Expert Systems With Applications 123 (2019) 246–255

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Review

Identifying turning points in animated cartoons

Chang Liu*, Mark Last, Armin Shmilovici

Department of Software and Information Systems Engineering, Ben-Gurion University, Beer Sheva, Israel

ARTICLE INFO

Article history:
Received 4 October 2018
Revised 31 December 2018
Accepted 2 January 2019
Available online 15 January 2019

Keywords:
Story's turning points
Story elements detection
Story understanding
Video analytics

ABSTRACT

Detecting key story elements such as protagonist, opponent, desire, turning points, battle, and victory, etc. is essential for various narrative work applications including content retrieval and content recommendation systems. The task of automatically identifying story elements is challenging because of its complexity and subjectiveness and currently, there are no available algorithms for this task. In this paper, we focus on identifying turning points in a story of a cartoon movie. The proposed methodology extends the novel two-clocks theory, originally validated on scripts of theatre plays, to video stories. The assumption behind the two-clocks theory is that the perception of time is different when some special event happens to a certain agent (e.g., time flows slower for a patient and quicker for a tourist). The story timeline is monitored with two clocks: an event clock, which measures the regular time flow of the story; and a weighted clock, which measures the timing of the story events. We have conducted an experiment on 28 episodes of a cartoon series and achieved promising results: 78.6% precision for turning points identification and 100% precision for key scene detection. The proposed approach is the first step towards development of intelligent systems for automated understanding of stories in narrative works such as cinema movies and even amateur videos uploaded to the Internet.

1. Introduction

With the widely spread of web accessibility and the development of video producing technologies, people are exposed to a massive amount of videos. Among them, many video narrative works (e.g., movies, TV series, cartoons, etc.) are made with shots and scenes presenting the plot of some story. Automated understanding of the stories told by such videos via analyzing the video content and structure can be beneficial for multiple tasks including video retrieval, video recommendation, and video annotation. This kind of analysis can also be utilized for educating students in film production or delivering preferable video content to users. It is natural for a human to perform video story analytics by watching the videos. However, the manual approach is time-consuming and not scalable to massive amounts of online videos as it contains repetitive efforts such as assigning annotation tags to the videos (Gomez-Uribe & Hunt, 2016; Soares & Viana, 2015). Therefore automatic computational methods are being developed for video content and story analytics. Unlike many of these methods, which concentrate on detecting detailed visual elements such as objects and actions or describing short video clips with simple sentences, we aim at developing methods to identify the key elements of a story in a video, such as the "hero" (the protagonist), the turning points, the roles of the characters, etc, and to understand how the elements advance the story. This paper focuses on detection of one key element – the turning points.

According to the widely used *three-act "Paradigm"*[1], a conceptual scheme of scriptwriting/story-writing, a good story is composed of three main acts and each of them plays a different role in the story (i.e., the set up, confrontation, and resolution) (Field, 2007), as shown in Fig. 1. There are different elements within the three-act structure, such as climaxes, midpoint, beginning, inciting incident, second thoughts, obstacles, disaster, wrap-up and end, which construct the main framework of a story. With the ultimate goal of understanding stories in videos, we start by identifying important story elements. In this paper, we build upon an innovative two-clocks theory from Lotker (2016) that is aimed at detecting one key event in a narrative work (e.g., a movie script or a theater play) to identify multiple turning points of cartoon stories. This is the first application of the two-clocks theory to

* Corresponding author.
E-mail addresses: liuc@post.bgu.ac.il (C. Liu), mlast@bgu.ac.il (M. Last), armin@bgu.ac.il (A. Shmilovici).

https://doi.org/10.1016/j.eswa.2019.01.003
0957-4174/

[1] The most notable contribution of the leading American screenwriter, Sydney Alvin Field. The structure of three-act was proposed in his first book *Screenplay: The Foundations of Screenwriting (Dell Publishing, 1979)*, and became popular among writers and Hollywood film producers as guideline and quality measurement.

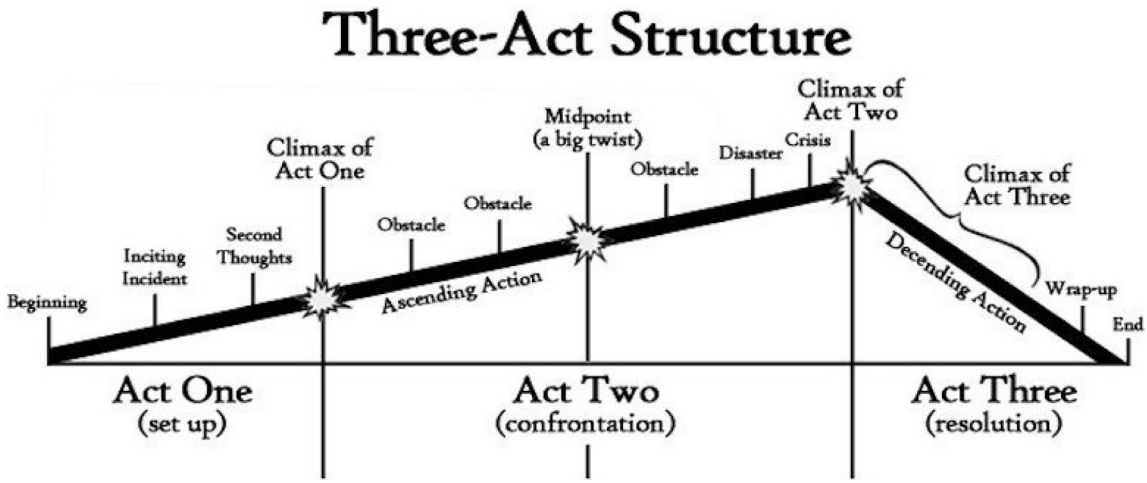


Fig. 1. Syd Field's *three-act "Paradigm"*.

Table 1
Elements validation of three episodes from the cartoon seires: *the Flintstones, Season 1*.

Elements	Episode-1	Episode-3	Episode-9	Episode-27
1st climax	✓	✓	✓	✓
2nd climax	✓	✓	✓	×
3rd climax	✓	✓	✓	✓
Beginning	✓	✓	✓	✓
Inciting incident	✓	✓	✓	✓
Second thought	×	×	×	×
Obstacles	✓	✓	✓	✓
Midpoint	✓	✓	✓	✓
Disaster	×	×	×	×
Crisis	✓	×	✓	×
Wrap-up	×	✓	×	✓
End	✓	✓	✓	✓

video analytics and this theory is demonstrated to be powerful for identifying turning points of a cartoon story by our experiments on a series of 28 animated cartoons of *the Flintstone Season 1*. In order to verify that the chosen cartoons follow the same story structure of three-act as well as the including elements, we arbitrarily chose four episodes (episodes 1, 3, 9 and 27) from this series and did elements identification. The results in Table 1 demonstrated that the analyzed cartoon stories generally follow the three-act structure with several elements missing (e.g., second thought, disaster, or wrap-up etc.), because the cartoon stories are not as long or complex as movie stories.

In this paper we present a prototype expert system for detecting key scenes in a movie. The proposed system is built upon integration of human perception of time studied by psychologists (Block & Grondin, 2014) and movie scriptwriting guidelines (Field, 2007). The proposed expert rules for detecting turning points and key scenes in a movie are based on these two knowledge sources.

Specifically, we extend *Lotker*'s two-clocks theory to video analytics and demonstrate evaluation results on the first season of the animated cartoon, *the Flintstones, Season 1* (28 episodes, around 24 min length per episode). Similar to what *Lotker* has done in his paper (Lotker, 2016) (i.e., detecting one key scene from a Shakespeare play), we conducted our experiments for identifying multiple turning points of the story in each cartoon episode. The rest of this paper is organized as follows: Section 2 introduces the two main trends in understanding video contents as well as the two-clocks theory from *Lotker*; Section 3 highlights the differences between our methodology and the original *Lotker*'s method, and provides the experimental design; Section 4 presents and discusses the evaluation results; Section 5 outlines further steps towards automated understanding of video stories.

2. Related works

Most works in video understanding are based on computer vision algorithms. Those algorithms perform well on basic video understanding tasks, such as recognizing actions in video clips (Peng & Schmid, 2016; Saha, Singh, Sapienza, Torr, & Cuzzolin, 2016; Sigurdsson, Divvala, Farhadi, & Gupta, 2016; Singh, Saha, & Cuzzolin, 2016) and generating captions for videos (Kaufman, Levi, Hassner, & Wolf, 2016; Rohrbach et al., 2017; Torabi, Pal, Larochelle, & Courville, 2015; Venugopalan et al., 2015). Most of these algorithms focus on analyzing short video clips (less than 30 s length), which makes them very suitable for exploring the detailed (or low-level) information in videos such as *"drinking"* or *"walking"* but very poor at understanding high-level events in those videos (*"enjoying a party"* or *"going home"*). This is likely caused by using the low-level features, e.g., optical-flow in the video frames (Varol, Laptev, & Schmid, 2017), VGG features (Sigurdsson et al., 2016) or the raw video RGB frames (Venugopalan et al., 2015). In addition to these low-level visual features, some works combined audio features (e.g., frequency spectrogram) for video understanding as well (Evangelopoulos et al., 2013; Lee, Abu-El-Haija, Varadarajan, & Natsev, 2018). Such visual and audio features are not enough to identify high-level activities, which need more abstract information such as emotions or intents. Moreover, there is a lack of quality labeled data. It is common to use supervised learning techniques with neural networks, on paired data of clip-action samples or clip-caption samples, for recognizing actions or generating captions respectively. In order to understand the high-level events, however, one needs such complex information as emotions (enjoy or disgust) or intents (stay or leave), which is much harder to be labeled. 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. As a complement to the computer vision algorithms, features from multiple modalities in a video (i.e., textual, visual, and aural) are utilized for applications such as movie summarization (Evangelopoulos et al., 2013), recommendation (Bougiatiotis & Giannakopoulos, 2018) and scene detection (Baraldi, Grana, & Cucchiara, 2017; Zhu & Liu, 2009). In terms of textual feature extrac-

tion, it is common to use typical methods such as word count (Baraldi et al., 2017), Bag-of-Words, topic modeling (Bougiatiotis & Giannakopoulos, 2018), and textual saliency (Evangelopoulos et al., 2013), then fuse them with features extracted from other modalities. Many efforts have been made to improve the quality of the fusion process. Though the additional information boosts the performance, the connection between the added features and the story development is still weak. In our work, we analyze the problem of video story understanding from a different angle, which is extracting knowledge representing the story structure, and thus making one-step forward towards an expert system that can understand a story like humans.

According to John Truby's 22 steps of scriptwriting[2], a good story is driven by the interactions between the hero and his opponents (Truby, 2008). As long as there are people involved in a story, there will be a social network among the involved human characters. In terms of a video narrative works (e.g., movies), the idea of movie character network is closer to human intuition than computer vision based algorithms. Weng, Chu, and Wu (2007) firstly attempted to build a character network for a movie based on the frequency of movie character's appearances in the movie scenes and detected sub-stories of the movie based on the character network. Later on, *Tran* and his team published a series of works on constructing character networks of movies (based on appearance time and co-occurrence of characters, named CoCharNet) (Tran, Hwang, Lee, & Jung, 2016; Tran & Jung, 2015) and the application of CoCharNet on movie summarization (Tran, Hwang, Lee, & Jung, 2017). Recently they solved several problems of story analytics including detecting turning points (Lee & Jung, 2018). They proposed the idea of affective character network, in which not only the appearance time of characters will be modeled, but also their affective relationships (modeled by emotion analysis). Such relationships are capable of reflecting the tension changes in stories, and the tension of the story is modeled as a function to the movie time. Based on the derivative of the tension to the movie time, they identified turning points of the movie story. They also detected communities among movie characters (a character community is a group of characters who have close or similar relationships to a leading character such as the hero's friends and families) and defined a story model, which reflects the structure of the movie story, using the affective character network.

As another attempt to identify key events in a narrative work, *Lotker* has proposed his theory of modeling a well known psychology concept *"Time perception"* as different clocks (Lotker, 2016). The general idea is that time flows differently in different situations. For example, when someone goes through a life threatening experience such as bungee jumping, the time flows much slower for the jumper than that for his spectators. Inspired by this, two clocks are modeled: one for the regular time (or the stage time) and another one for the time of each event. By identifying the "clock drift" between the two clocks, *Lotker* claims that the key events of a narrative work (e.g., a movie) can be detected. In Lotker (2016), his innovative theory is demonstrated to be effective on three Shakespeare play scripts. Here, we present the first attempt to extend it to the video analytics domain.

Compared to a previous attempt to identify turning points via changes in an affective character network (Lee & Jung, 2018), our approach is much simpler and does not depend on the characters' face detection, which may be challenging under poor lighting conditions prevailing in many movies.

[2] John Truby is an American screenwriter, director and scriptwriting teacher who is famous for his 22 steps theory for scriptwriting.

3. Methods

3.1. The two-clocks theory

As defined in Lotker (2016), the two clocks are: (1) *the event clock* – counting the character speeches (lines), and (2) *the weighted clock* – counting the number of words spoken by the character per speech. Specifically, the event clock is advanced by one tick every time a character speaks and the weighted clock is an accumulation of the number of words spoken by the characters every time. We use the same definition of two clocks in our implementation, formally:

Definition 1. Define $D = \{(d_1, \ldots, d_n) | d_i \in \mathbb{Z}, d_{i+1} - d_i = 1\}$ as the dialogue of a narrative work, which is a finite sequence of dialogue speech (line) ids, starting from $d_1 = 1$ and ending at $d_n = N$ where N is the total number of lines in the dialogue.

Definition 2. An event clock is defined as C_e: $[d_1, d_n] \rightarrow [e_1, e_n]$, and $C_e(i) = e_i = d_i$, i.e., an event clock is the sequence of dialogue line ids.

Definition 3. A weighted clock is defined as C_w: $[d_1, d_n] \rightarrow [w_1, w_n]$, and

$$C_w(i) = w_i = \sum_{i \in D} s(d_i)$$

where $s(d_i)$ is the number of words in the ith line d_i.

Lotker claimed that the key scene could be detected by looking for the biggest gap between these two clocks, i.e., the *"clock drift"*. The gap function is defined as:

$$\beta = argmax_{d_i} NC_w - NC_e \quad (1)$$

where NC_w and NC_e are the normalized weighted clock (C_w) and event clock (C_e), and β is the clock drift (min–max normalization is used). Note that both clocks range in a unit interval [0,1] after normalization, so the gap function actually finds the global maximum of the clock difference function $NC_w - NC_e$.

We extend the two-clocks theory to detection of key scenes in cartoon movies. The modifications we made to the original methodology are described in Section. 3.2.

3.2. Applying the two-clocks theory to cartoon movies

Several changes are made in our work comparing to Lotker (2016):

1. We use the movie subtitles as input instead of a play script.
2. We consider both the minimum and the maximum values of the difference between the clocks rather than the maximum values only.
3. We search for multiple turning points in a movie instead of a single one in a play.

Compared to the typical plays that have a small number of long scenes (lengthy conversations), the typical movies have large numbers of short scenes (brief conversations). This makes the stories of movies more complex and brings more dialogues. Considering that the scripts of many movies and other videos are not publicly available, whereas subtitles are much easier to find (even automatically generated by speech recognition tools), we take movie subtitles as input instead of scripts. In addition, the human audience watches a movie or a cartoon without reading the script, which may not accurately match the actual movie anyway, thus making subtitles an important and reliable source of information about the movie story. Moreover, we see our work as the first step towards understanding videos, which do not have written scripts at all. Therefore we choose subtitles as the input rather than scripts.

Table 2
Example metadata of *the Flintstones Season 1*.

Episodes	Duration	# of Responses	# of Scenes
1: The Flintstone Flyer	23:58	289	71

However, a subtitle file usually does not indicate the speaker of each line (i.e., one cannot know who says what by reading the subtitles only), while it is necessary to distinguish between the lines spoken by different characters in order to build the two clocks. To address this problem, one may preprocess a movie with automatic speaker identification tools. In our experiments, we solve this problem manually.

Based on the above definitions of the two-clocks theory and the processed subtitle files, we apply the two-clocks theory to an animated cartoon. We create the two clocks based on Definition 2 and 3 and normalize them into a unit interval [0,1]. Instead of detecting a single clock drift in a play as done by the original method Lotker (2016), we detect three clock drifts from each episode of the cartoons, by looking for the three local extrema of the clock difference function $NC_w - NC_e$. The intuition behind this choice is: for most of the stories, there is more than one turning point in the plots. According to the John Truby's 22 steps of scriptwriting, a good story should include three revelations and decisions that drive the whole story (Truby, 2008), and those revelations and decisions could be regarded as turning points of the story. Moreover, we assume that not only the maxima refer to the turning points but also the minima. In order to prove this assumption, we modify the gap function as:

$$\beta = argmax_{d_i} |NC_w - NC_e| \quad (2)$$

In order to detect three *"clock drifts"* from the two clocks, we apply the FindPeaks[] function in Mathematica[3] to the clock difference $NC_w - NC_e$.

4. Evaluation

4.1. Experiment design

Preliminary experiment was done in Lotker (2016) to detect one key scene in each of the three famous Shakespeare plays (Julius Caesar, Othello, Romeo and Juliet). In order to evaluate the two-clocks theory on the video analytics domain, we detected turning points in a cartoon series, *the Flintstones, Season 1*, which contains 28 episodes, with an average length of 23 min 45 s (excluding the part unrelated to the story, i.e., the beginning and ending parts of each episode). An example of the metadata of these series of cartoons is shown in Table 2, and the full metadata list of the first season is shown in Table 4, Appendix A. For each episode, we identified up to three candidate turning points. Scenes are the basic components of a cartoon, and therefore we referred the detected turning points to scenes instead of dialogue lines. In other words, one scene usually includes dialogue that has more than one lines, and the detection of any of these lines will be regarded as the same result, i.e., the scene it belongs to. In our experiments, each cartoon episode is split into scenes using an open source software, PySceneDetect[4]. In total, we obtained 1897 scenes from the 28 episodes, with an average of 68 scenes per episode.

Understanding story elements is a challenging and subjective task and it is difficult to find any benchmark annotated dataset of significant size. The Flintstones episodes contain relatively complex human-oriented stories, yet, the number of characters (and their voices) is small, (making it relatively easy for future automatic preprocessing for video analytics). Until the automatic preprocessing tools will be accurate enough to preprocess the cartoons, we asked a human evaluator to judge the quality of our implementation. The evaluator was asked to watch the cartoons and identify up to three scenes, which were considered the ground truth turning points (at least one scene). Then, we compared the turning points detected by our implementation to the ground truth, and considered both top-1 and top-3 accuracy for evaluating the quality. In addition to human evaluation, we also performed a random experiment on the same cartoon seasons as baseline results. Specifically, we chose three arbitrary scenes from each episode and calculated both top-1 and top-3 accuracy based on the randomly chosen scenes. In order to deeply explore our results, we asked the evaluator to watch the detected scenes that were not considered by him as turning points and decide whether they contain other key story elements. We also demonstrated the results of key story elements detection.

Table 3
Accuracy of turning points identification (a) and key story elements detection (b). We demonstrated both number of correct cases and the percentage accuracy.

(a) Turning points identification results

	Top-1		Top-3	
	correct	acc.(%)	correct	acc.(%)
Random	0	–	2	7.1
Our	8	28.6	22	**78.6**

(b) Turning points identification results

	Top-1		Top-3	
	correct	acc.(%)	correct	acc.(%)
Random	6	21.4	13	46.4
Our	20	**71.4**	28	**100**

4.2. Numerical results

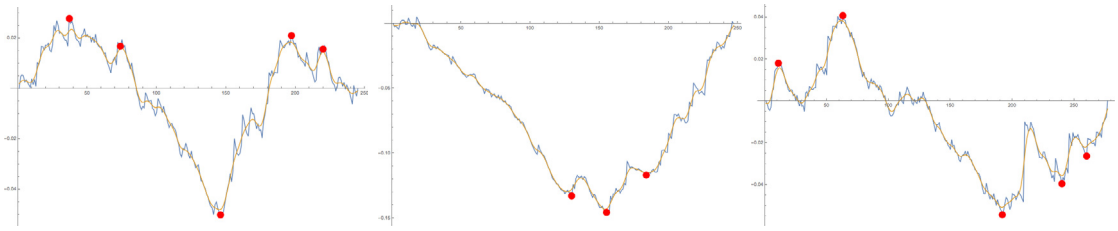
Our results of turning points identification and key story elements detection are shown in Table 3a and b, respectively. The accuracy was defined as the percentage of the number of correctly identified cases to the number of all cartoon episodes (i.e., 28). Specifically, for turning points identification (Table 3a), the random selection failed on all episodes for top-1 scene and successfully identified turning points from only 2 episodes considering top-3 scenes. On the other hand, our algorithm, based on the two-clocks theory, successfully found turning points in 8/22 episodes out of 28, considering top-1/top-3 candidates, respectively. Our experiments obtained the high accuracy of 78.6% with top-3 candidates for the task of turning points identification.

We also asked the evaluator to answer the question: for the detected scenes that are not considered as turning points, whether they correspond to other key story elements? The numerical results are shown in Table 3b (the turning points are considered as important scenes). Surprisingly, our algorithm can always find something meaningful considering the top-3 candidates (i.e., 100% accuracy), and even the top-1 accuracy reached 71.4% (20 episodes out of 28). Another impressive result was observed when we considered the scene level accuracy (i.e., the number of key story element scenes divided by the number of total detected scenes): we achieved 71.4% and 72.0% scene level accuracy for top-1 and top-3 candidates. The detailed detection results and ground truth turning points are presented in Appendix C.

4.3. Discussion

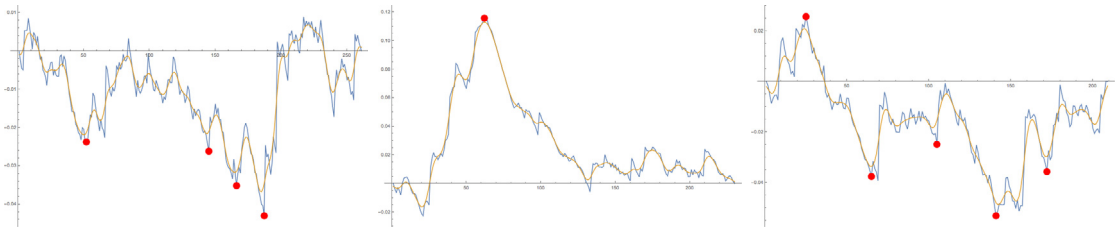
We visualized the clock difference using the *Second Moment Clocks Diagram* defined in Lotker (2016), where the x-axis is the

[3] https://reference.wolfram.com/language/ref/FindPeaks.html.
[4] https://pyscenedetect.readthedocs.io/en/latest/.



(a) Episode 3 (b) Episode 10 (c) Episode 18

Fig. 2. Success examples. The graphs of clock difference to the responses ids.



(a) Episode 8 (b) Episode 24 (c) Episode 28

Fig. 3. Failure examples. The graphs of clock difference to the responses ids.

first clock (the responses id) and the *y*-axis is the difference between the two normalized clocks (i.e., $NC_w - NC_e$). Because the two clocks are both normalized to [0,1] so that the above difference function always starts at 0 and ends at 0. This property of the diagram helps to easily detect the time when the two clocks have large difference and thus turning points could be identified. Fig. 2 shows three success examples from our experiment. In both Fig. 2a and b, the highest peaks/deepest valleys are exactly the turning points of both stories. There are some cases where the turning points are not detected as the highest peaks (deepest valleys) but as one of the top-3 peaks/valleys, such as Fig. 2c. As discussed before, a good story sometimes includes more than one turning points, and this is observed in the 3rd episode, as shown in Fig. 2a, where two turning points are identified in the top-3 peaks/valleys.

On the other hand, there are 7 out of 28 episodes where the turning points are not identified within the top-3 peaks/valleys (considered as failure cases), and three of them are shown in Fig. 3. For episode 24, as shown in Fig. 3b, the algorithm found only one peak from the graph, which is not the turning point. However, even though the algorithm sometimes failed to find the turning points, its discoveries are still meaningful. As an example, we will deeply explore the detection results of episode 27.

For all episodes in *the Flintstones* cartoons, there are story summaries available on Wikipedia[5]. For episode 27 (titled: *Room for Rent*), the story summary is:

> *Tired of hearing their husbands complain about finances, Wilma and Betty rent rooms to student musicians. Fred and Barney go along with the arrangement, unaware that their wives are providing the lodging in return for music and dancing lessons.*

Note that in the summary, Wilma, Betty, Fred and Barney are the four main characters throughout the entire cartoons; Wilma is Fred's wife and Betty is Barney's wife. We demonstrated the *Second Moment Clocks Diagram* graph of episode 27 annotated with all 7 peaks/valleys detected by the algorithm as well as the roles of their corresponding scenes in the story (the roles follow the definition in the *Three-act "Paradigm"* in Fig. 1), as shown in Fig. 4. The descriptions of the detected scenes can be concluded as follows (produced by a human evaluator):

1. **[Inciting incident] Scene 5:** The wives first time meet the student musicians and agree to rent them their rooms in return for music and dancing lessons.
2. **[Inciting incident] Scene 18:** The husbands come up with the same idea of renting their spare rooms for extra income before meeting their wives.
3. **[N/A] Scene 19:** This scene has no contribution to the story. It shows the interaction between Fred and his pet.
4. **[Obstacle] Scene 23:** The conflicts between the husbands and the students.
5. **[Midpoint] Scene 25:** The music lesson given by the student musicians. This scene is a midpoint because this lesson directly helped the wives to win 500 dollars from a music competition after the students left. In other words, the plot in this scene is a direct cause of the story's ending.
6. **[Turning point] Scene 30:** The students are leaving their rooms. This is the turning point because this is exactly when the husbands found their wives did not take money from the students and the wives told them the agreement about the music lesson. After this scene, the story moved to the resolution act (in Fig. 1) and then follows the ending.
7. **[Wrap-up] Scene 31:** The wives win the music competition. This scene is the preparation of the story ending.

Except for the 19th scene which is irrelevant to the main story plots, the other six scenes have their own contributions to the main story, and our algorithm successfully detected these story elements. This is impressive because the discoveries are made by a single algorithm, which is close to the fact when a human is watching a cartoon and understand the story. More examples of such analytics are demonstrated in Appendix B.

The conducted experiments show that the two-clocks theory is applicable to identifying turning points in various types of narrative works (such as movies), in addition to the originally evaluated Shakespeare play scripts. Thus, it appears that the two-clocks approach resembles the human perception of the time flow: the time flows differently in different situations. Moreover, our initial results indicate that the theory is able to discover more story elements than just turning points, e.g. inciting incidents, obstacles, midpoints, etc., which are defined by "the three-act structure". Based on our initial experiments, we expect the proposed approach to detect key events in professional movies, which do not necessar-

[5] https://en.wikipedia.org/wiki/List_of_The_Flintstones_episodes.

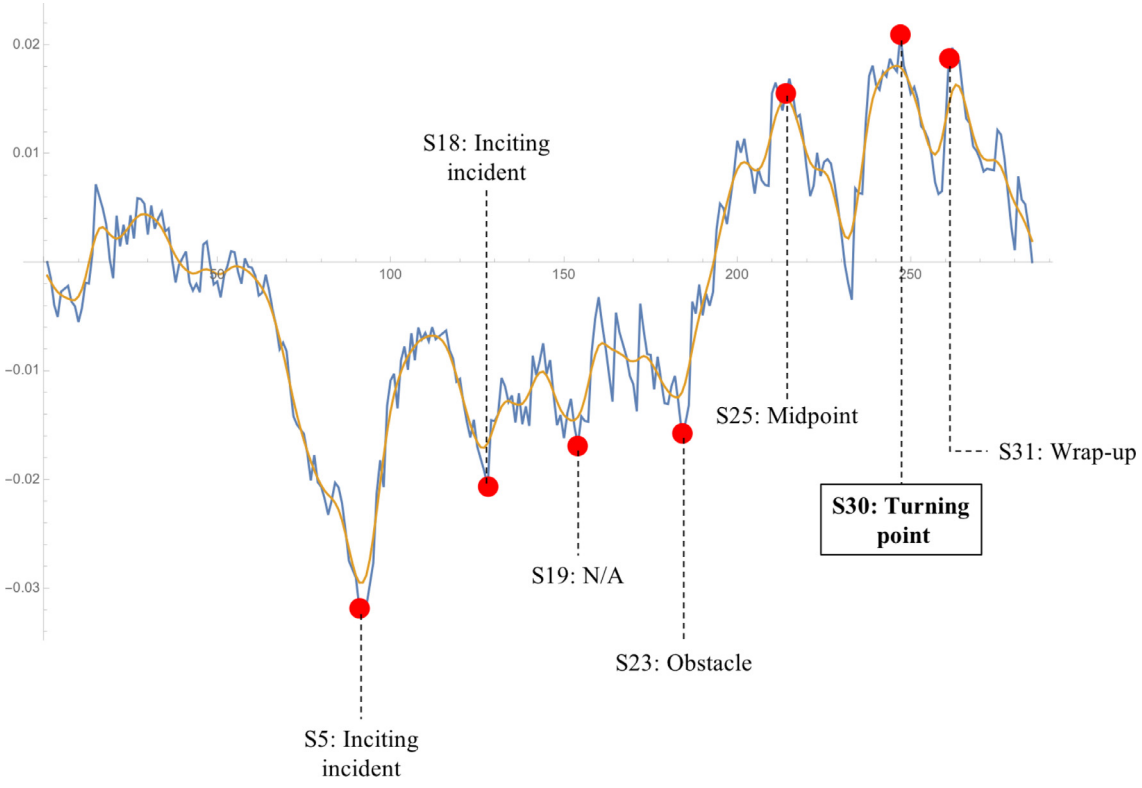


Fig. 4. The *Second Moment Clocks Diagram* of episode 27. 7 peaks/valleys are found for this episode, and their roles as story elements are marked.

ily adhere to the "three-act" structure, and even in amateur videos. This implies the potential of the proposed approach for movie understanding applications such as movie recommendation and summarization systems.

However, our current implementation of the two-clocks theory has some limitations: (1) Using only subtitles content is simple to implement, yet it ignores important visual and musical cues. In movie genres such as action or adventure, it is common to see intensive fast-paced scenes such as fighting or escaping from danger where nobody talks and thus no subtitles are available. (2) Another limitation is our assumption that the story timeline is linear. When there is a twist in the story timeline (i.e., the movie temporarily "flashbacks" to a scene that happened in the past), the algorithm may fail (this situation happened in our experiment because of a timeline twist in the 11th episode, see Table 5).

5. Conclusion

The main motivation of this work is to detect turning points (one of the story elements) as a first step towards understanding the entire movie. We adapt *Lotker*'s two-clocks theory to movie analytics and conduct evaluation experiments on 28 different stories. The demonstrated results show that the proposed approach is effective in identifying turning points and even additional story elements in a movie such as the starting point of the main story, using very limited information from the movie, i.e., the change of dialogues and the number of spoken words.

We differentiate our work from the related works in the following aspects:

1. Previously, the two-clocks theory has only been examined on three Shakespeare play scripts (Lotker, 2016). However, in this paper, we present a prototype expert system built upon the two-clocks theory and demonstrate its promising performance on a different type of narrative works, cartoon movies. With further evaluation on large-scale cinema movies, we believe that the outcomes of the proposed system can be useful for related applications such as movie summarization and recommendation.
2. Unlike the classic methods for text feature extraction (such as Bag-of-Words, topic extraction, text saliency scores etc.) used by related works (Baraldi et al., 2017; Bougiatiotis & Giannakopoulos, 2018; Evangelopoulos et al., 2013), our paper offers a novel way of extracting knowledge from the subtitles (or scripts) of a video by representing the subtitle text by its key story elements. The proposed system detects movie scenes, which contain turning points and other key story elements, thus bridging the gap between the raw data (subtitles) of a video and its story structure.
3. The proposed system is easy to implement and it is applicable to various types of narrative works, compared to the previous attempts of analyzing movie stories using affective character networks (Lee & Jung, 2018).

Some potential future directions of this work are: (1) The two-clocks theory could be expanded to other modalities (visual, audio etc.). For example, instead of constructing clocks based on subtitles and word counting, one can use the appearance times of characters in a scene and the duration of their presence, which are visual information, for clock construction. (2) Classify key story elements in a movie integrating the outcomes of the two-clocks theory with features extracted from other story modalities (such as the affective character network of Lee & Jung, 2018); (3) Combine (1) and (2) for intelligent applications such as story-based automated movie summarization and recommendation; (4) Adapt the approach to different types of narrative works.

Appendix A. the Flintstones Season 1 metadata

Table 4
Metadata of *the Flintstones Season 1*.

Episodes	Duration	# of Responses	# of Scenes
1: The Flintstone Flyer	23:58	289	71
2: Hot Lips Hannigan	23:44	264	78
3: The Swimming Pool	23:01	245	63
4: No Help Wanted	24:00	225	25
5: The Split Personality	23:52	227	36
6: The Monster from the Tar Pits	23:46	249	72
7: The Babysitters	23:44	225	142
8: At the Races	23:50	261	84
9: The Engagement Ring	24:02	251	47
10: Hollyrock, Here I Come	23:46	247	43
11: The Golf Champion	23:33	233	106
12: The Sweepstakes Ticket	23:55	222	50
13: The Drive-In	23:56	218	116
14: The Prowler	24:00	244	92
15: The Girls Night Out	24:03	246	54
16: Arthur Quarry's Dance Class	23:47	264	57
17: The Big Bank Robbery	23:29	272	78
18: The Snorkasaurus Hunter	23:57	277	87
19: The Hot Piano	23:23	297	94
20: The Hypnotist	23:38	194	30
21: Love Letters on the Rocks	23:46	266	54
22: The Tycoon	23:17	265	60
23: The Astra' Nuts	23:37	233	41
24: The Long, Long weekend	23:45	230	52
25: The Dough	23:55	296	104
26: The Good Scout	23:39	187	65
27: Room for Rent	23:48	285	36
28: Fred Flintstone Before & After	23:56	209	60

Appendix B. Examples of detailed story analytics

We demonstrate three more examples to show the capability of the two-clocks theory. Similar to the discussion in Section. 4.3, the analytics of episodes 1, 3, and 9 are shown below.

B.1. Episode 1

Fig. 5 shows the *Second Moment Clocks Diagram* of episode 1 and the brief story summary on Wikipedia is:

On a Sunday, Fred fakes illness so he and Barney can get out of taking their wives to the opera, as the night coincides with a Bowling Championship. With the use of Barney's homemade prehistoric helicopter as a means of escape, the two then join their bowling team for the tournament. They almost get away with their scheme, until loose-lipped Barney gives away their night's activities by using a fake mustache he and Fred used earlier to try to trick their wives at the bowling alley.

The detected scenes can be described as follows:

1. **[Inciting incident] Scene 8:** The first time Fred sees Barney's homemade helicopter and he is injured when he tries to fly it and fails to land.
2. **[Turning point] Scene 11:** After decided to attend the Bowling Championship, Fred and Barney are told that they will take their wives to the opera as planned one month ago. This is regarded to be the first turning point because after this scene, the story turns into the main part where Fred faked illness to fool the wives so that he and Barney can go to the Bowling Championship.
3. **[Midpoint] Scene 37:** On their way to the opera, the wives decide to call their husbands in the intermission because they sympathize their husbands for being unable to join.

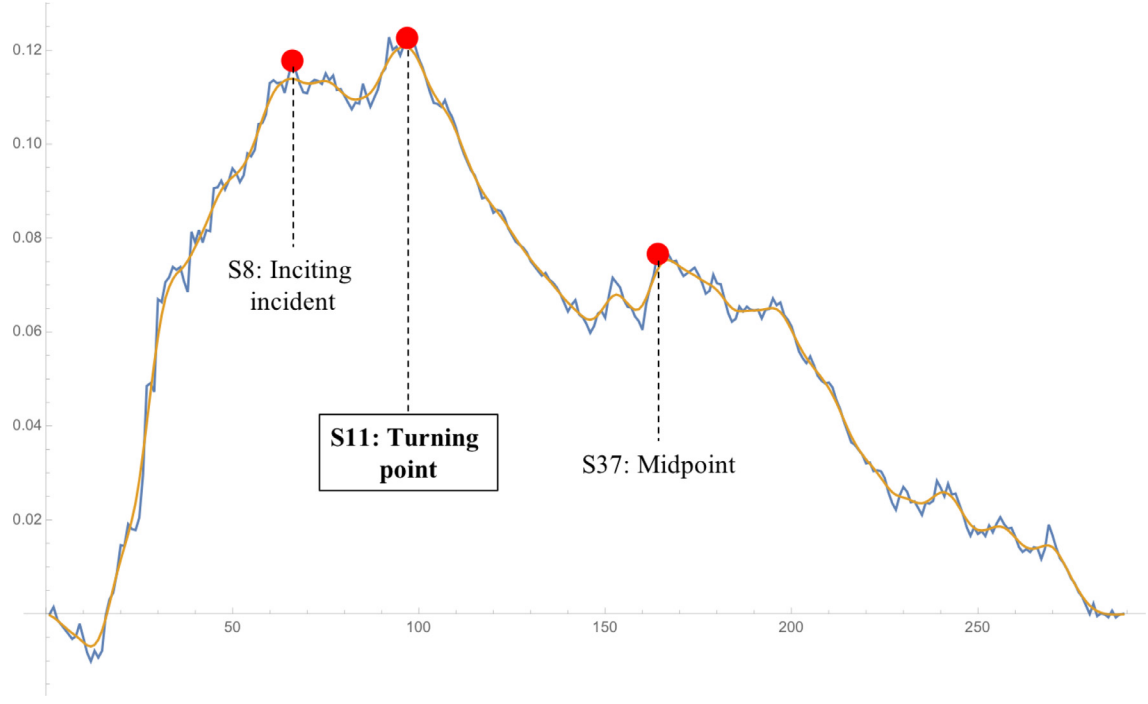


Fig. 5. The *Second Moment Clocks Diagram* of episode 1. 3 peaks are found for this episode.

B.2. Episode 3

Fig. 6 shows the *Second Moment Clocks Diagram* of episode 3 and the brief story summary on Wikipedia is:

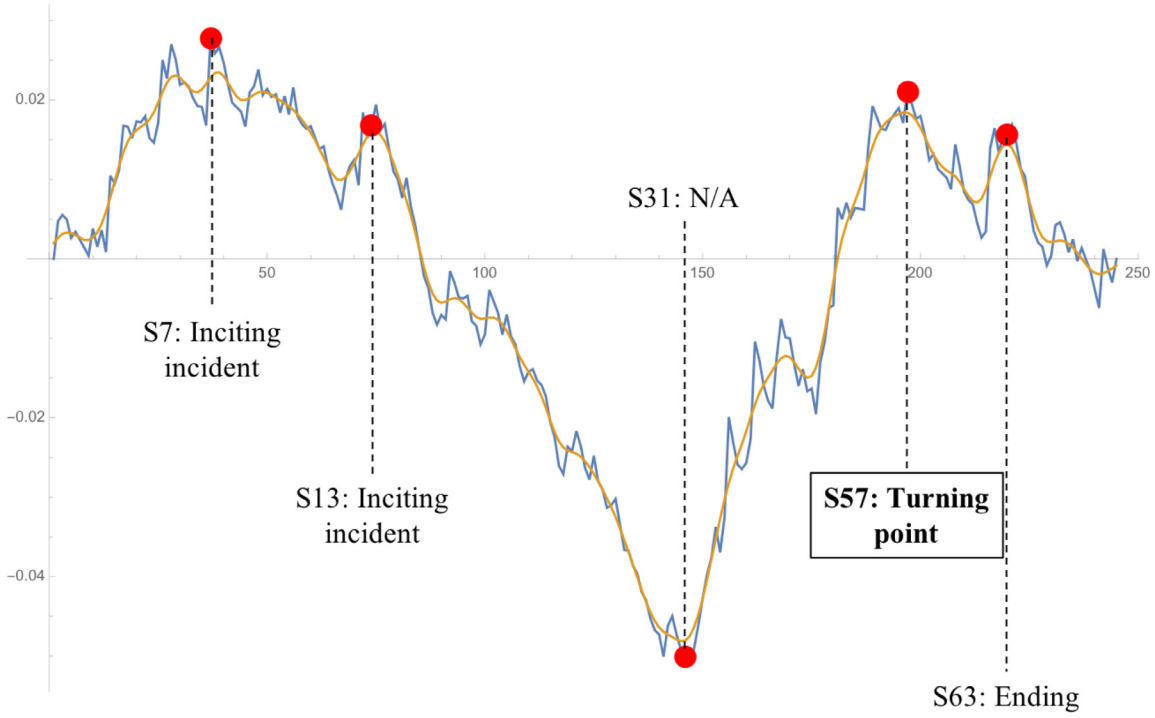


Fig. 6. The *Second Moment Clocks Diagram* of episode 3. 5 peaks/valleys are found for this episode.

Fred and Barney build a joint swimming pool in their backyards, leading to fights before Fred's surprise birthday party.

The detected scenes can be described as follows:

1. **[Inciting incident] Scene 7:** One of the fights between Fred and Barney.
2. **[Inciting incident] Scene 13:** Fred convinces Barney to build the swimming pool together.
3. **[N/A] Scene 31:** This scene has no contribution to the story. It is a irrelevant line spoken by an irrelevant character only for fun.
4. **[Turning point] Scene 57:** Fred opens the door and sees Barney with a birthday cake while singing a congratulation song. This is the turning point because after this scene Fred and Barney stopped fighting and enjoyed the birthday pool party together.
5. **[Ending] Scene 63:** They are too noisy during the pool party and one of their neighbors calls the police. Fred is arrested by the police, then Barney goes to the police office and helps Fred out.

B.3. Episode 9

Fig. 7 shows the *Second Moment Clocks Diagram* of episode 9 and the brief story summary on Wikipedia is:

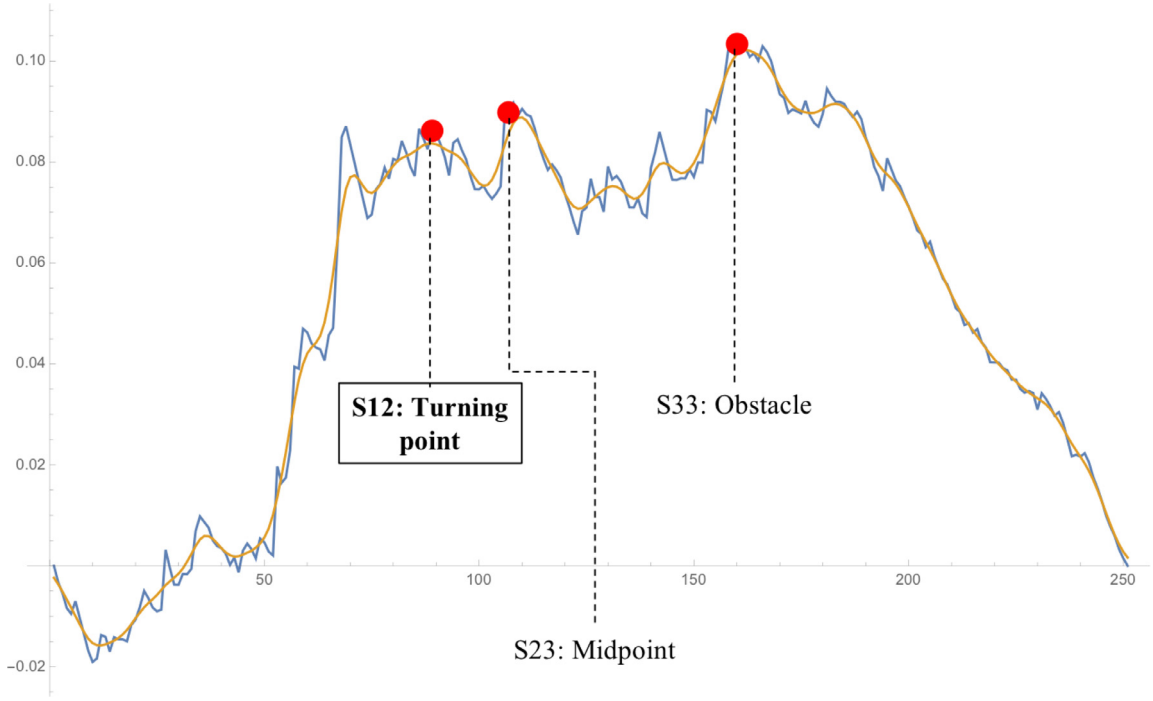


Fig. 7. The *Second Moment Clocks Diagram* of episode 9. 3 peaks are found for this episode.

Barney decides to surprise Betty with a belated engagement ring, which he gives to Fred for safekeeping but Wilma discovers the ring and assumes it is a gift for her. Not wanting to shatter her dream, Fred decides to buy a second ring but doesn't have the cash, so he cons Barney into going several rounds with a boxing champ in order to win a $500 prize.

The detected scenes can be described as follows:

1. **[Turning point] Scene 12:** Wilma coincidentally finds Barney's engagement ring and thinks that it is a surprise from Fred who dare not tell her the truth. This is the turning point because after this scene, the story turns into the main part where Fred and Barney tried to solve this problem.
2. **[Midpoint] Scene 23:** Fred and Barney come up with the idea that Fred will buy a second engagement ring for Barney to solve the problem.
3. **[Obstacle] Scene 33:** Considering Barney can never win the boxing champion, the wives try to convince the boxing manager to let Barney win the $500 prize which will be offered by the wives. The manager pretends to agree but actually intends to pocket the money by letting Barney lose.

Appendix C. Detailed results

Table 5
Detailed detection results of top-1 and top-3 candidates. The ground truth turning points decided by the human evaluator are shown as their scene ids in the second column. In the two *Detected Scene* columns, the scene ids of turning points are in square brackets and the ids of key story elements are in parentheses. The + and − in the *Result* columns indicates if there is turning points successfully detected (+) or not (−).

Ep.	Turning points	Top-1		Top-3	
		Detected Scene	Result	Detected Scene	Result
1	11, 58, 68	[11]	+	[11], (8), (37)	+
2	38, 45	[45]	+	[45], (61), (21)	+
3	13, 32, 57	31	−	31, (7), [57]	+
4	4, 10, 17	[10]	+	[10], (1), (20)	+
5	10, 33	(31)	−	(31), 17, [10]	+
6	56, 64	[64]	+	[64], 40, (30)	+
7	60, 111, 135	(97)	−	(97), (125), [60]	+
8	67	(64)	−	(64), 56, (49)	−
9	12, 32, 44	(33)	−	(33), (23), [12]	+
10	25, 36, 40	[25]	+	[25], (18), 26	+
11	91	(23)	−	(23), (68), 59	−
12	20, 35, 47	13	−	13, [35], (1)	+
13	24, 66, 103	[103]	+	[103], 70, (39)	+
14	44, 78	(56)	-	(56), 25, [44]	+
15	29, 44	[29]	+	[29], (34), (5)	+
16	28, 51, 54	[28]	+	[28], [51], 7	+
17	16, 59	(12)	−	(12), [59], 56	+
18	80	73	−	73, (27), [80]	+
19	79	(30)	−	(30), 41, [79]	+
20	16, 28	12	−	12, 5, [28]	+
21	10, 48	17	−	17, 35, [10]	+
22	38, 59	(13)	−	(13), 30, [38]	+
23	16, 31	12	−	12, (35), [31]	+
24	49	(12)	−	(12)	−
25	74, 98	29	−	29, (18), 9	−
26	55	61	−	61, (21), 36	−
27	22, 30	(5)	−	(5), [30], (18)	+
28	15, 55	(23)	−	(23), (5), (41)	−

Credit authorship contribution statement

Chang Liu: Data curation, Formal analysis, Investigation, Software, Validation, Writing - original draft, Visualization. **Mark Last:** Methodology, Writing - review & editing. **Armin Shmilovici:** Conceptualization, Writing - review & editing.

References

Baraldi, L., Grana, C., & Cucchiara, R. (2017). Recognizing and presenting the storytelling video structure with deep multimodal networks. *IEEE Transactions on Multimedia, 19*(5), 955–968. doi:10.1109/TMM.2016.2644872.

Block, R. A., & Grondin, S. (2014). Timing and time perception: A selective review and commentary on recent reviews. *Frontiers in Psychology, 5*, 648. doi:10.3389/fpsyg.2014.00648.

Bougiatiotis, K., & Giannakopoulos, T. (2018). Enhanced movie content similarity based on textual, auditory and visual information. *Expert Systems with Applications, 96*, 86–102. doi:10.1016/j.eswa.2017.11.050.

Evangelopoulos, G., Zlatintsi, A., Potamianos, A., Maragos, P., Rapantzikos, K., Skoumas, G., et al. (2013). Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention. *IEEE Transactions on Multimedia, 15*(7), 1553–1568. doi:10.1109/TMM.2013.2267205.

Field, S. (2007). *Screenplay: The foundations of screenwriting*. Delta.

Gomez-Uribe, C. A., & Hunt, N. (2016). The netflix recommender system: Algorithms, business value, and innovation. *ACM Transactions on Management Information Systems (TMIS), 6*(4), 13.

Kaufman, D., Levi, G., Hassner, T., & Wolf, L. (2016). Temporal tessellation for video annotation and summarization. arXiv.org.

Lee, J., Abu-El-Haija, S., Varadarajan, B., & Natsev, A. P. (2018). Collaborative deep metric learning for video understanding. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining KDD* (pp. 481–490). New York, NY, USA: ACM. doi:10.1145/3219819.3219856.

Lee, O.-J., & Jung, J. J. (2018). Modeling affective character network for story analytics. *Future Generation Computer Systems*. doi:10.1016/j.future.2018.01.030.

Lotker, Z. (2016). The tale of two clocks. In *Proceedings of the IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM)* (pp. 768–776). doi:10.1109/ASONAM.2016.7752325.

Peng, X., & Schmid, C. (2016). Multi-region two-stream R-CNN for action detection. In *Proceedings of the European conference on computer vision* (pp. 744–759). Springer.

Rohrbach, A., Torabi, A., Rohrbach, M., Tandon, N., Pal, C., Larochelle, H., et al. (2017). Movie description. *International Journal of Computer Vision, 123*(1), 94–120.

Saha, S., Singh, G., Sapienza, M., Torr, P. H., & Cuzzolin, F. (2016). Deep learning for detecting multiple space-time action tubes in videos. *CoRR* arXiv:1608.01529.

Sigurdsson, G. A., Divvala, S., Farhadi, A., & Gupta, A. (2016). Asynchronous temporal fields for action recognition. arXiv.org.

Singh, G., Saha, S., & Cuzzolin, F. (2016). *Online real time multiple spatiotemporal action localisation and prediction on a single platform* arXiv:1611.08563.

Soares, M., & Viana, P. (2015). Tuning metadata for better movie content-based recommendation systems. *Multimedia Tools and Applications, 74*(17), 7015–7036.

Torabi, A., Pal, C. J., Larochelle, H., & Courville, A. C. (2015). Using descriptive video services to create a large data source for video annotation research. *CoRR*. arXiv:1503.01070.

Tran, Q. D., Hwang, D., Lee, O.-J., & Jung, J. E. (2016). Exploiting character networks for movie summarization. *Multimedia Tools and Applications, 76*(8), 10357–10369.

Tran, Q. D., Hwang, D., Lee, O.-J., & Jung, J. J. (2017). A novel method for extracting dynamic character network from movie. In J. J. Jung, & P. Kim (Eds.), *Big data technologies and applications* (pp. 48–53). Cham: Springer International Publishing.

Tran, Q. D., & Jung, J. E. (2015). Cocharnet: Extracting social networks using character co-occurrence in movies. *Journal of Universal Computer Science, 21*(6), 796–815.

Truby, J. (2008). *The anatomy of story: 22 steps to becoming a master storyteller*. Farrar, Straus and Giroux.

Varol, G., Laptev, I., & Schmid, C. (2017). Long-term temporal convolutions for action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. PP 99.

Venugopalan, S., Rohrbach, M., Donahue, J., Mooney, R., Darrell, T., & Saenko, K. (2015). Sequence to sequence – video to text. In *Proceedings of the IEEE international conference on computer vision (ICCV)* (pp. 4534–4542). IEEE.

Weng, C. Y., Chu, W. T., & Wu, J. L. (2007). Movie analysis based on roles' social network. In *Proceedings of the IEEE international conference on multimedia and expo* (pp. 1403–1406). doi:10.1109/ICME.2007.4284922.

Zhu, S., & Liu, Y. (2009). Automatic scene detection for advanced story retrieval. *Expert Systems with Applications, 36*(3), 5976–5986. Part 2. doi: 10.1016/j.eswa.2008.07.009.