Narrative Abstraction Model for Story-oriented Video

Byunghee Jung, Taeyeong Kwak, Junehwa Song* and Yoonjoon Lee*

Division of Computer Science, KAIST (Korea Advanced Institute of Science and Technology),
#373-1, Guseong-dong, Yuseong-gu,

Daejon, Republic of Korea (305-701)
+82-2-781-5966

{bhjung, tykwak}@dbserver.kaist.ac.kr, {junesong, yjlee}*@cs.kaist.ac.kr

ABSTRACT

TV program review services, especially drama review services, are one of the most popular video on demand services on the Web. In this paper, we propose a novel video abstraction model for a review service of story-oriented video such as dramas. In a drama review service, viewers want to understand the story in a short time and service providers want to provide video abstracts at minimum cost. The proposed model enables the automatic creation of a video abstract that still allows viewers to understand the overall story of the source video. Also, the model has a flexible structure so that the duration of an abstract can be adjusted depending on the requirements given by viewers. We get clues for human understanding of a story from scenario writing rules and editorial techniques which are popularly used in the process of video producing. We have implemented the proposed model and successfully applied it to several TV dramas.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing – *abstracting methods*; H.5.1 [**Information Interfaces and Presentation**]: Multimedia Information Systems – *videos*

General Terms

Design and Human Factors

Keywords

Video abstraction, Narrative abstraction, Narrative structure, Film, Story-oriented, Story understanding

1. INTRODUCTION

Due to increases in network technology performance and the busy lives of viewers, TV program review services have become one of the most popular video on demand services on the Web. According to the click count of a broadcasting company website, the total click count for various TV program review services including video on demand, script and related information review services reaches near 37 million per day [9]. According to this report, drama is one of the most popular genres for the review service along with news and sports. In terms of the consumed bandwidth as well as the click count, drama occupies a large portion (up to 44 percent) as compared to other genres. This is largely due to the fact that users need to watch a drama for at least several minutes in order to understand what happens and why it happens in a given episode.

The viewers of the drama review service are usually willing to watch the episode thoroughly to understand its story. However, they can easily be interrupted by several events, such as business phone-calls, scheduled personal matters, and so on. A major reason for these kinds of inconveniences is that watching a drama episode completely generally requires quite a long time – close to an hour. Therefore, in order to enable viewers to understand the story in a shorter time, it is very helpful to provide them with abstracted versions of drama episodes called video abstracts. One critical problem in accomplishing this is that manual video abstraction requires much manpower and time. For example, a person approximately spends several hours when he makes a video abstract of one fifty-minute drama episode. Furthermore, since the duration of a video abstract is fixed, the service provider must regenerate video abstracts whenever the duration requirements change.

In this paper, we propose a video model that is applicable to the automatic generation of video abstracts of story-oriented video genres such as drama. As it targets story-oriented video, the generated abstract gives a summary of the video which still allows viewers to understand the overall story. Also, the model has a flexible structure so that the duration of the resulting video abstract can be adjusted depending on the time limitation given by viewers.

A video abstract is defined as a sequence of still or moving images that present the content of a video compactly with concise information about the content without losing the original message [14]. Because the characteristics of content determine the aim of viewers who want to watch a video abstract, the method of video abstraction should be different according to these characteristics. According to the characteristic of content, video can be classified into two types: story-oriented video and event-oriented video. Story-oriented video such as a movie, drama or sitcom has no pre-determined format and viewers can obtain information or understand the story when they watch the whole video. Event-oriented video such as sports or news has a pre-determined format

MM'04, October 10–16, 2004, New York, New York, USA.

and users can be entertained or obtain information even through watching only a few interesting parts.

The proposed model concentrates on that the aim of viewers to watch the drama review service is to understand the overall story not to feel the overall atmosphere or impression. The model enables automatic and duration-flexible abstraction that satisfies the viewer's aim – story understanding in specified time. In order to achieve this, we analyze scenario writing rules and editorial techniques in order to get clues about the human understanding of the story of a video. Taking such techniques under consideration, we observe that a story is expressed through a sequence of dramatic incidents and the progressions between dramatic incidents. Thus, concentrating on dramatic incidents that contribute more to the story is a good solution to making an understandable abstract of a story-oriented video.

Based on such observations, we establish our narrative abstraction model as a set of dramatic incidents and the progress of story between them. Also, the degree of progress between two dramatic incidents is measured as follows. The metric of the degree of story progress is based on the importance of characters, the intensity of interactions between the characters in dramatic incidents, and the frequency and intensity of the interrelationship between the dramatic incidents. The process of abstraction is selecting a sequence of dramatic incidents that show further progress of a story satisfying the given duration. To show the feasibility and performance of our model, we implement and apply it to TV dramas - the most typical example of a story-oriented video - and analyze the performance from several viewpoints. Using empirical evidence, we show that viewers are satisfied with such abstracts in terms of understanding the story and saving time. Furthermore, automatic abstraction can increase not only the service provider's productivity, *i.e.*, saving their manpower and time, but also the efficiency of a service's bandwidth utilization.

The rest of this paper is organized as follows. In Section 2 and 3, we consider the general scenario writing rules and editorial techniques that are the key ideas of our model and introduce related work. Our narrative abstraction model and its implementation are presented in Section 4 and experimental results are described in Section 5. Finally, we give concluding remarks and discuss further work in Section 6.

2. A FILM AS A FORMAL SYSTEM

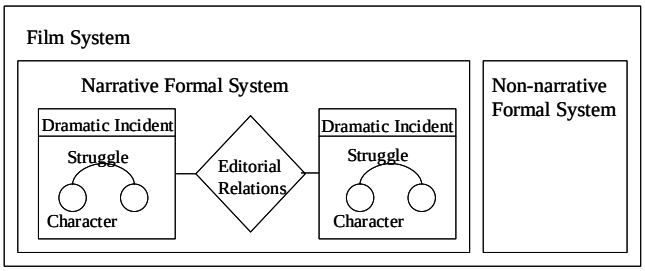


Figure 1. The general structure of a film system

To build up an abstraction method for a story-oriented video, we need semantic knowledge which helps viewers understand the progress of a story. Unlike text which has explicit structures such as terms or sentences, a video has no explicit structure from which we can easily extract semantic knowledge. Thus, we get clues for such semantic knowledge from scenario writing rules, which are popularly adopted in scenario writing to facilitate film-producing process. These rules are accumulated knowledge to effectively express semantics of a scene in the form of a film [11]. We can also obtain clues from editorial techniques which are frequently used to express a story explicitly when editors make a film. In this section, we describe the model that we build up from the techniques, as shown in Figure 1.

As in [6], we regard films or movies as a type of a formal system. A system, in a general sense, is a group of interacting, interrelated, or interdependent elements forming a complex whole. In the case of a film system, scenes in a video are related to each other to provide information or to make a story understandable.

A film system can be categorized into two types, namely, a narrative formal system and a non-narrative formal system [6]. A narrative formal system includes story-oriented videos such as drama and movie, whereas a non-narrative formal system includes event-oriented videos and special purpose videos such as documentaries, educational videos, political presentations, etc. Therefore, the focus of this paper is the narrative formal system. A narrative formal system consists of two elements: a narrative element and a stylistic element. A narrative element is a scene that contains a dramatic incident and a stylistic element is a scene that helps an audience understand a dramatic incident. For example, a background scene, a monologue scene and a retrospection scene are all stylistic elements. The narrative elements are the main constituents for the progress of a story

A narrative is a physical realization of a story [5], *i.e.*, a narrative is a series of dramatic incidents sequenced over a video's playing time in order to express a story. A dramatic incident contains interactions among two or more characters. An author leads the story through the interactions and the struggles caused by the interactions. According to the types of interactions, dramatic incidents can be classified into three categories as following [11].

- Action without dialog

- Action with dialog

- Dialog without action

Dramatic incidents in a story are causally related with each other. This relationship expresses the progress of a story. To show this relationship prominently in a film, the editors compose a film

using editorial techniques. This is the major difference between a scenario and a film. A scenario can directly describe the causal relationship using descriptions. However, in a film, audiences have to understand the causal relationship through imagination while watching the video. To have an audience follow the story during playtime, editors apply appropriate editorial techniques to express the causal relationship. This kind of editorial techniques is called *editorial relations*. Generally, four kinds of editorial relations have been widely used as shown below [6]. These editorial relations basically use repetition and contrast of features (e.g., color tone, background, shot duration, etc) to effectively express the progress of story.

- Graphic relations

The progress of a story can be expressed through contrast of important characters' action or contrast of luminance and color conditions between scenes.

- Rhythmic relations

To emphasize a scene, the editors usually compose the scenes with many short shots. Also, a radical progress of a story can be expressed by alternately placing a fast pacing scene with a slow pacing scene.

- Spatial relations

The progress of a story can be expressed as the change of spaces.

- Temporal relations

The progress of a story can be expressed as the change of time.

Given a pair of dramatic incidents, we consider several factors to evaluate the progress of story between dramatic incidents. First, the embedded editorial relations and their intensity are important metrics. To express more radical progress of a story, more intensive editorial relations are applied between dramatic incidents. In addition, the intensity of an interaction in each dramatic incident is important to evaluate the progress. We also note that the interaction is highly related with the importance of involved characters. Important characters generally involve more frequently and intensely in interactions. Therefore, the importance of characters can be another metric for evaluating the dramatic incident's contribution.

3. RELATED WORK

Story-oriented abstraction has been studied mainly in the field of text summarization [12]. A text abstract is given as a compact representation of terms and sentences related to a subject. For such abstraction, a subject is first selected, usually as a term which most frequently appears in the text. Then, related terms and sentences are selected. In novels or other story-telling writings, meaningful events are usually semantically related with a subject, and expressed via terms and sentences related with the subject.

The summary sequence approach shown in [1,16] clusters the shots/scenes with similar physical features, selects parts (*e.g.*, key frames) of each cluster, and combines them to make an abstract. Also, the summary sequence approaches in [13, 19] give reviewers quick impression of entire source video using control over video browsing, *e.g.*, playback speed, sampling rate control. These summary sequence approaches may give quick overall atmosphere of a video. However, they may not helpful in understanding the story. This is because viewers cannot infer clear semantics or causal relationships from watching parts of each dramatic incident or sequence of them. To understand a story, it can be more helpful to remove the clusters less important to the progress of a story than to minimize the size of each cluster.

Automatic speech recognition or languages understanding techniques have been adopted for the abstraction of story-oriented video [7,18]. This approach can take advantage of lexical semantic of terms as in the text summarization. However, it cannot be applied to a drama. Scenario of a drama consists of characters' dialogs and action descriptions. Thus, there may not be obvious subject terms. This makes the summarization difficult.

There have been some researches that applied film producing techniques for special purposes. The research in [3,8] applied camera position or composition of screen to detect boundaries of video segments. Also, the work in [2,17] utilized the relevance between the components in a shot *e.g.*, embedded color, motion and object features to detect the boundary of video segments or to retrieve specific video segments. Our approach is quite different in that we apply the knowledge acquired from the general scenario writing rules and editorial techniques, to scene-based video abstraction.

4. VIDEO ABSTRACTION MODEL

In this section, we describe our video abstraction model. Our approach mainly focuses on the generation of a video abstract which provides viewers with a good level of understanding even in a short-length abstract. Thus, we should select pairs of video segments whose relationship contributes to the progress of a story more than others. At the same time, the total duration of the selected segments should satisfy the target duration. The contribution of a pair of video segments to the progress of the whole story can be determined through three aspects: the intensity of editorial relations, the intensity of interaction within a dramatic incident, and importance of characters in a dramatic incident.

In Section 4.1, we present requirements of a good story-oriented video abstract. The discussion in Section 2 leads to our story-oriented video model, called *narrative abstraction model*, which is described in detail in Section 4.2. In the model, we introduce a function called *DoP* to determine the progress of a story between two video segments. Finally, a video abstraction process based on the proposed model is described in Section 4.3.

4.1 Requirements

We adapt the four criteria defined in [15] for e-learning video abstraction to the evaluation of abstracts from our model. Comparing to these 4C requirements, our requirements for a good

abstraction of story-oriented video are converted to the following 4C. Coverage and concise attributes are required for the correct extraction of essential segments and coherence is required to maintain the validity of interrelationships between segments. Coordination is required to adapt the abstract to different user's environments.

- Coverage

An abstract should include all the essential segments of a story.

- Conciseness

An abstract should be composed of only the necessary segments.

- Coherence

Each segment of an abstract should be interrelated with each other in terms of a story.

- Coordination

An abstract must be customizable to each viewer's different intention and environment. More specifically, the duration should be customizable to viewer's different time limit.

4.2 Narrative Abstraction Model

In this section, we briefly describe the proposed scene-based video model called *narrative abstraction model*. We first pull together the related basic concepts to describe the model. We then briefly describe the concept of *DoP* function. Our discussion in this section is rather brief just to convey the core concepts. More detailed description can be found in our follow-up paper.

Figure 2 shows the general structure of a story-oriented video as described in Section 2. As shown, a story-oriented video is a sequence of scenes. A scene can be a narrative or a stylistic element. Since the stylistic element is to assist viewers to better understand a narrative element, a story can be understood with an abstract composed only of narrative elements. In other words, the background or monologue help understanding the next dramatic incident, *i.e.*, where the incident takes place and why the persons do that. However, even though the monologue is removed, viewers can infer why it happens from causal relationship between the previous and the next dramatic incidents. The stylistic elements occupy very small portion in terms of total duration - about less than 10 %.

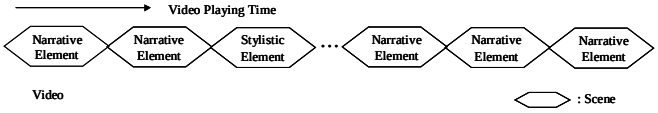


Figure 2. General structure of story-oriented video

The basic unit of our model is the narrative element. For simplicity, stylistic elements are omitted. As there is no interaction between characters and little variation in color and luminance in stylistic elements, there is progress of a story neither between two stylistic elements, nor between a stylistic element and a narrative element. For example, there is no causal relationship between the background scene and a dramatic incident. A narrative element includes at least one interaction between characters. Also, there exists at least one editorial relation between each pair of (consecutive or non-consecutive) narrative elements to show the progress of story. We refer to this existence of progress in terms of story as *narrative connection.* Because the progress can exist between the scenes that are not contiguous, narrative connections exist between the possible pairs of narrative elements in the model. We represent the degree of progress of a story in a narrative connection by its *DoP* value. Figure 3 illustrates our model.

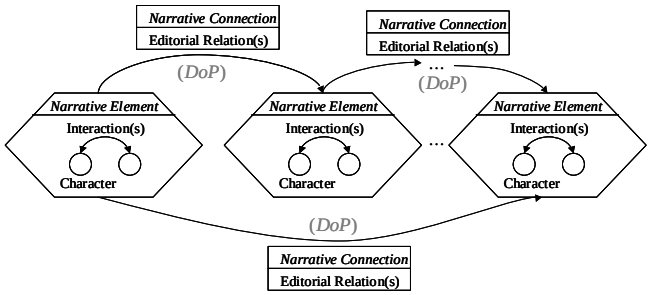


Figure 3. Framework of narrative abstraction model

4.2.1 Basic Notations

Table 1. Notations for the proposed model

Symbol	Description
V	Video
S	Set of shots in a video
$s[i]$	i-th shot in S
$s[c, j]$	j-th shot of the scene c in a video
C	Set of scenes in a video
$c[i]$	i-th scene in C
NE	Set of narrative elements in a video
$ne[i]$	i-th narrative element in NE
$NC(ne_i, ne_j)$	Narrative connection between narrative element ne_i and ne_j
O	Set of objects appearing in a video
$o[i]$	i-th newly appearing object in a video
O_{ne}	Set of objects appearing in narrative element ne

Table 1 summarizes the notations used to describe our narrative abstraction model formally. A *shot* is the elementary segment of a film that is separated by physical editing e.g. cut. A *scene* is a

sequence of shots that are taken in the same place, or constitute a dramatic incident in a story.

Definition 1. A story-oriented video V is a sequence of scenes, *i.e.*, $V = \langle c[1], c[2], c[3], \cdots, c[n] \rangle$.

Definition 2. A narrative element is a scene which is either dialog or action.

The action and dialog can be characterized as follows. In the case of dialog, the facial shots of each character involved in the dialog are repeated. In the case of action, the variation of distance between characters is great. If the facial shots are repeated and the variation of distance is great simultaneously, the scene is an 'action with dialog'. Hereafter, we call a character involved in an interaction an *object*.

Definition 3. A narrative connection NC is a function defined over $NE \times NE$, *i.e.*,
$NC : NE \times NE \rightarrow \{e_i \mid e_i \subseteq ER\}$,

where $ER = \{graphic, rhythmic, spatial, temporal\}$.

Each editorial relationship, *i.e.*, graphic, rhythmic, spatial, and temporal relation is defined in Section 4.2.2.

For any pair of narrative elements ne_i and ne_j , there is a narrative connection $NC(ne_i, ne_j)$, which is a subset of ER . For example, $NC(ne_i, ne_j) = \{graphic, temporal\}$ or

$NC(ne_i, ne_j) = \{graphic, rhythmic, temporal\}$.

Definition 4. DoP is a function such that

$$DoP : NC(ne_i, ne_j) \rightarrow \square ,$$

where $\square$ is a real number between 0 and 3.

A method to obtain the DoP value is described in Section 4.2.2.

Definition 5. A video abstract AV of a video V with duration d is a subset of NE such that DoP values of AV is the maximum among all possible subsets of NE satisfying the given duration d.

Then, the abstraction of a story-oriented video is the process of obtaining AV from the given V .

4.2.2 The DoP Function

Consider a pair of narrative elements ne_i and ne_j . The DoP function, $DoP(NC(ne_i, ne_j))$, represents the degree of progress of a story between the narrative elements, ne_i and ne_j , *i.e.*, on the narrative connection $NC(ne_i, ne_j)$. It is measured through the intensity of interactions in ne_i and ne_j , the intensity of the editorial relations in $NC(ne_i, ne_j)$, and the importance of the characters involved in the interactions. We denote each of them by $I(NC(ne_i, ne_j))$, $R(NC(ne_i, ne_j))$, $O(NC(ne_i, ne_j))$ respectively. Each metric is normalized between 0 and 1. Hence, the DoP function is computed by their weighted sum:

$$DoP(NC(ne_i, ne_j)) = I(NC(ne_i, ne_j)) \cdot w_I + R(NC(ne_i, ne_j)) \cdot w_R + O(NC(ne_i, ne_j)) \cdot w_O .$$

The intensity of interactions $I(NC(ne_i, ne_j))$ is measured differently according to their types, *i.e.*, an action or a dialog. The intensity of a dialog increases as the number of shots in the dialog increases and the average duration of shots decreases. Also, the intensity of an action increases with the sum of the differences of distances between two objects in consecutive frames. The intensity of interactions in a narrative connection is computed by adding the intensities of the actions and the dialogs for ne_i and ne_j .

The intensity of the editorial relations, $R(NC(ne_i, ne_j))$, is also measured differently for each type, *i.e.*, for a graphic, a rhythmic, a spatial, or a temporal relation as follows:

- Graphic relation

A graphic relation represents either a rapid change in action or a rapid change in light or color conditions. Its intensity is derived from (1) the difference of the object distances, (2) the difference of average R, G, B values (3) the difference of average luminance values between ne_i and ne_j .

- Rhythmic relation

A rhythmic relation alternates a fast-paced scene with a slow-faced one. Thus, its intensity is derived from the difference in the average shot durations of ne_i and ne_j .

- Spatial relation

A spatial relation represents the change in spaces, and its intensity is computed from the difference in the average R, G, B values of the background regions of ne_i and ne_j . The background regions are obtained by excluding the detected face regions of each object. This is feasible because, in a drama, most objects are shown in close-up views.

- Temporal relation

A temporal relation represents the change in times. The intensity is measured through the difference in the number of objects in ne_i and ne_j .

Lastly, we estimate the importance of an object via its coverage in a video. Then, given a narrative connection, the importance of objects involved in the interactions $O(NC(ne_i, ne_j))$ is computed by the average importance of all involved objects.

We show the effectiveness of the proposed DoP via an experiment in Section 5.

4.2.3 A Narrative Structure Graph

The proposed narrative abstraction model is represented as a weighted directed acyclic graph (WDAG), which we call a *Narrative Structure Graph (NSG).*

Definition 6. Given a story-oriented video, V, with a set of narrative elements, NE, a *narrative structure graph (NSG)* is a weighted directed acyclic graph $\langle Ve, E, W \rangle$

where $Ve = \{ne \mid ne \in NE\}$,
$E = \{NC(ne_i, ne_j) \mid ne_i, ne_j \in NE\}$,
$W = \{DoP(NC(ne_i, ne_j))\}$.

For simplicity, we represent each narrative element ne_i, narrative connection $NC(ne_i, ne_j)$, and its *DoP* value $DoP(NC(ne_i, ne_j))$ as a vertex v_i, an edge e_{ij}, and a weight w_{ij}, respectively.

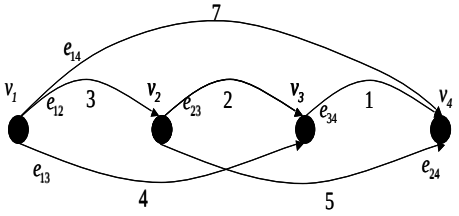
4.3 Video Abstraction Process

We have proposed a novel story-oriented video abstraction model in Section 4.2. In this section, we describe an abstraction process based on the proposed model. The abstraction goes through three steps, *i.e.*, preprocessing, modeling, and abstraction.

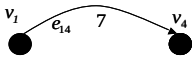
For preprocessing, we perform face recognition as well as shot detection and scene detection. The basic concept of our video model is based on the specific object's activity (interaction); therefore it is important to discriminate the object's identity. We use the method in [4] for the shot and scene change detection and the face recognition. After preprocessing, we model the input video through our narrative abstraction model. Three steps are involved: distilling the narrative elements from scenes, identifying each narrative connection, and calculating the *DoP* values for each narrative connection. Then, we construct a NSG from the model.

Given a NSG, the abstraction is the process of constructing a subgraph until the duration of the subgraph reaches the target duration. The final subgraph should have the maximum total weight among all possible candidates satisfying the given duration. This problem of identifying an optimal subgraph is NP-complete [10]. We use a simple greedy method: we first choose the edge with the highest weight, and add the head and tail vertices of the edge to the subgraph if the target duration is still satisfied. Any vertex which is already in the current subgraph is ignored. Then, all the edges connecting the vertices in the current subgraph are added. This process is repeated as long as the total duration of the subgraph is less than the target duration.

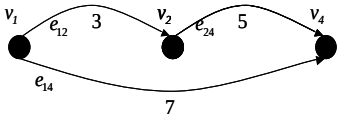
Figure 4(a) is the NSG of an input story-oriented video. The durations of each vertex, v_1, v_2, v_3, v_4 are 1 min, 2 min, 3 min and 2 min, respectively. The vertices are subscribed according to their temporal order in the source video. The abstracts with the target duration of 3min and 5min are shown in Figure 4(b) and 4(c), respectively. In Figure 4(b), the edge e_{14} which has the highest weight is selected and its head and tail - v_1 and v_4 - are added to the subgraph. Then, all edges connecting the vertices in the current subgraph (in this case, only edge e_{14}) are added. Total duration of the subgraph satisfies the target duration of 3 min. In Figure 4(c), the edge e_{12} is added which has the highest weight and satisfies the target duration 5min. Then, vertex v_2 and edge e_{12} and e_{24} are added to the subgraph.



(a) Narrative Structure Graph of an Input Video



(b) An abstract with target duration = 3 min



(c) An abstract with target duration = 5 min

Figure 4. Examples of an abstraction

5. EVALUATION

5.1 Experimental Environments

We evaluate the proposed model using TV drama videos. We chose following four episodes, two from the TV drama "Autumn Story" (AS-1 and AS-2, below) and the other two from "Body Guard" (B-1 and B-2). The duration of each episode is about 50 minutes. AS-1 and AS-2 mostly consist of dialogs, and B-1and B-2, actions. The selected videos are MPEG-1 compressed with the aspect ratio of 352×240.

- AS-1: the fourth episode of "Autumn Story" (41 scenes, 23 narrative elements)
- AS-2: the fifth episode of "Autumn Story" (41 scenes, 23 narrative elements)
- B-1: the nineteenth episode of "Bodyguard" (44 scenes, 20 narrative elements)
- B-2: the twentieth episode of "Bodyguard" (44 scenes, 25 narrative elements)

To evaluate our model, we manually produced the abstracts of the same videos and used them as "ground truths". For the ground-truth abstracts, we selected the scenes included in the scenario writer's written summary. For each episode, three ground-truth abstracts were prepared, with the duration of 5, 10, and 20 mins, respectively. To focus on the evaluation of the abstraction model, we manually corrected the errors in the shot and scene detection and the face recognition.

5.2 Evaluation Results

We have performed three different experiments.

Experiment 1: To evaluate the coverage and conciseness properties of our model, we measure the precision and recall of our abstracts compared to the ground truths. The result is summarized in Table 2. It shows high values for precision and recall and indicates that our abstraction satisfies the coverage and conciseness properties.

Table 2. Comparison of precision & recall

Video	Target Duration	# of Narrative Elements in Ground Truth	Precision	Recall
AS-1	5 min.	3	100 % (3/3)	100 % (3/3)
	10 min.	5	80 % (4/5)	80 % (4/5)
	20 min.	12	81 % (9/11)	75 % (9/12)
AS-2	5 min.	2	100 % (2/2)	100 % (2/2)
	10 min.	4	100 % (4/4)	100 % (4/4)
	20 min.	13	83 % (10/12)	76 % (10/13)
B-1	5 min.	4	75 % (3/4)	75 % (3/4)
	10 min.	6	83 % (5/6)	83 % (5/6)
	20 min.	11	83 % (10/12)	90 % (11/12)
B-2	5 min.	3	67 % (2/3)	67 % (2/3)
	10 min.	7	83 % (5/6)	71 % (5/7)
	20 min.	15	78 % (11/14)	73 % (11/15)

Experiment 2: To show that our model meets the properties of coherence and coordination, we measure the variation of the total *DoP* values, in percentage, while decreasing the target duration (Figure 5). The figure compares abstracts generated by three different methods, *i.e.,* the ground truth, our method, and the cluster-based method. (See experiment 3 for the cluster-based method we used.) The figure shows, using our model, the total *DoP* value slowly decreases with the target duration, showing that coherence and the coordination are well met.

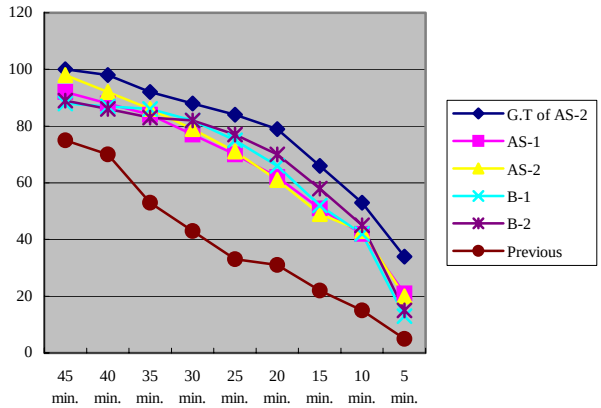


Figure 5. Variation of total *DoP* values

Experiment 3: We performed a subjective test as well. We selected a group of viewers and showed them a pair of abstracts for each episode and target duration, one generated by our method and the other by a cluster-based method similar to [1]. The cluster -based method that we implement is clustering the frames with similar visual features (e.g., color) and selecting the middle segment as representative segment of each cluster. The duration of each segment differs according the target duration. The scores in Table 3 are generated by comparing the similarity between the abstract written by a scenario writer and the one written by the test participants. We request the participants to describe the whole story after watching the abstracts.

Table 3. Comparison of story understandability

Video	Target Duration	Our Method	Cluster based Method	G.T. of AS-1
AS-1	5 min.	7.2	4.8	7.8
	10 min.	7.2	5.8	7.8
	20 min.	8.9	7.3	9.3
AS-2	5 min.	6.6	3.2	6.8
	10 min.	6.6	3.2	6.8
	20 min.	8.5	5.8	9.0
B-1	5 min.	8.5	4.2	8.9
	10 min.	8.5	4.2	8.9
	20 min.	9.5	6.7	9.5
B-2	5 min.	8.3	4.5	8.5
	10 min.	8.3	4.5	8.5
	20 min.	9.7	6.9	9.7

6. CONCLUSION & FUTURE WORK

In this paper, we presented a video abstraction model for story-oriented video. The model considers story-oriented video as a narrative formal system, which is composed of a sequence of dramatic incidents and their connections. According to this view, a story is delivered to viewers through the interactions among the characters in each dramatic incident as well as the editorial relations between dramatic incidents. For each pair of dramatic incidents, we estimate the degree of story progress achieved by them. Thus, this video model serves as a basis for the automatic generation of video abstracts. This paper provides a method to generate the abstracts which give the best understanding of the overall story under the given time limits.

It is difficult to build up a formal model which can aid the automatic abstraction of a story-oriented video. This is mainly due to the lack of semantic knowledge embedded in a story-oriented video. To overcome such difficulty, we look into the process of drama creation from scenario writing to film editing. We obtain useful clues about human understanding of a story of a video from general rules in scenario writing and editorial techniques in film production. Our initial experiments reveal that our model achieves good performance according to the four requirements of a good story-oriented video abstract, *i.e.,* coverage, conciseness coherence, and coordination.

At the current stage, the performance and precision of the proposed abstraction model are dependent on those of related technologies such as scene and face detection. Thus, an abstraction system should be built in a modular way, so that better technologies, if any, can easily be plugged into the system. In this way, the performance and precision will be continuously improved with the advance of those technologies. Another approach to overcome this limitation is to develop the system with an elaborate interaction design. The system should be designed so that users can effectively correct errors in detection with minimal interactions over the course of the whole generation process.

In this paper, we described our model in the context of scene-based abstraction of each episode. As a future work, we plan to extend the model to include other entities of video in different granularities. One direction is to extend the model to include multiple episodes of a drama. It will enable the model to provide abstracts for multiple episodes or even for an entire drama. Another direction is to refine the model to have selectivity in a more fine-grained level, e.g., in the shot level. In that way, an abstract can selectively include a whole scene or parts of a scene depending on their relative importance, giving more flexibility in meeting duration requirements.

7. ACKNOWLEDGEMENT

The authors would like to thank KBS (Korean Broadcasting System) BTRI (Broadcast Technical Research Institute) for its contributions in providing us with valuable data of video files and website access. The author, Byunghee Jung, is also a research engineer of KBS BTRI and she would like to express her gratitude to her company for the use of facilities and modules in her project at KBS BTRI.

8. REFERENCES

[1] A. Hanjalic, and H. Zhang, “An Integrated Scheme for Automated Video Abstraction Based on Unsupervised Cluster-Validity Analysis”, IEEE Transaction on Circuits and Systems for Video Technology, Vol. 9, No. 8, pp. 1280-1289, December, 1999.

[2] A. Hanjalic, R. L. Lagendijk, “Automated High-Level Movie Segmentation for Advanced Video-Retrieval Systems”, IEEE Transaction on Circuits and Systems for Video Technology, Vol. 9, No. 4, June 1999.

[3] A. Yoshitaka, T. Ishii, M. Hirakawa and T. Ichikawa, “Content-Based Retrieval of Video Data by the Grammar of Film”, IEEE , pp.310-317, 1997.

[4] B. Jung, M. Ha, H. Kim, K. Park and W. Kim, “A Component-based DCT/LDA Face Recognition Method for Character Retrieval in TV Programs”, Proceedings of International Workshop on Image Analysis for Multimedia Interactive Services, WedAmPo1, April, 2004.

[5] D. Arijon, “Grammar of The Film Language”, Silman-James Press, 1991.

[6] D. Bordwell, C. Thomson, “Film Art: An Introduction”, McGraw-Hill College, 1996.

[7] H. J. Zhang, J. H. Wu, D. Zhong, and S.W. Smoliar, "Video parsing, retrieval and browsing: an integrated and content-based solution," Proceedings of ACM Multimedia, pp. 15-24, November, 1995, New York, USA.

[8] H. Sundaram and S. F. Chang, “Computable Scenes and Structures in Films”, IEEE Transaction on Multimedia, vol. 4, no. 4, pp.482-491, December, 2002.

[9] http://access.kbs.or.kr/

[10] http://www.nada.kth.se/~viggo/wwwcompendium/node46.html

[11] J. I. Shin, W. S. Hwang, “Scenario Basic Formula”, Da-Bo Munhwa, 1987 (in Korean).

[12] J. Kupiec, J. Pedersen and F. Chen, “A Trainable Document Summarizer”, Proceedings of the 18th Annual International ACM SIGIR, pp. 68-73, July, 1995, Washington, USA.

[13] J. Nam amd A. H. Tewfik, “Video abstract of video”, IEEE Third Workshop on Multimedia Signal Processing, pp.117-122, Sep. 1999.

[14] S. Pfeiffer, R. Lienhart, S. Fischer, and W. Effelsberg, “Abstracting Digital Movies Automatically”, Journal of Visual Communication and Image Representation, Vol. 7, No. 4, December, pp. 345-353, 1996.

[15] L. He, E. Sanoki, A. Gupta, and J. Grudin, “Auto-Summarization of Audio-Visual Presentation”, Proceedings of ACM Multimedia, pp. 489-198, 1999, Orlando, Florida, United States.

[16] M. M. Yeung and B. L. Yeo, “Video visualization for compact presentation and fast browsing of pictorial content”, IEEE Transaction on Circuits and Systems for Video Technology, vol. 7, no. 5, Oct. 1997.

[17] M. M. Yeung, B. L. Yeo, W. Wolf and B. Liu, “Video Browsing using Clustering and Scene Transitions on Compressed Sequences”, Proceedings of IS&T SPIE Multimedia Computing and Networking, pp. 399-413, Feb, 1995, San Jose, California.

[18] M. Smith, T. Kanade, “Video skimming and characterization through the combination of image and language understanding technique”, Proceedings of 16th IEEE Conference on Computer Vision and Pattern Recognition, pp. 775-781, 1997.

[19] N.Vasconcelos, A.Lippman, “A Spatiotemporal Motion Model for Video Summarization”, Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Santa Barbara, 1998.