Video Recommendation Using Crowdsourced Time-Sync Comments

Qing Ping
College of Computing & Informatics
Drexel University
Philadelphia, PA, United States
qp27@drexel.edu

ABSTRACT

Most existing work on video recommendation focuses on recommending a video as a whole, largely due to the unavailability of semantic information on video shot-level. Recently a new type of video comments has emerged, called time-sync comments, that are posted by users in real playtime of a video, thus each has a timestamp relative to the video playtime. In the present paper, we propose to utilize time-sync comments for three research tasks that are infeasible or difficult to tackle in the past, namely (1) video clustering based on temporal user emotional/topic trajectory inside a video; (2) video highlight shots recommendation unsupervisedly; (3) personalized video shot recommendation tailored to user moods. We analyze characteristics of time-sync comments, and propose feasible solutions for each research task. For task (1), we propose a deep recurrent auto-encoder framework coupled with dictionary learning to model user emotional/topical trajectories in a video. For task (2), we propose a scoring method based on emotional/topic concentration in time-sync comments for candidate highlight shot ranking. For task (3), we propose a joint deep collaborative filtering network that optimizes ranking loss and classification loss simultaneously. Evaluation methods and preliminary experimental results are also reported. We plan to further refine our models for task (1) and (3) as our next step.

CCS CONCEPTS

• **Information systems** → *Information retrieval*; *Document representation*; *Content analysis and feature selection*;

KEYWORDS

Time-sync comments, video clustering, video highlight detection, personalized video shot recommendation, mood-aware recommendation

ACM Reference Format:
Qing Ping. 2018. Video Recommendation Using Crowdsourced Time-Sync Comments. In *Twelfth ACM Conference on Recommender Systems (RecSys '18), October 2–7, 2018, Vancouver, BC, Canada.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3240323.3240329

RecSys '18, October 2–7, 2018, Vancouver, BC, Canada

ACM ISBN 978-1-4503-5901-6/18/10.
https://doi.org/10.1145/3240323.3240329

1 INTRODUCTION

When recommending a video as a whole, it is often challenging to tackle the cold start problem [18] and the poor explainability of the recommendation list [4], which can be partially due to lack of data, such as explicit user preferences, user activity, and meta/content data of a video. When it comes to recommending a fragment of a video, namely a shot within a video, to user preferences, the task becomes even more challenging, since there is even less data that explicitly or implicitly indicates user preferences on each shot of the video. However, recommendation of video shots, if achieved, can be beneficial both theoretically and practically. For example, if we can model user preferences on each shot of a collection of videos, we can re-think the problem of overall video clustering or grouping, by taking into account temporal similarity of user preference trajectories between two videos. Moreover, if we can model preferences of a group of users on each shot, we can get a set of highlights or summarization of a video, which offers novel information for entire video retrieval.

Toward this goal, we propose to take advantage of a new type of data, namely crowdsourced time-sync video comments, for several tasks related to video shot recommendation. The "bullet comments" video sharing platform originated from Niconico in Japan [1], followed by Acfun [2] and Bilibilibi [3] in China, and has similar variants such as YouTube Live Stream Chats[4] and Twitch Chats [5] in United States (Figure 1). The crowdsourced time-sync comments, or "bullet comments", are a series of comments generated by a group of users for a video on-the-fly while the video is being played. In other words, each comment is associated with a unique user id and a unique playtime of the video, and thus can be mapped to a unique key frame in the video.

Appealing as time-sync comments are for the video shot recommendation tasks, there are also challenges embedded in this type of comments. First, the content of such time-sync comments contains noise. In other words, comments are not necessarily relevant to what is being played in the video. Second, the comments suffer from the temporal lagging issue. This issue refers to the phenomenon that some discussion of a current shot may go well beyond this shot and carry onto the subsequent shots, and eventually die down. If the comments are not aligned to their original shots, then the corresponding user preference modelling will lead to inaccurate recommendations.

[1] http://www.nicovideo.jp/
[2] http://www.acfun.cn/
[3] https://www.bilibili.com/
[4] https://www.youtube.com/channel/UC4R8DWoMoI7CAwX8_LjQHig
[5] https://www.twitch.tv/

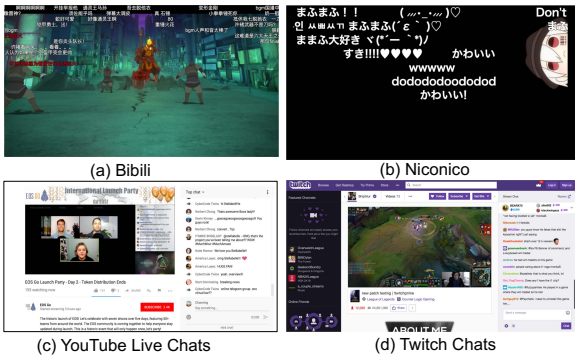

Figure 1: Examples of Video Sharing Platforms with Time-Sync Commenting Support.

Given both the opportunities and challenges, we define three research sub-tasks related to video shot recommendation utilizing crowdsourced time-sync comments, and propose our research plans for each of the sub-task.

Research task-1: Video clustering based on temporal user emotional trajectories inside each video. Instead of taking a video as an unbreakable unit, we consider a video as a long document and use its time-sync comments to induce temporal user emotional trajectories for video clustering. Such clustering will hopefully reflect more intrinsic characteristics of a video, such as genre, plot patterns, emotion tones, and so on. The by-product user emotional trajectories can also be served as natural explanations of such clustering.

Research task-2: Video highlight shots recommendation based on majority audience reactions. Given all time-sync comments generated by a group of users for a series of shots in a video, the task is to recommend a subset of shots that can be seen as "voted" highlights of the video. This recommendation can be categorized as non-personalized recommendation. However, since it is generated by majority reactions of audience, it can be a good starting point for cold-start problem when we have no knowledge of new users but have to provide them video highlight recommendation.

Research task-3: Mood-aware personalized video shots recommendation. Given a set of users, a set of shots and time-sync comments of users and shots, the task is to recommend new video shots to each user that are tailored to their preferences. In addition, this personalized recommendation can be mood-aware, namely we can recommend shots that not only meets user's past preference history, but also can be refined by the "mood" of audience. For example, a user may like a variety of video shots, yet right now, he or she may only want to watch shots that are "hilarious", "touching" or even "breath-taking surprising". Such task may also be beneficial for filming art students and producers, who may need to survey shots of films that induce certain emotional reactions from audience. The rest of the paper is organized as follows: Section 2 introduces related research for the three research tasks. In Section 3, the proposed methodologies are illustrated for each task in detail. In Section 4, we introduce evaluation plans for the three tasks. In Section 5, we report preliminary results for one of the task. Finally, conclusions are presented in Section 6.

2 RELATED WORK

2.1 Video Clustering

Clustering videos into intrinsic semantic groups is a critical task in video analysis [11]. Video clustering by low-level features such as visual and audio features alone is inherently limited due to the semantic gap between low-level visual features and high-level [1, 8]. Researchers have explored combining low-level features with textual information, such as video tags [13, 14, 20], speech transcripts [19] and so on. However, to our best knowledge, there is little if not none research that takes audience's reactional comments frame by frame as semantic data for video clustering.

To achieve this goal, we borrow inspiration from previous research work that models character and character relationship trajectories in book novels [2, 5, 7]. Earlier work propose to model character trajectories in novels by feeding character context sentences into a deep recurrent auto-encoder combined with dictionary learning, so that character trajectories can be derived as a series of topics learnt from the framework [7]. Later work takes into consideration of both character and character relationship context sentences, and also feeds them into a deep recurrent auto-encoder in a contrastive learning framework [2]. The present study plans to adapt this methodology to the task of deriving user emotional trajectories in a video using time-sync comments, and use the derived user emotional trajectories for video clustering.

2.2 Video Highlight Recommendation

First, following the definition in previous work [16], we define highlights as most memorable shots in a video with high emotion intensity. For highlight recommendation, some researchers propose to represent emotions in a video by a curve on the arousal-valence plane with low-level features such as motion, vocal effects, shot length, and audio pitch [6], color [10], mid-level features such as laughing and subtitles [17]. Nevertheless, due to the semantic gap between low-level features and high-level semantics, accuracy of highlight detection based on video processing is limited [8].

Some pioneer work has been done to utilize time-sync comments for video highlight recommendation. One work proposed to represent shots as latent topics by LDA, and used a centroid-diffusion algorithm to detect highlights [15]. Another work proposed to use pre-trained semantic vector of comments to cluster comments into topics, and find highlights based on topic concentration [9]. However, neither of existing work aimed to solve the ill-alignment of time-sync comments due to lagging. Moreover, highlights have to been trained by a classifier on manually labeled data, which may be inflexible and costly. In present study, we plan to bridge these research gaps. First, before highlight recommendation, we will perform lag-calibration to minimize inaccuracy due to comment lags. Second, we propose to represent each shot by the combination of topic and emotion concentration in unsupervised way.

2.3 Personalized Video Shot Recommendation

There is one pioneer work that also utilizes time-sync comments for personalized video shot recommendation [3]. In this work, user time-sync comments are first labelled as positive (including neutral) and negative using Stanford sentiment analysis toolkit. Then

the positive and negative labels are used as ground-truth data for training in a joint deep collaborative filtering module and an auto-encoder LSTM module.

However, this pioneer work might suffer a limitation due to its strong assumption that the "positive" or "negative" sentiments are reflections of user "likes" or "dislikes" of a shot. In reality, this is usually not true. For example, user can express "I'm so touched that I'm going to cry", "I already have tears", "I hate the bad guy!" in time-sync comments, but this does not necessarily mean that user dislikes a shot or the video in general. In present study, we plan to relax this assumption, and propose a deep collaborative filtering framework that are trained to predict a user's reactions to a video shot to be one of eight emotion categories or neutral.

3 RESEARCH METHODOLOGY

3.1 User Emotional Trajectory-based Video Clustering

User emotional trajectory refers to the emotional or topical trajectories extracted from user time-sync comments of a video. For example, the trajectories for a comedy video can be "lovely couple-hilarious incidents-funny conversations-unexpected sudden turn of events-sad for the characters-cheering up for the characters-eventually happy ending". Strictly speaking, the trajectories are not always emotional, and can be simply objectively topical.

To model the trajectory, we borrow inspiration from existing work that models character and character relationship trajectories from novels [2, 5, 7]. Given a series of time-sync comments for a set of videos, we propose to learn the latent topics shared by all videos with dictionary learning, similar to previous work. The dictionary learning is achieved by feeding each time-sync comment into the deep auto-encoder, with a look-up and weighing function of topic vectors as reconstruction of comment representations. The look-up weight vector can then be seen as a topic distribution of the current comment. After training, we will be able to generate a series of such topic distributions for each video, which are the so-called trajectories. These emotional(topical) trajectories will then be used for video clustering.

We also plan to maintain the latent topic embedding in the same space with the word embedding of original comments, so that topics can be "manifested" by looking up its top-N most similar words in the embedding space. This way we can provide semantically explicit explanation for the user emotional/topic trajectories, thus also explanations for our video clustering.

3.2 Video Highlight Recommendation

As discussed earlier, time-sync comments are inherently ill-aligned temporally and contain noise comments. In addition, time-sync comments are usually very short and populated with Internet slangs, which results in the semantic sparsity issue. To overcome these challenges, we propose the following series of steps as solutions.

First, a global word embedding should be trained on a very large corpus that covers sufficient semantic variance on Internet slangs in this type of video sharing platform. Second, in order to re-calibrate ill-aligned comments, we propose to cluster comments inside a video so that comments that are both semantically and temporally close to each other should be grouped together. Here semantical similarity between comments will be determined by word similarity in global embedding space in each comment. Then timestamps of comments can be rewound back to the earliest timestamps in a cluster. The underlying logic is to find the "starting point" of a chain of similar comments that go beyond their original shots, and re-align them to the earliest timestamp. Third, to model the "higlightness" of each video shot, we propose to take into account several factors, including the popularity of the shot (number of comments in this shot), the emotional concentration of a shot (whether expressed user emotions converge or diverge), and the topic concentration of a shot (whether topics in comments converge or diverge). By calculating a "highlightness" score for each shot, we will be able to rank all shots in a video and recommend the top-K as highlight shots.

We also propose a secondary task after we recommend the video highlights, which is to select a few comments from all comments of each highlight shot as a brief summary for this highlight. The objective of the selection will be to cover as complete diverse semantics as possible with limited summary sentences.

3.3 Mood-Aware Personalized Video Shot Recommendation

The objective of this research task is to predict which video shots are mostly likely to excite certain emotions (happy, sad, angry, surprised, fear, disgust, trust) of a user, given history of emotional reactions of this user expressed in sync comments in other video shots. This can break down to two joint tasks, namely predicting whether a user will have emotional reactions to a shot or not, and predicting what emotion the user may have for this shot. The former can be considered as an implicit ranking-based recommendation problem, where a positive example will be a shot with explicit user emotional comment, and a negative example will be a shot with no comments of this user. The latter can be considered as a multi-class classification problem to predict what type of emotion will most likely be the user's reaction.

Since there are no labels of emotions for time-sync comments, similar to previous work[3], we will first semi-automatically label comments with a customized emotion dictionary. We plan to construct the emotion dictionary incrementally, by first manually select a set of seed terms for each emotion, and then adding more terms to each emotion category by comparing similarity between seed words and candidate words in global embedding space.

For the two joint tasks, we plan to build a joint-objective deep collaborative filtering network, that predicts both whether a user and a video shot should be associated by an emotional comment, and what type of emotion will it be. Each user will be represented by abstraction of all previous comments this user has posted. Similarly, each video will be represented by abstractions all its comments excluding this user's (if there is any). We would also like to experiment whether we can exclude all comments to represent a video shot, and only use the collaborative embedding.

4 EVALUATION METHODOLOGY

In this section we introduce evaluation methods for each of the three research tasks.

4.1 Evaluation of Trajectory-Based Video Clustering

For this task, we plan to evaluate our model relying on the user-voted tags for each video. For example, for a detective video, tags can be "suspense", "logic-reasoning", "criminal", "action" and so on. Then if two videos share at least one tag together, we consider they belong to at least one micro-cluster. Our goal is to predict a list of videos that have as many shared tags as possible with the query video.

We will randomly select a set of videos and retrieve top-K most similar videos for each query video calculated by trajectory similarities generated by our model. Then we will evaluate with precision at rank 10 (P@10) and mean average precision (MAP).

4.2 Evaluation of Video Highlight Recommendation

For this task, we plan to evaluate how much "hits" our recommended highlight shots have relative to the ground-truth highlight shots. The "hits" are defined as whether there is a temporal overlap between any recommended shots and ground-truth shots or not. Precision, recall and F-1 are used to evaluate our recommendations based on the "hits".

As for the ground-truth video highlight shots, we plan to construct them from "mix-clip" videos uploaded by users. "Mix-clip" videos are usually self-edited user videos that put together the most favorite shots of a long video, subjective to each user's preferences. We plan to collect a set of such "mix-clips" for each long video, and pick the shots that are "voted" by multiple "mix-clips" as the ground-truth highlight shots.

4.3 Evaluation of Mood-Aware Personalized Video Shot Recommendation

For this task, we plan to evaluate our model with the help of crowd-sourced labeling. We will present each video and accompanying time-sync comments to judges, and ask them to label each comment from eight basic emotions plus neutral. The resulting labels will be used as ground-truth.

To evaluate whether or not a video shot meets a user's mood (so that it is very likely the user will post comment with this mood tune), we will use the Normalized Discounted Cumulative Gain (NDCG) and mean average precision (MAP). To evaluate the accuracy of multi-class classification, we will present the confusion matrix to evaluate how well our model can classify each user-shot pair to a certain emotion category.

5 PRELIMINARY RESULTS

We report our preliminary results on the second research task: video highlight shot recommendation [12]. We compare our proposed method with three baselines: (1) randomly selected shots; (2) uniformly selected shots and (3) popularity-based shots based on number of comments in the shots. We also compare modules of our method: popularity and lag-calibration alone (Popularity+L), popularity and emotional/topical concentration (Popularity +E+T), all together (Popularity+L+E+T). The results are reported in Table 1. From the results we observe that lag-calibration itself is crucial to the performance of the task, and when combined with emotional and topical concentration, the method performs even better.

Table 1: Comparison of Video Highlight Recommendation Methods.

Methods	Precision	Recall	F-1
Random-Selection	0.1578	0.1587	0.1567
Uniform-Selection	0.1775	0.1830	0.1797
Popularity-Selection	0.2594	0.2167	0.2321
Popularity +E+T	0.2796	0.2357	0.2500
Popularity + L	0.3125	0.2690	0.2829
Popularity +L+E+T	0.3099	0.3071	0.3066

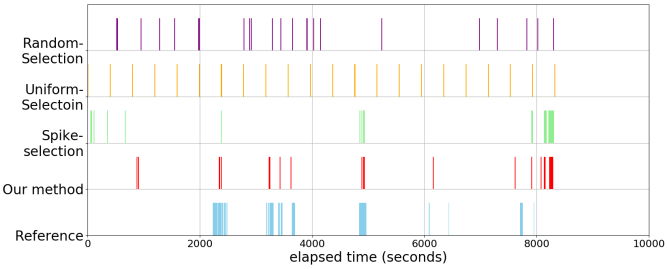

Figure 2: Visualization of Highlights of Ground-Truth and Comparative Methods for a Long Video.

We also visualize the ground-truth highlights of a long video, and the highlights recommended by baselines and our method as in Figure 2. From this visualization, we can see that highlights generated by our method (red bars) have a good overlap with ground-truth (blue bars). The popularity-based method (green bars) have a lot of false positives at the beginning and end of the video, majority of which are greetings and not relevant discussions.

6 CONCLUSIONS

In this paper, we introduce a novel type of data, namely the crowd-sourced time-sync comment, as a promising ingredient for video recommendation tasks. Information in such comments is rich and informative since it is a timely and direct expression of user's reactions to each shot of a video. We propose three research tasks: (1) video clustering with user emotional/topical trajectories, (2) video highlight recommendation and (3) mood-aware personalized video shot recommendation, all taking advantage of the time-sync comments. We analyze the opportunities and challenges inherent in time-sync comments, and propose corresponding feasible solutions for each task based on existing work and preliminary experiment results. Our next-step research plan focuses on the refined modeling and implementation of task (1) and (3). For task (1), we would like to explore other model options besides our proposed models, such as temporal topic modeling and memory network. For task (3), we would like to explore the best way to combine the implicit ranking and multi-class classification in the same framework effectively. Finally, we would like to investigate the explanability of our models for each task utilizing time-sync comments.

ACKNOWLEDGMENTS

This work is supported by the NSF project "A Visual Analytic Observatory of Scientific Knowledge" (NSF 1633286).

REFERENCES

[1] Kevin Beyer, Jonathan Goldstein, Raghu Ramakrishnan, and Uri Shaft. 1999. When is "nearest neighbor" meaningful?. In *International conference on database theory*. Springer, 217–235.

[2] Snigdha Chaturvedi, Mohit Iyyer, and Hal Daumé III. 2017. Unsupervised Learning of Evolving Relationships Between Literary Characters.. In *AAAI*. 3159–3165.

[3] Xu Chen, Yongfeng Zhang, Qingyao Ai, Hongteng Xu, Junchi Yan, and Zheng Qin. 2017. Personalized key frame recommendation. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 315–324.

[4] James Davidson, Benjamin Liebald, Junning Liu, Palash Nandy, Taylor Van Vleet, Ullas Gargi, Sujoy Gupta, Yu He, Mike Lambert, Blake Livingston, et al. 2010. The YouTube video recommendation system. In *Proceedings of the fourth ACM conference on Recommender systems*. ACM, 293–296.

[5] Lea Frermann and György Szarvas. 2017. Inducing semantic micro-clusters from deep multi-view representations of novels. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 1873–1883.

[6] Alan Hanjalic and Li-Qun Xu. 2005. Affective video content representation and modeling. *IEEE transactions on multimedia* 7, 1 (2005), 143–154.

[7] Mohit Iyyer, Anupam Guha, Snigdha Chaturvedi, Jordan Boyd-Graber, and Hal Daumé III. 2016. Feuding families and former friends: Unsupervised learning for dynamic fictional relationships. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 1534–1544.

[8] Keng-Sheng Lin, Ann Lee, Yi-Hsuan Yang, Cheng-Te Lee, and Homer H Chen. 2013. Automatic highlights extraction for drama video using music emotion and human face features. *Neurocomputing* 119 (2013), 111–117.

[9] Guangyi Lv, Tong Xu, Enhong Chen, Qi Liu, and Yi Zheng. 2016. Reading the Videos: Temporal Labeling for Crowdsourced Time-Sync Videos Based on Semantic Embedding.. In *AAAI*. 3000–3006.

[10] Chong-Wah Ngo, Yu-Fei Ma, and Hong-Jiang Zhang. 2005. Video summarization and scene detection by graph modeling. *IEEE Transactions on Circuits and Systems for Video Technology* 15, 2 (2005), 296–305.

[11] Juan Carlos Niebles, Hongcheng Wang, and Li Fei-Fei. 2008. Unsupervised learning of human action categories using spatial-temporal words. *International journal of computer vision* 79, 3 (2008), 299–318.

[12] Qing Ping and Chaomei Chen. 2017. Video Highlights Detection and Summarization with Lag-Calibration based on Concept-Emotion Mapping of Crowdsourced Time-Sync Comments. In *Proceedings of the Workshop on New Frontiers in Summarization*. 1–11.

[13] Arash Vahdat, Guang-Tong Zhou, and Greg Mori. 2014. Discovering video clusters from visual features and noisy tags. In *European Conference on Computer Vision*. Springer, 526–539.

[14] Jingya Wang, Xiatian Zhu, and Shaogang Gong. 2016. Video Semantic Clustering with Sparse and Incomplete Tags.. In *AAAI*. 3618–3624.

[15] Yikun Xian, Jiangfeng Li, Chenxi Zhang, and Zhenyu Liao. 2015. Video highlight shot extraction with time-sync comment. In *Proceedings of the 7th International Workshop on Hot Topics in Planet-scale mObile computing and online Social neTworking*. ACM, 31–36.

[16] Min Xu, Jesse S Jin, Suhuai Luo, and Lingyu Duan. 2008. Hierarchical movie affective content analysis based on arousal and valence features. In *Proceedings of the 16th ACM international conference on Multimedia*. ACM, 677–680.

[17] Min Xu, Suhuai Luo, Jesse S Jin, and Mira Park. 2009. Affective content analysis by mid-level representation in multiple modalities. In *Proceedings of the First International Conference on Internet Multimedia Computing and Service*. ACM, 201–207.

[18] Ming Yan, Jitao Sang, and Changsheng Xu. 2015. Unified youtube video recommendation via cross-network collaboration. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*. ACM, 19–26.

[19] Dong-Qing Zhang, Ching-Yung Lin, Shi-Fu Chang, and John R Smith. 2004. Semantic video clustering across sources using bipartite spectral clustering. In *Multimedia and Expo, 2004. ICME'04. 2004 IEEE International Conference on*, Vol. 1. IEEE, 117–120.

[20] Guang-Tong Zhou, Tian Lan, Arash Vahdat, and Greg Mori. 2013. Latent maximum margin clustering. In *Advances in Neural Information Processing Systems*. 28–36.