Toward a Structural and Semantic Metadata Framework for Efficient Browsing and Searching of Web Videos

Hyun-Hee Kim (김현희)*

Contents

ABSTRACT

This study proposed a structural and semantic framework for the characterization of events and segments in Web videos that permits content-based searches and dynamic video summarization. Although MPEG-7 supports multimedia structural and semantic descriptions, it is not currently suitable for describing multimedia content on the Web. Thus, the proposed metadata framework that was designed considering Web environments provides a thorough yet simple way to describe Web video contents. Precisely, the metadata framework was constructed on the basis of Chatman's narrative theory, three multimedia metadata formats (PBCore, MPEG-7, and TV-Anytime), and social metadata. It consists of event information, eventGroup information, segment information, and video (program) information. This study also discusses how to automatically extract metadata elements including structural and semantic metadata elements from Web videos.

Keywords: Multimedia, Structural Metadata, Semantic Metadata, MPEG-7, PBCore, TV-Anytime, Chatman's Narrative Theory, Social Metadata, Event, Segment, Content-based Search

* Professor, Department of Library and Information Science, Myongji University, Seoul, Korea (kimhh@mju.ac.kr)
논문접수일자: 2017년 1월 23일 최초심사일자: 2017년 1월 23일 게재확정일자: 2017년 2월 10일
한국문헌정보학회지, 51(1): 227-243, 2017. [http://dx.doi.org/10.4275/KSLIS.2017.51.1.227]

1. Introduction

The advances in information technology together with the rapid evolution of multimedia data are resulted in the huge growth of Web videos. Most recently, it was predicted that letters from Facebook would be replaced by photos and videos within next 5 years (http://fortune.com/2016/06/14/facebook-video-live/). Due to such rapid growth of the Web videos over the Internet, it is becoming very important to perform accurate and efficient content-based searches and to provide nonlinear access to any segment of them without viewing whole videos. Furthermore, there is a strong demand for short pieces of audio-visual (AV) content in the archive by media professionals (Huurnink et al. 2010), thus it may be beneficial to increase support for fine-grained access to AV content, for example, through automatic segmentation.

To do that, we need to employ a structural and semantic metadata framework in which the structural metadata is used to describe the structure of AV content in terms of video segments, frames, still and moving regions, and audio segments, and the semantic metadata is used to represent the objects, events, and notions from the real world that are captured by the AV content. MPEG-7 supports multimedia structural and semantic descriptions. However, it is known that MPEG-7 is not currently suitable for describing multimedia content on the Web (Arndt et al. 2007) and further it is too complicated to work with (List and Fisher 2004).

In this study, we proposed a structural and semantic metadata framework, which was designed considering Web environment. More specially, in order to investigate how to structure Web videos and which structural and semantic metadata elements are useful, we reviewed social metadata and three international multimedia metadata formats, such as PBCore 2.0, MPEG-7, and TV-Anytime (TVA). Then, we adapted Chatman's narrative theory (Chatman 1975) in which a narrative is any report of connected events, presented in a sequence of written words or videos (Teeter and Sandberg 2016; Chatman 1975).

On the basis of the narrative theory, we identified events in AV content. Comprehending an event depends on identifying the nature of its key action and the roles played by the people and objects in the action (Nowak, Plotkin and Jansen 2000; Klix 2001). Thus, we assume that an event is an instantaneous or temporally extended action or 'happening' that may involve a single agent object or object, an interaction between two or more agent objects, or all the agent objects (Shotton et al. 2002). Therefore, for the elements of an event specific metadata, we used object, agent object, action, place, time, and theme extracted from the analysis of a discourse such as dialogues, sound, and music. Our proposed metadata framework being constructed through

the above-mentioned steps can be utilized for applications such as content-based searches, fast browsing, and dynamic video summarization.

2. Multimedia Metadata: A Comparison of PB-Core, TV-Anytime, and MPEG-7

We below described three international multimedia metadata formats, such as PBCore 2.0 (http://pbcore.org/introducing-pbcore-2-0/), MPEG-7 (ISO/IEC 2002 - 2004) (http://mpeg.chiariglione.org/standards/mpeg-7), and TV-Anytime (TVA) (http://www.tv-anytime.org/) (Evain and Martínez 2007). We selected three multimedia metadata formats because they are international standards and used by many digital libraries. PBCore 2.0 was created by the public broadcasting community in the United States of America for use by public broadcasters. PBCore 2.0 that was built on the foundation of the Dublin Core (ISO 15836) is made up of 4 content classes, 15 containers, and 82 elements. The four content classes consist of intellectual content (descriptive metadata), intellectual property (creation and usage information), instantiation (technical metadata), and extension.

TVA includes content description metadata, instance description metadata, segmentation metadata, and consumer metadata. The content description metadata describes general information about a piece of content (e.g., title), whereas the instance description metadata does a particular instance of a piece of content (e.g., video format) (Lee et al. 2005). The segmentation metadata is used to define, access, and manipulate temporal intervals (e.g., segments) within an AV stream and the consumer metadata includes usage history data and user preferences for a personalized content service.

MPEG-7 consists of twelve parts including multimedia description schemes (MDS), audio, visual, and systems. Among them, three of MDS, audio, and visual are the most important parts of it. MDS are metadata structures for describing and annotating AV content. MDS include basic elements (schema tools, basic datatypes, link and media localization, and basic tools), content description, content management, content organization, navigation and access, and user interaction.

The content description represents the structure and semantics of AV content. That is, the content description tools represent perceivable information, comprising structural aspects (structure description tools) and conceptual aspects (semantic description tools). The structure description tools allow the description of the content in terms of spatio-temporal segments organized in a

hierarchical structure. The semantic description tools allow the description of the content from the viewpoint of real-world semantics and conceptual notions: objects, events, abstract concepts, and relationships. The semantic and structure description tools can be further related by a set of links, which allows AV content to be described on the basis of both content structure and semantics together. On the other hand, visual covers basic visual features such as color, texture, shape, motion, localization, and face recognition, whereas audio provides structures for describing audio content.

As described above, three metadata formats have in common in that they have many overlapping metadata elements, and that they do not consider social relations among users and social metadata. Despite many similarities among them, there are some differences in the following aspects. First, MPEG-7 enables to specify low-level AV features such as color and motion, and further to specify high-level AV features such as segments, objects, and events. Those features are useful for applications such as content-based querying and video summarization. Second, TVA focuses more on describing consumer profiles including search preferences to facilitate automatic filtering and acquisition of content by agents on behalf of the consumer, comparing to the other two metadata formats. Third, PBCore 2.0 provides a simple way to describe AV content.

We compared these three metadata formats in terms of creation and production information, media information, structural and semantic information, and usage information using key metadata elements (see Table 1). The creation and production information is related to data about the creation and production of AV content, media information is related to the media-specific characteristics of the content, and the usage information to the usage of the content, such as copyright and usage history. On the other hand, the structural information describes the AV content from the viewpoint of its structure, whereas the semantic information represents the content from the viewpoint of real-world semantics and conceptual notions.

〈Table 1〉 A Comparison of PB-Core (2.0), TVA (1.3), and MPEG-7 (10)

Creation and Production Information		
PB-Core(2.0) (2011)	TVA(1.3) (2003)	MPEG-7(10) (2001)
pbcoreAssetType pbcoreAssetDate pbcoreIdentifier pbcoreAssetTitle pbcoreCreator (creator, creatorRole) pbcoreContributor (contributor, contributorRole)	title mediaTitle shortTitle creditsList productionDate prductionLocation synopsis	title creator (role, agent [name, type]) creationCoordinates (creationLocation, creationDate) date genre subject

Creation and Production Information		
PB-Core(2.0) (2011)	TVA(1.3) (2003)	MPEG-7(10) (2001)
pbcorePublisher (publisher, publisherRole) pbcoreSuject pbcoreDescription pbcoreAnnotation pbcoreGenre pbcorecoverage (coverage, coverageType) pbcoreRelation (pbcoreRelationType, pbcoreRelationIdentifier)	programDescription keyword genre language captionLanguage signLanguage relatedMaterial	abstract language captionLanguage signLanguage relatedMaterial DS (publicationType, materialType, mediaLocator, mediaInformation, creationInformation, usageInformation)
Media Information		
PB-Core	TVA	MPEG-7
pbcoreInstantiation (instantiationIdentifier, instantiationLocation, instantiationDuration, instantiationMediaType, instantiationColors, etc.) instantiationEssenceTrack (essenceTrackType, essenceTrackAspectRatio, essenceTrackEncoding, essenceTrackFrameRate, etc.)	fileFormt fileSize system bitRate audioAttributes videoAttributes	fileFormt fileSize bitRate visualCoding audioCoding visual (color, texture, shape, motion, localization, and face recognition) audio (silence, spoken content, timbre, sound effects, melody contour, etc.)
Structural and Semantic Information		
PB-Core	TVA	MPEG-7
pbcorePart (pbcoreIdentifier, pbcoreTitle, pbcoreDescription, pbcoreSubject, etc.)	segment (title, synopsis, keyword, relatedMaterial, programRef, description, segmentLocator, keyFrameLocator) segmentGroup (programRef, groupType, description, groupInterval, segments, groups, keyFrameLocator)	structure DS segment DS (creation information, usage information, media information, textual annotation) segmentRelation DS semantic DS (agentObject DS, object DS, event DS, concept DS, semanticState DS, semanticPlace DS, semanticTime DS)
Usage Information		
PB-Core	TVA	MPEG-7
pbcoreAudienceLevel pbcoreAudienceRating pbcoreRightsSummary (rightsSummary, rightsLink, rightsEmbedded)	releaseInformation (releaseDate, releaseLocation) parentalGuidance userDescription (userPreference, usageHistory) awardsList copyrightNotice	release (country, date) target (market, age) usageInformation DS (rights, availability, usageRecord, etc.)

3. Related Studies

Let us start with studies on Web video use. Cunningham and Nichols (2008) suggested that the video queries submitted by many participants were driven by their mood or emotional state, and YouTube was the primary site consulted by them. Huurnink et al. (2010) mentioned that media professionals have a strong demand for short pieces of audiovisual material in the archive, and their queries predominantly consist of broadcast titles and of proper names.

Shotton et al. (2002) proposed a metadata classification schema for the characterization of items and events in cell biological videos that permits subsequent query by content. Following MPEG-7 nomenclature, they first defined metadata intrinsic to the information content of the video as either structural metadata or semantic metadata. Then, they showed how the semantic metadata types should be organized within a database.

Agnew, Kniesner, and Weber (2007) described the implementation of MPEG-7 within the Moving Image Collections (MIC), which is a union catalog of the world's moving images. The MIC Union Catalog was designed to utilize a core registry schema that is designed to map readily to any metadata schema used to describe moving images. They developed draft MPEG-7 to MIC and MIC to MPEG-7 maps. They also discussed issues with MPEG-7 as a descriptive metadata schema, as well as mapping and implementation issues identified in their study.

Benitez, Zhong, and Chang (2007) proposed two research prototype systems that demonstrate the generation and consumption of MPEG-7 structure and semantic descriptions in retrieval applications. The first prototype system allows to segment and model objects as a set of regions with corresponding visual features and spatiotemporal relations. The region-based model provides an effective base for similarity retrieval of objects. Whereas the second system enables to represent semantic and perceptual facts about the world using multimedia.

Lunn (2009) investigated three aspects of scholars' and students' information seeking behaviour in a television broadcast context, and the associated implications for design and construction of metadata elements in surrogate records in future broadcast retrieval systems. The three aspects of information seeking behaviour in focus are information need characteristics, preferred search entries, and application of relevance criteria.

His research results showed that 24 access points (metadata elements) are appropriate in relation to a future broadcast retrieval system. These are title, participant, author, date of production, channel, summary, spoken words, clips, and images. He mentioned that questionnaire respondents expressed the need for a further specification of roles for the author access point. Roles in such

a specification are director, actor, cameraman, commentator, and anchorperson.

Makkonen et al. (2010) proposed two algorithms that are used to find events by using video metadata. The first algorithm is about missing data compensation that harvests missing data values from textual descriptions in video metadata and the second one is a layered clustering method that divides videos into clusters, each of which is considered as an event. Their test results of the two methods showed that the missing data compensation yielded better results in terms of accuracy than using raw data, and that the clustering method can produce good quality clusters of events.

Algur, Bhat, and Jain (2014) constructed high level descriptive metadata for all the category of Web videos such as YouTube. They suggested that by using the high level descriptive metadata information, a user is facilitated to locate a specific video and is further able to comprehend rapidly the main concept of a video without the need to watch the whole of it. They proposed general descriptive metadata, object-specific metadata, and event-specific metadata. The general descriptive metadata have 13 elements including title, director, actors, theme, language, storage format, number of emotional scenes, and number of songs. On the other hand, the event-specific metadata consist of 7 elements including category, time, and starting frame, whereas the object-specific metadata have 7 elements such as object ID and motion of the object.

Cho (2014) analyzed the structure of EBBS (European Broadcasting Union) Core, PB Core, and KBS (Korean Broadcasting System) metadata, and found no big differences among the three formats. He suggested that for efficient browsing and reusing AV content, the AV content needs to be segmented based on shots or scenes rather than sequences, a series of connected scenes.

Taking into account the findings and issues from the previous studies, we can conclude that it may be beneficial to allow fine-grained access to AV material through automatic segmentation or content-based analysis. For this end, Shotton et al. proposed a good semantic metadata schema but it was designed for cell biological videos and thus domain-specific. On the other hand, Algur at al. described general descriptive metadata, object-specific metadata, and event-specific metadata for general Web videos. However, they did not give a framework for describing the relationships among these three metadata. Therefore, we think that it is needed to design a structural and semantic metadata framework for general Web videos in a more elaborate way.

4. Structural and Semantic Metadata Framework

4.1 Theoretical Model

For constructing an efficient structural and semantic metadata framework for video data, we need to examine how the narrative of a video is constructed and how its topic is determined by viewers. We reviewed Chatman's narrative theory (Chatman 1975) in order to precisely examine the concept of narrative. Chatman analyzed the narration structures of films and novels. He suggested that each narrative has two parts: a story, the content or chain of events (actions, happenings), plus what may be called the existents (characters, setting); and a discourse, that is, the expression, the means by which the content is communicated. Chatman mentioned that the story is the what in a narrative that is depicted, discourse the how. Leaving the role of discourse for a story aside, his argument regarding the structure of a story was quite simple, that is, a story equals events plus characters plus setting. As shown below, Chatman's model conceptualizes narrative transmission as follows:

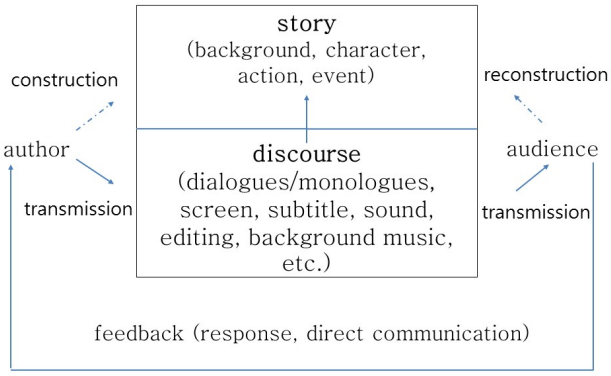
Actual author → [implied author → (narrator) → (narratee) → implied audience] → actual audience

Chatman's model describes the basic concepts that need to be represented in some formal way so that a story engine can build a coherent narrative regarding content and presentation form. Thus, Chatman's model can be adapted for automated story generation (Reijnders 2011).

Kim (2009) proposed a modified model of narrative communication introducing a notion of the space of communication, in front of which actual authors and audience posit themselves (see Figure 1). In his model, implied authors and implied audience by Chatman are nothing but psychological entities in the communication space which is mediated by audio-visual technologies. Producers should think over what audience want to watch and hear, while audience should do what producers would like to put into their works.

According to the model as shown in Figure 1, a sender (e.g., video producers) intends to deliver a message, its meaning is reinterpreted and reconstructed from a receiver's point of view during the process of communication. Thus, understanding the narrative structure of a video helps to construct an efficient method for abstracting visual information (e.g., moving images). For example, in order to enable receivers (users) to better understand the structural analysis of a video, we need to get information on the story and discourse of a narrative. That is, for constructing an

efficient video trailer, we need to extract shots from a video that include topic-related events, characters, and background. Additionally, in order to get information on discursive features, we need the shots that engage in narrative activities, such as dialogues and monologues of characters, commentary, camera work and movement, and music.



〈Figure 1〉 Model of Narrative Communication (Kim 2009; Chatman 1978)

4.2 Metadata Framework

We below described our proposed structural and semantic metadata framework that was constructed on the basis of Chatman's narrative theory, social data, and three multimedia metadata formats. Humans often claim to understand events when they manage to formulate a coherent story or narrative explaining how they believe the event was generated, thus narratives lie at foundations of our cognitive procedures. On the basis of Chatman's narrative theory, our metadata framework includes event and eventGroup information in order to divide a video into clusters, each of which is considered as an event. We also include both segment and video (program) information that are closely related to the event information.

We utilized social data, such as social metadata, social content, and social relation information being obtained from social network services, such as Facebook and YouTube. Such social data are useful in that they enable to automatically construct metadata and to use social relation information. However, social data have some disadvantages such as spam tags, thus, it is necessary to filter them before using.

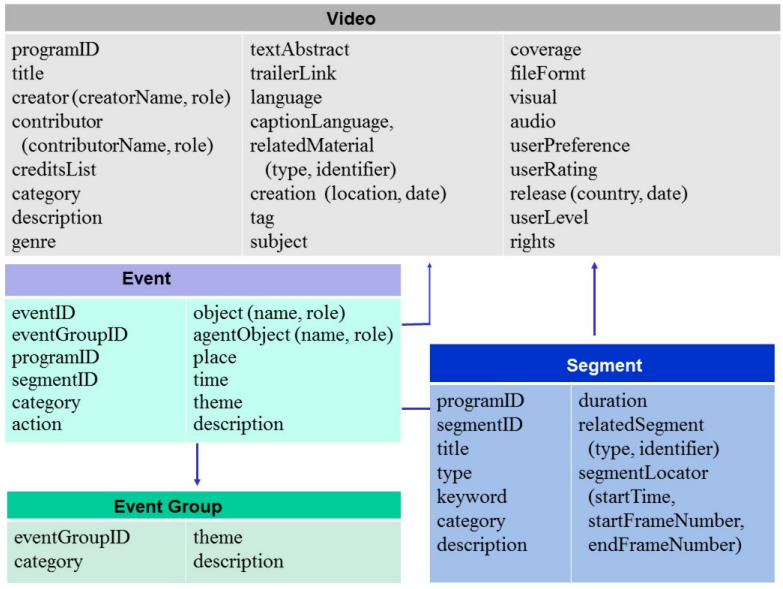
4.2.1 Event and EventGroup Information

Let's start with the event and eventGroup information. For the elements of an event specific metadata (see Figure 2 and Table 2), there are 12 elements including object (name, role), agentObject (name, role), action, place, time, and theme extracted from the analysis of a discourse such as dialogues, sound, and music.

Then, we made an eventGroup that denotes a collection of events that are grouped together, for a particular purpose or due to a shared property. An eventGroup contains events, or other eventGroups. For example, if there is a birthday party that is an important event in the overall story of a program. Then the agentObject, action, place, and time elements can be attached to the birthday party event. Depending on the type of an event, we can group events extracted from either a program or several programs through the eventGroup information.

4.2.2 Segment Information

We assume that a program can be segmented automatically based on a shot, a scene or a sequence. Any segment may be described by metadata elements that are used to represent a whole video because segments are parts of a video. Moreover, a specific feature for segment is also allowed. The specific feature is segmentLocator, which describes the location of a segment within a program in terms of start time, start frame number, and end frame number.



〈Figure 2〉 Structural and Semantic Metadata Framework

〈Table 2〉 Details of the Proposed Metadata Framework

Video		
Element	Description	Example
programID	An ID of a given AV material	PID: 012
title	A name or label relevant to a material	Land
creator (creatorName, role)	creatorName and role sub-elements identify the primary person (s), or organization (s) responsible for creating a material and the role played by the creator (s)	creator (H. Kim, director)
contributor (contributorName, role)	contributorName and role sub-elements identify person (s), or organization (s) that made substantial creative contributions to a material and the role played by the contributor (s)	contributor (H. Park, narrator)
creditsList	The list of credits (e.g. actors) for a material	Actors: S. kim …
category	Category of a material	Shows
description	Free-form text or a narrative to report general notes, or summaries	What you need to do to be happy …
genre	Genre of a material	Comedy
textAbstract	Free-form text to report abstract about the intellectual content of a video	The video explains why two …
trailerLink	A URI pointing to the dynamic video skim for a material	http://www...
language	Language of a material	ENG
captionLanguage	Caption language of a material	KOR
relatedMaterial (type, identifier)	Type sub-element describes the relationship between the AV material being described and any other AV material, and identifier sub-element indicates an ID of a related AV material or an URL that points to it	relatedMaterial (part-whole relations, PID1234)
creation (location, date)	Location and date sub-elements identify where and when a material was produced (created)	creation (Seoul, 2015)
tag	Keywords assigned by users	Library
subject	Topic headings or keywords that portray the intellectual content of a material	Powhatan Indians
coverage	The geographic location or the time period covered by an AV content	1607-1631
fileFormt	File format	AVI
visual/audio	Information for basic visual features (e.g., color, or object recognition) and audio content (e.g., melody contour)	visual and audio information
userPreference	User preference information, such as favorite actors or TV shows	Information about favorite video genre
userRating	A user's rating for an AV material	Number of likes and dislikes
release (country, date)	Country and date sub-elements identify where and when a material was released	release (Korea, 2015)
userLevel	A type of audience for whom a material is primarily designed or educationally useful	For educational use (K-2)
rights	Copyrights for an AV material	CC BY 3.0

Segment		
Element	Description	Example
programID	An ID of the AV material that includes a given segment	PID: 2354
segmentID	An ID of a segment	SID: 230
title	A name or label relevant to a segment	Birthday party
type	Type of video segmentation	Shots or scenes
keyword	Keywords for a segment	Library
category	Category for a segment	Shows
description	Summary for a segment	This segment describes how …
duration	Duration of a segment	00:56:46
relatedSegment (type, identifier)	Type sub-element describes the relationship between the segment being described and any other segment, and identifier indicates an ID of a related segment or an URL that points to it.	relatedSegment (part-whole relations, SID123)
segmentLocator (startTime, startFrameNumber, endFrameNumber)	Location of the segment within a program in terms of start time, start frame number, and end frame number	segmentLocator (00:05:54, 0, 183)
Event		
Element	Description	Example
eventID	An ID of the event described	EID: 34
programID	An ID of the AV material that includes the event	PID: 12
segmentID	An ID of the segment that includes the event	SID: 12
category	Category for the event	Entertainment
action	Action described in the event	Dancing
object	Object appeared in the event	Flowers
agentObject	Agent object appeared in the event	Six people
place	Place related to the event	Park
time	Time related to the event	01:12:2017
theme	Topic for the event	Social sports
description	Summary for the event	This event is …
Event Group		
Element	Description	Example
eventGroupID	An ID of grouped events	GEID: 23
category	Category for grouped events	Entertainment
theme	Topic for grouped events	Comedy
description	Summary for grouped events	This grouped events …

4.2.3 Relationships between Event Information and Segment Information

The event can be linked to segment, as described below. Suppose making a sandwich is an important event in a given program, then our proposed framework tools allow to search segment database to find the segment that is matched with the database and to link the event to the segment being retrieved through its segmentID. Events can occupy extended periods of time that need

not necessarily be contained within individual video segments. For this reason, as shown in Figure 2, events are also linked directly to videos.

4.2.4 Video Information

A whole video has 25 metadata elements that consist of creation and production information, media information, usage information, and social metadata. Among them, the relatedMaterial element contains two sub-elements, type and identifier. The type element is used to describe the relationship between the AV material being described and any other AV material; both AV materials can be related by part-whole relations or different versions of an original.

We also utilized social metadata (e.g., tags), social content, and social relation information. For the userPreference element that is used to describe a user's preferences for consumption of multimedia material, we can use two social recommendation methods: collaborative filtering approach and content-based approach (Bouadjenek, Hacid and Bouzeghoub 2016). The collaborative filtering approach intends to recommend items to a user based on other people who are found to have similar preferences or tastes, whereas the content-based approach that is based on recommending items that are similar to those in which the user has shown interest in the past.

For example, personalization of content access and content consumption can be done by using a user preference information, which will be matched with media descriptions and also done by using other people's preference information, if they have similar preference to the user.

On the other hand, the visual element is used to describe basic visual features such as color, texture, shape, motion, localization, face recognition, and object recognition, whereas the audio element to describe audio content, such as spoken content, timbre, and sound effects. These visual and audio elements are employed to facilitate content-based searching (e.g., query-by-humming).

5. Discussion and Conclusion

With the explosive growth of Web videos on the Internet, it becomes challenging to efficiently browse and search hundreds or even thousands of Web videos. In this study, we proposed a structural and semantic metadata framework for Web videos after reviewing three metadata formats (PBCore 2.0, MPEG-7, and TVA), social metadata, related studies, and Chatman's narrative theory.

Our metadata framework allows to structure a program (narrative) on the basis of events, and it consists of event information, eventGroup information, segment information, and video (program)

information. Our proposed metadata framework can be utilized for applications such as content-based searches, fast browsing, and video summarization. We need to do a further study on how structural and semantic metadata types should be organized within a database.

Social data being obtained from social network services are utilized for some elements of the video information. Our proposed metadata framework can be utilized to perform content-based searches, fast browsing, or relevance assessment, and to construct static video summaries (storyboards) or dynamic video skims (trailers).

The next research issue concerned will be how to automatically extract metadata elements. There have been many studies on automatic metadata construction (Christel 2009; Park and Lu 2009). Recently, some metadata elements such as duration and title can be generated by employing fully automated video content analysis systems such as InfoExtractor tool (http://infoextractor.org/), whereas video trailer can be generated automatically by using Azure Media Services (https://azure.microsoft.com/).

Video shot (or scene) boundary detection techniques that can be used to automatically segment videos has received attention in the field of computerized image processing, and the techniques have made big progress (Smeaton, Over and Doherty 2010; Chen, Delannay and Vleeschouwer 2011). The content features of a video such as faces and objects can be also identified automatically by using image processing techniques (Togawa and Okuda 2005; Yokoi, Nakai and Sato 2008). However, such content feature extraction techniques have not yet been developed to the point where they are robust or effective in dealing with large collections.

Therefore, it is necessary to review other techniques and theories utilized together with the image processing techniques. One such area is theories of cognitive psychology and cognitive neuroscience techniques. Mehmood et al. (2016) proposed an efficient human-attention model that combines both external (multimedia content) and internal information (viewer's neuronal responses) for video summarization. Their model allows the fusion of multimedia and neuronal signals, which provides a bridge that links the digital representation of multimedia with the viewer's perceptions. On the other hand, Behroozi, Daliri, and Shekarchi (2016) described that visual attention-related brain activities evoked by stimulus can be regarded as the signature of detection and identification of objects. That is, they investigated the possibility to identify conceptual representation based on the presentation of 12 semantic categories (e.g., animal and building) of objects using EEG signals in conjunction with a multivariate pattern recognition technique.

For the automatic construction of metadata, we also can use social metadata, social content, and social relation information in a more elaborate way. For example, according to the study

of Wang et al. (2012), event driven web video summarization was extracted using tag localization and key-shot identification. They localized the tags that are associated with each video into its shots, estimated the relevance of the shots with respect to the event query by matching the shot-level tags with the query, and identified a set of key-shots by performing near-duplicate key-frame detection. We need to further investigate studies on how to apply social data and cognitive neuroscience into efficient multimedia metadata construction of Web videos.

References

[1] Kim, Yong-Ho. 2009. "A Structural Model of Mediated Visual Communication in Narrative Movies: Focusing on Chatman and Bordwell's Controversy." *Korean Journal of Journalism & Communication Studies*, 53(1): 209-232.

[2] Cho, Young-Joon. 2014. "The Study on improvement of Broadcast Metadata about Clip Video at Broadcast Content Managements." In *Proceedings of 2014 Korean Society of Broadcast Engineers Summer Conference*, June 30-July 2, 2014, Jeju: Jeju National University: 59-63.

[3] Agnew, G., Kniesner, D. and Weber, M. B. 2007. "Integrating MPEG-7 into the Moving Image Collections Portal." *Journal of the American Society for Information Science and Technology*, 58(9): 1357-1363.

[4] Algur, S. P., Bhat, P. and Jain, S. 2014. "Metadata Construction Model for Web Videos: A Domain Specific Approach." *International Journal of Engineering and Computer Science,* 3(12): 9742-9748.

[5] Arndt, R. et al. 2007. "COMM: Designing a Well-Founded Multimedia Ontology for the Web." In *Proceedings of the 6th International Semantic Web Conference and the 2nd Asian Semantic Web Conference*, November 11-15, 2007, Busan: BEXCO.

[6] Behroozi, M., Daliri, M. R. and Shekarchi, B. 2016. "EEG Phase Patterns Reflect the Representation of Semantic Categories of Objects." *Medical & Biological Engineering & Computing*, 54(1): 205-221.

[7] Benitez, A. B., Zhong, D. and Chang, S. F. 2007. "Enabling MPEG-7 Structural and Semantic Descriptions in Retrieval Applications." *Journal of the Association for Information Science and Technology*, 58(9): 1377-1380.

[8] Bouadjenek, M. R., Hacid, H. and Bouzeghoub, M. 2016. "Social Networks and Information

Retrieval, How Are They Converging? A Survey, a Taxonomy and an Analysis of Social Information Retrieval Approaches and Platforms." *Information Systems*, 56(2016): 1-18.

[9] Chatman, S. 1975. "Towards a Theory of Narrative." *New Literary History*, 6(2): 295-318.

[10] Chen, F., Delannay, D. and De Vleeschouwer, C. 2011. "An Autonomous Framework to Produce and Distribute Personalized Team-Sport Video Summaries: A Basketball Case Study." *IEEE Transactions on Multimedia*, 13(6): 1381-1394.

[11] Christel, M. G. 2009. *Automated Metadata in Multimedia Information Systems: Creation, Refinement, Use in Surrogates, and Evaluation*. Synthesis Lecture on Information Concepts, Retrieval, and Services, 2. San Rafael, CA: Morgan & Claypool Publishers.

[12] Cunningham, S. J. and Nichols, D. M. 2008. "How People Find Videos." In *Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries,* June 16-20, 2008, Pittsburgh, PA: Omni William Penn Hotel: 201-210.

[13] Evain, J. P. and Martínez, J. M. 2007. "TV-Anytime Phase 1 and MPEG-7." *Journal of the American Society for Information Science and Technology*, 58(9): 1367-1373.

[14] International Organization for Standardization/International Electrotechnical Commission (ISO/IEC). 2002-2004. *ISO/IEC 15938: Part 1-8: Information Technology: Multimedia Content Description Interface (MPEG-7).* Geneva: International Organization for Standardization.

[15] Huurnink, B. et al. 2010. "Search Behavior of Media Professionals at an Audiovisual Archive: A Transaction Log Analysis." *Journal of the Association for Information Science and Technology*, 61(6): 1180-1197.

[16] Klix, F. 2001. "The Evolution of Cognition." *Journal of Structural Learning and Intelligence Systems*, 14: 415-431.

[17] Lee, H. K. et al. 2005. "Personalized TV Services and T-Learning Based on TV-Anytime Metadata." In *Proceedings of the 6th Pacific-Rim Conference on Multimedia,* November 13-16, 2005, Jeju: Ramada Plaza Jeju Hotel: 212-223.

[18] List, T. and Fisher, R. B. 2004. "CVML-An XML-based Computer Vision Markup Language." In *Proceedings of the 17th International Conference on Pattern Recognition,* August 23-26, 2004, Cambridge: 789-792.

[19] Lunn, B. K. 2009. *Towards the Design of User based Metadata for Television Broadcasts*. Saarbrücken: VDM Verlag.

[20] Makkonen, J. et al. 2010. "Detecting Events by Clustering Videos from Large Media Databases." In *Proceedings of the 2nd ACM International Workshop on Events in Multimedia,* October 25, 2010, Firenze: 9-14.

[21] Mehmood, I. et al. 2016. "Divide-and-Conquer based Summarization Framework for Extracting Affective Video Content." *Neurocomputing*, 174(A): 393-403.

[22] The Moving Picture Experts Group (MPEG). [n.d.]. *MPEG*. Villar Dora: The Moving Picture Experts Group. [online] [cited 2016. 9. 11.] <http://mpeg.chiariglione.org/>

[23] Nowak, M. A., Plotkin, J. B. and Jansen, V. A. 2000. "The Evolution of Syntactic Communication." *Nature*, 404(6777): 495-498.

[24] Park, J. R. and Lu, C. 2009. "Application of Semi-Automatic Metadata Generation in Libraries: Types, Tools, and Techniques." *Library & Information Science Research*, 31(4): 225-231.

[25] PBCore. [n.d.]. *PBCore*. [online] [cited 2016. 9. 3.] <http://pbcore.org/introducing-pbcore-2-0/>

[26] Reijnders, K. 2011. *Suspense Tours: Narrative Generation in the Context of Tourism*. Amsterdam: Universiteit van Amsterdam.

[27] Shotton, D. M. et al. 2002. "A Metadata Classification Schema for Semantic Content Analysis of Videos." *Journal of Microscopy*, 205(1): 33-42.

[28] Smeaton, A. F., Over, P. and Doherty, A. R. 2010. "Video Shot Boundary Detection: Seven Years of TRECVid Activity." *Computer Vision and Image Understanding*, 114(4): 411-418.

[29] Teeter, P. and Sandberg, J. 2016. "Cracking the Enigma of Asset Bubbles with Narratives." *Strategic Organization*, 15(1): 91-99.

[30] Togawa, H. and Okuda, M. 2005. "Position-Based Keyframe Selection for Human Motion Animation." In *Proceedings of 11th International Conference on Parallel and Distributed Systems,* July 20-22, 2005, Fukuoka: 182-185.

[31] TV-Anytime Forum. 2005. *TV Anytime Forum.* [online] [cited 2016. 5. 16.] <http://www.tv-anytime.org/>

[32] Wang, M. et al. 2012. "Event Driven Web Video Summarization by Tag Localization and Key-Shot Identification." *IEEE Transactions on Multimedia*, 14(4): 975-985.

[33] Yokoi, K., Nakai, H. and Sato, T. 2008. "Toshiba at TRECVID 2008: Surveillance Event Detection Task." In *Proceedings of the TRECVID 2008 Workshop,* November 17-18, 2008, Gaithersburg, MD: National Institute of Standards and Technology.