>
<tr>
<td></td>
<td>Audio-Video</td>
<td>1.60</td>
<td>1.24</td>
<td>0.82</td>
<td>95.37</td>
<td>0.97</td>
</tr>
<tr>
<td>Movie</td>
<td>Audio</td>
<td>1.18</td>
<td>2.03</td>
<td>2.06</td>
<td>1.73</td>
<td>93.00</td>
</tr>
<tr>
<td></td>
<td>Video</td>
<td>2.12</td>
<td>1.71</td>
<td>1.27</td>
<td>0.9</td>
<td>94.00</td>
</tr>
<tr>
<td></td>
<td>Audio-Video</td>
<td>1.16</td>
<td>0.79</td>
<td>1.49</td>
<td>0.93</td>
<td>95.63</td>
</tr>
</tbody>
</table>