An Approach to Personalized Video Summarization Based on User Preferences Analysis

Maria Miniakhmetova and Mikhail Zymbler
South Ural State University, Faculty of Computational Mathematics and Informatics
Chelyabinsk, Russia
miniakhmetovams@susu.ru, mzym@susu.ru

***Abstract*—Video summary is a sequence of still or moving pictures that represents the content of a video. Personalized summary provides a person with brief information reflecting essential message of the video according to his/her interests. Existing methods of discovering user's personal interests often demands from the user either extra efforts or extra equipment, e.g. manually setting up relative preferences or camera to capture of eyes movement. The paper presents an approach to constructing personalized video summary utilizing user's "like/neutral/dislike" estimations of videos watched beforehand. Summary is built as a sequence of scenes extracted from the video, which are most influencing the user. Most influencing scene contain a set of objects detected on video, which are in range of user's interest. Formal definitions of most influencing scene and range of interest are given and mathematical model of constructing personalized video summary is described.**

***Index Terms*—video data mining, personal video summarization, scene extraction**

I. Introduction

Video summary is a sequence of stills or moving pictures presenting the content of a video [1]. Personalized summary provides a person with brief information reflecting essential message of the video in accordance with his/her interests. Existing methods of discovering user's personal interests often demands from the user either extra efforts, e.g. setting up relative preferences [2], manual selection of keyframes [3] or extra equipment, e.g. to measure user physiological response [4], or to capture of eyes movement [5].

In this paper we present an approach to constructing personalized video summary, which utilizes user's "like-neutral-dislike" estimations of videos watched beforehand. Summary is built as a sequence of scenes extracted from the video, which are most influencing the user. Most influencing scene contain a set of objects detected on video, which are in range of user's interest. The rest of the paper is organized as follows. Section II discusses related work. Section III gives the formal definitions and describes the mathematical model of constructing personalized video summary. Section IV contains concluding remarks and directions for future research.

This work was financially supported by the Ministry of education and science of Russia ("Research and development on priority directions of scientific-technological complex of Russia for 2014-2020" Federal Program, contract No. 14.574.21.0035).

II. Related Work

A substantial amount of work has been done in the area of video data mining including the task of video summarization [1], [6]. *Video summarization* (or video abstracting) is the mechanism that allows the user to gain certain perspectives of a video document without watching/addressing the video in its entirety [7]. According to this research video summary could be either static or dynamic.

Static video summary is a simple set of keyframes which are extracted from the underlying video source. *Dynamic video summary*, also called video skim, consists of a collection of video segments (and corresponding audio) extracted from the original video [7].

Depending on the sources of information used for summary construction there are three types of video summarization techniques [6].

Internal summarization techniques in which technical information is used as the source of data for video summarization. This could be image features extracted from video frames, audio features, text analysis information, even codec-specific video file's metadata. In [8] an approach to create static video summaries by means of color feature extraction from video frames based upon k-Means clustering algorithm [9]. In [10] similar method for constructing dynamic video summary is described.

External summarization techniques analyse external (user-based) information during any stage of the video lifecycle [4]. This approach uses information about user's physiological responses to determine memorable or emotionally engaging video content for a given user. It is assumed that these data then could be used for constructing video summaries specifically for this user. Another interesting study [11] presents an approach to video summarization based on analysis of facial activity.

Hybrid summarization techniques use both internal and external information in video summarization process. For example, in [5] audio-visual cues and textual annotation to detect important/informative events were used. For personalized video summary construction then eye movement and operation of remote controller of video player capturing has been used to analyse viewer's behavior while watching a video.

As it was mentioned above there are summarization techniques that allows to create personalized video summaries which are the summaries constructed for the particular person according to his/her personal preferences [4], [11]. While analyzing existing approaches to personalized video summarization we have divided them into several classes.

Techniques based upon usage of extra devices involve special equipment to collect some physiological data about particular user while he/she is watching video [4], [11]. This could be electrocardiogram, electroencephalogram or just a picture of user's face which then used for facial activity recognition or eye movement tracking. After some preprocessing these data then become source of information about user's interests.

Techniques that involve the user into the interests estimation process. There are studies devoted to personalized video summarization using information about user's preferences defined by him/her manually. In [3] user have been asked to select several keyframes in a given video sequence, then those keyframes are used to perform the automatic temporal segmentation using the analysis of inter-frame similarity to the keyframes. After that the video summarization problem has been reduced to the knapsack problem. In [2] user has to define personal preferences in the application's settings.

Techniques based on automatic user's preferences analysis. This approach supposes automatic estimation of user interests using covert tracking his/her activity [5].

III. Personalized Video Summarization Based on User Preferences

In this section we describe an approach to personalized video summary construction by means of analysis of user's preferences. The approach is based on the following ideas.

User's estimations of videos in video database. Video database or video-on-demand system (e.g. YouTube) allows a user to rate videos providing he/she with simple binary "like/dislike" scale, which can be expanded by the "neutral" estimation if we additionally consider watched but unrated videos. Let us denote the expanded set of a user's estimations as follows:

$$E = \{e^+, e^0, e^-\} \tag{1}$$

Preprocessing of video is an activity devoted to extracting the metadata from video file while uploading it to the database. Preprocessing consists of two stages, namely video structuring and object detection.

Video structuring stage is performed using one of various scene detection techniques [12]. If we denote video as V, its shot as s and number of frames as F, then there is the following relationship between video and scene:

$$\begin{aligned} &V = \{s_j\}_{i=1}^{n}; \\ &0 \le n \le F \end{aligned} \tag{2}$$

In these terms *video summary* $r(V)$ is a subset of video scenes, which represents content of the whole video and its duration $d(r(V))$ does not exceed some predefined value L.

$$\begin{aligned} &r(V) = \{s_j\}_{j=1}^{t}, s_j \subseteq V; \\ &d(r(V)) = \sum_{j=1}^{t} d(s_j); \\ &d(r(V)) \le L \end{aligned} \tag{3}$$

Object detection stage is based upon one of various object detection techniques (e.g. [13], [14], etc.). Let us denote a set of all video objects detected inside the video database as O and the number of these objects as M.

$$O = \{o_i\}_{i=1}^{M} \tag{4}$$

Let us assume that there is a number of videos watched by the user, i.e. each one of these videos has got some estimation from the E set. Paying attention to the fact that we have information about the structure (2) and objects detected on these videos, we are now able to evaluate the *importance* of each object for the user based upon the probability theory.

Let us consider the fact of appearence of the object $o_i \in O$ on the scene of the watched video as an A_i event. Additionally, we denote the number of scenes in the videos that are "liked", "disliked" and "neutrally estimated" by the user as L, D and N respectively. Then the number of scenes that contain o_i object and belong to the videos that are "liked", "disliked" and "neutrally estimated" by the user can be analogically denoted as L_i, D_i and N_i. The total number of all videos that are watched by the user we denote as W.

Due to E is a full set of mutually exclusive events the value of statistical probability that the o_i object has affected the user's evaluation of the video can be calculated as follows:

$$\begin{aligned} &P(A_i) = P(A_i|e^+) \cdot P(e^+) + P(A_i|e^0) \cdot P(e^0) + \\ &+ P(A_i|e^-) \cdot P(e^-) = P(A_i \cap e^+) + \\ &+ P(A_i \cap e^0) + P(A_i \cap e^-) \end{aligned} \tag{5}$$

where

$$P(A_i|e^+) = \frac{L_i}{L}; \quad P(A_i|e^0) = \frac{N_i}{N}; \quad P(A_i|e^-) = \frac{D_i}{D}$$

and

$$P(e^+) = \frac{L}{W}; \quad P(e^0) = \frac{N}{W}; \quad P(e^-) = \frac{D}{W}$$

We denote the importance of the o_i object as $Imp(o_i)$. The value of $Imp(o_i)$ can be estimated using the following equation:

$$Imp(o_i) = P(A_i) \cdot \frac{L_i - D_i}{\max(1, N_i)} \tag{6}$$

All the objects that *significantly* affect user's estimations of videos we call *user's range of interest (ROI)*. We consider the o_i object to have a significant impact if $|Imp(o_i)| \ge minimp$

where $minimp$ is a predetermined threshold and $minimp \in R^+ \cup 0$:

$$ROI = \left\{ o_i \in O \;\middle|\; |Imp(o_i)| \geq minimp \right\} \tag{7}$$

After we have found the user's ROI, we able to evaluate an impact of each scene s_j in any video, which has not yet been watched by the user, using the following equation.

$$\begin{aligned} Imp(s_j) = sgn\Big(\underset{o_i \in s_j}{argmax}\big(|Imp(o_i)|\big)\Big) \cdot \\ \cdot \max_{o_i \in s_j} |Imp(o_i)| \cdot \sum_{o_i \in s_j} Imp(o_i) \end{aligned} \tag{8}$$

Now we can define *personalized video summary* $pr(V)$ as an ordered sequence of scenes, which are detected on the video during the preprocessing and have the greatest impact on user's estimations. Paying attention to the limitation on the duration of a video summary, we consider that the duration of this sequence should not exceed a predefined limit $maxd > 0$.

$$\begin{aligned} &pr(V) = \bigcup_{i=1}^{t} s_i | s_i \subseteq V; \\ &\sum_{i=1}^{t} d(s_i) \leq maxd; \\ &\forall s_i \subseteq pr(V), s_j \not\subseteq pr(V) : |Imp(s_i)| \geq |Imp(s_j)| \end{aligned} \tag{9}$$

In (8) the sgn function gives the sign of the impact (positive or negative) of the most important object on the scene. This basically shows which of the estimates (1) of the full video we should expect and will be used in future work for the evaluation of the quality of the constructed video summary.

IV. Conclusion

In this paper we have described an approach to constructing personalized video summary utilizing user's "like/neutral/dislike" estimations of videos watched beforehand. Summary is built as a sequence of scenes extracted from the video, which are most influencing the user. Most influencing scene contain a set of objects detected on video, which are in range of user's interest. Formal definitions of most influencing scene and range of interest are given and mathematical model of constructing personalized video summary is described.

As future work we plan to implement a prototype of personal video summarization system based upon the described mathematical model.

References

[1] V. Vijayakumar and R. Nedunchezhian, "A study on video data mining," *IJMIR*, vol. 1, no. 3, pp. 153–172, 2012. [Online]. Available: http://dx.doi.org/10.1007/s13735-012-0016-2

[2] G. Ghinea, R. Kannan, S. Swaminathan, and S. Kannaiyan, "A novel user-centered design for personalized video summarization," in *2013 IEEE International Conference on Multimedia and Expo Workshops, Chengdu, China, July 14-18, 2014*. IEEE, 2014, pp. 1–6. [Online]. Available: http://dx.doi.org/10.1109/ICMEW.2014.6890642

[3] B. Han, J. Hamm, and J. Sim, "Personalized video summarization with human in the loop," in *IEEE Workshop on Applications of Computer Vision (WACV 2011), 5-7 January 2011, Kona, HI, USA*. IEEE Computer Society, 2011, pp. 51–57. [Online]. Available: http://dx.doi.org/10.1109/WACV.2011.5711483

[4] A. G. Money and H. W. Agius, "Analysing user physiological responses for affective video summarisation," *Displays*, vol. 30, no. 2, pp. 59–70, 2009. [Online]. Available: http://dx.doi.org/10.1016/j.displa.2008.12.003

[5] A. Yoshitaka and K. Sawada, "Personalized video summarization based on behavior of viewer," in *Eighth International Conference on Signal Image Technology and Internet Based Systems, SITIS 2012, Sorrento, Naples, Italy, November 25-29, 2012*, K. Yétongnon, R. Chbeir, A. Dipanda, and L. Gallo, Eds. IEEE Computer Society, 2012, pp. 661–667. [Online]. Available: http://dx.doi.org/10.1109/SITIS.2012.100

[6] A. G. Money and H. W. Agius, "Video summarisation: A conceptual framework and survey of the state of the art," *J. Visual Communication and Image Representation*, vol. 19, no. 2, pp. 121–143, 2008. [Online]. Available: http://dx.doi.org/10.1016/j.jvcir.2007.04.002

[7] B. T. Truong and S. Venkatesh, "Video abstraction: A systematic review and classification," *TOMCCAP*, vol. 3, no. 1, 2007. [Online]. Available: http://doi.acm.org/10.1145/1198302.1198305

[8] S. E. F. de Avila, A. P. B. Lopes, A. da Luz Jr., and A. de Albuquerque Araújo, "VSUMM: A mechanism designed to produce static video summaries and a novel evaluation method," *Pattern Recognition Letters*, vol. 32, no. 1, pp. 56–68, 2011. [Online]. Available: http://dx.doi.org/10.1016/j.patrec.2010.08.004

[9] S. P. Lloyd, "Least squares quantization in PCM," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–136, 1982.

[10] H. Zhou, A. H. Sadka, M. R. Swash, J. Azizi, and U. A. Sadiq, "Feature extraction and clustering for dynamic video summarisation," *Neurocomputing*, vol. 73, no. 10-12, pp. 1718–1729, 2010. [Online]. Available: http://dx.doi.org/10.1016/j.neucom.2009.09.022

[11] H. Joho, J. Staiano, N. Sebe, and J. M. Jose, "Looking at the viewer: analysing facial activity to detect personal highlights of multimedia contents," *Multimedia Tools Appl.*, vol. 51, no. 2, pp. 505–523, 2011. [Online]. Available: http://dx.doi.org/10.1007/s11042-010-0632-x

[12] M. del Fabro and L. Böszörményi, "State-of-the-art and future challenges in video scene detection: a survey," *Multimedia Syst.*, vol. 19, no. 5, pp. 427–454, 2013. [Online]. Available: http://dx.doi.org/10.1007/s00530-013-0306-4

[13] T. Althoff, H. O. Song, and T. Darrell, "Detection bank: An object detection based video representation for multimedia event recognition," *CoRR*, vol. abs/1405.7102, 2014. [Online]. Available: http://arxiv.org/abs/1405.7102

[14] A. J. C. Campilho and M. S. Kamel, Eds., *Image Analysis and Recognition - 11th International Conference, ICIAR 2014, Vilamoura, Portugal, October 22-24, 2014, Proceedings, Part II*, ser. Lecture Notes in Computer Science, vol. 8815. Springer, 2014. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-11755-3