A VIDEO SUMMARIZATION ALGORITHM USING THE SEMANTIC SCORE METHOD

Yasushi TAKAHASHI*, Kazuo SUGIYAMA**, Makoto WATANABE**

* Sony Corporation, Takanawa Court 3-13-1Takanawa, Minato-ku, Tokyo, 108-0074 Japan
** Faculty of Engineering, Chiba University, Yayoi-cho 1-33, Inage-ku, Chiba 263-8522 Japan

Abstract: This report proposes a video summarization algorithm that generates a video summary using the Semantic Score Method for the purpose of automatically generating a video clip. The Semantic Score Method enables to quantitatively describe narrative structure of content through its cognitive evaluation by viewers. The paper derives the algorithm through a comparative analysis of 15 movies and their trailers, and demonstrates the effectiveness of the algorithm by evaluating the quality of the video summaries it produces. This algorithm makes it possible to produce various video clips according to viewer characteristics based on the Semantic Scores of a large number of viewers collected at movies and TV drama previews. It also makes it possible to intuitively select program titles or browse content from moving images generated by the algorithm in a storage-based broadcasting environment.

Keywords: video summary, content evaluation, movie, broadcast, internet

1. Introduction

In the age of multi-channel digital TV broadcasting and broadband networks, video summaries that enable viewers intuitively to comprehend a movie or TV program's content are extremely crucial for marketing such titles and influencing viewer selections. A great deal of research on the automation of video summaries has been carried out using image and audio signal processing technologies[1][2], but it has been difficult to obtain a summary of narrative videos, such as movies and TV dramas, based on their meaning. In response, the authors of this paper proposed the Semantic Score (SS) Method for describing a story using a rating scale of complication and resolution[3]. Through a comparative analysis of 15 movies and their trailers, the paper derives a video summarization algorithm that combines the SS Method and signal processing technology. It then demonstrates the effectiveness of the algorithm by evaluating the quality of the video summaries it produces. Within these pages, the term "video summarization algorithm" refers to a series of video processing procedures that extract from a movie the optimum scenes and shots required for the desired video summary and then cut and order the necessary video shots. This technology targets automated processing in the future.

2. Video Summary Requirements and the Narrative Structure of Movies

2.1. Video Summary Requirements Suited to Viewer Needs

Recently, countless numbers of trailers and promotional videos are shown at movie theaters and on TV. However, in the broadband age where consumers independently select what they want from a vast amount of content, it is no surprise that they are not always satisfied with the advertising video that is pushed upon them. For instance, even when selecting a title, consumers want to search in a number of ways, including by plot summary, highlights, and scenes with their favorite actors. There are also varied demands regarding the length of summary videos. Automatic or semi-automatic video summarization technology using algorithms is indispensable to responding to such diverse needs. R. Lienhart and his fellow researchers have so far proposed a
video abstracting technique that employs a heuristic approach based on movie production know-how[4]. It does not, however, have a quantitative algorithm, which makes it difficult to efficiently respond to a variety of video summary needs. Creating a video summary that ascertains a movie’s outline and generates interest in the movie requires a technique based on the characteristic of viewer narrative cognition, and the Semantic Score Method discussed hereafter proves vitally useful for that.

2.2. Narrative Structure and the Semantic Score (SS) Method

The SS Method is a technique using a rating scale of complication (+) and resolution (-) by scene for the quantitative evaluation of the process of narrative understanding. It is based on typical questions and comments that viewers have such as “What is this?” or “Now I see!” while they are watching a movie or other narrative video.

The evaluation standards of complication and resolution employed by the SS Method correspond to the phases used to express the stages of dramatic narrative plots since the ancient plays of Greece. The 19-century German drama critic G. Freitag diagrammed these narrative processes in what is known as Freitag’s Triangle (Figure 1)[5]. B. Laurel revised Freitag’s Triangle from the perspective of information characteristics by quantifying it in units (events)[5].

We applied this concept to movies and decided to carry out evaluations by scene using the previously mentioned narrative cognition standard. The term “scene,” as used in this paper, is the narrative unit into which a story is divided, and it consists of incidents and events that advance the story. For example, the movie Speed can be divided into 246 scenes, resulting in the semantic scores shown in the Score column of Table 1. These scores are values between +1.0 and -1.0 that evaluate complication and resolution per scene. The degree of complication in the right column is the integrated value of those scores, and it indicates the complexity of the story up to the respective scene. The semantic graph (Figure 2), in which the x-axis represents the scene and the y-axis the degree of complication, expresses the narrative structure as understood by the viewers. This is called the “story shape.” The highest peak in the graph corresponds to the story’s climax. Speed consists of 10 episodes, and we can detect local peaks that correspond nearly to them. P1 through P10 in Figure 2 are peaks in which the height difference between the peak and the valleys to its left and right exceeds 1.4. That these scenes are the major turning point of each episode is apparent from their content.

The SS Method has the ability to quantitatively describe narrative structure in this manner.

3. A Video Summarization Algorithm Using Comparative Analysis of Movies and Their Trailers

This section presents a comparative analysis of 15 movies and their trailers using the SS Method to derive a video summarization algorithm, while also giving attention to the production process of both movies and trailers.

3.1. Comparative Analysis of Speed and Its Trailer

We began with a pre-analysis of the movie Speed as our example. We evaluated Speed using the SS Method described in Section 2, which yielded the movie’s story shape (Figure 2). For its trailer, we divided it into shots to elucidate the details, and then analyzed the content of the video, corresponding movie scenes and their semantic scores, and structural components of the video (Table 2). We categorized the

<table>
<thead>
<tr>
<th>Scene</th>
<th>Content</th>
<th>Score</th>
<th>Complication</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>An elevator goes down.</td>
<td>0.3</td>
<td>0.3</td>
</tr>
<tr>
<td>2</td>
<td>A security guard examines a stranger.</td>
<td>0.5</td>
<td>0.8</td>
</tr>
<tr>
<td>3</td>
<td>Introduces himself as a repair man.</td>
<td>-0.3</td>
<td>0.5</td>
</tr>
<tr>
<td>4</td>
<td>Let him show a work order.</td>
<td>0.3</td>
<td>0.8</td>
</tr>
<tr>
<td>5</td>
<td>A stranger kills the gurdman.</td>
<td>0.8</td>
<td>1.6</td>
</tr>
<tr>
<td>6</td>
<td>A group of executives run inside an elevator.</td>
<td>0.1</td>
<td>1.7</td>
</tr>
<tr>
<td>7</td>
<td>The terrorist pushes a button.</td>
<td>0.5</td>
<td>2.2</td>
</tr>
<tr>
<td>8</td>
<td>An explosion crushes the elevator.</td>
<td>0.6</td>
<td>2.8</td>
</tr>
<tr>
<td>9</td>
<td>Emergency brakes stop the elevator.</td>
<td>-0.7</td>
<td>2.1</td>
</tr>
<tr>
<td>10</td>
<td>The team specialists rush into the building.</td>
<td>-0.3</td>
<td>1.8</td>
</tr>
<tr>
<td>11</td>
<td>Thirteen passengers are shut in an elevator.</td>
<td>0.2</td>
<td>2.0</td>
</tr>
<tr>
<td>12</td>
<td>A bomb is set up anywhere.</td>
<td>0.5</td>
<td>2.5</td>
</tr>
</tbody>
</table>

Table 1: Semantic Score of the movie Speed

<table>
<thead>
<tr>
<th>No.</th>
<th>Trailer shots</th>
<th>Movie scenes</th>
<th>Components</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>N: The latest movie news.</td>
<td>21 They found a bomb.</td>
<td>-0.4</td>
</tr>
<tr>
<td>2</td>
<td>N: The best work, got ahead of “The hard”.</td>
<td>22 The work looks pretty solid.</td>
<td>-0.5</td>
</tr>
<tr>
<td>3</td>
<td>N: World wide sales exceeded 20,000,000.</td>
<td>23 The work looks pretty solid.</td>
<td>-0.5</td>
</tr>
<tr>
<td>4</td>
<td>N: A group of executives run inside.</td>
<td>24 A group of executives run inside.</td>
<td>0.1</td>
</tr>
<tr>
<td>5</td>
<td>N: Thirteen passengers are shut in an elevator.</td>
<td>25 A crane is tied to the elevator.</td>
<td>-0.5</td>
</tr>
</tbody>
</table>

Table 2: Shot analysis of the trailer Speed
structural components of the video into five types (A through E) as follows:

A. Title: Relates the movie’s title through a logo or narration.

B. Story: Relates the movie’s story by depicting its outline and subject matter by means of video, narration, or a telep.

C. Symbolic images: Represents the work’s images through symbolic video or video fragments.

D. Cast and crew introduction: Introduces the cast as well as the director, screenwriter(s), and producer(s).

E. Publicity information: Emphasizes and publicizes the movie’s features such as production cost, awards won, and production technology.

3.1.1. Analysis of the Story Shape of Speed

Ninety-five out of the 117 total shots that comprise Speed’s trailer were mapped to the 66 movie scenes. The vertical lines in Figure 2 indicate the position of scenes used in the trailer. A detailed inspection of the distribution reveals a large number of scenes were extracted from the introductory portion of the story, while none were extracted from the conclusion.

Now if we say scene extraction can be quantified as a percent of 100 by dividing the number of scenes used in the trailer from specific sections of the movie by the number of scenes in that section, then in the introduction, the scene extraction rate would reach 39.1%, while in the story development, turning point, and conclusion it would be 23.9%, 22.0%, and 0% respectively. Here, “story development” refers to a series of episodes that develop the story, and “turning point” refers to important events that lead to a climax. The graph peaks are major scenes representing such episodes, and the probability of their extraction in the trailer is thought to be high. In the case of Speed, six out of 10 peak scenes (underlined peak numbers in Figure 2) were used in the trailer. This is remarkably high compared to the scene extraction rate of 20.7% throughout the movie.

3.1.2. Analysis of the Trailer Shots of Speed

Figure 3, which corresponds to Figure 2, plots the trailer shots that proceed from top to bottom on the vertical axis onto the movie scenes on the horizontal axis. The dotted
lines at the beginning and end of the graph represent symbolic image, publicity, and cast/crew introduction-related video, and they tend to jump wildly all over the graph. On the other hand, the solid lines represent the video of the story, and it is clear that the editing was done mostly along the lines of the movie's approximate temporal order.

3.1.3. Characteristics of scene extraction for the trailer Speed

If one studies the scene extraction rate by score, one notices that this trailer makes extensive use of scenes with high absolute value scores. In addition, the main cast members of full-shot or closer images and visual/audio special effects are found extensively in scenes with high absolute value scores, while scenic locations are found extensively in scenes with low absolute value scores (Figure 4). Figure 4 divides scene scores into five blocks, and it shows the percentage of shots that belong to six image characteristics.

Based on the above-mentioned analysis, in the next section, we present an analysis of 15 titles.

![Figure 4 Image Characteristics of the Trailer Speed](image)

3.2. Comparative Analysis of 15 Titles and Their Trailers

We selected three titles each from the five major movie genres (action, drama, comedy, romance, and fantasy) for a total of 15 titles[6], and then analyzed the relationship between them and their trailers based on the following six approaches.

3.2.1. Structural Components of Trailer Videos

We divided the trailer into the five structural components of title, story, symbolic images, cast and crew introduction, and publicity information, and then tabulated the components in seconds, resulting in a time breakdown by genre (Figure 5). The combined average indicates that the top structural component in a trailer is story, and it accounts for 66.0% of its time. The next most prevalent structural component after story is title at 13.7%. This is followed by cast and crew introduction, symbolic images, and publicity information in that order.

This reveals that an algorithm for accurately extracting story is the most important part of video summarization. By adding title and cast/crew introduction shots to this, we can form a video summarization framework. The publicity information component is influenced by the distributor's publicity strategy, and it consists of a lot of video that is not found in the movie itself. The symbolic image component is influenced by the creator. These last two components are not suited for automatic extraction using an algorithm.

In light of this, we decided on an algorithm that extracts the story, title, and cast/crew introduction components from a movie.

![Figure 5 Five structural components of trailer videos](image)

3.2.2. Scene Extraction Rates for a Movie's Introduction, Story Development, Turning Point, and Conclusion

An analysis of the trailer scene extraction rate for a movie's introduction, story development, turning point, and conclusion showed that at 7.9%, conclusion had the lowest rate (Figure 6). This is due to the fact that not giving away the ending raises interest and anticipation in a movie. Introduction was only a mere three points higher than story development and turning point, but it is particularly important to video summaries as a component introducing the background story and characters.

Based on the above-mentioned observations, we decided on an algorithm that did not include scenes from the conclusion and raised the scene extraction rate for the introduction to a level higher than that for story development and turning points.

![Figure 6 Scene extraction rates for movie's introduction, story development, turning point, and conclusion](image)
3.2.3. Scene Extraction Rate by Score

We arranged scene extraction rates into five blocks of scores in order to find the relationship between the scores of extracted scenes and scene extraction rates (Figure 7). This revealed a tendency for higher scene extraction rates for absolute value 0.7 to 1.0 blocks.

Story development and turning point scenes with strong complication increase anticipation of what lies ahead, but those with strong resolution tend to include information that shows how an episode concludes. This is not observed in introduction scenes because they do not convey a sense of conclusion.

Based on the above observations, we decided on an algorithm that extracts introduction scenes with high absolute values and story development and turning point scenes with high complication scores.

![Image](Figure 7 Scene extraction rates in the score block)

3.2.4. Extracted Scene Scores and Image Characteristics

An analysis of the relationship between the scores of extracted scenes and image characteristics of shots taken from those scenes and used in the trailers revealed that shots containing the main cast members of full-shot or closer images and those with visual and audio special effects accounted for 73.1% scenes with a score of absolute value 0.7 or more, while scenic shots accounted for 29.1% of scenes with a score of absolute value 0.2 or less (Figure 8).

Based on this, we decided on an algorithm that extracts shots containing the main cast members of full-shot or closer images and those with visual and audio special effects from high-scoring scenes and scenic shots from low-scoring scenes.

3.2.5. Scene Extraction Rate from Peaks

A comparison of scene extraction rates from peaks and from the entire feature for action movies and comedies reveals that the rate from peaks is markedly higher (Figure 9). Other genres were omitted because each of them had a title with a peak extraction rate of zero.

Based on this, we decided on an algorithm that adds peak scenes to video summaries.

![Image](Figure 9 Scene extraction rates from peaks)

3.2.6. Average Number of Seconds in Trailer Shots

We determined the average length of trailer shots and average length of the movie's scenes in seconds by genre (Figure 10). The combined average shot length for trailers was 1.8 seconds, and this result also demonstrated a strong correlation between average shot length for trailers and average length of the movie scenes.

Accordingly, we decided on an algorithm that set an approximate standard of 1.8 seconds for the average shot length, and allowed some variety in shot length in order to incorporate the tendencies of various genres.

![Image](Figure 10 The average length of trailer shots and movie's scenes in seconds)

3.3. Video Summarization Algorithm

We used the analysis results from Section 3.2. to develop our video summarization algorithm as follows:

1) Extraction of high-score scenes: The algorithm extracts high score scenes with absolute value of 0.7 or higher from the introduction and high score scenes with positive values of 0.7 or more from story development and turning point sec-
tions.

2) Extraction of peak scenes: The algorithm extracts peak scenes and their subsequent scene from the introduction and peak scenes from story development and turning point sections.

3) Extraction of high-impact shots: In high-score scenes and peak scenes, the algorithm extracts shots that contain full-shot or closer images of the main characters or shots with striking audio/visual special effects.

4) Extraction of scenic shots: The algorithm extracts scenic shots from scenes with low absolute value scores.

5) Shot length: We set the standard for shot length to between 1.5 and 2 seconds.

6) Shot arrangement: We arrange the extracted shots in the order they appear in the movie, add the title at the beginning, and introduce the cast members at the end.

The next section presents the production and evaluation of video summaries using our algorithm.

4. Creating and Evaluating Video Summaries

4.1. Creating Algorithm-Generated Video Summaries

To prove the effectiveness of high-score, peak, and scenic scenes that constitute the core of the algorithm and see the impact of scene arrangement and sound effects, we established seven different video summarization algorithms for producing 1- to 2-minute video summaries (Table 3).

Type A in Table 3 assembles video using simply cut shots, while Type B considers sound effect according to set rules to eliminate the sense of being abruptly cut off. With Type B, shots where dialogue gets cut off midway through have the cut dialogue superimposed onto the following shot using a voiceover, and the theme music is then added to all the shots as background music. Type 2 and 3 subsamples high-score scenes, so that there will be no major variation in scene number between Types 1, 2, and 3.

4.2. Evaluation of Video Summaries by Paired Comparison

We evaluated the relative merits of the seven types of summary videos though a paired comparison of Type A and a paired comparison of Type B. We divided 26 student test subjects into two groups, varied the sample presentation order, and added dummy samples in between the real test samples in order to eliminate the effect of the order in which we presented the samples. Once we collected the data, we gave one point to samples that paired up well, tabulated the pairs, and then used the eigenvector method to normalize the root sum of scale to one. This resulted in the order A3>A2>A1>A4 and B3>B2>B1, starting from the best for Type A and Type B, respectively (Table 4).

Based on the above evaluation, we determined: (1) Type 3, which consists of high-score scenes, peak scenes, and scenic scenes, to be the best for both Type A and B; and (2) that arranging shots in the same order as the movie is better than a random arrangement.

Table 4 Evaluation of video summaries by paired comparison

<table>
<thead>
<tr>
<th>A paired comparison of Type A</th>
<th>A1</th>
<th>A2</th>
<th>A3</th>
<th>A4</th>
<th>Σn</th>
<th>Σn²</th>
<th>Normalize</th>
<th>Order</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>13</td>
<td>14</td>
<td>10</td>
<td>10</td>
<td>47</td>
<td>2209</td>
<td>0.21</td>
<td>3</td>
</tr>
<tr>
<td>A2</td>
<td>12</td>
<td>13</td>
<td>11</td>
<td>20</td>
<td>56</td>
<td>3136</td>
<td>0.29</td>
<td>2</td>
</tr>
<tr>
<td>A3</td>
<td>16</td>
<td>15</td>
<td>13</td>
<td>22</td>
<td>66</td>
<td>4356</td>
<td>0.41</td>
<td>1</td>
</tr>
<tr>
<td>A4</td>
<td>8</td>
<td>6</td>
<td>4</td>
<td>13</td>
<td>31</td>
<td>961</td>
<td>0.09</td>
<td>4</td>
</tr>
<tr>
<td>Sum 10662</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>A paired comparison of Type B</th>
<th>B1</th>
<th>B2</th>
<th>B3</th>
<th>Σn</th>
<th>Σn²</th>
<th>Normalize</th>
<th>Order</th>
</tr>
</thead>
<tbody>
<tr>
<td>B1</td>
<td>13</td>
<td>12</td>
<td>4</td>
<td>29</td>
<td>841</td>
<td>0.17</td>
<td>3</td>
</tr>
<tr>
<td>B2</td>
<td>14</td>
<td>13</td>
<td>5</td>
<td>32</td>
<td>1024</td>
<td>0.20</td>
<td>2</td>
</tr>
<tr>
<td>B3</td>
<td>22</td>
<td>21</td>
<td>13</td>
<td>56</td>
<td>3136</td>
<td>0.63</td>
<td>1</td>
</tr>
<tr>
<td>Sum 5001</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

4.3. Creating an Additional Summary Video by Free Editing

In our comparison of Type A3 and B3 with existing trailers, we decided to add a video summary freely edited from a movie. We did this because existing trailers include images that are not in the movie and also employ sound effects, which rules out a simple comparison with summary videos made by algorithm. On the other hand, freely edited video summaries, made using only shots from the movie, enable comparable evaluation with summary videos made by algorithm. The freely edited type was a 1-minute and 6-second-long video summary that was created by a student under the concept of conveying the allure of an action movie.
4.4. 5-Point Evaluation for Summary Videos and Existing Trailers

We conducted a 5-point evaluation for the following four questions using four samples: Type A3 and Type B3 created by an algorithm, an existing trailer (T), and a freely edited video summary (F). The possible responses to the following questions were "Unequivocal yes," "Yes," "Undecided," "Not really" and "Not at all."

- (a) Do you want to see this movie?
- (b) Can you imagine the movie's content?
- (c) Do you like this trailer?
- (d) Is the trailer itself any good?

The test subjects were comprised of 38 students. We divided them into two groups, and then presented the four samples to them, ordering differently by group. The points for responses were from 5 for "Unequivocal yes" down to 1 for "Not at all." The highest total score was for the existing trailer (T), which had an average of 4.1. This was followed by the sound-effect type (B3) at 3.3, freely edited type (F) at 3.0, and simply cut shot type (A3) at 2.6 (Figure 11). The significant probability is 0.000 for B3 and A3 and 0.090 for B3 and F. This makes it clear that algorithms that add sound effects are: (1) superior to the simply cut shot type and (2) by no mean inferior to freely edited video summaries, thereby proving the effectiveness of our algorithm.

Based on these results, we added the following to our algorithm:

7) Sound effects: Shots where dialogue gets cut off midway through will have the cut dialogue superimposed onto the following shot using a voiceover, and then the theme music is added to all the shots as background music.

5. Conclusions and Future Issues

5.1. Effectiveness of Semantic Score-Based Algorithm

The SS Method effectively adds the following four summary video parameters to our proposed algorithm:

1) Video from and around peaks for conveying the overall image of the story.
2) Scenes with high absolute value scores for making a strong impact.
3) Background scenes with low absolute value scores, usually scenic shots, for eliminating the undesirable montage effect that can be created by automatic assembly of scenes.
4) Control of scene extraction rate by introduction, story development, turning point, and conclusion in order to increase anticipation of the movie.

This algorithm can flexibly generate video summaries suited to specific purposes by adjusting the threshold values of the above-mentioned parameters and the filters for extracting shots. For instance, it can create outline, highlight, or cast member-specific summary videos of the desired length. In this way, our algorithm is effective for generating video summaries that ascertain narrative structure.

5.2. Broad Applications Using Metadata and a Video Summarization Engine

The video summarization algorithm with the Semantic Score Method opens wide application areas. For instance, delivering semantic scores as content metadata[7] over broadcast channels and the Internet and installing video summarization engines on user terminals or servers would enable the provision of a content browser anytime, anywhere using video summaries that fit user demands (Figure 12). As an example, we prototyped a content browser envisioned for application with TVs that can store program data. Using this, viewers will be able to directly select movie or drama content while browsing a video clip, and will be able to watch a movie with an easily searchable and easy to understand interface (Figure 13).
5.3. Future Issues

Improvements are being made to our technique so that it can easily collect semantic scores at movie and TV program previews[8][9]. Meanwhile, research and development that integrates results and findings in many fields is required to establish image and audio signal processing technology modeled after our algorithm. Our algorithm, which is based on viewer content evaluation, can also serve as a tool for visually representing viewer excitement and emotion. We are interested in further exploring such wide-ranging applications.

References