THE SEGMENTATION OF NEWS VIDEO INTO STORY UNITS

Lekha Chaisorn *Tat-Seng Chua* *Chin-Hui Lee*

School of Computing, National University of Singapore, Singapore 117543
{lekhacha, chuats, chl}@comp.nus.edu.sg

ABSTARCT

The segmentation 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 locate the 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, and over 89% in F_1 measure on scene/story boundary detection.

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 video is to segment it into small, single-story units and classify these units according to their semantics. Research on segmenting an input video into shots is well established. 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 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 [1] such as meeting, sunset, etc, and organized hierarchically to facilitate browsing.

In order to alleviate the high dimensionality and the corresponding data sparseness problems, this paper proposes a two-level, multi-modal framework to segment an input news video into 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, and employ a Decision Tree [2] to classify the video shots into one of the pre-defined categories. The result of shot level analysis is a set of shots tagged with one of the predefined categories. We next perform the HMM [3] analysis at the scene level 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, followed by higher-level analysis at the sentence level [4]. This paper discusses the design of our approach together with experimental results.

2. RELATED WORKS

Current research on video scene segmentation can be divided into two broad categories. The first category considers mostly visual and temporal similarity as basis to group sequences of similar shots into scene [5][6][7][8]. These methods are able to capture scenes that take place in same time and space or are modeled using simple cinematic rules. However, they are not able to capture most scenes formed on the basis of semantic similarity.

The second category attempts to segment or classify video scenes based on semantics. Because of the difficulty and often subjective nature of this approach, most early works focused on structured domains such as sports or news. Most reported works considered a small set of features and focused on a few selected classes in news such as Speech/report, Anchor, Walking, Gathering, and Computer graphics shots [9] or as news, weather reporting, commercials, basketball, and football [10][11]. These methods provide only partial, intermediate solution to general news video organization problem.

The main contributions of this paper are two-fold. First, it proposes a complete solution for news videos organization. It considers all possible categories of shots and scenes to cover all types of news video. Second, in order to overcome the data sparseness problem, it performs the story segmentation analysis at two levels, similar to that employed successfully in NLP. Existing applications circumvent this problem by considering only a subset of feature and classes.

3. THE MULTI-MODAL TWO-LEVEL FRAMEWORK

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 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 the 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. Within a station, the visual contents of each news category, like the anchor -person shots, finance and weather reporting etc., tends to be highly similar. Hence, it is possible to adopt a learning-

0-7803-7304-9/02/$17.00 C2002 IEEE

based approach to train a system to recognize and segment the contents of each category within each broadcast station.

The classification of news video contents is difficult because there are many categories that are highly similar and can only be differentiated by using an appropriate combination of features. The segmentation of news video into story units is even more difficult as it requires both visual and semantic information. To tackle the problem effectively, we adopt the strategies successfully employed in NLP in partitioning the problem into two simpler sub-problems of performing the classification followed by segmentation. The main stages as shown in Figure 1 are as follows. (a) We perform shot segmentation using mature technique and extract a suitable set of features to model the contents of each shot; (b) We employ a learning-based approach that uses multi-modal features to classify the shots into the set of well-defined categories. (c) We use a combination of shot content features, categories, and temporal features to identify story boundaries using the HMM technique.

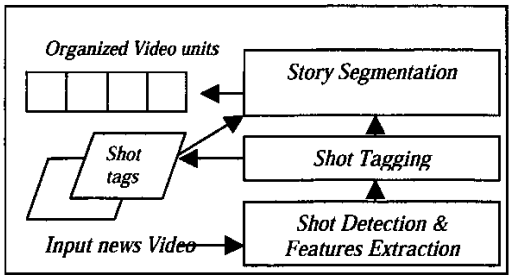

Figure 1: Overall system components

4. THE CLASSIFICATION OF VIDEO SHOTS

The first step is to divide an input news video into shots. Here, we employ the multi-resolution analysis technique [12] that can effectively locate both abrupt and gradual transition boundaries. After the video is segmented, there are several ways in which the contents of each shot can be modeled. In our work, we adopt the approach that use both key frame and feature trajectories 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.

The second step is to determine an appropriate and complete set of categories to cover all the shot types. The categories must be meaningful so that the category tag assigned to each shot reflects its contents 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. 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.

The third step is to determine a suitable set of features to facilitate shot classification. 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. The features selected can be divided into three groups.

(a) Low-level Visual Content Feature: The main feature in this group is the 256-Color histogram, which can be used to: (i) identify those shot types with similar visual contents, such as the weather and finance reporting; and (ii) measure the changes in background between successive shots, which provides important clues to determining a possible change in shot category or story.

(b) Temporal Features: The main features under this group are the Background Scene Change; Speaker Change; Audio Type; Motion Activity; and, Shot Duration. The Background Scene Change and Speaker Change are used respectively to measure whether there is a change of background or speaker between the current and previous shots. Audio is very effective in identifying categories that have noisy background (such as Sports and Live-reporting) or with accompanying background music (such as Intro/Highlight shots). Here, we adopt an algorithm similar to that discussed in [13] to classify audio into the broad categories *of speech, music, noise, speech and noise, speech and music, or silence.* We classify Motion Activity into *low, medium, high,* or *no* motion in order to identify shots with a lot of motion (such as Sports) or those with little motion (such as the Anchor-person shots). Lastly, we use Shot Duration to differentiate shots with short (such as Sports), medium, or long duration (such as Anchor-person shots).

(c) High-level Object-based features: This group includes features such as Human Faces or Videotext, which are essential to model the semantic contents of shots. For faces, we adopt the algorithm developed in [14] to detect the number of mostly frontal faces and their sizes in the key frame. For Videotext, we employ the algorithm developed in [15] to detect the number of lines of text appearing in the key frame. From the size of faces, we estimate shot type as closed-up, medium, or long. From the Videotext feature, we identify whether the text is centralized or not. In general, a text-scene shot (such as the scene summarizing the results of a soccer game) typically contains multiple lines of centralized text, which is different from normal shots.

In step four, we represent the contents of each shot using a color histogram vector and a feature vector. The feature vector of a shot is of the form:

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

where "*a*" is the type of audio, "*m*" *is* the motion activity level, "*d*" *is* the shot duration level, "*f*" *is* the number of faces, "*s*" *is* the shot type, "*t*" *is* the number of lines of text in the scene, and "*c*" *is* set to true if the videotext is centralized. Note that at this stage we do not include the scene change and speaker change features in the feature vector as they are not important for shot classification. They will be included in story boundary detection using HMM.

Finally, we perform the classification of video shots in three stages as follows. We first remove the commercials using a heuristic approach developed in our lab that incorporates the presence of black frames, still frames, cut rate, and/or audio silence. Our studies showed that we are able to achieve a high detection accuracy of over 97%. Next, we identify the Weather and Finance shots using the histogram-matching algorithm developed in our lab to compute the shot-category

similarity that takes into consideration of perceptually similar colors. Third, we employ a Decision Tree (DT) to perform the classification of the rest of the shots using a learning-based approach.

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 locate the story boundaries. 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 (further details of HMM can be found in [3]). We use the shot sequencing information, tagged category and appropriate features of the shots to perform the analysis. We model each shot by: (a) its tagged category represented by tag id (#1-12, see Table 1); (b) scene/location change (c for change and u for unchanged); and, (c) speaker change (c or u). Thus, each feature vector of a shot is represented by 1 of the 12 possible categories of shots (commercial category is not included), 1 out of 2 possible scene change feature, and 1 out of 2 possible speaker change feature. This gives a total of 12x2x2 = 48 distinct vectors (symbols) for modeling using the HMM framework. In our preliminary study, we employ an ergodic HMM to perform the analysis. Our initial test indicates that the number of state equals to 4 gives the best result. Figure 2 illustrates an ergodic HMM with 4 hidden states.

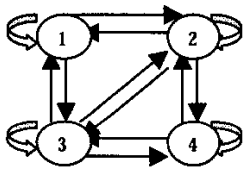

Figure 2: The ergodic HMM with 4 hidden states

When 4 states are used, we need to estimate 4x48 = 192 probabilities for emission of symbol k at state j. An example of an observation sequence of input feature vectors and its corresponding output state sequence is presented in Figure 3.

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

Figure 3: Example of an observation sequence and its corresponding output state sequence

6. TESTING AND RESULTS

We use two days of news 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 segmented shots. After the removal of commercials, the training and the test data sets contain 200 and 183 shots respectively. The numbers of story/scene boundaries are respectively 39 and 40 for the training and test data sets.

6.1. Shot-level classification results

The results from the DT classification are presented in Table 1. The diagonal entries in the Table show the number of shots correctly classified into the respective category, while the off-diagonal entries show those wrongly classified. The overall results show that we could achieve classification accuracy of over 95%. The analysis of DT results also indicates that face and audio are the most important features for shot-level classification.

Classified as->	1	2	3	4	5	6	7	8	9	10
1) Intro/highlight	26					1				
2) Anchor		16					4			
3) 2anchor			2							
4) Gathering				13						
5) Still image					1					
6) Live-reporting						82		1		
7) Speech		1					11			
8) Sport						1		8		
9) Text-scene									6	
10) Special										5

Note: 11) Weather and 12) Finance are not included here

Table1: The classification result from the DT

In order to ascertain the effectiveness of the selected features, we perform a separate experiment by successively adding in features based on their order of importance derived from the DT analysis. The analysis indicates that all the features are essential in order to achieve the desired performance.

6.2. Scene/News Story Segmentation Results and Evaluation

One method to measure the performance of the Information Retrieval systems is to use precision (P) and recall (R) values. The formulas are expressed as below:

$$P = NC/(NC+FP) \quad (2)$$
$$R = NC/(NC+FN) \quad (3)$$

Here, NC is the number of correct boundaries detected, FN is the number of false negative (missed), and FP is the number of false positive (not a boundary but is detected as a boundary). By giving equal weights to precision and recall, we can derive an F_1 value to measure the overall system performance as:

$$F_1 = 2*R*P/(R+P) \quad (4)$$

We set up three experiments (Tests I, II, and III) for scene/story boundary detection. As described in Section 5, our experiments indicate that the number of states equals to 4 gives the best result, thus we set the number of states to 4 in these three HMM tests.

Test	NB	NC	FN	FP	R (%)	P (%)	F_1 (%)
I	40	37	3	2	94.9	92.5	93.7
II	40	35	5	3	***87.5***	***92.1***	***89.7***

Note: NB = total number of correct boundaries

Table 2: The results from HMM analysis

For **Test I**, we assume that all the shots are correctly tagged. We perform the HMM to locate the story boundaries and we could achieve a F_1 value of 93.7%. This experiment demonstrates that HMM is effective in news story boundary detection. **Test II** is similar to **Test 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 F_1 measure of 89.7%. The results of both tests are detailed in Table 2.

In **Test 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 F_1 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.

In order to evaluate the importance of each feature used in **Test 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. From the experimental results, by using only the Tag-ID feature, the system could achieve an F_1 measure of 86.4%. On the other hand, the use of the second and third feature alone return low F_1 measures of 41.7 and 33.3 respectively. However, by combining the last two features with the Tag-ID feature, the system's F_1 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.

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. 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) boundary. Our results demonstrate that our two-level framework is effective and we could achieve an accuracy of over 95% for shot classification, and of over 89% in scene/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.

8. REFERENCES

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

[2] J. R. Quinlan, "Induction of Decision Trees "Machine Learning", vol. 1, pp. 81-106, 1986.

[3] L. Rabiner and B. Juang, "Fundamentals of Speech Recognition", Prentice-Hall, 1993.

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

[5] John R Kender and Boon-Lock Yeo, "Video Scene Segmentation Via Continuous Coherence", 1998 Conference on Computer Vision and Pattern recognition, Santa Babara.

[6] Tong Lin and Hong-Jiang Zhang, "Automatic Scene Extraction by Shot Grouping", 15th internaional Conference on Pattern Recognition, Barcelona, Spain, Sep 2-8, 2000.

[7] Silvia Pfeiffer, Rainer Lienhart, and Wolfgang Effelsberg. " Scene determination Based on Video and Audio Features", Proceeding of the IEEE conference on Multimedia Computing and System, Volume I, pp. 59-81, 1998,

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

[9] Ichiro Ide, Koji Yamamoto, and Hidehiko Tanaka "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, 1998.

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

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

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

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

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

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