Higher-level video interpretation using shot transition and motion activity features

Tung Sing Leung†, Ankush Mittal† and Loong Fah Cheong†
National University of Singapore

Abstract — Semantic video indexing has generally been restricted to the specific domains like news. A lot of semantic information for a large number of classes is contained in shot editing and motion features. In this paper, we use the information, derived from these features, integrated over a larger timescale than a shot length time to form rules for the extraction of semantic video classes. Such formation of rules yields high representational accuracy of the classes as shown by the experiments conducted on 3 hours of video data. Key applications of the approach include media classification, commercial removal and sports highlight extraction.

Keywords — content-based retrieval, editing techniques, high-level descriptors, commercial extraction, media classifier

I. Introduction

Today, a large amount of video data is produced and stored in the analog format. The digital form of video data enables content-based retrieval of videos so that the video databases can be searched with visual queries. In order to achieve this, all the video data in the database must be indexed and video data abstractions must be created.

To make use of digital videos effectively, semantic level indices should be furnished to the users. Currently, there is a great need of framework that operates on semantic level with broad strategies. Much of the work on semantic level is confined to a limited domain. [2][3][4] proposed schemes for the parsing of soccer, basketball and tennis video, respectively. The video data is parsed to recognize the features specific to the particular sport such as center circle, a goal post in the soccer field, court-line, player location and textural information to determine events happening during the game. The sports parser algorithms above can only be applied in a specific sport domain and a new set of indexes must be defined for each kind of sport. A number of commercials detection algorithms [5] [6] [7] [8] had been proposed, but experiments were done only to extract commercials from the news sequence.

We design a classification process by studying “the film grammar”, which is an accumulation of knowledge and conventions for expressing certain semantics of a scene more effectively [9]. Many of the film grammars have been developed for the camera techniques and the editing techniques. Editing techniques consist of the cuts and gradual transitions such as wipe, dissolve, fade-in and fade-out. Each of them does convey some meaning also. For example a fast cut to evoke tension and a dissolve to suggest the passage of time. Therefore editing techniques seem to provide some important cues for video indexing and should not be omitted.

In this paper, a video indexing system, which is based mainly on those cues given by the editing techniques and the visual dynamics of shots is proposed. The input of the system is digitized video data of various types of TV programs. The segmentation module is used to segment the input video into shots and identify all the editing techniques used. In addition, features like static shot and black frames, which are the distinct feature in TV news programs and commercials, respectively, are also extracted. The second module is used to classify and group all the shots into different categories: MTVs, News, Sports and commercials. It also locates some higher level of indices such as the replay shots in the sports sequence.

The paper is organized as follows. Section 2 presents the features used in our approach. Design of the rule-based system is considered in Section 3. Experimental results are presented in Section 4. Conclusion follows in Section 5.

II. The syntactic features

The low level features that we utilized for the video classification system are the following: 5 transition types (cut, dissolve, fade-out, fade-in, wipe), static shot and black frames. A shot is an unbroken sequence of frames from one camera. A scene is defined as a collection of one or more adjoining shots that focus on an object or the object of interest.

A. Static Shot detection

A static shot is a shot consisting of very little objects and camera movements. This feature provide strong cues in the News detection simply because the news consist of many static shots (e.g. anchorperson shots) comparing with all other TV programs. A shot will be declared as static shot if the following conditions are fulfilled:
1. the average frame-to-frame difference within a shot is smaller than a very small threshold, $T_{diff}$.
2. the shot length of the shot is longer than $T_{length}$

III. Design of the rule-based system

A rule-based system is needed to map the low level measures to the corresponding high-level video characteristics based on some predefined rules. For example, in a news program, the anchor person segments are usually charac-
terized by very low activities compared to the other segments.

First of all, the rule-based system need to convert the temporal locations of each low-level features in a set of row vectors, one for each feature. The length of each vector is equal to the number of frames in the video sequence. Each element of the vector corresponding to one frame can only be either 0 or 1, representing “FEATURE NOT EXIST” and “FEATURE EXIST”, respectively. We would like to call this row vector as the feature vector and is denoted by $F_i(k)$, where $F_i(k) \in \{0, 1\}$, with $i = \{commercial, sport, MTV, news, black frame, cut, dissolve, wipe, fade, gradual, fast cut\}$ and $k$ can be any value between 0 and total number of frames in the video sequence. For example,

$$F_{\text{wipe}}(k) = \begin{cases} 1, & 2 \leq k \leq 5 \\ 0, & \text{otherwise} \end{cases}$$

means a wipe section is detected in the video sequence from the frame 2 to 5.

### A. Identifying sections with a high cutting rate

Generally, the average shot lengths of MTVs, and commercial sequences are relatively shorter than that of the sports and the news sequence. Since cut and gradual transitions are random events, they are subjected to the semantic content of the video, the video editor’s style, etc. Therefore statistical concept can be used to study the pattern of shot length in various sequences. Fig. 1 shows the normal plots of the shot length for all the four TV programs after the natural log is applied. Since all the four plots are almost linear, therefore we can assume that the shot length of all the four types follow log normal distribution. Table I shows the statistical properties of shot length for each category of the TV program. Their PDFs in Fig. 2a show that the log shot length in MTVs and commercials follow very similar distribution, which is quite different from that of the Sports and the News. However, if MTV and commercial sequences are classified as Type 1, and the sports and the news sequences are classified as Type 2, then these two types of videos will have significant difference in their distributions as shown in Fig. 2b. A new feature vector, $F_{\text{fastcut}}$ can be defined to represent Type 1 video as follows:

$$F_{\text{fastcut}}(k) = \begin{cases} 1, & S(j)_{\text{start}} \leq k \leq S(j)_{\text{end}} \\ 0, & \text{otherwise} \end{cases}$$

where $\{S(1), S(2), \ldots, S(n)\}$ are the $n$ shots in the video sequence and having corresponding shot length $\{L_1, L_2, \ldots, L_n\}$; $T_{\text{shot length}}$ is set to $e^{4.84} \approx 57$ frames based on Fig. 2b.

### B. Detecting Commercials

During the commercial break of any TV programs, the TV station broadcast a commercial block, which is a sequence containing several consecutive commercials (or spots). This sequence normally has a high scene change rate to attract the audiences’ attention. Short dissolve, fade-in and fade-out are used frequently, while wipe is very rarely used. Most importantly, a single black frame has to be inserted between 2 successive commercials to distinguish one spot from others.

Based on these observations, a segment is likely to be a commercial block if it contains a number of black frames, and each of them are separated by a duration shorter than $T_c$ times the mean duration between the black frames in the entire sequence. That is, assuming the location of all the $m$ black frames in the sequence is denoted by $\{B_1, B_2, \ldots, B_m\}$, then the mean distance between the black frames is denoted by

$$b_{\text{black frame}} = \frac{1}{m} \sum_{i=1}^{m-1} (B_{i+1} - B_i)$$

Commercial blocks are hypothesized over those regions

<table>
<thead>
<tr>
<th>TV Programs</th>
<th>Mean of log shot length</th>
<th>Standard deviation of log shot length</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sports</td>
<td>4.7598</td>
<td>0.8039</td>
</tr>
<tr>
<td>News</td>
<td>4.7111</td>
<td>0.8257</td>
</tr>
<tr>
<td>MTVs</td>
<td>3.3658</td>
<td>0.6188</td>
</tr>
<tr>
<td>Commercials</td>
<td>3.2813</td>
<td>0.7175</td>
</tr>
</tbody>
</table>
having concentrated black frames by

\[ F_{\text{commercial}}(k) = \begin{cases} 1, & B_i \leq k \leq B_{i+1} \\ 0, & \text{otherwise} \end{cases} \]

where \( 1 \leq i \leq m \) and \( T_{\text{c}} \) is set to 1.7. Short hypothesized non-commercial sections less than 35 seconds long are merged with their following segment. Short hypothesized commercial sections less than 28 seconds long are merged into the following segment. Because the scene changes can also occur rapidly in a non-commercial content, therefore after the merging steps described above, any hypothesized commercial section containing no black frame at all will be removed.

C. Detecting Music video (MTV)

MTV clips normally have a high scene change rate, gradual transitions like dissolve, fade-in and fade-out are used frequently and most of them last for 2 minutes and above. Based on these observations, the hypothesized MTV sections can be obtained by:

\[ F_{\text{mtv}}(k) = h(F_{\text{fastcut}}(k) - F_{\text{commercial}}(k) - F_{\text{wipe}}(k)) \]

where \( h(x) = \begin{cases} 1, & x \geq 1 \\ 0, & x \leq 0 \end{cases} \)

Two positive \( F_{\text{mtv}}(k) \) sections that are not longer than 40 seconds are merged. Those positive \( F_{\text{mtv}}(k) \) sections that are shorter than 80 seconds are then removed. These merging and removing steps are essential because as shown in Fig. 2b, MTVs and commercials can be classified as Type I with maximum 86% accuracy and there are 14% probability in misclassifying sports and news if all the left over are labeled as MTVs.

D. Detecting News

Generally, a news program has a low cutting rate and static shots appear frequently. Furthermore, gradual transitions like dissolve and wipes are often used especially in BBC news.

This is very difficult to distinguish news broadcast from the sports sequence because they have very similar characteristics such as shot length distributions, the types of gradual transitions used, and the existence of static shots. Static shot usually occur in sports (e.g. soccer) when a long panning shot is used to track the ball. The shot lengths of these types of shots are typically long and the spatial changes within the shot are small because the players and the ball appear to be very small within the long shot.

Based on these observations and problems, sections of video sequence that are not marked as Type I are labeled as the News, if there are many static shots found within the sections and the sections contain no wipe transitions. There are two steps in identifying the news sequences. First step is to ensure that no Type I regions will be declared as News. This can be done by:

\[ F_{\text{static}}(k) = h(F_{\text{static}}(k) - F_{\text{fastcut}}(k)) \]

Two positive \( F_{\text{static}}(k) \) sections that are not longer than 2 minutes are then merged. Those positive \( F_{\text{static}}(k) \) sections which are shorter than 80 seconds are then removed. The second step prevents any sports region from being declared as News. Therefore the hypothesized TV Broadcast News sections can be obtained by:

\[ F_{\text{news}}(k) = h(F_{\text{static}}(k) - F_{\text{wipe}}(k)) \]

However wipe transition is normally used to separate two location video sequences of different stories in the BBC news. If wipes are excluded, the system may not be able to capture certain portions of the news and reduce the recall rate. Experimental results show that much more gain in precision rate can be obtained instead. This is because the wipe transition is used more frequently in sports compared to the BBC News. Furthermore, those missing out portions usually can be filled by the merging procedure at the end of the News detection.

E. Detecting Sports

A sport program normally has a low scene change rate; wipe and dissolve transitions are used frequently to enclose replay sequences, but fade-in and fade-out transitions will not be used at all. As mentioned before, static shots often appear when a long shot is used. Based on these observations, the system will label all the unlabeled Type 2 sections that consist of a high concentration of dissolve and wipe transitions. After the merging, any of these sections which last for less than 80 seconds will be unlabeled.

The hypothesized sport sections can be obtained by:

\[ F_{\text{sports}}(k) = h(F_{\text{dissolve}}(k) + F_{\text{wipe}}(k) - F_{\text{news}}(k) - F_{\text{fastcut}}(k)) \]

Two separated sports sections are merged if they are less than 80 seconds apart. Since short sport dip is not common, those sports sections which are shorter than 60 seconds are removed.

F. Fine-timing of boundaries based on specific temporal structure

Until this stage, all the 4 categories of TV programs should be more or less located by the corresponding feature vectors but the exact boundaries of each sequence have not been determined and the rule-based system has a great difficulty in recovering them exactly based on the 7 low-level features that we have chosen. These error can be minimized if the temporal structure of the each program category is taken into account.

The feature vectors of each program categories are converted into corresponding indexes as illustrated by the example shown in Fig. 3. Notice that in this example, the first positive pulse (a false detection) in the \( F_{\text{sport}}(k) \) is removed by the \( F_{\text{news}}(k) \) region because the duration in the portion of \( F_{\text{sport}}(k) \) is much shorter than that of \( F_{\text{news}}(k) \); therefore it is likely to be a false detection. After the conversion, let \( b_1, b_2, ..., b_N \) denote \( N \) hypothesized boundaries between any two distinct TV programs. The initial positions of these boundaries are set in the following manners:
1. Any hypothesized boundaries besides a commercial segment are moved to the first and the last black frame respectively. This is illustrated by the boundaries $b_1$ and $b_2$ in Figure 3. The purpose of doing this is to maximize the precision rate of commercial detection.

2. As shown by boundary $b_2$ in the Figure 3, a hypothesize boundary $b_i$ is located in the middle of the unlabeled region.

3. If there is no overlap or unlabeled regions in between two successive segments, the boundary $b_i$ will be moved to the end of the preceding segment like $b_3$ in the figure.

4. If the unlabeled gap between two labeled segments is shorter than 4000 frames, the hypothesized boundary $b_i$ will be located in the middle of the unlabeled region. However if the unlabeled gap is longer than 4000 frames, the hypothesize boundary will be put at the beginning of the unlabeled segment. This is because these long unclassified regions are likely belonging to some other TV program categories, they are not ignored so that the rule-based system can be expanded in the future more easily to handle other TV programs. This is illustrated by $b_5$ in the figure.

After all the initial locations for the hypothesized boundaries are determined, these hypothesized boundaries are fine-tuned based on the temporal structures of the two segments, $SEG(k)$ and $SEG(k+1)$, that enclose the boundary. If $SEG(k)$ is a News or MTVs segments, then $b_i$ are snapped to the nearest fade-out. This is because News and MTVs sequences usually end with a fade-out transition. If $SEG(k)$ and $SEG(k+1)$ belong to different types of sequences, for example $SEG(k)$ is Type 1 while $SEG(k+1)$ is Type 2, the system will snap $b_i$ to the nearest Type 1 boundary. If $b_i$ is located beside the commercial segments, the system will assume the duration of the first and the last spot of a commercial sequence to be 30 seconds and adjust $b_i$ accordingly.

G. Detecting replay sequences in sports

As mentioned before, a sport replay is normally enclosed by a pair of dissolve transition. Under this situation, we will denote all the $n$ dissolve transitions in the sports sequence by $\{D(1), D(2), ..., D(n)\}$. Those shots between $D(i+1)$ and $D(i)$ are declared as replays, if

$$D(i+1) - D(i) < 0.3 \times \mu_{dissolve}$$ (8)

where $1 \leq j \leq n$ and

$$\mu_{dissolve} = \frac{1}{n-1} \sum_{i=1}^{n-1} (D(i+1) - D(i))$$

Eq(7 & 8) assume the duration of the replay sequences are shorter than that of the non-replay sequences. This assumption is invalid when two replay sequences are too close to each other. Under that situation, these two replays will be simply merged into one replay sequence. This situation is uncommon in normal TV sports programs. Although soccer is used as an example here, these rules can also be applied to the other types of sports like volleyball, cricket, etc.

H. Degree of interesting of a replay sequence

After a replay sequence is identified, the rule-based system can further “measure” how important the replay is. This can be achieved by counting the number of cuts or dissolves within the replay sequence. For example, after a goal in a soccer match, a replay will be presented to show how the goal was made from different camera angles. Each of these is normally separated by a dissolve (and cut may be used for this purpose as well).

IV. Experimental Evaluation

A. Set Up

A 3-hour long video sequences has been used to test the rule-based system. That sequence contains 11 minutes of MTV, 18 minutes of commercials, three full-length news (BBC and Singapore TCS news) lasting for 1.25 hours and 1.2 hours of sports clips.

B. Performance parameters

Two parameters, cover recall and cover precision, are used to evaluate the system. They are defined by:

$$\text{cover recall} = \frac{L_c}{L_a} \times 100\%$$ (9)

$$\text{cover precision} = \frac{L_c}{L_d} \times 100\%$$ (10)

where $L_c$ is the length of correct detection, $L_a$ is the length of actual detection and $L_d$ is the length of declaration that was made by the system.

C. Final Results after fine-tuning & Discussion

Table II shows the system accuracy before the fine-tuning and the final performance of the system after the fine-tuning is shown in Table III. The results in Table II show
that the rule-based system performs well in both classifying different video types and locate the boundaries between two segments. The system performs best in locating the News and commercials. Therefore, the system will provide the best result in locating the commercial block within the TV news broadcast.

### TABLE II

**Results of video classification before fine-tuning.**

<table>
<thead>
<tr>
<th></th>
<th>$l_c$ (m/s)</th>
<th>$l_d$ (m/s)</th>
<th>$l_d$ (m/s)</th>
<th>Recall</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sports</td>
<td>73'50'</td>
<td>80'42'</td>
<td>80'52'</td>
<td>91.1</td>
<td>90.9</td>
</tr>
<tr>
<td>Comm</td>
<td>16'55'</td>
<td>20'59'</td>
<td>16'55'</td>
<td>76.7</td>
<td>100.0</td>
</tr>
<tr>
<td>MTV</td>
<td>11'25'</td>
<td>11'25'</td>
<td>17'32'</td>
<td>100.0</td>
<td>65.1</td>
</tr>
<tr>
<td>News</td>
<td>64'48'</td>
<td>70'29'</td>
<td>65'45'</td>
<td>91.9</td>
<td>98.6</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>165'49</strong></td>
<td><strong>183'37</strong></td>
<td><strong>180'16</strong></td>
<td><strong>90.3%</strong></td>
<td><strong>92.0%</strong></td>
</tr>
</tbody>
</table>

### TABLE III

**Results of video classification after boundaries fine-tuning.**

<table>
<thead>
<tr>
<th></th>
<th>$l_c$ (m/s)</th>
<th>$l_d$ (m/s)</th>
<th>$l_d$ (m/s)</th>
<th>Recall</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sports</td>
<td>18'8'</td>
<td>80'42'</td>
<td>83'10'</td>
<td>96.8</td>
<td>93.9</td>
</tr>
<tr>
<td>Comm</td>
<td>19'35'</td>
<td>21'01'</td>
<td>19'55'</td>
<td>93.2</td>
<td>98.4</td>
</tr>
<tr>
<td>MTV</td>
<td>81'7'</td>
<td>11'25'</td>
<td>81'7'</td>
<td>72.6</td>
<td>100.0</td>
</tr>
<tr>
<td>News</td>
<td>69'11'</td>
<td>70'29'</td>
<td>72'15'</td>
<td>98.1</td>
<td>95.8</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>175'11</strong></td>
<td><strong>183'37</strong></td>
<td><strong>180'16</strong></td>
<td><strong>95.4%</strong></td>
<td><strong>95.4%</strong></td>
</tr>
</tbody>
</table>

The reason for the system to give a poor precision rating with MTV data set is that it makes one serious false detection, which last for about 2 minutes in total. The false detection is due to misclassifying a portion of the sports as MTV.

The system also makes a wrong declaration by misclassifying a portion of MTV (about 2.5 minutes) as sport. This is because that small portion of MTV sequence consists of a slow cutting rate which is lesser than the threshold, $T_{shotlength}$, for Type I.

By comparing the results of Table II and III, the accuracies are improved though the exact location of each boundary still remains unknown. It can be noticed that the fine-tuning have the effect of balancing the performance of the recall and precision. Actually the fine-tuning is used to move the boundaries, $b_i$, forward and backward along the temporal axis. From the results, it can be noticed that the fine-tuning part play important roles in increasing the recall performance without affecting much on the precision performance and vice versa. For example, in the commercial category, after the fine-tuning, a large gain of 16.5% recall performance is obtained by giving up 1.6% of precision performance. Similarly for MTV, an achievement of about 35% improvement in precision rating is obtained by giving up 27.4% of recall rating. Same things happen for the news category, where 6.2% of improvement in recall with a trade-off of losing 2.8% of precision rating. The reduction of the performance in one side is always less than the gain of performance.

### V. Conclusions

Our system performs video classification and replay shot detection based mainly on the editing techniques that are identified by the integrated algorithm and the visual dynamics of shots. The rules in the system are designed based on the film grammar and some prior knowledge about the editing techniques used in various TV programs. The experimental results have shown that the rule-based system can effectively classify various TV programs into four categories: MTVs, News, sports and commercials. The results successfully demonstrate that by identifying various editing techniques in a given video stream, it is possible to efficiently extract high-level semantic information and sufficiently support the retrieval of many meaningful semantic contents such as “show 10 most interesting replay sequences of all the sports programs today”.

### Acknowledgment

The second author is greatly indebted to Dr. P. V. Krishnan, I.I.T. Delhi, for his kind support throughout the project. The basic foundation to pursue proper research was given by him.

### References