MovieCuts: A New Dataset and Benchmark for Cut Type Recognition

Alejandro Pardo[1], Fabian Caba Heilbron[2], Juan León Alcázar[1], Ali Thabet[1,3], and Bernard Ghanem[1]

[1] King Abdullah University of Science and Technology, KAUST
{alejandro.pardo,juancarlo.alcazar,ali.thabet,bernard.ghanem}@kaust.edu.sa
[2] Adobe Research {caba}@adobe.com
[3] Facebook Reality Labs {thabetak}@fb.com

Abstract. Understanding movies and their structural patterns is a crucial task in decoding the craft of video editing. While previous works have developed tools for general analysis, such as detecting characters or recognizing cinematography properties at the shot level, less effort has been devoted to understanding the most basic video edit, *the Cut*. This paper introduces the Cut type recognition task, which requires modeling multi-modal information. To ignite research in this new task, we construct a large-scale dataset called MovieCuts, which contains 173,967 video clips labeled with ten cut types defined by professionals in the movie industry. We benchmark a set of audio-visual approaches, including some dealing with the problem's multi-modal nature. Our best model achieves 47.7% mAP, which suggests that the task is challenging and that attaining highly accurate Cut type recognition is an open research problem. Advances in automatic Cut-type recognition can unleash new experiences in the video editing industry, such as movie analysis for education, video re-editing, virtual cinematography, machine-assisted trailer generation, machine-assisted video editing, among others. Our data and code are publicly available: https://github.com/PardoAlejo/MovieCuts.

Keywords: Video Editing, Cut-types, Recognition, Shot transition, Cinematography, Movie Understanding.

1 Introduction

Professionally edited movies use the film grammar [1] as a convention to tell visual stories. Through the lenses of the film grammar, a movie can be deconstructed into a hierarchical structure: a string of contiguous frames form a shot, a sequence of shots build a scene, and a series of scenes compose the movie. Typically, scenes portray events that happen in single locations using shots recorded with a multi-camera setup [28]. Like punctuation in the written grammar, careful transition between shots is also an important component of the film grammar. Indeed, shot transitions can be viewed as the most basic video editing device [30,52]. They create changes of perspective, highlight emotions, and help to advance stories [30,52]. Several types of *soft* shot transitions like wipes, or fades are

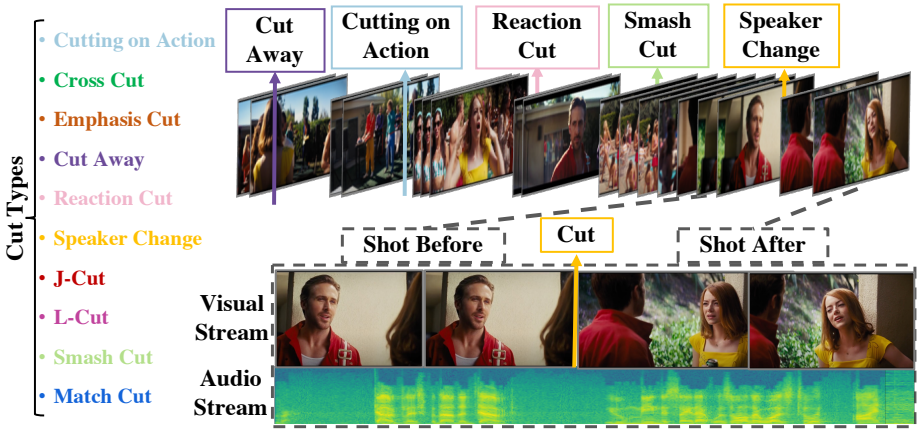


Fig. 1: **Cut Type Recognition Task.** Within a movie scene, most of the shot transitions use the *Straight Cut*, or Cut. Cuts are designed to preserve audio-visual continuity [45] across time, space, or story, and can be classified by their the semantic meaning [46]. This figure illustrates a scene with different cut types happening one after the other. Towards the end of the scene, there is a dialogue portrayed by cutting when the active speaker changes. Understanding this task requires audio-visual analysis. The visual stream helps detect the camera change to focus on one of the two actors. The audio stream exhibits a clear change of frequencies when the cut happens. Combining these two cues allows us to predict the cut type: *Speaker Change*.

commonly used between scenes. However, within a scene, the most used transition between shots [13] is the Cut, which simply joins two shots without any special effect.

Cuts in professionally edited movies are not random but rather have language, structure, and a taxonomy [1]. Each cut in a movie has a purpose and a specific meaning. Thus, to understand movie editing styles, one has to understand the cuts. Tsivian *et al.* introduced a platform called Cinemetrics to analyze movies by analysing their cut frequency [13]. While Cinemetrics is helpful in characterizing the rhythm and pace of cuts, analyzing and understanding the semantic meaning of these cuts remains a rather difficult task. In the computer vision community, recent works have tackled the problem of analyzing different cinematography components at the shot level [56,58,24,41] for automatic film analysis. However, only a few works, have focused on shot transitions for film analysis [57] and continuity editing [37,16]. We argue that automatically recognizing and understanding cut types would make an important step towards computationally characterize the principles of video editing, enabling new experiences for movie analysis for education, video re-editing, virtual cinematography, machine-assisted trailer generation, and machine-assisted video editing. We showcase one example of the latter in Section 4.5.

Figure 1 illustrates the cut type recognition task introduced in this work. ***A Cut is composed of two adjacent shots and the transition between them.*** Cuts are not only made of frames but also of their time-aligned sound stream. In many situations, sounds and speech drive the cut and shape its meaning. Our goal is then to recognize the cut type by analyzing the clip's audio-visual information across shots. Multiple research challenges emerge from this new multi-shot video understanding task. First, there is a need for a high-level understanding of visual and audio relationships over time to identify the intended cut type. To identify Speaker Change in Figure 1, one needs a detailed audio-visual inspection to associate the sounds before and after the cut to the corresponding actor's voice. Although it sounds trivial, small changes in the signal can change the cut type. For instance, if the speakers are in different locations, the cut type would no longer be Speaker Change but rather a Cross Cut (see Figure 2). If the active speaker does not change after the cut, the cut type would be Reaction Cut instead. Thus, it is essential to understand the fine-grained details of both signals. We argue that these challenges can promote the development of new techniques to address the multi-modal and multi-shot nature of the problem.

Understanding the audio-visual properties of movies has a long-standing track of interest [32,17,42,53]. The community has developed methods to recognize characters and speaker [36,15,9], events and actions [33,17,21], story-lines [2,53,24], shot-level cinematography properties such as shot-scale and camera motion [40,11], and mine shot-sequencing patterns [58,57,52]. While these approaches have set an initial framework for understanding editing in movies, there is still a lack of automated tools that understand the most basic and used editing technique, the Cut.

This work aims to study and bootstrap research in Cut type recognition. To do so, we introduce MovieCuts, a new large-scale dataset with manually curated Cut type annotations. Our new dataset contains $173,967$ clips (with cuts) labeled with ten different cut categories taken from the literature [5,51] and movie industry [50]. We hired professional and qualified annotators to label the cut type categories. MovieCuts offers the opportunity to benchmark core research tasks such as multi-modal analysis, long-tailed distribution learning, and multi-label classification. Furthermore, the study of this task, might benefit other areas like machine-assited video editing. While we observe improvements by leveraging recent techniques for audio-visual blending [55], there is ample room for improvement, and the task remains an open research problem.

Contributions. Our contributions are threefold: **(1)** We introduce the cut type recognition task. To the best of our knowledge, our work is the first to address and formalize the task from a machine learning perspective. **(2)** We collect a large-scale dataset containing qualified human annotations that verify the presence of different cut types. We do an extensive analysis of the dataset to highlight its properties and the challenges it presents. We call this dataset MovieCuts (Section 3). **(3)** We implement multiple audio-visual baselines and establish a benchmark in cut type recognition (Section 4).

2 Related Work

The Anatomy of a Movie Scene. Scenes are key building blocks for storytelling in film. They are built from a sequence of shots to depict an event, action, or element of film narration. Extensive literature in film studies has analyzed and characterized the structure of a scene. It includes (among others) the properties and categorization of shots and, to the interest of our work, the type of shot transitions [1,10,35]. There are four basic shot transitions: the wipe, the fade, the dissolve, and the cut. Each one of these four transitions has its purpose and appropriate usage. For instance, soft transitions like wipe, fade, and dissolve, are commonly used to transition between scenes and evoke a passage of time or change in location. Our work studies the cut, the instantaneous (hard) change from one shot to another, which is arguably the most frequently used.

Film theory has developed multiple taxonomies to organize the types of cuts [1,48,5]. Case in point, Thompson and Bowen [48] divide the types of cuts (or edits) into five different categories: action edit, screen position edit, form edit, concept edit, and combined edit. While such categorization provides a high-level grouping, it is too coarse. The categorization centers around the emotional aspects of the edit rather than the audio-visual properties of the cut. Film courses [51] and practitioners [50] have also developed a taxonomy of cut types. These tend to be more specific and closely describe the audio-visual properties of the shot pair forming the cut. We choose our list of Cut types based on the existing literature and narrow it down to categories that video editors recognize in their daily routine.

Edited Content in Video Understanding. Edited video content such as movies has been a rich source of data for general video understanding. Such video sources contain various human actions, objects, and situations occurring in people's daily life. In the early stages of action recognition, the Hollywood Human Actions (HOHA) [32] and HMDB51 [31], introduced human action recognition benchmarks using short clips from Hollywood movies. Another group of works used a limited number of films to train, and test methods for character recognition [43], human action localization [14], event localization [33], and spatio-temporal action and character localization [4]. With the development of deep-learning techniques and the need for large-scale datasets to train deep models, Gu *et al.* proposed the AVA dataset [17]. AVA is a large-scale dataset with spatio-temporal annotations, actors, and actions, whose primary data sources are movies and TV shows. Furthermore, other works have focused on action, and event recognition across shots [33,21]. Finally, Pavlakos *et al.* leverage information across shots from TV shows to do human mesh reconstruction [38]. Instead of leveraging movie data to learn representations for traditional tasks, we propose a new task to analyze movie cut types automatically.

Stories, Plots, and Cinematography. Movies and TV shows are rich in complexity and content, which makes their analysis and understanding a challenging task. Movies are a natural multi-modal source of data, with audio, video, and even transcripts being often available. Several works in the literature focus on the task of understanding movie content. Recent works have addressed the task

of movie trailer creation [27,20,62,44], TV show summarization [6,7], and automated video editing [37]. Moreover, Vicol *et al.* proposed MovieGraphs [53], a dataset that uses movies to analyze human-centric situations. Rohrbach *et al.* presented a Movie Description dataset [42], which contains audio narratives and movie scripts aligned to the movies' full-length. Using this dataset, a Large Scale Movie Description Challenge (LSMDC) has hosted competitions for a variety of tasks, including Movie Fill-In-The-Blank [34], and movie Q&A [47], among others. Like LSMDC, MovieNet [24] and Condensed Movies [2] are big projects that contain several tasks, data, and annotations related to movie understanding. MovieNet includes works related to person re-identification [26,60,22,23], Movie Scene Temporal Segmentation [41], and trailer and synopsis analysis [61,25]. All these works have shown that movies have rich information about human actions, including their specific challenges. However, only a few of them have focused on artistic aspects of movies, such as shot scales [40,11,3], shot taxonomy and classification [54], and their editing structure and cinematography [56,57,58]. These studies form the foundations to analyze movie editing properties but miss one of the most used techniques, the Cut. Understanding cuts is crucial for decoding the grammar of the film language. Our work represents a step towards that goal.

3 The MovieCuts Dataset

3.1 Building MovieCuts

Cut Type Categories. Our goal is to find a set of cut type categories often used in movie editing. Although there exists literature in the grammar of film language [1,46] and the taxonomy of shot types [54,11,3], there is no gold-standard categorization of cuts. As mentioned earlier in the related work, there exist categorization of cut types [48] but it focuses on the emotional aspects of the cuts rather than the audio-visual properties of the shots composing the cut. We gathered an initial taxonomy (17 cut types) from film-making courses (*e.g.* [51]), textbooks [5], and blogs [50]. We then hired ten different editors to validate the taxonomy. All the editors studied film-making, two have been nominated for the Emmys, and most have over 10 years of experience. Some of the original categories were duplicated and some of them were challenging to mine from movies. Our final taxonomy includes 10 categories. Figure 2 illustrates each cut type along with their visual and audio signals:

1. **Cutting on Action:** Cutting from one shot to another while the subject is still in motion.
2. **Cross Cut:** Cutting back and forth within locations.
3. **Emphasis Cut:** Cut from wide to close within the same shot, or the other way around.
4. **Cut Away:** Cutting into an insert shot of something and then back.
5. **Reaction Cut:** A cut to the reaction of a subject (facial expression or single word) to the comments / actions of other actors, or a cut after the reaction.
6. **Speaker Change:** A cut that changes the shot to the current speaker.

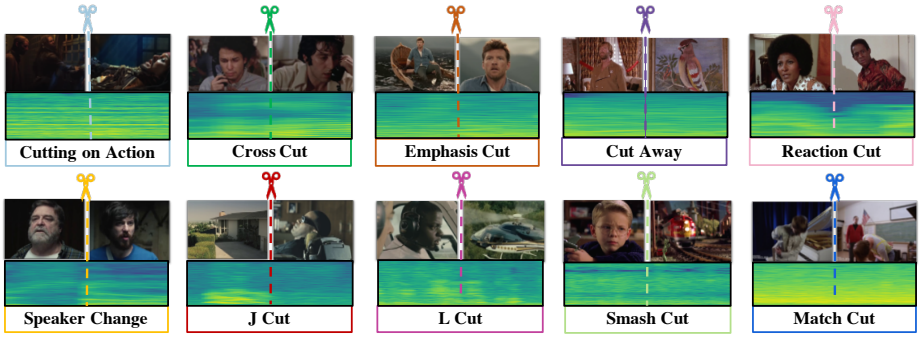


Fig. 2: **MovieCuts Dataset.** MovieCuts contains 173,967 video clips labeled with 10 different Cut types. Each sample in the dataset is composed of two shots (with a cut) and their audio spectogram. Our cuts are grouped into two major categories, visual (top) and audio-visual (bottom) driven.

7. **J Cut:** The audio of the next shot begins before you can see it. You hear what is going on before you actually see what is going on.
8. **L Cut:** The audio of the current shot carries over to the next shot.
9. **Smash Cut:** Abrupt cut from one shot to another for aesthetic, narrative, or emotional purpose.
10. **Match Cut:** Cut from one shot to another by matching a concept, an action or a composition of both.

Video Collection and Processing. We need professionally edited videos containing diverse types of cuts. Movies are a perfect source to gather and collect such data. As pointed out by Bain *et al.* [2], there are online video channels[4] that distribute *individual* movie scenes, thus facilitating access to movie data for research. We downloaded 9,363 scenes, which come from 1,986 movies. However, these movie scenes come untrimmed, and further processing is required to obtain cuts from them. We automatically detect all the cuts in the dataset with a highly accurate shot boundary detector [18] (97.3% precision and 98.5% recall), which yields a total of 195,000 candidate cuts for annotation.

Human Annotations and Verification. Our goal at this stage is to collect expert annotations for 195,000 candidate cuts. To do so, we hired Hive AI to run our annotation campaign. We choose them given their experience in labeling data for the entertainment industry. Annotators did not necessarily have a film-making background, but they had to pass a qualification test to participate in the labeling process. At least three annotators reviewed each cut/label candidate pair, and only the annotations with more than two votes were kept. The annotators also have the option to discard cuts due to: (i) errors in the shot

[4] MovieClips YouTube Channel is the source of scenes in MovieCuts.

boundary detector, and (ii) the cut not showing any of the categories in our taxonomy. We also build a handbook in partnership with professional editors to include several examples per class and guidelines on addressing edge cases. We discarded 21,033 cuts, which left us with a total of 173,967 cuts to form our dataset. From the 21,073 discarded clips, we found that 12,090 did not have enough consensus, which leads to an inter-annotator agreement of 93.8%. Given that inter-annotator agreement does not account for missing labels, five professional editors labelled a small subset of two thousand cuts and created a high-consensus ground truth. We found that our annotations exhibited a 90.5% precision and 88.2% recall when contrasted with such a gold standard.

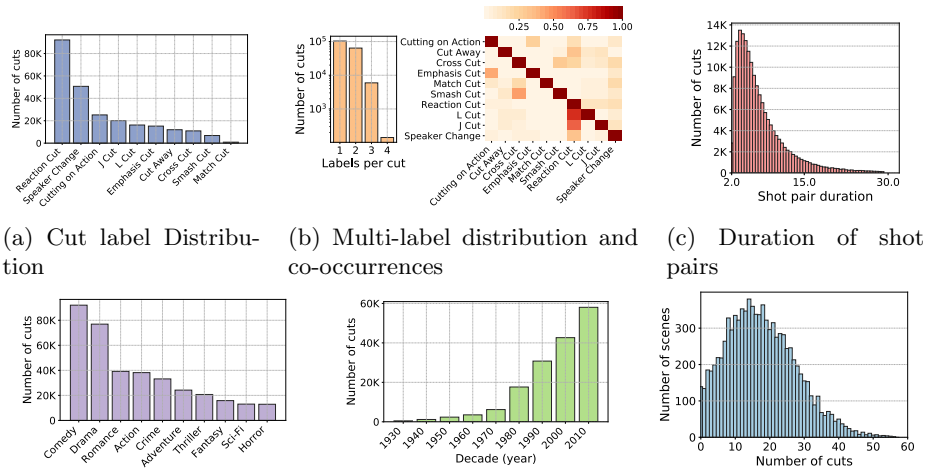
3.2 MovieCuts Statistics

Cut label distribution. Figure 3a shows the distribution of cut types in MovieCuts. The distribution is long-tailed, which may reflect the editors' preferences for certain types of cuts. It is not a surprise that *Reaction Cut* is the most abundant label given that emotion and human reactions play a central role in storytelling. Beyond human emotion, dialogue and human actions are additional key components to advance movie storylines. We observe this in the label distribution, where *Speaker Change* and *Cutting on Action* are the second and third most abundant categories in the dataset. While classes such as *Smash Cut* and *Match Cut* emerge scarcely in the dataset, it is still important to recognize these types of cuts, which can be considered the most creative ones. We also show the distribution of cut types per movie genre in the *supplementary material*.

Multi-label distribution and co-occurrences. We plot in Figure 3b the distribution of labels per cut and the co-occurrence matrix. On one hand, we observe that a significant number of cuts contain more than one label. On the other, we observe that certain pair of classes co-occur more often, *e.g. Reaction Cut / L Cut*. The multi-label properties of the dataset suggest that video editors compose and combine cut types quite often.

Duration of shot pairs. We study the duration of the shot pairs that surround (and form) the cuts. Figure 3c shows the distribution of such shot pair duration. The most typical length is about 3.5 seconds. Moreover, we observe that the length of shot pairs ranges from 2 seconds to more than 30 seconds. In Section 4, we study the effect of sampling different context around the cut.

Cut genre, year of production, and cuts per scene. Figures 3d, 3e, 3f show statistics about the productions from where the cuts are sampled. First, we observe that the cuts are sampled across a diverse set of genres, with Comedy being the most frequent one. Second, we sourced the cuts from old and contemporary movie scenes. While many cuts come from the last decade, we also scouted cuts from movie scenes from the 1930's. Finally, we observe that the number of cuts per scene roughly follows a normal distribution with a mean of 15 cuts per scene. Interestingly, few movie scenes have a single cut, while others may contain more than 60 cuts. These statistics highlight the editing diversity in MovieCuts.



(a) Cut label Distribution

(b) Multi-label distribution and co-occurrences

(c) Duration of shot pairs

(d) Cuts-per-scene by genre

(e) Cuts-per-scene by production year

(f) Cuts-per-scene distribution

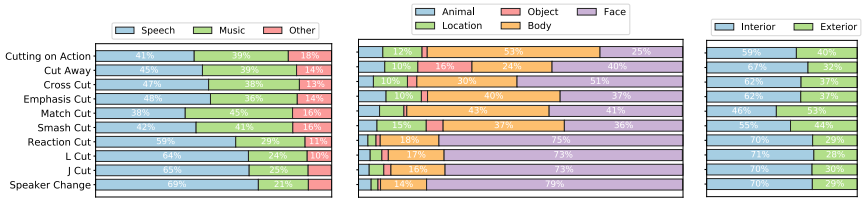
Fig. 3: **MovieCuts statistics.** Figure 3a shows the number of instances per cut type. Labels follow a long-tail distribution. Figure 3b indicates that a large number of instances contain more than a single cut type. Moreover, certain pairs of cut types co-occur more often. Figure 3c plots the distribution of lengths (in seconds) of all the dataset instances. Figure 3d summarizes the production properties of the movie scenes and cuts used in our study. Figure 3e shows the distribution based on year of production. Finally, Figure 3f shows the distribution of number of cuts per scene.

3.3 MovieCuts Attributes

Sound attributes. We leverage an off-the-shelf audio classifier [12] to annotate the sound events in the dataset. Figure 4a summarizes the distribution of three super-groups of sound events: Speech, Music, and Other. Dialogue related cuts such as *Speaker Change* and *J Cut* contain a large amount of speech. Contrarily, visual-driven cuts *e.g. Match Cut* and *Smash Cut* hold a larger number of varied sounds and background music. These attributes suggest that, while analyzing speech is crucial for recognizing cut types, it is also beneficial to model music and general sounds.

Subject attributes. We build a zero-shot classifier using CLIP [39], a neural network trained on 400M image-text pairs, to tag the subjects present in our dataset samples (4b).

Interestingly, dialogue and emotion-driven cuts (*e.g. Reaction Cut*) contain many face tags, which can be interpreted as humans framed in medium-to-close-up shots. Contrarily, Body is the most common attribute in the *Cutting on Action* class, which suggests editors often opt for framing humans in long shots when actions are occurring.



(a) Sound (b) Subject (c) Location

Fig. 4: **MovieCuts attributes.** MovieCuts contains diverse sounds (4a), subjects (4b), and locations (4c). Some sounds co-occur more often in particular cut types. For instance, Speech is the predominant sound for dialogue related cut types such as Speaker Change or J Cut. Similarly, there exists correlation between cut types and the subjects depicted in the movie clip.

Location Attributes. We reuse CLIP [39] to construct a zero-shot classifier of locations on our dataset. Figure 4c summarizes the distribution of locations (Interior/Exterior) per cut type. On one hand, we observe that most cut types contain instances shot in Interior locations 60%-70% of the time. On the other hand, *Match Cuts* reverse this trend with the majority (53%) of cuts shot in Exterior places. The obtained distribution suggests that stories in movies (as in real-life) develop (more often) in indoor places.

4 Experiments

4.1 Audio-visual Baseline

Our base architecture is shown in Figure 5. Similar to [55], it takes as input the audio signal and a set of frames (clip), which are then processed by a late-fusion multi-modal CNN. We use a visual encoder and an audio encoder to extract audio and visual features per clip. Then, we form an audio-visual feature by concatenating them. Finally, a Multi Layer Perceptron (MLP) computes the final predictions on the audio-visual features. We optimize a binary cross-entropy (BCE) loss $\mathcal{L}$ per modality and for their combination in a one-vs-all manner to deal with the problem's multi-label nature. Our loss is summarized as:

$$\text{loss} = \omega_a \mathcal{L}\left(\hat{y}_a, y\right) + \omega_v \mathcal{L}\left(\hat{y}_v, y\right) + \omega_{av} \mathcal{L}\left(\hat{y}_{av}, y\right), \tag{1}$$

where ω_a, ω_v, and ω_{av} are the weights for the audio, visual, and audio-visual losses, respectively. Using this architecture, we propose several baselines:

MLP classifier. We use the backbone as a feature extractor for each stream and train an MLP to predict on top of them and their concatenation.
Encoder fine-tuning. We train the whole backbone starting from Kinetics-400 [29] weights for the visual stream and from VGGSound [12] weights for the audio.

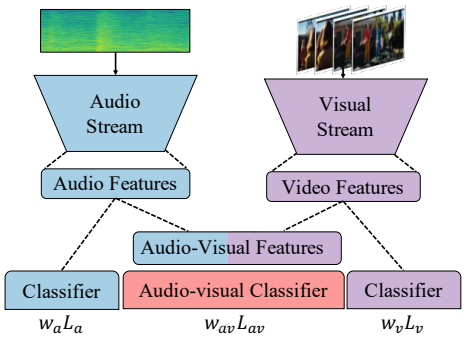


Fig. 5: **Audio-visual pipeline.** A late-fusion multi-modal network processes audio and visual streams. We train both networks jointly using audio loss L_a, visual loss L_v, and audio-visual loss L_{av}, weighted with w_a, w_v, and w_{av}, respectively.

Modality variants. We train our model using the different modalities: audio only, visual only, and audio-visual. For audio-visual, we combine the losses in a naive way giving each one of them the same weight, *i.e.* $\omega_a = \omega_v = \omega_{av}$.

Modality blending. To combine losses from multiple modalities in a more effective way, Wang *et al.* [55] proposed *Gradient Blending* (GB), a strategy to compute the weight of each modality loss at training time. We use the offline GB algorithm to calculate ω_a, ω_v, and ω_{av}. For further details, refer to Algorithms 1 and 2 of the paper [55].

4.2 Experimental Setup

Dataset summary. We divide our dataset into training, validation, and test sets by using 70%, 10%, and 20% percent of the data, respectively. Thus, we use 121,423 clips for training, 17,505 clips for validation, and 35,039 clips for testing. We make sure that the sets are *i.i.d.*w.r.t. the movie genre, we show distributions per genre for each split in the supplementary material. We report all experiments on the validation set unless otherwise mentioned.

Metrics. Following [59], we choose the mean Average Precision (mAP) across all classes and per-class AP to summarize and compare baseline performances. This metric helps to deal with MovieCuts' multi-label nature. We also report Precision-Recall curves in supplementary material.

Implementation details. For all experiments, we use ResNet-18 [19] as the backbone for both visual and audio streams. **For the audio stream**, we use a ResNet with 2D convolutions pre-trained on VGGSound [12]. This backbone takes as input a spectrogram of the audio signal and processes it as an image. To compute the spectogram we take the audios of each pair of clips, and apply consecutive Fourier Transforms with 512 points windows with 353 overlapping points between them. If the audio is longer than 10 seconds we trim it to 10 seconds only. **For the visual stream**, we use a ResNet-(2+1)D [49] pre-trained on Kinetics-400 [29]. We sample 16 frames from a window centered around the cut as the input to the network Using single streams, *i.e.* only audio or only visual, we use the features after the average pooling followed by an MLP composed

of a 512×128 Fully-Connected (FC) layer followed by a ReLU and a $128 \times N$ FC-layer, where N is the number of classes ($N = 10$). Using two streams, we concatenate the features after the first FC-layer of the MLP to obtain an audio-visual feature per clip of size $128 \times 2 = 256$. Then, we pass it through a second FC-layer of size $256 \times N$ to compute the predictions. We train using SGD with momentum 0.9 and weight decay of 10^{-4}. We also use a linear warm-up for the first epoch. We train for 8 epochs with an initial learning rate of 3×10^{-2}, which decays by a factor of 10 after 4 epochs. We use an effective batch-size of 112 and train on one NVIDIA A100 GPU.

4.3 Results and Analysis

As described in Section 4.1, we benchmark the MovieCuts dataset using several combinations of modalities. Results are reported in Table 1.

Linear Classifier *vs.* Fine-tune: We evaluate the performance of using frozen *vs.* fine-tuned features. As one might expect, the fine-tuning of the backbone shows consistent improvement over all the classes regardless of the modality used. For instance, the Audio-Visual backbone performance increases from 30.82% to 46.57%. These results validate the value of the dataset for improving the audio and visual representations encoded by the backbones.

Modality Variants: Consistently across training strategies, we observe a common pattern in the results: the visual modality performs better (43.98%) at the task than its audio counterpart (27.24%). Nonetheless, combining both modalities still provides enhanced results for several classes and the overall mAP (46.57%). We observe that cuts driven mainly by visual cues, such as Cutting on Action, Cut Away, and Cross Cut, do not improve their performance when audio is added. However, the rest of the classes improve when using both modalities. In particular, L Cut, J Cut, and Speaker Change improve drastically, since these types of cuts are naturally driven by audio-visual cues.

Gradient Blending: The second-to-last row in Table 1 shows the results of using the three modalities combined with the GB weights $\omega_a = 0.08$, $\omega_v = 0.57$ and $\omega_{av} = 0.35$. GB performs slightly better (47.43%) than combining the losses naively (46.57%), where $\omega_a = \omega_v = \omega_{av}$.

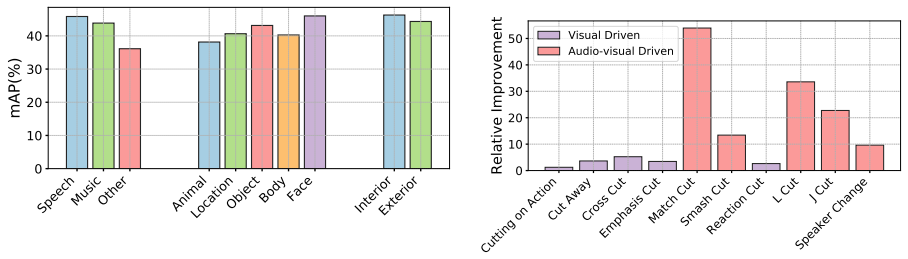
Scaled Gradient Blending: By experimenting with the Gradient Blending weights, we found that scaling them all by a constant factor can help. We empirically found that scaling the weights by a factor of 3 ($\omega_a = 1.31$, $\omega_v = 4.95$ and $\omega_{av} = 2.74$) improves the results to 47.91% mAP.

Frame Sampling: In addition to these experiments, we explore how to pick the frames to feed into the visual network. For all the previous experiments and as mentioned, we use *Fixed Sampling* by sampling frames from a window centered around the cut. However, this is not the only strategy to sample frames. We explore two other strategies: *Uniform Sampling* that takes sample frames from a uniform distribution across the two shots forming the cut, and *Gaussian Sampling*, which samples the frames from a Gaussian centered around the cut. We fit both audios up to 10 seconds into the audio stream.

	Model	mAP	CA	CW	CC	EC	MC	SC	RC	LC	JC	SC
Linear	Audio(A)	23.9	36.7	14.8	11.8	14.6	1.5	10.7	65.3	15.3	18.4	50.0
	Visual(V)	28.8	53.8	36.3	16.9	19.4	1.1	13.3	69.7	12.9	16.2	48.0
	AV	30.8	55.5	32.8	16.0	20.3	1.7	13.2	73.7	17.4	21.6	56.0
Fine-tune	Audio	27.2	42.6	19.0	14.6	15.8	1.5	12.9	69.4	18.5	21.3	56.7
	Visual	44.0	64.8	60.8	33.2	30.7	1.5	21.5	81.2	33.7	42.0	70.3
	AV	46.6	65.2	62.5	31.1	30.5	2.0	22.3	82.9	43.3	50.0	75.1
	AV+GB	47.4	64.8	62.4	32.5	31.6	1.8	23.8	83.1	45.6	51.0	77.4
	AV+SGB	**47.9**	**65.6**	**63.0**	**34.9**	**31.8**	**2.3**	**24.3**	**83.3**	**45.0**	**51.6**	**77.1**

Table 1: **Baseline comparison on MovieCuts**. We show the results of our different baselines using visual, audio, and audio-visual modalities. The last two rows use both modalities combined with Gradient Blending (GB) [55] and Scaled Gradient Blending (SGB). All the reported results are % AP. We observe three key findings. (1) Fine-tuning on MovieCuts provides clear benefits over the linear classifier trained on frozen features. (2) Audio-visual information boosts the performance of the visual only stream. (3) Gradient Blending provides further performance gains by an optimal combination of both modalities. Showing classes: Cutting on Action (**CA**), Cut Away (**CW**), Cross Cut (**CC**), Emphasis Cut (**EC**), Match Cut (**MC**), Smash Cut (**SC**), Reaction Cut (**RC**), L Cut (**LC**), J Cut (**JC**), Speaker Chance (**SC**).

Fixed Sampling gives the best results with 47.91% mAP, followed by *Gaussian Sampling* with 47.44%, and *Uniform Sampling* gives the lowest mAP among them with 47.17%. These results suggest that the most critical information lays around the cut. We hypothesize that the model is not good enough at handling context, architectures better at handling sequential inputs may benefit from the context of the *Gaussian* or *Uniform* sampling.



(a) Performance breakdown per attribute. (b) Audio-visual improvements per type.

Fig. 6: **Performance breakdown.** Here, we showcase a detailed performance analysis. Figure 6a shows the performance breakdown according to attributes of MovieCuts, such as type of sound, subjects, locations, duration per clip, and production year. Figure 6b shows the performance gain of the audio-visual model versus the visual-only model. For this analysis, we group the cut classes into visual-driven and audio-visual driven.

Test set results. After obtaining the best-performing model on the validation set, we evaluate this model on the test set. We obtain 47.70% mAP, which is slightly lower than the results on the validation set 47.91% mAP. For the full test set results, and experiments using Distribution-Balanced loss [59], refer to the **supplementary material**.

4.4 Performance Breakdown

Attributes and dataset characteristics. Figure 6a summarizes the performance of our best audio-visual model from Table 1 for different attributes and dataset characteristics. In most cases, the model exhibits robust performance across attributes. The largest performance gap is observed between Speech and Other sounds. We associate this result with the fact that cuts with complex audio editing, *e.g.* those that include sound effects, often employ abstract editing such as Smash Cuts and Match Cuts, which are harder for the model to recognize. We also observe that the model is better at classifying cuts when there are faces in the scene, which aligns with the fact that it was trained on movies, which are mainly human-centered. These findings showcase the multi-modal nature of MovieCuts. Thus, the results can be improved by studying better audio-visual backbones.

Audio-visual improvements per cut type. Figure 6b shows the relative improvement of the audio-visual model w.r.t. using the visual stream only. It highlights whether the type of cut is driven by visual or audio-visual information. We observe that the audio-visual driven cuts generally benefit the most from training a joint audio-visual model. Match cuts show a relative 50% improvement when adding audio. These types of cuts use audiovisual concepts to match the two shots. The second-largest gains are for cuts related to dialogue and conversations (L cut, J cut). For instance, L Cuts improve by 30%; this class typically involves a person on screen talking in the first shot while only their voice is heard in the second shot. By encoding audio-visual information, the model disambiguates predictions that would otherwise be confused by the visual-only model. Finally, all classes show a relative improvement w.r.t. the visual baseline. This suggests that the GB [55] strategy allows the model to optimize modality weights. In the worst-case scenario, GB achieves slight improvements over the visual-only baseline. In short, we empirically demonstrate the importance of modeling audio-visual information to recognize cut types.

4.5 Machine-assisted Video Editing with MovieCuts

We argue that recognizing cut types can enable many applications in video editing. In this section, we leverage the knowledge of our Audio-Visual model to attempt automated video editing. Inspired by [37], we use EditStock[5] to gather raw footage of edited movies and perform video editing on them. Specifically, we use our model to create cuts (shot transitions) between two long sequences

[5] EditStock.com

of shots. Further details can be found in **supplementary material**. We measure qualitatively and quantitatively our model's editing by comparing it with different automated editing methods: (1) **Random baseline frame by frame RF:** Every frame, we perform a cut with a probability of 0.5. (2) **Random baseline snippet by snippet RS:** Similar to how the model is trained, every 16 frames snippet we cut with a probability of 0.5. This restriction allows each shot to be on screen for at least 16 frames. (3) **Biased Random BR:** Similar to (2), we cut every 16 snippets. However, this time we use the expected number of cuts prior. Thus, we ensure that the number of random cuts is the same as the ground truth. (4) **MovieCuts' AV model AV:** We use our audio-visual model's scores to score all possible cuts between the two sequences. Then, we use the top-k cuts, where k is given by the expected number of cuts. (5) **Human Editor GT:** From EditStock we collect the actual edited sequence edited by professional editors. We use these sequences as a reference for the quantitative study, and ground-truth for the qualitative evaluation.

In the qualitative evaluation we ask 63 humans to pick between our method and all the other methods. We observe that users picked our method (AV) over the human editor 38% of the times while the BR was picked 34% of the times – RF and RS were picked only 15.7% and 1.8%, respectively. Furthermore, for the quantitative results we use the human edit as ground-truth and measure Purity, Coverage, and F1 for each method. These metrics were implemented by [8] and measure the similarity between the segmentation of two different sequences. The results are consistent with the qualitative study and show that MovieCuts' edits have an F1 of 81% while BR, RS, and RF have only 77%, 63% 17%, respectively. A more in-depth analysis of this study can be found in **supplementary material**. This simple experiment suggests that the MovieCuts dataset allows the model to learn about video editing by learning cut-type recognition. Thus, we argue that further improvement in Cut-type recognition tasks can translate into advances in tasks related to machine-assisted video editing.

5 Conclusion

We introduced the cut-type recognition task in movies and started research in this new area by providing a new large-scale dataset, MovieCuts accompanied with a benchmark of multiple audio-visual baselines.. We collect $173,967$ annotations from qualified human workers. We analyze the dataset diversity and uniqueness by studying its properties and audio-visual attributes. We propose audio-visual baselines by using learning approaches that address the multi-modal nature of the problem. Although we established a strong research departure point, we hope that more research pushes the envelope of cut-type recognition by leveraging MovieCuts.

Acknowledgements. This work was supported by the King Abdullah University of Science and Technology (KAUST) Office of Sponsored Research through the Visual Computing Center (VCC) funding.

References

1. Arijon, D.: Grammar of the film language. Focal Press London (1976)
2. Bain, M., Nagrani, A., Brown, A., Zisserman, A.: Condensed movies: Story based retrieval with contextual embeddings (2020)
3. Benini, S., Svanera, M., Adami, N., Leonardi, R., Kovács, A.B.: Shot scale distribution in art films. Multimedia Tools and Applications **75**(23), 16499–16527 (2016)
4. Bojanowski, P., Bach, F., Laptev, I., Ponce, J., Schmid, C., Sivic, J.: Finding actors and actions in movies. In: Proceedings of the IEEE international conference on computer vision. pp. 2280–2287 (2013)
5. Bordwell, D., Thompson, K., Smith, J.: Film art: An introduction, vol. 7. McGraw-Hill New York (1993)
6. Bost, X., Gueye, S., Labatut, V., Larson, M., Linarès, G., Malinas, D., Roth, R.: Remembering winter was coming. Multimedia Tools and Applications **78**(24), 35373–35399 (2019)
7. Bost, X., Labatut, V., Linares, G.: Serial speakers: a dataset of tv series. arXiv preprint arXiv:2002.06923 (2020)
8. Bredin, H.: pyannote.metrics: a toolkit for reproducible evaluation, diagnostic, and error analysis of speaker diarization systems. In: Interspeech 2017, 18th Annual Conference of the International Speech Communication Association. Stockholm, Sweden (August 2017), http://pyannote.github.io/pyannote-metrics
9. Brown, A., Huh, J., Nagrani, A., Chung, J.S., Zisserman, A.: Playing a part: Speaker verification at the movies. arXiv preprint arXiv:2010.15716 (2020)
10. Burch, N.: Theory of film practice. Princeton University Press (2014)
11. Canini, L., Benini, S., Leonardi, R.: Classifying cinematographic shot types. Multimedia tools and applications **62**(1), 51–73 (2013)
12. Chen, H., Xie, W., Vedaldi, A., Zisserman, A.: Vggsound: A large-scale audio-visual dataset. In: International Conference on Acoustics, Speech, and Signal Processing (ICASSP) (2020)
13. Cutting, J.E.: The evolution of pace in popular movies. Cognitive research: principles and implications **1**(1), 30 (2016)
14. Duchenne, O., Laptev, I., Sivic, J., Bach, F., Ponce, J.: Automatic annotation of human actions in video. In: 2009 IEEE 12th International Conference on Computer Vision. pp. 1491–1498. IEEE (2009)
15. Everingham, M., Sivic, J., Zisserman, A.: Hello! my name is... buffy"–automatic naming of characters in tv video. In: BMVC. vol. 2, p. 6 (2006)
16. Galvane, Q., Ronfard, R., Lino, C., Christie, M.: Continuity editing for 3d animation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 29 (2015)
17. Gu, C., Sun, C., Ross, D.A., Vondrick, C., Pantofaru, C., Li, Y., Vijayanarasimhan, S., Toderici, G., Ricco, S., Sukthankar, R., et al.: Ava: A video dataset of spatio-temporally localized atomic visual actions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6047–6056 (2018)
18. Gygli, M.: Ridiculously fast shot boundary detection with fully convolutional neural networks. In: 2018 International Conference on Content-Based Multimedia Indexing (CBMI). pp. 1–4. IEEE (2018)
19. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)

20. Hesham, M., Hani, B., Fouad, N., Amer, E.: Smart trailer: Automatic generation of movie trailer using only subtitles. In: 2018 First International Workshop on Deep and Representation Learning (IWDRL). pp. 26–30. IEEE (2018)
21. Hoai, M., Zisserman, A.: Thread-safe: Towards recognizing human actions across shot boundaries. In: Asian Conference on Computer Vision. pp. 222–237. Springer (2014)
22. Huang, Q., Liu, W., Lin, D.: Person search in videos with one portrait through visual and temporal links. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 425–441 (2018)
23. Huang, Q., Xiong, Y., Lin, D.: Unifying identification and context learning for person recognition. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
24. Huang, Q., Xiong, Y., Rao, A., Wang, J., Lin, D.: Movienet: A holistic dataset for movie understanding. In: The European Conference on Computer Vision (ECCV) (2020)
25. Huang, Q., Xiong, Y., Xiong, Y., Zhang, Y., Lin, D.: From trailers to storylines: An efficient way to learn from movies. arXiv preprint arXiv:1806.05341 (2018)
26. Huang, Q., Yang, L., Huang, H., Wu, T., Lin, D.: Caption-supervised face recognition: Training a state-of-the-art face model without manual annotation. In: The European Conference on Computer Vision (ECCV) (2020)
27. Irie, G., Satou, T., Kojima, A., Yamasaki, T., Aizawa, K.: Automatic trailer generation. In: Proceedings of the 18th ACM international conference on Multimedia. pp. 839–842 (2010)
28. Katz, E., Klein, F.: The film encyclopedia. Collins (2005)
29. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al.: The kinetics human action video dataset. arXiv preprint arXiv:1705.06950 (2017)
30. Kozlovic, A.K.: Anatomy of film. Kinema: A Journal for Film and Audiovisual Media (2007)
31. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T.: Hmdb: a large video database for human motion recognition. In: 2011 International conference on computer vision. pp. 2556–2563. IEEE (2011)
32. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–8. IEEE (2008)
33. Liu, X., Hu, Y., Bai, S., Ding, F., Bai, X., Torr, P.H.: Multi-shot temporal event localization: a benchmark. arXiv preprint arXiv:2012.09434 (2020)
34. Maharaj, T., Ballas, N., Rohrbach, A., Courville, A., Pal, C.: A dataset and exploration of models for understanding video data through fill-in-the-blank question-answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6884–6893 (2017)
35. Murch, W.: In the Blink of an Eye, vol. 995. Silman-James Press Los Angeles (2001)
36. Nagrani, A., Zisserman, A.: From benedict cumberbatch to sherlock holmes: Character identification in tv series without a script. arXiv preprint arXiv:1801.10442 (2018)
37. Pardo, A., Caba, F., Alcazar, J.L., Thabet, A.K., Ghanem, B.: Learning to cut by watching movies. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 6858–6868 (October 2021)
38. Pavlakos, G., Malik, J., Kanazawa, A.: Human mesh recovery from multiple shots. arXiv preprint arXiv:2012.09843 (2020)

39. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. arXiv preprint arXiv:2103.00020 (2021)
40. Rao, A., Wang, J., Xu, L., Jiang, X., Huang, Q., Zhou, B., Lin, D.: A unified framework for shot type classification based on subject centric lens. In: The European Conference on Computer Vision (ECCV) (2020)
41. Rao, A., Xu, L., Xiong, Y., Xu, G., Huang, Q., Zhou, B., Lin, D.: A local-to-global approach to multi-modal movie scene segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10146–10155 (2020)
42. Rohrbach, A., Torabi, A., Rohrbach, M., Tandon, N., Pal, C., Larochelle, H., Courville, A., Schiele, B.: Movie description. International Journal of Computer Vision **123**(1), 94–120 (2017)
43. Sivic, J., Everingham, M., Zisserman, A.: "who are you?"-learning person specific classifiers from video. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1145–1152. IEEE (2009)
44. Smith, J.R., Joshi, D., Huet, B., Hsu, W., Cota, J.: Harnessing ai for augmenting creativity: Application to movie trailer creation. In: Proceedings of the 25th ACM international conference on Multimedia. pp. 1799–1808 (2017)
45. Smith, T.J., Henderson, J.M.: Edit blindness: The relationship between attention and global change blindness in dynamic scenes. Journal of Eye Movement Research **2**(2) (2008)
46. Smith, T.J., Levin, D., Cutting, J.E.: A window on reality: Perceiving edited moving images. Current Directions in Psychological Science **21**(2), 107–113 (2012)
47. Tapaswi, M., Zhu, Y., Stiefelhagen, R., Torralba, A., Urtasun, R., Fidler, S.: MovieQA: Understanding Stories in Movies through Question-Answering. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
48. Thompson, R., Bowen, C.J.: Grammar of the Edit, vol. 13. Taylor & Francis (2009)
49. Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M.: A closer look at spatiotemporal convolutions for action recognition. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 6450–6459 (2018)
50. Http://www.cuvideoedit.com/types-of-edits.php
51. Https://filmanalysis.yale.edu/editing/#transitions
52. Tsivian, Y.: Cinemetrics, part of the humanities' cyberinfrastructure (2009)
53. Vicol, P., Tapaswi, M., Castrejon, L., Fidler, S.: Moviegraphs: Towards understanding human-centric situations from videos. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
54. Wang, H.L., Cheong, L.F.: Taxonomy of directing semantics for film shot classification. IEEE transactions on circuits and systems for video technology **19**(10), 1529–1542 (2009)
55. Wang, W., Tran, D., Feiszli, M.: What makes training multi-modal classification networks hard? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12695–12705 (2020)
56. Wu, H.Y., Christie, M.: Analysing cinematography with embedded constrained patterns. In: WICED-Eurographics Workshop on Intelligent Cinematography and Editing (2016)
57. Wu, H.Y., Galvane, Q., Lino, C., Christie, M.: Analyzing elements of style in annotated film clips. In: WICED 2017-Eurographics Workshop on Intelligent Cinematography and Editing. pp. 29–35. The Eurographics Association (2017)

58. Wu, H.Y., Palù, F., Ranon, R., Christie, M.: Thinking like a director: Film editing patterns for virtual cinematographic storytelling. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) **14**(4), 1–22 (2018)
59. Wu, T., Huang, Q., Liu, Z., Wang, Y., Lin, D.: Distribution-balanced loss for multi-label classification in long-tailed datasets. In: European Conference on Computer Vision. pp. 162–178. Springer (2020)
60. Xia, J., Rao, A., Xu, L., Huang, Q., Wen, J., Lin, D.: Online multi-modal person search in videos. In: The European Conference on Computer Vision (ECCV) (2020)
61. Xiong, Y., Huang, Q., Guo, L., Zhou, H., Zhou, B., Lin, D.: A graph-based framework to bridge movies and synopses. In: The IEEE International Conference on Computer Vision (ICCV) (October 2019)
62. Xu, H., Zhen, Y., Zha, H.: Trailer generation via a point process-based visual attractiveness model. In: Twenty-Fourth International Joint Conference on Artificial Intelligence (2015)