Personalized Video Summarization Based Exclusively on User Preferences

Costas Panagiotakis[1,2(✉)], Harris Papadakis[3], and Paraskevi Fragopoulou[2,3]

[1] Department of Management Science and Technology, Hellenic Mediterranean University, 72100 Agios Nikolaos, Greece
cpanag@hmu.gr
[2] Institute of Computer Science, FORTH, Heraklion, Greece
fragopou@ics.forth.gr
[3] Department of Electrical and Computer Engineering, Hellenic Mediterranean University, 71004 Heraklion, Greece
adanar@hmu.gr

Abstract. We propose a recommender system to detect personalized video summaries, that make visual content interesting for the subjective criteria of the user. In order to provide accurate video summarization, the video segmentation provided by the users and the features of the video segments' duration are combined using a Synthetic Coordinate based Recommendation system.

Keywords: Recommender system · Video summarization

1 Introduction

Video summarization is an application of recommender systems [9,13] that generally aims at providing users with targeted information about items that might interest them. Recommender systems are also used to provide users with suggestions for various entities such as e-shop items, web pages, news, articles, movies, music, hotels, television shows, books, restaurants, friends, etc.

In this work, we study the problem of personalized video summarization without an priori knowledge of the video categories. According to our knowledge, this is the first work that solves the personalized video summarization based exclusively on user preferences for a given dataset of videos. In order to solve this problem, we propose a video segmentation method that yields global video segments. The main contribution of this work is the proposed video segmentation method and the efficient combination of the video segments' duration attribute with the Synthetic Coordinate based Recommendation system (SCoR) [12] without the use complex audiovisual features.

2 Related Work

The problem of content recommendation can be described as follows. Given a set U of users, a set I of items and a set R of user ratings for items, we

J. M. Jose et al. (Eds.): ECIR 2020, LNCS 12036, pp. 305–311, 2020.
https://doi.org/10.1007/978-3-030-45442-5_38

need to predict ratings for user-item pairs which are not in R. One of the main recommender system techniques is similarity-based Collaborative Filtering [1]. Such algorithms are based on a similarity function which takes into account user preferences and outputs a similarity degree between pairs of users. Another important approach in recommender systems is Dimensionality Reduction. Each user or item in the system is represented by a vector. A user's vector is the set of his ratings for all items in the system (even those that have not been rated by the specific user). The Matrix Factorization method [5] that characterizes both items and users by vectors of latent factors inferred from item rating patterns, is also a Dimensionality Reduction technique. High correlation between item and user factors leads to a recommendation.

In [12], the SCoR recommender system has been proposed that assigns synthetic coordinates to users and items (nodes). SCoR assigns synthetic coordinates (vectors) to users and items as proposed in [2], but instead of using the dot product, SCoR uses the Euclidean distance between a user and an item in the Euclidean space, so that, when the system converges, the distance between a user-item pair provides an accurate prediction of that user's preference for the item. SCoR has been also successfully applied to the distributed community detection problem [11] and to the interactive image segmentation problem [10].

A video summary usually includes the most important scenes and events from a video, with the shortest possible description. Many traditional video summarization approaches, which are not personalized, [8,16] find a global optimal representation of a given video taking into account only its audiovisual features. As the given, video synopsis datasets and annotations increase, the computer vision community realized that the problem of video summarization can be also defined and solved separately for each user taking into account his preferences. Thus, the research on personalized video summarization is gaining increased attention recently [19].

There exist supervised methods based on complex audiovisual features that can become personalized by training on annotations coming from a single user [18]. Other personalized methods use text queries [17]. They suffer, however, from the cold start problem, not being able to provide recommendations for users that are not in the training set. In addition, only a small number of examples per user are often available. This limits the class of possible methods to simple models that can be trained from a handful of examples [6]. More recent methods use a ranking formulation, where the goal is to score interesting video segments higher than non-interesting ones [4,6,14,19] while combining audiovisual representation and user preferences. In [19], a novel pairwise deep ranking model is proposed that employs deep learning in order to learn the relationship between highlighted and non-highlighted video segments. A two-stream network structure is developed by representing video segments from complementary information on the appearance of video frames and temporal dynamics across frames for video highlight detection. Rather than training one model per user, the model proposed in [6] is personalized via its inputs, which allows to effectively adapt its predictions, given only a few user-specific examples. To train this model, a large-

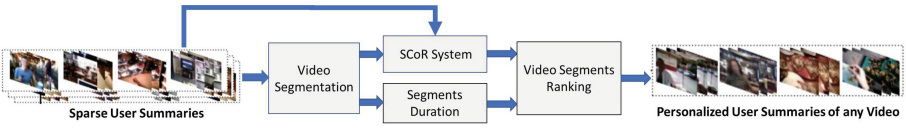


Fig. 1. The schema of the proposed system architecture.

scale dataset of users and GIFs is created, providing an accurate indication of their interests. In this work, we use the same dataset and a ranking formulation.

3 Personalized Video Summarization

In this Section, the proposed personalized video summarization method is described. Figure 1 depicts the two stages of the proposed framework. In the first stage, each video is segmented into non overlapping segments according to the preferences of the users. In the second stage, the personalized rankings of the video segments are provided.

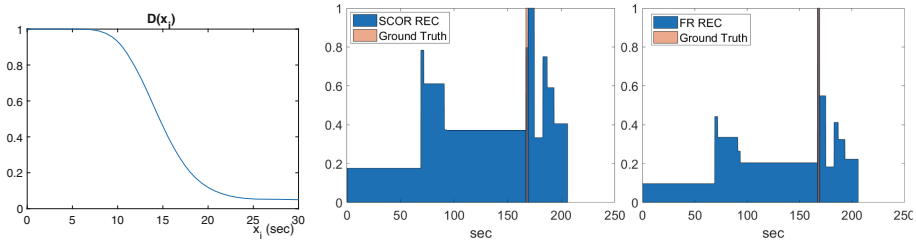
3.1 Video Segmentation

The goal of video segmentation is to provide the candidate video segments that are included in the video summarization, significantly reducing the problem search space from the set of frames to the set of video segments. The simplest video segmentation is to use fixed segments (e.g. of 5 s duration) [6]. Several audiovisual based video summarization methods use shot detection [3] or other more complex temporal segmentation approaches [7,19] to provide accurate (non-overlapping) video segmentation. In this work, since the audiovisual data are not taken into account, we take advantage of the user preferences in the training set to derive the video segmentation.

Let F_v be the union of segment borders (frames) in ascending order, that the users provide in the training set according to the proposed video highlights of video v. As the number of users increases, the frames of F_v correspond to an over-segmentation of the given video v. So, in this work we simplify set F_v, so that there is a minimum duration for each video segment, e.g. at least 1 s. To do so, we repetitively remove the frame f from F_v according to Eq. 1, until the minimum segment length is at least 1 s.

$$f = \underset{i \in \{1,2,\ldots,|F_v(i)|\}}{\arg\min} \; min(\delta_v(i), \delta_v(i+1)) + \frac{1}{|v|} \cdot max(\delta_v(i), \delta_v(i+1)) \qquad (1)$$

where $\delta_v(i) = |F_v(i) - F_v(i-1)|$ corresponds to the duration of video segment $[F_v(i), F_v(i-1)]$ and $|v|$ is the video length. This equation selects the frame that corresponds to the shortest segment. In order to decide which of the two border frames of a segment should be eliminated, we also take into account the size of the longest neighbor segment $(max(\delta_v(i), \delta_v(i+1)))$, so that the frame in between the two shorter in duration video segments is selected.



(a) (b)

Fig. 2. (a) An example of the ranking function $D(x_i)$. (b) An example of $SCOR_u(i)$ (left) and $FR_u(i)$ (right) recommendations on a given video.

3.2 Video Segments Duration

Generally, it holds that the users select short video segments to be included in the proposed video synopsis (e.g. less than 20 s). In this work, we apply a statistical analysis approach with personalized components taking into account the average segment duration of a user (d_u), of a video (d_v), for dataset (d) and the standard deviation of the video segment duration in dataset (σ). So, for a user u and an unseen (for that user) video v, the ranking function $D(x_i)$ (see Fig. 2(a)) is computed, where x_i denotes the duration of segment $[F_v(i), F_v(i-1)]$.

$$D(x_i) = (1-\lambda) \cdot (1 - CDF_{\mu,\sigma}(x_i)) + \lambda \tag{2}$$

where $CDF_{\mu,\sigma}$ is the Cumulative Gamma distribution function with mean value $\mu = \frac{d_u+d_v+d}{3} + 3 \cdot \sigma$ and standard deviation σ. The popular two-parameter Gamma distribution is selected, since it is defined only for positive values, such as the duration attribute. The positive parameter λ (e.g. $\lambda = 0.05$) and the addition of $3 \cdot \sigma$ is used to relax the effect of the duration attribute to the whole ranking process, since it is a complementary feature in the final decision process.

3.3 Ranking Video Segments

In the final stage of the proposed method, the video segments are ranked by combining the segment duration based on the ranked function $D(x_i)$ and the ranking of video segments provided by the SCoR system.

Similarly to [12], in order to train SCoR, we get all video segments (see Sect. 3.1) of each video v that have been summarized by user u. Let $[F_v(i), F_v(i-1)]$ be the video segment i of video v, then the recommendation $R_u(i)$ of user u for this segment, that is used to train the SCoR, is given by the percentage of the video segment frames $[F_v(i), F_v(i-1)]$ that belong to the video summary that user u provides. This means that $R_u(i) \in [0, 1]$.

SCoR [12] assigns synthetic coordinates to users and items (video segments), so that the distance between a user and a video segment provides an accurate prediction of the user preference for that video segment. The lowest ranking value

(recommendation) is assigned a distance of 1, whereas the highest ranking value is assigned a distance of 0. When the system converges, users and video segments have been placed in the same multi-dimensional Euclidean space. Let $p(u)$ and $p(i)$, be the position of user u and video segment i in this space. Then, for a pair of user u and video segment i, SCoR is able to provide a recommendation $SCOR_u(i) = max(0, 1 - ||p(u) - p(i)||_2)$. The final personalized recommendation $FR_u(i) \in [0, 1]$ is given by the product of SCoR and the duration based recommendations:

$$FR_u(i) = \frac{SCOR_u(i) \cdot D(x_i)}{max_j SCOR_u(j) \cdot D(x_j)} \tag{3}$$

The denomination of Eq. 3, normalizes the final recommendation $FR_u(i)$ so that its maximum value is one. Figure 2(b) depicts an example of $SCOR_u(i)$ (left) and $FR_u(i)$ (right) recommendation for a given video.

4 Experimental Results

In our experimental results, we included the proposed method ($SCOR-D$) and two methods from the literature ($PHD-CA+SVM-D$ [6] and $Video2GIF$ [4]) and the following three variants of the proposed method:

- $SCOR$: The variant of the proposed method that only uses the SCoR system.
- $SCOR-FIX$: The variant of the proposed method that combines SCoR with fixed length (5 s, as proposed in [6]) video segmentation.
- $RANDOM$: Random summaries based on the proposed video segmentation.

To obtain personalized video highlight data, we have used the large scale dataset proposed in [6], that contains *13,822 users and 222,015 annotations on 119,938 YouTube videos*. Due to the fact that our method is only based on user preferences, we keep users and videos with at least five annotations in order to be able to provide recommendations (cold start problem). The resulting dataset consists of *1822 users and 6347 annotations on 381 videos* with 129,890 candidate video segments under the proposed video segmentation with variable segment lengths, and 199,462 video segments with fixed, 5 s, segment length. The dataset was randomly separated into training and test sets, as proposed in [6]. In the test set, we included annotations from 191 users concerning their last (191) annotated videos (50% of the given videos).

To evaluate the performance of the video summarization methods, we report the mean Average Precision (mAP) [14] and the Normalized Meaningful Summary Duration ($NMSD$) [6]. $NMSD$ rates how much of the video has to be watched before the majority of the ground truth selection is shown, given that the frames in the video are re-arranged in descending order of their predicted recommendation scores. In addition, we report the F_1 *score* that is computed by comparing the ground truth selection with the video summary of the same length ($recall = precision$) that is created by adding frames in descending order of their predicted recommendation scores. Thus, the F_1 *score* measures the percentage of the video summary that belongs to the ground truth selection.

Table 1. Comparison with the state-of-the-art comparison

Criteria	PHD-CA + SVM-D	Video2GIF	SCOR-D	SCOR	SCOR-FIX	RANDOM
mAP	16.68	15.86	**21.65**	15.71	10.22	9.67
$nMSD$	40.26	42.06	**28.82**	42.48	44.52	55.96
F_1 *score*	–	–	**18.32**	9.51	5.72	4.69

Table 1 presents the average mAP, $nMSD$ and F_1 *score*. It holds that the proposed method $SCOR-D$ clearly outperforms all the remaining methods under any evaluation metric. The importance of the duration attribute and the proposed variable length video segmentation is verified by comparing the results of the proposed method against $SCOR$ and $SCOR-FIX$, respectively. The F_1 *score* of the proposed method is 9% and 13% higher than the F_1 *score* of $SCOR$ and $SCOR-FIX$, respectively. $SCOR$ is the second method in performance, while $SCOR-FIX$ is the third one, under any evaluation metric. Finally, it should be noted that the performances of $PHD-CA+SVM-D$ and $Video2GIF$ have been obtained in the whole dataset of [6], so they are not directly comparable with the other methods.

5 Conclusions

In this work, we presented a methodology to detect personalized video highlights without taking into account audiovisual features. The proposed method efficiently uses known user preferences to derive a video segmentation and it combines the segment duration attribute with the SCoR recommender system [12], yielding accurate personalized video summarization. According to our experimental results, the proposed system outperforms other variants and methods from literature. The proposed methodology can be extended to include rich audiovisual features [15], in order to be able to provide personalized user summaries even for unseen videos.

Acknowledgements. This research has been co-financed by the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH - CREATE - INNOVATE (project code: T1EDK-02147).

References

1. Adomavicius, G., Kwon, Y.: Improving aggregate recommendation diversity using ranking-based techniques. IEEE Trans. Knowl. Data Eng. **24**(5), 896–911 (2012)
2. Gorrell, G.: Generalized Hebbian algorithm for incremental singular value decomposition in natural language processing. In: EACL 2006, 11st Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference, 3–7 April 2006, Trento, Italy (2006)

3. Gygli, M.: Ridiculously fast shot boundary detection with fully convolutional neural networks. In: 2018 International Conference on Content-Based Multimedia Indexing (CBMI), pp. 1–4. IEEE (2018)
4. Gygli, M., Song, Y., Cao, L.: Video2gif: automatic generation of animated gifs from video. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1001–1009 (2016)
5. Koren, Y., Bell, R., Volinsky, C.: Matrix factorization techniques for recommender systems. Computer **42**(8), 30–37 (2009)
6. del Molino, A.G., Gygli, M.: Phd-gifs: personalized highlight detection for automatic gif creation. arXiv preprint arXiv:1804.06604 (2018)
7. Pal, G., Acharjee, S., Rudrapaul, D., Ashour, A.S., Dey, N.: Video segmentation using minimum ratio similarity measurement. Int. J. Image Min. **1**(1), 87–110 (2015)
8. Panagiotakis, C., Doulamis, A., Tziritas, G.: Equivalent key frames selection based on iso-content principles. IEEE Trans. Circuits Syst. Video Technol. **19**(3), 447–451 (2009)
9. Panagiotakis, C., Papadakis, H., Fragopoulou, P.: Detection of hurriedly created abnormal profiles in recommender systems. In: International Conference on Intelligent Systems (2018)
10. Panagiotakis, C., Papadakis, H., Grinias, E., Komodakis, N., Fragopoulou, P., Tziritas, G.: Interactive image segmentation based on synthetic graph coordinates. Pattern Recogn. **46**(11), 2940–2952 (2013)
11. Papadakis, H., Panagiotakis, C., Fragopoulou, P.: Distributed detection of communities in complex networks using synthetic coordinates. J. Stat. Mech: Theory Exp. **2014**(3), P03013 (2014)
12. Papadakis, H., Panagiotakis, C., Fragopoulou, P.: Scor: a synthetic coordinate based system for recommendations. Expert Syst. Appl. **79**, 8–19 (2017)
13. Ricci, F., Rokach, L., Shapira, B.: Recommender systems: introduction and challenges. In: Ricci, F., Rokach, L., Shapira, B. (eds.) Recommender Systems Handbook, pp. 1–34. Springer, Boston (2015). https://doi.org/10.1007/978-1-4899-7637-6_1
14. Sun, M., Farhadi, A., Seitz, S.: Ranking domain-specific highlights by analyzing edited videos. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8689, pp. 787–802. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10590-1_51
15. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4489–4497 (2015)
16. Truong, B.T., Venkatesh, S.: Video abstraction: a systematic review and classification. ACM Trans. Multimedia Comput. Commun. Appl. (TOMM) **3**(1), 3 (2007)
17. Vasudevan, A.B., Gygli, M., Volokitin, A., Van Gool, L.: Query-adaptive video summarization via quality-aware relevance estimation. In: Proceedings of the 2017 ACM on Multimedia Conference, pp. 582–590. ACM (2017)
18. Xu, J., Mukherjee, L., Li, Y., Warner, J., Rehg, J.M., Singh, V.: Gaze-enabled egocentric video summarization via constrained submodular maximization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2235–2244 (2015)
19. Yao, T., Mei, T., Rui, Y.: Highlight detection with pairwise deep ranking for first-person video summarization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 982–990 (2016)