# A NOVEL USER-CENTERED DESIGN FOR PERSONALIZED VIDEO SUMMARIZATION

*Gheorghita Ghinea*[1], *Rajkumar Kannan*[2], *Sridhar Swaminathan*[2], *Suresh Kannaiyan*[2]

[1]Information Systems and Computing, Brunel University, Middlesex UB8 3PH, United Kingdom
george.ghinea@brunel.ac.uk

[2]Department of Computer Science, Bishop Heber College (Autonomous), Tiruchirappalli, India
rajkumar@bhc.edu.in, sridarah@gmail.com, sureshk.naga@gmail.com

## ABSTRACT

In the past, several automatic video summarization systems had been proposed to generate video summary. However, a generic video summary that is generated based only on audio, visual and textual saliencies will not satisfy every user. This paper proposes a novel system for generating semantically meaningful personalized video summaries, which are tailored to the individual user's preferences over video semantics. Each video shot is represented using a semantic multinomial which is a vector of posterior semantic concept probabilities. The proposed system stitches video summary based on summary time span and top-ranked shots that are semantically relevant to the user's preferences. The proposed summarization system is evaluated using both quantitative and subjective evaluation metrics. The experimental results on the performance of the proposed video summarization system are encouraging.

***Index Terms***— Video summarization, video semantics, semantic multinomial, personalization, user preferences, multimedia information systems, multimedia retrieval and browsing

## 1. INTRODUCTION

Video summarization aims at producing compact version of a full-length video while preserving the significant content of the original video. For generating a video summary, most of the *automatic video summarization* methodologies detect interesting or significant segments of a video based on certain criteria which are mostly audio, visual and textual saliencies [1]. However users always try to summarize videos based on the semantic content of a video, rather than saliencies alone. So the well-known semantic gap problem is thus shown to also exist in video summarization.

Generic video summarization will not be sufficient when the users' needs and interests change over a time. Users are seldom satisfied by a common video summary produced by the video summarization system as the produced video summary may not contain content of particular semantic concept or genre liked by a user. So the criterion used to summarize a video should be the user's preferences and interests over the video semantics. Personalized video summarization is a useful technique for producing the customized video summaries to the users based on their interests over video semantics.

Therefore our hypothesis is that a generic video summary does not satisfy every user and multiple video summaries for the same video should be generated depending on the preferences and needs of individual users.

The following are the important requirements for the proposed summarization system:

- To support an efficient video content management and personalized video browsing in large scale video information system for individual users
- To support effective resource management such as internet bandwidth in large scale web video information system, based on whether user's interest is to watch an entire video or not.

The aim of this paper is to propose a novel system for the personalized video summarization that produces the customized video summaries by adapting to the user's interests. The contributions of this paper are,

- A novel summarization methodology that generates semantically meaningful personalized video summaries using semantic multinomial and user preferences.
- A novel personalized video summarization system that tailors the summaries based on individual user's preferences over video semantics.

## 2. RELATED WORK

A wide number of contributions can be found in the area of video summarization. Various user attention based models have been proposed for summarization to make use of the users' perceptual response to low-level audio, visual and textual features [1-4], and the users' response while watching a video [5-7]. A multimodal saliency curve is constructed for movie summarization using a spatio-temporal saliency model, an AM-FM speech model and Part of Speech (POS) tagging for computing visual, aural and textual saliencies respectively [1]. Video features that easily attract users' attention and influence human perception, such as motion, contrast, special scenes and statistical rhythm, are extracted and modelled for summarization [2]. Audio, visual and linguistic attention models were used to generate the attention curve of a video for both static and dynamic video summarization [3]. The authors utilized both low-level attention models such as motion, static, camera and audio attention models and mid-level attention models such as face, speech and music attention models. Attention scores are computed using a motion attention model, and are attached to the scene, clusters, shots and subshots in a temporal graph for video summarization [4].

Most of the attention (or saliency) based video summarization methods utilize users' task-independent response or attention to audio, visual and textual modalities of a video. A major limitation with this class of approaches is that, they often fail to work well on videos with semantically rich content (like movie and sports videos). As they do not consider high-level video semantics, summaries generated based on saliency might not, however, contain semantically interesting or significant content of a video. For goal-oriented, task-specific video summarization, it is necessary to consider the semantics underlying a video and the users' requirements on summarization.

Variations in user's eye movement, blink, and head motion are considered for identifying interesting segments of a video [5]. Affective segments of videos and Regions of interest (ROIs) are discovered by analyzing the viewers' eye-gaze [6]. The authors in [7] presented an affective video summarization approach based on the facial expressions of viewers while watching the video. Facial expressions were analyzed to infer affective scenes from videos.

The effectiveness of these methods often depends upon the ability to capture users' responses and mapping of such responses to the corresponding video segments. Also, this class of methods always needs the controlled summarization setups. Same as the attention based video summarization methods; these methods do not consider the video semantics and users' requirements and thus suffer from less generalizability.

High-level video semantics can be used as preferences to the users for personalized video summarization [8-11]. Users' Degree of Interest on event, person, and object were used for personalized summarization of life-log videos in a multi-camera office environment [8]. This approach totally relies on manual annotation of events such as *working*, *eating*, *printing*, *meeting*, etc. High-level semantic concepts such as *humans*, *explosion*, *indoor*, *outdoor*, *close-up*, *zoom-in*, *moving objects*, etc were automatically detected from videos for personalized summarization [9]. The authors used a constrained optimization problem for selecting shots that are relevant to the users' preferences. The importance of a video segment is measured using users' constraints and preferences over audio-visual semantic concepts [10]. IBM research has proposed a personalized video summarization system for pervasive mobile devices such as PDA [11]. The User, device and transmission profiles were used for adaptive personalized video summarization and transmission; however, the system allows only a single visual semantic concept as binary preference at a time.

Most of these approaches explicitly obtain users' preferences over video semantics for personalized video shot summarization. The limitation with these approaches is that they support only binary valued preferences. Binary valued preferences might not be adequate when the user wants to relatively prioritize the preferences. Also computation of the similarity between video semantics and users' preferences, and the selection of ranked shots were marginally discussed in these works.

In contrast to the previous personalized video summarization systems, the proposed system supports multiple real valued preferences at a time. Also, the proposed summarization system provides a wide variety of semantic concepts as preferences to the users and uses efficient similarity measures as well as constrained shot selection scheme for personalized video summarization.

## 3. SYSTEM OVERVIEW

The architecture of the proposed video summarization system is shown in figure 1. The system consists of three modules: *pre-processing*, *user interface* and *video summarization*. Here, the database contains a collection of videos and their metadata.

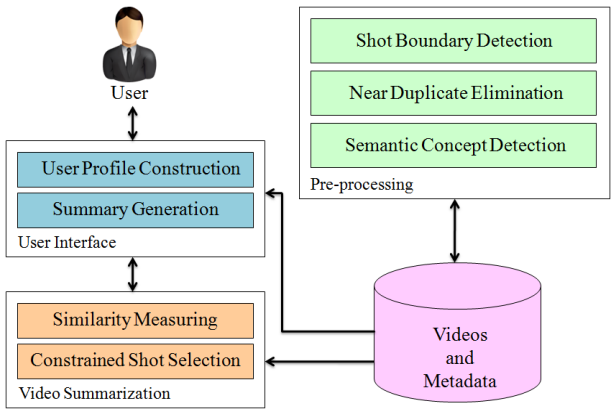


**Fig. 1.** Architecture of personalized video summarization system.

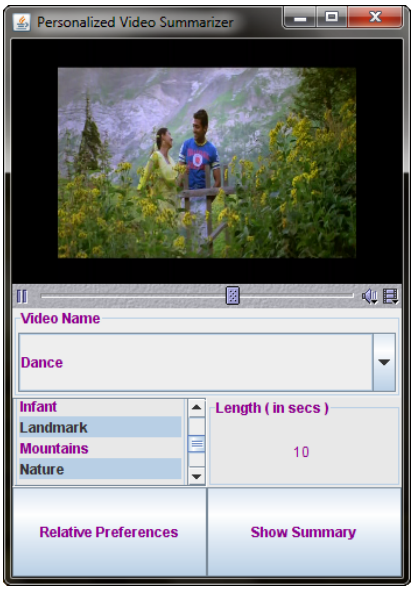


**Fig. 2.** User interface of the video summarization system.

The proposed system uses IBM's *Multimedia Analysis and Retrieval* System (*IMARS*) [12] for pre-processing. First, videos are segmented into set of shots using a combination of audio-visual features. This ensures smooth audio-visual transitions in the summary presentation. Since a keyframe can represent a shot, a single keyframe is extracted from each shot. Near duplicate shots in a video are identified using these keyframes. In a set of visually similar shots, shot that appears first in the video is kept and other near duplicate shots are removed, so that the summary will not contain more than one shot with similar visual content. Twenty five semantic concepts such as *beach*, *flower scene*, *indoors, etc.,* are detected from each keyframe. The *relevance scores* to each video shot for a set of semantic concepts are assigned. This score is a posterior concept probability ranging from -1 to +1 that shows the relevance between a shot and a particular semantic concept, where -1 implies *highly irrelevant* and +1, *highly relevant*.

The vector of semantic weights (i.e. relevance scores), denoted as the *semantic multinomial*, represents each image in a semantic space. The IMARS provides a diverse set of semantic concept detectors for visual scene categories that covers *places*, *people*, *objects*, *settings*, *activities* and *events*. So, the semantic concepts used are sufficient enough to represent a keyframe in a semantic space. Hence, the proposed summarization system can be used with videos of any genre from any domain.

Figure 2 depicts the user interface of the proposed video summarization system. The user interface assists the users to specify their preferences over 25 semantic concepts and to give a preferred summary length for a video. The set of semantic concepts preferred by a user will be considered as a *user profile*. The users can specify preferences using either *list* or *slider*. When using list, all chosen preferences

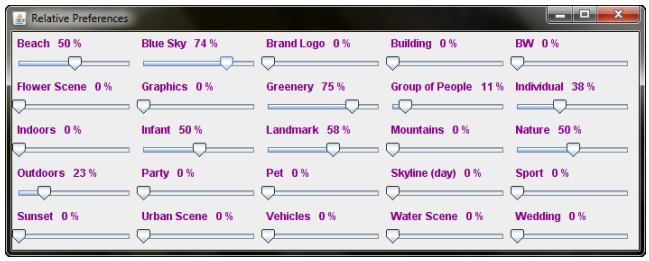


**Fig. 3.** Relative user preference panel.

will assume a numeric value 1 and rest will assume a numeric value 0. Since list allows only the binary preferences, sliders can be used to select real values from 0 to 1 (Figure 3).

In the video summarization module, *Dot Product* and *Cosine Similarity* measures are employed to determine likeliness between user's profile and semantic multinomial. Summary skimmed using top-ranked video shots and shown to the user.

## 4. SUMMARIZATION METHODOLOGY

Let video $V = \{u_i , 1 \le i \le n\}$ consists of $n$ shots, where each shot $u_i$ has duration $d_i$ seconds. Let $C = \{c_j , 1 \le j \le m \}$ denote a set of $m$ semantic concepts. For a shot $u_i$ , let $R_i = [r_{i1}, r_{i2}, r_{i3},\ldots, r_{im}]$ denote the *semantic multinomial*, containing the relevance scores of $m$ semantic concepts. Let $P= [p_1, p_2, p_3,\ldots, p_m]$ denote a vector (i.e. *user profile*), which consist of a set of weights for semantic concepts that are selected as preferences by the user. Each preference $p_j$ takes a value between 0 and 1. Let $T$ denote a summary time (in seconds) given by the user. Let $S_i$ be a *similarity score* computed for the shot $u_i$.

### 4.1. Shot Ranking

The similarity between the semantic multinomial and user preferences can be computed using any vector similarity measure. In this paper we compared both inner product similarity (Dot Product) and angular similarity (Cosine Similarity). Shots are ranked based on the similarity between the relevance scores and a user profile using either *Dot Product* or *Cosine Similarity*. For a shot $u_i$, the *Dot Product* between the semantic multinomial $R_i$ and a preference vector $P$ gives a similarity score $S_i$ . That is,

$$S_i = \vec{R}_i \cdot \vec{P} \tag{1}$$

For a shot $u_i$, *Cosine Similarity* is computed between the semantic multinomial $R_i$ and the preference vector $P$. That is,

$$S_i = \frac{\vec{R}_i \cdot \vec{P}}{\left\|\vec{R}_i\right\| \times \left\|\vec{P}\right\|} \tag{2}$$

### 4.2. Shot Selection

The objective of shot selection is to select shots that maximize the cumulative similarity score for the summary while not exceeding the time constraint $T$. This can be considered as an instance of 0-1 knapsack problem which is defined as,

$$\max \sum_{i=1}^{n} S_i \cdot x_i$$
$$\text{subject to } \sum_{i=1}^{n} d_i \cdot x_i \leq T \quad (3)$$

Here, $S_i$ is the similarity score computed for shot $i$ and $x_i$ is a binary decision variable that takes value 1 if $u_i$ is selected for the summary, otherwise 0. For quantitative evaluation, the selected shots are ordered decreasingly based on their similarity scores. For subjective evaluation, the selected shots are ordered by following the original video order. A summary is skimmed by concatenating the selected shots, and showed to the user with their corresponding audio.

## 5. EXPERIMENTS AND RESULTS

The proposed summarization system is evaluated both quantitatively and subjectively with song videos. Experiments were conducted with 10 song videos of total duration of 52 minutes which were collected from various web sources. A total of 1240 video shots were manually labelled for validation.

### 5.1. Experimental Setup

Since the relevance scores are in a range [-1,+1], and the system allows multiple preferences, a higher negative relevance score for a preferred semantic concept will reduce the similarity score in the Dot Product, even though a shot has many higher positive scores for other semantic concepts (*false negatives*). This will also increase the chances for the shots with lower negative relevance scores of semantic concepts to enter in the summary (*false positives*). As a solution, relevance scores in the range [-1,+1] can be normalized into a range [0,1]. Since this *range normalization* is the linear transformation of values, the effect will still remain the same. In order to solve this problem, negative relevance scores are ignored and are assumed to be zero. So, four types of summarization techniques were experimented. They are,

- Dot Product without negative relevance score (DP+)
- Dot Product with negative relevance score (DP-)
- Cosine Similarity without negative relevance score (CS+)
- Cosine Similarity with negative relevance score (CS-)

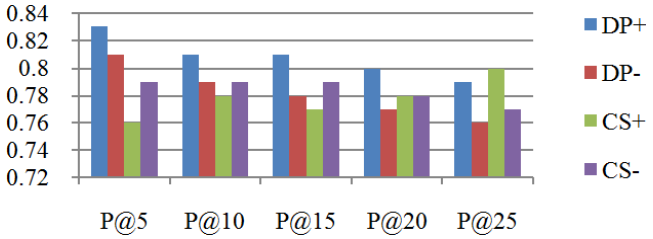


**Fig. 4.** Average of precisions for different single preferences.

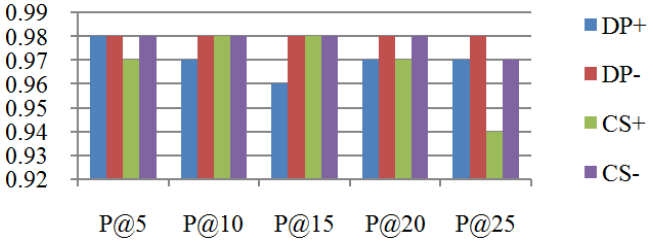


**Fig. 5.** Average of precisions for different multiple preferences.

Techniques 2 and 4 do not use negative relevance scores and set to zero.

### 5.2. Quantitative Evaluation

For each video, four different queries (single preference or multiple preferences) with average query complexity were considered. For this quantitative evaluation, selected shots are ordered decreasingly based on their similarity scores. Results are evaluated using *Ranked Precision* where precisions at intervals *n* are averaged for all the queries.

Figure 4 shows the comparison of the different similarity measures and their average of precisions when using a single semantic concept as preference. Dot Product similarity without negative score (DP+) performs better than others. The average of precisions when using multiple semantic concepts as preferences is shown in figure 5. It shows that Dot Product with negative relevance (DP-) score performs better than the other summarization techniques.

The Performance of the different summarization techniques can be directly assessed from the results of the quantitative experiments. When using a single preference in the summarization, Dot Product without negative relevance score (DP+) performs better than the others.

Different multiple preferences for all test videos were given to the system by using similarity measure as DP-. Averages of precisions at different average recalls are calculated for the system (Figure 6). Precision falls abruptly when recall reaches 0.3. To measure the ranking efficiency, the top ranked shots are manually graded using a scale of 0-3 to measure the average of *Normalized Discounted Cumulative Gain* (*nDCG*) at various positions. Figure 7 shows that, as the result set increases, nDCG decreases gradually.

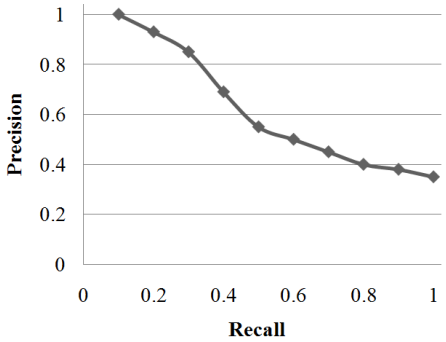


**Fig. 6.** Precision-Recall curve for different multiple preferences.

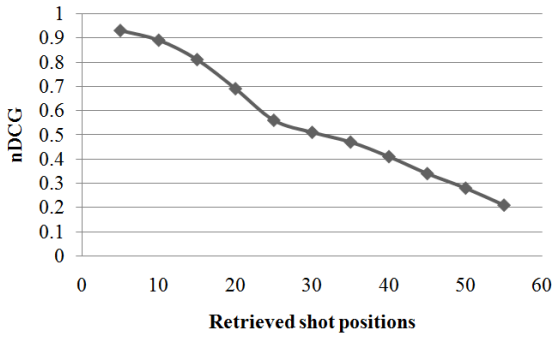


**Fig. 7.** Normalized Discounted Cumulative Gain for different multiple preferences.

Figure 8 shows the keyframes of the top 25 shots that are retrieved from a test video for a user preference "*flower scene*". It also shows some of the false positives in the resultant summary. This wrong prediction happens because, the color values and distribution of the actors' costume in that particular keyframe somehow resembles flowers.

### 5.3. Subjective Evaluation

The performance of the system was subjectively evaluated using a questionnaire with 20 test subjects (12 male and 8 female). For comparison, generic summaries consisting of significant shots for each video were created for a length of 1 minute. The subjects were asked to use the system to generate summaries of 1 minute length for the 10 test videos with different preferences. The summarization methodology was set as DP-. The subjects were not informed about the methodology used for summarization. For this subjective evaluation, the selected shots are temporally ordered, so that they will follow the original video order. When using the system, subjects were also shown generic summaries of the test videos. A questionnaire was prepared to comparatively evaluate the tailored and generic video summaries. The questions asked were:

- How informative was these summaries?
- How enjoyable was these summaries?
- Are these summaries relevant to your interests?
- How willing would you be to accept these summaries?

**Fig. 8.** Keyframes of the top 25 shots for the preference 'flower scene'.

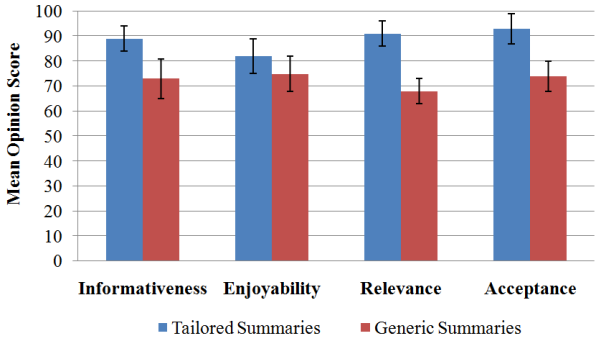


**Fig. 9.** Mean opinion scores for tailored and generic summaries.

These questions evaluate the summarization performance measures namely *informativeness*, *enjoyability*, *relevance* and *acceptance* respectively. For each measure, subjects were asked to rate the summaries of a specific type on a scale of 1-100.

Figure 9 shows the *Mean Opinion Score* (MOS) given by the subjects for the tailored and generic summaries under each qualitative evaluation measure. Since the proposed system meets the users' requirements for summarization, the proposed tailored summarization method performs better than the generic summarization under all the qualitative performance measures.

Quality of summaries was assessed by two *Quality of Perception* measures, which are *Quality of Perception - Information Assimilation* (*QoP-IA*) and *Quality of Perception - Satisfaction* (*QoP-S*) [13]. For these experiments, QoP-IA denotes the user's ability to assimilate information from a video summary, and QoP-S implies the user's satisfaction from a video summary. QoP-IA is measured by averaging the scores for the relevance and the information acquired from the summaries. QoP-S is calculated by averaging the scores of enjoyability and acceptance. For tailored summaries, QoP-IA is 90.4 and QoP-S is 87.1. For generic summaries, QoP-IA is 70.2 and QoP-S is 74.9. So, it can be seen that the tailored summaries are both informative and satisfactory than the generic summaries.

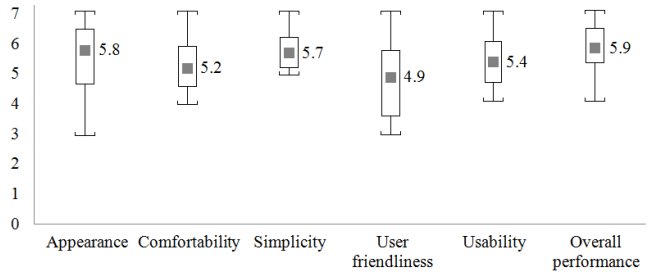


**Fig. 10.** Subjects' opinion on summarization system usability.

The usability of the system was subjectively evaluated with the test subjects. The *Computer System Usability Questionnaire* (CSUQ) [14] was used to measure subjects' Degree of Satisfaction (on a scale of 1 to 7) for the system and the user interface. The questions used are:

- The interface of this system is pleasant.
- I feel comfortable using this system.
- It was simple to use this system.
- It is easy to find the information I needed.
- Overall, I am satisfied with how easy it is to use this system.
- Overall, I am satisfied with this system.

These questions evaluate the system's usability under the criteria *appearance*, *comfortability*, *simplicity*, *user friendliness*, *usability* and *overall performance* respectively. Figure 10 shows the box plot of the Mean Opinion Scores given by the subjects on system usability. The results show that the system performs fairly well under all the system usability criteria.

## 6. CONCLUSION AND FUTURE WORK

This paper presented a novel preference aware video summarization system that produces semantically meaningful personalized video summaries by adapting to individual user's interests. Utilizing high-level feature extraction techniques reduces the manual effort in video semantics annotation. Video shot similarity measuring and constrained selection scheme guarantees efficient ranking and selection of relevant video segments for customization. The experimental results on the personalized video summarization demonstrate the effectiveness of the proposed system and the need for personalization.

The proposed user-centered design for personalized summarization can be extended to other specific domain of videos by incorporating domain specific video semantic annotation techniques. Though the system generates video summaries based on visual semantics, sometimes users may also be interested in choosing high-level audio and textual semantics as preferences. In future, the system would also consider audio and textual semantics as user preferences for personalized multimodal video summarization. Also, performance of the proposed system will be assessed with videos from more diverse range of genres.

## 7. REFERENCES

[1] G. Evangelopoulos, A. Zlatintsi, A. Potamianos, P. Maragos, K. Rapantzikos, G. Skoumas, and Y. Avrithis, "Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention," *IEEE Transactions on Multimedia*, vol. 15, no. 7, pp. 1553-1568, Nov. 2013.

[2] J. You, G. Liu, L. Sun, and H. Li, "A multiple visual models based perceptive analysis framework for multilevel video summarization," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no. 3, pp. 273-285, Mar. 2007.

[3] Y.F. Ma, X.S. Hua, and L. Lu, and H.J. Zhang, "A generic framework of user attention model and its application in video summarization," *IEEE Transactions on Multimedia*, vol. 7, no. 5, pp. 907–919, Oct. 2005.

[4] C.W. Ngo, Y.F. Ma, and H.J. Zhang, "Video summarization and scene detection by graph modeling," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, no. 2, pp. 296-305, Feb. 2005.

[5] W.T. Peng, W.T. Chu, C.H. Chang, C.N. Chou, W.J. Huang, W.Y. Chang, and Y.P. Hung, "Editing by viewing: automatic home video summarization by viewing behavior analysis," *IEEE Transactions on Multimedia*, vol. 13, no. 3, pp. 539-550, Jun. 2011.

[6] H. Katti, K. Yadati, M. Kankanhalli, C. Tat-Seng, "Affective video summarization and story board generation using pupillary dilation and eye gaze," *In IEEE International Symposium on Multimedia*, pp. 319-326, 2011.

[7] H. Joho, J.M. Jose, R. Valenti, and N. Sebe, "Exploiting facial expressions for affective video summarisation," *In Proceedings of ACM CIVR 2009*, Article No 31, 2009.

[8] H.S. Park, and S.B. Cho, "A personalized summarization of video life-logs from an indoor multi-camera system using a fuzzy rule-based system with domain knowledge," *Information Systems*, vol. 36, no. 8, pp. 1124-1134, 2011.

[9] W.N. Lie, and K.C. Hsum, "Video summarization based on semantic feature analysis and user preference," *In IEEE International Conference on Sensor Networks, Ubiquitous and Trustworthy Computing*, pp. 486-491, 2008.

[10] V. Parshin, and L. Chen, "Video summarization based on user-defined constraints and preferences," *In Proceedings of RIAO*, pp. 18-24, 2004.

[11] B.L. Tseng, C.Y. Lin, and J.R. Smith, "Video summarization and personalization for pervasive mobile devices," *Proceedings SPIE 4676, Storage and Retrieval for Media Databases 2002*, pp. 359-370, 2002.

[12] IBM's Multimedia Analysis and Retrieval System (IMARS), http://researcher.watson.ibm.com/researcher/view_project.php?id=877, accessed on Feb. 2014.

[13] S.R. Gulliver, and G. Ghinea, "Defining the users perception of distributed multimedia quality," *ACM Transactions on Multimedia Computing, Communications, and Application*, vol. 2, no. 4, pp. 241-257, Nov. 2006.

[14] J.R. Lewis, "IBM computer usability satisfaction questionnaires: psychometric evaluation and instructions for use," *International Journal of Human-Computer Interaction*, vol. 7, no. 1, pp. 57-78, 1995.