Plots to Previews: Towards Automatic Movie Preview Retrieval using Publicly Available Meta-data

Bhagyashree Gaikwad, Ankita Sontakke, Manasi Patwardhan,
Niranjan Pedanekar and Shirish Karande
TCS Research
{bhagyashree.gaikwad,ankita.sontakke,manasi.patwardhan,n.pedanekar,shirish.karande}@tcs.com

Abstract

'Preview', a concept popularized by Netflix, is a contiguous scene of a movie or a TV show highlighting its story, characters, and tone, thus helping viewers to make quick viewing decisions. To create previews, one needs scene-level semantic annotations related to the story, characters, and tone. Soliciting such annotations is an involved exercise and these are expensive to generate automatically. Instead, we aim at creating previews by availing readily available scene meta-data, while avoiding dependency on semantic scene-level annotations. We hypothesize that movie scenes that best match publicly available IMDb plot summaries can make good previews. We use 51 movies from the MovieGraph dataset, and find that a match of the plot summaries with scene dialogues, available through subtitles, is adequate to create usable movie previews, without the need for other semantic annotations. We validate the hypothesis by comparing ratings for scenes selected by the proposed method to those for scenes selected randomly, obtained from regular viewers as well as an expert. We report that even with this 'minimalist' approach, we can select at least one good preview scene for 26 out of 51 movies, as agreed upon by a critical expert judgment. Error analysis of the scenes indicates that features related to the plot structure might be needed to further improve the results.

1. Introduction

With a wide variety of content available on video streaming platforms such as Netflix and Amazon Prime Video, viewers need a browsing mechanism that provides them with enough information to make a decision about whether to watch given video content. Given the short attention span of modern viewers, such a mechanism needs to catch their attention in less than 90 seconds[1]. To cater to this need, Netflix popularized the concept of 'Previews', which provide a "quick highlight of the story, characters, and tone of a title", thus enabling viewers to make a viewing decision in a short time[2].

Previews are not the same as trailers. They are smaller in duration, usually below 90 seconds, designed to assist viewers in quicker decision-making. Generating trailers is a creative process that requires editing skills and expert interventions [31]. Automatic trailer generation thus is a complex problem [14] requiring various aspects to be taken into considerations such as, (i) shot selection based on cinematography [44, 30], scene-level [8, 22], character-centric [28] and affective [12, 6, 31] features, (ii) shot ordering maximizing attractiveness[41], affect [13], aesthetics [30], saliency [6], etc., along with minimum disclosure of the spoilers [30, 31] and (iii) selection of affectively coherent theme music [13]. On the contrary, automatic preview generation is comparatively a simpler scene retrieval problem, and hence, easier to automate. The automation would further facilitate preview generation at scale to cater to the needs of a variety of viewers. In this paper, we suggest a mechanism for automatic preview scene retrieval for movies.

To automate the task of preview scene retrieval, one would need an input that provides highlights about the story, characters, and the tone of the movie. Also, to select a movie scene as a preview, one needs to use the text-audio-visual features of the movie scene which depict its theme. Thus, scene-level features in the form of annotations are needed, or they need to be extracted using audio-visual techniques. Either of these approaches are expensive and time consuming. Instead, the emphasis of this paper is to suggest a mechanism for automatic retrieval of preview scene for movies by using readily available meta-data. In short, we define a 'minimalist' approach for preview scene retrieval which relies only on readily available meta-data for performing the task.

The 'Plot summaries' provided by the IMDb website[2],

[1] https://about.netflix.com/en/news/new-netflix-tv-experience-includes-video-previews-that-speed-your-next-selection

[2] https://help.imdb.com/article/contribution/titles/plots/G56STCKTK7ESG7CP

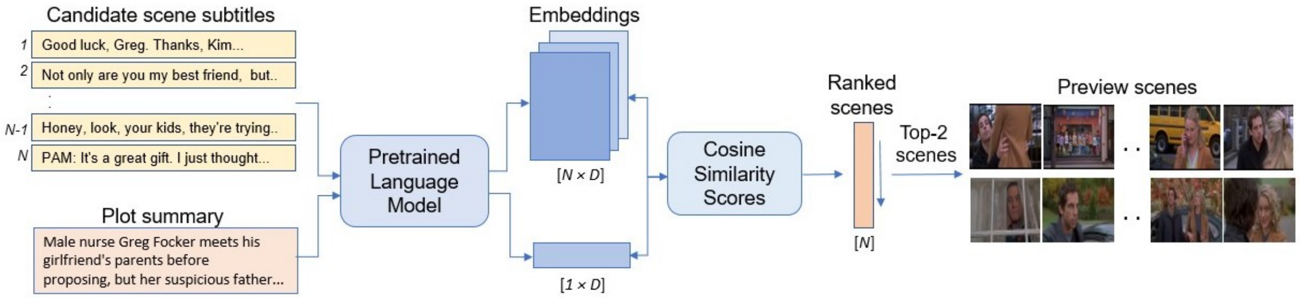

Candidate scene subtitles
1 Good luck, Greg. Thanks, Kim...
2 Not only are you my best friend, but..
N-1 Honey, look, your kids, they're trying..
N PAM: It's a great gift. I just thought...
Plot summary
Male nurse Greg Focker meets his girlfriend's parents before proposing, but her suspicious father...
Pretrained Language Model
Embeddings
[N × D]
[1 × D]
Cosine Similarity Scores
Ranked scenes
[N]
Top-2 scenes
Preview scenes

Figure 1. A schematic diagram of the proposed method to automatically extract preview scenes using candidate scene subtitles and plot summary. N is the number of candidate scenes from a movie and D is the dimension of output embeddings given by a Language model.

briefly describe the story of the movie by avoiding spoilers and talk about the characters. They also often mention the events highlighting some dramatic situation in the main characters' life, which creates curiosity in the mind of the viewers about the movie. For example, in the plot summary of the movie *Meet the Parents* as mentioned in figure 2, the event *Greg Focker meeting his girlfriend's parents and the father being suspicious* is important and the story of the movie revolves around this event.

We hypothesize that the movie scene that best matches the plot summary is a good preview scene. To check if scene dialogues, readily available through subtitles, are adequate to create usable movie previews, we use (i) semantic scene annotations provided by the MovieGraph dataset (including scene subtitles) [38] and (ii) only the scene subtitles as features to match the scene with the plot summary. We perform user studies, where viewers rate the suitability of scenes as previews. We examine how the scenes recommended by our method following both (i) and (ii) above, are rated as compared to randomly selected scenes. We find that the scenes selected using our method are better preview scenes as compared to the randomly selected movie scenes. Also, using scene subtitles as the only scene feature for recommending previews provides comparable performance to using multiple semantic scene-level features. This signifies the potential of using a 'minimalist' approach for preview scene retrieval. We find that this approach allows us to select at least one good preview scene for 26 out of 51 movies, as per a critical expert judgment. Error analysis of the scenes indicates that features related to the plot structure might be needed to further improve the results of this approach.

2. Related Work

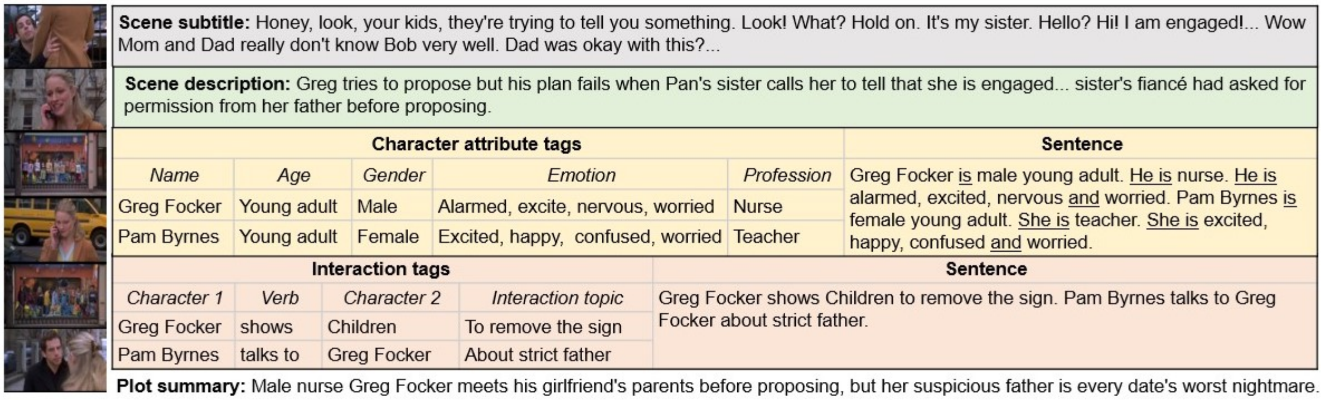
Movies have been summarized considering various aspects, such as main character centrality using script analysis [28], audio-visual saliency captured with perceptual and computational attention modeling [5] and using film domain knowledge, motion analysis and audio features based semantic context detection features [2]. In [15], authors propose a method to generate personalized summaries by measuring similarity between user preferences and high-level features of shots and scenes extracted semi-automatically. The summaries created using these approaches retain the chronological order of shots in the movie as opposed to the trailers. They also include the spoilers as opposed to the trailers or the previews.

Automatic trailer generation is an active application area of research for over a decade [31, 39, 13]. Works addressing shot selection for the purpose treat the problem as anomaly detection [33], maximization of attractiveness and saliency [41], and video highlight detection [39]. [13] focuses on shot reordering based on affective impact and theme music selection. [33] follows the chronological ordering of selected shots. [31] addresses the problem of trailer generation in an end-to-end manner with machine-human collaboration. It emphasizes that trailer generation being a creative process often requires expert intervention. The concept of movie previews is relatively new and differs from traditional trailers. Because we treat the preview creation as a simple scene selection problem, complete automation of this task may be more feasible, as compared to the trailer generation problem.

Browsing for scenes within movies has been performed via task of plot alignment in [23], where a given query is matched with the sentence(s) in the plot that is aligned with the movie scene script. Apart from this, other works on plot alignment emphasize on applications such as story-based content retrieval and semantic video summarization [35, 34, 36], which have different motivation than the preview scene selection task.

3. Dataset

There are several movie-related datasets in the literature that cater to movie understanding and summarization. Some datasets such as Trailer Moment Detection [39], Large-Scale Movie and Trailer [10] are not publicly available. Datasets such as MSA [40], Cognimuse [46] and MovSum [45] either provide a limited number of annota-



Scene subtitle: Honey, look, your kids, they're trying to tell you something. Look! What? Hold on. It's my sister. Hello? Hi! I am engaged!... Wow Mom and Dad really don't know Bob very well. Dad was okay with this?...

Scene description: Greg tries to propose but his plan fails when Pan's sister calls her to tell that she is engaged... sister's fiancé had asked for permission from her father before proposing.

Character attribute tags					**Sentence**
Name	*Age*	*Gender*	*Emotion*	*Profession*	Greg Focker is male young adult. He is nurse. He is alarmed, excited, nervous and worried. Pam Byrnes is female young adult. She is teacher. She is excited, happy, confused and worried.
Greg Focker	Young adult	Male	Alarmed, excite, nervous, worried	Nurse	
Pam Byrnes	Young adult	Female	Excited, happy, confused, worried	Teacher	

Interaction tags				**Sentence**
Character 1	*Verb*	*Character 2*	*Interaction topic*	Greg Focker shows Children to remove the sign. Pam Byrnes talks to Greg Focker about strict father.
Greg Focker	shows	Children	To remove the sign	
Pam Byrnes	talks to	Greg Focker	About strict father	

Plot summary: Male nurse Greg Focker meets his girlfriend's parents before proposing, but her suspicious father is every date's worst nightmare.

Figure 2. Example of MovieGraph annotations for a scene and Plot summary of the movie *Meet the Parents*. Character attributes and interaction tags are converted into meaningful natural language sentences by adding underlined supporting words.

tions per movie, or have taken less number of movies into consideration. There are datasets that provide annotations specific to a task such as video [17] and movie question answering [37], generating audio descriptions [27], scene understanding [25], shot cinematography analysis [44], movie inference [20], movie retrieval [18, 1], etc.

The MovieNet dataset [9] contains a large number of multi-modal annotations for 1,000 movies including cinematographic style, character bounding box and id, scene boundary, action tag, place tag, and plot synopsis manually aligned to movie segments. However, not all these annotations are available for all the movies in the dataset. The annotations about cinematographic style such as shot type and camera angle are not relevant to our task. MovieGraphs dataset [38] contains richer annotations for 51 English language movies along with the scene boundaries for the movies. Each movie has an average of 150 scenes with 43 seconds of average scene duration. As compared to the MovieNet dataset, we find the scene-level annotations provided in the MovieGraph such as scene subtitles, scene description, character attributes, and interactions, to be more relevant for our task. As a result, we use the MovieGraph dataset for our task. We have examined the following scene-level annotations in the dataset for their relevance to our task: (i) Scene subtitles, (ii) Scene description which is a natural language summary of the scene spanning multiple sentences, (iii) Character attribute tags such as the age, gender, profession and emotional states of a character in a particular scene, (iv) Character interaction tags per scene along with its topic and the reason, (v) Character relationship tags, (vi) Scene situation which depicts high-level topic of the scene, (vii) Place tag depicting the scene location, and (viii) Summary of the character interaction. An example of scene annotations provided by MovieGraph for scene subtitles, scene description, character attributes tags and interaction tags are illustrated in figure 2.

4. Approach

In this section we elaborate on our hypothesis and the methodology.

4.1. Hypothesis

Our observations of the existing preview scenes on the Netflix website and the corresponding plot summaries as indicated in section 1, substantiate our intuition that the IMDb plot summaries can be a good source of information to select preview scenes from movie videos. Based on this intuition, we define our hypothesis as 'a movie scene that best matches the IMDb plot summary is a good preview scene'. To the best of our knowledge, as there are no available baselines for automatic preview scene retrieval task in the literature, we compare our method with a random baseline. More formally, we state our **Null Hypothesis** as: Out of the candidate movie scenes, any randomly sampled scene is a good preview scene and the **Alternative Hypothesis** as: Out of the candidate movie scenes, a movie scene that is ranked higher (Top-2) based on its similarity with the IMDb plot summary is a good preview scene.

As mentioned in the alternative hypothesis, we rank the candidate movie scenes using a scoring function that computes the score of the j^{th} candidate scene of the i^{th} movie s_{ij}, with the plot summary of the i^{th} movie p_i. The scoring function is defined using equation 1, where 'sim' is the similarity function and F is a (set-of) feature(s) of the scene. 'E' is an embedding technique used to convert the feature and plot summary into a vectorized representation, which can be further used for similarity computation.

$$Score(s_{ij}) = sim\{E[F(s_{ij})], E[p_i]\} \tag{1}$$

There can be a distinct set of scene features and embedding mechanisms that can be used for the above computation. We conduct pilot studies, described in section 5.1, to find if publicly available scene subtitles are sufficient as the only feature to score the candidate preview scenes. We perform user studies described in section 4.3, to evaluate the scenes sampled from two populations described by the null and alternative hypotheses, and perform hypothesis testing to check if we can reject the null hypothesis.

4.2. Candidate Scenes Selection

Instead of considering all possible scenes in a movie for preview selection, we identify a subset of scenes by filtering some candidate scenes based on our observations about the Netflix preview scenes and insights from movie literature. Screenplays follow a basic linear structure that can be divided into three acts, of which the first act typically establishes the story, situation, characters, and their relationship [7]. We consider the first 1/3$^{\text{rd}}$ part of a movie to be the first act. The first act also contains dramatic actions that play a key role for the audience to determine whether they like the movie [7]. Hence the scenes from the first act of the movie are likely to be more engaging and informative and stand a better chance to be considered as preview scenes. Based on this intuition, we take the scenes from the first 1/3$^{\text{rd}}$ part of the movie as candidate preview scenes. We further observe that the scenes of duration less than 30 seconds are less informative. Motivated by this observation and the attention span of a viewer mentioned in section 1, we further filter the scenes having the duration ranging from 30 to 90 seconds as candidate preview scenes. With this mechanism, out of 150 average scenes per movie, we consider only an average 24 scenes per movie as candidates for the null and alternative hypotheses.

As a pre-processing step to our method (alternative hypothesis), we apply an additional filtering criterion for these candidate scenes. As per the definition of preview scenes, the scenes are required to feature the main characters. We define the main characters as the ones who occur most frequently across all the scenes in the movie. We use the following procedure to identify the main characters: (i) use the character annotations per scene from MovieGraph dataset, (ii) score a character based on the number of scenes the character has appeared in, (iii) use this scoring mechanism to rank the characters, (iv) consider top-3 characters as the main characters since typically a screenplay has 3 main characters[3]. With this information we further filter out only the scenes having the main characters to feed as an input to our method. Thus, after filtering out of 24 average candidates scenes per movie, we consider on an average 23 scenes per movie as an input to our method. Using character annotations fits in our 'minimalist' approach as the automated techniques to identify characters [42, 29] are not very involved and also provide close to 98% accuracy.

4.3. User Studies

As we do not have a dataset of annotated movie preview scenes, we perform user studies to validate our hypothesis. We evaluate top-2 preview scenes selected by our method (alternative hypothesis) and 2 random scenes (null hypothesis) for all 51 movies in the dataset. Effectively, 102 scenes are sampled for each hypothesis. Prior to allocating these scenes to the raters, we shuffle the scenes to avoid comparative bias, which may get introduced if the scenes from the same movie appear one after another. Based on the definition of the preview scenes[4], we provide the raters with the following set of statements to rate each scene on: (i) **Story**: This scene gives the viewer a hint about the story of the movie; (ii) **Character**: This scene gives the viewer a hint about the characters in the movie; (iii) **Feel**: This scene gives the viewer a hint about the feel of the movie and (iv) **Decision**: This scene helps the viewer in deciding whether to watch the movie.

We provide each rater a 20-minute tutorial to explain each statement along with a guidelines manual. Here the 'feel' of a movie talks about its tone being funny, romantic, scary, thriller, dramatic, etc. For each statement, the raters are asked to give responses based on the scene's compliance with the statement on a 5-point Likert scale [19] ranging from 'Strongly disagree' (1) to 'Strongly agree' (5). We also get an additional response for the question, 'Have you watched the movie this scene is from?' to analyze the possible bias that may get introduced when a rater has already watched the movie.

We test our hypothesis by using three methods: (i) Mann–Whitney U test [21], to check if the median ratings for the scenes retrieved by our method are significantly greater than the median ratings for the random scenes. Mann–Whitney U test is an appropriate hypothesis test for us as it works for the ordinal data like the Likert scale variables [32]. (ii) Visualizing the histograms with the frequency of ratings plotted against each unit of the Likert scale. More agreements on scenes selected by our method as compared to the random scenes serve as a validation of our method. (iii) We also use mean ratings to compare the two methods. A method with a higher mean rating of scenes proves to be better suited for the preview selection task.

5. Experiments

Prior to giving out the scenes for external user evaluation at a larger scale, we conduct internal pilot studies to identify if readily available scene feature like scene subtitles can give us comparable performance with the method using more semantic scene features. We then conduct the main experiment by using scene subtitles as the scene features to

[3] https://freshmenscreenplay.com/how-many-characters-can-my-screenplay-have/

[4] https://about.netflix.com/en/news/new-netflix-tv-experience-includes-video-previews-that-speed-your-next-selection

validate our hypothesis, where the scenes are evaluated by regular viewers and also by an expert.

5.1. Pilot Experiments

For the pilot experiments, we use two sets of feature combinations as described below. In each experiment, scenes selected by our method and the random method are evaluated with the user study described in section 4.3. For the pilot experiments, two of the authors evaluate 204 scenes in total with 102 scenes coming from each hypothesis, to solicit one response per scene.

5.1.1 Plot Summary with Multiple Features (PS-MF)

We start with all the scene features provided by the MovieGraph dataset as described in section 3. We use Sentence-BERT [26] pre-trained on meaningful natural language sentences, to compute the similarity score of scene features with the plot summary as depicted by equation 1. Some of the MovieGraph features described in section 3 such as character attributes and interaction tags are not in the form of sentences. Hence as a pre-processing step, we form meaningful sentences out of these features. For example, the interaction sentence *Pam Byrnes talks to Greg Focker about strict father* is formed by integrating the interaction *talks to*, the topic *about strict father* and characters *Pam Byrnes* and *Greg Focker*. We also integrate character attribute tags by using some supporting words such as 'He/She is', 'and', to form sentences, as shown in figure 2. We follow a similar procedure detailed in supplementary material (appendix B) to convert character relationship, summary interaction, situation, and place tags to natural language sentences. We also compile the subtitles provided for a scene in the dataset into a single paragraph by removing the timestamps present in the raw file.

By observation, we find some of the annotations to be redundant for our task. The sentences formed with character relationship and summary interaction tags consist of two character names joined with a relation and interaction verb respectively. For example, summary interaction statement, *Greg Focker asks Pam Byrnes.* is formed by using two character names (*Greg Focker, Pam Byrnes*) joined by summary interaction verb (*ask*). Similarly for the relationship statement, two character names are joined by a relation tag. Whereas, character information is already covered by the character attribute sentence. The summary interaction is also covered by the interaction statement. Information about a situation or a place is rarely mentioned in the plot summary (3 out of 51 movies). These observations allow us to filter out redundant features such as character relation and summary interaction, and only consider the remaining four scene features for this experiment, viz., scene description, interaction statement, character attribute statement, and the scene subtitles, which cover most of the aspects of preview definition.

We compute the cosine similarity scores between embeddings of plot summary and each of the four scene features using equation 1, resulting in 4 distinct similarity scores. We get embedding vectors of all the scene features except scene subtitles, by using pre-trained Sentence-BERT [26]. Since subtitles are longer than other features often forming a paragraph, we use paraphrase-mpnet-base-v2 [5] tuned to paragraphs to get the embedding vectors of subtitles. To find out the cosine similarity score, we use the same embedding mechanism for the plot summary, as the one used for the corresponding scene feature. Based on the similarity scores, we form four different ranked lists of movie scenes. The ranks are then aggregated with MC4 rank aggregation algorithm [4]. We pick the top 2 ranked scenes as the preview scenes for a movie selected by our method (alternative hypothesis).

5.1.2 Plot Summary with Scene Subtitles (PS-SS)

For practical applications, the above approach would be limited only to movies with annotations available in the dataset. To completely automate the preview retrieval task, we need to use annotations that are readily available. Though as a part of the pre-processing step, we are using scene boundary and character detection annotations from MovieGraph, these annotations are usually available as meta-data or can be easily solicited using existing methods available in the literature. For example, [25, 3], are the state-of-the-art scene boundary detection methods, while [29] is the state-of-the-art person re-identification method with over 98% rank-1 accuracy, which can be used for character identification. Movie subtitles can be easily obtained online for free, independent of any dataset annotations. Whereas, the other three semantic annotations considered in PS-MF, viz. scene description, character interaction, and character attributes, require human inputs or more involved automated mechanisms. Also, the performance of methods for automatic extraction of these features is not up to the mark as per the current literature [43, 24]. Hence, we conduct our second pilot experiment (PS-SS) by using scene subtitles as the only scene feature for preview scene selection. We apply the same pre-processing and embedding extraction mechanism as described in section 5.1.1 for scene subtitles and plot summaries. We then use the cosine similarity score and rank the scenes, as per the equation 1. Top-2 scenes with the highest similarity score are considered as preview scenes. The method is illustrated in figure 1.

The results of Mann–Whitney U test for both the pilots demonstrate that the median evaluation scores for PS-MF

[5] https://www.sbert.net/docs/pretrained_models.html

and PS-SS are significantly greater ($p < 0.01$) than that of the random method for all the four evaluation criteria viz. story, characters, feel, and decision making (table 1). PS-MF and PS-SS have more agreements (Agree, Strongly agree) and fewer disagreements (Strongly disagree, Disagree, Neither agree nor disagree) as compared to the random method for all four evaluation criteria, viz. story, character, feel, and decision. We provide the histograms of pilot experiments in the supplementary material (appendix A).

Pilot	Story	Characters	Feel	Decision
PS-MF	3.25^{*}	3.68^{*}	3.39^{*}	3.19^{*}
PS-SS	3.67^{**}	4.09^{**}	3.67^{**}	3.51^{**}
PS-SS - Ex	3.20^{**}	$3.55'$	3.62^{*}	2.95^{**}
PS-SS - RV	3.18^{**}	NA	3.74^{**}	3.50^{**}

Table 1. Mean evaluation scores against the four statements (higher is better) for pilot experiments, PS-MF and PS-SS, and for the main experiment PS-SS evaluated by the expert (- Ex) and by regular viewers (- RV). Median scores for preview scenes by each of our method are significantly greater than scores for random scenes (**$p < 0.001$, *$p < 0.01$, $'p < 0.05$)

To analyze if the PS-SS method achieves comparable performance with PS-MF, we compare the mean evaluation scores, calculated for 102 scenes sampled for the alternative hypothesis for both the pilots. It is observed that, PS-SS performs comparably with PS-MF in terms of the mean scores for all four statements (table 1). With this, we infer that scene subtitles as the only scene feature suffice for this task.

5.2. Main Experiment

As explained in the prior section, in our main experiments we use the PS-SS method for preview scene selection. We perform following user studies as a part of our main experiment: (i) An expert as a rater: As we do not have ground truth annotations of the preview scenes, this user study helps us to solicit the same. Here the expert rater is a screenplay and script writer. (ii) Regular viewers as raters: In real-world applications, previews are created for regular viewers to help them in decision-making. Thus, it is important to evaluate our hypothesis with the ratings provided by regular viewers.

In the PS-SS pilot experiment, we observe that the character statement, ‘This scene gives the viewer a hint about the characters in the movie.’ has more than 82% of ‘Agree’ and ‘Strongly agree’ ratings for all scenes coming from both the hypotheses. This can happen because there is a high chance of every movie scene containing at least one main character. This observation points us to the possibility of the character statement not providing any discriminatory signal for the task at hand. Based on this observation, as well as to reduce the cognitive load of the evaluation process, we decide to drop the character statement for the user study involving regular viewers as the raters. However, we keep this statement in the expert user study to validate the results of the pilot in order to get more insight into it.

With regular viewers as raters, we solicit 3 responses per scene following the method elaborated in section 4.3. We invite 10 participants, 7 males and 3 females with age range 22 to 30 years as regular viewers. Each viewer evaluates around 60 scenes uniformly sampled from the PS-SS method and the random method. We pay the participants approximately 10 cents per scene to ensure engagement. The task took around 27 man-hours. We solicit only 1 response per scene from the expert, but we also solicit plain-text comments to justify the rating he has provided. The results of the main experiments are presented in the next section.

6. Results and Discussion

The Mann–Whitney U test conducted on the ratings by regular viewers and the expert, re-validates our pilot observations. The median evaluation scores for scenes by our method (PS-SS) are significantly greater than scores for the random method, both as per the expert and regular viewer’s ratings (table 1). For frequency computation of the histograms for the experiment involving regular viewers, following [11, 31], we consider all the 3 ratings provided for a scene as distinct inputs (figure 3). Consistent with our pilot experiments, it is observed that PS-SS has more agreements as compared to the random method as per the evaluations of both regular viewers (figure 3) and the expert rater (figure 4), with respect to all the statements. We further analyze the results to evaluate the pros and cons of our method.

6.1. Result Analysis

As per the definition of the preview scenes, to validate if the story, characters, and feel of a scene serves as a hint to the viewer for deciding whether to watch a movie, we find out correlation of the expert ratings between each of the Story, Characters, and Feel statements and the ratings of the Decision statement, using Spearman’s correlation coefficient. For this analysis, we consider all scenes evaluated by the expert for both hypotheses. The correlation between Story-Decision is high (0.72), whereas that between Feel-Decision and Character-Decision are moderate (0.67 and 0.62 respectively). This demonstrates that the story element of the scene may have slightly more impact on decision-making as compared to the characters or feel of the scene. As the ratings provided for the Decision statement are more crucial to the task, and the ratings of the other statements about Story, Characters and Feel are correlated with the Decision statement, we focus on ratings provided to the Decision statement for further analysis.

To analyze the effect of the watched movies on preview scene decision making by the regular viewers and the ex-

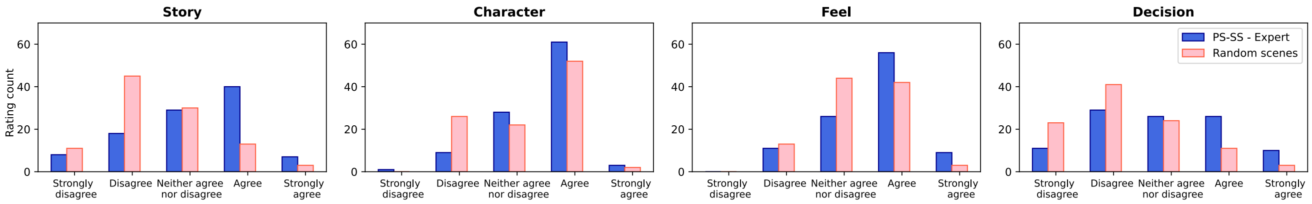


Figure 3. Histograms for Main experiment, Plot Summary with Scene Subtitles (PS-SS) rated by the expert

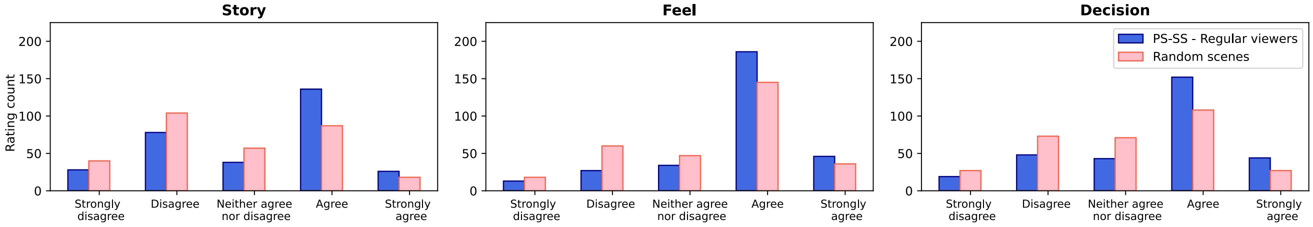


Figure 4. Histograms for Main experiment, Plot Summary with Scene Subtitles (PS-SS) rated by regular viewers

pert, we form two populations of all the scenes based on the answers provided by the raters to the question ‘Have you watched the movie this scene is from?’ as ‘yes’ (P1) or ‘no’ (P2). We observe that the median rating by the regular viewers for the scenes belonging to the movies which are watched is significantly higher than the median rating provided for the scenes of the movies which are not watched ($p < 0.01$). Whereas for the expert ratings, there is no significant difference between medians of the two populations ($p > 0.01$). This shows that whether the regular viewers have watched the movie or not, affects their ratings. They tend to agree more on the scenes they have watched to be preview scenes. On the other hand, no such bias seems to affect the expert ratings.

Within each population P1 and P2, defined above, we further form sub-populations based on the random scenes and the scenes selected by PS-SS. We observe that for the scenes belonging to the movies a regular viewer has not watched, there is a significant difference in the medians of the ratings provided for the sub-populations ($p < 0.01$), whereas for the scenes which belong to the movies a regular viewer has watched, there is no significant difference between the populations ($p > 0.01$). (The relevant histograms are provided in the supplementary material, appendix A) This shows that the regular viewers can discriminate the scenes better as preview or non-preview, for the movies they have not watched. This may be because they are not biased by the prior understanding of the movie. This observation also indicates that our method (PS-SS) can be effective in enabling decisions about movies the viewers have not watched, which is the main purpose of preview scenes.

To get the idea of overall agreements, we analyze the scenes that are rated as ‘Agree’ or ‘Strongly Agree’ for the decision statement. Out of the scenes sampled using the PS-SS method, 64.1% and 35.3% of scenes are agreed upon by regular viewers and the expert, respectively. Whereas, out of the scenes sampled from the random method, 44.1% and 13.7% of scenes are agreed upon by regular viewers and the expert, respectively. Though the overall agreement is higher for the PS-SS method in comparison with the random method by both the expert and the regular viewers, it is still not high in absolute terms, particularly for the expert rater. The lower % agreements for the PS-SS especially by expert and the relatively higher % agreements for the random method by the regular viewers indicate that there is still scope for improvement in the PS-SS method. Also, the inter-rater reliability for 3 ratings provided by distinct raters is low (Krippendorff’s alpha coefficient [16] $=$ 0.27 considering interval scale). This can be because of the inherent differences in the likes and dislikes of the raters for the subjective task at hand, leading to variations in the ratings. However, the low inter-rater reliability further encourages us to analyze the limitations of our work.

6.2. Error Analysis

We observe that the mean expert rating for both the null and alternate hypothesis for the decision statement is low (2.63) as compared to the mean rating of the regular viewers (3.31). This demonstrates that the expert is more critical and conservative in terms of agreeing on a scene being a preview scene. This allows us to use the preview scenes agreed by the expert as the ground truth preview scenes. We find for 26 out of 51 movies, the expert agreed upon one or more preview scenes selected by our method. For 17 movies, the expert, being critical, disagrees with all the (4) scenes sampled from both hypotheses (2 from each). For the remaining 8 movies, we find that the expert has disagreed on any one of the scenes suggested by our method, but has agreed on at least one of the scenes from the random population. We identify 11 such scenes from the random scene population and analyze them since these are the good preview scenes missed by our approach. Our analysis indicates that the errors can be listed in the following categories:

1. **Presence of multiple and varied plot events in the plot summary** For example, The IMDb Plot summary of the movie *Forrest Gump* is *The presidencies of Kennedy and Johnson, the Vietnam War, the Watergate scandal and other historical events unfold from the perspective of an Alabama man with an IQ of 75, whose only desire is to be reunited with his childhood sweetheart*, which mentions more than one event that can not be captured in a single scene or corresponding dialogues.

2. **Scenes creating affect and curiosity** For example, a scene from the movie *The Sixth Sense* where a boy shoots a person, is rated high because of the affect, tension, and curiosity created by the audio-visual features which are not captured by our method.

3. **Scene missed due to absence of main characters** One scene from movie *Harry Potter and the Sorcerer's Stone* showing a wizard performing magic found interesting by the expert is filtered out by our method as there is no main character in this scene.

4. **Subtitles not carrying important information visually shown in the scene** For example, a highly rated scene from the movie *Milk* about gay rights, depicts the relevance to the theme by showing two men kissing each other being warned by a third person. But the subtitles do not carry the information seen in the visuals, and our method does not select it.

5. **Better verbal match with the subtitles of other scenes** For example, the movie *Ocean's Eleven* is about robbing Las Vegas casinos. An expert-agreed scene talks about the high-level security at the casino setting up the curiosity of the viewer. Whereas, subtitles of the top 2 scenes selected by our method talk about the robbery at the Las Vegas Casinos leading to a better match to the plot summary than this scene.

6. **Erroneous detection of subtitles in the dataset** For example, For *Horrible Bosses* and *Bad Santa*, actual dialogues do depict conflict making the scenes interesting. However, these scenes are not selected by our method due to erroneous detection of subtitles in the dataset itself.

In short, since our approach heavily relies on readily available scene meta-data, the scenes having more match between the subtitles and the plot summary are ranked higher, and we tend to miss the scenes creating the feel of the movie with affect and tension using audio-visuals signals. Our dependency on plot summary may lead to errors, where it provides generic information about the movie rather than highlighting important events.

7. Conclusion and Future Work

In this work, we focus on a recent and novel application of creative video scene retrieval, that is automatic preview scene selection, to help viewers for efficient browsing of movies. We explore the possibility of preview scene retrieval by exclusively using readily available meta-data information such as scene subtitles and IMDb plot summary. Being simple, makes this approach scalable to large scale video streaming platforms.

We hypothesize that 'a movie scene that best matches the IMDb plot summary is a good preview scene'. We define a match by using a similarity-based scoring mechanism with pre-trained language model embeddings of the inputs. We validate our hypothesis by performing user studies with both regular viewers and an expert rater. As per their ratings, our method performs superior as compared to a random baseline on movie story, character, feel, and decision based evaluation criteria. With the 'minimalist' approach we have taken, we achieve an encouraging performance in terms of a critical expert selecting at least one of the scenes by our method as a preview for 26 out of 51 movies. Though the analysis of the scenes our approach missed (only for 8 out of 51 movies) indicates that it can benefit from audio-visual and high-level semantic information, we intend to provide usable preview scenes by employing readily available meta-data. This means we aim at high-precision scene preview recommendation rather than recall of all scenes which can be considered good previews.

Though our method performs better than randomly generated previews, we would like to use the ground truth preview scenes to check the efficacy of our method and further enhance it by addressing the limitations. We plan to solicit such ground truth preview annotations either by taking consensus among multiple experts or by considering the movie or TV show scenes by 'Netflix' as ground truth preview scenes. As per the analysis performed on the results, it would be possible to enhance the performance of this method by taking into consideration an input that talks about one or a few exciting events in the movie hinting to the story. We also aim at exploiting the 'logline'[6] structure of the plot summary, which includes features such as protagonist, inciting incident, goal, and conflict. In future, we plan to see if we can identify such events and use them to select preview scenes.

[6] https://www.masterclass.com/articles/screenwriting-tips-how-to-write-a-logline

References

[1] Max Bain, Arsha Nagrani, Andrew Brown, and Andrew Zisserman. Condensed movies: Story based retrieval with contextual embeddings. In *Proceedings of the Asian Conference on Computer Vision*, 2020.

[2] Hsuan-Wei Chen, Jin-Hau Kuo, Wei-Ta Chu, and Ja-Ling Wu. Action movies segmentation and summarization based on tempo analysis. In *Proceedings of the 6th ACM SIGMM international workshop on Multimedia information retrieval*, pages 251–258, 2004.

[3] Shixing Chen, Xiaohan Nie, David Fan, Dongqing Zhang, Vimal Bhat, and Raffay Hamid. Shot contrastive self-supervised learning for scene boundary detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9796–9805, 2021.

[4] Cynthia Dwork, Ravi Kumar, Moni Naor, and Dandapani Sivakumar. Rank aggregation methods for the web. In *Proceedings of the 10th international conference on World Wide Web*, pages 613–622, 2001.

[5] Georgios Evangelopoulos, Konstantinos Rapantzikos, Alexandros Potamianos, Petros Maragos, A Zlatintsi, and Yannis Avrithis. Movie summarization based on audiovisual saliency detection. In *2008 15th IEEE International Conference on Image Processing*, pages 2528–2531. IEEE, 2008.

[6] Georgios Evangelopoulos, Athanasia Zlatintsi, Alexandros Potamianos, Petros Maragos, Konstantinos Rapantzikos, Georgios Skoumas, and Yannis Avrithis. Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention. *IEEE Transactions on Multimedia*, 15(7):1553–1568, 2013.

[7] Syd Field. *Screenplay: The foundations of screenwriting*. Delta, 2005.

[8] Huaiyi Huang, Yuqi Zhang, Qingqiu Huang, Zhengkui Guo, Ziwei Liu, and Dahua Lin. Placepedia: Comprehensive place understanding with multi-faceted annotations. In *European Conference on Computer Vision*, pages 85–103. Springer, 2020.

[9] Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. Movienet: A holistic dataset for movie understanding. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 709–727, Cham, 2020. Springer International Publishing.

[10] Qingqiu Huang, Yuanjun Xiong, Yu Xiong, Yuqi Zhang, and Dahua Lin. From trailers to storylines: An efficient way to learn from movies. *arXiv preprint arXiv:1806.05341*, 2018.

[11] Bogdan Ionescu, Patrick Lambert, Didier Coquin, Laurent Ott, and Vasile Buzuloiu. Animation movies trailer computation. In *Proceedings of the 14th ACM international conference on Multimedia*, pages 631–634, 2006.

[12] Go Irie, Kota Hidaka, Takashi Satou, Akira Kojima, Toshihiko Yamasaki, and Kiyoharu Aizawa. Latent topic driving model for movie affective scene classification. In *Proceedings of the 17th ACM international conference on Multimedia*, pages 565–568, 2009.

[13] Go Irie, Takashi Satou, Akira Kojima, Toshihiko Yamasaki, and Kiyoharu Aizawa. Automatic trailer generation. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 839–842, 2010.

[14] Vivekraj V. K., Debashis Sen, and Balasubramanian Raman. Video skimming: Taxonomy and comprehensive survey. *ACM Comput. Surv.*, 52(5), Sept. 2019.

[15] Rajkumar Kannan, Gheorghita Ghinea, and Sridhar Swaminathan. What do you wish to see? a summarization system for movies based on user preferences. *Information Processing & Management*, 51(3):286–305, 2015.

[16] Klaus Krippendorff. Computing krippendorff's alpha-reliability. 2011.

[17] Jie Lei, Licheng Yu, Tamara Berg, and Mohit Bansal. TVQA+: Spatio-temporal grounding for video question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8211–8225, Online, July 2020. Association for Computational Linguistics.

[18] Jie Lei, Licheng Yu, Tamara L. Berg, and Mohit Bansal. Tvr: A large-scale dataset for video-subtitle moment retrieval. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 447–463, Cham, 2020. Springer International Publishing.

[19] Rensis Likert. A technique for the measurement of attitudes. *Archives of psychology*, 1932.

[20] Jingzhou Liu, Wenhu Chen, Yu Cheng, Zhe Gan, Licheng Yu, Yiming Yang, and Jingjing Liu. Violin: A large-scale dataset for video-and-language inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10900–10910, 2020.

[21] Nadim Nachar et al. The mann-whitney u: A test for assessing whether two independent samples come from the same distribution. *Tutorials in quantitative Methods for Psychology*, 4(1):13–20, 2008.

[22] Apostol Natsev, John R Smith, Jelena Tešić, Lexing Xie, and Rong Yan. Ibm multimedia analysis and retrieval system. In *Proceedings of the 2008 international conference on Content-based image and video retrieval*, pages 553–554, 2008.

[23] Amy Pavel, Dan B. Goldman, Björn Hartmann, and Maneesh Agrawala. Sceneskim: Searching and browsing movies using synchronized captions, scripts and plot summaries. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & amp; Technology*, UIST '15, page 181–190, New York, NY, USA, 2015. Association for Computing Machinery.

[24] Jesus Perez-Martin, Benjamin Bustos, and Jorge Pérez. Improving video captioning with temporal composition of a visual-syntactic embedding. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3039–3049, 2021.

[25] Anyi Rao, Linning Xu, Yu Xiong, Guodong Xu, Qingqiu Huang, Bolei Zhou, and Dahua Lin. A local-to-global approach to multi-modal movie scene segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10146–10155, 2020.

[26] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China, Nov. 2019. Association for Computational Linguistics.

[27] Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Christopher Pal, Hugo Larochelle, Aaron Courville, and Bernt Schiele. Movie description. *International Journal of Computer Vision*, 123(1):94–120, 2017.

[28] Jitao Sang and Changsheng Xu. Character-based movie summarization. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 855–858, 2010.

[29] Charu Sharma, Siddhant R Kapil, and David Chapman. Person re-identification with a locally aware transformer. *arXiv preprint arXiv:2106.03720*, 2021.

[30] Alan F Smeaton, Bart Lehane, Noel E O'Connor, Conor Brady, and Gary Craig. Automatically selecting shots for action movie trailers. In *Proceedings of the 8th ACM international workshop on Multimedia information retrieval*, pages 231–238, 2006.

[31] John R Smith, Dhiraj Joshi, Benoit Huet, Winston Hsu, and Jozef Cota. Harnessing ai for augmenting creativity: Application to movie trailer creation. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1799–1808, 2017.

[32] Gail M Sullivan and Anthony R Artino Jr. Analyzing and interpreting data from likert-type scales. *Journal of graduate medical education*, 5(4):541–542, 2013.

[33] Domen Tabernik, Alan Lukezic, and Klemen Grm. movie2trailer: Unsupervised trailer generation using anomaly detection.

[34] Makarand Tapaswi, Martin Bäuml, and Rainer Stiefelhagen. Story-based video retrieval in tv series using plot synopses. In *Proceedings of International Conference on Multimedia Retrieval*, pages 137–144, 2014.

[35] Makarand Tapaswi, Martin Bäuml, and Rainer Stiefelhagen. Aligning plot synopses to videos for story-based retrieval. *International Journal of Multimedia Information Retrieval*, 4(1):3–16, 2015.

[36] Makarand Tapaswi, Martin Bauml, and Rainer Stiefelhagen. Book2movie: Aligning video scenes with book chapters. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1827–1835, 2015.

[37] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[38] Paul Vicol, Makarand Tapaswi, Lluis Castrejon, and Sanja Fidler. Moviegraphs: Towards understanding human-centric situations from videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8581–8590, 2018.

[39] Lezi Wang, Dong Liu, Rohit Puri, and Dimitris N Metaxas. Learning trailer moments in full-length movies with co-contrastive attention. In *European Conference on Computer Vision*, pages 300–316. Springer, 2020.

[40] Yu Xiong, Qingqiu Huang, Lingfeng Guo, Hang Zhou, Bolei Zhou, and Dahua Lin. A graph-based framework to bridge movies and synopses. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4592–4601, 2019.

[41] Hongteng Xu, Yi Zhen, and Hongyuan Zha. Trailer generation via a point process-based visual attractiveness model. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

[42] Xuan Zhang, Hao Luo, Xing Fan, Weilai Xiang, Yixiao Sun, Qiqi Xiao, Wei Jiang, Chi Zhang, and Jian Sun. Alignedreid: Surpassing human-level performance in person re-identification. *arXiv preprint arXiv:1711.08184*, 2017.

[43] Ziqi Zhang, Yaya Shi, Chunfeng Yuan, Bing Li, Peijin Wang, Weiming Hu, and Zheng-Jun Zha. Object relational graph with teacher-recommended learning for video captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13278–13288, 2020.

[44] Bolei Zhou and Dahua Lin. A unified framework for shot type classification based on subject centric lens. 2020.

[45] Athanasia Zlatintsi, Petros Koutras, Niki Efthymiou, Petros Maragos, Alexandros Potamianos, and Katerina Pastra. Quality evaluation of computational models for movie summarization. In *2015 Seventh International Workshop on Quality of Multimedia Experience (QoMEX)*, pages 1–6. IEEE, 2015.

[46] Athanasia Zlatintsi, Petros Koutras, Georgios Evangelopoulos, Nikolaos Malandrakis, Niki Efthymiou, Katerina Pastra, Alexandros Potamianos, and Petros Maragos. Cognimuse: A multimodal video database annotated with saliency, events, semantics and emotion with application to summarization. *EURASIP Journal on Image and Video Processing*, 2017(1):1–24, 2017.