MovieNet: A Holistic Dataset for Movie Understanding

Qingqiu Huang*, Yu Xiong*, Anyi Rao, Jiaze Wang, and Dahua Lin

CUHK-SenseTime Joint Lab, The Chinese University of Hong Kong
{hq016, xy017, ra018, dhlin}@ie.cuhk.edu.hk
jzwang@link.cuhk.edu.hk

Abstract. Recent years have seen remarkable advances in visual understanding. However, how to understand a story-based long video with artistic styles, *e.g.* movie, remains challenging. In this paper, we introduce MovieNet – a holistic dataset for movie understanding. MovieNet contains 1,100 movies with a large amount of multi-modal data, *e.g.* trailers, photos, plot descriptions, *etc.*. Besides, different aspects of manual annotations are provided in MovieNet, including $1.1M$ characters with bounding boxes and identities, $42K$ scene boundaries, $2.5K$ aligned description sentences, $65K$ tags of place and action, and $92K$ tags of cinematic style. To the best of our knowledge, MovieNet is the largest dataset with richest annotations for comprehensive movie understanding. Based on MovieNet, we set up several benchmarks for movie understanding from different angles. Extensive experiments are executed on these benchmarks to show the immeasurable value of MovieNet and the gap of current approaches towards comprehensive movie understanding. We believe that such a holistic dataset would promote the researches on story-based long video understanding and beyond. MovieNet will be published in compliance with regulations at https://movienet.github.io.

1 Introduction

"You jump, I jump, right?" When Rose gives up the lifeboat and exclaims to Jack, we are all deeply touched by the beautiful moving love story told by the movie *Titanic*. As the saying goes, "Movies dazzle us, entertain us, educate us, and delight us". Movie, where characters would face various situations and perform various behaviors in various scenarios, is a reflection of our real world. It teaches us a lot such as the stories took place in the past, the culture and custom of a country or a place, the reaction and interaction of humans in different situations, *etc.*. Therefore, to understand movies is to understand our world.

It goes not only for human, but also for an artificial intelligence system. We believe that movie understanding is a good arena for high-level machine intelligence, considering its high complexity and close relation to the real world.

* Equal contribution

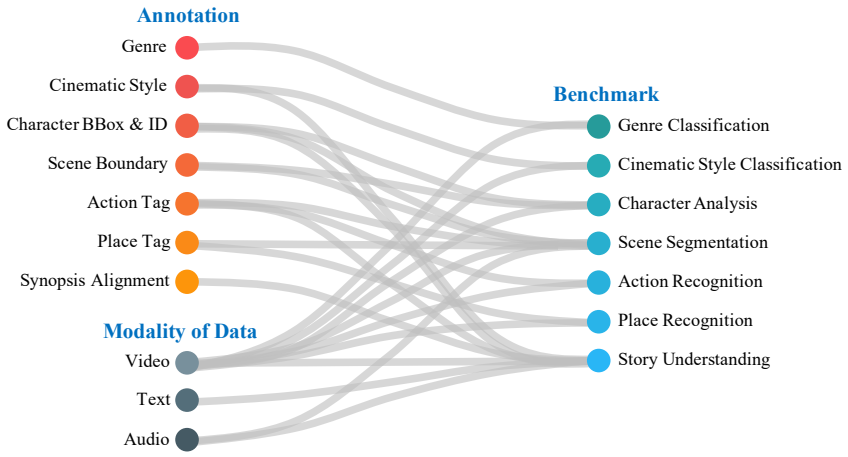


Fig. 1: The data, annotation, benchmark and their relations in MovieNet, which together build a holistic dataset for comprehensive movie understanding.

What's more, compared to web images [15] and short videos [7], the hundreds of thousands of movies in history containing rich content and multi-modal information become better nutrition for the data-hungry deep models.

Motivated by the insight above, we build a holistic dataset for movie understanding named *MovieNet* in this paper. As shown in Fig. 1, MovieNet comprises three important aspects, namely *data*, *annotation*, and *benchmark*.

First of all, MovieNet contains a large volume of data in multiple modalities, including movies, trailers, photos, subtitles, scripts and meta information like genres, cast, director, rating *etc.*. There are totally $3K$ hour-long videos, $3.9M$ photos, $10M$ sentences of text and $7M$ items of meta information in MovieNet.

From the annotation aspect, MovieNet contains massive labels to support different research topics of movie understanding. Based on the belief that middle-level entities, *e.g.* character, place, are important for high-level story understanding, various kinds of annotations on semantic elements are provided in MovieNet, including character bounding box and identity, scene boundary, action/place tag and aligned description in natural language. In addition, since movie is an art of filming, the cinematic styles, *e.g.*, view scale, camera motion, lighting, *etc.*, are also beneficial for comprehensive video analysis. Thus we also annotate the view scale and camera motion for more than $46K$ shots. Specifically, the annotations in MovieNet include: (1) $1.1M$ characters with bounding boxes and identities; (2) $40K$ scene boundaries; (3) $65K$ tags of action and place; (4) $12K$ description sentences aligned to movie segments; (5) $92K$ tags of cinematic styles.

Based on the data and annotations in MovieNet, we exploit some research topics that cover different aspects of movie understanding, *i.e.* genre analysis, cinematic style prediction, character analysis, scene understanding, and movie segment retrieval. For each topic, we set up one or several challenging bench-

Fig. 2: MovieNet is a holistic dataset for movie understanding, which contains massive data from different modalities and high-quality annotations in different aspects. Here we show some data (in blue) and annotations (in green) of *Titanic* in MovieNet.

marks. Then extensive experiments are executed to present the performances of different methods. By further analysis on the experimental results, we will also show the gap of current approaches towards comprehensive movie understanding, as well as the advantages of holistic annotations for throughout video analytics.

To the best of our knowledge, MovieNet is the first holistic dataset for movie understanding that contains a large amount of data from different modalities and high-quality annotations in different aspects. We hope that it would promote the researches on video editing, human-centric situation understanding, story-based video analytics and beyond.

2 Related Datasets

Existing Works. Most of the datasets of movie understanding focus on a specific element of movies, *e.g.* genre [66,49], character [1,3,26,39,51,19,29], action [32,18,37,5,6], scene [43,11,25,40,14,41] and description [47]. Also their scale is quite small and the annotation quantities are limited. For example, [19,51,3] take several episodes from TV series for character identification, [32] uses clips from twelve movies for action recognition, and [40] exploits scene segmentation with only three movies. Although these datasets focus on some important aspects of movie understanding, their scale is not enough for the data-hungry learning paradigm. Furthermore, the deep comprehension should go from middle-level elements to high-level story while each existing dataset can only support a single task, causing trouble for comprehensive movie understanding.

MovieQA. MovieQA [54] consists of $15K$ questions designed for 408 movies. As for sources of information, it contains video clips, plots, subtitles, scripts, and DVS (Descriptive Video Service). To evaluate story understanding by QA is a good idea, but there are two problems. (1) Middle-level annotations, *e.g.*, character identities, are missing. Therefore it is hard to develop an effective approach towards high-level understanding. (2) The questions in MovieQA come from the wiki plot. Thus it is more like a textual QA problem rather than story-based video understanding. A strong evidence is that the approaches based on textual plot can get a much higher accuracy than those based on "video+subtitle".

LSMDC. LSMDC [45] consists of 200 movies with audio description (AD) providing linguistic descriptions of movies for visually impaired people. AD is quite different from the natural descriptions of most audiences, limiting the usage of the models trained on such datasets. And it is also hard to get a large number of ADs. Different from previous work [54,45], we provide multiple sources of textual information and different annotations of middle-level entities in MovieNet, leading to a better source for story-based video understanding.

AVA. Recently, AVA dataset [24], an action recognition dataset with 430 15-min movie clips annotated with 80 spatial-temporal atomic visual actions, is proposed. AVA dataset aims at facilitating the task of recognizing atomic visual actions. However, regarding the goal of story understanding, the AVA dataset is not applicable since (1) The dataset is dominated by labels like *stand* and *sit*, making it extremely unbalanced. (2) Actions like *stand, talk, watch* are less informative in the perspective of story analytics. Hence, we propose to annotate semantic level actions for both action recognition and story understanding tasks.

MovieGraphs. MovieGraphs [55] is the most related one that provides graph-based annotations of social situations depicted in clips of 51 movies. The annotations consist of characters, interactions, attributes, *etc.*. Although sharing the same idea of multi-level annotations, MovieNet is different from MovieGraphs in three aspects: (1) MovieNet contains not only movie clips and annotations, but also photos, subtitles, scripts, trailers, *etc.*, which can provide richer data for various research topics. (2) MovieNet can support and exploit different aspects of movie understanding while MovieGraphs focuses on situation recognition only. (3) The scale of MovieNet is much larger than MovieGraphs.

Table 1: Comparison between MovieNet and related datasets in terms of data.

	# movie	trailer	photo	meta	script	synop.	subtitle	plot	AD
MovieQA[54]	140						✓	✓	
LSMDC[45]	200				✓				✓
MovieGraphs[55]	51								
AVA[24]	430								
MovieNet	1,100	✓	✓	✓	✓	✓	✓	✓	

Table 2: Comparison between MovieNet and related datasets in terms of annotation.

	# character	# scene	# cine. tag	# aligned sent.	# action/place tag
MovieQA[54]	-	-	-	15K	-
LSMDC[45]	-	-	-	128K	-
MovieGraphs[55]	22K	-	-	21K	23K
AVA[24]	116K	-	-	-	360K
MovieNet	1.1M	42K	92K	25K	65K

3 Visit MovieNet: Data and Annotation

MovieNet contains various kinds of data from multiple modalities and high-quality annotations on different aspects for movie understanding. Fig. 2 shows the data and annotations of the movie *Titanic* in MovieNet. Comparisons between MovieNet and other datasets for movie understanding are shown in Tab. 1 and Tab. 2. All these demonstrate the tremendous advantage of MovieNet on both quality, scale and richness.

3.1 Data in MovieNet

Movie. We carefully selected and purchased the copies of 1,100 movies, the criteria of which are (1) colored; (2) longer than 1 hour; (3) cover a wide range of genres, years and countries.

Metadata. We get the meta information of the movies from IMDb and TMDb[1], including title, release date, country, genres, rating, runtime, director, cast, storyline, *etc.*. Here we briefly introduce some of the key elements, please refer to supplementary material for detail: (1) Genre is one of the most important attributes of a movie. There are total $805K$ genre tags from 28 unique genres in MovieNet. (2) For cast, we get both their names, IMDb IDs and the character names in the movie. (3) We also provide IMDb ID, TMDb ID and Douban ID of each movie, with which the researchers can get additional meta information from these websites conveniently. The total number of meta information in MovieNet is $375K$. Please note that each kind of data itself, even without the movie, can support some research topics [31]. So we try to get each kind of data as much

[1] IMDb: https://www.imdb.com; TMDb: https://www.themoviedb.org

as we can. Therefore the number here is larger than $1,100$. So as other kinds of data we would introduce below.

Subtitle. The subtitles are obtained in two ways. Some of them are extracted from the embedded subtitle stream in the movies. For movies without original English subtitle, we crawl the subtitles from YIFY[2]. All the subtitles are manually checked to ensure that they are aligned to the movies.

Trailer. We download the trailers from YouTube according to their links from IMDb and TMDb. We found that this scheme is better than previous work [10], which use the titles to search trailers from YouTube, since the links of the trailers in IMDb and TMDb have been manually checked by the organizers and audiences. Totally, we collect $60K$ trailers belonging to $33K$ unique movies.

Script. Script, where the movement, actions, expression and dialogs of the characters are narrated, is a valuable textual source for research topics of movie-language association. We collect around $2K$ scripts from IMSDb and Daily Script[3]. The scripts are aligned to the movies by matching the dialog with subtitles.

Synopsis. A synopsis is a description of the story in a movie written by audiences. We collect $11K$ high-quality synopses from IMDb, all of which contain more than 50 sentences. Synopses are also manually aligned to the movie, which would be introduced in Sec. 3.2.

Photo. We collect $3.9M$ photos of the movies from IMDb and TMDb, including poster, still frame, publicity, production art, product, behind the scene and event.

3.2 Annotation in MovieNet

To provide a high-quality dataset supporting different research topics on movie understanding, we make great effort to clean the data and manually annotate various labels on different aspects, including character, scene, event and cinematic style. Here we just demonstrate the *content* and the *amount* of annotations due to the space limit. Please refer to supplementary material for details.

Cinematic Styles. Cinematic style, such as view scale, camera movement, lighting and color, is an important aspect of comprehensive movie understanding since it influences how the story is telling in a movie. In MovieNet, we choose two kinds of cinematic tags for study, namely view scale and camera movement. Specifically, the view scale include five categories, *i.e. long shot*, *full shot*, *medium shot*, *close-up shot* and *extreme close-up shot*, while the camera movement is divided into four classes, *i.e. static shot*, *pans and tilts shot*, *zoom in* and *zoom out*. The original definitions of these categories come from [22] and we simplify them for research convenience. We totally annotate $47K$ shots from movies and trailers, each with one tag of view scale and one tag of camera movement.

Character Bounding Box and Identity. Person plays an important role in human-centric videos like movies. Thus to detect and identify characters is a foundational work towards movie understanding. The annotation process of character bounding box and identity contains 4 steps: (1) Some key frames, the

[2] https://www.yifysubtitles.com/

[3] IMSDb: https://www.imsdb.com/; DailyScript: https://www.dailyscript.com/

number of which is $758K$, from different movies are selected for bounding box annotation. (2) A detector is trained with the annotations in step-1. (3) We use the trained detector to detect more characters in the movies and manually clean the detected bounding boxes. (4) We then manually annotate the identities of all the characters. To make the cost affordable, we only keep the top 10 cast in credits order according to IMDb, which can cover the main characters for most movies. Characters not belong to credited cast were labeled as "others". In total, we got $1.1M$ instances of $3,087$ unique credited cast and $364K$ "others".

Scene Boundary. In terms of temporal structure, a movie contains two hierarchical levels – shot, and scene. Shot is the minimal visual unit of a movie while scene is a sequence of continued shots that are semantically related. To capture the hierarchical structure of a movie is important for movie understanding. Shot boundary detection has been well solved by [48], while scene boundary detection, also named scene segmentation, remains an open question. In MovieNet, we manually annotate the scene boundaries to support the researches on scene segmentation, resulting in $42K$ scenes.

Action/Place Tags. To understand the event(s) happened within a scene, action and place tags are required. Hence, we first split each movie into clips according to the scene boundaries and then manually annotated place and action tags for each segment. For place annotation, each clip is annotated with multiple place tags, *e.g.*, {deck, cabin}. While for action annotation, we first detect sub-clips that contain characters and actions, then we assign multiple action tags to each sub-clip. We have made the following efforts to keep tags diverse and informative: (1) We encourage the annotators to create new tags. (2) Tags that convey little information for story understanding, *e.g.*, *stand* and *talk*, are excluded. Finally, we merge the tags and filtered out 80 actions and 90 places with a minimum frequency of 25 as the final annotations. In total, there are $42K$ segments with $19.6K$ place tags and $45K$ action tags.

Description Alignment Since the event is more complex than character and scene, a proper way to represent an event is to describe it with natural language. Previous works have already aligned script [37], Descriptive Video Service (DVS) [45], book [67] or wiki plot [52,53,54] to movies. However, books cannot be well aligned since most of the movies would be quite different from their books. DVS transcripts are quite hard to obtain, limiting the scale of the datasets based on them [45]. Wiki plot is usually a short summary that cannot cover all the important events of the movie. Considering the issues above, we choose synopses as the story descriptions in MovieNet. The associations between the movie segments and the synopsis paragraphs are manually annotated by three different annotators with a coarse-to-fine procedure. Finally, we obtained $4,208$ highly consistent paragraph-segment pairs.

4 Play with MovieNet: Benchmark and Analysis

With a large amount of data and holistic annotations, MovieNet can support various research topics. In this section, we try to analyze movies from five as-

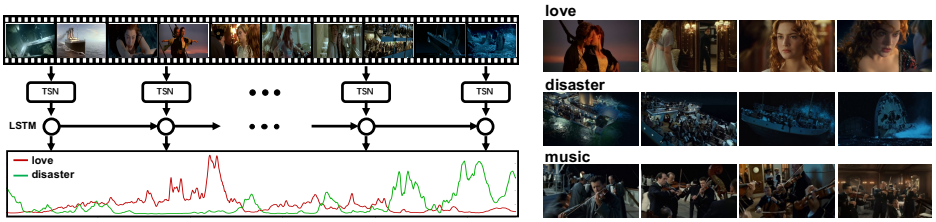
Table 3: (a). Comparison between MovieNet and other benchmarks for genre analysis. (b). Results of some baselines for genre classification in MovieNet

(a)

	genre	movie	trailer	photo
MGCD[66]	4	-	1.2K	-
LMTD[49]	4	-	3.5K	-
MScope[10]	13	-	5.0K	5.0K
MovieNet	**21**	**1.1K**	**68K**	**1.6M**

(b)

Data	Model	r@0.5	p@0.5	mAP
Photo	VGG16 [50]	27.32	66.28	32.12
	ResNet50 [27]	**34.58**	**72.28**	**46.88**
Trailer	TSN-r50[57]	17.95	**78.31**	43.70
	I3D-r50 [9]	16.54	69.58	35.79
	TRN-r50 [65]	**21.74**	77.63	**45.23**



(a) **(b)**

Fig. 3: (a). Framework of genre analysis in movies. (b). Some samples of genre-guided trailer generation for movie *Titanic*.

pects, namely *genre*, *cinematic style*, *character*, *scene* and *story*. For each topic, we would set up one or several benchmarks based on MovieNet. Baselines with currently popular techniques and analysis on experimental results are also provided to show the potential impact of MovieNet in various tasks. The topics of the tasks have covered different perspectives of comprehensive movie understanding. But due to the space limit, here we can only touched the tip of the iceberg. More detailed analysis are provided in the supplementary material and more interesting topics to be exploited are introduced in Sec. 5.

4.1 Genre Analysis

Genre is a key attribute for any media with artistic elements. To classify the genres of movies has been widely studied by previous works [66,49,10]. But there are two drawbacks for these works. (1) The scale of existing datasets is quite small. (2) All these works focus on image or trailer classification while ignore a more important problem, *i.e.* how to analyze the genres of a long video.

MovieNet provides a large-scale benchmark for genre analysis, which contains $1.1K$ movies, $68K$ trailers and $1.6M$ photos. The comparison between different datasets are shown in Tab. 3a, from which we can see that MovieNet is much larger than previous datasets.

Based on MovieNet, we first provide baselines for both image-based and video-based genre classification, the results are shown Tab. 3b. Comparing the result of genre classification in small datasets [49,10] to ours in MovieNet, we find that the performance drops a lot when the scale of the dataset become larger. The newly proposed MovieNet brings two challenges to previous methods. (1) Genre classification in MovieNet becomes a long-tail recognition problem where the label distribution is extremely unbalanced. For example, the number of "Drama" is 40 times larger than that of "Sport" in MovieNet. (2) Genre is a high-level semantic tag depending on action, clothing and facial expression of the characters, and even BGM. Current methods are good at visual representation. When facing a problem that need to consider higher-level semantics, they would all fail. We hope MovieNet would promote researches on these challenging topics.

Another new issue to address is how to analyze the genres of a movie. Since movie is extremely long and not all segments are related to its genres, this problem is much more challenging. Following the idea of learning from trailers and applying to movies [30], we adopt the visual model trained with trailers as shot-level feature extractor. Then the features are fed to a temporal model to capture the temporal structure of the movie. The overall framework is shown in Fig. 3a. With this approach, we can get the genre response curve of a movie. Specifically, we can predict which part of the movie is more relevant to a specific genre. What's more, the prediction can also be used for genre-guided trailer generation, as shown in Fig. 3b. From the analysis above, we can see that MovieNet would promote the development of this challenging and valuable research topic.

4.2 Cinematic Style Analysis

As we mentioned before, cinematic style is about how to present the story to audience in the perspective of filming art. For example, a *zoom in* shot is usually used to attract the attention of audience to a specific object. In fact, cinematic style is crucial for both video understanding and editing. But there are few works focusing on this topic and no large-scale datasets for this research topic too.

Based on the tags of cinematic style we annotated in MovieNet, we set up a benchmark for cinematic style prediction. Specifically, we would like to recognize

Table 4: (a). Comparison between MovieNet and other benchmarks for cinematic style prediction. (b). Results of some baselines for cinematic style prediction in MovieNet

(a)

	shot	video	scale	move.
Lie 2014 [4]	327	327		✓
Sports 2007 [62]	1,364	8	✓	
Context 2011 [61]	3,206	4	✓	
Taxon 2009 [56]	5,054	7		✓
MovieNet	46,857	7,858	✓	✓

(b)

Method	scale acc.	move. acc.
I3D [9]	76.79	78.45
TSN [57]	84.08	70.46
TSN+R^{3}Net[16]	**87.50**	**80.65**

Table 5: Datasets for person analysis.

	ID	instance	source
COCO[35]	-	262K	web image
CalTech[17]	-	350K	surveillance
Market[64]	1,501	32K	surveillance
CUHK03[33]	1,467	28K	surveillance
AVA[24]	-	426K	movie
CSM[28]	1,218	127K	movie
MovieNet	**3,087**	**1.1M**	movie

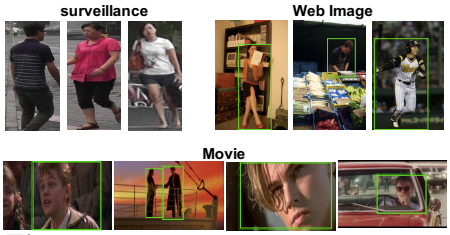


Fig. 4: Persons in different data sources

Table 6: Results of (a). Character Detection and (b).Character Identification

(a)

Train Data	Method	mAP
COCO[35]	FasterRCNN	81.50
Caltech[17]	FasterRCNN	5.67
CSM[28]	FasterRCNN	89.91
MovieNet	FasterRCNN	92.13
	RetinaNet	91.55
	CascadeRCNN	**95.17**

(b)

Train Data	cues	Method	mAP
Market[64]	body	r50-softmax	4.62
CUHK03[33]	body	r50-softmax	5.33
CSM[28]	body	r50-softmax	26.21
MovieNet	body	r50-softmax	32.81
	body+face	two-step[36]	63.95
	body+face	PPCC[28]	**75.95**

the view scale and camera motion of each shot. Comparing to existing datasets, MovieNet is the first dataset that covers both view scale and camera motion, and it is also much larger, as shown in Tab. 4a. Several models for video clip classification such as TSN [57] and I3D [9] are applied to tackle this problem, the results are shown in Tab. 4b. Since the view scale depends on the portion of the subject in the shot frame, to detect the subject is important for cinematic style prediction. Here we adopt the approach from saliency detection [16] to get the subject maps of each shot, with which better performances are achieved, as shown in Tab. 4b. Although utilizing subject points out a direction for this task, there is still a long way to go. We hope that MovieNet can promote the development of this important but ignored topic for video understanding.

4.3 Character Recognition

It has been shown by existing works [55,58,36] that movie is a human-centric video where characters play an important role. Therefore, to detect and identify characters is crucial for movie understanding. Although person/character recognition is not a new task, all previous works either focus on other data sources [64,33,35] or small-scale benchmarks [26,3,51], leading to the results lack of convincingness for character recognition in movies.

We proposed two benchmarks for character analysis in movies, namely, character detection and character identification. We provide more than $1.1M$ instances from $3,087$ identities to support these benchmarks. As shown in Tab. 5,

Table 7: Dataset for scene analysis.

	scene	action	place
OVSD [46]	300	-	-
BBC [2]	670	-	-
Hollywood2 [37]	-	1.7K	1.2K
MovieGraph[55]	-	23.4K	7.6K
AVA [24]	-	360K	-
MovieNet	**42K**	**45.0K**	**19.6K**

Table 8: Datasets for story understanding in movies in terms of (1) number of sentences per movie; (2) duration (second) per segment.

Dataset	sent./mov.	dur./seg.
MovieQA [54]	35.2	202.7
MovieGraphs [55]	408.8	44.3
MovieNet	83.4	428.0

Table 9: Results of Scene Segmentation

Dataset	Method	AP(↑)	M_{iou}(↑)
OVSD [46]	MS-LSTM	0.313	0.387
BBC [2]	MS-LSTM	0.334	0.379
MovieNet	Grouping [46]	0.336	0.372
	Siamese [2]	0.358	0.396
	MS-LSTM	**0.465**	**0.462**

Table 10: Results of Scene Tagging

Tags	Method	mAP
action	TSN [57]	14.17
	I3D [9]	20.69
	SlowFast [20]	**23.52**
place	I3D [9]	7.66
	TSN [57]	**8.33**

MovieNet contains much more instances and identities comparing to some popular datasets about person analysis. The following sections will show the analysis on character detection and identification respectively.

Character Detection. Images from different data sources would have large domain gap, as shown in Fig. 4. Therefore, a character detector trained on general object detection dataset, *e.g.* COCO [35], or pedestrian dataset, *e.g.* CalTech [17], is not good enough for detecting characters in movies. This can be supported by the results shown in Tab. 6a. To get a better detector for character detection, we train different popular models [44,34,8] with MovieNet using toolboxes from [13,12]. We can see that with the diverse character instances in MovieNet, a Cascade R-CNN trained with MovieNet can achieve extremely high performance, *i.e.* 95.17% in mAP. That is to say, character detection can be well solved by a large-scale movie dataset with current SOTA detection models. This powerful detector would then benefit research on character analysis in movies.

Character Identification. To identify the characters in movies is a more challenging problem, which can be observed by the diverse samples shown in Fig. 4. We conduct different experiments based on MovieNet, the results are shown in Tab. 6b. From these results, we can see that: (1) models trained on ReID datasets are inefficient for character recognition due to domain gap; (2) to aggregate different visual cues of an instance is important for character recognition in movies; (3) the current state-of-the-art can achieve 75.95% mAP, which demonstrates that it is a challenging problem which need to be further exploited.

4.4 Scene Analysis

As mentioned before, scene is the basic semantic unit of a movie. Therefore, it is important to analyze the scenes in movies. The key problems in scene

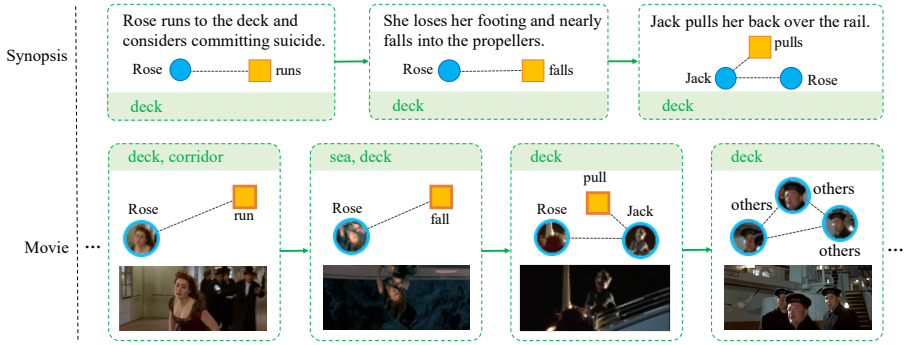


Fig. 5: Example of synopses paragraph and movie segment in MovieNet-MSR. It demonstrate the spatial-temporal structures of stories in movies and synopses. We can also see that character, action and place are the key element for story understanding.

understanding is probably *where is the scene boundary* and *what is the content in a scene*. As shown in Tab. 7, MovieNet, which contains more than 43K scene boundaries and 65K action/place tags, is the only one that can support both *scene segmentation* and *scene tagging*. What's more, the scale of MovieNet is also larger than all previous works.

Scene Segmentation We first test some baselines [46,2] for scene segmentation. In addition, we also propose a sequential model, named Multi-Semtantic LSTM (MS-LSTM) based on Bi-LSTMs [23,42] to study the gain brought by using multi-modality and multiple semantic elements, including audio, character, action and scene. From the results shown in Tab. 9, we can see that (1) Benefited from large scale and high diversity, models trained on MovieNet can achieve better performance. (2) Multi-modality and multiple semantic elements are important for scene segmentation, which highly raise the performance.

Action/Place Tagging To further understand the stories within a movie, it is essential to perform analytics on the key elements of storytelling, *i.e.*, place and action. We would introduce two benchmarks in this section. Firstly, for action analysis, the task is multi-label action recognition that aims to recognize all the human actions or interactions in a given video clip. We implement three standard action recognition models, *i.e.*, TSN [57], I3D [9] and SlowFast Network [20] modified from [63] in experiments. Results are shown in Tab. 10. For place analysis, we propose another benchmark for multi-label place classification. We adopt I3D [9] and TSN [57] as our baseline models and the results are shown in Tab. 10. From the results, we can see that action and place tagging is an extremely challenging problem due to the high diversity of different instances.

4.5 Story Understanding

Web videos are broadly adopted in previous works [7,60] as the source of video understanding. Compared to web videos, the most distinguishing feature of

movies is the story. Movies are created to tell stories and the most explicit way to demonstrate a story is to describe it using natural language, *e.g.* synopsis. Inspired by the above observations, we choose the task of movie segment retrieval with natural language to analyze the stories in movies. Based on the aligned synopses in MovieNet, we set up a benchmark for movie segment retrieval. Specifically, given a synopsis paragraph, we aim to find the most relevant movie segment that covers the story in the paragraph. It is a very challenging task due to the rich content in movie and high-level semantic descriptions in synopses. Tab. 8 shows the comparison of our benchmark dataset with other related datasets. We can see that our dataset is more complex in terms of descriptions compared with MovieQA [54] while the segments are longer and contain more information than those of MovieGraphs [55].

Generally speaking, a story can be summarized as "*somebody do something in some time at some place*". As shown in Fig. 5, both stories represented by language and video can be composed as sequences of {character, action, place} graphs. That being said, to understand a story is to (1) recognize the key elements of story-telling, namely, character, action, place *etc.*; (2) analyze the spatial-temporal structures of both movie and synopsis. Hence, our method first leverage middle-level entities (*e.g.* character, scene), as well as multi-modality (*e.g.* subtitle) to assist retrieval. Then we explore the spatial-temporal structure from both movies and synopses by formulating middle-level entities into graph structures. Please refer to supplementary material for details.

Using middle-level entities and multi-modality. We adopt VSE [21] as our baseline model where the vision and language features are embedded into a joint space. Specifically, the feature of the paragraph is obtained by taking the average of Word2Vec [38] feature of each sentence while the visual feature is obtained by taking the average of the appearance feature extracted from ResNet [27] on each shot. We add subtitle feature to enhance visual feature. Then different semantic elements including character, action and cinematic style are aggregated in our framework. We are able to obtain action features and character features thanks to the models trained on other benchmarks on MovieNet, *e.g.*, action recognition and character detection. Furthermore, we observe that the focused elements vary under different cinematic styles. For example, we should focus more on actions in a full shot while more on character and dialog in a close-up shot. Motivated by this observation, we propose a cinematic-style-guided attention module that predicts the weights over each element (*e.g.*, action, character) within a shot, which would be used to enhance the visual features. The experimental results are shown in Tab. 11. Experiments show that by considering different elements of the movies, the performance improves a lot. We can see that a holistic dataset which contains holistic annotations to support middle-level entity analyses is important for movie understanding.

Explore spatial-temporal graph structure in movies and synopses. Simply adding different middle-level entities improves the result. Moreover, as shown in Fig. 5, we observe that stories in movies and synopses persist two important structure: (1) the temporal structure in movies and synopses is that the story

Table 11: Results of movie segment retrieval. Here, G stands for global appearance feature, S for subtitle feature, A for action, P for character and C for cinematic style.

Method	Recall@1	Recall@5	Recall@10	MedR
Random	0.11	0.54	1.09	460
G	3.16	11.43	18.72	66
G+S	3.37	13.17	22.74	56
G+S+A	5.22	13.28	20.35	52
G+S+A+P	18.50	43.96	55.50	7
G+S+A+P+C	18.72	44.94	56.37	7
MovieSynAssociation [59]	**21.98**	**51.03**	**63.00**	**5**

can be composed as a sequence of events following a certain temporal order. (2) the spatial relation of different middle-level elements, *e.g.*, character co-existence and their interactions, can be formulated as graphs. We implement the method in [59] to formulate the above structures as two graph matching problems. The result are shown in Tab. 11. Leveraging the graph formulation for the internal structures of stories in movies and synopses, the retrieval performance can be further boosted, which in turn, show that the challenging MovieNet would provide a better source to story-based movie understanding.

5 Discussion and Future Work

In this paper, we introduce MovieNet, a holistic dataset containing different aspects of annotations to support comprehensive movie understanding. We introduce several challenging benchmarks on different aspects of movie understanding, *i.e.* discovering filming art, recognizing middle-level entities and understanding high-level semantics like stories. Furthermore, the results of movie segment retrieval demonstrate that integrating filming art and middle-level entities according to the internal structure of movies would be helpful for story understanding. These in turn, show the effectiveness of holistic annotations.

In the future, our work would go on in two aspects. (1) **Extending the Annotation.** In the future, we would further extend the dataset to include more movies and annotations. (2) **Exploring more Approaches and Topics.** To tackle the challenging tasks proposed above, we would explore more effective approaches. Besides, there are more meaningful and practical topics that can be addressed with MovieNet, such as movie deoldify, trailer generation, *etc.*

Acknowledgment This work is partially supported by the SenseTime Collaborative Grant on Large-scale Multi-modality Analysis (CUHK Agreement No. TS1610626 & No. TS1712093), the General Research Fund (GRF) of Hong Kong (No. 14203518 & No. 14205719), and Innovation and Technology Support Program (ITSP) Tier 2, ITS/431/18F.

References

1. Arandjelovic, O., Zisserman, A.: Automatic face recognition for film character retrieval in feature-length films. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE (2005) 4
2. Baraldi, L., Grana, C., Cucchiara, R.: A deep siamese network for scene detection in broadcast videos. In: 23rd ACM International Conference on Multimedia. pp. 1199–1202. ACM (2015) 11, 12
3. Bauml, M., Tapaswi, M., Stiefelhagen, R.: Semi-supervised learning with constraints for person identification in multimedia data. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2013) 4, 10
4. Bhattacharya, S., Mehran, R., Sukthankar, R., Shah, M.: Classification of cinematographic shots using lie algebra and its application to complex event recognition. IEEE Transactions on Multimedia **16**(3), 686–696 (April 2014) 9
5. 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 (2013) 4
6. Bojanowski, P., Lajugie, R., Bach, F., Laptev, I., Ponce, J., Schmid, C., Sivic, J.: Weakly supervised action labeling in videos under ordering constraints. In: European Conference on Computer Vision. Springer (2014) 4
7. Caba Heilbron, F., Escorcia, V., Ghanem, B., Carlos Niebles, J.: Activitynet: A large-scale video benchmark for human activity understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 961–970 (2015) 2, 12
8. Cai, Z., Vasconcelos, N.: Cascade r-cnn: Delving into high quality object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6154–6162 (2018) 11
9. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6299–6308 (2017) 8, 9, 10, 11, 12
10. Cascante-Bonilla, P., Sitaraman, K., Luo, M., Ordonez, V.: Moviescope: Large-scale analysis of movies using multiple modalities. arXiv preprint arXiv:1908.03180 (2019) 6, 8, 9
11. Chasanis, V.T., Likas, A.C., Galatsanos, N.P.: Scene detection in videos using shot clustering and sequence alignment. IEEE transactions on multimedia (2008) 4
12. Chen, K., Pang, J., Wang, J., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Shi, J., Ouyang, W., Change Loy, C., Lin, D.: Hybrid task cascade for instance segmentation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019) 11
13. Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Xu, J., Zhang, Z., Cheng, D., Zhu, C., Cheng, T., Zhao, Q., Li, B., Lu, X., Zhu, R., Wu, Y., Dai, J., Wang, J., Shi, J., Ouyang, W., Loy, C.C., Lin, D.: MMDetection: Open mmlab detection toolbox and benchmark. arXiv preprint arXiv:1906.07155 (2019) 11
14. Del Fabro, M., Böszörmenyi, L.: State-of-the-art and future challenges in video scene detection: a survey. Multimedia systems (2013) 4
15. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. Ieee (2009) 2

16. Deng, Z., Hu, X., Zhu, L., Xu, X., Qin, J., Han, G., Heng, P.A.: R3net: Recurrent residual refinement network for saliency detection. In: Proceedings of the 27th International Joint Conference on Artificial Intelligence. pp. 684–690. AAAI Press (2018) 9, 10
17. Dollar, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian detection: An evaluation of the state of the art. IEEE transactions on pattern analysis and machine intelligence **34**(4), 743–761 (2011) 10, 11
18. Duchenne, O., Laptev, I., Sivic, J., Bach, F.R., Ponce, J.: Automatic annotation of human actions in video. In: Proceedings of the IEEE International Conference on Computer Vision (2009) 4
19. Everingham, M., Sivic, J., Zisserman, A.: Hello my name is... buffy – automatic naming of characters in tv video. In: BMVC (2006) 4
20. Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 6202–6211 (2019) 11, 12
21. Frome, A., Corrado, G.S., Shlens, J., Bengio, S., Dean, J., Ranzato, M., Mikolov, T.: Devise: A deep visual-semantic embedding model. In: Advances in neural information processing systems. pp. 2121–2129 (2013) 13
22. Giannetti, L.D., Leach, J.: Understanding movies, vol. 1. Prentice Hall Upper Saddle River, New Jersey (1999) 6
23. Graves, A., Schmidhuber, J.: Framewise phoneme classification with bidirectional lstm and other neural network architectures. Neural networks **18**(5-6), 602–610 (2005) 12
24. 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) 4, 5, 10, 11
25. Han, B., Wu, W.: Video scene segmentation using a novel boundary evaluation criterion and dynamic programming. In: IEEE International conference on multimedia and expo. IEEE (2011) 4
26. Haurilet, M.L., Tapaswi, M., Al-Halah, Z., Stiefelhagen, R.: Naming tv characters by watching and analyzing dialogs. In: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE (2016) 4, 10
27. 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 (2016) 8, 13
28. 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) (2018) 10
29. 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) 4
30. 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) 9
31. 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: Proceedings of the European Conference on Computer Vision (ECCV) (2020) 5
32. Laptev, I., Marszałek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society (2008) 4

33. Li, W., Zhao, R., Xiao, T., Wang, X.: Deepreid: Deep filter pairing neural network for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 152–159 (2014) 10
34. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017) 11
35. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014) 10, 11
36. Loy, C.C., Lin, D., Ouyang, W., Xiong, Y., Yang, S., Huang, Q., Zhou, D., Xia, W., Li, Q., Luo, P., et al.: Wider face and pedestrian challenge 2018: Methods and results. arXiv preprint arXiv:1902.06854 (2019) 10
37. Marszałek, M., Laptev, I., Schmid, C.: Actions in context. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society (2009) 4, 7, 11
38. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013) 13
39. Nagrani, A., Zisserman, A.: From benedict cumberbatch to sherlock holmes: character identification in tv series without a script. BMVC (2017) 4
40. Park, S.B., Kim, H.N., Kim, H., Jo, G.S.: Exploiting script-subtitles alignment to scene boundary dectection in movie. In: IEEE International Symposium on Multimedia. IEEE (2010) 4
41. Rao, A., Wang, J., Xu, L., Jiang, Xuekun, H.Q., Zhou, B., Lin, D.: A unified framework for shot type classification based on subject centric lens. In: Proceedings of the European Conference on Computer Vision (ECCV) (2020) 4
42. 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) 12
43. Rasheed, Z., Shah, M.: Detection and representation of scenes in videos. IEEE transactions on Multimedia (2005) 4
44. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in neural information processing systems. pp. 91–99 (2015) 11
45. Rohrbach, A., Rohrbach, M., Tandon, N., Schiele, B.: A dataset for movie description. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3202–3212 (2015) 4, 5, 7
46. Rotman, D., Porat, D., Ashour, G.: Optimal sequential grouping for robust video scene detection using multiple modalities. International Journal of Semantic Computing **11**(02), 193–208 (2017) 11, 12
47. Shao, D., Xiong, Y., Zhao, Y., Huang, Q., Qiao, Y., Lin, D.: Find and focus: Retrieve and localize video events with natural language queries. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 200–216 (2018) 4
48. Sidiropoulos, P., Mezaris, V., Kompatsiaris, I., Meinedo, H., Bugalho, M., Trancoso, I.: Temporal video segmentation to scenes using high-level audiovisual features. IEEE Transactions on Circuits and Systems for Video Technology **21**(8), 1163–1177 (2011) 7
49. Simões, G.S., Wehrmann, J., Barros, R.C., Ruiz, D.D.: Movie genre classification with convolutional neural networks. In: 2016 International Joint Conference on Neural Networks (IJCNN). IEEE (2016) 4, 8, 9

50. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014) 8
51. Tapaswi, M., Bäuml, M., Stiefelhagen, R.: knock! knock! who is it? probabilistic person identification in tv-series. In: IEEE Conference on Computer Vision and Pattern Recognition. IEEE (2012) 4, 10
52. Tapaswi, M., Bäuml, M., Stiefelhagen, R.: Story-based video retrieval in tv series using plot synopses. In: Proceedings of International Conference on Multimedia Retrieval. p. 137. ACM (2014) 7
53. Tapaswi, M., Bäuml, M., Stiefelhagen, R.: Aligning plot synopses to videos for story-based retrieval. International Journal of Multimedia Information Retrieval **4**(1), 3–16 (2015) 7
54. Tapaswi, M., Zhu, Y., Stiefelhagen, R., Torralba, A., Urtasun, R., Fidler, S.: Movieqa: Understanding stories in movies through question-answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition (2016) 4, 5, 7, 11, 13
55. Vicol, P., Tapaswi, M., Castrejon, L., Fidler, S.: Moviegraphs: Towards understanding human-centric situations from videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2018) 4, 5, 10, 11, 13
56. 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) 9
57. Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Van Gool, L.: Temporal segment networks: Towards good practices for deep action recognition. In: European conference on computer vision. pp. 20–36. Springer (2016) 8, 9, 10, 11, 12
58. Xia, J., Rao, A., Xu, L., Huang, Q., Wen, J., Lin, D.: Online multi-modal person search in videos. In: Proceedings of the European Conference on Computer Vision (ECCV) (2020) 10
59. Xiong, Y., Huang, Q., Guo, L., Zhou, H., Zhou, B., Lin, D.: A graph-based framework to bridge movies and synopses. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4592–4601 (2019) 14
60. Xu, J., Mei, T., Yao, T., Rui, Y.: Msr-vtt: A large video description dataset for bridging video and language. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5288–5296 (2016) 12
61. Xu, M., Wang, J., Hasan, M.A., He, X., Xu, C., Lu, H., Jin, J.S.: Using context saliency for movie shot classification. In: 2011 18th IEEE International Conference on Image Processing. pp. 3653–3656. IEEE (2011) 9
62. Yang, Y., Lin, S., Zhang, Y., Tang, S.: Statistical framework for shot segmentation and classification in sports video. In: Computer Vision – ACCV 2007. pp. 106–115. Springer Berlin Heidelberg (2007) 9
63. Yue Zhao, Yuanjun Xiong, D.L.: Mmaction. https://github.com/open-mmlab/mmaction (2019) 12
64. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: A benchmark. In: Proceedings of the IEEE international conference on computer vision. pp. 1116–1124 (2015) 10
65. Zhou, B., Andonian, A., Oliva, A., Torralba, A.: Temporal relational reasoning in videos. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 803–818 (2018) 8
66. Zhou, H., Hermans, T., Karandikar, A.V., Rehg, J.M.: Movie genre classification via scene categorization. In: Proceedings of the 18th ACM international conference on Multimedia. pp. 747–750. ACM (2010) 4, 8

67. Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., Fidler, S.: Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In: Proceedings of the IEEE international conference on computer vision. pp. 19–27 (2015) 7