A Video Semantic Annotation System Based on User Attention Analysis

Jin-Young Moon and Changseok Bae
BigData Software Research Laboratory
Electronics and Telecommunications Research Institute
Daejeon, KOREA
{jymoon, csbae}@etri.re.kr

Wan-Chul Yoon
Department of Industrial and Systems Engineering
Korea Advanced Institute of Science and Technology
Daejeon, KOREA
wcyoon@kaist.ac.kr

***Abstract*—Automatic semantic annotation of videos is a crucial to the success of video search and summarisation based on content semantics. In contrast to broadcast news and sports, automatic semantic annotation for non-commercial videos generated by ordinary people suffers from lack of semantic data like subtitles and webcast text. It is, however, impracticable to expect that most video shooters or owners of the non-commercial contents do semantic annotation of their videos by using annotation tools manually before video dissemination. This paper proposes a video semantic annotation system that automatically analyzes and annotates a video element with the user attention state. The attention state includes the attention target, attention degree and emotional states by using gaze and Electroencephalography data from a user watching the video. To show the benefits achieved by the proposed system, the paper describes a promising application scenario of video summarisation using semantic annotations based on user attention. The use of annotations generated by the proposed system enables the summarisation system to enrich the possible summary types.**

Keywords-video annotation; semantic analysis; user attention; human factors; video summarisation

I. INTRODUCTION

Due to the huge rise in smartphone usage, shooting and sharing videos by ordinary people have been dramatically increasing nowadays. For example, the number of videos uploaded on YouTube has increased from 35 hours per minute in 2010 to 48 hours per minute in 2011. The number of videos uploaded in two months exceeds that of videos created by big three U.S. television networks (ABC, CBS, and NBC) in six decades [1]. Therefore, automatic or semi-automatic semantic annotation of videos that assigns semantics to various video elements, which can be a whole video, a scene, a shot, an objects or a regions, without a large burden to users, is vital to success of semantic video search and reconstruction among videos. Using the semantic annotation, the videos can be effectively shared by social media or re-created by the users to satisfy their personal needs.

Several techniques have been proposed to extract semantics automatically from broadcast news and sports by combining semantic data like subtitles, text from score box and webcast text [2]-[8]. It is infeasible not only to apply these techniques to non-commercial videos with the lack of those kinds of semantic data as it is but also to expect ordinary people who generate videos to do voluntarily semantic annotation of the videos by using semantic annotation tools before the users upload the video to social media, like YouTube or Facebook.

Therefore, we propose a video semantic annotation system that automatically assigns semantics related to the state of user attention to a specific object or region contained in a video, which a user focuses on, while the user watches the video. The proposed system analyzes an attention target by calculating the gaze position and recognizing the attended target, the attention degree, and emotional state by processing Electroencephalography(EEG) data from an EEG headset. The proposed system generates an annotation element on an attended target per user gaze together with the attention degree and emotional state when the user is attending to the target.

The rest of this paper is organized as follows. Section II provides an overview of the related work. In Section III, we show an overall architecture of the proposed system and describe main functionalities of system modules individually. Section IV focuses on an application scenario adopting the proposed system in order to show its feasibility and anticipated benefits. Finally, we conclude this paper with future work.

II. RELATED WORK

There have been proposed several manual semantic annotation tools for videos. They were introduced in detail and summarized in [6]. The tools enable users to do annotation delicately on the whole video, video segment during a specific time interval, a frame at a specific time, or a region in the form of a rectangle or a polygon within a frame in order to express its concept or the relationship with other elements. They are, however, inadequate for ordinary users because annotating with the manual tools is a difficult and time-consuming process to them without a definite advantage. Therefore, only automatic or semi-automatic semantic annotation techniques will be presented for consideration. In addition, we also examine user-related semantics like emotion and attention as well as content-related semantics.

A. From the perspective of content-related semantics

Extracting semantics for any videos only by computer vision and image processing is very challenging up to the present. Therefore, there have been some proposed techniques to detect events and do annotation automatically for sports and news videos because those kinds of videos have relatively affluent source of content-related semantics, for example video/audio channels and webcast text. Liu *et al.* [2] made use of periodicity of the video shot content and audio keywords of a racket sports video to detect rally events. Xu and Zhang *et al.* [3][4] and Refaey *et al.* [5] utilized a webcast text feature as well as video/audio features to increase the accuracy rate of event detection in soccer and baseball videos. Bagdanov *et al.* [6] created subtitles by using linguistic and dynamic visual ontologies with reasoning for a soccer video. Messina *et al.* [7] segmented a story by speaker and shot clustering and classified its subjects by using speech-to-text in the audio/video stream of news videos. Mezaris *et al.* [8] proposed a fusion technique to combine visual, audio and text analysis results. However, the proposed techniques are insufficient for most non-commercial videos, which do not follow standard patterns, like the rules of the games, and do not contain easily extractable semantic data, for example score boxes, subtitles and headlines, which are located in given positions in a sports or news video.

B. From the perspective of user-related semantics

Both limitation of extracting content-related semantics and necessity of user-related semantics for personalized contents search gave rise to work on extracting on user-related semantics and utilizing them on video reconstruction like emotion annotation on videos and user response-based video summarisation. While watching videos, users show their feelings unconsciously through facial expressions, eye movement and brain waves as a response of the video. Joho *et al.* [9] devised three pronounced levels of facial expression and the change rate of the expressions by analyzing a recorded video of users for affective video summaries. Money and Harry [10] analyzed physiological user responses including electro-dermal response, respiration amplitude, respiration rate, blood volume pulse, and hear rate. They revealed that most entertaining sub-segments within a video bring on an intense response of a user. Peng *et al.* [11] proposed Interest Meter to measure the viewing interest of a user by analyzing facial expressions, saccadic movements, blink, and head movement of the user for home video summarisation. Davis *et al.* [12] recorded EEG signals from EEG headsets as an emotional user feedback of video content related to explicit user tagging. Although the analyzed user-related semantics in [9]-[12] can be used for affective video summarisation, they are unable to be widely used for video reconstruction because the user state has no relationship with content-related semantics.

III. THE PROPOSED SYSTEM

To overcome shortcomings of both content-related and user-related semantics extraction, this paper proposes a video semantic annotation system that automatically analyzes user attention and annotates the recognized attention target within a video with the degree of the user attention and the emotional states. The system is largely composed of the module of annotation generation that analyzes user response related to attention and the module of annotation provision

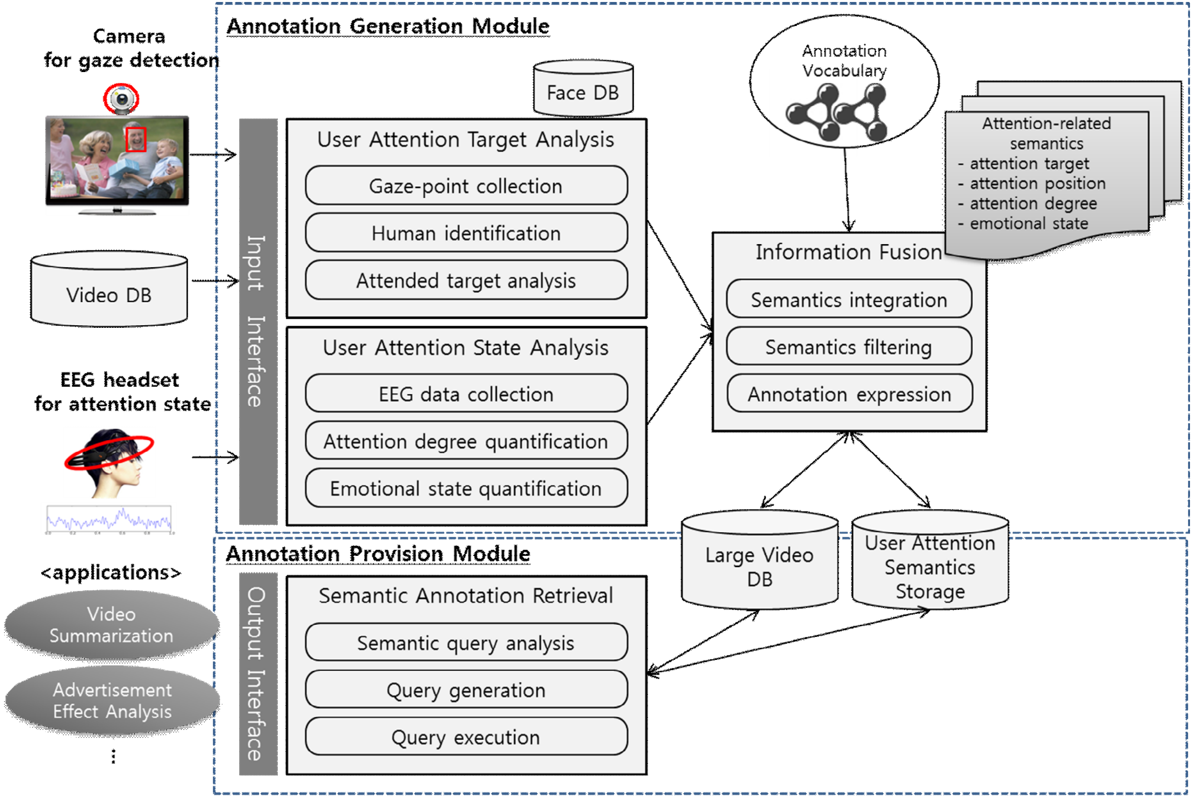

Figure 1. Architecture of the proposed system

module that provides a retrieval method of semantic annotation, as depicted in Figure 1.

The system selects the gaze point measured from a camera observing a user and the attentional and emotional user states analyzed by EEG data from a commercial EEG headset. There are other various user-related semantics, which include eye movement data, like fixation, saccadic movement and blink, and facial expressions recorded by a camera to observe the user. There are three reasons to use both gaze and brainwaves to analyze user attention. First, gaze and brainwaves reflect the delicate mental state of a user. For example, the emotion state of a user is not shown clearly in the face of the user generally. Second, in contrast to only using eye movement, the combined use of gaze and brainwaves can increase the accuracy of user attention analysis. Because mental imagination without external physical stimuli from videos can arouse eye movement, the other attention measure is necessary to screen the false alarm of attention data. Last of all, in contrast to only using brain waves, the combined use of gaze and brainwaves can assign a specific video element to user-related semantics by recognizing the attended target within videos.

In the module of annotation generation, the sub-modules of user attention target analysis and user attention state analysis collect raw data from a camera observing a user and an EEG headset in real-time and analyze the collected data in parallel. The sub-module of information fusion integrates attention data obtained from the two sub-modules of analysis to generate meaningful semantics related to the user attention and to filter redundant or false attention data. By using the annotation vocabularies, like ontologies or MPEG-7 metadata, the generated semantic annotations are expressed and stored with structural constructs in the annotation storage.

In the module of annotation provision, the stored annotations are provided to external applications to search for a video or a video segment according to the video semantics or to reconstruct videos, like video summarisation. The sub-module of semantic query analysis transforms user requests for annotations, annotated videos, or video segments into native queries that can be executed directly in the annotation storage.

IV. APPLICATION OF THE SYSTEM

To differentiate the effects of generated semantic annotations related to user attention combining content-related and user-related semantics from previous semantic annotation techniques related to either of them, this paper describes an application scenario that adopts the proposed system and summarizes a video or multiple videos from a video DB.

Shown in Figure 2, the application scenario is followed. A user watches a video wearing an EEG headset in front of a monitor equipped a camera observing the gaze of the user in order to generate semantic annotations by the proposed system. The every annotation includes the time-stamp to watch the video, the play-time within a video, an attended identified target of an object or a person within a frame, and the degree of attention and emotional state. We assume that the user owns the user's own face DB that contains friends or family who are shown frequently in the home video or public figures like celebrities. After automatic semantic annotation, the annotated video is analyzed to obtain overall user attention in it, like the general feelings of the video and the list of the most attended objects or people that draw interest of the user. The analyzed significant abstract information on user interest can be stored separately in the user interest DB.

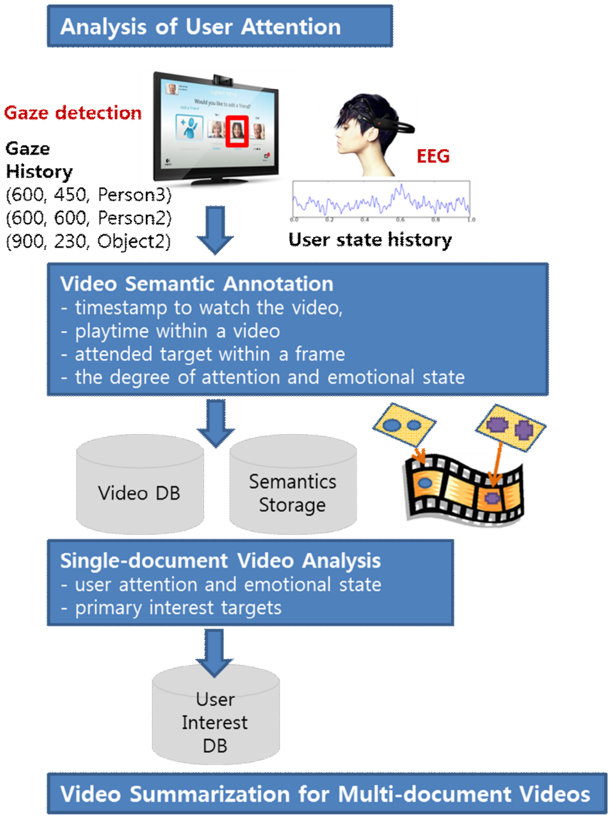
Analysis of User Attention
Gaze detection
Gaze History
(600, 450, Person3)
(600, 600, Person2)
(900, 230, Object2)
EEG
User state history
Video Semantic Annotation
- timestamp to watch the video,
- playtime within a video
- attended target within a frame
- the degree of attention and emotional state
Video DB
Semantics Storage
Single-document Video Analysis
- user attention and emotional state
- primary interest targets
User Interest DB
Video Summarization for Multi-document Videos

Figure 2. Application Scenario of Video Summarisation

The generated semantic annotations and analyzed information on a single video enable the video summarisation system to provide the new types of video summaries. Previous techniques for emotion annotation [9]-[12] can support summarisation of summary type 1-1 because their techniques of emotion annotations do not have any relationship with a person shown in the videos. The accumulated relationship with a person and the user state as a response corresponding to the person enriches the possible types of the video summary. First, summarisation condition can be set exquisitely by combining a person and its corresponding emotional properties like summary type 1-3. Second, the annotated video can be re-summarized at the current viewpoint, like summary type 2-1 and 2-3. Last but not least, the video can be summarized without annotating process by current interest obtained from annotations recently generated by the proposed system if some people are already recognized in advance, like summary type 3-1 and 3-3. Table 1 enumerates all the possible types of video summary by adopting semantic annotations generated by the proposed system.

TABLE I. POSSIBLE TYPES OF VIDEO SUMMARY

Time	**Condition by user attention**		
	Attended target	***User attention state***	***Both of them***
	summary reflecting user state when watching the video		
$T_A \approx T_S$	Type 1-1: Containing attended people that attracted the interest of the user at that time	Type 1-2: Containing parts where the user in a particular mood with concentration at that time	Type 1-3: Containing attended people that attracted the interest of the user in a particular mood at that time
	summary reflecting current user state		
$T_A \Rightarrow T_S$	Type 2-1: Containing attractable people included in the current interesting targets	N/A	Type 2-3: Containing attractable people included in the current interesting targets in a particular mood now
	summary reflecting current user state without annotation		
only T_S	Type 3-1: Containing attractable people included in the current interesting targets if the people are identified in advance	N/A	Type 3-3: Containing attractable people included in the current interesting targets in a particular mood if the people are identified in advance

(T_A: Time to annotate, T_S: Time to summarize, ⇒: precede in time, N/A: Not Applicable)

If the video summarisation system is used for home video that archives the daily life of a user with the user's family or special events for the family, the system can generate various video summaries of videos that were attracted at that time or are attractable now.

V. CONCLUSION AND FUTURE WORK

The proposed system was designed to analyze automatically the state of user attention with attended target, the degree of attention, and the emotional state. Compared to previous related work, the proposed system produces the relationship with a person shown in a video and its properties related to the user attention. The use of annotations generated recently by the proposed system enables the summarisation system to create a new type of a video summary from the current viewpoint. In addition, it enables the summarisation system to generate a new type of a video summary that the user has never watched if some people shown in video are identified in advance. That is how the proposed system enriches the summary types.

In the future, we will propose a novel algorithm to integrate and filter data related to user attention state and show its feasibility by the experiment of human subjects.

ACKNOWLEDGMENT

This work was supported by the Global Frontier R&D Program on <Human-centered Interaction for Coexistence> funded by the National Research Foundation of Korea grant funded by the Korean Government(MEST) (NRF-M1AXA003-2011-0028371).

REFERENCES

[1] Search Eingine Watch arcitlce, "New YouTube Statics: 48 Hours of Video Uploaded Per Miute; 3 Billion Vies Per Day", <http://searchenginewatch.com/article/2073962/New-YouTube-Statistics-48-Hours-of-Video-Uploaded-Per-Minute-3-Billion-Views-Per-Day> , 14. 02. 2012.

[2] Chunxi Liu, Qingming Huang, Shuqiang Jiang, Liyuan Xing, Qixiang Ye, and Wen Gao, "A framework for flexible summarization of racquet sports video using multiple modalities," Computer Vision and Image Understanding, Volume 113, Issue 3, pp. 415-424, ISSN 1077-3142, March 2009.

[3] Changsheng Xu, Jinjun Wang, Hanqing Lu, and Yifan Zhang, "A Novel Framework for Semantic Annotation and Personalized Retrieval of Sports Video," Multimedia, IEEE Transactions on, vol. 10, no. 3, pp. 421-436, April 2008.

[4] Yifan Zhang, Changsheng Xu, YongRui, Jinqiao Wang, and Hanqing Lu, "Semantic Event Extraction from Basketball Games using Multi-Modal Analysis," Multimedia and Expo, 2007 IEEE International Conference on, pp. 2190-2193, 2-5 July 2007.

[5] Mohammed A. Refaey, Wael Abd-Almageed, and Larry S. Davis, "A Logic Framework for Sports Video Summarization Using Text-Based Semantic Annotation," Semantic Media Adaptation and Personalization, SMAP '08. Third International Workshop on, pp. 69-75, 15-16 Dec. 2008.

[6] Andrew D. Bagdanov, Marco Bertini, Alberto Del Bimbo, Giuseppe Serra, and Carlo Torniai, "Semantic annotation and retrieval of video events using multimedia ontologies," Semantic Computing, 2007. ICSC 2007. International Conference on, pp. 713-720, 17-19 Sept. 2007.

[7] A. Messina, R Borgotallo, G. Dimino, D. Airola Gnota, and L. Boch, "ANTS: A Complete System for Automatic News Programme Annotation Based on Multimodal Analysis," Image Analysis for Multimedia Interactive Services, 2008. WIAMIS '08. Ninth International Workshop on, pp. 219-222, 7-9 May 2008.

[8] Vasileios Mezaris, Spyros Gidaros, Walter Kasper, Jörg Steffen, Roeland Ordelman, Marijn Huijbregts, Franciska de Jong, Ioannis Kompatsiaris, and Michael G. Strintzis, "A system for the semantic multimodal analysis of news audio-visual content," EURASIP J. Adv. Signal Process 2010, Article 47, February 2010.

[9] Hideo Joho, Joemon M. Jose, Roberto Valenti, and Nicu Sebe, "Exploiting facial expressions for affective video summarisation," In Proceedings of the ACM International Conference on Image and Video Retrieval (CIVR '09). ACM, New York, NY, USA, Article 31, 2009.

[10] Arthur G. Money and Harry Agius, "Analysing user physiological responses for affective video summarisation, Displays, Volume 30, Issue 2, pp. 59-70, April 2009.

[11] Wei-Ting Peng, Wei-Ta Chu, Chia-Han Chang, Chien-Nan Chou, Wei-Jia Huang, Wen-Yan Chang, and Yi-Ping Hung, "Editing by Viewing: Automatic Home Video Summarization by Viewing Behavior Analysis," Multimedia, IEEE Transactions on, vol. 13, no. 3, pp. 539-550, June 2011.

[12] S. Davis, E. Cheng, I. Burnett, and C. Ritz, "Multimedia user feedback based on augmenting user tags with EEG emotional states," Quality of Multimedia Experience (QoMEX), 2011 Third International Workshop on, pp. 143-148, Sept. 2011.