MND: A New Dataset and Benchmark of Movie Scenes Classified by their Narrative Function

Chang Liu, Armin Shmilovici, Mark Last

liuc@post.bgu.ac.il, {armin, mlast}@bgu.ac.il
Department of Software and Information Systems Engineering
Ben-Gurion University of the Negev (Israel)

Abstract. The success of Hollywood cinema is partially attributed to the notion that Hollywood filmmaking constitutes *both an art and an industry*: an artistic tradition based on a standardized approach to cinematic narration. Film theorists have explored the narrative structure of movies and identified forms and paradigms that are common to many movies – *a latent narrative structure*. We raise the challenge of understanding and formulating the movie story structure and introduce a novel story-based labeled dataset—the Movie Narrative Dataset (MND). The dataset consists of 6,448 scenes taken from a manual annotation of 45 cinema movies, by 119 distinct annotators. The story-related function of each scene was manually labeled by at least six different human annotators as one of 15 possible key story elements (such as *Set-Up*, *Debate*, *Midpoint*) defined in screenwriting guidelines.

To benchmark the task of scene classification by their narrative function, we trained a XGBoost classifier that uses simple temporal features and character co-occurrence features to classify each movie scene into one of the story beats. With five-fold cross validation over the movies, the XGBoost classifier produced a F1 measure of 0.31 that is statistically significant above a static baseline classifier.

These initial results indicate the ability of machine learning approaches to detect the narrative structure in movies. Hence, the proposed dataset should contribute to the development of story-related video analytics tools, such as automatic video summarization and movie recommendation systems.

Keywords: Computational Narrative Understanding, Movie Understanding, Movie Analytics, Plot Points Detection, Scene Classification

1 Introduction

1.1 Background

The success of Hollywood cinema is partially attributed to the notion that Hollywood filmmaking constitutes *both an art and an industry*: an artistic tradition based on a standardized approach to cinematic narration [1,3]. This artistic system had influenced other cinemas, creating a sort of international film language. Hollywood has developed some fairly explicit principles for how stories can be told effectively. For example, some scenes will typically contain unresolved issues that demand settling further along [5]. Sometimes a film puzzles or frustrates us, when we cannot identify character goals or clear-cut lines of cause and effect [6]. Some manuals of screenwriting have picked up on the principles, turning them into explicit rules [2].

Popular movies present stories – narratives – in an audio-visual manner. Narrative is a core mechanism that human beings use to find meaning that helps them to understand their world [8]. *Narratology* is the study of stories and story structure and the ways these effect our perception, cognition, and emotion [9]. Most research focus on the story as it is physically communicated, not on the story, as it is understood. Here, we explore how the story form has been intended by its filmmakers, though implicitly, to engage its spectators.

Narratology is well developed in the “text worlds” (e.g., literature [28] – earliest known work on Narratology is Aristotle’s’ *Poetica*). The recent progress in Natural Language Processing has increasingly focused on developing computational models that reason about stories: in Computational Narrative Understanding, theoretical frameworks in narratology are used to develop computational models that can reason about stories [8,10]. A related problem is *narrative scene detection***,** which attempts to observe the spatial, temporal, and agential boundaries between story segments [12,13]. At a higher level, *narrative plotline detection*, is the act of assembling scenes into more general narrative units defined by agents who may range over both time and space [8,14].

Most works in video understanding are based on computer vision algorithms. Those algorithms perform well on basic, fact-based video understanding tasks, such as recognizing actions in video clips [38,40], question-answering about the video contents [21,39], and generating captions for videos [22,37]. However, most of these algorithms focus on analyzing *short video clips* (less than 30 seconds), which makes them very suitable for exploring the detailed (or low-level) information in videos such as *”playing soccer”* or *”running”* but very poor at understanding high-level events in those videos (e.g., *”enjoying a party”* or *”going home”*), due to the casual and temporal relationships between events, which can be complex and are often implicit.

The huge gap between the state-of-the-art computer vision algorithms and story analytics seems hard to be bridged, and therefore, novel approaches to understanding the video stories are needed. **The main contribution of this paper is that we formally raise a novel research task in the field of computational narrative understanding - identifying the narrative function of movie scenes according to screenwriting guidelines.** To this end, we collect and assemble a new benchmark dataset of movie scenes labeled by their narrative function – the Movie Narrative dataset (MND[1]). As a complement to the computer vision algorithms, features from the latent story structure can be utilized to enhance applications such as movie summarization [20,32,36], and movie recommendation [41].

---

[1] The MND dataset will be available via a github web-site

1.2 Research Objectives and Contributions

Our first objective is to provide a dataset of movie scenes labeled with high level concepts such as *Debate* – the internal conflict of the protagonist whether to return to her “comfort zone” after facing a *Catalyst event* that knocked her out of it. A crowd-sourcing experiment is used for constructing the dataset. The collected labels are analyzed to verify the following two hypotheses: (1) most movies in our dataset adhere fairly well to the latent story structure described by the screenwriting book [15] and (2) even non-experts can identify the scenes’ narrative function after reading the annotation guidelines and watching a movie.

Our second objective is to provide a lightweight solution for the challenging task of classifying movie scenes by their narrative function, with the use of relatively simple features and a supervised machine learning algorithms. Although a fully automated pipeline is desired, at this stage we use the manually annotated features from the MovieGraphs dataset [16] to avoid the errors prevalent in the current scene splitting, character identification and other movie annotation tools, which often fail in common cases such as darkly illuminated scenes.

The original contributions of our paper to the domain of computational narrative understanding in movies are two-fold: a) We introduce an new task for movie analytics – learning the latent narrative function of each scene. b) We introduce the first benchmark dataset of movie scenes labeled by their narrative function that will be released to the research community; and c), We demonstrate that a movie’s latent story structure can be automatically detected using machine learning with some relatively simple scene features, outperforming a strong temporal distribution baseline.

The rest of the paper is organized as follows: Section 2 presents some background and some related work; Section 3 describes the elements of the latent story model that we use; Section 4 describes the construction and the features of the MND dataset; Section 5 presents an MND task of movie scenes classification by their narrative function; and finally, Section 6 concludes the paper. The appendices in the supplementary document provide some technical details about the data collection and pre-processing.

2 Background and Related Work

2.1 Introduction to Story Models

The architecture of a typical movie at its highest level has four 25-35 minutes long acts—*Setup*, *Complication*, *Development*, and *Climax*—with two optional shorter subunits of *Prolog* and *Epilog* [1,9]. At a middle level, there are typically 40-60 scenes. The scenes develop and connect through short-term chains of cause and effect. Characters formulate specific plans, react to changing circumstances, gain or lose allies, and otherwise take specific steps toward or away from their goals [4]. At the third a level of organization audiovisual patterning carries the story along bit by bit. For example, within a scene, we often find patterns of cutting—an establishing shot introduces the setting, reverse camera angles meshed with the developing dialogue and close-ups cue the relation between the characters [11]. This paper focus on the middle level.

Aristotle's *Poetics* is the earliest surviving work on dramatic and literary theories in the West. His three-act form—as applied to movies, has come to mean *Act One*, *Act Two*, and *Act Three*. Each act has its own characteristics: Act One introduces the character(s) and the premise—what the movie is about; Act Two focuses on confrontation and struggle; Act Three resolves the crisis introduced in the premise. The three act structure was extended by [2] with various plot devices – or story "beats". In the film development terms, a "beat" refers to a single story event that transforms the character and story at a critical point in time. Beats such as *Inciting Incident* (typically in the first act), *Disaster* and *Crisis (must appear at least once)*, intend to intensify conflict, develop characters, and propel the plot forward. Beats can be also considered as "checkpoints" along the way, which will complete the story and reveal the movie's structure. Turning Points (TP) are moments that direct the plot in a different direction, therefore separate between acts. [19] attempts to identify the 5 TP that separate between the acts in feature length screenplays by projecting synopsis level annotations. There are also rules of thumb indicating where to expect each TP. For example, the 1st TP - the *Opportunity*, separates between the *Setup* act and the *New Situation* acts and is expected to occur after the initial 10% of the movie duration. A comparison between seven different story models is available [17].

In this paper, we decided to use the recent Save the Cat!® theory [15] which is popular among scriptwriters. It suggests that a good story is like music, which has beats that control the rhythm and flow. The theory defines for the writers 15 story beats we introduce in section 3, that play different roles in the story development. The main advantage of this model is that it can also incorporate two common deviations from the main story: a) *B-Story* - a plot device that carries the theme of the story, but in a different way with different characters. b) *Fun and Games* - scenes that are purely for the enjoyment of the audience (e.g., action scenes such as car chase, funny scenes, romantic scenes – depending on the genre). Thus, this story model can handle more complex movies that have side-stories and scenes that do not advance the plot.

2.2 Related Datasets and Research

Only recently, large movie datasets have been constructed for the purpose of movie understanding tasks. However, due to movie copyright issues, some large movie datasets do not contain all the scenes of a movie [27,38], or due to annotation difficulties, only some of the scenes are fully annotated [26]. Most datasets focus on a specific aspect of movies, such as, genre [43,44], Question-Answering [21,39], generating textual descriptions [22,37], or integrating vision and language relations [23]. Obtaining quality human annotation for a full movie is challenging, therefore, some of the annotations are generated from text, e.g. from synchronizing the movie with its script or synopsis. Following is the description of datasets that are sufficiently large, full (scene wise) and contain some high-level story elements labels, therefore, are most related.

MovieNet - 1,100 movies, 40K scenes, many modalities, (e.g., cinematic styles) quality annotations, and metadata [18]. *SyMoN* – 5,193 video summaries [48]. The most important aspect for story analytics is the alignment between a movie and its script and synopsis.

TRIPOD - 122 movies, 11,320 scenes [19,20], metadata, The most important aspect for story analytics is the annotation of the 5 Turning Points in a movie (via textual analysis of their scripts and synopsis).

MovieGraphs - 51 movies, 7.2K scenes, high quality manual annotation of the relationships and the interactions between movie characters [16,24]. We used some of those quality annotations for the construction of our dataset.

FSD – 60 episodes of the *Flintstones* cartoons, about 26 minutes long, 1,569 scenes [25]. Each scene was manually labeled with the 9 labels from the story model of [2]. A classifier was trained to predict the label for each scene. This is the most similar dataset to ours. The main differences are that they use a less elaborate story model than ours [15], for much shorter movies (only about 25% long), inaccurate scene splitting, with the same characters in each episode, while we have built a heterogeneous dataset of 45 movies with more quality annotations and features (e.g., manual scene splitting and character identification).

3 The Story Model: 15 Story Beats

What is a story beat? In the film development terms, a “beat” refers to a single story event which transforms the character and story at a critical point in time. For each beat, there is a suggested time for it to arrive in the story. In Table 1, we present the detailed definitions of the 15 story beats as well as their suggested appearance time within a typical 110 minutes movie [15]. The recommended position of each beat is proportional to the movie length. In our dataset, the function of each movie scene in the progression of the story is labeled by one of those 15 story beats. In addition, a “*None*” label is used for scenes that do not progress the plot significantly.

Table 1. Definition of the 15 story beats of [15]

- **Opening Image (minute 1):** It presents the first impression and sets the tone, mood, type and scope of the movie. It is an opportunity to give the audience a starting point of the hero, before the story begins.
- **Theme Stated (minute 5):** A character (often not the main character) will pose a question or make a statement (usually to the main character) that is the theme of the movie. It will not be obvious. Instead it will be off-hand conversational remark that the main character does not get at the time, but it will mean a lot later on.
- **Set-Up (minute 1-10)**: Sets-up the hero, the stakes, and the goal of the story. It is also where every character in the “A” story (the main story) is introduced or at least hinted at. Additionally this is the time when the screenwriter starts to hint every character’s tic, behavior, and flaws that needs to be addressed, showing why the hero will need to change later on. There could be scenes that present the hero

in his home, work, and "play" environments. Typically, the hero is presented in a comfortable state of stagnation or "inner death".

- **Catalyst (minute 12)**: A catalyst moment knocks the hero out of his or her "before" world that was shown in the set-up. The hero loses the safety of its current state.
- **Debate (minute 12-25)**: The debate section must answer some question about how to deal with the catalyst. Debate shows us that the hero declares, "This is crazy!" and is conflicted by the options to resolve the dilemma: "should I go?" "Dare I go?" "Stay here?" The best action will most likely involve overcoming an obstacle, and therefore will result in the beginning change in the hero's character.
- **Break into Two (minute 25)**: The events cannot draw the hero into Act Two. The hero takes an action because he wants something. The hero MUST proactively decide to leave the old world and enter a new world because he wants something. This is the point where we leave "the way things were" and enter into an upside down version of it. "The Before" and "The After" should be distinct, so the movement into "The After" should also be definite.
- **B Story (minute 30-55)**: It is a different story (such as a love story) where the hero deals with its emotional side, perhaps the hero is even nurtured, energized, and motivated. The B story carries the theme of the story, but in a different way with different characters. The characters are often polar opposites of the characters in Act One, the "upside down versions" of them.
- **Fun and Games (minute 30-55)**: This is where the hero explores the upside down world he/she has entered into. The "Fun and Games" scenes are purely for the enjoyment of the audience – depending on the genre, it could have action scenes (such as car chase); funny scenes, romantic scenes, etc. During "Fun and Games", we aren't as concerned with the plot moving forward and the stakes won't be increased here.
- **Midpoint (minute 55)**: This is where the fun and games are over. The midpoint is where the stakes are raised (no turning back) so that it's either a (false) victory where the hero thinks that everything is fixed and he obtained his goal; or it seems like a (false) defeat for the hero. Sometimes a public display of the hero (such as in a big party). Our hero still has a long way to go before he learn the lessons that really matter.
- **Bad Guys Close In (minute 55-75)**: The (internal or external) forces that are aligned against the hero tighten their grip. As an opposite of the midpoint, If the hero had a false victory in the midpoint and the bad guys seem temporarily defeated, it is during Bad Guys Close In that the bad guys regroup and the hero's overconfidence and jealousy within the good guy team start to undermine all that they accomplished. This is because the hero hasn't fully learned the lesson he or she is supposed to learn, and the bad guys haven't completely been vanquished. As a result, our hero is headed for a huge fall. If it was a false defeat in the midpoint, now the here is hope.
- **All Is Lost (minute 75)**: The hero losses what he wants and feels the smell of death (or defeat). Most often, it is a false defeat. All aspects of the hero's life

are in a mess. If the midpoint was a false victory, then this is the low point for the hero when he or she has no hope.

- **Dark Night of the Soul (minute 75-85)**: This is the darkness night before the dawn, when the hero is forced to admit his or her humility and humanity, yielding control to "fate" or to "the universe". It is just before the hero digs deep down and pulls out that last best idea that will save the hero and everyone else. However, at this very moment this idea is no where in sight.
- **Break into Three (minute 85)**: The hero takes an action because he needs some- thing. At this point both the A story (which is the external, obvious story) and the B story (the internal, emotional story) meet and intertwine. The characters in the B story, the insights gleaned during their conversations discussing the theme, along with the hero striving for a solution to win against bad guys all comes together to reveal the solution to the hero. The hero has passed every test, dug down to find the solution. Now he or she just needs to apply it.
- **Finale (minute 85-110)**: The bad guys are defeated in ascending order, meaning that first the weaker enemy loose, then the middle men, and finally the top enemy. The source of "the problem" must be completely and absolutely defeated for the new world to exist. It is more than just the hero winning; the hero must also change the world.
- **Final Image (minute 110)**: In a happy conclusion, the final image is the opposite of the opening image and acts as proof that change has happened in the hero, and that the change is real. The B-story is resolved. In a sad ending, the hero rejects the change.

For example, in the movie *Silver Lining Playbook*[2], in the *Opening Image* (minutes 1-2) we see the main male character *alone* in a Psychiatric Facility, talking to his absent wife. In the *Midpoint* scenes (minute 58) we see the main male character dancing for the first time with the main female character, then he runs away and we later see him on his bed, in a turmoil of desire and guilt. In the *Final Image* (minutes 114-116) the whole family is back together in the house. The love of the main male and the female characters has not completed them alone; it has completed the greater family circle.

4 The MND Dataset

4.1 Data Collection- Movies and Scenes

We constructed a labeled dataset of scene categories (story beats) to facilitate the use of supervised machine learning algorithms. We limited our movie selection to the 51 movies used in the MovieGraphs dataset [16]. This dataset was constructed for the purpose of understanding human interactions in movies [24], and therefore, it is heavily biased towards realistic stories (many romantic comedies while almost no horror movies, fantasies or science fiction movies). The reason why we chose to use

[2] https://savethecat.com/beat-sheets/the-silver-linings-playbook-beat-sheet

this dataset is that it offers rich and accurate manual annotations of low-level movie features, such as shot/scene splitting, character identification, character attributes and actions etc., which are difficult to extract automatically and accurately with the current video-processing techniques. Our previous studies that used state-of- the-art video analytics software such as Microsoft's *Video Indexer*[3], suffered from errors in identifying characters, especially in darkly illuminated scenes, and our previous work showed the negative consequences of using fully automated, but inaccurate, scene splitting software for narrative understanding. Therefore, as a preliminary study, we base our work on the MovieGraphs dataset, use its provided scene boundaries and extract features using its high-quality annotations. We hope that in the near future, more accurate video processing tools will be developed, allowing a significant extension of this dataset with a minimal manual effort.

We had to discard 6 movies, either because we could not obtain the same version of the movies used in the MovieGraphs dataset, or because of an irrelevant movie style from the plot point of view (e.g. a biographical movie), and eventually labeled 45 movies contains 6,448 labeled scenes. Nine of the movies have their "gold-standard story beats" summarized by professional writers/scriptwriters who are proficient in the Save the Cat!® theory[4]. The gold standard story beats are used to evaluate the quality of the collected labels.

4.2 Collecting Story Beat Labels for the Scenes

We received the departmental ethics committee approval for using student volunteers as annotators. Each volunteer, which completed its task, received two bonus grade points for a data-science class. We selected 119 human annotators (out of 180 applicants, all senior undergraduate students in the Information Systems Engineering Department), based on their English proficiency level and their level of interest in watching movies (refer to the supplementary information for more detail). During the annotation process, we ensured that: (1) each annotator was assigned at least 3 movies (including one of the 9 movies with "gold-standard" annotation); and (2) each movie was annotated by at least 5 different annotators (in practice, except for one, all movies were assigned to 6 or more annotators). The annotators were provided with the guidelines that described in detail the background concepts, definitions of story beats, required workflows and accepted criteria. They were asked to choose one most appropriate story beat for each scene, out of 16 options described in Table 1. The annotation experiment was performed on the *Moodle* teaching platform using the H5P interactive video application[5]. The annotators were allowed to jump forward and backward without limitation so that they can skip any scene at first and label it later if needed. To evaluate the annotator's attention during the task, we used the knowledge about

[3] https://azure.microsoft.com/en-us/services/media-services/video-indexer/

[4] The "gold-standard story beats" are provided in the manner of movie deconstruction articles, an example can be found here: https://savethecat. com/beat-sheets/the-silver-linings-playbook-beat-sheet

[5] https://h5p.org/documentation

the characters which participate in each scene to automatically generate ten simple quizzes about the participation of a specific character in a specific scene. The quizzes were inserted at the end of randomly selected scenes in each movie.

For choosing the single best scene label from the annotations, we follow a two steps process: a) Select the scene label which received the most votes; b) If tie, use the time dependent label distribution (Figure 1 right) to select the label with highest likelihood from the tied ones.

4.3 Dataset Analysis

Evaluation Metrics: Watching and labeling a full length movie is a tough commitment for the crowdsourcing workers. We measure the quality of the collected labels in three ways: (1) *Fleiss' Kappa* [33,34] was used to measure the inter annotator agreement; and (2) Visualization of the labels distribution along the normalized movie duration time were used to verify the compatibility of the collected label with the theory; (3) the similarity between the collected story beats and the gold-standard story beats. A summary of the statistics of the dataset is presented in Table 2

Table 2. The Movie Narrative Dataset Summary

	Min	Max	Avg.	Median
Movie duration (hh:mm:ss)	01:35:31	02:47:59	02:03:15	01:58:06
# of scenes	51	279	143	142
Scene duration (mm:ss)	00:01	11:10	00:45	00:37
Kappa	0.11	0.67	0.25	0.24

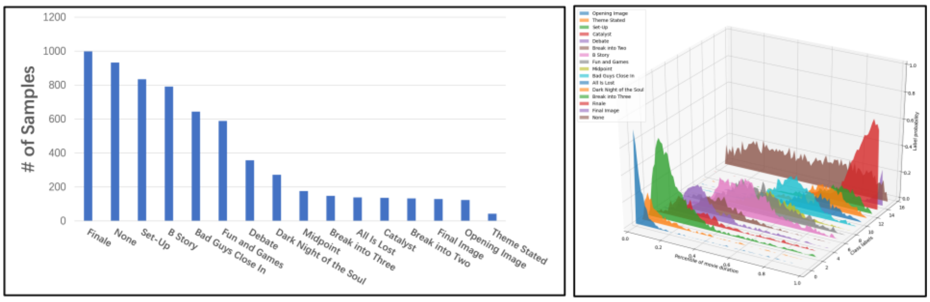

Fig. 1. Label distribution of the MND dataset**. Left**: label histogram. **Right**: the temporal label distribution over all movies. On the class label axis, each tick represents a label: 0 - Opening Image, 1 - Theme Stated, 2 - Set-Up, 3 - Catalyst, 4 - Debate, 5 - Break into Two, 6 - B Story, 7 - Fun and Games, 8 - Midpoint, 9 - Bad Guys Close In, 10 - All Is Lost, 11 - Dark Night of the Soul, 12 - Break into Three, 13 - Finale, 14 - Final Image, 15 - None.

Considering the graphs of the normalized temporal label distributions (Figure 1, right), the peaks of the distributions correspond fairly well with what is expected from Table 1 (e.g., the *Midpoint* is expected at about 55/100=0.50 of the movie).

For each movie, we check if there exists an “outlier” annotator and remove this annotation to increase the agreement. An “outlier” annotator is defined as the annotator whose annotation reduces the agreement the most. For example, the movie *Jerry Maguire* had an initial Kappa of 0.09 (6 annotators) and after removal of the outlier annotator, its Kappa increased to 0.12. The Kappa score presented in Table 2 is the improved score after the outlier removal step. The movies *Dallas Buyers Club* and *Pulp Fiction* obtained the highest Kappa (0.67 and 0.54, respectively), while the movies *Forrest Gump* and *Ocean's Eleven* obtained the lowest Kappa of 0.11. Considering the large number of scenes, categories, movies, and annotators, we consider median Kappa 0.24 as fairly high for such a subjective annotation task. The Kappa score corroborates the hypothesis that most movies adhere fairly well to the latent story structure described by the screenwriting book [15]. A low Kappa score may indicate a movie which is difficult to interpret or may not comply with the story model (e.g., no single main story such as *Crash*, or many *Flash Backs* such as *Forrest Gump*).

By comparing the collected story beats to the 9 movies with Gold-Standard story beats (i.e., evaluation by experts), we can evaluate how well the annotators understood their task. Figure 2 presents the visualization of label distribution along normalized movie durations for the gold standard labels (right) compared to the collected labels (left). Overall, the two visualizations are similar to each other in the shapes, locations and peaks of the label distribution, therefore we can infer that even non-experts can identify the scenes’ narrative function after watching a movie. Therefore, indicating a relatively good quality of the collected labels, and that the entire dataset is reliable and consistent.

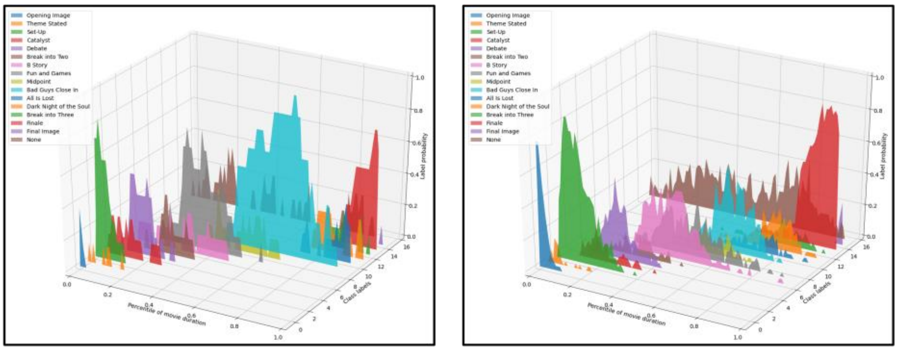

Fig. 2. Visualization of the label distribution for the 9 movies with gold standard story beats. Left: visualization of the gold standard story beats; Right: visualization of the collected story beats.

The main differences are a) The *None* category: the annotators classified many more scenes as *None*; b) more differences are in the labels that their definition is difficult to understand and their location in Table 1 is in a 20-25 minutes range: *Bad Guys Close In* and *Fun & Games* are selected more in the gold standard annotations. The distribution difference measures we used to compare the gold standard annotation with the non-expert annotation (KL-divergence, Bhattacharyya Distance and Earth Movers' distance) also indicate the difference in those labels

5 A MND Task: Movie Scenes Classification by their Narrative Function

In this section, we demonstrate that we can use the MND to automatically detect a movie's latent story structure (represented by the scene labels) using machine learning with some relatively simple movie features, outperforming a strong temporal distribution baseline.

5.1 Data Pre-processing

The MND dataset consists of 6,448 scenes from 45 full-length cinema movies. The scene boundaries are provided by the MovieGraphs dataset [16], and we use the scene boundaries in order to avoid unnecessary noise caused by scene splitting errors. In the dataset, we annotated one label (out of 16 labels introduced in 3) for each scene. Since there are some infrequent class labels, in order to solve the data imbalance problem, we applied label set reduction by merging 4 labels with other labels based on their definition and their temporal neighborhoods: *Opening Image* and *Theme Stated* are merged with *Set-Up*, *All Is Lost* is replaced by *Dark Night of the Soul*, and *Final Image* is merged with *Finale*. Eventually, each of the 6,448 scenes are labeled by one out of 12 labels. The label histograms before and after reduction are presented in Figure 3 (left-up and left-down). In the dataset we also all keep the raw annotations, for future research.

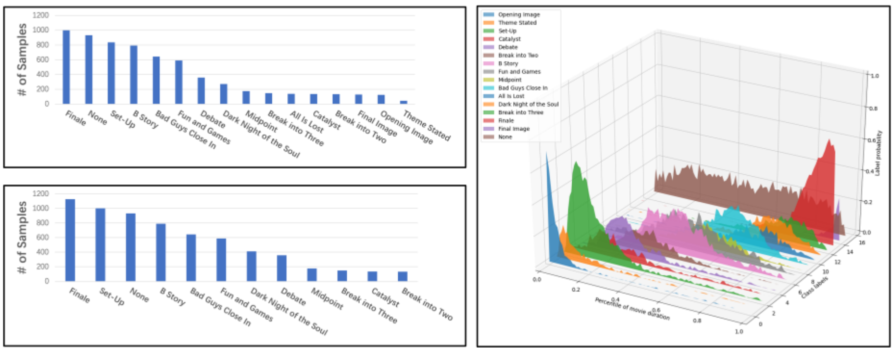

Fig. 3. Label **Reduction and temporal distribution of the MND Dataset.** Left-up: label histogram before reduction; Left-down: label histogram after reduction; Right: Temporal Distribution of the 16 labels (before reduction).

5.2 Feature Engineering

We use the same feature sets that were introduced in [25]. Specifically, we construct the following two sets of features: The basic feature set is inspired by the localized temporal distribution of most labels (Figure 3, right), and it contains 5 features that are computed from the position of the scene in the movie, were *i* is the scene counter. set [30,31] is used to capture the protagonist's importance in a scene. We expect the protagonist to participate in most scenes that advance the story – the story beats. The features are described in Table 3. The features' Information Gain (I.G.) column indicates that the most influential feature, *r_t_loc*, is the normalized position of the start of each scene in the movie. Considering that each story bead in Table 1 has a recommended position, this hints that many movies follow that recommendation.

Please note that the MND dataset contains potentially many more features than needed to develop a more elaborate and accurate model. For example, we could use the two-clock feature of [29,45] that attempts to detect turning point scenes. We could extract musical cues from the audio, lighting cue and cinematic cue (e.g., camera motions and close-ups) from the video [18,26] and possible use metadata information such as the movie genre. We wanted to develop a fairly simple model that may be used to benchmark further improvements in story analytics.

Table 3. Full feature list used for story-based scene classification on full length movies.

Feature Set	Name	Type	Description	I.G.
Basic	*dur*	float	scene duration (seconds)	0.81
	t	float	start time (seconds)	0.76
	close_beg_id	float	proximity to the beginning, 1/*i*	0.83
	r_id_loc	float	*i*/*n*, *n* is number of scenes in the movie	1.75
	r_t_loc	float	*t*/*len*, *len* is the duration of the movie	4.40
Character network	*protagonist appear*	bool	1: protagonist appears in this scene	0.99
	average scores	float	the average character scores	0.95

5.3 Baseline Approach

We present two simple baseline approaches: (1) *Majority rule* and (2) *Maximum likelihood* labeling (temporal label distribution baseline). The majority rule baseline is simply labeling all scenes as the most frequent label (e.g., a single label). The maximum likelihood labeling baseline is computed for each given scene based on its normalized location in the movie. In Figure 3 (right), we present the label temporal distribution along the movie time, which is used for the maximum likelihood labeling baseline. Specifically, for a given scene, (e.g., starts at minute 29) we firstly computed the percentile of movie time (*t*) this scene appeared (e.g., 24%), and from the demonstrated distribution, select the label with maximum probability at *t* percent. If the rare labels are selected, they are replaced by the more frequent labels described above.

5.4 Classification Experiment and Baseline Results

We applied five-fold cross validation approach over the Movie Narrative Dataset. In each validation fold, we keep 9 movies for testing and 36 movies for training. We used XGBoost [35] as the classification algorithm and typical evaluation matrices (precision, recall, accuracy and F1 measure) were used for quantitative evaluation. The XGBoost classifier won most of the recent data-mining competitions before the introduction of Deep Neural Networks. The implementations of the algorithms used in this work are based on the distributed Python implementation of XGBoost, and scikit-learn, a widely used machine learning library for Python. We used the default parameter settings provided by the implementations.

As presented in Table 4, the maximum likelihood labeling baseline significantly outperformed the majority rule baseline, and our classification approach with XGBoost classifier and the features introduced in Table 3 improved the accuracy by 0.03 and the F1 measure by 0.05. The improvements are tested to be statistically significant by *t-test*. Although the accuracy results on this small dataset are still low, they indicate that (1) The idea of automated scene classification in full-length cinema movies is feasible; (2) The proposed prototype approach and suggested features can be useful for various, more complex story models. We can expect better performance with more advanced feature engineering and larger amounts of annotated data.

Table 4. Baseline classification results on the MND dataset.

Methods	Precision	Recall	Accuracy	F1
Majority Rule	0.14	0.08	0.17	0.02
Maximum Likelihood Labeling	0.26	0.27	0.45	0.26
XGBoost	**0.31**	**0.34**	**0.48**	**0.31**

We further experimented with the *two-clocks* feature of [45] and with more character co-occurrence network features [30]. They did not contribute significantly to the classification performance.

We explored the possibility that different movies in our dataset belong to different story structures, thus reducing the performance of the classification algorithm. Considering the temporal label distribution presented in Fig 3 (right), we observe that there exist some story elements with a wide temporal distribution (such as *B-Story* and *Bad Guys Close In*). Possible reasons for such wide distributions might be (1) multiple label occurrences per movie (e.g., a story may have more than one *B-Story*); (2) difficult concepts (e.g., the annotators had a hard time understanding the concept of *Theme Stated*), and (3) a movie not conforming to the story model (some movies with very low Kappa scores indicate that the annotators are confused, meaning that the movie itself may not fit our story model well). It might be that the classification performance would improve if the confusing classes were removed from the label set and only the most critical elements were kept (such as *Inciting incident*, and *Debate*), however, label removal or merging might obscure some of the fine elements necessary for understanding the movie.

6 Conclusion and Future Research

In this paper, we defined a novel task for movie analytics: movie scene classification by their narrative function. It is an important step towards understanding the latent story structure within narrative videos such as movies, TV series or animated cartoons. We constructed a novel benchmark labeled scene dataset, the Movie Narrative Dataset (MND). From the manually annotated scene/shot boundaries and character identifications, provided in the MovieGraphs dataset, we constructed two sets of features for 45 movies. The extracted features include the basic information about the movie itself (such as duration, number of scenes etc.), character network features and temporal character appearance features. The features represent the aspects of the movie stories from different angles. The classification and feature selection experiments demonstrated the use of machine learning algorithms for the scene classification task. The evaluated algorithms were able to discover the sequential character of the key elements in the story model we used, which has been further verified by the temporal label distribution baseline and results with the basic feature set. The scene classifiers can serve as a benchmark for future research.

The value of scene classification by their narrative function for movie understanding is in *extracting the high level abstract concepts* associated with each story type, e.g., *Debate*, *All is Lost*. These can provide some automatic understanding about the protagonist's character traits and the motivations that drive him in facing his challenges. While the F1 measure of our classification model is relatively small, it might be enhanced with the addition of more elaborate features such as character emotions [46], character interactions [24], dialogue (subtitles) analysis and cinematic mood cues such as shots and camera movements, music and lighting [47]. As a first attempt to learn the story structure of a narrative video, we believe that this work has opened a promising direction for video story understanding.

Future research can naturally follow our results by (1) adding low level story-related features such as automatically detected characters, objects, actions, place, and emotions, etc.; (2) adding cinematic cue features such as shots and camera angle, illumination, music and voice; (3) generating a story-aware summary (in a video or text format) of a given narrative video, using the most important story elements and scenes, (e.g., following the work in [20]); (4) Use the MND in an attempt to detect even higher level narrative concepts such as the 36 dramatic situations of [28]. (5) utilizing story-related scene classification to boost the performance of other story-related tasks, such as movie question-answering; (6) exploring alternative, more elaborate movie story structures; (7) generating additional benchmark collections of story-based video annotations.; (8) constructing a fully automated pipeline that can process a video from start (e.g., scene cutting) to end (the detected story elements) and use it to annotate a large dataset such as [18].

Acknowledgement

This research was partially supported by the Israeli Council for Higher Education (CHE) via the Data Science Research Center, Ben-Gurion University of the Negev, Israel

References

1. Kristin T., Storytelling in the New Hollywood: Understanding Classical Narrative Technique Paperback – November 5, 1999
2. Field S. Screenplay: The foundations of screenwriting. Delta; 2007
3. Andrew D., Bazin A., (New York: Oxford University Press, 1978)
4. Bordwell D., Staiger, J.,;Thompson K., The Classical Hollywood Cinema. New York: Columbia University Press., 1985.
5. Iglesias K.,, “8 Ways to Hook the Reader,” *Creative Screenwriting* 13, 4 (2006), 48–49.
6. Iglesias K., Writing for Emotional Impact (Livermore, CA: Wingspan, 2005),
7. Bordwell, D. & Thompson, K., Film Art: An Introduction. New York: McGraw-Hill (2010).
8. Piper A., Jean So R., Bamman D., “Narrative Theory for Computational Narrative Understanding,” Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP). (2021)
9. Cutting J.E., Narrative theory and the dynamics of popular movies, Psychonomic Bulletin & Review, 23(6), 2016.
10. Valls-Vargas J., Zhu J., Ontañón S., Narrative Information Extraction with Non-Linear Natural Language Processing Pipelines, 2017, PhD thesis at Drexel University.
11. Arijon D., Grammar of the Film Language, Los Angeles : Silman-James Press ; Hollywood, CA : Distributed by Samuel French Trade, 1991.
12. Delmonte R., Marchesini G., A semantically-based computational approach to narative structure. In IWCS 2017—12th International Conference on Computational Semantics—Short papers.
13. Mikhalkova E., Protasov T., Sokolova P., Bashmakova A., Drozdova A., Modelling narrative elements in a short story: A study on annotation schemes and guidelines. In Proceedings of The 12th Language Resources and Evaluation Conference, pages 126–132, 2020
14. Wallace., Multiple narrative disentanglement: Unraveling Infinite Jest. In Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1–10.
15. Snyder B. Save the Cat!: The Last Book on Screenwriting You’ll Ever Need. Cinema/Writing. M. Wiese Productions; 2005. Available from: https://books.google.co.il/books?id=I1VjmAEACAAJ.
16. Vicol P., Tapaswi M., Castrejón L., Fidler S., MovieGraphs: Towards Understanding Human-Centric Situations from Videos. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2018. ISBN 9781538664209., doi: 10.1109/CVPR.2018.00895.
17. Huntley C., How and Why Dramatica is Different from Six Other Story Paradigms; July 2007 (accessed June 2019). Available from:

http://dramatica.com/articles/how-and-why-dramatica-is-different-from-sixother-story-paradigms.
18. Huang Q., Xiong Y., Rao A., Wang J., Lin D., MovieNet: A Holistic Dataset for Movie Understanding. In: Vedaldi A., Bischof H., Brox T., Frahm JM. (eds) Computer Vision – ECCV 2020. ECCV 2020. Lecture Notes in Computer Science, vol 12349. Springer, Cham. https://doi.org/10.1007/978-3-030-58548-8_41
19. Papalampidi P., Keller F., Lapata M., Movie plot analysis via turning point identification. EMNLP-IJCNLP 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference, pages 1707–1717, 2020. doi: 10.18653/v1/d19-1180.
20. Papalampidi P., Keller F., Lapata M., Film Trailer Generation via task Decomposition, 2021, https://doi.org/10.48550/arXiv.2111.08774
21. Garcia N., Nakashima Y., Knowledge-Based Video Question Answering with Unsupervised Scene Descriptions. In: Vedaldi A., Bischof H., Brox T., Frahm JM. (eds) Computer Vision – ECCV 2020. ECCV 2020. Lecture Notes in Computer Science, vol 12363. Springer, Cham. https://doi.org/10.1007/978-3-030-58523-5_34
22. Zhong Y., Wang L., Chen J., Yu D., Li Y., Comprehensive Image Captioning via Scene Graph Decomposition. In: Vedaldi A., Bischof H., Brox T., Frahm JM. (eds) Computer Vision – ECCV 2020. ECCV 2020. Lecture Notes in Computer Science, vol 12359. Springer, Cham. https://doi.org/10.1007/978-3-030-58568-6_13
23. Cao J., Gan Z., Cheng Y., Yu L., Chen YC., Liu J., Behind the Scene: Revealing the Secrets of Pre-trained Vision-and-Language Models. In: Vedaldi A., Bischof H., Brox T., Frahm JM. (eds) Computer Vision – ECCV 2020. ECCV 2020. Lecture Notes in Computer Science, vol 12351. Springer, Cham. https://doi.org/10.1007/978-3-030-58539-6_34
24. Kukleva A., Tapaswi M., Laptev I., Learning interactions and relationships between movie characters. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 9849–9858, 2020.
25. Liu C., Shmilovici A., Last M., Towards Story-based Classification of Movie Scenes. PLoS ONE, 2020. ISSN 19326203. doi: 10.1371/journal.pone.0228579.
26. Bain M., Nagrani A., Brown A., Zisserman A.; Condensed Movies: Story Based Retrieval with Contextual Embeddings, Proceedings of the Asian Conference on Computer Vision (ACCV), 2020, https://openaccess.thecvf.com/content/ACCV2020/html/Bain_Condensed_Movies_Story_Based_Retrieval_with_Contextual_Embeddings_ACCV_2020_paper.html
27. Cascante-Bonilla P., Sitaraman K., Luo M., Ordonez V., Moviescope: Large-scale Analysis of Movies using Multiple Modalities, https://doi.org/10.48550/arXiv.1908.03180
28. FiggisM., The Thirty-Six Dramatic Situations. Faber & Faber, 2017.
29. Chang Liu, Mark Last, and Armin Shmilovici., Identifying Turning Points in Animated Cartoons. Expert Systems with Applications, 123:246–255, jun 2019.

ISSN 09574174. URL https://linkinghub.elsevier.com/retrieve/pii/S0957417419300041. doi: 10.1016/j.eswa.2019.01.003.

30. Lee O.J, Jung J.J., Modeling affective character network for story analytics. Future Generation Computer Systems, 92:458–478, 2019.
31. Lee O.J, You E.S., Kim J.T.,. Plot structure decomposition in narrative multimedia by analyzing personalities of fictional characters. Applied Sciences, 11(4), 2021. ISSN 2076-3417. doi: 10.3390/ app11041645. URL https://www.mdpi.com/2076-3417/11/4/1645.
32. Tran Q.D., Hwang D., Lee O.J., Jung J.E.. Exploiting character networks for movie summarization. Multimedia Tools and Applications. 2017; 76(8):10357–10369. https://doi.org/10.1007/s11042-016-3633-6
33. Fleiss J.L. Measuring nominal scale agreement among many raters. Psychological bulletin. 1971; 76 (5):378. https://doi.org/10.1037/h0031619
34. Landis JR, Koch GG. The measurement of observer agreement for categorical data. biometrics. 1977; p. 159–174. https://doi.org/10.2307/2529310 PMID: 843571
35. XGBoost Python Package; last accessed August 2021. Available from: https://xgboost.readthedocs.io/ en/latest/python/index.html.
36. Evangelopoulos G, Zlatintsi A, Potamianos A, Maragos P, Rapantzikos K, Skoumas G, et al. Multimodal Saliency and Fusion for Movie Summarization Based on Aural, Visual, and Textual Attention. IEEETransactions on Multimedia. 2013; 15(7):1553–1568. https://doi.org/10.1109/TMM.2013.2267205
37. Rohrbach A, Torabi A, Rohrbach M, Tandon N, Pal C, Larochelle H, et al. Movie Description. International Journal of Computer Vision. 2017; 123(1):94–120. https://doi.org/10.1007/s11263-016-0987-1
38. Rohrbach A., Rohrbach M., Tandon N., Sciele B., A Dataset for Movie Description, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3202-3212, 2015.
39. Tapaswi M, Zhu Y, Stiefelhagen R, Torralba A, Urtasun R, Fidler S. MovieQA: Understanding Stories in Movies Through Question-Answering. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016.
40. Carreira J, Zisserman A. Quo vadis, Action ecognition? a new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2017. p. 6299–6308.
41. Ji J., Krishna R., Fei-Fei L., Niebles J.C.. Action genome: Actions as compositions of spatio-temporal scene graphs. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2020
42. Carlos A. Gomez-Uribe and Neil Hunt. The netflix recommender system: Algorithms, business value, and innovation. ACM Trans. Manage. Inf. Syst., 6(4), December 2016. ISSN 2158-656X. doi: 10.1145/2843948. URL https://doi.org/10.1145/2843948
43. Jeong A W.I., Soojin J., YoungBin KIM, Poster-Based Multiple Movie Genre Classification Using Inter-Channel Features IEEE Access PP(99):1-1, April 2020, DOI:10.1109/ACCESS.2020.2986055
44. Movie genre classification via scene categorization. MM'10 - Proceedings of the ACM Multimedia 2010 International Conference, pages 747–750, 2010. doi: 10.1145/1873951.1874068.

45. Lotker Z. The tale of two clocks. In: 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM); 2016. p. 768–776.
46. Mitta T.l, Mathur P., Bera A., Manocha D., Affect2mm: Affective analysis of multimedia content using emotion causality. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5661–5671, 2021.
47. Avgerinos C., Nikolaidis N., Mygdalis V., Pitas I., Feature extraction and statistical analysis of videos for cinemetric applications. 2016 Digital Media Industry and Academic Forum, DMIAF 2016 - Proceedings, pages 172–175, 2016. doi: 10.1109/DMIAF. 2016.7574926.
48. Sun, Yidan, Qin Chao, and Boyang Li., Synopses of Movie Narratives: a Video-Language Dataset for Story Understanding, arXiv preprint arXiv:2203.05711 2022.