A Multi-Modal Approach To Story Segmentation for News Video

Lekha Chaisorn, Tat-Seng Chua, and Chin-Hui Lee
The School of Computing, National University of Singapore
Singapore 117543 Tel: +65 6874 2505, 6874-4302 Fax: +65 6774-4401
[Lekhacha, chuats, chl]@comp.nus.edu.sg

Abstract This research proposes a two-level, multi-modal framework to perform the segmentation and classification of news video into single-story semantic units. The video is analyzed at the shot and story unit (or scene) levels using a variety of features and techniques. At the shot level, we employ Decision Trees technique to classify the shots into one of 13 pre-defined categories or mid-level features. At the scene/story level, we perform the HMM (Hidden Markov Models) analysis to locate story boundaries. Our initial results indicate that we could achieve a high accuracy of over 95% for shot classification and over 89% in $F_1$ measure on scene/story boundary detection. Detailed analysis reveals that HMM is effective in identifying dominant features, which helps in locating story boundaries. Our eventual goal is to support the retrieval of news video at story unit level, together with associated texts retrieved from related news sites on the web.

Keywords: News Story Segmentation; Shot Classification; Multi-modal Approach; Learning-based Approach;

1. INTRODUCTION

The rapid advances in computing, multimedia, and networking technologies have fueled the wide spread adoption of World Wide Web (WWW). This in turn has resulted in the production and distribution of large amount of multimedia data, especially digital video. These videos are used mostly in web-based applications such as Distance Learning [13], Web Digital library [3], Television News Archive [23] in order to support a wide variety of users of different interests to browse, interact, and retrieve multimedia data.

In particular, news video is one of the essential information that users will access daily. To facilitate user personalization, we need to understand users' interests and the structure of news in general. Some users may want to access to business or finance news, some may want to check weather forecasting, and some may only want to view the score of football match, etc. Most importantly, users
normally want to access appropriate footages of news video in conjunction with the corresponding text news available in various news web sites. To support these features effectively, we must be able to break a long sequence of news video into stories, and organize the video hierarchically based on stories with appropriate text of these news extracted from the web. Such organization will facilitate user access and support personalization of news video.

Research on segmenting an input video into shots, and using these shots as the basis for video organization, is well established [24][18]. A shot represents a contiguous sequence of visually similar frames. It, however, does not usually convey any coherent semantics to the users. As users remember video contents in terms of events or stories but not in terms of changes in visual appearances as in shots, it is necessary to organize the video contents in terms of small, single-story unit that represents the conceptual chunks in users' memory. These video units can further be classified according to their semantics [5] such as meeting, interview, sunset, etc, and organized hierarchically to facilitate browsing.

This paper aims at developing a system to automatically segment and classify news video into semantic units using a learning-based approach. It is well known that the learning-based approaches are sensitive to feature selection and often suffers from data sparseness problems due to difficulties in obtaining annotated data for training. One approach to tackle the data sparseness problem is to perform the analysis at multiple levels as is done successfully in natural language processing (NLP) research [10]. For example, in NLP, it has been found to be effective to perform the part-of-speech tagging at the word level, before the phrase or sentence analysis at a higher level. Thus in this research, we propose a two-level, multi-modal framework to tackle the news story boundary detection problem. The video is analyzed at the shot and story unit (or scene) levels using a variety of features. At the shot level, we use a set of low-level and high-level features to model the contents of each shot. We employ the Decision Tree to classify the video shots into one of the 13 pre-defined categories. At the story level, we perform HMM analysis [22] to identify news story boundaries.

To focus our research, we adopt the news video domain as such video is more structured and has clearly defined story units. Our eventual goal is to support the retrieval of story unit level, together with the associated text news stories retrieved from various sites on the web.

Briefly, the content of this paper is organized as follows. Section 2 describes related work and Section 3 discusses the design of the multi-modal two-level classification framework. Section 4 presents the details of shot level classification and Section 5 discusses the details of story/scene segmentation. Section 6 presents the experimental results, and Section 7 contains our conclusion and discussion of future work.
2. RELATED WORK

Video classification is a hot topic of research for many years and much interesting research has been done. Because of the difficulty and often subjective nature of video classification, most early works examined only certain aspects of video classification in a structured domain such as sports or news.

Ide et al. [15] tackled the problem of news video classification and used video text, motion, and face as the features. They first segmented the video into shots and used multiple techniques including clustering to classify each shot into one of the five classes: Speech/report, Anchor, Walking, Gathering, and Computer Graphics. Their classification technique seems effective for this restricted class of problems. Zhou et al. [26] examined classification of basketball videos into a set of restricted categories of Left-court, Middle-court, Right-court, and Closed-up. They considered only motion, color, and edges as the features and employed a rule-based approach to classify each video shot (represented using a key frame). Chen and Wong [7] also used a rule-based approach to classify news video into six classes of news, Weather, reporting, commercials, basketball, and football. They used the feature set of motion, color, text caption, and cut rate in the analysis.

Another category of techniques incorporated information within and between video segments to determine class transition boundaries using mostly the HMM approaches. Eickeler et al. [12] considered 6 features, deriving from the color histogram and motion variations across the frames, and employed HMM to classify the video sequence into the classes of Studio Speaker, Report, Weather Forecast, Begin, End, and the editing effect classes. Huang et al. [14] employed audio, color, and motion as the features and classified the TV programs into the categories of news report, Weather forecast, commercials, basketball games, and football games. Alatan et al. [1] aimed to detect dialog and its transitions in fiction entertainment type videos. They modeled the shots using the features of audio (music/silence/speech), face, and location change, and used HMM to locate the transition boundary between the classes of Establishing, Dialogue, Transition, and Non-dialogue.

In summary, most reported works considered only a limited set of classes and features, and provided only partial, intermediate solutions to the general video organization problem. In our work, we want to consider all possible categories of shots and scenes to cover potentially all types of news video. Another major difference between our approach and existing works is that we perform the story segmentation at two levels, similar to the approach successfully employed in NLP research.
3. THE MULTI-MODAL TWO-LEVEL FRAMEWORK

3.1 Structure of News

Before we discuss about the design of our approach, we shall describe the structure of general news video structure. Most news videos have rather similar and well-defined structures. The news video typically begins with several Intro/Highlight shots that give a brief introduction of the upcoming news to be reported. The main body of news contains a series of stories organized in terms of different geographical interests (such as international, regional and local) and in broad categories of social political, business, sports and entertainments. Each news story normally begins with Anchor-person shots. Most news ends with reports on Sports, Finance, and Weather. In a typical half an hour news, there will be at least one period of commercials, covering both commercial product and self-advertisement by the broadcast station. Figure 1 illustrates the structure of a typical news video.

![Diagram of news video structure]

**Figure 1:** The structure of local news video under study

Although the ordering of news items may differ slightly from broadcast station to station, they all have similar structure and news categories. In order to project the identity of a broadcast station, the visual contents of each news category, like the anchor person shots, Finance and Weather reporting etc., tends to be highly similar within a station, but differs from that of other broadcast stations. Hence, it is possible to adopt a learning-based approach to train a system to recognize the contents of each category within each broadcast station.

3.2 The design of a news classification and segmentation system

Although news video is structured, it presents great challenges in classifying them and in identifying story boundaries. The classification is difficult because there are many categories that are highly similar and can only be differentiated by using an appropriate combination of features. Examples of similar and ambiguous categories include: (a) the speech, interview, and meeting shots; (b) certain live reporting and sports; and (c) between different types of sports. For example, we might need a combination of face, text caption, visual background and audio features to differentiate between anchor-person, interview and meeting shots. The
identification of story boundaries is even more difficult as it requires both visual and semantic information.

To tackle the problem effectively, we must address three basic issues. First, we need to identify the suitable units to perform the analysis. Next, we need to extract an appropriate set of features to model and distinguish different categories. Third, we need to adopt an appropriate technique to perform the classification and identify the boundaries between stories. To achieve this, we adopt the following strategies as shown in Figure 2:

![Diagram of overall system components]

*Figure 2: Overall system components*

a) We first segment the input video into shots using a mature technique. Here, we employ the temporal multi-resolution analysis [TMRA] technique developed in our lab [2][18].
b) We extract a suitable set of features to model the contents of shots. The features include low level visual and temporal features, and high-level features like faces. We select only those features that can be automatically extracted in order to automate the entire classification process.

c) We employ a learning-based approach that uses multi-modal features to classify the shots into the set of well-defined categories.

d) Finally, given a sequence of shots in respective subcategories, we use a combination of shot content features, categories, and temporal features to identify story boundaries using the HMM technique.

The detailed design of the system and the choice of features are discussed in Sections 4 and 5.

4. THE CLASSIFICATION OF VIDEO SHOTS

This section describes the details of shot classification, including: shot segmentation; choice of appropriate shot categories and feature set; and the classification process.

4.1 Shot Segmentation and Key Frame Extraction

The first step in news video analysis is to segment the input news video into shots. For this step, we employ the temporal multi-resolution analysis (TMRA) technique developed in our lab [2] that can effectively locate both abrupt and gradual transition boundaries effectively. The TMRA framework also selects a best representative key frame for each shot.

After the video is segmented, there are several ways in which the contents of each shot can be modeled. We can model the contents of the shot: (a) using a representative key frame; (b) as feature trajectories [6]; or, (c) using a combination of both. In this research, we adopt the hybrid approach as a compromise to achieve both efficiency and effectiveness. Most visual content features will be extracted from the key frame while motion and audio features will be extracted from the temporal contents of the shots. This is reasonable as we expect the visual contents of shots to be relatively similar so that a key frame is a reasonable representation.

4.2 Shot Categories

4.2.1. The Selection of Shot Categories

The next step is to determine an appropriate and complete set of categories to cover all shot types. The categories must be meaningful so that the category tag assigned to each shot is reflective of its content and facilitates the subsequent stage of segmenting and classifying news stories. We studied the set of categories
employed in related works, and the structures of news video in general and local news in particular. We arrive at the following set of shot categories as shown in Figure 3. They are Intro/Highlight, Anchor, 2Anchor, Meeting/Gathering, Speech/Interview, Live-reporting, Still-image, Sports, Text-scene, Special, Finance, Weather, and Commercial. These 13 categories cover all essential types of shots in typical news video. Some categories are quite specific such as the Anchor or Speech categories. Others are more general like the Sports, Special or Live-reporting categories.

![Image of shot categories]

*Figure 3: Examples of the predefined categories and example shots*

### 4.2.2 Characteristics of shots in each category

Here we discuss the characteristics of each shot category.

- **Commercials.** Commercials are used to present messages or to sell products. Because it is expensive to air commercials especially during prime hours, the commercials tend to be short and packed with product-oriented information. Thus commercials typically contain fast changing shots, and end with still images showing the company’s logos or products. In most countries, it is mandatory to air several black frames preceding and after a block of commercials. However, this is not always the case in many countries, like in Singapore. Nevertheless, our studies show that commercial boundaries can normally be characterized by the presence of black frames, still frames and/or audio silence [17].

- **Intro/Highlight.** At the beginning of news reports, most broadcast stations air several Intro/Highlight shots to give the highlights of the upcoming news to be reported. It runs through the important news to be covered in condensed and entertaining form. These shots contain speech with background music,
and in shot durations. In the news under study, the duration of each of these shots is less than 10 second on average.

- **Anchor.** This is the most typical shot type in news reports with one anchor person appearing in fixed background. Thus, the shot is normally of closed up or medium distance shot containing one detected frontal face. The shot is usually of closed-up or medium distance shot containing one detected frontal face. Also, at the beginning of the anchor shots, there are usually one to two lines of text at the bottom stating the anchor person's name and title. The shot duration is usually long, and the motion (only the face region) is low.

- **2Anchor.** Shots in this category containing two anchor persons and are usually shown as part of transition from one topic to the next. They are usually medium shots with two detected (frontal) faces. The shot duration is usually short (switch topics), and the motion (only the faces region) is again low.

- **Meeting/gathering.** This type of shots usually contains at least two detected faces with similar sizes in a frame. However, the accuracy is dependent very much on face detection algorithm. If the faces in the frame are not large enough, then some of these shots will be misclassified. Shot duration of this type varies from medium to long.

- **Speech/Reporting.** When a person is giving a speech, or a reporter is reporting from a relay spot, there is usually one person speaking in the middle of a frame. This type of shot is similar to Anchor shot, thus additional technique is required to solve the ambiguity. One technique is to perform clustering to identify the background of Anchor person shots since anchor person appears much more frequently in news with almost identical background than the person in Speech.

- **Still image.** If one or more faces are detected with or without text captions and there is no motion in the shot, it is possible that the shot comprises still images.

- **Finance.** This type of shots is characterized by image content. Based on our video samples, Finance shots normally appear in the middle of the news. The visual contents of these shots, in terms of colors, position of text, music, etc., are almost the same. Thus, we can effectively identify such shots simply by using the image similarity between an unseen shot (represented by its key frame) and the representative key frame of Finance shots.

- **Weather.** This type of shots is similar to Finance shots in such a way that the visual contents of these shots are very similar. Thus we can again use image similarity measure to identify these shots.

- **Sports.** In our training and testing samples, there are a few types of sport that are commonly reported: football, basketball, and tennis. These shots normally contain background noise and high motion. In this work, we simply use these
characteristics to classify Sport shots as a general category. Additional techniques are required in order to further classify the shots into specific types of sport, such as the football, basketball, etc.

- **Text-scene.** This type of shots is used to present the summary of events such as the results of a sport game, or latest currency exchange rates etc. They typically contain only multiple lines of normally centralized text.

- **Special.** This is a special type of category for the news under study. It is used to present typically light hearted events. For example, in our sample, one such shot is about a dolphin giving birth to a baby. They thus comprise of some speech in music background.

- **Live-Reporting.** Live-reporting shots typically involve the reporting of an event, follow by footages of the event. The event can be of any types and in any environment. Thus it is hard to detect such shots. In this research, we simply classify those shots that do not fit into the above categories as Live-reporting shots.

4.2.3 Relationship between shot categories and story units

In general, news video contains a series of stories organized in terms of different geographical interests (such as international, regional and local) and in broad categories of social political, business/finance, entertainment, sports and weather. Each story may contain shots of multiple categories beginning with an Anchor shot, followed by several Meeting, Live-reporting (not shown), and/or Speech shots as can be seen in Figure 4.

![Diagram showing the relationship between shot categories and story units](image)

*Figure 4: The relationship between shot categories and story units*
Different news stories may contain similar combination of shot categories. However, the patterns, order, and the frequencies of shot categories are different. For example, general news reporting, e.g. world news will contain mostly anchor person shots, with some meeting, live-reporting, and a few on interview shots. On the other hand, the sports story will contain mostly the actual sports footages, possibly with some interview and Text-scene shots.

4.3 Choice and extraction of features for shot representation

The choice of suitable features is critical to the success of most learning-based classification systems. Here, we aim to derive a comprehensive set of features that can be automatically extracted from MPEG video to facilitate shot classification.

4.3.1 Low-level Visual Content Feature

**Color Histogram**: Color histogram models the visual composition of the shot. It is particularly useful to resolve two scenarios in shot classification. First, it can be used to identify those shot types with similar visual contents such as the Weather and Finance categories. Second, the color histogram can be used to model the changes in background between successive shots, which provides important clues to determining a possible change in shot category or story. Here, we represent the content of key frame using a 256-color histogram.

4.3.2 Temporal Features

**Background scene change**: We use the background scene change feature to measure the difference between the color histogram of the current and previous shots. It is represented by ‘c’ if there is a change and ‘u’ otherwise.

**Audio**: This feature is very important especially for Sport and Intro/Highlight shots. For Sport shots, its audio track includes both commentary and background noise, and for Intro/Highlight shots, all the narrative is accompanied by background music. Here, we adopt an algorithm similar to that discussed in Lu et al. [20] to classify audio into the broad categories of speech, music, noise, speech and noise, speech and music, or silence.

**Speaker change**: Similar to background scene change feature, this feature measures whether there is a change of speaker between the current and previous shot. It takes the value of ‘u’ for no change, and ‘c’ if there is a change. The later condition also applies to shots that do not contain speech but when there is a change from the previous speech to non-speech shot or vice versa. The detection of non-speech shot from speech shot can be done by detecting the shot’s audio type as described earlier.

**Motion activity**: For MPEG video, there is a direct encoding of motion vectors, which can be used to indicate the level of motion activities within the shot. We
usually see high level of motion in sports and certain live reporting shots such as the rioting scenes. Thus, we classify the motion into low (like in an Anchor-person shot where only the head region has some movements), medium (such as riot shots), high, or no motion (for still frame shots).

**Shot duration:** For Anchor-person or Interview type of shots, the duration tends to range from 20 to 50 seconds. For other types of shots, such as the Live-reporting or Sports, the duration tends to be much shorter, ranging from a few seconds to about 10 seconds. The duration is thus an important feature to differentiate between these types of shots. We set the shot duration to short (if it is less than 10 seconds), medium (if it is between 10 to 20 seconds), and long (for shot greater than 20 seconds in duration).

### 4.3.3 High-level Object-based features

**Face:** Human activities are one of the most important aspects of news videos, and many such activities can be deduced from the presence of faces. Many techniques have been proposed to detect faces in an image or video. In our study, we adopt the algorithm developed in Chua et al. [9] to detect mostly frontal faces in the key frame of each shot. We extract in each shot the number of faces detected as well as their sizes. Figure 5 shows the result of detecting mostly frontal faces.

**Shot type:** We use the camera focal distance to model the shot types, which include close-up, medium-distance or long-distance shot etc. Here, we simply use the size of the detected faces to estimate the shot type.

![Figure 5: The result of face detection of example shots](image)

**Videotext:** Videotext is another type of object that appears frequently in news video and can be used to determine video semantics. We employ the algorithm developed in Zhang and Chua [25] to detect videotexts. For each shot, we simply determine the number of lines of text appearing in the key frame. Figure 6 presents the result of text detection in sample video shots.
Centralized Videotext: We often need to differentiate between two types of shots containing videotexts. The normal shot where the videotexts appearing at the top or bottom of a shot to indicate its contents; and the Text-scene shot where only a sequence of texts is displayed to summarize an event, such as the results of a soccer game. A text-scene shot typically contains multiple lines of centralized text, which is different from normal shots that may also contain multiple lines of text but normally un-centralized. Hence, we include this feature to identify Text-scene shots. It takes the value “true” for centralized text and “false” otherwise.

4.4 Shot representation

After all features are extracted, we represent the contents of each shot using a color histogram vector and a feature vector. The histogram vector is used to match the content of a shot with the representative shot of certain categories, while the feature vector is used by the classifier to categorize the shots into one of the remaining categories.

The feature vector of a shot is of the form:

\[ S_t = (a, m, d, f, s, t, c) \]  

where

- \( a \) the class of audio, \( a \in \{ \text{speech}, \text{music}, \text{silence}, \text{noise}, \text{nm= speech + noise, \text{mn= music+noise}} \} \)
- \( m \) the motion activity, \( m \in \{ \text{l=low, m=medium, h=high} \} \)
- \( d \) the shot duration, \( d \in \{ \text{s=short, m=medium, l=long} \} \)
- \( f \) the number of faces, \( f \in \mathbb{N} \)
- \( s \) the shot type, \( s \in \{ \text{c= closed-up, m=medium, l=long, u=unknown} \} \)
- \( t \) the number of lines of text in the scene, \( t \in \mathbb{N} \)
- \( c \) the text centralized feature. It is set to “true” if the videotexts present are centralized, \( c \in \{ \text{t=true, f=false} \} \)
For example, the feature vector of an Anchor-person shot may be \((t, 1, 1, 1, c, 2, f)\). Note that at this stage we did not include the scene change and speaker change features in the feature set. These two features are not important for shot classification and will be included in story boundary detection using HMM.

### 4.5 The classification of video shots

There are three steps in the shot classification process. They are commercials removal, Weather and Finance shots filtering, and shot classification using decision tree.

#### 4.5.1 Commercials removal

As we described earlier, commercial boundaries can normally be characterized by the presence of black frame, still frames and audio silence, or any combination of all of them. We therefore employ a heuristic approach to identify the presence of commercials and detect the beginning and ending of the commercial blocks [17]. Our tests on six news videos (180 minutes) obtained from the MediaCorp of Singapore demonstrate that we are able to achieve a higher detection accuracy of over 97%. Figure 7 shows the details of the commercial removal process.

![Diagram of commercial removal process](image)

**Figure 7:** The commercial removal process

#### 4.5.2 Weather and Finance shots Detection

For the remaining shots, we first identify the shot types that have very similar visual features. Examples of these shot types include the Weather and Finance reports. For these shot types, we simply extract the representative histogram of the respective categories and employ the histogram-matching algorithm developed in Chua & Chu [8] to compute the shot-category similarity that takes into consideration the perceptually similar colors. We employ a high threshold of 0.8 to determine whether a given shot belongs to the Weather or Finance
category. By simply using this measure, we could achieve a very high classification accuracy of over 95% for these two categories.

4.5.3 Classification using Decision Trees

After the Commercial, Weather and Finance shots are filtered out, we employ a Decision Tree [16] to classify the rest of the shots using a learning-based approach. Decision Tree (DT) is one of the most widely use methods in machine learning. The Decision Tree has the advantages that it is robust to noisy data, capable of learning disjunctive expression, and the training data may contain missing or unknown values [21]. The Decision Tree has been successfully employed in many multi-class classification problems [11][26]. We thus select the Decision Tree for our shot classification problem.

Let \( n_b \) the number of instances in branch \( b \), \( n_{bc} \) the total number of positive instance on branch \( b \) of class \( c \), and \( n_p \) the total number of instances in all branches. The formulas for computing the entropy and average entropy are expressed as follows.

\[
\text{Entropy} = \sum_c -\left( \frac{n_{bc}}{n_p} \right) \log_2 \left( \frac{n_{bc}}{n_p} \right)
\]

(2)

\[
\text{AverageEntropy} = \sum_b \left( \frac{n_b}{n_p} \right) \sum_c -\left( \frac{n_{bc}}{n_p} \right) \log_2 \left( \frac{n_{bc}}{n_p} \right)
\]

(3)

At each stage of the learning process, the Decision Tree will calculate the average entropy of each attribute, and select the attribute with the lowest entropy as the branching attribute to determine the classes of the remaining training samples. The process stops when all the leaf nodes have only one individual class. The Decision Tree created during training process is used in the testing stage to classify the remaining shots.

5. SCENE/STORY BOUNDARY DETECTION

After the shots have been classified into one of the pre-defined categories, we employ HMMs to detect scene/story boundaries. We use the shot sequencing information, and examine both the tagged categories and appropriate features of the shots to perform the analysis. This is similar to the idea of part-of-speech (POS) tagging problem in NLP that uses a combination of POS tags and lexical information to perform the analysis.

HMM is a powerful statistical tool first successfully utilized in speech recognition research. HMM contains a finite set of states, each of which is associated with a probability distribution. Transitions among the states are governed by a set of probabilities called transition probabilities. In a particular state, an outcome or observation can be generated according to the associated
probability distribution. We can express the HMM parameters as the followings.
Each HMM is modeled with a set \( \lambda = (\pi, A, B) \). Here, \( \pi = (\pi_1, \ldots, \pi_n) \) is the initial state probability; and \( N \) is the number of states represented by \( Q = \{1, 2, \ldots, N\} \). \( A = \{a_{ij}\} \) is the state transition probability matrix where \( a_{ij} \) is the probability of moving from state \( i \) to state \( j \). \( B = \{b_{jk}\} \) is the observation probability distribution matrix with \( b_{jk} \) is the emission of symbol \( k \) at state \( j \); and \( M \) is the number of symbols. Finally, \( V \) is the set of symbols (feature vectors), with \( V = \{v_1, v_2, \ldots, v_M\} \).

Given a HMM \( \lambda \), and the observation sequence, \( O = (O_1, O_2, \ldots, O_T) \), \( P(O | \lambda) \) is the probability that HMM \( \lambda \) produces the observation sequence \( O \). \( P(O | \lambda) \) can be computed by:

\[
P(O | \lambda) = \sum_{Q} P(O | Q, \lambda) P(Q | \lambda)
\]

(4)

Further details on HMM can be found in [22].

In our approach, we represent each shot by: (a) its tagged category; (b) scene/location change (c= change, u = unchanged); and, (c) speaker change (c = change, u = unchanged). We use the tag id as defined in Figure 3 to denote the category of each shot. Since the commercial category is not used here, so there are 12 categories. Each shot \( i \) is thus represented by a feature vector given by:

\[
S_i = (t_i, p_i, c_i)
\]

(5)

where \( t_i \) is the tag id of shot \( i \); \( p_i \) is the speaker change indicator (c or u); and \( c_i \) is the scene change indicator (c or u). Thus, each output symbol is represented by 1 of the 12 possible categories of shots, 1 out of 2 possible scene change feature, and 1 out of 2 possible speaker change feature. This gives a total of \( 12 \times 2 \times 2 = 48 \) distinct vectors to be modeled using the HMM framework.

In our preliminary experiments, we employ the ergodic HMM framework. We perform the experiments by varying the number of states from 4 to 9 to evaluate the results. As we have a small training data set, our initial experiments indicates that the number of state equals to 4 gives the best result. Figure 8 illustrates an ergodic HMM with 4 states used in our approach. When 4 states are used, we need to estimate \( b_{jk} \) for \( 4 \times 48 = 192 \) probabilities.

![Figure 8: The ergodic HMM with 4 hidden states](image-url)
6. TESTING AND RESULTS

This section discusses the experimental setups and results of the shot-level classification and scene/story boundary detection.

6.1 Training and Test Data

We use two days of news video (one from May 2001, the other from June 2001) obtained from the MediaCorp of Singapore to test the performance of our system. Each day of news video is half an hour in duration. One day is used for training, and the other for testing. We use the temporal multi-resolution analysis (TMRA) [2] to segment the input video sequence into shots. In order to eliminate indexing errors, we manually index the features of all the shots. After the removal of commercials, the training data set contains 200 shots and testing data set contains 183 shots. The numbers of story/scene boundaries are respectively 39 and 40 for the training and test data sets.

In Information Retrieval, there are several methods to measure the performance of the systems. The most popular methods are the precision (P), Recall (R), and F1 measure. Here we use F1 value and it can be derived by giving equal weights to precision and recall.

\[ F_1 = \frac{2RP}{R+P} \]  \hspace{1cm} (6)

\[ P = \frac{NC}{NC + FP} \]  \hspace{1cm} (7)

\[ R = \frac{NC}{NC + FN} \]  \hspace{1cm} (8)

where NC is the number of correct boundaries detected, FN is the number of boundaries missed or false negatives, and FP is the number of false positives (not a boundary but is detected as a boundary).

6.2 Shot Level Classification

6.2.1 Results of shot-level classification

The results of shot-level classification using the Decision Tree are presented in Table 1. The diagonal entries in Table 1 show the number of shots correctly classified into the respective category, while the off-diagonal entries show those wrongly classified. It can be seen that the largest classification error occurs in the Anchor category where a large number of shots is misclassified as Speech. This is because their contents are quite similar, and we probably need additional features, like background or speaker change, and the context of neighboring shots to differentiate them.
Table 1: The classification results from the Decision Tree

<table>
<thead>
<tr>
<th>Classified as:</th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th>e</th>
<th>f</th>
<th>g</th>
<th>h</th>
<th>i</th>
<th>j</th>
</tr>
</thead>
<tbody>
<tr>
<td>a) Intro/highlight</td>
<td>26</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>b) Anchor</td>
<td>16</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>c) 2Anchor</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>d) Meeting/gathering</td>
<td>13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>e) Still image</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>f) Live-reporting</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>82</td>
<td></td>
<td></td>
</tr>
<tr>
<td>g) Speech</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>11</td>
</tr>
<tr>
<td>h) Sport</td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>i) Text-scene</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>j) Special</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5</td>
<td></td>
</tr>
</tbody>
</table>

Overall, our initial results indicate that we could achieve a classification accuracy of over 95%.

One useful by-product of performing the Decision Tree analysis is the set of rules generated by the Decision Tree. These rules may be used to provide better insight into how certain decisions were made. Figure 9 shows the set of 14 rules extracted from our training data.

Figure 9: The set of rules extracted from the Decision Tree

Rule 1: shot type in [a, i], audio = speech, class => Live-reporting ** (0.684)
Rule 2: face = 1, shot type in [m, c], captions <= 1, audio = speech, motion = 1
   Class => Anchor (0.56)
Rule 3: audio = speech and noise, Class => Sport (0.686)
Rule 4: audio in {music and speech, noise}, Class => Intro/highlight (0.796)
Rule 5: face = 1, captions > 1, motion = 1, Class => Speech/Interview (0.59)
Rule 6: captions <= 1, audio = m, Class => Intro/highlight (0.894)
Rule 7: face >= 3, Class => Meeting/Gathering (0.807)
Rule 8: face = 2, shot type in [m, c], Captions > 1, Class => Anchor (0.581)
Rule 9: captions Position Center = yes, Class => Text-only (0.832)
Rule 10: captions > 1, centralized captions = no, audio = music
   Class => Special (0.773)
Rule 11: face = 1, motion in [l-m], shot duration in [m-h],
   Class => Speed/Interview (0.771)
Rule 12: face = 1, motion in [l-n], Class => Still image (0.763)
Rule 13: face in [>=2], shot type in [m, c], Captions <= 1
   Class => Meeting/Gathering (0.593)
Rule 14: face = 1, audio = speech, motion in [m-h], shot duration = 1
   Class => Live-reporting (0.708)
6.2.2 Effectiveness of the features selected

In order to ascertain the effectiveness of the set of features selected, we perform separate experiments by using different number of features. As face is found to be the most important feature, we use the face as the first feature to be given to the system. With the face feature alone, the system returns an accuracy of only 59.6%. If we include the audio feature, the accuracy increases rapidly to 78.2%. However, this accuracy is still far below the accuracy that we could achieve by using all the features. When we successively add in the rest of features in the order of shot type, motion, videotext, text centralization, and shot duration, the performance of the system improves steadily and eventually reaches an accuracy of 95.1%. The analysis indicates that all the features are essential in shot classification. Figure 10 shows a summary of the analysis.

![Figure 10: Summary of features analysis](image)

6.3 Scene/Story Boundary Detection

6.3.1 Results of scene/story boundary detection

We have performed three experiments (Ex I, II, and III) for scene/story boundary detection. As explained in Section 5, our initial experiments indicate that the number of states equals to 4 gives the best results. Thus, we set the number of states to 4 in these three HMM tests.

For Ex I, we assume that all the shots are correctly tagged. We perform the HMM to locate the story boundaries and we could achieve a F1 value of 93.7%. This experiment demonstrates that HMM is effective in news story boundary detection.
Ex II is similar to Ex I except that we perform the HMM analysis on the set of shots tagged using the earlier shot classification stage with about 5% tagging error. The test shows that we are able to achieve an F1 measure of 89.7%.

The results of both tests are detailed in Table 2.

**Table 2: The results of HMM analysis in tests I & II**

<table>
<thead>
<tr>
<th>Ex</th>
<th>NB</th>
<th>NC</th>
<th>FN</th>
<th>FP</th>
<th>R (%)</th>
<th>P (%)</th>
<th>F1 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>40</td>
<td>37</td>
<td>3</td>
<td>2</td>
<td>94.9</td>
<td>92.5</td>
<td>93.7</td>
</tr>
<tr>
<td>II</td>
<td>40</td>
<td>35</td>
<td>5</td>
<td>3</td>
<td>87.5</td>
<td>92.1</td>
<td>89.7</td>
</tr>
</tbody>
</table>

Note: NB = total number of correct boundaries

In Ex III, we want to verify whether it is necessary to perform the two-level analysis in order to achieve the desired level of performance. We perform HMM analysis on the set of shots with their original feature set but without the category information. We vary the number of features used from the full feature set to only a few essential features. The best result we could achieve is only 37.6% in F1 value. This test shows that although in theory a single stage analysis should perform the best, in practice, because of data sparseness, the 2-level analysis is superior.

6.3.2 Effectiveness of the features selected for HMM analysis

In order to evaluate the importance of each feature used in Ex II, we perform another set of experiments using only the individual feature one at a time, and by adding the second and third feature to the Tag-ID feature. Figure 11 (a) shows the precision and recall values of the results when using different combinations of features. Figure 11 (b) presents the corresponding F1 measures.

From Figure 11 (b), it can be seen that when using only the Tag-ID feature, the system could achieve an F1 measure of 86.4%. On the other hand, the use of the second and third feature alone returns low F1 measures of 41.7 and 33.3 respectively. However, by combining the last two features with the Tag-ID feature, the system’s F1 performance improves gradually from 86.4% (with Tag-ID as the only feature) to 88.9% (Tag-ID +Sp), and reaches 89.7% when all the three features are included (Tag-ID +Sp +Sc). The analysis indicates that the first feature (Tag-ID) is the most important feature for scene/story boundary detection. It further confirms that shot classification facilitates the detection of news boundaries, and therefore our two-level approach is effective.
6.3.3 Analysis of HMM results

Here we analyze how the HMM framework detects the story boundaries. Figure 12 lists two examples of the output state sequences resulting from the HMM analysis with 4 states. The figure indicates that State 4 signals the transition from one news topic to the next. Thus, by detecting State 4, we can locate the boundary of the current news topic.

Observation sequence 1: 1cc 1uu 1cu ... 1uu 2cc 4cc 4uu 6uu 6uu ...
Output states sequence 1: 3 3 3 ... 3 4 1 1 1 1 ...

Observation sequence 2: 6uc 6uu 6uu ... 6uu 10uu 10uu 10uu 10uu ...
Output states sequence 2: 1 1 1 ... 1 4 3 3 3 ...

Figure 12: Two examples of observation sequences and their corresponding output state sequences

In order to compare the distribution of the observation symbols for the training and test data sets, we further analyze the distribution of the observation symbols in each of the 4 states as shown in Figures 13(a) - 13(d). From the Figures, we can see that for each state, the distribution is dominated by a few (sometimes only one) symbols, and that the frequency distribution of the same symbol is similar in
both the training and test data sets. For example, in Figure 13(a), which shows the
distribution of State 1, the distribution is dominated by symbols 8 with the
frequency of about 50 out of 183 shots for both the training and test data sets. The
corresponding observation vector for this symbol is 6uu which represents a Live-
reporting shot. Figure 13(b) shows that the distribution of symbols in State 2 is
dominated by symbol 27. It corresponds to feature 8uu that represents a sport
shot. Figure 13(c) shows that the distribution of State 3 is dominated by symbols
2 and 4, which correspond to feature vectors 1uc and 1uu. Both vectors represent
Intro/highlight shots. Finally, the distribution of symbols in State 4 as shown in
Figure 13(d) is dominated by symbol 5. It corresponds to feature vector 2ce that
represents an Anchor person shot.

The analysis shows that HMM is effective in identifying dominant features
which help in locating the transition states. The analysis in Figures 13 suggests
that the Anchor-Person feature provides a strong clue to identify news story
boundaries.

![Histogram analysis of State 1](image1)

![Histogram analysis of State 2](image2)

![Histogram analysis of State 3](image3)

![Histogram analysis of State 4](image4)

Figure 13 (a)-13 (d): The distributions of the observed symbols of the 4 states

### 7. CONCLUSION AND FUTURE WORK

We have developed a two-level framework that can automatically segment an
input news video into story units. Given an input video stream, the system
performs the analysis at two levels. The first is shot classification, which
classifies the video shots into one of 13 pre-defined categories or mid-level
features using a combination of low-level, temporal and high-level features. The second level builds on the results of the first level and performs the HMM analysis on the mid-level features to locate story (or scene) boundaries. Our results demonstrate that our two-level framework is effective and we could achieve an accuracy of over 95% for shot classification, and over 89% in scene/story boundary detection. Our detailed analysis also indicates that HMM is effective in identifying dominant features that can be used to locate story transitions.

As our training data is rather sparse, our conclusion is preliminary. Although in theory one level analysis should yield better results, the two-level analysis that requires less training data has been found to be superior. This conclusion is reinforced in NLP research. Nevertheless, we need to do further tests by using a large set of training data, and by using news from different broadcast stations and countries. We will also incorporate speech to text data obtained from the audio track and use text segmentation technique to help identify story boundaries. We hope to fuse information from multiple sources in order to develop a robust and reliable story boundary detection model for news and other types of video.

Currently, we are working on incorporating question-answering techniques to associate video news stories to text news stories available on the web. The web-based approach not only provides supplementary information to enhance user insight to news video, it also provides the extracted information to: (a) expand users free text query; and (b) improve quality of video feature especially the speech to text data.

We are also willing on performing story summarization using a combination of text-based and utility-based techniques. Our eventual goal is to convert an input news video into a set of news stories together with their classification and summaries along with associated text-based news stories from web sites. This will bring us a major step towards supporting personalized news video for general users.

ACKNOWLEDGMENTS

The authors would like to acknowledge the support of the National Science & Technology Board and the Ministry of Education of Singapore for the provision of a research grant RP3960681 under which this research is carried out. The authors would also like to thank Rudy Setiono, Wee-Kheng Leow, and Gao Sheng for their comments and fruitful discussions on this research, and Chandrashekara Anantharamu for his help in programming techniques.
REFERENCES


