INTERNATIONAL ORGANISATION FOR STANDARDISATION
ORGANISATION INTERNATIONALE DE NORMALISATION
ISO/IEC JTC1/SC29/WG11
CODING OF MOVING PICTURES AND AUDIO INFORMATION

ISO/IEC JTC1/SC29/WG11
MPEG 99/M5481
December, 1999, Maui, USA

Title: Weight DS Extension for Affect-based Content Characterization
Source: Sony Corporation
Purpose: proposal and discussion

Authors: Yoshiaki Shibata, Yasushi Takahashi, Mikio Kamada, Ei-ichi Osawa, Hitoshi Kimura and Masayoshi Miura

Date: December 6, 1999

## 1. Introduction

In this document, we propose an extension of the Weight Description Scheme (Weight DS) so that it can treat the high-level weightings identified through our affect-based content characterization. In fact, the current definition of the Weight DS [M4866] already accommodates our proposal due to its high flexibility. However, it is too much flexible to be called a standard because it can accommodate any type of weight and weight-like functionality without strict definition. Hence we also propose a constraint on the description of the Weight DS in order to achieve an interoperability among applications based on the MPEG-7 Weight DS.

The Weight DS, which was initially introduced by Jin-Soo Lee, et al. [P674], is a component of the Generic Audio Visual Description Scheme (Generic AVDS)[N2966]. The Weight DS can be attached to Segment, Region, Event, Object, their Relation Graphs and so on, to describe the relative importance of each entity. However, the meaning of importance is ambiguous, and there have been various types of weighting or weighting-like functionality proposed so far [P674, M4866, M4867, M5003, M5101, M5200, M5267]. In order to clarify what the differences and commonalties of the proposals are, the Core/Validation Experiment (CE/VE) of the Weight DS was established in Melbourne [N2972].

The original proposal [M4866] addresses two types of weighting: One for low-level (syntactic) weighting and another one for high-level (semantic) weighting. While the original proposal addresses these two types of weightings only for descriptors, the classification is applicable to any weighting. Sanghoon Sull, et al. [M5101] also classified their fidelity (now, regarded as one of the weightings) into two categories: the quantitative fidelity that corresponds to the low-level weighting and qualitative one for high-level weighing. While there are various proposals for the low-level weighting, no proposal has been made for the high-level weighting except that a human being assigns the value arbitrarily.

Similar argument may be applied not only to the high-level weighting but also to any other high-level features. According to the MPEG-7 Requirement [N2996],

> *MPEG-7 shall support multimedia description using various types of features,*

and the subjective feature is regarded as one of the important examples. In addition, MPEG-7 Requirements also states that:

> *MPEG-7 shall support description allowing to pre-view information content in order to aid users to overcome their unfamiliarity with the structure and/or types of information, or to clarify their undecided needs,*

as one of the Functional Requirements. Although a mechanism that allows efficient browsing based on given pieces of content has been extensively studied in the Summary DS [N2969], study on how to choose pieces of content for the summary (especially for the high-level summary) is still insufficient to fulfill the requirements.

Signal processing algorithms based on sophisticated models, on the other hand, have been studied to capture the high-level content features including high-level weighting. Although some of them, such as scene detection, seem to work well in some cases, there still exists a large gap between the results obtained through such automatic tools and how an audience actually interprets and/or responds to content. Furthermore, it is unlikely in the near future that it will be possible to extract higher-level features such as how a story develops, where an audience is interested, excited, thrilled, saddened, and so on, by just analyzing content at the signal level.

This document consists of five sections and one appendix. In Section 2, our work on affect-based content characterization is briefly introduced. In Section 3, the current Weight DS is reviewed to identify the issue to be standardized. Then, a constraint on the Weight DS specification is proposed to achieve the interoperability as a standard in Section 4, followed by the summary and conclusions of our proposal in Section 5. Appendix-A supplies more information as well as concrete examples for affect-based content characterization.

## 2. Summary of Affect-based Content Characterization

In this section, we will briefly explain our work on affect-based content characterization. For more information and concrete examples, please see the Appendix-A attached to this document.

Our approach to content characterization consists of the following two techniques:

- Semantic Score Method
- Physiological Measurement and Analysis

The former is a well-defined analytical evaluation by human being where a score (called Semantic Score) is given to each pre-determined scene. The evaluation criterion is based on one proposed by Freytag [Freytag98, Laurel93] which characterizes how a story develops in terms of complication and resolution. A scene that exhibits complication (raises questions in the viewer's mind) is scored with a positive value between 0 and 1, depending on the degree of complication. On the other hand, a scene that exhibits resolution (answers to questions) is scored with a negative value between 0 and –1. The cumulative sum of the score values from the first scene to a certain scene gives the net intensity of complication for this scene which reflects context. This results in a visual representation of the story shape as shown in Figure 1. In this figure, the vertical and horizontal axes denote the net intensity of complication and time of the whole program, respectively. Note that the story shape reflects how the story develops, i.e., the development of a story.

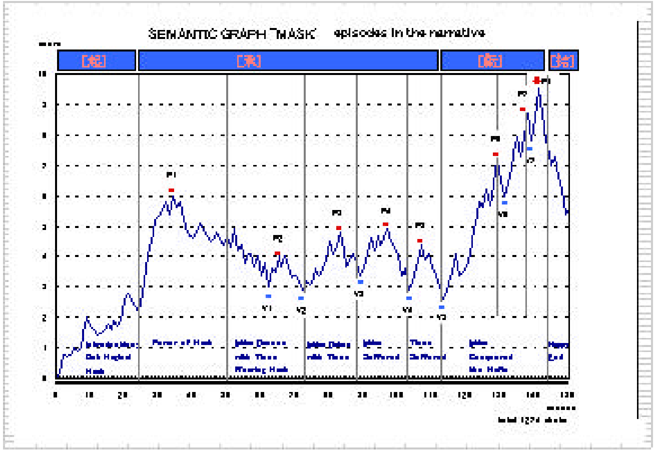


**Figure 1: The story shape based on Semantic Score Method**

While the Semantic Score Method can characterize how a story develops, it is an interpretation-based analytical approach conducted consciously. This means that it is difficult with this technique to capture fresh impressions of viewers when they watch content for the first time. Thus, another tool that captures this kind of information is necessary. We introduce physiological measurement and analysis for that purpose. We have conducted experiments that measured various kinds of physiological responses from an audience watching a video. Our experimental results show that there are many promising physiological responses that indicate where an audience feels interested and/or excited. In addition, some responses reflect specific emotions such as happiness, anger, sadness, and so on. Therefore, physiological data, after appropriate post processing, can be used to identify how content influences an audience emotionally.

In both cases, the result is expected to be a weighting value attached to predefined video segment. An attribute to specify the type of weighting such as "StoryComplication" for the Semantic Score Method, "Interested", "Excited", "Happy", "Angry", and so on, for the Physiological Measurement and Analysis, is also required. Figure 2 schematically shows the weighting flow for the type of weighting as "StoryComplication".

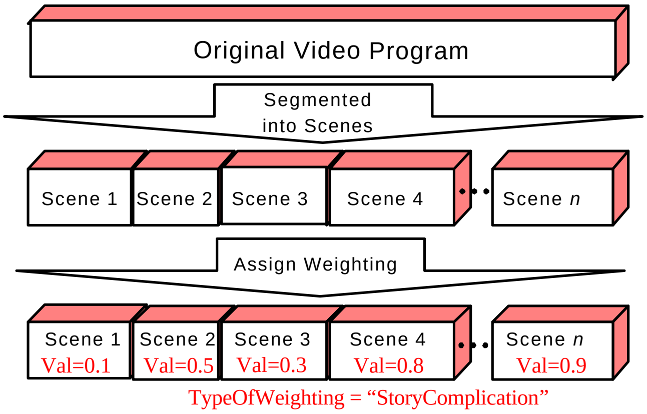


**Figure 2: Scene weighting based on Semantic Score Method**

Note that these techniques provide valuable information that cannot be obtained through conventional signal processing based technology. Using the content characterizing data expressed as the weighting of video segments, various kinds of high-level video summaries as well as video retrievals are realized. For example, we can generate an outline-type video summary that reflects the development of a story. We can also retrieve a part of video in terms of interest, excitation, and various specific emotions.

## 3. Review of the Current Weight DS

While the current Weight DS is under Core/Validation Experiment and its Description Definition Language (DDL) definition is still not fixed, the latest DDL description of the Weight DS available [M5005] is:

```
<!-- ################################################ --!>
<!-- Definition of Weight DS as Primitive DS          --!>
<!-- ################################################ --!>

<DescType name=“Reference”>
       <attrDecl name=“idref” required=“true”>
       <datatypeRef name=“IDREF”/> </attrDecl>
       <empty/>
<DescType>

<DSType name=“WeightValueDS”>
       <attrDecl name=“value” required=“true”> <datatypeRef name=“real”/>
</attrDecl>
       <DescTypeRef name=“Reference”/>
</DSType>

<DSType name=“WeightDS”>
       <attrDecl name=“id”> <datatypeRef name=“ID”/> </attrDecl>
       <attrDecl name=“type”> <datatypeRef name=“string”/> </attrDecl>
       <attrDecl name=“reliability”> <datatypeRef name=“real”/> </attrDecl>
       <DSTypeRef name=“WeightValueDS” minOccur =“1” maxOccur=“*”/>
</DSType>
```

In this definition, “Reference” denotes a descriptor that references the target entity to be weighted. “WeightValueDS” is a DS with which the weighting value is described. “type” provides the semantics of the weighting value (called the weight type, hereafter), and “reliability” is used to describe how reliable the weighting values are.

The Weight DS defined above can be applied to any kind of weighting because the weight type is defined as a string. This means that any arbitrary text description to specify the semantics of the weighting may be given. In this sense, our proposals introduced in Section 2 can be also accommodated within the existing Weight DS.

However, because of its high flexibility, it is likely to occur that the weight type specification and its interpretation will vary among applications and thus fail to achieve interoperability, which is one of the main goals of MPEG-7. For example, suppose that an application, say Application-A, specifies the weight type as “importance” and gives weighting values to its certain entities according to their relative importance, there is no guarantee that another application, Application-B, exactly interprets the weight type “importance” as Application-A expects. Furthermore, even though Application-A and Application-B are in the same application category, say image retrieval, the weighting value optimized for Application-A is unlikely the same as the optimized value for Application-B especially when image retrieval algorithms adopted in Application-A and Application-B are different. Therefore, only defining the weight type as a string is not sufficient to be standardized, and a mechanism, which will result in imposing a certain constraint on the textual description for the weight type, should be considered.

## 4. Proposal – Structured Weight Type Description

In this section, we will propose one possible mechanism for weight type description in order to achieve the interoperability: organizing weights into a taxonomy.

### 4.1. Weight Type Taxonomy for Subjective Weighting

In Section 2, we propose using weight types such as “StoryComplication”, “Interested”, “Excited”, “Happy”, and “Angry”, to accommodate our content characterization. The analysis of these weight types indicates that all the weight types can be grouped into one category: “Subjective”. Also, in our content characterization, the weight type “StoryComplication”, for which the Semantic Score is one of the specific candidates, is classified into “Analysis” category while the rest of weight types fall into either “Mood” or “Emotion” category. Consequently all the weight types can be classified so as to form a hierarchical structure.

Strictly speaking, however, even though the weight type “StoryComplication” is refined by “SemanticScore”, this weight type description is not sufficient to specify all relevant information on the weighting values attached to video segments. As is explained in Section 2, the Semantic Score is used to depict the story shape in a video, and thus if evaluators or group of evaluators with different cultural background are asked to evaluate a video, then the resulting story shape would not be the same. Furthermore, the difference will be larger when the weighting in the “Mood” and “Emotion” categories are performed. For example, it often happens that a comedy, which is amusing for Americans, does not always make Japanese laugh. In such a case, the weighting value obtained from each group of evaluators will show a large discrepancy[1]. Hence, an application which analyzes, e.g., the Semantic Score itself demands information on “Who’s Semantic Score?” because the information of evaluators profile often plays a crucial role in such an analysis.

The proposed weight type taxonomy is illustrated in Figure 3. Note that “Arbitrary” is added to “Analysis” category because, as is often suggested [M4866, M5101], the weighting value arbitrarily assigned by specific people such a creator should be taken into account. “FilmRating” is also added to “Analysis” so that the film rating can be conducted for the video segment by assigning a pre-determined weighting value to each segment. On the other hand, human fundamental moods as well as emotions are listed up and tentatively classified into two categories: “Mood” and “Emotion”, though further consideration regarding, e.g., appropriate classification, their extractability, and so on, is still necessary.

The dotted lines in Figure 3 imply that the weight type can be further refined into sub-categories according to its additional information. For example, when information of evaluator or the group of evaluators is taken into consideration, “SemanticScore” may be further classified into the one by a certain individual (e.g. Yoshi Shibata), by evaluators of a certain gender and ages (e.g. male & thirties), or by evaluators of a certain nation (Japanese)

[1] This does not mean these weighting values are useless. On the contrary, the data will imply valuable information on the question, “How they are different?”. One of the goals of our content characterization is to provide analytical tools for such an affect-based study.

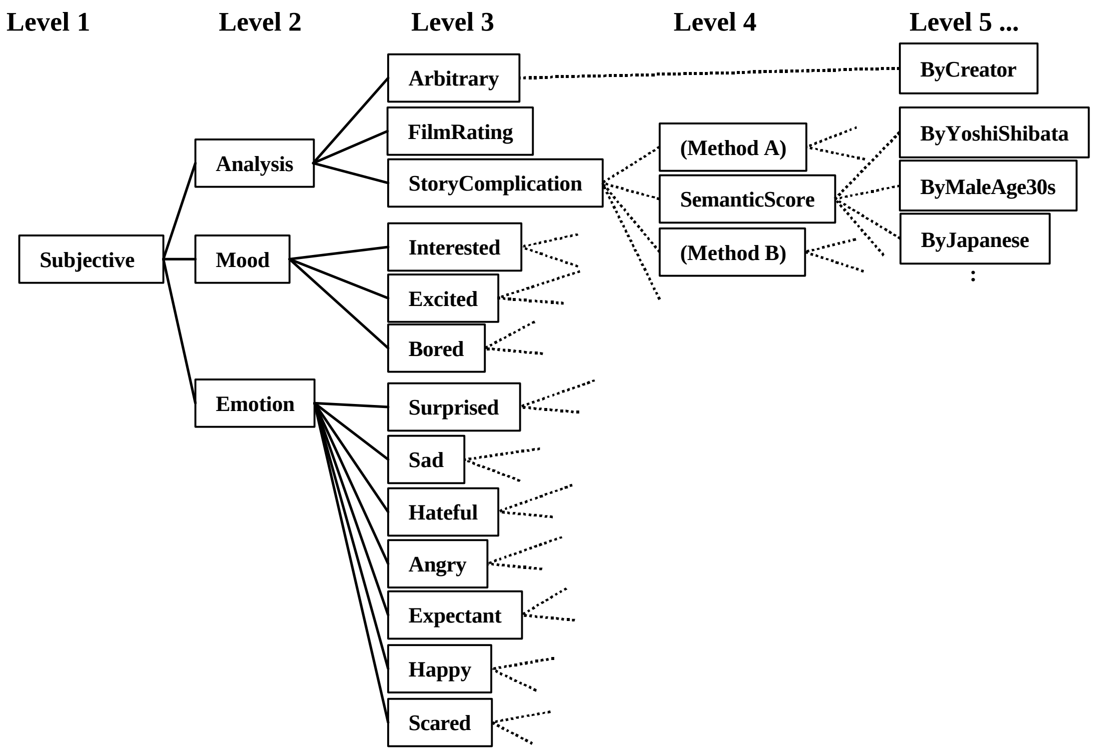


**Figure 3: Weight type taxonomy for subjective weighting**

## 4.2. Weight Type Description for Interoperability

Based on the taxonomy developed above, the proposed constraint imposed on the weight type description is an expression, which reflects the hierarchy in the taxonomy. For example:

```
weightType=”Subjective.Analysis.StoryComplication.xxx.yyy…”;
```

With this expression, if “xxx.yyy…” includes all the relevant information for the weighting value without ambiguity, then the interoperability among various applications is perfectly realized. In reality, however, this is not feasible, because the additional information to be concatenated as “xxx.yyy…” strongly depends on applications and/or purpose of applications and thus it would be impractical to identify all the possible additional information relevant to the weighting.

Note, however, that once the weight type description is expressed in such a fashion, an application can use the specific functionality of the weighting according to its level of *knowledge,* i.e., the level of depth in the taxonomy. For example, an application, which can understand the weight type of level four in the taxonomy or “SemanticScore”, will enjoy the “SemanticScore” functionality regardless of its extraction details. Furthermore, if an application can recognize only the first level of the taxonomy or “Subjective”, then the application still utilize the weighting functionality as a relative significance of video segment from a subjective point of view[2].

MPEG-7 should standardize up to a certain level of such taxonomy as reserved keywords for the weight type description in order to achieve interoperability. In addition, MPEG-7 should also allow specific applications to share detailed description like “SemanticScore.ByYoshiShibata” to keep interoperability within the applications. However, which level of taxonomy is best to be standardized in MPEG-7 is open to discussion. In the case of our subjective weighting taxonomy in Figure 3, our current expectation for the standardization is the level three, which are indicated by solid lines in the figure. But, further study is needed to identify the most appropriate level of the taxonomy for the standardization.

[2] This adds to additional requirement for the taxonomy construction, i.e., the lower level of the weight type in the taxonomy should be a refinement of the higher one so that an application can *approximately* use the higher level weight type when detailed information is not necessary.

### 4.3. Weight Type Taxonomy for Fidelity

The proposed approach is applicable to any weighting functionality when the taxonomy of the functionality is carefully analyzed. Taking fidelity as another example, the weight type description with proposed constraint is demonstrated.

Fidelity, proposed by Sanghoon Sull, et al. [M5101], is the notion of how well a representative item describes a given set of items. As they mainly focus on the fidelity for key frame in the key frame hierarchical tree structure used for multi-level video abstraction, we also discuss the fidelity for key frame. In this application, the fidelity value is attached to each node in the tree-structure to indicate how well a key frame at the node represent all the frames in video segment assigned to the node. Those fidelity values are used to allow efficient searching or browsing for a video content. Note that although the fidelity was originally proposed as one of attributes of HighlightLevel DS, its functionality is currently regarded as a kind of weighting [N2972] and the Weight DS accommodates fidelity because an arbitrary textual description can be specified for the weight type.

According to Sanghoon Sull, et al., the fidelity is divided into two categories: qualitative fidelity and quantitative fidelity. While the former is assigned by a human being, the latter is a value obtained through similarity measurement between relevant frames. As one of the example of the quantitative fidelity, they took an inverse of maximum frame distance. However, there is no reason that inverse of *minimum* frame distance or inverse of *average* frame distance cannot be used for the fidelity. In other words, the quantitative fidelity can be divided into those three categories. Furthermore, since there are various metrics for the distance between frames such as L1 norm, L2 norm and so on, each category can be further divided into more specific ones according to the distance metric it adopts.

Based on above analysis, the weight type taxonomy for the fidelity is illustrated in Figure 4.

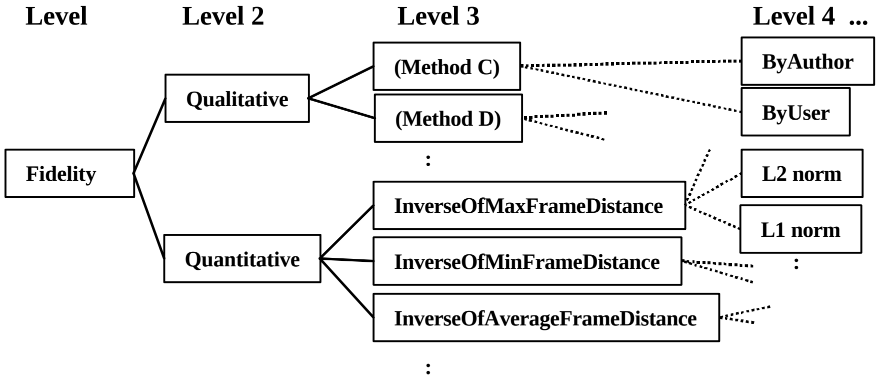


**Figure 4: Weight type taxonomy for fidelity**

The weight type description for the fidelity is then expressed, for example, as

```
weightType=”Fidelity.Quantitative.InveerseOfMaxFrameDistance.L2norm”
```

to specify its functionality without ambiguity. Note that, for an application that only needs relative magnitude of the fidelity values, the weight type description with “Fidelity” might be sufficient. However, for other applications that treat the fidelity value quantitatively, more specific description of the weight type for the fidelity is inevitable. For example, an application that needs to judge whether the fidelity of a given key frame is high or low cannot determine the fidelity threshold without detailed information.

Note that the same argument will be applied to the fidelity defined as an attribute of HighlightLevel DS. In this case, a technique for the weight type description explained above cannot be used because there is no *fidelity type* specification. To achieve interoperability in the quantitative level, therefore, MPEG-7 should standardize the definition of the fidelity with its extraction algorithm or with its strict property to be preserved at its extraction and usage[3].

## 5. Summary and Conclusions

In this document, we propose an extension of the Weight DS to treat the high-level weightings obtained through the affect-based content characterization. The current definition of the Weight DS already accommodates the high-level weighting due to its high flexibility. However, it is too flexible to be called a standard because any type of weighting functionality can be specified without strict definition. Therefore, a constraint for the weight type specification is proposed in order to achieve interoperability. The constraint consists of the weight type taxonomy and the weight type description that reflects the taxonomy. Although MPEG-7 should standardize up to a certain level of the taxonomy as reserved keywords for the weight type description, the level of taxonomy to be standardized is open to discussion. Furthermore, all the weight or weight-like functionality should be carefully analyzed to determine the taxonomy necessary to realize interoperability.

## 6. References

**[P674]** Jin Soo Lee, Heon Jun Kim, Jung Min Song, Sung Bae Jun and Hee Youn Lee, “Description Scheme for Image Retrieval System using Relevance Feedback”, MPEG-7 Proposal P674 at Lancaster, Feb. (1999)

**[N2966]** Multimedia Description Scheme Group, “MPEG-7 Generic AV Description Schemes ver. 0.6”, ISO/IEC JTC1/SC29/WG11 MPEG99/N2966, Melbourne, Oct. (1999)

**[M4866]** Jin-Soo Lee, Heon-Jun Kim, Kyoungro Yoon and Hee-Yeon Lee, “Weight DS in Generic Description Scheme for MPEG-7”, ISO/IEC JTC1/SC29/WG11 MPEG99/M4866, Vancouver, July. (1999)

**[M4867]** Jin-Soo Lee, Heon-Jun Kim, Kyoungro Yoon and Hee-Yeon Lee, “Automatic Weight Learning in Relevance Feedback and Weight Updating Method using Similarity Score Normalization”, ISO/IEC JTC1/SC29/WG11 MPEG99/M4867, Vancouver, July. (1999)

**[M5003]** S. Sekiguchi, Y.Isu, H.Nishikawa and K.Asai, “Proposal of the Description Usage DS”, ISO/IEC JTC1/SC29/WG11 MPEG99/M5003, Melbourne, Oct. (1999)

**[M5101]** Sanghoon Sull, Jung-Rim Kim and Yuman Kim, “Efficient and effective search and browsing using fidelity”, ISO/IEC JTC1/SC29/WG11 MPEG99/M5101, Melbourne, Oct. (1999)

**[M5200]** Koichi Emura and Toshihiko Munetsugu, “Proposal of Pointofview DS”, ISO/IEC JTC1/SC29/WG11 MPEG99/M5200, Melbourne, Oct. (1999)

**[M5267]** Riccardo Leonardi, Giovanni Paltenghi and Lorenzo Rossi, “Limitations of the MPEG-7 Generic DS: Reorganizing the Syntactic/Semantic DS’s”, ISO/IEC JTC1/SC29/WG11 MPEG99/M5267, Melbourne, Oct. (1999)

**[N2972]** Jin-Soo Lee, Sanghun Sull, S.Sekiguchi, Koichi Emura, Riccardo and Leonardi, “MPEG-7 Core/Validation Experiment on the Weight DS”, ISO/IEC JTC1/SC29/WG11 MPEG99/N2972 Melbourne, Oct. (1999)

**[N2972]** Jin-Soo Lee, Sanghun Sull, S.Sekiguchi, Koichi Emura, Riccardo and Leonardi, “MPEG-7 Core/Validation Experiment on the Weight DS”, ISO/IEC JTC1/SC29/WG11 MPEG99/N2972 Melbourne, Oct. (1999)

**[N2996]** Requirements Group, “MPEG-7 Requirements Document V.10”, ISO/IEC JTC1/SC29/WG11 MPEG99/N2996 Melbourne, Oct. (1999)

**[N2969]** MDS Group, “Validation of the MPEG-7 Summary DS”, ISO/IEC JTC1/SC29/WG11 MPEG99/N2969 Melbourne, Oct. (1999)

**[Freytag98]** Gustav Freytag, “Technique of the Drama”, 2ed, ed. Translated by Elias J. MacEwan, Chicago: Scott, Foresman, 1898.

**[Laurel93]** Brenda Laurel, “Computers as Theatre”, Addison-Wesley, 1993.

**[M5005]** Jin-Soo Lee, Jungmin Song, Heon-Jun Kim, Kyoungro Yoon and Hee-Yeon Lee, “Weight Description Scheme in Generic Audio Visual Description Scheme”, ISO/IEC JTC1/SC29/WG11 MPEG99/M5005, Melbourne, Oct. (1999)

[3] One of possible properties to be imposed on the fidelity value might be linearity. For example, assign value 0 for minimum or no fidelity, value 1 for maximum or perfect fidelity, and guarantee that the fidelity of value 2*x* is twice as faithful as the fidelity of value *x* for all *x* between 0 and 0.5.

## Appendix-A

# Affect-based Content Characterization

## A.1. Introduction

As digital content distribution has become popular with the increasing number of channels together with a drastic expansion of capacity in home storage devices, technology for efficient search and retrieval of desired content is in high demands. To meet this demand, various technologies based on signal processing analysis of digital bit stream have been extensively studied [Kanade96].

What a user wants to retrieve, on the other hand, depends strongly on his/her situation. For visual object retrieval based on its attributes such as shape, color and texture, signal processing based technology may be able to meet the user's need. However, there is another kind of higher level and widely anticipated request such as "I want to watch the highlight scene of this program", "where is the most interesting part of this movie" and so on. Unfortunately, it is unlikely in the near future that it will be possible to extract such high-level features as where in a content an audience is interested, excited, thrilled, laughing, and so on, by just analyzing content at the signal level.

In addition, if we need a video summary from which we can grasp the essence of the content, then signal processing based technology often encounters problems. Though some of the video summarization steps, such as scene identification, can be realized using sophisticated algorithms, more difficult and important process exists in the next step, i.e., which scene is to be chosen for the video summary. This is possible only when the information on how a story in a video develops is fully understood, which is also likely to unattainable using only signal processing based technology.

In order to describe this sort of high-level feature, text annotation attached to a video segment is often suggested [Kim96]. Because of its high flexibility, the method remains a supplement tool for a content description. However, there are obvious drawbacks to this method:

- It is an extra and time-consuming task to annotate each video segment, especially for old content already created and stored in an archive,
- Sophisticated natural language processing is required to extract the high-level feature from text annotation.

In short, the drawbacks of text annotation are its insufficient machine friendliness.

Another possible solution might be letting an author and/or a creator give arbitrary an "weighting value" to each video segment indicating its relative importance. The word "importance" is, however, quite ambiguous: an important video segment in one context often becomes unimportant in another context. Moreover, there is a risk that an author's conscious scoring largely deviates from what an audience interprets and/or feels.

On the basis of above considerations, we have studied content characterization based on affective information. Since all content is eventually *consumed* by human beings in the form of being watched, listened, felt, and perceived, the reaction and response of an audience are valuable information to describe the content. In our current work, we mainly focus on video characterization especially for storied video such as movies. But similar techniques can be applied to other kind of content such as TV drama, sports, music and video games.

In this appendix, an attempt is made to characterize a video based on affective information. Our approach to video characterization consists of the following two techniques:

- Semantic Score Method
- Physiological Measurement and Analysis

The former is a well-defined subjective evaluation of video by human being based on Freytag's theory [Freytag98, Laurel 93]. We also have conducted experiments where we measured various kinds of physiological responses from an audience watching a video. Although the analysis of physiological data is still in progress, we found that many physiological responses can be used to detect where in a video an audience feels interested and/or excited.

In addition, some physiological responses may even indicate the emotions of the audience during video appreciation.

This appendix is composed of five sections. In Section A.2, the Semantic Score Method is described together with some concrete examples of its application. The preliminary results of the physiological measurement and analysis are then introduced in Section A.3. In Section A.4, we will address some expected arguments against our approach as discussion, followed by the summary and conclusions of this appendix in Section A.5.

## A.2. Semantic Score Method

In this Section, the Semantic Score Method, a new technique to characterize video content based on subjective evaluation, is introduced.

### A.2.1. Freytag's Triangle

Gustav Freytag, a German critic and playwright, suggested in 1863 that the action of a play could be represented graphically when the pattern of emotional tension created in its audience was evaluated. Since tension typically rises during the course of a play until the climax of the action and falls thereafter, he modeled this pattern in a triangle form, referred to as "Freytag's Triangle" shown in Figure A.1. In this figure, the left side of the triangle indicates the rising action that leads up to a climax or turning point while the right side of the triangle is the falling action that happens from the climax to the conclusion. The horizontal axis of the graph is time; the vertical axis is *complication*. According to Brenda Laurel [Laurel93], the complication axis of the Freytag's triangle represents the *informational attributes* of each dramatic incident. An incident that raises questions is part of the rising action, which increases *complication*, while one that answers questions is part of falling action, resulting in decreasing the complication, i.e., *resolution*.

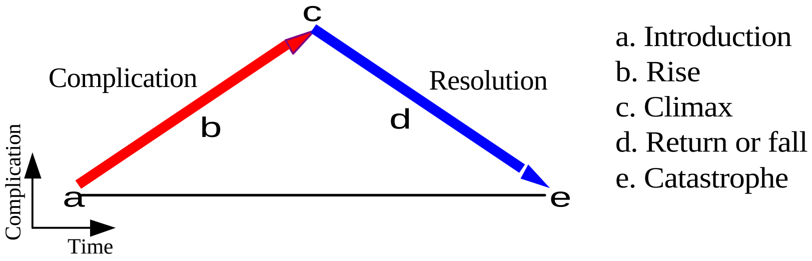


**Figure A.1: Freytag's triangle [Laurel93]**

In reality, however, things are more complicated than in Freytag's idealized model. One dramatic incident may raise some questions and answer others simultaneously. Hence, the *degree of complication* is introduced to specify the net increase of complication caused by an incident. The degree of complication is represented by a positive value when complication caused by an incident's raising questions overwhelms resolution by answering questions and a negative value vice versa. The *cumulative complication* for an incident is defined as the cumulative sum of the degree of complication for all incidents preceding this incident, representing the net intensity of complication for this incident. Note also that because of a fractal-like property of a play where whole story is composed of several sub-stories that can be further divided into various dramatic incidents, the shape of a practical play is characterized in more irregular and jagged form than shown in Figure A.1.

In order to help readers' understanding, we reproduce an example from [Laurel93]. Assume the following background situation: a group of strangers have been invited by an anonymous person to spend the weekend in a remote mansion. During the night, one member of the group (Brown) has disappeared. Some of the remaining characters are gathered in the drawing room expressing concern and alarm. The butler (James) enters and announces that Brown has been found. The following are conversations made among those people.

**James:** I'm afraid I have some rather shocking news.
**Smith:** Spit it out, man.
**Nancy:** Yes, can't you see my nerves are absolutely shot? If you have any information at all, you must give it to us at once.

**James:** It's about Mr. Brown.
**Smith:** Well?
**James:** We've just found him on the beach.
**Smith:** Thank heavens. Then he's all right.
**James:** I'm afraid not, sir.
**Smith:** What's that?
**James:** Actually, he's quite dead, sir.
**Nancy:** Good God! What happened?
**James:** He appears to have drowned.
**Smith:** That's absurd, man. Brown was a first-class swimmer.

The informational components raised in the above dialog are summarized as:

**a.** James has shocking news.
**b.** The news concerns Brown.
**c.** Brown has been found.
**d.** Brown is dead.
**e.** Brown has drowned.
**f.** Brown was a good swimmer.

Then, each component is evaluated based on the degree of complication (between 0 and +/-1). Possible scoring result is shown in Table A.1.

| Informational Component | Degree of Complication | Cumulative Complication |
|---|---|---|
| **a.** James has shocking news. | +0.4 | 0.4 |
| **b.** The news concerns Brown. | +0.5 | 0.9 |
| **c.** Brown has been found. | -0.7 | 0.2 |
| **d.** Brown is dead. | +0.9 | 1.1 |
| **e.** Brown has drowned. | -0.4 | 0.7 |
| **f.** Brown was a good swimmer. | +0.8 | 1.5 |

**Table A.1: Complication/Resolution based evaluation**

In this table, the component **c** and **e** are evaluated as *negative* complication (resolution). The former provides an answer to the puzzle that "Brown had disappeared", while the latter gives an answer to the question that "how Brown died" raised in the component **d**. The third column in the table denotes the cumulative sum of the degree of complication from the component **a**. Assume that each component in the table is a dramatic incident occurring sequentially. Then, since the degree of complication at each incident indicates the increase of complication at each incident, the cumulative complication in the table reflects the net complication at each moment resulting from preceding incidents since the initial one. The cumulative complication is then used to visualize the story shape for the dialog as shown in Figure A.2.

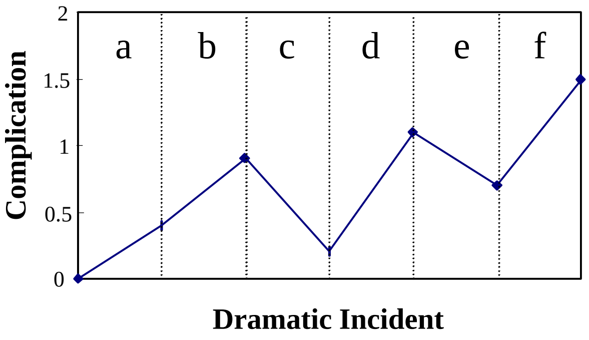


**Figure A.2: The story shape for the dialog example**

### A.2.2. Semantic Score Method

Based on the Freytag's play analysis, we have developed a subjective evaluation method for storied video named as the "Semantic Score Method". According to Brenda Laurel [Laurel93], an implicit assumption was made in the Freytag analysis that there is a direct relationship between what we *know* about the dramatic incident and how we *feel* about it. Our method, however, mainly focuses on the former aspect, i.e., we have developed this method as an analytical tool for a video characterization with *interpretation-based* subjective evaluation. In short, the evaluators are asked to give a positive (negative) value to each pre-determined scene according to the degree of complication (resolution). The evaluators are expected to *interpret* what happens in the scene and *analyze* dramatic incidents involved in the scene to represent the scene with a single value (which we call the Semantic Score) from the complication/resolution viewpoint.

In order to obtain reliable data from the general audiences, we have developed the following items as supporting tools of the method:

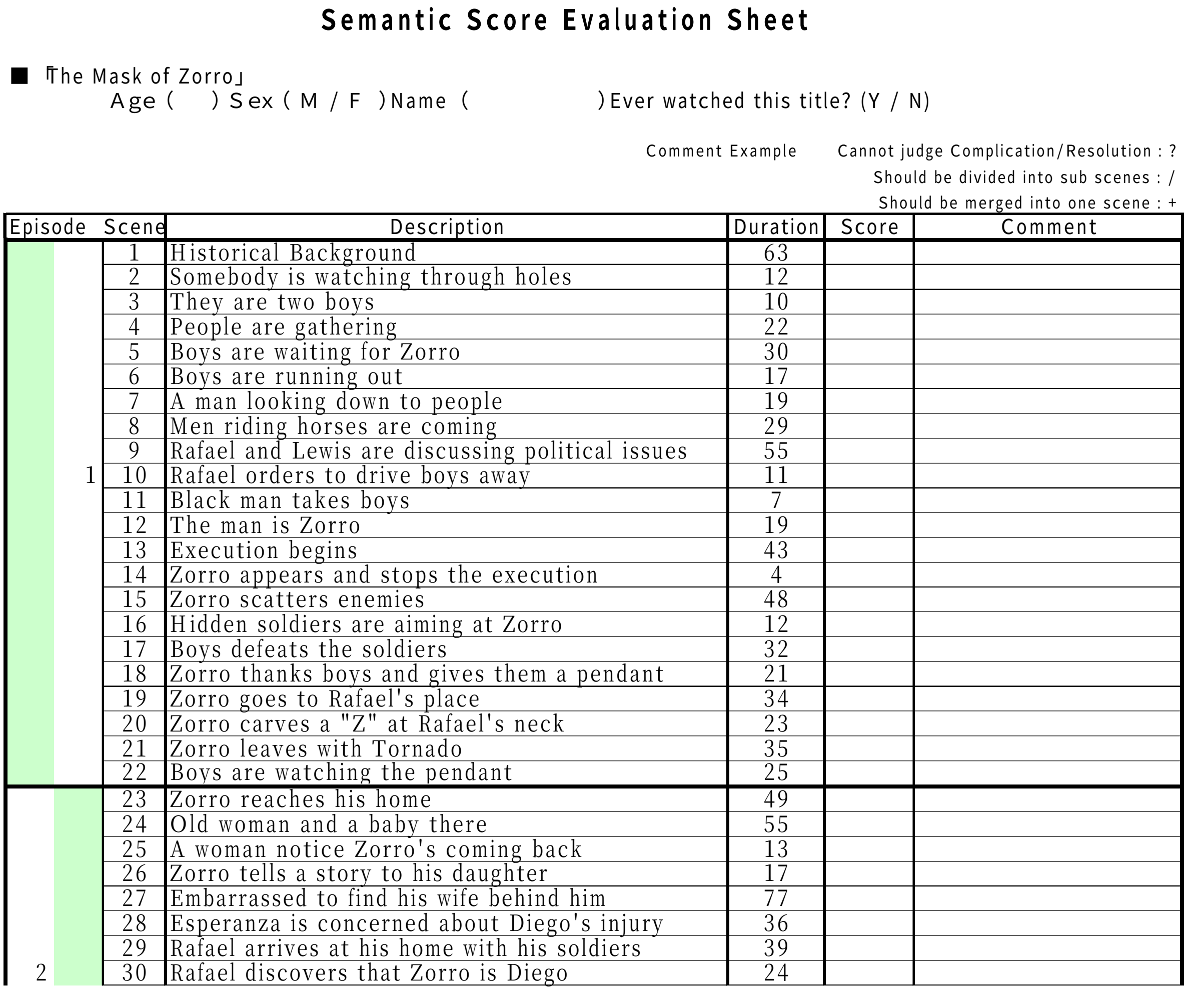
1. Instruction video and booklet
2. Test material (target movie)
3. Specially designed score sheet

The instruction video and booklet are provided to explain the purpose of the evaluation, the evaluation procedure and the evaluation criterion (the complication and the resolution). The instruction also includes a concrete evaluation example: using a short storied video composed of seven scenes, we demonstrate how each scene should be scored step by step. This example is expected to provide evaluators with a common yardstick with which how a certain scene is to be scored. The score is assigned within a range between –5 and 5 by steps of one. If necessary, however, it is allowed to score a scene with a fractional value as well.

The test material is a movie whose story shape is to be characterized. Since evaluators are asked to score scenes one by one, the video is slightly modified from its original form, i.e., the end of each scene is marked by showing the final frame as a still for a few seconds, superimposed with its scene number. Using this test material, evaluators can recognize each scene easily, resulting in smooth evaluation of the scenes.

One of the controversial issues is the scene definition. In our method, we define a scene as a video segment that has minimum semantics as a story component with monotonic complication or resolution. The boundary between scenes is identified when a situation is drastically changed. Here, the situation includes time, place, character, context (e.g. in dialog), particular dramatic incident, and so on. This implies that one scene may be composed of several shots or that a long shot may be divided into several scenes. For example, when a long video shot has both question raising and answering sequentially, the shot should be divided into two concatenated scenes. According to our scene detection conducted for twenty movies in five genres, one movie is typically divided into 100 - 250 scenes, resulting in each scene lasting 30 – 60 seconds. The scene length depends on its genre: an action type movie tends to have shorter scenes while the one regarded as a love story tends to have longer scenes than other genres.

In addition, a specially designed score sheet is provided to record the scores the evaluators assigns, a part of which is shown in Figure A.3. In this score sheet, each row corresponds to a scene, which is composed of the scene number, a short scene description, duration of the scene, and a cell to be filled with the complication value. Supplemental information is provided for the sake of evaluators' convenience, i.e., evaluators can easily recognize where they are evaluating at any moment. In addition, several consecutive scenes are grouped to form an episode, the second level story component that can be identified without ambiguity. In this score sheet, the boundary of the episode is represented with a thick solid line. Although the yardstick evaluators keep in mind may vary during the evaluation, a request to keep the consistency of scoring within whole video is often hard to achieve. Evaluators are asked, therefore, to at least keep the consistency within the episode.



Semantic Score Evaluation Sheet

■ 「The Mask of Zorro」
Age (  ) Sex ( M / F ) Name (  ) Ever watched this title? (Y / N)

Comment Example Cannot judge Complication/Resolution : ?
Should be divided into sub scenes : /
Should be merged into one scene : +

| Episode | Scene | Description | Duration | Score | Comment |
|---|---|---|---|---|---|
| 1 | 1 | Historical Background | 63 | | |
| | 2 | Somebody is watching through holes | 12 | | |
| | 3 | They are two boys | 10 | | |
| | 4 | People are gathering | 22 | | |
| | 5 | Boys are waiting for Zorro | 30 | | |
| | 6 | Boys are running out | 17 | | |
| | 7 | A man looking down to people | 19 | | |
| | 8 | Men riding horses are coming | 29 | | |
| | 9 | Rafael and Lewis are discussing political issues | 55 | | |
| | 10 | Rafael orders to drive boys away | 11 | | |
| | 11 | Black man takes boys | 7 | | |
| | 12 | The man is Zorro | 19 | | |
| | 13 | Execution begins | 43 | | |
| | 14 | Zorro appears and stops the execution | 4 | | |
| | 15 | Zorro scatters enemies | 48 | | |
| | 16 | Hidden soldiers are aiming at Zorro | 12 | | |
| | 17 | Boys defeats the soldiers | 32 | | |
| | 18 | Zorro thanks boys and gives them a pendant | 21 | | |
| | 19 | Zorro goes to Rafael's place | 34 | | |
| | 20 | Zorro carves a "Z" at Rafael's neck | 23 | | |
| | 21 | Zorro leaves with Tornado | 35 | | |
| | 22 | Boys are watching the pendant | 25 | | |
| 2 | 23 | Zorro reaches his home | 49 | | |
| | 24 | Old woman and a baby there | 55 | | |
| | 25 | A woman notice Zorro's coming back | 13 | | |
| | 26 | Zorro tells a story to his daughter | 17 | | |
| | 27 | Embarrassed to find his wife behind him | 77 | | |
| | 28 | Esperanza is concerned about Diego's injury | 36 | | |
| | 29 | Rafael arrives at his home with his soldiers | 39 | | |
| | 30 | Rafael discovers that Zorro is Diego | 24 | | |

**Figure A.3: Score Sheet for the Semantic Score Method (originally in Japanese)**

## A.2.3. Video Characterization using Semantic Score Method

### A.2.3.1. Evaluation Procedure

An experiment for video characterization based on the Semantic Score Method has been conducted. As the Test Materials, we have used the following three commercial movie titles[4],

“THE MASK OF ZORRO”,

“AWAKENINGS”

“IT COULD HAPPEN TO YOU”.

Twelve university students from two universities were asked to be evaluators for each title. Thus, a total of 36 evaluators were involved in the experiment.

The evaluation procedure adopted in the experiment is summarized as follows:

1. Instruction was given to the evaluators using the instruction video and booklet
2. Evaluators completed watching the designated title in a normal fashion.
3. Rewatched the test material and evaluated using the score sheet after reviewing the instructions.
4. Evaluators were asked to answer some questionnaires and interviewed.

[4] Officially permitted to use these titles as test materials for this work by Sony Pictures Entertainment.

It should be noted that in step 2 all evaluators were asked to watch the assigned title before actual evaluation in step 3. The reason for this is to let evaluators know the content in advance so that evaluators can evaluate the title calmly. What the method aimed at is not an identification of exciting part of a video where evaluators might lose themselves but to characterize how a story in the title develops. Thus, excitation caused by unexpected development of a story and/or audiovisual effects should be carefully controlled. In other words, if evaluators were really excited with watching the title, they might even forget the evaluation itself.

It is obvious, on the other hand, that a fresh impression on first viewing provides valuable information for other aspects of video characterization. An attempt to treat these aspects are made based on physiological measurement and analysis, which will be discussed in Section A.3.

### A.2.3.2. Semantic Graph

The Semantic Score, the movie characterizing data directly obtained through the evaluation, were integrated to form the story shape, which we call "Semantic Graph". Figure A.4 shows the Semantic Graph of "THE MASK OF ZORRO" for four evaluators.

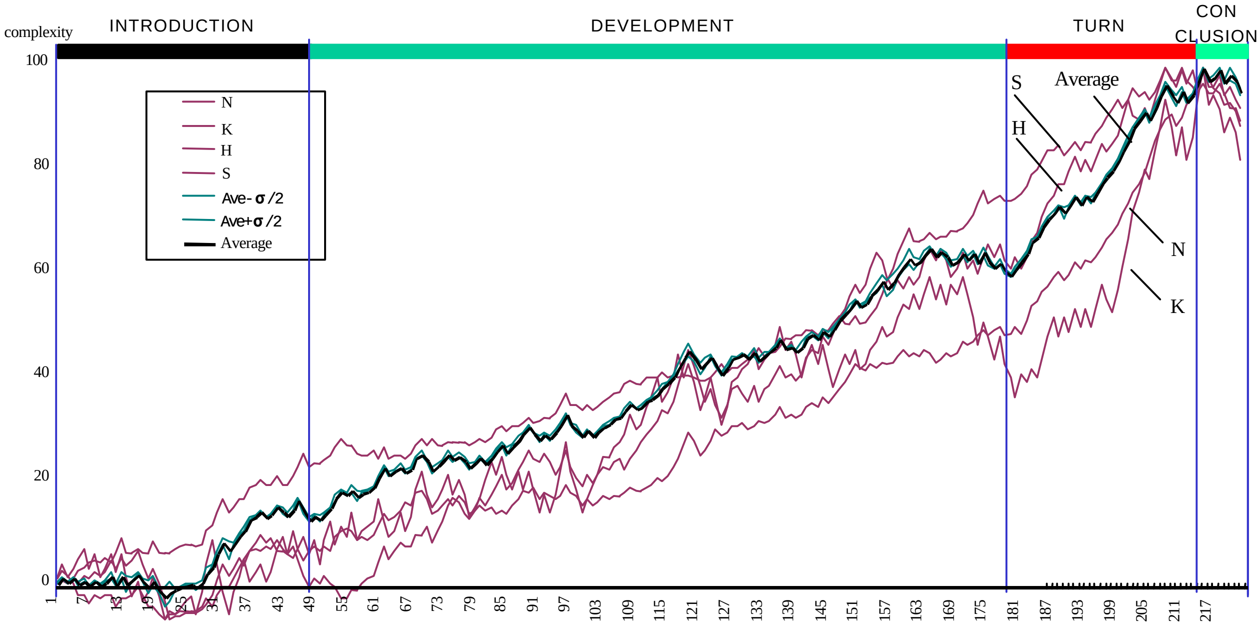


**Figure A.4: Semantic Graph of "THE MASK OF ZORRO"**

In Figure A.4, the vertical axis denotes complication while the horizontal axis is the scene number (instead of time of the whole movie). All data are normalized with repsect to their maximum peak value so that direct comparison among them is available. It is also noted that a whole story is divided into four regions based on a conventional Japanese story model known as Ki-Sho-Ten-Ketsu. According to the model, a story is composed of four fundamental parts called Ki, Sho, Ten and Ketsu which corresponds to Introduction, Development, Turn, and Conclusion, respectively.

Although six evaluators were asked to evaluate the title at this round, two evaluator's results were excluded from this graph because they deviate greatly from others from a statistical point of view. We had an interview with one of them and found that the evaluator misunderstood the criteria for complication/resolution. Furthermore, the evaluator was found to dislike action type movies and had no interest on the title at all.

Nevertheless, the Semantic Graphs obtained from the rest of four evaluators agree well with each other qualitatively. Although the *dynamic range* of the Semantic Score seems to depend strongly on the evaluator's personality, local salient structures in particular scenes such as scene 121, which represents strong complication followed by resolution, were identified regardless of evaluators.

The thick line in the graph was obtained by combining the rest of four Semantic Graphs. Note that a simple averaging operation does not work well in this analysis because there are cases where a scene is scored with both high positive and negative values. Although the case clearly suggests that evaluators recognize something in the scene, a simple averaging operation may diminish the information.

Since all four evaluators are students in the same department of the university and their ages are almost the same, they are expected to have similar backgrounds. Hence, the thick line in the graph represents the *average interpretation* of the title among them. This implies that when the evaluations are conducted by groups composed of evaluators having similar backgrounds such as gender, age, culture, living environment, and so on, the discrepancy of the Semantic Graphs among the groups reflects the difference of movie interpretation caused by different backgrounds. In other words, how audiences with a certain profile interpret a movie can be identified by careful selection of evaluators according to their background.

### A.2.4. Applications – Trailer-type Video Summary

The Semantic Score Method itself may be regarded as an analytical tool for video evaluation. However, once a video is characterized using the Semantic Score, there are many potential applications based on the video characterizing data. We, as an initial trial, have applied the method to identify how to create an attractive movie trailer.

A trailer is a short movie clip consisting of small pieces of film mainly taken from an original movie. It is mainly used to advertise a new movie and therefore a trailer often includes a video segment, telop, narration, and so on, that does not appear in the original move in order to enhance its effectiveness. Strictly speaking, a trailer is not a so-called video summary: it is rare that we can grasp the whole outline of a movie by just watching its trailer, but should be attractive enough to make many people feel like to watch the movie. Although the trailer creation itself is a highly refined artistic work, it is interesting to investigate how a skilled and/or talented creator creates an attractive trailer from a video segment selection viewpoint.

Based on the Semantic Score obtained from many movies together with their originally created trailers, we analyzed which scenes are to be chosen for each trailer. Borrowing the conventional Japanese story model, the scene selection strategy we found is summarized as follows:

**Introduction (Ki):**

- Choose both complication and resolution scenes whose absolute Semantic Score are higher than a given threshold,
- Choose scenes at local peaks in the Semantic Graph and the following scene,

**Development (Sho):**

- Choose complication scenes whose Semantic Score are higher than a given threshold,
- Choose scenes at local peaks in the Semantic Graph,

**Turn (Ten):**

- Same as those in the Development,

**Conclusion (Ketsu):**

- No scene should be chosen.

This analysis clearly indicates that the Semantic Score provides those who want to tackle this sort of high-level problem with valuable information which is hard to characterize using conventional signal processing based technology.

In order to simulate a practical trailer creation, further analysis is necessary because the scenes used in the Semantic Score Method typically last for thirty to sixty seconds and thus simply concatenating selected scenes gives a long video clip which is too long to be a trailer. Therefore, we have studied how to identify a shot within the selected scene. We found that the shot selection strategy should consider the following criteria:

- ✓ Upper body image of main actor/actress
- ✓ Whole body image of main actor/actress
- ✓ Visual effect (CG, telop, dissolve, etc.)
- ✓ Sound effect (climax of BGM, explosion, scream, etc.)
- ✓ Speech

- ✓ High activity (of visual object and/or camera work)
- ✓ Shot length (slow motion more than several seconds)
- ✓ Camera zoom-in/out

By giving an appropriate weighting value to each shot within the scenes according to the criteria listed above, a shot candidate for a trailer can be determined at each scene. Although we have conducted this shot selection manually, this is the step where signal processing based technology can play an important role for its efficiency and, if possible, achieve it fully automatically.

Based on the strategy identified above, we have simulated trailer creation for "THE MASK OF ZORRO" using the Semantic Score obtained in the experiment in Section A.2.3. The *simulated* trailer was then evaluated by six evaluators. According to the evaluators, the simulated trailer received evaluation score of more than 60 point compared with an original trailer having 100 points as a basis. Note that the simulated trailer was created only using video segments in the original movie, i.e., none of extra technique such as taking special video segments and/or advertising narration that is not included in the original movie is considered. Hence, this result clearly indicates that the strategy we identified actually reflects a certain essence of an attractive trailer creation. Although it is obvious that a trailer created in such way cannot overwhelms the one created by a skilled and/or talented creator, we hope that this sort of approach will bring some reference materials that stimulate his/her creativity.

## A.3. Physiological Measurement and Analysis

In this Section, the physiological measurement and analysis is introduced to characterize a video. Since an attempt has just begun and the analysis is still under progress, we will show some preliminary results of our experiment conducted for the video characterization.

### A.3.1. Why Physiological Approach?

While the analytical approach described in Section A.2 characterizes how a story develops in a video, it cannot tell exactly where an audience is interested, excited, thrilled and so on, when he/she watches a video for the first time. This is because that the Semantic Score Method is an interpretation-based analytical approach that should be conducted consciously.

On the other hand, recent development of sensor technology and brain science have shown the potential to reveal human emotional and/or mental states through physiological measurement and analysis. This suggests that applying physiological technology to content characterization will provide valuable information that is hard to identify through the analytical approach. In addition, the physiological approach has some unique characteristics that the Semantic Score Method cannot achieve:

- Physiological approach can be done in real-time, resulting in real-time content characterization,
- Physiological approach can be done automatically, resulting in automatic content characterization,
- Physiological approach can be done with high accuracy in time, resulting in high-resolution content characterization.

It should be noted, however, that this does not mean that the analytical approach can be replaced with the physiological approach. Since we still have very little knowledge about the mechanism of human brain and physiological property, it is infeasible in the near future to estimate what he/she is thinking through physiological measurement and analysis. Furthermore, even if human thought estimation becomes possible, it cannot be used for content characterization if it imposes too heavy burden on audiences, e.g. with many electrodes directly attached to human bodies.

Our current hypothesis is that both the analytical and the physiological approach are complementary. Each approach highlights different aspect of content characteristic and thus both of them should be used to achieve the best content characterization.

## A.3.2. Preliminary Results of Physiological Measurement and Analysis

Based on the consideration described above, we have conducted an experiment on physiological measurement and analysis for content characterization.

Since this kind of experiment is regarded as medical, we have conducted the experiment with a help of a medical doctor in the medical department of a university. We also asked the Internal Review Board (IRB) of the depertment for permission to the experiment and it was officially accepted.

In the experiment, we have measured various physiological responses such as Electrocardiogram (ECG), Electromyogram (EMG) at the forehead, eye-blinking rate, and so on, from subjects while they watched a video. After the informed consent was obtained, several sensors and electrodes were attached on their bodies. They were then asked to sit down in front of an 80-inch screen on which the test material was shown. During the movie playback, physiological signals of the subjects were automatically recorded in real-time. Subjects' eye images were also taken on video to detect eye-blinking rate. Although they were forced to sit down during the playback, their motion was unconstrained.

The test materials used in the experiment were the same as used in Section A.2.3. But, because what we wanted to identify were fresh physiological responses from a subject caused by fresh stimuli, subjects who had watched the title before were filtered out in advance. This is the main difference between the physiological approach and the analytical approach. As is explained in Section A.2.3.1, evaluators are asked to watch the title *before* actual evaluation in the Semantic Score Method. This should be avoided in the physiological approach because subjects' experience to watch the title before would make them lose their sensitivity[5].

Since our attempt has just begun and the detailed analysis of the physiological data is still in progress. To demonstrate the usefulness of the physiological approach, however, we will show a couple of preliminary results of our analysis.

Figure A.5 shows time dependent EMG activities obtained from three subjects for a certain scene in "THE MASK OF ZORRO". The EMG activity is the electrical activity of muscles recorded by an electromyograph. In this measurement, electrodes were placed on the forehead of subjects, and the voltage difference between the electrodes was continuously recorded. Hence, when a muscle gave a particular movement, then the movement was indirectly detected as a change of the electric signal.

As is seen in Figure A.5, there is a spike in the EMG activity and, more notably the spikes in the EMG activities from three subjects coincided. In fact, this is the moment when all subjects smiled at the bang sound in the movie. In order to explain what happened in the scene, two images are extracted before and after the bang sound and shown under the graph. At the image before the bang sound, Zorro had attempted to jump off a wall onto his waiting horse (see left image).  But just before he lands, the horse moves forward and Zorro ends up on the ground. The bang sound occurrs at this moment. Since he was supposed to mount the horse, he was embarrassed after the bang sound (see right image). Subjects also expected that Zorro could mount the horse smoothly, the unexpected happening led subjects to smile.

As is demonstrated, the EMG activity can be used to detect smiles of subjects. Strictly speaking, what we detect was a particular muscle movement at the forehead, and it could happen not only for smiles but also for other emotions. Therefore, electromyography is a promising tool to capture some emotions of human being through his/her muscle activity. Note that, once we obtain information regarding the moment of specific emotion in the movie, the moment of smiling in this case, then it is worthwhile to save the information as data about the movie (metadata), so as to be used to retrieve the smiling parts in the movie.

[5] In our experiment, some subjects in the experiment were also asked to evaluate the title based on the Semantic Score Method. In this case, physiological measurement was first conducted, and then the subjective evaluation was asked to perform.

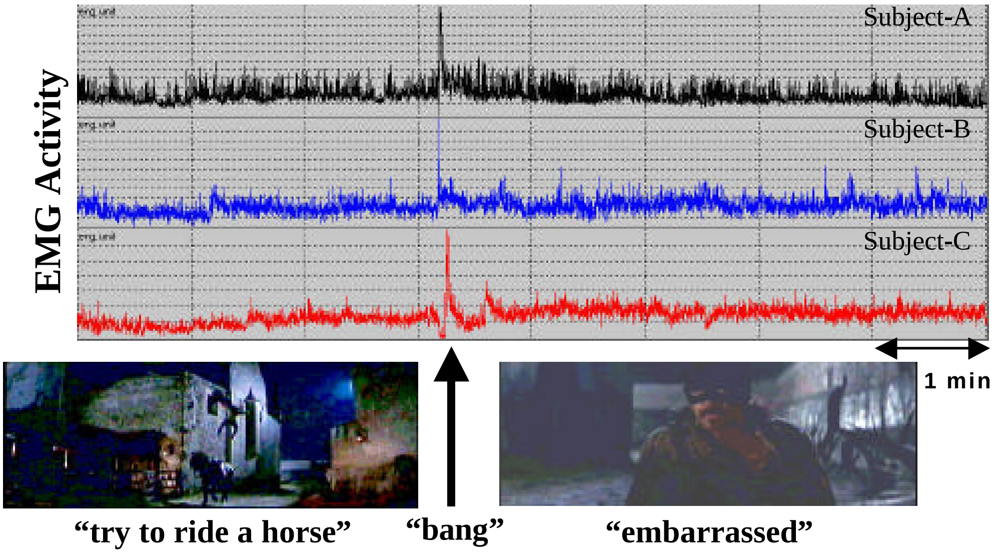


**Figure A.5: Spikes in Electromyogram (EMG) caused by smiling in "THE MASK OF ZORRO"**

Another example shown in Figure A.6 concerns highlight scene identification through the analysis of non-blinking periods. Video image of subject's eye was analyzed using image processing technology and eye-blinking points in time were extracted. Then non-blinking periods were measured as a time difference between two eye-blinking points. Figure A.6 shows non-blinking periods along the entire movie. In the graph of Figure A.6, the horizontal axis denotes time over whole the movie while the vertical axis is non-blinking period in second. This graph was created as follows: when a non-blinking period is given, then a regular square whose height (the vertical coordinate) is the same as the period is aligned on the horizontal period. Hence, non-blinking periods can be easily observed from the graph.

According to the graph, we found that there are long periods with non-blinking. The notable point again is that these long non-blinking periods correspond well to the highlight scenes in the movie. Here, we defined the highlight scene as the one audience pays special attention to. In order to show what happened at each highlight scene, an image is extracted from each highlight scene and shown around the graph with an arrow pointing to the corresponding non-blinking period. Simple text annotation was also attached to each frame to describe the scene.

Figure A.6 clearly suggests that the detection of non-blinking period can be a tool to identify the highlight scene in a movie. This is qualitatively explained by the fact that, as a natural property of human being, we tend to open our eyes wide without blinking when we watch something that attracts our attention.

Note again that, once we can detect the highlight scenes in such a way, the time data that references to the highlight scene should be preserved so as to be used for the highlight scene retrieval in the movie and/or the highlight scene based movie summarization.

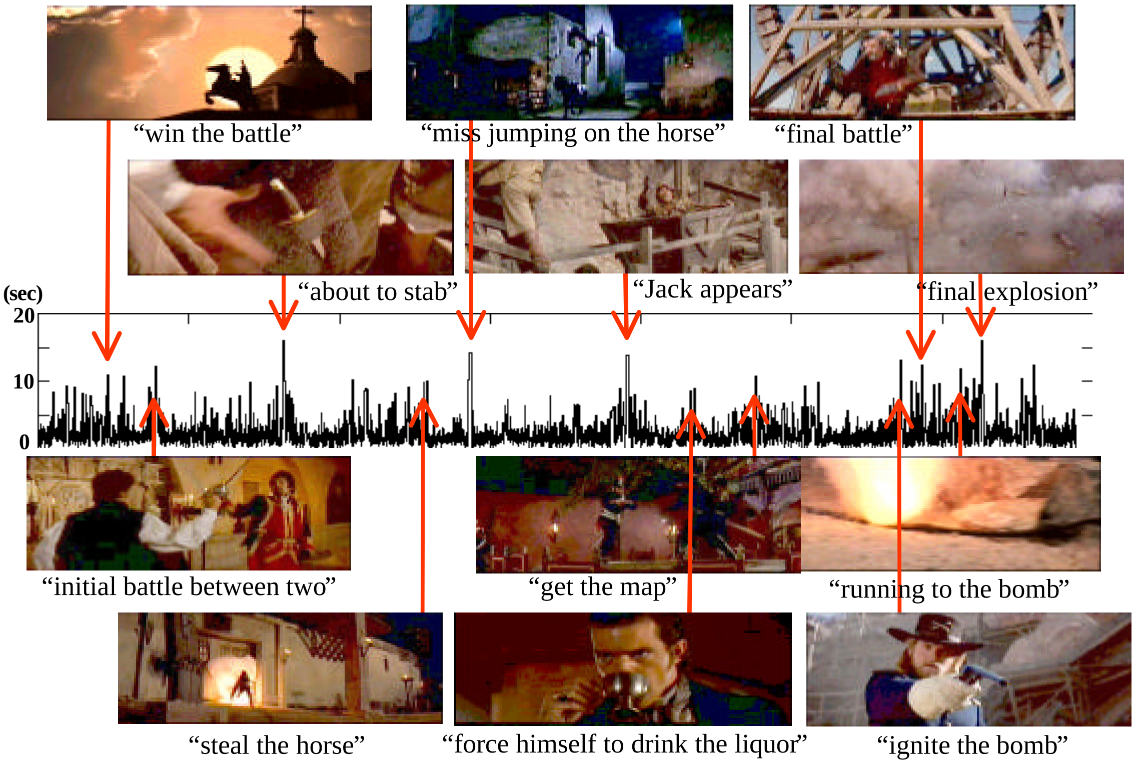


**Figure A.6: Highlight scenes identified by non-blinking periods in "THE MASK OF ZORRO"**

As we have demonstrated with above two examples, the physiological measurement and analysis is a promising tool for content characterization, which conventional technology cannot achieve before. Further analysis of the physiological data will be reported in the near future.

## A.4. Discussion

So far, we have introduced affect-based content characterization. Since there exist a couple of arguments against such an approach, we will address some of them here.

### A.4.1. Usefulness of Subjective Indexing

One argument often raised is that how useful such a subjective data for general indexing is. This argument is due to the fact that the sense of values and even interpretation varies among people, i.e. depends on their personalities. Moreover, it is true that a physiological response from a human being also strongly depends on his/her physiological characteristics, e.g., some people tend to blink while others not.

This is an essential problem involved in any high-level indexing since no subjective information is uniquely determined. Our understanding is, however, that if majority shares the same information and/or concept, then it is useful for us to use the information as index. For example, we cannot give an index "beautiful" to color "red" but we may give an index "warmness" to it. This is because whether we regards red color as beautiful or not would depends on our personal preferences, while we can easily associate red color with warmness (at least in Japan). Hence, giving an index "warmness" to color "red" is useful for color retrieval system in color database.

How to determine such indices is another problem to be solved. One possible solution may be based on so-called "common sense". In order to obtain reliable indexing, however, some sort of statistical approach such as the one used in the market research should be necessary.

In this sense, our content characterization also should be conducted by taking account of statistic point of view especially when the content characterizing data that reflects a general trend of audiences is demanded[6].

### A.4.2. Efficiency of Affect-based Content Characterization

Another argument we should address is the efficiency of content characterization or, say, metadata extraction. Automatic metadata extraction, where everything is conducted within a machine without any human assistance, is needed mainly because of its high efficiency. Our affect-based approach, on the other hand, has a much higher cost compared with such an ideal approach: the affect-based approach is basically based on human-content interaction and thus human plays an essential role. In addition, whether it is analytical or physiological one, an evaluator is bound by the evaluation at least within the running time of content.

However, if the benefit of metadata usage exceeds its extraction cost, then the extraction is worthwhile. In particular, when the metadata is described in accordance with a standardized metadata description such as MPEG-7, the metadata can be reused elsewhere. This suggests that the metadata is not only used for internal purpose but also distributed on demand for a fee, which compensates the extraction cost.

It is our belief that the content characterizing data obtained through our affect-based approach is worthwhile enough to be extracted regardless of its extraction cost when its unique and valuable characteristics and its wide range of potential applications are taken into consideration.

Furthermore, if we want to characterize newly created content rather than already existing content, the content characterization may be easily incorporated into the content production process. Usually, new content (once created) is initially previewed and evaluated by limited number of people before it is officially released to public. The evaluation may be conducted by interviewing the evaluators and even observing evaluators' behavior while their watching. The Semantic Score Method and the physiological approach provide an additional means for content evaluation, which provides information that cannot be obtained by the conventional method. Therefore, introducing our affect-based approach into the evaluation process is itself beneficial and, once the approach is incorporated in the evaluation, the content characterizing data will be obtained as a by-product.

### A.4.3. Affect-based Approach vs. Conventional Approach

One of our motivations for this work comes from the fact that there are aspects that are hardly treated to characterize only using conventional approach. We have demonstrated that information, which is hard to extract through signal processing based technology, can be obtained through the affect-based approach. However, this does not mean that our affect-based approach can substitute for the signal processing based approach. The affect-based approach highlights only high-level aspects of content characterization while there are many other aspects for a content especially low-level aspects which can be automatically characterized using signal processing based technology. On the other hand, many attempts have been and will be made to achieve such a high-level content characterization automatically based on signal processing technology, e.g., using sophisticated probabilistic models artificial intelligence, neural networks and so on, and it may be successfully realized in the future. For these attempts, the affect-based approach will supply with useful reference information and content characterizing data in a machine-friendly form, because one of the goals in our approach is to describe the high-level content characteristic in a quantitative fashion.

In conclusion, our affect-based approach, which includes the Semantic Score Method and the physiological measurement and analysis, and signal processing based approach are complementary to each other. We are now studying how to utilize these approaches adequately to achieve the best content characterization.

## A.5. Summary and Conclusions

In this appendix, we have introduced our attempt to characterize a video from the affective viewpoint. Our approach consists of two technique: the well-defined analytical approach named as the Semantic Score Method and the physiological measurement and analysis where various physiological response are measured from

[6] There is another aspect of our approach. Since the content characterizing data obtained from a certain individual will reflect how he/she interprets/feels the content is, this specific data may be useful for a certain application especially when he/she is a particular person, i.e., a celebrity.

subjects watching a video. The former technique is used to characterize how a story in a video develops, i.e., the story shape, while the latter for identifying where in a video audience feels interested, excited, and so on, as well as some specific emotions during appreciation. We have demonstrated the Semantic Score Method with its definition, its evaluation procedure as well as its applications. A couple of the preliminary results regarding the physiological measurement and analysis have been also introduced in order to demonstrate its usefulness. The content characterizing data obtained through the techniques is worthwhile even though it needs higher cost than the conventional signal processing based approach, especially when the data is described based on metadata description standard like MPEG-7 and become reusable. Our approach mainly focuses on high-level aspect of content characterization while signal processing based technology plays an important role especially for the low-level aspect. Therefore, appropriate usage of the techniques is demanded to achieve the best content characterization.

## A.6. Acknowledgements

We thank to Sony Pictures Entertainment for giving permission to use their commercial movie titles for this work. The Semantic Score Method has been developed as a collaborative study with Chiba University, and the physiological measurement has been conducted with a help of Niigata University School of Medicine. We highly acknowledges Prof. Kazuo Sugiyama, Prof. Makoto Watanabe in Chiba University and Prof. Takehiko Bando in Niigata University for their valuable help, advice and encouragement. We finally thank to all evaluators and subjects involved in our experiment for their serious cooperation.

## A.7. References

**[Kanade96]** Takeo Kanade and Shin-ichi Sato, “Informedia: CMU Digital Video Library Project”, IPSJ Magazine, Vol.37, No.9, pp.841-847 (1996) (in Japanese)

**[Kim96]** Y. B. Kim, M. Shibata, “Content-based video indexing and retrieval - A natural language approach -”, IEICE Trans. on Information and Systems, Vol. E79-D, No.6, pp.695-705 (1996)

**[Freytag98]** Gustav Freytag, “Technique of the Drama”, 2ed, ed. Translated by Elias J. MacEwan, Chicago: Scott, Foresman, 1898.

**[Laurel93]** Brenda Laurel, “Computers as Theatre”, Addison-Wesley, 1993.