PERSONALIZED VIDEO SUMMARIZATION BY HIGHEST QUALITY FRAMES

Kaveh Darabi,Gheorghita Ghinea

Brunel University, London
{cspgkkd1,George.ghinea}@brunel.ac.uk

ABSTRACT

In this work, a user-centered approach has been the basis for generation of the personalized video summaries. Primarily, the video experts score and annotate the video frames during the enrichment phase. Afterwards, the frames scores for different video segments will be updated based on the captured end-users (different with video experts) priorities towards existing video scenes. Eventually, based on the pre-defined skimming time, the highest scored video frames will be extracted to be included into the personalized video summaries. In order to evaluate the effectiveness of our proposed model, we have compared the video summaries generated by our system against the results from 4 other summarization tools using different modalities.

Index Terms— Video summarization, Personalization, Upgrading frames scores.

1. INTRODUCTION

The growing amount of multimedia content has imposed the need for development of systems which are able to summarize videos of different genres automatically. Consequently, a considerable research effort has been allocated to this topic and various abstraction techniques have been developed. Broadly, two basic types of video summaries exist, static key-frames abstracts and dynamic video skims [1]. As a result of advanced audio-visual capturing tools, developing effective techniques to generate dynamic video skimming is becoming increasingly popular [2]. Generally, video summarization techniques comprise two phases: firstly, video segmentation in which a system aims to detect video shot boundaries; secondly, selection of the most important segments using their representative key-frames [3]. However, applying the regular video summarization methods will result in generation of identical video summaries for all viewers. It is important, though, to capture the user's interests and modify the video summary in a way that meets the user's requirements – in other words, to generate personalized video summaries. A personalized video summarization system then is designed to generate a shorter version of a video based on the user's preferences and interests while it retains the significant semantic content of the original video stream [4].

In this context, generating useful metadata and extracting the most valuable user's preferences and applying them to generate the video abstracts to address the end-users' expectations are the most important areas of research. Furthermore, exploiting appropriate summarization techniques that can be adopted alongside personalization methods to produce effective summaries based on the learned user's profiles represent another challenging topic. In this paper, we address this challenge and propose a framework to produce personalized video summaries based on video experts' assigned scores to video frames. Accordingly, the structure of this paper is as follows: Section II presents work related to our efforts; our approach is then detailed in Sections III and IV, whilst Section V presents evaluation results. Lastly, conclusions are drawn and opportunities for future work are identified in Section VI.

2. RELATED WORK

Previously, various automatic and semi-automatic approaches have been proposed for video summarization purposes. Considering the difficulties associated with understanding the semantic content of videos, most automatic video summarization methods rely on the saliency of low level visual [5], auditory [6] and textual features [7]. In semi-automatic approaches, human involvement is the determining factor in saliency detection of the video segments [8]. However, these techniques do not address the end-users' different priorities and expectations. Therefore, personalized video summarization topic is becoming increasingly popular in recent years. Work closely related to ours [9] employs MPEG-7 metadata, as well as user profiling alongside a supervised learning algorithm in order to generate personalized content. However, the effectiveness of this approach is limited to the availability of MPEG-7 metadata. In [10], a fuzzy rule-based system to approximate the human decision making process was applied for personalized summary generation task. The users inputted their degrees of interest in each event, person, and object so that the system could retrieve the target video. However, the knowledge domain data was provided through questionnaires which are extremely expensive time-wise.

The behaviour of the viewers is the determining factor in selection of the personalized content in [11]. Here, the attention level of users was measured while they were watching the videos by monitoring their operations on the remote controller of the video player as well as their eye movements. Human physiological responses such as respiration rate and blood volume pulse can also be the basis for generating personalized video summaries. Accordingly, the mentioned factors were monitored in order to measure changes in a user's affective state. Video segments which elicit significant physiological responses in the users are more likely to be interesting to a specific user and thus be included in the summary [12]. In both of the aforementioned methods, external distractive factors can deteriorate the end product dramatically, however. In a semi-automatic, manifold embedding-based approach, human subjects were asked to choose their preferred key-frames in an input video sequence, in order to overcome the barriers against detection of semantically rich video frames. Then, the visual summaries were constructed based on the inter-frame visual similarity to the pre-selected key-frames [13]. However, visually similar video segments might be semantically different. In a recent work, sketches have been the basis to represent the personalized summaries of the videos. Using an interactive selection method, users select the subjects in any frame and the visually similar key-frames are extracted from the video [14]. In a resource-allocation-based framework, playback speed and perceptual comfort have been the key elements for generation of personalized video summaries [15]. After segmentation of each video into segments and clips, a number of candidate sub-summaries were generated for each segment by assigning different combinations of playback speeds (from a set of discrete options) to each of contributing clips. The benefit for each sub-summary was computed by calculation of the base benefits of the corresponding clips and extra gain through satisfying specific preferences (inclusion of the user's favourite object, time duration, and story continuity). As several sub-summaries can be generated for each segment, the procedure of selecting the best sub-summaries can be computationally very expensive.

3. VIDEO SUMMARIZATION BY GROUP SCORING

In a proposed model in [16], a user-centered approach to video summarization based on a group scoring was suggested, in which original video frames are scored by a number of video scorers (experts) and the assigned scores averaged to produce a singular value for each frame. A group of frames with the highest average scores are then chosen to be inserted into the final summary. In this approach, the required number of video experts could be varied based on the different use-case scenarios. The proposed method was evaluated and shown to achieve promising results (vis. a vis. machine-generated approaches) in 6 different video categories. However, the generated summaries for all of the end-users were identical and their individual preferences were not envisaged in the summarization process. In this paper, we develop a model to personalize the final summaries in accordance to the individual end-user's expectations, and thus to produce a better user experience.

3.1. Video segments enrichment

For enrichment and scoring purposes a semi-automatic model has been applied in our framework. In the first step, the original videos are segmented into a number of scenes (group of semantically and visually similar frames). Later, each scene is enriched with a group of audio and visual tags and the appointment of a representative key-frame.

3.1.1. Shot boundary detection

AVCutty [17] as a typical scene boundary detection tool has been adopted to determine the timestamps for each contributing scene. It should be reminded that each scene in the context of a complete video plays the same role as a paragraph in a whole text. Therefore, there should be a semantic and visual correlation and cohesion between the existing frames of a particular scene. The mentioned tool utilizes the color and motion features of the video frames for scene change detection purposes. The required minimum time length for each scene has been set to 3 seconds. Thus, any identified video scene with shorter length will be added to the next scene. This facilitates scoring and annotating of the original video by reducing the number of unnecessary pauses for the enrichment task.

3.1.2. Video scenes annotation and scoring

In this stage, video experts are asked to score and enrich the video segments based on the auditory, visual and textual content of the video. As shown in Fig. 1, experts score the video frames 'on the fly' in a range between 0-10 using the Slider tool. Using the identified timestamps for the scene boundaries, the videos will be paused automatically at the end of each scene and the video experts immediately will be prompted to annotate the video scene using the provided graphical user interface shown in Figure 2 (while the scoring process is stopped). The video scorers can optionally enrich the video scenes while the videos are halted, by assigning audio and visual tags to each scene. These tags could contain information regarding the significant events, objects and any activities in the corresponding video scene. The video scorers have the possibility to choose the previously assigned tags (by former scorers) or to add new ones based on their personal perception and priorities to the scenes. Once the annotation process for one scene is finished, the scorers will then be engaged in scoring the video frames for the following scene

using the Slider tool. By re-starting the video, the initial frames from the upcoming scene is likely to be scored by unwanted grades. This is due to a predictable minor delay from the time in which video experts have to observe and evaluate the contextual significance of the opening frames (of the following scene) till the point they can actually start scoring. Therefore, to minimize the negative effect of this lag, a new pre-set value was dynamically calculated and assigned to the Slider tool each time that a scene starts. In order to produce this value, a score was computed for each scene, by averaging the previously assigned scores from the former experts to the whole frames of that particular scene. Any recent assigned scores from new scorers will update these computed average scores.

3.1.3. Key-frame selection for scenes

During the scene enrichment stage, the annotators (experts) are also presented with a set of 3 candidate key frames at the end of each scene. The video experts are asked to elect the one that they personally perceive as the highest quality to represent and summarize the semantic and visual content of that scene (shown in Fig.2). For extraction of these three nominated key frames, each video scene has to be fragmented into three equal shots in the first place, and each shot will be represented by a key frame (to improve the coverage rate of any visual content changes in whole scene). In order to select a key frame for each of these 3 identified video shots, two criteria should be considered. First, the frame has the highest assigned score between all the existing frames of that scene. Second, the candidate frame is temporally located in the middle of each shot. Therefore, between all the previously highest scored frames of each shot, the frame which is temporally closer to the center of that shot will be introduced as a potential key frame for that video shot (to increase the likelihood of extracting more visually significant and stable frames). These 3 nominee frames from each scene are then compared against each other from two different perspectives. Firstly, their visual content attractiveness and richness should be considered. Secondly, their capabilities in reflection of the semantic concepts of the corresponding video scene have to be taken into account. Finally, for each scene, the candidate frame that has the highest selection rate by different annotators will be selected as the representative key frame.

3.2. Capturing users' priorities

This phase is responsible for capturing an end-user's priorities in a particular video. As a result, prior to the generation of any final summary, end users will be provided with some visual and textual information regarding the content of the existing video scenes.

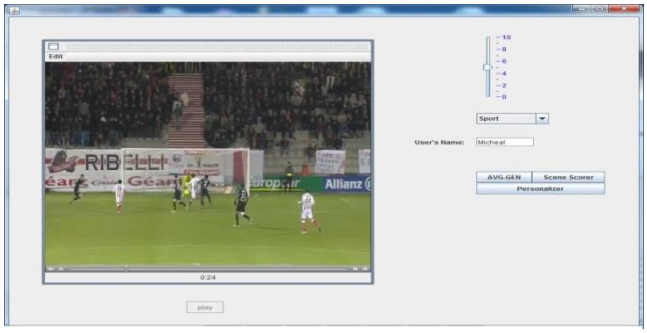

Fig.1. Interface for video experts to score the video frames

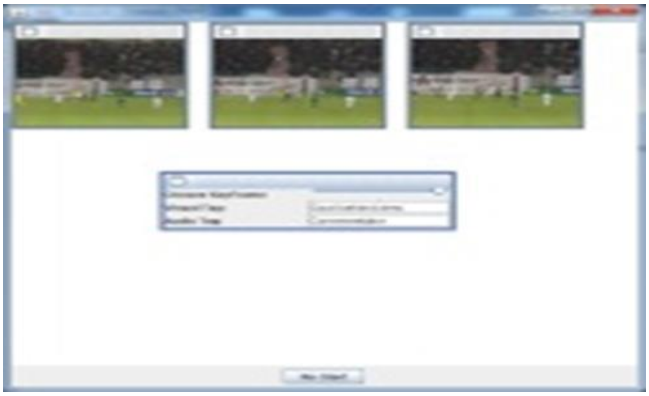

Fig.2. Interface to select a key-frame and annotate the scene

The goal here is to prioritize the video segments based on the users 'preferences and superiorities. Therefore, a list of representative key frames with their associated visual and audio tags is presented to the end users. Each of the displayed representative frames corresponds to a single video scene (these are the delegate key frames chosen by most of the video experts in the previous stage), while attached auditory and visual information to each key-frame correspond to the mostly verified tags for that scene by different video scorers (one audio content tag and one visual content tag per each scene). The end users will be asked to express their level of interest to each video scene, based on the displayed video frames and tags, using the provided slider tool (Fig.3). The users could choose 3priority levels for each scene. Level 0 has been considered for the scenes with the lowest level of significance to them, while level 1 is for the scenes with higher importance which were preferred to be included into final abstract. Level 2 designates the scenes that users found the most attractive and should be included with the highest priority into the final summary.

3.3. Updating the frames scores

In this phase, the initial generated average scores of the frames, assigned by the video scorers are updated based on the previously captured personal interests for each end-user.

Fig.3. Interface for end-users to prioritize the scenes

Therefore, based on the selected priority level for each scene by the end users, the primary average scores are updated. The scores of frames belonging to the scenes by the level 0 of interest will not be altered at all. However, in the scenes with a level 1 priority, the grades for the frames which their primary assigned scores are the highest among the frames of that scene, will be increased by 20 percent (to the maximum value of 12). This is done in order to potentially escalate the probability of incorporation of the highest quality frames of those scenes into the eventual video digest. The updated mark for the frames belonging to the scenes with the highest level of priority for a particular end-user will be recalculated in a different format. The grades for the frames which preliminary were scored the highest in each scene, will be upgraded to the maximum possible value (12). In fact, this would increase the chance of definite inclusion of the highest quality segments of those particular scenes (with level 2 priority) in the final summary. However, the marks for the frames of these scenes whose scores are not the highest but nonetheless manage to exceed the respective scene's average scores will be boosted by 20 percent as well (to the maximum of 12). The scores for the remaining frames of these scenes will remain unchanged.

4. PERSONALIZED SUMMARY GENERATION

In the final step, the personalized video summaries are produced based on the updated frames scores. In accordance to the summarization method based on group scoring, [16] the highest scored frames alongside the audio and textual content are selected and inserted into the final video digest. Considering the required number of frames, those highest scored frames will be selected to be added to a final list and to be sorted based on their time order in the original video. ReqNO calculates the required number of frames for extraction while TarVidTime shows the required video summary time:

$$ReqNO = \text{TarVidTime}(\text{seconds}) \times \text{FramesFrequencyScale} \quad (1)$$

So, if K represents the frame number in the original video, L is a list of chosen frames.

$$L = \left\{ \mathbf{F_K} \;/\; \mathbf{0 < K < ReqNO} \;\&\; \boldsymbol{AvgFra} \geq \boldsymbol{AvgFrame}_{\cup_{\mathbf{i=1}}^{\mathbf{N-ReqNo}} \mathbf{L'(i)}} \right\} \quad (2)$$

$$SortedFrames = \left\{ F_j \,\middle|\, 0 < j < ReqNo \;\&\; T_{F_j} > T_{F_{j-1}} \right\} \quad (3)$$

Using this sorted list, the temporally corresponding audio and text segments with those elected frames will be copied from the original tracks into the summary video. Considering that semantically and temporally close frames are usually similarly scored, the number of sudden cuts in the generated summary could drop significantly and video consistency and continuity are improved. As a result, more meaningful auditory and visual contents can be included in the final digest.

5. EXPERIMENTAL RESULTS

A group of short videos (2 minutes each) from 6 different video categories comprising, Movie, Sport, Documentary, Advertisement, Music and News genres were used to investigate the effectiveness of the proposed approach. 10 operators (vide experts) with different demographic details (5 Female and 5 Male within age range of 25-45) were asked to watch each of these 6 videos and to score and enrich the different segments of the videos based on their personal perceptions and preferences. As was mentioned in the last section, the experts have the option to select the previously assigned tags or to skip the annotation stage. However, they had to score the frames and to choose the representative key frame of each scene. The assigned scores for each frame were then averaged to generate a singular value for that frame. In order to produce personalized summaries, we adopted 30 end-users (15 Female and 15 Male within the age range of 20-60) to understand their priorities towards different scenes within the original videos. These users were of course different to the 10 experts who scored the videos initially. As explained, each of the users was provided with a collection of key-frames (each representing one scene) and the corresponding visual and audio tags per each video. Then, they were asked to select their level of interest in each scene of that particular video using the Slider tool. The previously generated scores by video experts were then customized based on the end-user's choices and personalized summaries were built for them based on the updated scores.

5.1. Analysis of the generated summaries

In order to assess the quality of our personalized video summarization approach, the generated results have been compared against the video abstracts produced by 4 other systems 3 of these tools summarize the videos automatically by assessment of different modalities and applying statistical and mathematical algorithms while the fourth tool, functions semi-automatically based on human involvement. In the first technique, the shots' semantic significance was measured by analyzing the audio-visual features. Therefore, audio, face and text importance analysis were carried out on each contributing shot. The results were then further enhanced by employing other factors such as camera motion, object motion and temporal motion coherence [18]. The second tool [6] abstracts the videos based on a saliency curve. The auditory, visual and textual information of the video segments were measured independently and were fused into a multi-modal saliency curve. Considering a predefined skimming percentage, the most salient video segments were inserted into final abstract. The third system, however, simply utilized the low-level visual features to produce the summaries [20]. Face detection (using Viola Jones algorithm), frames saliency detection (based on Ltti saliency and local entropy) and adjacent frames similarity measurement were applied to analyze spatiotemporal saliency of video segments. The fourth tool generates the summaries by averaging the assigned scores to the video frames (from a panel of video scorers) and selecting the highest scored ones considering the time constraint [16]. The 6 original videos alongside their 5 summary versions created by 5 existing tools (including the personalized summaries generated for each specific user using our proposed technique) were presented to the same 30 end-users on the basis of whose inputs their personalized summaries were created. These 5 summaries from each category were shown to the users in a random order so as to minimize order effects. Moreover, no information regarding the corresponding adopted summarization tools for each of the summary versions was revealed to participants. After watching the original video and the summaries the users were asked to score each of the generated abstracts awarding marks between 0 (worst video summary possible) to 10 (best video summary possible), from 4 different perspectives consisting of *Recall* (Re), *Precision* (Pe), *Timing* (Ti) and *Overall Satisfaction* (OS). Using *Recall*, the competence of the system in regards to coverage of the whole video was measured. In other words, it represents the extent to which the system reflects all the existing scenes from the original videos into the summaries. The *Precision* factor was adopted to calibrate the capability of the tools in selection of high quality segments from the original videos. As different end-users are likely to have different attitudes and opinions towards different scenes in a particular video, this criterion is tightly linked to the personalization aspects of the summarization process. *Timing* was utilized to test the level of temporal proximity of these built abstracts to the required summary length. Finally, *Overall Satisfaction* represents the overall users' experience and satisfaction from a number of perspectives including visual and aural coherency, continuity and adjustability. The given scores for each of these measures were averaged over 30 users and their mean values for each of the video categories are given in Table 1. S1, S2, S3, S4 and S5 indicate the average achieved scores by, respectively, the first, second, third, fourth and our proposed personalized systems. It should be reminded that the assigned scores are highly dependent on the visual, audio and contextual quality and characteristics of the original video. Therefore, the lower average grades for some videos are not necessarily tied to less efficiency of the summarization tools in those categories.

5.2. Validation of statistical results

Our proposed method has been scored highest from the *Precision* and *Overall satisfaction* point of views across all 6 existing categories. High Precision scores can justify the effectiveness of our method in producing the personalized results. As it can indicate that the video segments with higher priorities to each individual end-user have been identified to be inserted into the final digest at a considerable extent. Our model managed to deliver the best quality video digest among all 6 categories based on the average *Overall Satisfaction* marks. In order to validate the statistical significance of the assigned scores for our new proposed tool a t-test analysis has been adopted. These two main indicators were compared pairwise against the achieved scores by the other 4 systems and the results are displayed in Table 2. The outcome of this test highlights statistically significant differences (at the $p=0.05$ level) between the scored obtained by S5 (our new tool) and the other 4 summarization systems across these two measures. Generally, the S1 tool generates some good results in terms of *Recall* and *Precision*, however, the nature of this method leads to lower grades in terms of *Overall Satisfaction*. Summarizing the audio and video tracks separately and concatenation of static key-frames to generate slide shows thus have a negative effect on the general experience of end-users. The second method could achieve some good results for particular categories including the Movie and Music Video. However, the performance is considerably domain-dependent. The results for the fourth system enjoy acceptable user ratings over 6 different categories. However, lower scores for *Precision* and *Overall Satisfaction* are due to the inability of this method to actually generate personalized content.

3. Conclusion

In this paper, a new method for producing video summaries which can contribute in generation of a personalized video information system has been proposed. Experimental results

	S1				S2				S3				S4				S5			
	Re	Pi	Ti	OS	Re	Pi	Ti	OS	Re	Pi	Ti	OS	Re	Pi	Ti	OS	Re	Pi	Ti	OS
MOV	7.8	7.6	9.1	4.1	7.0	7.5	7.8	6.3	4.3	4.4	6.5	4.0	7.1	6.8	10	7.2	6.5	8.3	10	7.9
ADV	7.5	7.7	9.0	3.9	6.0	5.6	7.2	5.4	6.8	6.5	6.3	4.1	7.7	8.2	10	7.8	7.5	8.7	10	8.3
DOC	7.7	7.1	9.1	4.3	7.3	6.9	7.9	5.8	5.1	6.1	6.7	4.5	6.7	7.1	10	7.2	6.8	7.9	10	8.0
NEW	4.3	6.7	8.6	2.0	6.1	5.8	7.7	3.4	5.3	5.1	5.9	1.9	6.4	6.7	10	6.1	6.6	7.5	10	7.1
SPO	6.9	6.0	8.3	3.4	5.8	5.8	7.8	5.4	4.5	3.8	5.7	4.1	6.9	7.4	10	6.9	6.5	7.8	10	7.4
MUS	7.7	6.8	8.5	3.1	6.8	6.4	7.9	5.4	5.8	5.7	6.2	3.5	6.5	6.8	10	6.3	6.2	7.6	10	7.2

Table 1. Evaluation of our proposed tool against the other 4 systems

	S5-S4				S5-S3				S5-S2				S5-S1			
	Pe		OS		Pe		OS		Pe		OS		Pe		OS	
	T	P	T	P	T	P	T	P	T	P	T	P	T	P	T	P
SPO	2.3	0.012	2.34	0.025	13.88	1.2E-14	15.14	1.3E-15	7.68	9.04E-9	6.02	7.5E-7	3.23	0.0015	12.87	8.07E-14
DOC	3.18	0.0017	3.37	0.0010	5.57	2.5E-6	13.25	3.8E-14	3.68	0.0004	5.93	9.6E-7	2.11	0.021	15.00	1.6E-15
NEW	3.31	0.0012	4.11	0.0001	6.39	3.5E-7	14.98	1.7E-15	4.96	1.3E-5	11.48	1.2E-12	2.06	0.024	16.5	1.2E-16
ADV	2.46	0.009	2.64	0.006	6.89	7.1E-8	12.5	1.4E-13	10.84	5.0E-12	8.02	3.7E-9	4.25	9.9E-5	11.67	8.7E-13
MUS	4.32	8.2E-5	3.88	0.0002	5.4	4.1E-6	12.51	1.6E-13	3.19	0.0016	6.22	4.3E-7	2.10	0.022	9.91	3.9E-11
MOV	4.96	1.4E-5	2.91	0.0034	12.64	1.2E-13	10.14	2.3E-11	2.14	0.0203	4.05	0.00017	2.15	0.0196	11.03	3.3E-12

Table 2. Statistical significant test analysis from *Precision* and *Overall Satisfaction* point of views

indicate the effectiveness of this approach in delivering superior outcomes comparing to our previously proposed method and 3 other automatic summarization tools. However, proposing a method which requires a less end-user involvement is a topic for our future work.

7. REFERENCES

[1] W. Ren and Y. Zhu. "Video summarization approach based on machine learning". IEEE, Intelligent Information Hiding and Multimedia Signal Processing, pp.450-453, August 15-2009

[2] X. Li, "Image Annotation by Large Scale Content Based Image Retrieval", Proc. ACM Int'l Conf. Multimedia, pp. 607-610, 2006.

[3] R. Datta, "Content-Based Image Retrieval—Approaches and Trends of the New Age", Proc. ACM Multimedia Workshop Multimedia Information Retrieval, pp. 253-262, April 2005

[4] Y. Takahashi, N. Nitta and N. Babaguchi, "Automatic Video Summarization of Sports Videos Using Metadata", Advances in Multimedia Information Processing , Vol. 3332, pp 272-280, 2005

[5] R.M. Jiang, A.H. Sadka, and D. Crookes, "Hierarchical video summarization in reference subspace", Consumer Electronics, IEEE Transactions on, vol.55, no.3, pp.1551-1557, August 2009.

[6] G. Evangelopoulos, G. 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, Mar 2013

[7] C. Xu, Y. Zhang, G. Zhu, Y. Rui, Y. Lu and Q. Huang," Using Webcast Text for Semantic Event Detection in Broadcast Sports Video", Multimedia, IEEE Transactions on , vol.10, no.7, pp. 1342-1355, Nov 2008.

[8] M.M. Yeung and R.I. Yeo, "Video visualization for compact presentation and fast browsing of pictorial content", IEEE Transactions on Circuits and Systems for Video Technology and Systems for Video Technology, vol. 7, no. 5, pp. 771-785, October1997

[9] A. Jaimes, T. Echigo, M. Teraguchi and F. Satoh, "Learning personalized video highlights from detailed MPEG-7 metadata," Image Processing, 2002. Proceedings. 2002 International Conference on , vol.1, no., pp.I-133,I-136 vol.1, 2002

[10] H. Park and S. 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, Volume 36, Issue 8, Pages 1124–1134, December 2011

[11] A. Yoshitaka and K. Sawada, "Personalized Video Summarization Based on Behavior of Viewer," Signal Image Technology and Internet Based Systems (SITIS), 2012 Eighth International Conference on , vol., no., pp.661,667, Nov 2012.

[12] A. Money and H. Agius, "Analyzing User Physiological Responses for Affective Video Summarization." Displays (Elsevier), vol. 30, no 2, pp 59-70, 2009

[13] H. Bohyung, H. Jihun and J. Sim, "Personalized video summarization with human in the loop," Applications of Computer Vision (WACV), 2011 IEEE Workshop on , vol., no., pp.51,57, 5-7 Jan. 2011

[14] Y. Zhang, C. Ma, J. Zhang, D. Zhang, and Y. Liu," An interactive personalized video summarization based on sketches". In Proceedings of the 12th ACM SIGGRAPH International Conference on Virtual-Reality Continuum and Its Applications in Industry (VRCAI '13). ACM, New York, USA, 249-258. 2013

[15] F. Chen, C. Vleeschouwer and A. Cavallaro, "Resource Allocation for Personalized Video Summarization," Multimedia, IEEE Transactions on , vol.16, no.2, pp.455,469, Feb. 2014

[16] K. Darabi and G. Ghinea, "Video summarization based on group scoring", In proceeding of the 4th IEEE International Conference on Multimedia computing and Systems, Marrakech, 2014, PP. xxx-xxx

[17] http://www.avcutty.de/english/ (Accessed 25 December 2013)

[18] J. You, M. Hannuksela and M. Gabbouj, "Semantic audio-visual analysis for video summarization", IEEE Region 8 EUROCON , 2009

[19] M. Beom, L. Williem, and I. Park, Spatiotemporal Saliency-Based Video Summarization on a Smartphone, JBE, vol. 18, no. 2, pp.185-195, March 2013