Video Summarization by Group Scoring

Kaveh Darabi, Gheorghita Ghinea
School of Computing and Information Systems
Brunel University
London, United Kingdom
{cspgkkd1,george.ghinea}@brunel.ac.uk

Abstract— **In this paper a new model for user-centered video summarization is presented. Involvement of more than one expert in generating the final video summary should be regarded as the main use case for this algorithm. This approach consists of three major steps. First, the video frames are scored by a group of operators. Next, these assigned scores are averaged to produce a singular value for each frame and, lastly, the highest scored video frames alongside the corresponding audio and textual contents are extracted to be inserted into the summary. The effectiveness of this approach has been evaluated by comparing the video summaries generated by this system against the results from a number of automatic summarization tools that use different modalities for abstraction.**

Keywords- Video summarization, Saliency detection, Highest scored frames, user-centred

I. INTRODUCTION

The growing amount of multimedia content has imposed the need for development of systems which are able to summarize the videos of different genres automatically. Consequently, a considerable amount of research 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 skims 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] Various techniques have been applied for key-frame extraction purposes. Different summarization systems are being developed which use automatic and semi-automatic approaches. In all of these, an importance score should be computed for each segment by analyzing their various attributes including visual, audio and textual features [4]. These computed scores are used to rank the segments of videos and to select the most significant ones as a video digest. Sequential clustering algorithms [5], dynamic programming techniques like MINMAX and Iso-content [6,7] and motion patterns [8] are among the approaches that have been utilized for key-frame extraction using low-level characteristics. However, in some other studies, instead of adopting low-level features for key-frame selection, frame content is semantically analyzed and semantic context within each frame is modelled. While in the mentioned methods, the scores for different segment of the video are calculated automatically by systems based on the captured low-level or high-level features of the scenes, in semi-automatic approaches, human involvement is necessary for the assessment of the video scenes saliency. In this paper, a semi-automatic approach for generating the video summaries will be presented. The best scored frames by a group of users will be chosen to be inserted into the final summary as the representative key-frames.

II. RELATED WORK

. To create a perfect summary, some content-based summarization techniques have been developed to extract semantics of the video, [9]. However, understanding the semantic content of the video is far from the capabilities of today's intelligent systems. Therefore most of the current methods rely on low-level feature extraction methods [10] including color histogram features, edge histogram, visual, textual and aural features. [1]

In some works, low-level visual features alongside mathematical concepts like graphs and clustering for video summarization have been adopted. Different clustering algorithms were applied to cluster the video shots to create static or dynamic video summaries. Yeo [11] clustered the shots based on their temporal adjacency while Uchishahi [12] used YUV color histograms to cluster the shots using supervised clustering algorithms. In both methods representative frames from the highest scored segments were chosen using a frame packing algorithm. In contrast, [13] presents an attempt to use unsupervised clustering techniques, in which each video frame was labelled by a compressed chromaticity signature and a multi-level hierarchical clustering algorithm in conjunction with a trained Hidden Markov Models were used for the key-frame extraction purpose. In [14], a hierarchical video structure summarization using a Laplacian Eigenmap has been proposed. Considering Laplacian Eigenmap as an efficient way of information extraction [15], a reference frame subspace approach will be applied to the selected frames to measure their dissimilarity. Thus, the dissimilarity between any two frames can be modelled based on the difference of their dissimilarity vectors (a vector representing the dissimilarity between an image and all of

selected reference images) in Laplacian subspace. In another attempt [16], color histogram and gradient distribution were used as the image descriptors for each frame. Each video is divided into 2 second sequences with 1.84s overlap. In the pre-processing stage any video sequence containing any unwanted frame (determined based on the set of the unwanted descriptors) will be discarded. From the remaining shots the concatenation of mean and standard deviation of the feature vector for existing frames in each shot will be calculated and it will form the feature vector for that shot. Principle component analysis is applied to each shot to reduce the dimensionality and K-mean clustering algorithm will be adopted to cluster the visually similar shots. Changing the K the desired length of the video summary can be determined considering the equal length of each video shot. Finally, for each cluster the nearest video shot to the centroid of the cluster will be chosen as the representative for that cluster.

In other research, aural features have been the key elements to produce the summary. Using various speech recognition systems, transcripts of the video are retrieved and an inverted word index, phonetic index alongside a phrase glossary index will be created. In this system the audio pauses instead of shot boundaries have been used to segment the video. The importance score for each video segment can be computed by applying information retrieval techniques [17]. In another work, sport videos are abstracted using the audio features considering that, interesting events will lead to change in the speech excitement level. The percentage of excited speech in each audio segment will be computed alongside the energy level so the system will be able to compute the importance level of each video segment. [18] However, the noisy signal can produce some negative results.

Textual content has been another available resource for summarization purposes. In a novel work for sport video abstraction, a probabilistic latent semantic analysis (PLSA) is applied to cluster the webcast text into different categories. Later, words with the highest number of occurrence in each category are chosen as keywords to represent the event types. Sentences containing these keywords are text events [19].

Based on a method with human involvement, the user's spontaneous behaviors while viewing a video are captured to generate home video summaries. This system determines the importance level of different shots by measuring camera motion level plus movement of eyes and facial expression of the users while they are interacting with the video segments. [20]. However, this solution is beset by limitations in the accuracy of eye and face tracking techniques.

In the Click2SMRY framework, crowdsourcing has been the basis for the video summarization [21]. Here, each video is partitioned into equally-sized sub-segments (5 seconds each) and thereafter, the video workers were asked to identify the potential video highlights by holding the SPACE key on the keyboard while they were watching the original video. Therefore, each click was assigned to one corresponding sub-segment. Finally, based on the required length of summary, a number of these sub-segments with the highest selection rates by different workers were extracted to be inserted into the final summary. However, segmentation of a video shot solely based on the time element can increase the possibility of generating false results. This is due to the inability of this method to address the dramatic change in the visual and semantic content within each sub-segment.

III. VIDEO SUMMARIZATION

In our work, a human-based collaborative approach has been adopted to find the most valuable video segments to be included into final summary. Thus, considering the shortcomings of the existing fully-automatic summarization techniques to detect semantic concepts to a satisfactory level and their drawbacks in terms of domain dependencies, user-based methods should be regarded as the best available option for determining the most salient video segments. This, due to human's capabilities to determine semantically meaningful video content. Furthermore, the human's brain has the ability to assess and compare the quality of each section of the video in the context of whole video as opposed to most of the automatic summarization systems reviewed in the previous section. However, the personal inclinations and preferences of different people can be dramatically different, which will have a direct influence on their content selection. Therefore, adopting a group of operators can increasingly reduce the effect of subjectivity of sole actors and it will smooth the final video summary towards more satisfactory results for a wider range of audiences. Moreover, there are a number of scenarios in which, more than one expert is required and engaged in the video abstraction process. In order to create a solo video abstract, the video summaries generated by each of these experts should be compared to each other and a third party has to select the overlapping segments which can be very time-consuming. For instance, in the "Match of the Day" TV show in which highlights of the English Premier League football matches are shown, football pundits extract the most interesting scenes of a football match and include them in a summary. However, each of these pundits can produce their own version of summaries based on their personal interests and perceived significance of different sections of the video. Generating a single final summary in which the views and choices of different experts have been contemplated and reflected aggregately is another application of this method.

A. Frames Saliency Detection

In our approach a group of short videos from different categories are presented to M different operators (experts). In the first instance, the operators watch the videos with the sole purpose of familiarizing themselves with the subject matter and do not score them. In the next step, the same people are asked to score those videos whilst watching them. To do this, they indicate the scores using a slider with the value range of 0 to 10. The operators are asked to score the video frames based on their personal interests and the perceived significance of the content they were watching. Fig. 1 shows the scoring process of the video frames. This group was also advised to consider the different available modalities (audio, visual and textual) for scoring purposes. Therefore, per each N available frames in the original video there will be N assigned scores

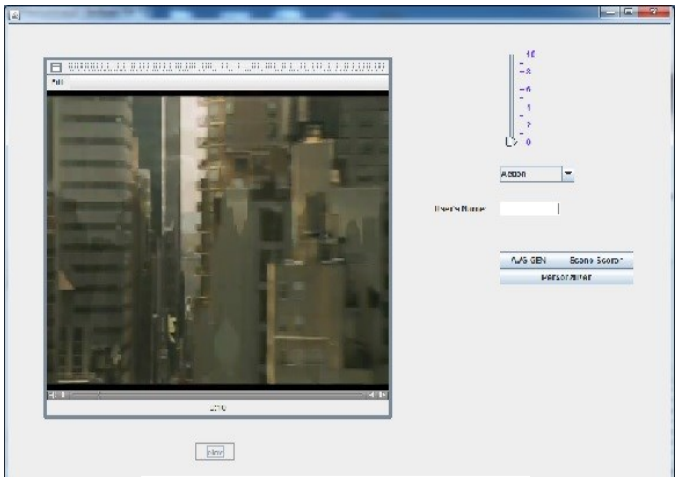


Figure1. Video Frame Scoring Interface

between 0-10 per each operator. The most satisfying frames will be scored with 10 and the least important sections are graded 0. $FrameScore_{NM}$ represents the value allocated to the N^{th} frame of the video by the M^{th} scorer. As opposed to the Click2SMRY framework [21] in which the video sub-shots had to be categorized as either highlights or non-highlights, in our proposed model, the panel of scorers is able to express their perceived importance level of each video frame.

B. Summary Generation

In the next step, the scores generated by all operators for all frames will be averaged and a single value will be computed for each single frame inside the original video. This represents the overall perceived quality of that particular frame across all M operators. $AvgFrame_N$ is computed as:

$$AvgFrame_N = \frac{\sum_{N=1}^{M} FrameScore_{NM}}{M} \quad (1)$$

The averaging process thus employed to smooth the frame scores towards a less biased result by reducing the effect of dramatic differences in assigned scores to a particular frame. The target video summary time and the video frames frequency scale (number of frames in 1 second) are the elements to determine the number of extracted frames. $ReqNO$ calculates the required number of frames for extraction while TarVidTime shows the required video summary time.

$$ReqNO = \text{TarVidTime}(\text{ seconds}) \times \text{FramesFrequencyScale} \quad (2)$$

In the final stage, the highest scored frames alongside the audio and textual content are selected and inserted into the final video digest. Thus, all the frames are sorted based on their ReqNO values. 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. So, if K represents the frame number in the original video, L is a list of chosen frames.

$$L = \left\{ F_K \mid 0 < K < ReqNO \;\&AvgFra \geq AvgFrame_{\bigcup_{i=1}^{N-ReqNo} L'(i)} \right\} \quad (3)$$

$$SortedFrames = \left\{ F_j \mid 0 < j < ReqNo \;\&\; T_{F_j} > T_{F_{j-1}} \right\} \quad (4)$$

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 usually are scored considerably similar, the number of sudden cuts in the generated summary will drop significantly. As a result, more meaningful auditory and visual contents could be included in the final digest.

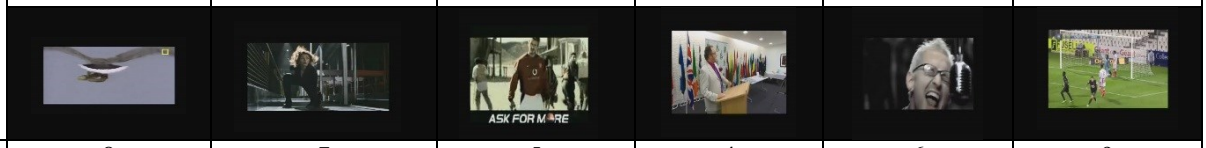
IV. EXPERIMENTS AND EVALUATION

A group of short videos (2 minutes each) from 6 different video categories comprising, *Movie, Sport, Documentary, Advertisement, Music* and *News* genres were utilized to assess the quality of the proposed method. 10 operators with different demographic details were asked to watch each of these 6 videos first and to score different sections of the videos based on their personal interests and preferences. User-assigned scores for each frame were then averaged and a 30 seconds video summary was generated by aggregating the top-scoring frames of the respective video. In Table I sample frames from each of the video clips alongside the corresponding assigned scores assigned by the first three operators have been presented.

A. Analysis of theGenerated Summaries

To measure the quality of the generated summaries, a comparison method has been adopted. So, these summaries were compared against the abstraction results of the same videos which were built by 3 automatic video summarization systems. These systems perform the video abstraction task by analyzing different modalities and employing different algorithms. In the first technique [22] summarization is based on audio-visual analysis. Shots are semantically measured using the semantic audio importance analysis; this is complemented by face and text importance. . Hence, other factors including camera motion, object motion and temporal motion coherence are also taken into account to build a semantic shot importance model. In the second system [17, 23] saliency of auditory, visual and textual information will be analyzed separately and will be integrated into a multi-modal saliency curve. Then most salient audio and video sub-clips based on a predefined skimming percentage will be chosen for inclusion in final summary. However, in the third system [24], low-level visual features are adopted solely for abstraction purposes. The similarity between adjacent frames, face region, and frame saliency are computed to analyze the spatiotemporal saliency in a video clip. The spatial saliency is calculated based on ltti saliency and local entropy of the video and Face detection measurement using Viola Jones algorithm.

TABLE I. ASSIGNED SCORES TO A SAMPLE FRAME BY 3 USERS

	DOC	MOV	ADV	NEWS	MUS	SPO
						
User1	8	7	5	4	6	9
User2	3	2	9	1	6	7
User3	7	7	5	5	4	8

Four created summaries of each of these 6 video categories were represented to 20 end users (10 Female and 10 Male within the age range of 25-55). These users were different to the initial 10 operators used to create the user-centric summaries and had no prior knowledge regarding any of the systems which had produced the summaries. After watching the original video and the summaries (the summaries were presented in a randomized order to participants, to avoid order effects), they will be asked to score each of these abstracts from 4 different perspectives consisting of *Recall* (Re), *Precision* (Pe), *Timing* (Ti) and *Overall Satisfaction* (OS). The best score will be 10 and the lowest is 0. In the first place, *Recall* measures the capability of the system in the terms of the full coverage of the whole video or in the other term, the extent that the system can reflects all the existing scenes from the original videos into the summaries. Secondly, *Precision* evaluates the ability of these systems in insertion of the most important scenes of the initial videos into the summaries. *Timing* explains the level of temporal proximity of these built abstracts to the required summary length. Finally, the *Overall Satisfaction* score represents the extent to which the end users are satisfied with the summaries from different points of views including visual and aural coherency, continuity and adjustability.

The given scores for each of these measures are averaged over 20 users and the final figures for each of the video categories are given in Table II. S1, S2, S3 and S4 show the results generated by, respectively, first, second, third and our proposed model.

B. Validation of the Statistical Results

To check the statistical significance between the grades achieved by our system and the other 3 tools, a t-test analysis were adopted. The results from this test (Table III) show that there are statistically significant differences between user ratings given to the summaries created by S4 (our user-centric summarization tool presented in this paper) and those created by the 3 other automatic summarization tools. Generally, *Recall*, *Precision* and *Timing* rates for the first system across all 6 categories have been high. However, the *Overall Satisfaction* has been the lowest between all 6 videos. It could be due to the nature of this method in which the audio and video are summarized separately. However, the extracted static key-frames are concatenated in a slide-show style and then will be combined by summarized audio later. The second method could achieve some good results for particular categories including for the Movie and Music Video; however the performance was considerably domain-dependent. The results generated by our proposed method scored the highest marks in terms of *Overall Satisfaction* and *Timing* in all 6 categories in spite of some average *Recall* results.

V. CONCLUSION AND FUTURE WORK

In this paper, a number of existing summarization methods were reviewed and a novel, user-centric model for video summarization was proposed. In our work, a group of operators are employed to score the video scenes as they are watching them.in the next step, their scores will be combined to come up with single value for each video frame. The highest scored frames by considering the 30 seconds time constraint will be extracted to be inserted into the final summary. The proposed method was evaluated by employing 20 end-users to compare its generated results against the summaries created by three existing automatic summarization systems. Experimental results indicated that the proposed approach is capable of delivering superior outcomes in terms of Overall Satisfaction and Precision indicating the usefulness of involving user input in the video summarization process. However, the proposed system scores low in terms of Recall rate, and we will seek to address this shortcoming in the future. Moreover, we also identify the production of personalised summaries as part of our future endeavours.

TABLE II. COMPARISON OF GENERATED SUMMARIES

	S1				S2				S3				S4			
	Re	**Pe**	**Ti**	**OS**	**Re**	**Pe**	**Ti**	**OS**	**Re**	**Pe**	**Ti**	**OS**	**Re**	**Pe**	**Ti**	**OS**
DOC	8.1	7.5	9.3	**4.15**	7	6.55	8.1	**5.7**	6.3	6	6.7	**4.8**	6	6.9	10	**7**
MOV	8.2	8.7	9.2	**4.45**	7.5	7.2	7.7	**6.1**	4.3	4.1	6.2	**4.5**	7.1	7.1	10	**7.3**
ADV	7.6	7.6	9.3	**4.2**	6.6	5.9	8.2	**6.1**	7.4	7	6.3	**5.1**	7.6	8.6	10	**8.4**
NEW	7.3	7.1	9.1	**2.15**	6.5	6.2	7.6	**3.8**	6.1	5.2	6.4	**2.4**	5.9	6.9	10	**6.3**
MUS	7.1	7.6	8.6	**2.7**	6.7	6.4	7.4	**5.3**	6	6.1	5.8	**3.5**	6.4	6.8	10	**6.4**
SPO	7.7	6.7	8.6	**3**	6.1	5.8	7.9	**5.2**	4.5	3.5	6	**3.4**	6.7	7	10	**6.8**

TABLEIII. OVERALL SATSISFACTION – SOLUTIONS COMPARISON

	S4-S1 (OS)		**S4-S2 (OS)**		**S4-S3 (OS)**	
DOC	T=10.72179	P= 1.69317E-09	T=4.950904	P=8.87201E-05	T=7.783993	P=2.51058E-07
MOV	T=7.024623	P= 1.09232E-06	T=2.534272	P=0.020221946	T=9.227575	P=1.89038E-08
ADV	T=11.15973	P= 8.73996E-10	T=5.510495	P=2.57535E-05	T=9.265785	P=1.77158E-08
NEW	T=14.6364	P= 8.4683E-12	T=9.561425	P=1.0785E-08	T=15.5836	P=2.80727E-12
MUS	T=7.241166	P= 7.12538E-07	T=3.942772	P=0.000873194	T=6.468593	P=3.36785E-06
SPO	T=11.10292	P= 9.51224E-10	T=5.051034	P=7.09458E-05	T=12.49933	P=1.29746E-10

ACKNOWLEGMENT

The authors wish to appreciate the help and support of all users involved in the scoring and evaluation stages of the project.

REFERENCES

[1] Y. Guo , Y. Zhu , F. Liu, C. Song and Z. Zhou , "Multi-view video summarization, Multimedia", IEEE Transactions on multimedia , Issue: 7 ,PP. 717 – 729, November 2010.

[2] C. Ngo, Y. Ma and H. Zhang, "Video summarization and scene detection by graph modelling", Vol 15 , Issue: 2 ,PP. 296 – 305, Feb. 2005.

[3] 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.

[4] C.M. Taskiran, Z. Pizlo and D. Delp, "Automated video program summarization using speech transcripts", Multimedia, IEEE, Vol 8, Issue 4, pp. 775 – 791, Aug. 2006.

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

[6] 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.

[7] J. Hays and A. Efros, "Scene Completion Using Millions of Photographs", Proc. ACM SIGGRAPH, 2007.

[8] J. Li, and J. Wang, "Automatic Linguistic Indexing of Pictures by Statistical Modeling Approach", IEEE Transaction on Pattern Analysis and Machine Intelligence, vol. 25, no. 9, pp. 1075-1088, Sept. 2003.

[9] Y. Takahashi, N. Nitta and N. Babaguchi, "Video Summarization for Large Sports Video Archives", Multimedia and Expo, ICME 2005, pp. 1170 – 1173, July 2005.

[10] S.L. King, and M.R. LYU, "Video summarization by video structure analysis and graph optimization", ICME, pp.1959-1962, July 2004.

[11] 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, vol. 7, no. 5, pp. 771-785, October1997.

[12] S. Uchihashi, J. Foote, A. Girgenson and J. Boreczky, "Generating semantically meaningful video summaries", 99 Proceedings of the seventh ACM international conference on Multimedia (Part 1), pp. 383-392, Nov 1999.

[13] C. Lu, M. Drew and J. Au, "Classification of summarized videos using hidden Markov Models on compressed chromaticity signatures", Proceeding of 9th ACM international conference on multi-media , PP. 479-482, Oct 2001.

[14] 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.

[15] M. Belkin and P. Niyogi, "Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering", Advances in Neural information Processing Systems, pp.14, 2001.

[16] P. Zemcik, J. Potucek, S. Sumec, A. Herout, M. Beran, V. Chmela, L. Lanik and M. Mlich, "Video summarization at Brno University of Technology", ACM,In prceeding of TRECVID 2007 rushes video summarization, SEP 2007.

[17] 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.

[18] R. Radhakrishnan, M. Siracusa, A. Divakaran and K. Hidetoshi, "An Enhanced Video Summarization System Using Audio Features for a Personal Video Recorder", Isao Otsuka, Mishima TR2006-024 February 2006.

[19] 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.

[20] W.T. Peng, "A User Experience Model for Home Video Summarization", Advances in Multimedia Modeling, 2009.

[21] S. Wu, R. Thawonmas ang K. Chen, "Video Summarization via Crowdsourcing", Proceeding CHI '11 Extended Abstracts on Human Factors in Computing Systems, pp. 1531-1536, May 2011.

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

[23] G. Evangelopoulos, A. Zlatintsi, A. Potamianos, P. Maragos , K. Rapantzikos, G. Skoumas and Y. Avrithis, "Video Event Detection and Summarization Using Audio, Visual and Text Saliency", Proc. IEEE Int'l Conf. on Acoustics, Speech and Signal Processing (ICASSP-09), Taipei, Taiwan, Apr. 2009.

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