THE SEGMENTATION AND CLASSIFICATION OF STORY BOUNDARIES IN NEWS VIDEO

Lekha Chaisorn and Tat-Seng Chua
The School of Computing, National University of Singapore
Singapore 117543
{Lekhacha, chuats}@comp.nus.edu.sg

Abstract The segmentation and classification of news video into single-story semantic units is a challenging problem. This research proposes a two-level, multi-modal framework to tackle this problem. 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 a Decision Tree to classify the shot into one of 13 pre-defined categories. At the scene level, we perform the HMM (Hidden Markov Models) analysis to eliminate shot classification errors and to locate story boundaries. We test the performance of our system using two days of news video obtained from the MediaCorp of Singapore. Our initial results indicate that we could achieve a high accuracy of over 95 % for shot classification. The use of HMM analysis helps to improve the accuracy of the shot classification and achieve over 89% accuracy on story segmentation.

Keywords: News Video classification; multi-modal approach; learning-based approach; Story Segmentation

1. INTRODUCTION

The effective management of the ever-increasing amount of broadcast news video is essential to support a variety of user-oriented functions, including the browsing, retrieval and personalization of news video. One effective way to organize the video is to segment the video into small, single-story units and classify these units according to their semantics. Research on segmenting an input video into shots is well established [Zhang et al 1993, Lin et al 2000]. A shot represents a contiguous sequence of visually similar frames. It is a syntactical representation and does not usually convey any coherent semantics to the users. In order to overcome this problem, recent works have been done to group related sequence of shots into scenes [Chang & Sundaram 2000, Wang et al. 2001]. The other challenge is to classify the shots and scenes created into well-defined categories.

X. Zhou et al. (eds.), *Visual and Multimedia Information Management*

Our research aims to develop a system to automatically segment and classify news video into semantic units. We propose a two-level, multi-modal framework to tackle this problem. We adopt the domain of news video in this study because news video is more structured and has clearly defined story units. 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 then employ a Decision Tree to classify the video shots into one of the 13 pre-defined categories. The result of shot level analysis is a set of shots tagged with one of the predefined categories. As the classification is performed independently at shot level, errors and ambiguity in shot tagging will occur. To overcome this problem, we perform HMM analysis [Rabiner & Juang 1993] at the scene level in order to eliminate the classification errors and to identify news story boundaries. Our approach is similar to that employed in natural language processing (NLP) research in performing part-of-speech tagging at the word level, and higher level analysis at the sentence level [Dale 2000].

Briefly, the content of this paper is organized as follows. *Section 2* describes related research 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* discusses the experiment results. *Section 7* contains our conclusion and discussions of future works.

2. RELATED WORKS

Video classification is a hot topic of research for many years and many 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 the sports or news.

Ide et al. [1998] tackled the problem of news video classification and used videotext, motion and face as the features. They first segmented the video into shots and used clustering techniques to classify each shot into one of the five classes of: Speech/report, Anchor, Walking, Gathering, and Computer graphics shots. Their classification technique is quite simple and seems effective for this restricted class of problems. Zhou et al. [2000] examined the classification of basketball video into the 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 [2001] also used a rule-based approach to classify news video into the 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 methods incorporated information within and between video segments to determine class transition boundaries using mostly the HMM approach. Eickeler et al. [1997] 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. [1999] 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. [2001] 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 changed, and used HMM to locate the transition boundaries 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 aim to consider all possible categories of shots and scenes to cover all types of news video. Another major difference between our approach and existing works is that we plan to perform the story segmentation analysis at two levels, similar to that employed successfully in NLP.

3. THE MULTI-MODAL TWO-LEVEL FRAMEWORK

3.1 Structure of News

Most news videos have rather similar and well-defined structures. Figure 1 illustrates the structure of a typical news video. The news video typically begins with several Intro/Highlight shots that give a brief introduction of the up coming news to be reported. The main body of news contains a series of stories organized in term 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 and ends with Anchor-person shots and several in between Live-reporting 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.

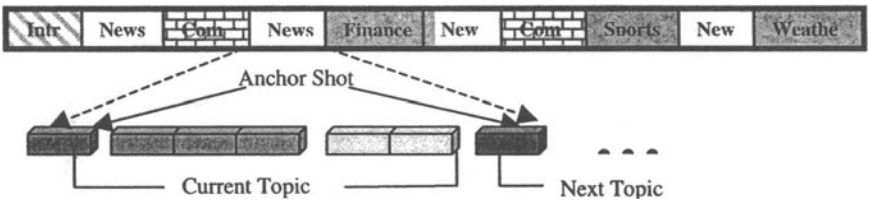

Figure 1: The structure of local news video under study

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 particular identifying the 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:

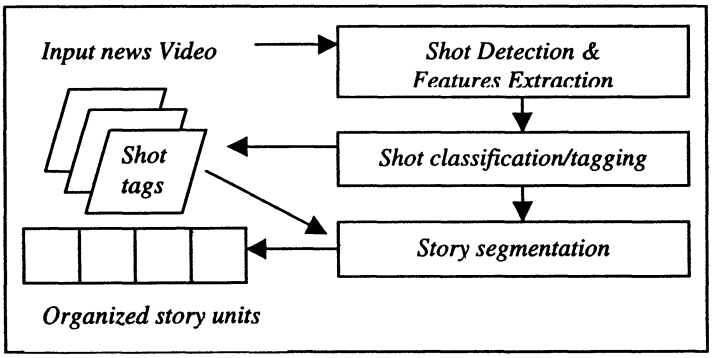

Figure 2: Overall system component

a) We first divide the input video into shots using a mature technique.
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 subcategories.

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.

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. We employ the multi-resolution analysis technique developed in our lab [Lin et al. 2000] that can effectively locate both abrupt and gradual transition boundaries effectively.

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; 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. We select the I-frame that is nearest to the center of the shot as the key frame.

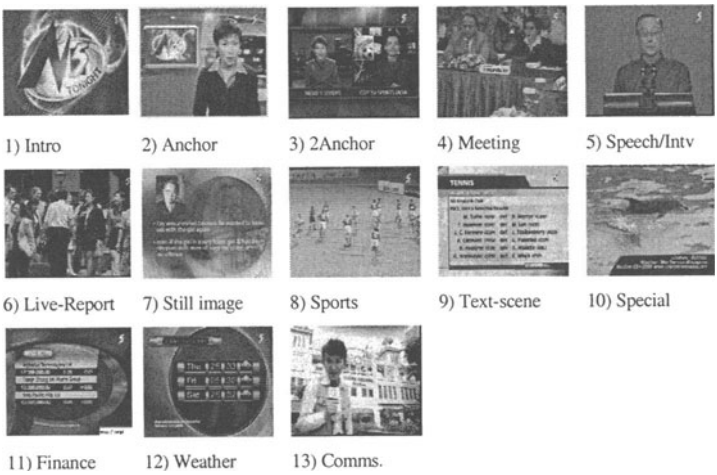

Figure 3: Examples of the predefined categories and example shots

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: *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. Figure 3 shows a typical example in each category.

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.

4.3(a) 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 reporting. 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(b) Temporal Features

Background scene change: Following the discussions on color histogram, we include 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.

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, 'c' if there is a change, and 'n' if this feature is not applicable to the shots that do not contain speech.

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. [2001] to classify audio into the broad categories of speech, music, noise, speech and noise, speech and music, or silence

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 those shots with people walking), *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(c) High-level Object-based features

Face: Human activities are one of 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. [2000] 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. The size of the face is used to estimate the shot types.

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

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 [2000] to detect videotexts. For each shot, we simply determine the number of lines of text appear in the key frame.

Centralized Videotext: We often need to differentiate between two types of shots containing videotexts. The normal shot where the videotexts appear at the top or bottom of a shot to indicate its contents. 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:

$$S_i = (a, m, d, f, s, t, c) \quad (1)$$

where

- a the class of audio, $a \in$ {t=speech, m=music, s=silence, n =noise, tn = speech + noise, tm= speech + music, mn=music+noise}
- m the motion activity, $m \in$ {l=low, m=medium, h=high}
- d the shot duration, $d \in$ {s=short, m=medium, l=long}
- f the number of faces, $f \in$ N
- s the shot type, $s \in$ {c= closed-up, m=medium, l=long, u=unknown}
- t the number of lines of text in the scene, $t \in$ N
- c set to "true" if the videotexts present are centralized, c $\in$ {t=true, f=false}

For example, the feature vector of an Anchor-person shot may be (t, l, l, 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

We first remove the commercials before performing the classification of the remaining shots using a learning-based approach.

In most countries, it is mandatory to air several black frames preceding or proceeding a block of commercials. However, this is not always the case in many countries, like in Singapore. Our studies have shown that commercial boundaries can normally be characterized by the presence of black frames, still frames and/or audio silence [Koh and Chua 2000]. We thus employ a heuristic approach to identify the presence of commercials and detect the beginning and ending of the commercials blocks. 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%.

We break the classification of remaining shots into two sub-tasks. 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 [1998] 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.

For the rest of the shots, we employ a decision tree to perform the classification in 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 [Quinlan 1986]. The decision tree approach has been successfully employed in many multi-class classification problems [Dietterich & Bakiri 1995, and Zhou et al. 2000]. We thus select the Decision Tree for our shot classification problem.

5. STORY/SCENE SEGMENTAION

After the shots have been classified into one of the pre-defined categories, we employ HMM to model the context of shot sequences in order to correct the shot classification errors and identify story boundaries. We use the shot sequencing information, and examine both the tagged category and appropriate features of the shots to perform the analysis. This is similar to the part-of-speech (POS) tagging problem in NLP that uses a combination of POS tags and lexical information to perform the analysis.

We model each shot by: (a) its tagged category; (b) scene/location change (c= change, u = unchanged); and, (c) speaker change (c = change, u = unchanged, and

n = not applicable). We use the tag id as defined in Figure 3 to denote the category of each shot. For example, an input shot represented by **(1 c c)** means that this is an Intro/highlight shot with changes in background/location and speaker from the previous shot. Example of an input sequence passed to HMM is illustrated in Figure 4.

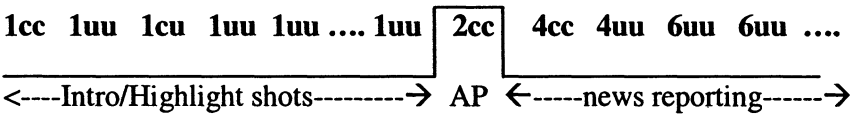

Figure 4: Example of an input sequence (AP=Anchor-person shot)

HMM is a power statistical tool first successfully utilized in speech recognition research. HMM contains a finite set of *states*, each of which is associated with a (generally multidimensional) 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. Video is a time sequencing data similar to that of speech, we thus employ HMM to perform the news video story segmentation.

In our priliminary experiments, we simply employ a left-to-right HMM with 4 to 9 internal states to model the news video.

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. In this study, in order to eliminate indexing errors, we manually index all the features of the shots segmented using the multi-resolution analysis algorithm [Lin et al. 2000]. After the removal of commercials, the training data set contains 200 shots and testing data set contains 183 shots. The numbers of stories/scene boundaries are respectively 39 and 38 for the training and test data sets.

6.2 Shot level classification

6.2(a) Results of shot-level classification using Decision Tree

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

Table 1: The classification results from the Decision Tree

Classified as->	a	b	c	d	e	f	g	h	i	j	k	l
a) Intro/highlight	26					1						
b) Anchor		16					4					
c) 2anchor			2									
d) Gathering				13								
e) Still image					1							
f) Live-reporting						82				1		
g) Speech		1					11					
h) Finance								*				
i) Weather									*			
j) Sport						1				8		
k) Text-scene											6	
l) Special												5

Figure 5 presents the partial Decision Tree learned from the training data. As can be seen from the tree generated, face is the most important feature for shot-level classification followed by audio and shot type.

6.2(b) 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 the accuracy of 95.10%. The results of the feature analysis are summarized in Figure 6. The analysis indicates that all the features are essential in shot classification.

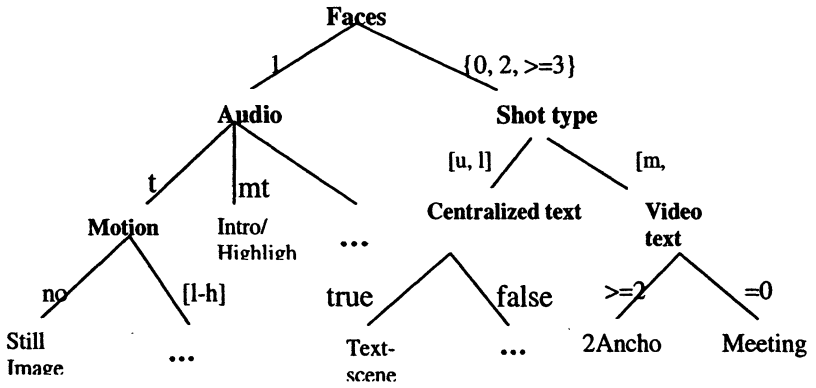

Figure 5: Part of the decision tree created from the training sample

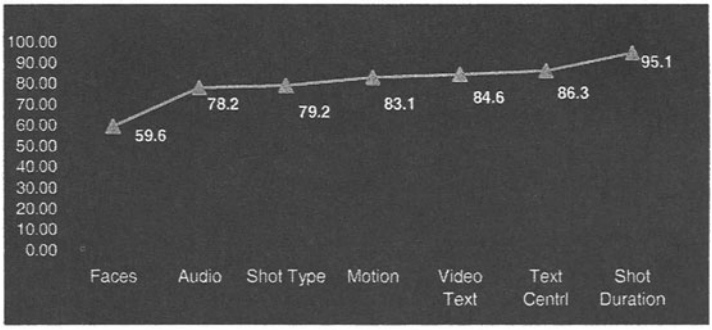

Figure 6: The results of features base-line analysis

6.3 Scene/News Story Segmentation

As stated in Section 5, we employ the left-to-right HMM to model news video sequences. We performed several experiments by setting the number of states ranging from 4 to 9. Our experiments showed that the number of states equals to 4 gives the best results.

In the first test, we assume that all the shots are correctly tagged and perform the HMM to locate story boundaries. The results indicate that out of the 38 correct story boundaries in the test set, we could achieve about 92% accuracy with two missing and one wrongly identified story boundary. This experiment demonstrates that HMM is effective in news story boundary detection.

Next, we perform HMM analysis on the set of shots tagged using the earlier shot classification stage with about 5% tagging error. The analysis aims to correct the shot tagging errors and locate story boundaries. The results demonstrate that: (a) we are able to correct 5 wrongly tagged shots to push the shot classification

accuracy to over 97%; and (b) we are able to achieve over 89% accuracy on story boundary detection.

The results show conclusively that our two-level framework is effective in detecting and classifying story boundaries in news video.

7. CONCLUSIONS AND FUTURE WORKS

We have developed a multi-modal two-level framework that can automatically segment an input news video into story units. We use a set of low-level and high-level features to perform the analysis. 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 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 to locate story (or scene) boundaries and classify news stories. Our results demonstrate that our two-level framework is effective and we could achieve an accuracy of over 89% on scene or story boundary detection,

We are now in the process of incorporating speech to text feature to enhance the system performance. Our eventual goal is to convert an input news video into a set of news stories together with their classification. 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 Chin-Hui Lee, Rudy Setiono and Wee-Kheng Leow for their comments and fruitful discussions on this research.

REFERENCES:

A. Aydin Alatan, Alin N. Akansu, and Wayne Wolf (2001). "Multi-modal Dialog Scene Detection using Hidden Markov Models for Content-based Multi-media Indexing". *Multimedia Tools and applications*, 14, pp 137-151.

Shiu-Fu Chang and Hari Sundaram (2000). " Structural and semantic analysis of video", IEEE International Conference on Multimedia and Expo (II): pp. 687-

Y. Chen and E. K. Wong (2001). " A knowledge-based Approach to Video Content Classification", Proceeding of SPIE Vol. 4315, pp.292-300.

Tat-Seng Chua and Chunxin Chu (1998). Color-based Pseudo-object for image retrieval with relevance feedback. International Conference on Advanced Multimedia Content Processing '98. Osaka, Japan, Nov. 148-162.

Tat-Seng Chua, Yunlong Zhao and Mohan S. Kankanhalli (2000). "An Automated Compressed-Domain Face Detection Method for Video Stratification", Proceedings of Multimedia Modeling (MMM'2000), USA, Nov, World Scientific, pp 333-347.

Robert Dale, Hermann Moisl, and Harold Somers (2000). "Handbook of natural language processing", Imprint New York: Marcel Dekker.

T. G. Dietterich, and G. Bakiri (1995). "Solving Multi-class Learning Problems via Error-Correcting Output Codes", Journal of Artificial Intelligence Research, pp 263-286

Stefan Eickeler, Andreas Kosmala, Gerhard Rigoll (1997). "A New Approach To Content-based Video Indexing Using Hidden Markov Models", IEEE workshop on Image Analysis for Multimedia Interactive Service (WIAMIS), pp 149-154.

J. Huang, Z. Liu, Y. Wang (1999). "Integration of Multimodal Features for Video Scene Classification Based on HMM", IEEE signal processing Society workshop on Multimedia Signal processing, Denmark, pp 53-58.

Ichiro Ide, Koji Yamamoto, and Hidehiko Tanaka (1998). "Automatic Video Indexing Based on Shot Classification", Conference on Advanced Multimedia Content Processing (AMCP'98), Osaka, Japan. S. Nishio, F. Kishino (eds), Lecture Notes in Computer Science Vol.1554, pp 87-102.

Michael I. Jordan (1998) (Eds). "Learning in Graphical Models", MIT Press.

Chun-Keat Koh and Tat-Seng Chua (2000). "Detection and Segmentation of Commercials in News Video", Technical report, The School of computing, National University of Singapore.

Yi Lin, Mohan S Kanhanhalli, and Tat-Seng Chua (2000). "Temporal Multi-resolution Analysis for Video Segmentationtion", Proceedings of SPIE (Storage and Retrieval for Media Databases)., San Jose, USA. Jan 2000, Vol 3972, pp 494-505.

Zhu Liu , Jingcheng Huang, and Yao Wang (1998). "Classification of TV Programs Based on Audio Information using Hidden Markov Models", IEEE Signal Processing Society, Workshop on Multimedia Signal Processing, Los Angeles, California, USA, pp 27-31.

Lie Lu, Stan Z. Li and Hong-Jiang Zhang (2001). "Content-based Audio Segmentation using Support Vector Machine", IEEE International Conference on Multimedia and Expo (ICME 2001), Japan, pp 956-959.

J. R. Quinlan (1986). "Induction of Decision Trees. Machine Learning" vol. 1, pp. 81-106.

L. Rabiner and B. Juang (1993). "Fundamentals of Speech Recognition", Prentice-Hall.

Jihua Wang, Tat-Seng Chua, and Liping Chen (2001). "Cinematic-based Model for Scene boundary detection", to appear in Proc. of Multimedia Modeling conference (MMM'01), Amsterdam, Netherlands.

Hong-Jiang Zhang, A. Kankanhalli and S.W. Smoliar (1993). "Automatic Partitioning of Full-motion Video", *Multimedia Systems*, 1(1), pp 10-28.

Yi Zhang and Tat-Seng Chua (2000). "Detection of Text Captions in Compressed domain Video". Proceedings of ACM Multimedia'2000 Workshops (Multimedia Information Retrieval), California, USA. Nov, pp 201-204.

WenSheng Zhou, Asha Vellaikal, and C–C Jay Kuo (2000). "Rule-based Classification System for basketball video indexing", Proceedings of ACM Multimedia'2000 Workshops (Multimedia Information Retrieval), California, USA. Nov, pp 213-216.

BIOGRAPHIES

Lekha Charson received her B.Sc. in Mathematics in 1982 from the Prince of Songkhla University, Thailand, and her MSc in Information and Computer Science in 1996 from the National University of Singapore. Presently, she is a PhD. candidate at the School of Computing,, National University of Singapore. She has worked as a senior lecturer at the Computer Science Department,

in the Prince of Songkhla University, Thailand, since 1989. Her current research interests include news video segmentation and classification, and speech and audio classification.

Tat-Seng Chua obtained his PhD from the University of Leeds, UK. He joined the School of Computing, National University of Singapore, in 1983. He was the Acting and Founding Dean of the School of Computing from 1998-2000. He spent three years as a research staff member at the Institute of Systems Science (now KRDL) in late 1980s. Dr Chua's main research interest is in multimedia information processing, in particular, on video and text retrieval and information extraction. Dr Chua has organized and served as program committee member of numerous international conferences in the areas of computer graphics and multimedia. He is the conference co-chair of: Multi-Media Modeling (MMM) '93, '95, '97 and '03; Computer Graphics International (CGI) '90, '00 and '01; and Pacific Graphic '98. He serves in the editorial boards of: IEEE Transactions of Multimedia (IEEE); The Visual Computer (Springer-Verlag); and Multimedia Tools and Applications (Kluwer).