Trailer Generation via a Point Process-Based Visual Attractiveness Model

Hongteng Xu[1,2], **Yi Zhen**[2], **Hongyuan Zha**[2,3]
[1]School of ECE, Georgia Institute of Technology, Atlanta, GA, USA
[2]College of Computing, Georgia Institute of Technology, Atlanta, GA, USA
[3]Software Engineering Institute, East China Normal University, Shanghai, China

Abstract

Producing attractive trailers for videos needs human expertise and creativity, and hence is challenging and costly. Different from video summarization that focuses on capturing storylines or important scenes, trailer generation aims at producing trailers that are attractive so that viewers will be eager to watch the original video. In this work, we study the problem of automatic trailer generation, in which an attractive trailer is produced given a video and a piece of music. We propose a surrogate measure of video attractiveness named fixation variance, and learn a novel self-correcting point process-based attractiveness model that can effectively describe the dynamics of attractiveness of a video. Furthermore, based on the attractiveness model learned from existing training trailers, we propose an efficient graph-based trailer generation algorithm to produce a max-attractiveness trailer. Experiments demonstrate that our approach outperforms the state-of-the-art trailer generators in terms of both quality and efficiency.

1 Introduction

With the proliferation of online video sites such as Youtube, promoting online videos through advertisements is becoming more and more popular and important. Nowadays, most advertisements consist of only key frames or shots accompanied by textual descriptions. Although these advertisements can be easily produced using existing video summarization techniques, they are oftentimes not attractive enough. Only a small portion of videos, e.g., Hollywood movies, are promoted by highly attractive trailers consisting of well-designed montages and mesmerizing music. Nevertheless, the quality of trailers generated by video summarization techniques are far from satisfactory, and it is reasonable because the goal of trailer generation is to maximize the video attractiveness, or equally, to minimize the loss of attractiveness, whereas video summarization aims at selecting key frames or shots to capture storylines or important scenes. To produce highly attractive trailers, human expertise and creativity are always needed, making trailer generation procedures costly. In this paper, we study how to produce trailers automatically and efficiently, and our approach may be applied to potentially millions of online videos and hence lower the cost substantially.

Trailer generation is challenging because we not only need to select key shots but also re-organize them in such a coherent way that the whole trailer is most attractive. Nevertheless, attractiveness is a relatively ambiguous and subjective notion, and it may be conveyed through several factors such as the shots, the order of shots, and the background music. In this paper, we propose an automatic trailer generation approach which consists of three key components: *1) A practical surrogate measure of trailer attractiveness; 2) An efficient algorithm to select and re-organize shots to maximize the attractiveness; 3) An effective method to synchronize shots and music for improved viewer experience.* Specifically, we learn an attractiveness model for movie trailers by leveraging self-correcting point process methodology [Isham and Westcott, 1979; Ogata and Vere-Jones, 1984]. Then, we position the trailer montages by exploiting the saliency information of the theme music. Finally, based upon the montage information and the attractiveness model, we construct a shot graph and generate a trailer by finding the shortest path that is equivalent to maximizing the attractiveness.

We summarize our contributions as follows: 1) We propose an effective surrogate measure of video attractiveness, namely, fixation variance. With this measure, we study the dynamics of video attractiveness and the properties of movie trailer. 2) Although point processes have been widely used for modeling temporal event sequences, such as earthquakes [Ogata and Vere-Jones, 1984] and social behaviors [Zhou *et al.*, 2013], to the best of our knowledge, our methodology is the first to model video attractiveness using self-correcting point processes, jointing a small number of existing works for leveraging point process methodology for vision problems. 3) We investigate the influence of music on trailer generation, and propose a graph-based music-guided trailer generation algorithm. Compared to the state-of-the-art methods, our method achieves significantly better results while using much less computational resource.

Related Work. Video summarization techniques have attracted much research interest and many works have been proposed in the past. Early works [Gong and Liu, 2000; Li *et al.*, 2001] extract features of frames and cluster frames accordingly, but their performance is limited. Besides vi-

sual frames, other information has been taken into consideration in video summarization. For example, viewer attention model based summarization methods are proposed [Ma *et al.*, 2005; You *et al.*, 2007]. User interaction is incorporated in a generic framework of video summarization [Li *et al.*, 2006]. Textual information is used to achieve transfer learning based video summarization [Li *et al.*, 2011]. Audio and visual analysis is performed simultaneously [Jiang *et al.*, 2011]. Recently, a joint aural, visual, and textual attention model is proposed for movie summarization [Evangelopoulos *et al.*, 2013]. Moreover, semantic information of videos has been exploited, including the saliency maps of images [Yan *et al.*, 2010], special events [Wang *et al.*, 2011; 2012], key people and objects [Lee *et al.*, 2012; Khosla *et al.*, 2013], storylines [Kim *et al.*, 2014]. The external information sources such as web images have also been demonstrated to be useful [Khosla *et al.*, 2013; Kim *et al.*, 2014]. Focusing on trailer generation problem, so far as we know, Ma *et al.* proposed the first user attention model for trailer generation [Ma *et al.*, 2002]. Irie *et al.* [Irie *et al.*, 2010] built a trailer generator that combines a topic model based on Plutchik's emotions [Hospedales *et al.*, 2009; Irie *et al.*, 2009] with the Bayesian surprise model [Itti and Baldi, 2009], and their system is reported to achieve the state-of-the-art performance. However, the system does not incorporate the relationships between a trailer and its music, and the causality between the surprise degree and video attractiveness is questionable. In all the above works, the definition and measures of the video attractiveness are largely overlooked, which turns out to be rather critical for trailer generation.

2 Properties of Movie Trailer

Suppose that we have a set of K training trailers $\{\mathcal{T}_k\}_{k=1}^K$, and in total N shots $\{\mathcal{C}_i\}_{i=1}^N$. Each shot comes from one trailer and consists of a set of frames. Let $\mathcal{C}_i \in \mathcal{T}_k$ indicate that shot $\mathcal{C}_i$ comes from the trailer $\mathcal{T}_k$, and $\mathcal{C}_i = \{f_j^{(i)}\}_{j=1}^{n_i}$ indicate that there are n_i frames in $\mathcal{C}_i$ and $f_j^{(i)}$ is the jth frame of $\mathcal{C}_i$. Similarly, we use $\mathcal{T} \subset \mathcal{V}$ to indicate that $\mathcal{T}$ is the trailer of the video $\mathcal{V}$. We also represent a video $\mathcal{V}$ or a trailer $\mathcal{T}$ as a sequence of shots, denoted as $\{\mathcal{C}_i\}_{i=1}^N$, in the sequel. We use the index of a frame as the time stamp of the shot (trailer and movie) for convenience. The beginning and the ending of a video or a trailer are denoted as $L_0 = 0$ and $L_N = \sum_{i=1}^N n_i$. The position of montage between $\mathcal{C}_i$ and $\mathcal{C}_{i+1}$ is denoted as $L_i = L_{i-1} + n_i = \sum_{j=1}^i n_j$. The *trailer generation problem* is: given a video $\mathcal{V}$ and a piece of music $\mathbf{m}$, we would like to generate a trailer $\mathcal{T} \subset \mathcal{V}$ that is the most attractive.

2.1 Measure and Dynamics of Attractiveness

We might observe such a common phenomenon: when attractive scenes such as handsome characters and hot actions appear, viewers will look at the same area on the screen; on the other hand, when boring scenes such as the cast of characters and tedious dialogues appear, viewers will no longer focus on the same screen area. In other words, the attractiveness of a video is highly correlated with the attention of viewers when they watch the video. Therefore, we propose a surrogate measure of attractiveness based on viewers' eye-movement, whose efficacy is validated by the following experiments. Specifically, we invite 14 (6 female and 8 male) volunteers to watch 8 movie trailers, which contain 1,083 shots[1]. We further record the motions of their gazes and calculate the mapped fixation points in each frame using Tobii T60 eye tracker. Denote the locations of gaze on the screen, namely, the fixation points, in the jth frame of $\mathcal{C}_i$ as $[\mathbf{x}_j^{(i)}, \mathbf{y}_j^{(i)}]$, where $\mathbf{x}_j^{(i)} \in \mathbb{R}^{14}$ (resp. $\mathbf{y}_j^{(i)} \in \mathbb{R}^{14}$) is the vector of the horizontal (resp. vertical) coordinates of the fixation points of the 14 volunteers. For the jth frame, we define the *fixation variance* as the determinant of the covariance matrix of the fixation points:

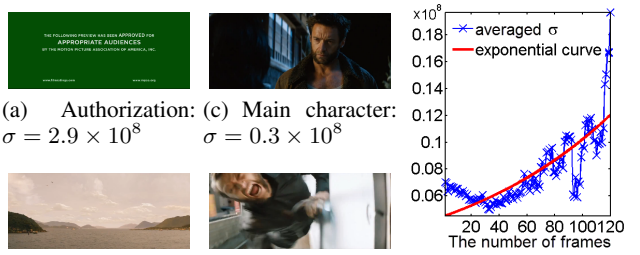
$$\sigma_j^{(i)} = \det\left(\mathrm{cov}([\mathbf{x}_j^{(i)}, \mathbf{y}_j^{(i)}])\right), \tag{1}$$

We average $\{\sigma_j^{(i)}\}_{j=1}^{n_i}$ for the frames belonging to the same shot. The averaged fixation variance reflects the spread of attention when watching the shot. Following the above reasoning and definition, we expect that the boring shots (e.g., background) should have large fixation variance whereas the attractive shots (e.g., hot action scenes, characters) should have small fixation variance. To verify this, we label these two types of shots manually and calculate the statistics of their fixation variance. The results are summarized in Table 1.

Table 1: The statistics of normalized fixation variance ($\times 10^8$)

	mean(σ)	median(σ)	variance(σ)
Boring shots	1.19	0.45	0.03
Attractive shots	**0.60**	**0.22**	0.01

It is easy to observe that both the mean and the median of the fixation variance of boring shots are about twice larger than those of the attractive shots. The variance is very small, meaning that our proposed fixation variance is stable in both boring and attractive shot groups. These results show that fixation variance is negatively correlated with the video attractiveness — it measures the loss of attractiveness accurately and robustly. Fig. 1(a-d) further shows typical examples.



(a) Authorization: $\sigma = 2.9 \times 10^8$ (c) Main character: $\sigma = 0.3 \times 10^8$

(b) The background: $\sigma = 1.9 \times 10^8$ (d) Action scenes: $\sigma = 0.1 \times 10^8$ (e) Dynamics of fixation variance.

Figure 1: (a-d) Four shots in trailer "*The Wolverine 2013*" and their fixation variances. (e) Dynamics of fixation variance calculated from training trailers.

[1]In this paper, we segment video into shots using a commercial software "CyberLink PowerDirector".

The dynamics of fixation variance. Given N shots $\{\mathcal{C}_i\}_{i=1}^N$, the averaged fixation variance in the jth frame is calculated as $\bar{\sigma}_j = \frac{1}{N_j}\sum_{i=1}^{N_j}\sigma_j^{(i)}$, where N_j is the number of shots having at least j frames. In Fig. 1(e), we find that the change of $\bar{\sigma}_j$ over time can be approximated by an increasing exponential curve. In other words, within a shot, its attractiveness decreases over time. The inter-shot dynamics can be modeled as the stitching of the fitted exponential curves for adjacent shots. It means that although the attractiveness within one shot decreases over time, the montage between shots increases attractiveness.

2.2 The Dynamics of Music

Similar to the dynamics of attractiveness, we empirically find that the dynamics of music are also highly correlated with the montages between shots. To see this, we first detect the saliency points of the music associated with a trailer as follows. 1) Using the saliency detection algorithm in [Hou *et al.*, 2012], we extract the saliency curve of a piece of music as

$$\widehat{\mathbf{m}} = \mathrm{G}((\mathrm{idct}(\mathrm{sign}(\mathrm{dct}(\mathbf{m}))))^2), \tag{2}$$

where $\mathrm{dct}(\cdot)$ and $\mathrm{idct}(\cdot)$ are a DCT transformation pair, $\mathrm{sign}(\cdot)$ is the sign function that returns 1, -1, and 0 for positive, negative, and zero inputs, respectively. $\mathrm{G}(\cdot)$ is a Gaussian filter. 2) After re-sampling $\widehat{\mathbf{m}}$ with the number of frames, we detect the peaks in $\widehat{\mathbf{m}}$. Regarding these peaks as the saliency points of the music, we investigate their correlations with the montages in the trailer as follows. For each peak, we label the peak as a correct indicator of the location of the montage when a montage appears within ± 6 frames (about 0.25 second). On the 8 sample trailers, we find that the time stamps of the saliency points are highly correlated with those of the montages (with accuracy 84.88%). Fig. 2 presents an example: in high-quality trailers, the montages of shots are synchronized with the rhythm of the background music.

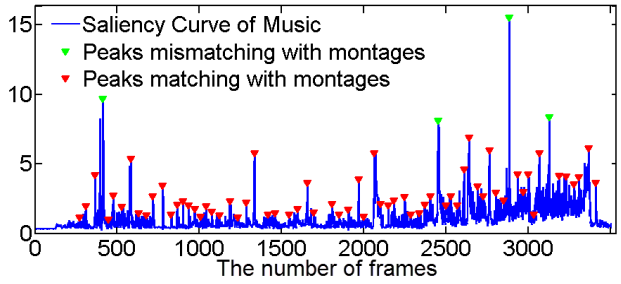


Figure 2: The Saliency Points of Music v.s. Montage Positions (Trailer of "*The Bling Ring*").

According to the analytic experiments above, we summarize three properties of movie trailer as follows: **Property 1.** *The loss of attractiveness can be approximated by a surrogate measure, namely, fixation variance.* **Property 2.** *Within each shot, the loss of attractiveness increases exponentially and this tendency is corrected when a new shot appears.* **Property 3.** *The self-correction of attractiveness, the rhythm of the music in the video, and the montages between shots are highly correlated.*

3 Point Process-based Attractiveness Model

3.1 Motivation

Our modeling assumption is that the fixation variance is highly correlated with the number of viewers losing their attention on the screen, which directly connects the notion of the attractiveness of a video with a specific point process model we will discuss below. To this end, suppose that there are V viewers watching a movie. We define a sequence $E_v(t)$ of the event "whether the viewer loses her attention or not at time t" for each viewer v:

$$E_v(t) = \begin{cases} 1, & \text{viewer } v \text{ loses her attention at time } t, \\ 0, & \text{otherwise.} \end{cases}$$

Although we do not observe the event sequence directly, we assume that the fixation variance is proportional to the number of viewers losing their attention. Therefore, given the fixation variance of training trailers, we can approximate the aggregated observations of viewers' events $\sum_v E_v(t)$, and model viewers' events as a temporal point process. Specifically, we propose an attractiveness model based on a specific point process, i.e., self-correcting point process.

3.2 Self-Correcting Point Process

A self-correcting point process is a point process with the following intensity function:

$$\lambda(t) = \frac{\mathbb{E}(dN(t)|\mathcal{H}_t)}{dt} = \exp\left(\alpha t - \sum\nolimits_{i:t_i<t}\beta\right), \tag{3}$$

where $N(t)$ is the number of events occurred in time range $(-\infty, t]$, $\mathcal{H}_t$ denotes the historical events happened before time t, and $\mathbb{E}(dN(t)|\mathcal{H}_t)$ is the expectation of the number of events happened in the interval $(t, t+dt]$ given historical observations $\mathcal{H}_t$. The intensity function in Eq. (3) represents the expected instantaneous rate of future events at time t.

The intensity function of the self-correcting point process increases exponentially with rate α and this tendency can be corrected by the historical observations via rate β. Note that the intensity function exactly matches the dynamics of attractiveness described in Property 2. Therefore, given a video $\mathcal{V} = \{\mathcal{C}_i\}_{i=1}^N$, for each shot $\mathcal{C}_i$, we define the *local* intensity function in its time interval $(0, n_i]$ as

$$\lambda_{\mathcal{C}_i}(t) = \exp\left(\alpha_i H^i t - \beta_i D_t^i\right), \tag{4}$$

where $t \in (0, n_i]$ and

$$H^i = \begin{cases} H(\hat{f}_1^{(1)}), & i = 1, \\ D(\hat{f}_{N_{i-1}}^{(i-1)} || \hat{f}_1^{(i)}), & i > 1, \end{cases}$$

$$D_t^i = \sum_{j=2}^{\lfloor t \rfloor} D(\hat{f}_{j-1}^{(i)} || \hat{f}_j^{(i)}),$$

where $\hat{f} = \mathrm{G}((\mathrm{idct2D}(\mathrm{sign}(\mathrm{dct2D}(f))))^2)/C$ is the normalized saliency map of frame f [Hou *et al.*, 2012]. C is a l_1 normalizer that guarantees $\hat{f}$ to be a distribution. For the first shot $\mathcal{C}_1$ of $\mathcal{V}$ ($i = 1$), $H^i = H(\hat{f}_1^{(1)})$ represents the entropy

of the first frame in $\mathcal{C}_1$, which is the *initial stimulus* given by $\mathcal{C}_1$. For the following shots ($i > 1$), $H^i = D(\hat{f}^{(i-1)}_{N_{i-1}} || \hat{f}^{(i)}_1)$ represents the KL-divergence between the last frame of $\mathcal{C}_{i-1}$ and the first frame of $\mathcal{C}_i$, which is the *initial stimulus* given by $\mathcal{C}_i$. Similarly, $D(\hat{f}^{(i)}_{j-1} || \hat{f}^{(i)}_j)$ represents the KL-divergence between the adjacent frames in $\mathcal{C}_i$, which is the *supplementary stimulus*. $D^i_t = \sum_{j=2}^{\lfloor t \rfloor} D(\hat{f}^{(i)}_{j-1} || \hat{f}^{(i)}_j)$ is the accumulative influence caused by the supplementary stimulus in the time interval $(0, t]$, where $\lfloor t \rfloor$ is the largest integer smaller than t.

The intensity function in Eq. (4) imitates the loss of attractiveness — it increases exponentially with the initial stimulus till the supplementary stimulus corrects this tendency. For each shot $\mathcal{C}_i$, its intensity function has two non-negative parameters (α_i, β_i). To summarize, we define our attractiveness model as a self-correcting point process with a *global* intensity function for time interval $(0, \sum_{i=1}^{N} n_i]$, which stitches N local intensity functions as $\lambda_{\mathcal{V}}(t) = \sum_{i=1}^{N} \lambda_{\mathcal{C}_i}(t - L_{i-1})$, where $L_0 = 0$, $L_i = L_{i-i} + n_i = \sum_{j=1}^{i} n_j$.

3.3 Model Learning

Given K training trailers $\{\mathcal{T}_k\}_{k=1}^{K}$ consisting of N shots, our goal is to learn parameters of the N *local* intensity functions. Similar to [Zhou *et al.*, 2013], we achieve this goal by pursuing an maximum likelihood estimation (MLE) of the parameters, and the likelihood function can be written as:

$$\mathsf{L} = \prod_{i=1}^{N} \mathsf{L}_i = \prod_{i=1}^{N} \Bigg(\Big(\prod_{t=1}^{n_i} (\lambda_{\mathcal{C}_i}(t))^{\sigma_{\mathcal{C}_i}(t)} \Big) \exp\left(-\sigma_m \int_0^{n_i} \lambda_{\mathcal{C}_i}(s) ds \right) \Bigg), \tag{5}$$

where L_i denotes the local likelihood for $\mathcal{C}_i$. $\sigma_{\mathcal{C}_i}(t)$ is the fixation variance of the tth frame of shot $\mathcal{C}_i$. σ_m is the maximum of fixation variance.

Besides capturing the *global* event dynamics by maximizing Eq. (5), we also require the proposed model to fit *local* event information in each time interval. Therefore, we further propose to minimize a novel data fidelity loss function to correlate the local intensity and fixation variance in each frame: $\sum_{i=1}^{N} \sum_{t=1}^{n_i} |\log(\gamma \lambda_{\mathcal{C}_i}(t) / \sigma_{\mathcal{C}_i}(t))|^2$, where γ is shared by all frames. This term encourages the local intensity (scaled by γ) to be equal to the fixation variance in each frame. To sum up, we learn our model by solving the following problem:

$$\begin{aligned} &\min_{\boldsymbol{\alpha}, \boldsymbol{\beta}, \gamma} -\log(\mathsf{L}) + \mu \sum_{i=1}^{N} \sum_{t=1}^{n_i} \left| \log\left(\frac{\gamma \lambda_{\mathcal{C}_i}(t)}{\sigma_{\mathcal{C}_i}(t)} \right) \right|^2, \\ &s.t.\ \boldsymbol{\alpha} \geq \mathbf{0},\ \boldsymbol{\beta} \geq \mathbf{0},\ \gamma > 0, \end{aligned} \tag{6}$$

where $\boldsymbol{\alpha} = [\alpha_1, .., \alpha_N]$ and $\boldsymbol{\beta} = [\beta_1, .., \beta_N]$ represent the parameters of N local intensity functions.

We develop an alternating algorithm to solve Eq. (6). To be specific, given the initial values of $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \gamma)$, we first solve the following subproblem for each shot with γ fixed:

$$\begin{aligned} &\min_{\alpha_i, \beta_i} -\log(\mathsf{L}_i) + \mu \sum_{t=1}^{n_i} \left| \log\left(\frac{\gamma \lambda_{\mathcal{C}_i}(t)}{\sigma_{\mathcal{C}_i}(t)} \right) \right|^2 \\ &s.t.\ \alpha_i \geq 0,\ \beta_i \geq 0. \end{aligned} \tag{7}$$

The objective of Eq. (7) can be written as follows:

$$\begin{aligned} I_i = \sum_{t=1}^{n_i} \Bigg(& \sigma_{\mathcal{C}_i}(t)(\beta_i D^i_t - \alpha_i H^i t) \\ & + \sigma_m \frac{e^{\alpha_i H^i t} - e^{\alpha_i H^i (t-1)}}{\alpha_i H^i e^{\beta_i D^i_t}} \\ & + \mu |\log\gamma + \alpha_i H^i t - \beta_i D^i_t - \log\sigma_{\mathcal{C}_i}(t)|^2 \Bigg). \end{aligned}$$

We can solve the subproblems using gradient-based methods in a parallel manner. With $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ such learned, γ can be updated using the following equation:

$$\gamma = \exp\left(\frac{\sum_{i=1}^{N} \sum_{t=1}^{n_i} \log(\sigma_{\mathcal{C}_i}(t) / \lambda_{\mathcal{C}_i}(t))}{\sum_{i=1}^{N} n_i} \right). \tag{8}$$

Algorithm 1 summarizes our learning algorithm.

Algorithm 1 Learning Proposed Attractiveness Model

Input: Training shots $\{\mathcal{C}_i\}_{i=1}^{N}$, the maximum number of iteration $M = 500$, the parameter $\mu = 0.5$. The gradient descent step size: $\delta_\alpha = 10^{-4}$, $\delta_\beta = 10^{-5}$.
Output: Parameters of our model $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, γ.
Initialize $\boldsymbol{\alpha}^0$, $\boldsymbol{\beta}^0$ and γ^0 randomly.
for $m = 1 : M$ **do**
  for $i = 1 : N$ **do**
    $\alpha_i^m = \left(\alpha_i^{m-1} - \delta_\alpha \frac{\partial I_i}{\partial \alpha_i} \big|_{\alpha_i = \alpha_i^{m-1}} \right)_+$.
    $\beta_i^m = \left(\beta_i^{m-1} - \delta_\beta \frac{\partial I_i}{\partial \beta_i} \big|_{\beta_i = \beta_i^{m-1}} \right)_+$.
    $(\cdot)_+$ sets negative value to be 0.
  end for
  γ^m is calculated by Eq. (8).
end for
$\boldsymbol{\alpha} = \boldsymbol{\alpha}^m$, $\boldsymbol{\beta} = \boldsymbol{\beta}^m$, $\gamma = \gamma^m$.

4 Trailer Generation

After learning the attractiveness model from the training set, we are able to generate an attractive trailer given a new testing video $\mathcal{V}$ and a piece of music $\mathbf{m}$ with maximum attractiveness, or equally, minimum loss of attractiveness. The problem of trailer generation can be formulated as:

$$\min_{\mathcal{T}} \int_{L_0}^{L_N} \lambda_{\mathcal{T}}(s) ds, \qquad s.t.\ \mathcal{T} \subset \mathcal{V}, \tag{9}$$

where $L_0 = 0$ and $L_N = \sum_{i=1}^{N} n_i$ are the beginning and the ending of the trailer, respectively. n_i is the number of frames in $\mathcal{C}_i$ and N is the number of candidate shots selected from $\mathcal{V}$. $\{n_i\}$ and N are the positions of the montages. According to Property 3 in Section 2.2, they are determined by the saliency points of the music, which is detected from the $\widehat{\mathbf{m}}$ in Eq. (2). The interval length between adjacent saliency points determines n_i, and the number of saliency points determines N. Since Eq. (9) is a combinatorial problem and NP-hard, we have to resort to approximate solutions. Inspired by [Xu *et*

al., 2014], we propose a graph-based method to solve Eq. (9) approximately and efficiently.

Step 1: Candidate Selection. We rewrite Eq. (9) as:

$$\min_{\mathcal{T}} \sum_{i=1}^{N} \int_{0}^{n_i} \lambda_{\mathcal{C}_i}(s)ds, \quad s.t.\ \mathcal{T} = \{\mathcal{C}_i \in \mathcal{S}_i\}_{i=1}^{N}, \qquad (10)$$

where $\mathcal{S}_i$, $i = 1, ..., N$, is the set of s ($s = 5$ in our experiments) candidate shots selected from $\mathcal{V}$ for $\mathcal{C}_i$. Each selected shot satisfies the following two constraints: 1) the length is not shorter than n_i; 2) it does not appear in $\mathcal{S}_{i-1}$. $\mathcal{S}_N$ contains only one shot, which corresponds to the title of the trailer given in advance.

Step 2: Parameter Assignment. For a candidate shot $\mathcal{C}_i$ in the new video, we do not know the parameters of $\lambda_{\mathcal{C}_i}(t)$ in advance. In this paper, we first extract the feature of $\mathcal{C}_i$ as

$$\mathbf{f}_{\mathcal{C}_i} = [H(\hat{f}_1^{(i)}), D(\hat{f}_1^{(i)} || \hat{f}_2^{(i)}), ...] \in \mathbb{R}^{64}. \qquad (11)$$

We fix the length of feature vector as 64 in this work — if the length of $\mathcal{C}_i$ is shorter than 64, we pad zeros in the end of $\mathbf{f}_{\mathcal{C}_i}$; if $\mathcal{C}_i$ is longer than 64, we cut the end of $\mathbf{f}_{\mathcal{C}_i}$. Then, we select the matching shot in the training set for $\mathcal{C}_i$ by comparing the features of training shots with those of candidate shots. The matching criterion is the Euclidean distance. The parameters of the matching shot are assigned to $\lambda_{\mathcal{C}_i}(t)$.

Step 3: Graph-based Stitching. Eq. (10) is still a complicated combinatorial optimization problem. To see this, we note that the selection of $\mathcal{C}_{i-1}$ has recursive influence on the selection of subsequent shots $\{\mathcal{C}_i, \mathcal{C}_{i+1}, ...\}$. As we know, the initial stimulus in $\lambda_{\mathcal{C}_i}(t)$ is the KL-divergence between the last frame of $\mathcal{C}_{i-1}$ and the first frame of $\mathcal{C}_i$. Hence, if we change $\mathcal{C}_{i-1}$, the intensity function $\lambda_{\mathcal{C}_i}(t)$ will change, and so dose the selection of $\mathcal{C}_i$.

The problem will be efficiently solved if we only consider the pairwise relationships between the shots in the adjacent candidate sets. Given $\{\mathcal{S}_i\}_{i=1}^{N}$, we can construct a trellis graph $\mathcal{G}$ with $N + 1$ layers. The nodes in the ith layer are the candidate shots from $\mathcal{S}_i$. The edge weights in the graph can be defined as follows,

$$w_{p,q}^{i,i+1} = \int_{0}^{n_i} \lambda_{\mathcal{C}_{p,i}}(s)ds + \int_{0}^{n_{i+1}} \lambda_{\mathcal{C}_{q,i+1}}(s)ds, \qquad (12)$$

where $w_{p,q}^{i,i+1}$ is the weight connecting the pth candidate in $\mathcal{S}_i$ with the qth candidate in $\mathcal{S}_{i+1}$. We calculate all the weights independently: the initial stimulus in $\lambda_{\mathcal{C}_{p,i}}(t)$ is the entropy of the first frame of $\mathcal{C}_{p,i}$, which is independent of shot selection in the former layers; on the other hand, $\mathcal{C}_{p,i}$ only influences $\mathcal{C}_{q,i+1}$ through the initial stimulus in $\lambda_{\mathcal{C}_{q,i+1}}$ that is the KL-divergence between the last frame of $\mathcal{C}_{p,i}$ and its first frame. Influence of $\mathcal{C}_{p,i}$ will not propagate to the following layers. In other words, Eq. (10) can be solved approximately by finding the shortest path [Dijkstra, 1959] in the graph $\mathcal{G}$ (from the first layer to the last one). We summarize our algorithm in Algorithm 2 and illustrate it in Fig. 3.

Algorithm 2 Graph-based Trailer Generation Algorithm

Input: a video $\mathcal{V}$, a piece of music $\mathbf{m}$, training shots with features and learned parameters.

Output: a movie trailer $\mathcal{T}$.

1. Segment $\mathcal{V}$ to shots $\{\mathcal{C}_j\}$ and extract features.
2. For each $\mathcal{C}_j$, find the matching shot in the training set and assign parameters accordingly.
3. Detect saliency points of $\mathbf{m}$ by Eq. (2).
4. Construct candidate set $\mathcal{S}_i$ and a hierarchical graph $\mathcal{G}$.
5. Calculate the weight of edge by Eqs. (4,12).
6. Find the shortest path in $\mathcal{G}$.
7. $\mathcal{T}$ is constructed by the sequence of shots corresponding to the shortest path associated with the music $\mathbf{m}$.

5 Experiments

We conduct two groups of experiments (objective and subjective) to empirically evaluate the proposed method. Our data set consists of 16 publicly available movies including 3 animation movies, 2 fantasy movies, 2 action movies, 5 fiction action movies and 4 dramas in 2012 and 2014. We also collect the movies, their theme music and official trailers. The experimental settings are as follows: we first select 8 of the trailers as the training set and collect the fixation data from 14 volunteers. Then, we learn our attractiveness model as described in Section 3.3. Finally, based on the attractiveness model, we produce trailers for the remaining 8 movies following Section 4. All movies and their trailers are with frame size 640×480.

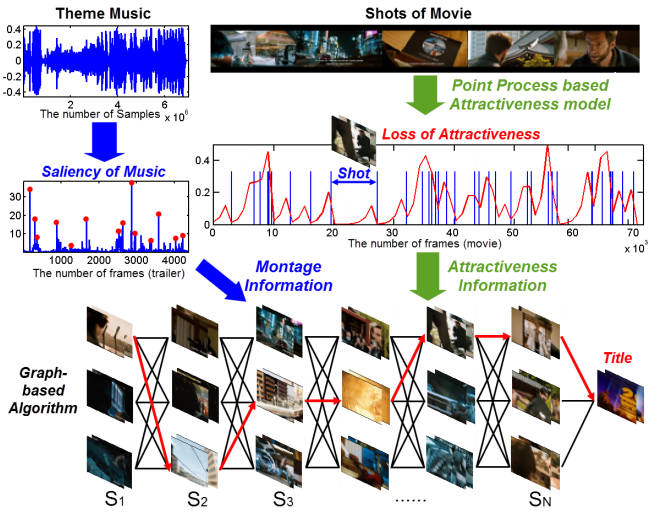


Figure 3: The scheme of our trailer generator.

We compare our method ("***Ours***") with the following four competitors[2]: *i*) the trailer generator "***V2T***" in [Irie *et al.*, 2010]; *ii*) the commercial video summarization software "***Muvee***[3]"; *iii*) the real official trailers ("***RT***") generated by professionals; *iv*) the real trailers without speech information ("***RTwS***"). For fair comparison, for "*V2T*", the training and testing sets are the same as we described above; for "*Muvee*", we generate trailers for the testing movies only as there

[2]Representative trailers generated by all methods are on the website: https://vimeo.com/user25206850/videos.

[3]http://www.muvee.com/

is no training phase. We also note that, in this work, we focus on the visual attractiveness model and its contribution on trailer generation, and hence we use "*RTwS*" as a baseline. Moreover, we are only able to implement the shot selection and arrangement algorithm of "*V2T*" since other steps such as feature extraction is omitted in the reference.

5.1 Objective Evaluation

Loss of Attractiveness. An important criterion for trailers is the loss of attractiveness, which can be approximately measured by the proposed fixation variance. We invite the 14 volunteers to watch the testing trailers generated by all of the five methods mentioned above, and record their fixation points in each frames by an eye tracker. For each method, we calculate the fixation variance σ in each frame for all 8 testing trailers, and hence obtain 32,309 σ's. The statistics of σ's reflect the overall loss of attractiveness. Specifically, larger σ indicates more loss of attractiveness. We present the mean, the median and the standard deviation of σ's for each method in Fig. 4.

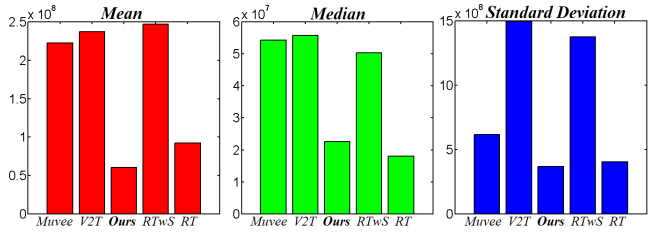


Figure 4: The mean, the median and the standard deviation of the fixation variance σ for various methods.

It is easy to observe that both the mean and the median obtained by our method are the smallest compared with its counterparts. Specifically, the results of our method are comparable to those of "*RT*" whereas the results of "*Muvee*", "*V2T*" and "*RTwS*" are much higher than those of "*RT*". Last but not least, the standard deviation results show that the attractiveness of our trailers is the most stable, and "*RT*" trailers are the second best in this aspect. In our opinion, the superiority of our method is mainly based on the utilizing of fixation variance and the proposed point process model. Firstly, fixation variance is real feature from viewers, which reflects the attractiveness of video better than the learned feature from video. Secondly, the proposed point process model captures the dynamics of attractiveness well, which provides us with a useful guidance to generate trailer.

Table 2: Comparison on computational cost.

	Training cost	Testing cost
V2T [Irie *et al.*, 2010]	0.0024 sec/frame	0.2676 sec/frame
Ours	**0.0014** sec/frame	**0.0113** sec/frame

Computational Cost. Computation cost is a key factor for trailer generation. As we mentioned in previous sections, it is very promising to have efficient automatic trailer generators that may be potentially applied to millions of online videos. We note that the training and testing computational complexities of "*V2T*" are both $O(N^3)$, whereas ours are $O(N)$ and $O(N^2)$ for training and testing, respectively.

Table 2 compares empirical training and testing cost of our method and "*V2T*". Both methods are implemented by MATLAB and run on the same platform (Core i7 CPU @3.40GHz with 32GB memory). Specifically, the training cost is calculated as the model learning time per frame for training set, and the testing cost is calculated as the trailer generation time per frame for the generation result. Table 2 validates that our method needs much less training and testing cost than "*V2T*".

5.2 Subjective Evaluation

In this subsection, we evaluate our method as well as the baselines through subjective experiments. Similar to [Irie *et al.*, 2010], for each testing trailer generated by different methods, we invited 14 volunteers to evaluate it by answering the following 3 questions: ***Rhythm***: "How well does the montage match with the rhythm of background music?" ***Attractiveness***: "How attractive is the trailer?" ***Appropriateness***: "How close is the trailer to an real trailer?" For each question, the volunteers were asked to provide an integer score in the range of 1 (lowest) to 7 (highest). Fig. 5 shows the overall results for all 8 testing movies.

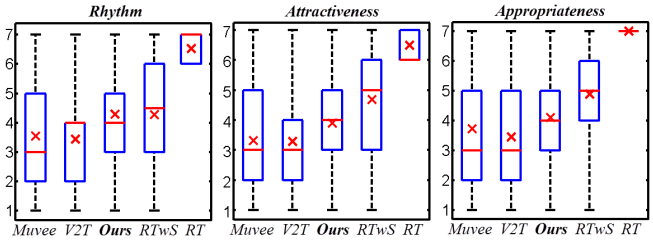


Figure 5: The box plots of scores for various methods on three questions. The red crosses are means and the red bars are medians.

Consistency of Objective and Subjective Evaluation. Similar to objective evaluation, we find in Fig. 5 that our method is better than "*V2T*" and "*Muvee*" in all three questions, indicating that our attractiveness model based on fixation variance and self-correcting point processes is reasonable, and is able to generate trailers that satisfy our subjective feelings. On the other hand, our method are inferior to "*RTwS*" and "*RT*", which is different from the result of objective evaluation. The reasons for the difference may be attributed to that we only use visual information to learn the attractiveness model and produce trailers. On the contrary, official trailers of "*RT*" often provide speeches, subtitles and special effects of montages, which impress viewers a lot. Similarly, although the trailers of "*RTwS*" does not have speech information, they still contain subtitles and special effects of montages. Since the information of these factors contribute to raising the attractiveness of the video, volunteers feel that the real trailers are better than our trailers. This observation points out a future extension of our work, that is, in addition to visual and music information, we should also enrich our model with information sources such as speech, subtitles, and montage effects to capture holistic movie attractiveness.

6 Conclusion and Future Work

In this work, we studied a challenging problem, namely, automatic trailer generation. To generate attractive trailers, we proposed a practical surrogate measure of video attractiveness called fixation variance, and made the first attempt to use point processes to model the attractiveness dynamics. Based upon the attractiveness model, we developed a graph-based music-guided trailer generation method. In the future, we are interested in extending our method to utilize other information such as speeches and subtitles. We would also like to explore parallel algorithms to further improve the scalability of our method.

Acknowledgement. This work is supported in part by NSF DMS-1317424 and NSFC-61129001/F010403.

References

[Dijkstra, 1959] Edsger W Dijkstra. A note on two problems in connexion with graphs. *Numerische mathematik*, 1(1):269–271, 1959.

[Evangelopoulos *et al.*, 2013] G. Evangelopoulos, A Zlatintsi, A Potamianos, P. Maragos, K. Rapantzikos, G. Skoumas, and Y. Avrithis. Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention. *Multimedia, IEEE Transactions on*, 15(7):1553–1568, 2013.

[Gong and Liu, 2000] Yihong Gong and Xin Liu. Video summarization using singular value decomposition. In *CVPR*, volume 2, pages 174–180, 2000.

[Hospedales *et al.*, 2009] Timothy Hospedales, Shaogang Gong, and Tao Xiang. A markov clustering topic model for mining behaviour in video. In *ICCV*, pages 1165–1172, 2009.

[Hou *et al.*, 2012] Xiaodi Hou, Jonathan Harel, and Christof Koch. Image signature: Highlighting sparse salient regions. *PAMI*, 34(1):194–201, 2012.

[Irie *et al.*, 2009] Go Irie, Kota Hidaka, Takashi Satou, Akira Kojima, Toshihiko Yamasaki, and Kiyoharu Aizawa. Latent topic driving model for movie affective scene classification. In *ACM Multimedia*, pages 565–568, 2009.

[Irie *et al.*, 2010] Go Irie, Takashi Satou, Akira Kojima, Toshihiko Yamasaki, and Kiyoharu Aizawa. Automatic trailer generation. In *ACM Multimedia*, pages 839–842, 2010.

[Isham and Westcott, 1979] Valerie Isham and Mark Westcott. A self-correcting point process. *Stochastic Processes and Their Applications*, 8(3):335–347, 1979.

[Itti and Baldi, 2009] Laurent Itti and Pierre Baldi. Bayesian surprise attracts human attention. *Vision research*, 49(10):1295–1306, 2009.

[Jiang *et al.*, 2011] Wei Jiang, Courtenay Cotton, and Alexander C Loui. Automatic consumer video summarization by audio and visual analysis. In *ICME*, pages 1–6, 2011.

[Khosla *et al.*, 2013] Aditya Khosla, Raffay Hamid, Chih-Jen Lin, and Neel Sundaresan. Large-scale video summarization using web-image priors. In *CVPR*, pages 2698–2705, 2013.

[Kim *et al.*, 2014] Gunhee Kim, Leonid Sigal, and Eric P Xing. Joint summarization of large-scale collections of web images and videos for storyline reconstruction. In *CVPR*, 2014.

[Lee *et al.*, 2012] Yong Jae Lee, Joydeep Ghosh, and Kristen Grauman. Discovering important people and objects for egocentric video summarization. In *CVPR*, pages 1346–1353, 2012.

[Li *et al.*, 2001] Ying Li, Tong Zhang, and Daniel Tretter. An overview of video abstraction techniques. *HP Laboratories Palo Alto*, 2001.

[Li *et al.*, 2006] Ying Li, Shih-Hung Lee, Chia-Hung Yeh, and C-CJ Kuo. Techniques for movie content analysis and skimming: tutorial and overview on video abstraction techniques. *Signal Processing Magazine, IEEE*, 23(2):79–89, 2006.

[Li *et al.*, 2011] Liangda Li, Ke Zhou, Gui-Rong Xue, Hongyuan Zha, and Yong Yu. Video summarization via transferrable structured learning. In *WWW*, pages 287–296. ACM, 2011.

[Ma *et al.*, 2002] Yu-Fei Ma, Lie Lu, Hong-Jiang Zhang, and Mingjing Li. A user attention model for video summarization. In *ACM Multimedia*, pages 533–542, 2002.

[Ma *et al.*, 2005] Yu-Fei Ma, Xian-Sheng Hua, Lie Lu, and Hong-Jiang Zhang. A generic framework of user attention model and its application in video summarization. *Multimedia, IEEE Transactions on*, 7(5):907–919, 2005.

[Ogata and Vere-Jones, 1984] Y Ogata and D Vere-Jones. Inference for earthquake models: a self-correcting model. *Stochastic processes and their applications*, 17(2):337–347, 1984.

[Wang *et al.*, 2011] Zheshen Wang, Mrityunjay Kumar, Jiebo Luo, and Baoxin Li. Sequence-kernel based sparse representation for amateur video summarization. In *Proceedings of the 2011 joint ACM workshop on Modeling and representing events*, pages 31–36, 2011.

[Wang *et al.*, 2012] Meng Wang, Richang Hong, Guangda Li, Zheng-Jun Zha, Shuicheng Yan, and Tat-Seng Chua. Event driven web video summarization by tag localization and key-shot identification. *Multimedia, IEEE Transactions on*, 14(4):975–985, 2012.

[Xu *et al.*, 2014] Hongteng Xu, Hongyuan Zha, and Mark Davenport. Manifold based dynamic texture synthesis from extremely few samples. In *CVPR*, pages 3019–3026, 2014.

[Yan *et al.*, 2010] Junchi Yan, Mengyuan Zhu, Huanxi Liu, and Yuncai Liu. Visual saliency detection via sparsity pursuit. *Signal Processing Letters, IEEE*, 17(8):739–742, 2010.

[You *et al.*, 2007] Junyong You, Guizhong Liu, Li Sun, and Hongliang Li. A multiple visual models based perceptive analysis framework for multilevel video summarization. *CSVT, IEEE Transactions on*, 17(3):273–285, 2007.

[Zhou *et al.*, 2013] Ke Zhou, Hongyuan Zha, and Le Song. Learning social infectivity in sparse low-rank networks using multi-dimensional hawkes processes. In *AISTATS*, pages 641–649, 2013.