# STORY BOUNDARY DETECTION IN NEWS VIDEO USING GLOBAL RULE INDUCTION TECHNIQUE

*[1]Lekha Chaisorn and [2]Tat-Seng Chua*

[1]Media Division,
Institute for Infocomm Research ($I^2R$), Singapore 119613
clekha@i2r.a-star.edu.sg
[2] School of Computing, National University of Singapore,
Singapore 117543
chuats@comp.nus.edu.sg

## ABSTRACT

Global rule induction technique has been successfully used in information extraction (IE) from text documents. In this paper, we employ the technique to identify story boundaries in news video. We divide our framework into two levels: shot and story levels. We use a hybrid algorithm to classify each input video shot into one of the predefined genre types and employ the global rule induction technique to extract story boundaries from the sequence of classified shots. We evaluate our rule induction based system on ~120-hours of news video provided by TRECVID 2003. The results show that we could achieve an $F_1$ accuracy of over 75%.

## 1. INTRODUCTION

A two-level multi-modal framework for story segmentation in news video based on Hidden Markov Models (HMM) was proposed in Chaisorn et al. in [5]. The framework composes of two levels: shot and story levels. It employed a hybrid approach and the HMM for the analysis at shot and story levels respectively. The system was evaluated on the TRECVID 2003 data set [3] and it achieved the best performance in the evaluations under story segmentation task [3,6]. But the disadvantages of the system are that it is computational expensive and not easy to scale up to larger and new corpuses. As the amount of news video is rapidly increasing, employing the HMM framework for the analysis will greatly affect the system performance.

Though several methods on story segmentation in news video have been proposed, their systems have some limitations. For examples, works in [1,2] employed heuristic approach, and their systems yielded good results on their test data. However, their methods rely heavily on features derived from news transcripts, and their performance will be severely affected if news transcripts are not available. Recent work reported in TRECVID workshops [3,4] used multi-modal features and employed mostly machine-learning based approaches. However, frameworks based on machine-learning approaches often encounter data sparseness problem due to insufficient amount of training data. In order to alleviate the data sparseness problem, one way is to divide the framework into multi-layered tasks such as that employed in NLP research [9] that analyzes text documents at word, phrase and sentence levels. This multi-layered approach was successfully adopted in [5,6,7].

In this research, we propose the use of a global rule induction technique to tackle the story segmentation problem. The technique is based on GRID (Global Rule Induction for Text Documents) [10], which we extended to induce rules to extract story boundary information from video sequences based on a two-level framework. The reasons for adopting the global rule induction technique are: (1) we observed that when employing HMM for the analysis at the story level, there are embedded pattern rules in the output; (2) we want to offer an effective system that requires less computational cost and complexity to cope with the dynamic real world problems in news story segmentation and indexing; and (3) we want to demonstrate the generality and quality of the proposed two-level framework [5].

## 2. THE SYSTEM FRAMEWORK

Briefly, the system composes of the shot and story levels as shown in Figure 1. A shot is a continuous sequence of frames taken in a continuous camera shooting, usually takes place in a physical location. A news story is defined as a segment of news broadcast with a coherent focus, which contains at least two independent, declarative clauses [3]. In our system, at the lower shot level, our shot tagger assigns each of the input shots a unique tag_ID. Each tag_ID represents one of the predefined shot categories. At the

higher story level, the story extractor extracts pattern rules (that constitute story boundaries) from a sequence of the tagged shots.

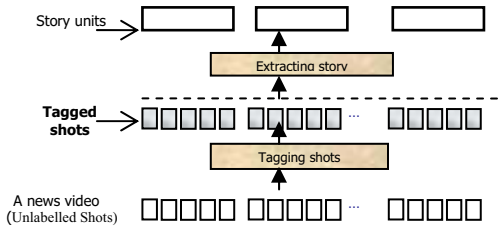


**Figure 1: Overview of the System Framework**

The following sections discuss the details of the system framework.

### 2.1 Selection of Shot Categories and Features

Most news video has similar structure. It covers news reports on regional and world news, weather, business, sports, commercials, etc. Thus, to reflect this structure and semantics of the shots' contents, we select the following set of shot categories: *Intro/highlight, LEDS, Anchor, 2Anchor, People, Speech/Interview, Live-reporting, Sports, Text-scene, Commercials, Weather, Finance, Special, Program logos* (such as *top s*tory, *health* and *sports, etc.* found in TRECVID data set).

| Type | Features; Description or format |
|---|---|
| Low-level | ***Color Histogram*;** 176 Luv color histogram |
| Temporal-level | **-Scene change**;** 1-changed, 0 -unchanged<br>***-Audio class;*** Speech, music, noise, speech+music, peech+noise, and silence;<br>***-Motion activity;*** Low, medium, high, or no motion;<br>***-Shot duration;*** Short, medium, long |
| Mid-level | ***-Face;*** number of faces<br>***-Shot type;*** Closed-up, medium, long, unknown;<br>***-Videotexts;*** Lines of text & 1 - center; 0 - /non-centered;<br>**-Cue phrases**;** 1 – presence; 0 - not presence; |

Note: Scene change and cue phrase will only be used in story level analysis

**Table 1: A set of features selected for the analysis.**

To facilitate the classification of shots into the categories, we first select a set of prominent features as shown in table 1. These are the features that are useful to discriminate one shot category from the others. For example, audio class (cheering noise) and motion activity are dominant features for "sports" category, background color and face features are dominant features for "anchor" and "2anchor" categories, etc.

After the features are selected, we next group the categories into three clusters based on their characteristics. These clusters are: *commercial/heuristic* (CM); *visual similarity-based* (VS)*; and machine-learning-based* (ML). The first cluster contains all commercial shots. The *visual similarity-based* cluster is composed of those categories where shots that are visually similar either within a broadcast station or within the same news video. The last cluster, ML cluster, contains the remaining shot categories that require more sophisticated machine learning techniques for their classification. The selected categories were demonstrated in [6,7] to be useful for story segmentation process. A subset of these categories was also successfully utilized in [8].

The analysis and more details of the selected categories and features can be found in [6,7].

### 2.2 The Tagging of Shots into Pre-defined Categories

Based on these three clusters, we employ a hybrid approach to perform the shot classification. Briefly, we use heuristic approach to perform commercial detection based on blank frames, audio silence, high cut rate and low confident value in ASR result. Next, we employ image similarity and clustering algorithm to identify shots of *visual similarity-based* type based on face and color histogram features. Finally, we employ a decision tree to classify the remaining shots in ML cluster using the features in Table 1. The accuracy of our shot classification algorithm tested on a subset of TRECVID training set [3] is about 90% [6, 7].

## 3. STORY BOUNDARY EXTRACTION USING GLOBAL RULE INDUCTION TECHNIQUE

### 3.1 Introduction to Global Rule Induction

Global rule induction aims to induce a set of rules in a form similar to if-then-else syntax from a set of examples to perform a specific task. Here, we employ GRID system developed by [10] to extract the rules from the training data. GRID emphasizes the use of global feature distribution in all of the training examples in order to make better decision on rule induction. For every training instance, GRID generates a context feature vector centered around the tagged slot (TS). In this case, the tagged slot denotes the boundary of a known story. The context feature vector is of the general form:

$$<c_{-k}> \ldots <c_{-2}> <c_{-1}> \boldsymbol{TS} <c_{+1}> <c_{+2}> \ldots <c_{+k}> \quad (1)$$

$$[<(f_{-k}^{1})..(f_{-k}^{m})>..<(f_{-1}^{1})..(f_{-1}^{m})> \boldsymbol{TS} <(f_{1}^{1})..(f_{1}^{m})>..<(f_{k}^{1})..(f_{k}^{m})>] \quad (2)$$

Where $k$ is the number of context units (shots) around the tagged slot *TS*. Each instance thus contains *2k+1* context shots (one tagged slot TS plus k shots to the left and k shots to the right). In Eq (1), $<c_i>$, $\{i = -k \text{ to } +k\}$ is the context shot $i$ of the tagged slot TS. Let m be the number of features in each context shot $<c_i>$ , the context feature vector for a single instance can therefore be represented as in Eq (2). By arranging all the instances as in Eq (2), we obtain a global context feature representation for the training set as shown in Figure 3.

Before we discuss of how the rules are generated, we shall discuss how to measure quality of the generated rule. Here we use *Laplacian* measure: *Laplacian* (r) = (M +1)/(N+1), where M denotes the number of negative training instances (errors) covered by rule r; and N gives the number of positive and negative instances covered by rule r.

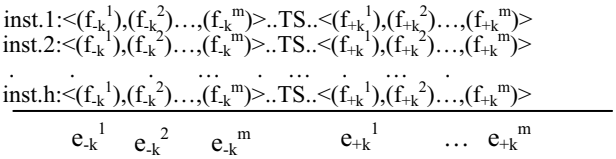


**Figure 3: Global distribution of *h* instances. The occurrences of the common element features at position *g* are cumulated as $e_g^i$.**

In order to understand how GRID learns the pattern rules, we present an example (taken from [10]) to extract the start time (<stime>) of seminar announcements in text documents as shown in Table 2. The task aims to extract the semantic slot <stime> (context position 0) which indicates the "start time" of a seminar announcement. In this example, we only consider two context units (at positions $LH_{-1}$ and $LH_{-2}$) and a subset of feature representations.

| Training instances | context position | | |
|---|---|---|---|
| | -2 | -1 | 0 |
| inst 1 | Time | : | <stime> 3:30 PM </stime> |
| inst 2 | Time | : | <stime> 2 p.m. </stime> |
| inst 3 | Time | : | <stime> 4 p.m. </stime> |
| inst 4 | start | at | <stime> 10 am </stime> |
| inst 5 | begin | from | <stime> 11:30 AM </stime> |

**Table 2: An example for extracting slot "starting time".**

The example contains 5 positive instances where the desired slots <stime> are tagged at context position 0. By examining the feature frequency for the context elements at positions $LH_{-1}$ and $LH_{-2}$ around the tagged slot (context position 0), the tagged slot NP_Time (the semantic representation of all time instances) at context position 0 appears most frequently (it occurs 5 times). Thus, it has the highest coverage in the training example pool. This feature is then selected. The generated rule is

**"NP_Time → NP_Time** is starting time**"**

This rule, however, does not satisfy the *Laplacian* measure as NP_Time also appears in all instances that are not "starting" time. Thus, more contextual information must be included to constraint the rule. From the Table, the token **":"** at the $LH_{-1}$ context position appears 3 times for all positive instances with feature NP_Time in slot 0, it is therefore selected next, and the rule is now constrained as:

**": NP_Time → NP_Time** is starting time**"**

For the CMU seminar announcement corpus, this rule is sufficient to satisfy the *Laplacian* measure, and thus we obtained "Rule 1" (see below). Once this rule is obtained, the 3 instances are removed from the positive training example pool. The above process is iterated on the remaining two positive examples and finally we obtain another two rules ("Rule 2" and "Rule 3"). Thus, all the generated rules that satisfied the *Laplacian* measure are:

**Rule 1: ": NP_Time → NP_Time** is starting time"
**Rule 2:** "**start at** NP_Time → NP_Time is starting time"
**Rule 3**: "**begin from** NP_Time → NP_Time is starting time"

During testing, if any of these rules applies, the tags <stime> and </stime> will be inserted beside the NP_Time's boundaries. Further details of GRID can be found in [10].

### 3.2 Extension of GRID to Story Boundary Extraction

We extend GRID system to incorporate the ability to extract the information in news video. To accomplish the story boundary extraction task, we model each shot (context unit) using the following features: (a) its tagged category represented by unique ID (obtained from the shot tagging process), (b) scene/location change (change or unchanged), and (c) cue-phrase at the beginning of story (presence or absence of cue-phrase).

A simple example to illustrate how GRID learns the pattern extraction rules from news video data is given in Table 3, which shows 5 positive instances selected from our training data. For simplicity, we use only one feature (shot category) per context unit. The desired slots are tagged as "<>", a dummy slot to indicate story boundary.

| Instances | Context position | | | | |
|---|---|---|---|---|---|
| | $LH_{-2}$ | $LH_{-1}$ | $C_0$ | $RH_{+1}$ | $RH_{+2}$ |
| 1 | SP | LR | <> | LEDS | LR |
| 2 | sport | LR | <> | LEDS | sport |
| 3 | LR | LR | <> | LEDS | sport |
| 4 | sport | sport | <> | LEDS | LR |
| 5 | sport | TS | <> | LEDS | SP |

**Note:** SP – Speech; TS – text-scene; LR – Live-reporting; LH – left context, RH – right context, and $C_0$ or <> tagged slot (story boundary)

**Table 3: An example for extracting slot "<>" in news video when k =2 (2 context shots to the left and 2 to the right of $C_0$).**

We can see from the Table that "LEDS" (leads-in category) appears most frequently at $RH_{+1}$ position, and it thus has the highest coverage in the training example pool. This feature is then selected, and the generated rule is:

**"<> LEDS → <> is story boundary"**

This rule, however, does not satisfy the *Laplacian* measure as we found that there are many LEDS shots in the negative training instances that are not the successor of story breaks. So the rule has to be constrained. We see that the token "LR" appears 3 times at $LH_{-1}$ position and is therefore selected next. The rule is now constrained as:

**"LR <> LEDS → <> is story boundary"**

In our corpus this rule is sufficient to meet the *Laplacian* measure, and thus the first rule found is:

**Rule 1: "LR <> LEDS → <> is story boundary"**

Once we obtain this rule, we remove the 3 instances from the positive training example pool. The algorithm then iterates the above process on the remaining two positive examples and finally obtains another two rules as follows:

**Rule 2: "sport <> LEDS → <> is story boundary"**
**Rule 3: "TS <> LEDS → <> is story boundary"**

After all rules are extracted that cover all the instances in the positive training pool, the resulting set of rules will be used to match the patterns in the testing instances.

## 4. EXPERIMENTAL RESULTS

The training and testing data sets provided by TRECVID [3] contains ~120-hours of CNN and ABC news video. Same to our previous work, we used ~60 hours for training and the rest for testing. We set up several experiments to evaluate the effectiveness of the selected features and context unit length *k*. The results are presented in the following sections.

### 4.1 Effect of contextual length *k* and feature sets

We conduct a series of experiments with different context length *k* and feature combinations. The four feature combinations used are: (a) Shot category (Sc); (b) Shot category + Scene change (Sc+Cc); (c) Shot category + cue-phrase (Sc+Cu); and (d) combination of all. From the experiments, we found that the best result is obtained when k =2 and using only *shot category* (Sc) as the feature, as shown in Table 4. The reason is because shot category is able to capture the main semantics of the shot and thus helps in finding story boundaries, which in fact are the transitions from one shot category to another.

| Features | $F_1$-value |
|---|---|
| Sc | **75.05** |
| Sc + Cu | 62.3 |
| Sc + Cc | 59.6 |
| Sc + Cu +Cc | 65.0 |

$F_1 = (2*R*P)/(R+P)$;
R - Recall;
P - Precision

**Table 4: The performance when using different feature sets and *k* =2.**

### 4.2 Evaluation Results

Table 5 presents the best results from GRID on the TRECVID data set [3].

| Data | Precision | Recall | $F_1$ |
|---|---|---|---|
| ABC | 71.95 | 84.51 | **77.72** |
| CNN | 76.76 | 68.47 | 72.38 |
| **Average** | 74.36 | 76.49 | 75.05 |

**Table 5: Results from GRID on the TRECVID data set.**

The Table shows that GRID performs better on ABC news video. This is because, CNN data contains variety of programs and the structure is more dynamic. Overall, the average performance of GRID reaches $F_1$ of 0.75. Although the results of GRID is slightly lower that that achieved by HMM of about 0.77 in $F_1$ measure as reported in [6, 7], GRID offers the advantage of requiring less computational cost and complexity; the cost (excluding feature extraction) based on the training data set is about 10 times lower as compared to the HMM. Figure 4 shows some examples of the detected story units in the test instances. Overall, about 60% of stories are detected by the common rule: "<> Anchor-Shot" (Figure 4a). Many story boundaries are also detected by rules: "<> Health-Shot" (4b), "Sport <> play-shot" (4c) and "Anchor <> Weather-Shot" (4d). More analysis of the results and the generated rules can be found in [6,7].


**Figure 4: Examples of the detected stories.**

## 5. CONCLUSION AND FUTURE WORK

We proposed a framework to extract story boundary information in news video. We employed a global rule induction technique using GRID system to perform the extraction task. We tested the generality and quality of our framework on the TRECVID data set [3]. Although the HMM yields slightly better results, GRID offers the advantage of requiring less computation cost and easier maintenance and scale-up to larger and new corpuses. As we can see, here shot genre is useful for story boundary extraction. It is also useful for retrieval of particular shots from news videos, such as queries on "weather", "sport", etc. Future work is to test the system on other types of videos such as the documentary or movie. We also plan to extend the system to generate news threads to support news retrieval and tracking task.

## 6. REFERENCES

[1] A. Merlino, D. Morey and M. Maybury. *Broadcast news navigation using story segmentation*. Proc. of the fifth ACM international conference on Multimedia, Seattle, Washington, United States, 1997.

[2] A.G. Hauptmann and M. J. Witbrock. *Story Segmentation and Detection of Commercials in Broadcast News Video.* Proc. of Int'l Conference on Advances in Digital Libraries (ADL), California, USA, 1998.

[3] //www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html#2003

[4] //www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html#2004

[5] L. Chaisorn, T.-S Chua and C.-H Lee, "The Segmentation of News Video into Story Units", *Proceedings of IEEE Int'l Conf. on Multimedia and Expo*, Lausanne, Switzerland, August 2002.

[6] L. Chaisorn, T.-S. Chua, and C.K Koh, Y-L. Zhao, H.-X. Xu, H-.M. Feng. *Story Segmentation and Classification for News Video*. Proceedings of TRECVID 2003, 17-18 November, Washington D.C., USA.

[7] L. Chaisorn. "A Hierarchical Multi-modal Approach to Story Segmentation in News Video". PhD thesis, School of Computing, National University of Singapore, 2005.

[8] W. H. Hsu and S.-F. Chang. *Visual Cue Cluster Construction via Information Bottleneck Principle and Kernel Density Estimation.* Intl' Conference on Image and Video Retrieval (CIVR) 2005, Singapore.

[9] Dale, H. Moisl, and H. Somers. *Handbook of Natural Language Processing.* Marcel Dekker, New York USA, 2000.

[10] J. Xiao, T. -S. Chua and J. Liu. *A Global Rule Induction Approach to Information Extraction.* Proc. of the 15th IEEE International Conference on Tools with Artificial Intelligence (ICTAI-03), 2003.