E-StorySpace: Explore Movie Story with Emotion in Space

Author1 Author2 Author3

Abstract

In this paper, we propose E-StorySpace, a spatial visualization for understanding and exploring story in movies. We first automatically extracted a series of elements of a story such as scenes, characters and dialogues with movie script processing algorithm. And we extracted multi-modal emotional features of a movie by deep neural network. Then we organized these elements of a story into feature vectors and clustered them at different levels in 2D space. So we got the spatial distribution of story content that can help users to explore and compare different parts of a story that they are interested in. This is the first time that a multi-modal emotional feature has been used to analyze a movie. We also provide multiple views for users to analyze stories from an emotional perspective. The novelty and effectiveness of E-StorySpace have been demonstrated by user study.

Keywords: Story Visualization, Emotion Analysis

1 Introduction

The key to understand a story is to analyze its components that express where, when, what and who in the story. For example, the way we usually tell stories is that people did something in somewhere, or what's happened in somewhere. Moreover, in a detailed way, we may describe how people feel and the relationship in stories. So the components throughout a story such as scenes, events, characters and emotion are important for analyzing and understanding a story.

Story visualization provides an intuitive way for understanding it and guiding user to explore stories. Many related work in recent years focus on different aspects of story visualization, including character's line, narrative style and scene partition [9, 18, 20]. They expressed and visualized stories in a chronological or narrative order. This kind of method can help users to understand the narration of a story. However, these methods lack a comparative analysis of details in a story and user can not know the relationship of different parts in the story easily. For example, user may want to know how different events that occur in the same scene are distributed in a story and how a particular group of characters meet and interact in stories. Most of work on story visualization are set to make a story more readable through visualization by a linear order. This may reduce initiative of users, when they aim to compare specific content of a story, because following a linear order can't help them to compare similar events quickly. Thus, we propose a spatial form to express content of a story, which maps events of a story in three aspects. These three aspects contain scenes, characters and emotion of each event and we map them by their similarity. We also create connections between different aspects. When user choose one of these aspects to explore, they can also see corresponding content of other aspects. Another problem in story visualization is that little attention was paid on using emotion to analyze a story. Emotion, as an important part of the whole story, is often used to analyze the audience's subjective feelings. Moreover, the change of emotion in a story determines the development and highlight of a story. However, in the field of story visualization and analysis, emotion has not been fully excavated and utilized.

To better address these requirements, we proposed E-StorySpace, a spatial form for exploring content of a story and analyzing details on emotion. We divided a story in events by movie script processing and mapped a story in three aspects (scene, character, emotion) respectively. Emotion, as a trait of E-StorySpace, is an important element for understanding and exploring content of stories. We extracted multi-modal emotion features from movie and recognized emotion type and its intensity automatically. We analyzed emotional content of a story using results recognition.

Taking the film Titanic for example, we analyze story of a movie with our E-StorySpace. Our primary contribution is using a spatial form to express stories with emotional analysis from multi-modal data. We automatically extracted kinds of elements in a story and mapped the story in 2D space with three forms, scenes, characters and emotion. The paper is structured as follows, Section 2 reviews the relevant work in the field of story visualization and emotion analysis. Section 3 describes how to map a story in spatial form and how we use emotion to analyze a story. Section 4 verifies the novelty and effectiveness our system through user study from 20 participates. Section 5 discusses the limitations of our framework, summarizes our work and talks about future work.

2 Related Work

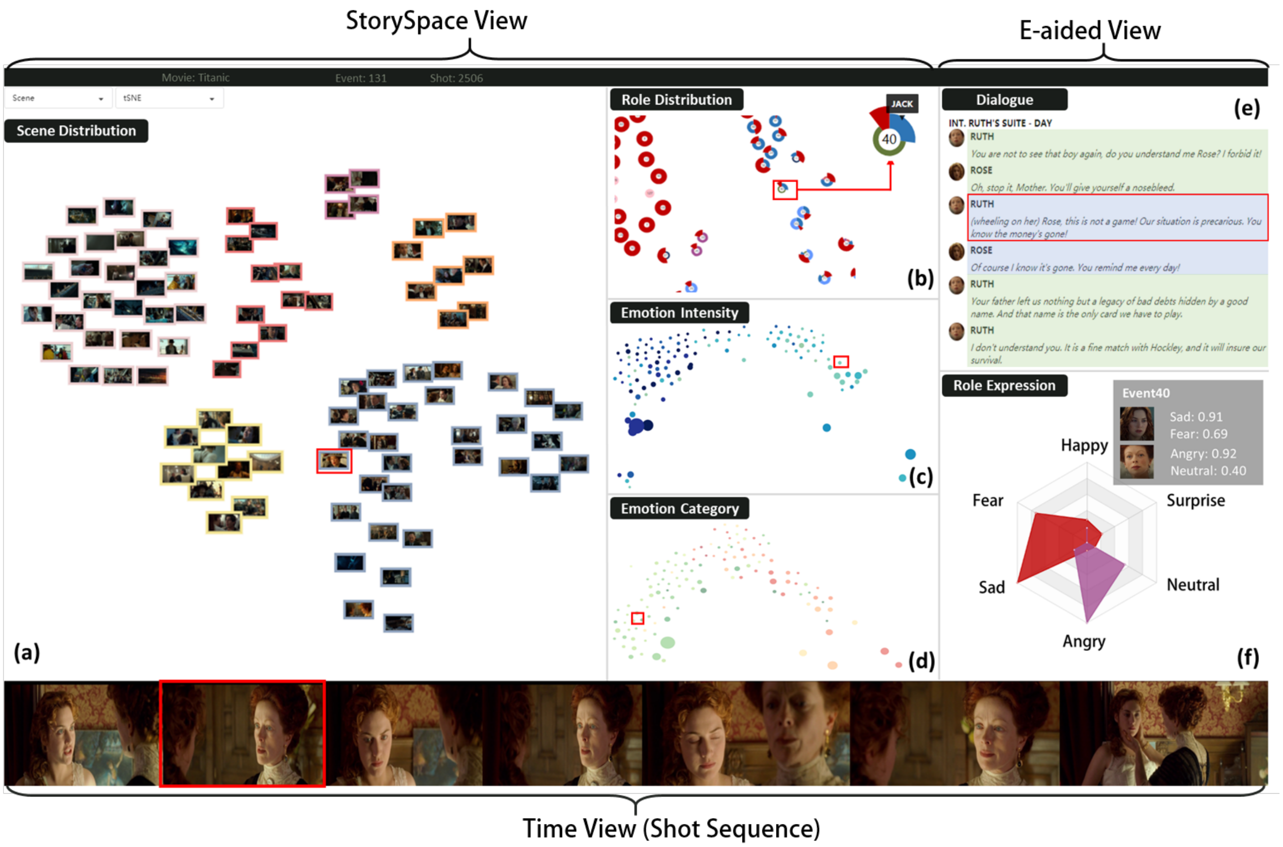
2.1 Story Visualization

Story visualization aims to present stories in an intuitive way for conveying information and understanding content of stories. Most of work use movies as resource and they usually use storytelling as a technique to present stories. Lu et al. proposed a hierarchical plot visualization method called StoryCake to explore the plot of a movie and they used fan-shaped visualization view to express the hierarchical relationship in different part of movie [18]. B. Bach et al. introduced TimeCurve as a general approach for visualizing patterns of evolution in temporal data and they used time node clustering and curve folding to express content of video stories [1]. The latest work StoryCurves [9] proposed a visualization technique for communicating nonlinear narratives and constructed a story curve from a sequence of events in a narrative order and chronological order. In addition, K. Kurzhals et al. proposed a system for visual analysis of film content including characters and scenes [10]. They used timeline as well as storyboard representation to support an effective summary of stories. However, these forms of visualization mainly based on timeline that may limit user engagement, as users must follow the linear order so they can't explore a story freely. Representing a story in spatial form can fully inspire user's motivation to explore and compare the content of a story.

Spatial forms are more perceptible than linear time for people, and it is more flexible as well. People can compare and locate content of a story quickly by clustering. So we designed E-StorySpace with a spatial form for story visualization, because we think people are more familiar with exploring in space and this form can increase memorability of a story as well.

2.2 Emotion Visualization and Analysis

The semantic meaning of a story is ambiguous, as the content of it can be perceived in many different ways. There are two different basic levels of content perception, cognitive level and affective level. Understanding the emotional content is an important dimension for story analysis [5]. With the development of the affective computing,



INT. RUTH'S SUITE - DAY

RUTH
You are not to see that boy again, do you understand me Rose? I forbid it!

ROSE
Oh, stop it, Mother. You'll give yourself a nosebleed.

RUTH
(wheeling on her) Rose, this is not a game! Our situation is precarious. You know the money's gone!

ROSE
Of course I know it's gone. You remind me every day!

RUTH
Your father left us nothing but a legacy of bad debts hidden by a good name. And that name is the only card we have to play.

RUTH
I don't understand you. It is a fine match with Hockley, and it will insure our survival.

Figure 1: Overview of E-StorySpace: E-StorySpace contains three parts of views. StorySpace view is the first part, including a main view and three switchable smaller view. Scenes, characters and emotion of stories are mapped into StorySpace view respectively and users can toggle and zoom in to see different views depending on their needs. The second part is the E-aided view at the right side that displays emotion of characters' dialogues and expressions. The time view at the bottom is corresponding to the StorySpace view based on users' operation.

related work about emotion analysis has a sustained growth. Most approaches are concentrated in using computer vision method to analyze emotional features of a movie. They extracted a range of features from video key-frames and learned the features related to emotion [8, 26]. Besides, recent work has focused more on learning emotional features with heterogeneous data or specific domain knowledge [3, 23, 24]. Considering the multi-modal state of movies, a lot of work also created multi-modal emotional analysis frameworks [17,25]. And some work create emotional summary for movie analysis. Zhao et al. presented movie affective content in the form of a summary by learning audio-visual features [27]. Lan et al. created a movie summary from emotion perspective [11]. Current work related to movie emotional analysis are all about feature learning, including the features from image, audio or text. However, their work have few detailed analysis of emotional content combining with scenes, events and characters. They just added emotion as a label of the whole story, without visualization of emotion either. There is still a lack of emotion-oriented work for movie visualization and analysis. So we propose E-StorySpace, using emotion to analyze and explore stories. A contribution of our work is that we extract emotion from a movie by multi-modal features recognition methods automatically and map emotional content of a story in 2D space.

3 Framework of E-StorySpace

As shown in Fig.1, our E-StorySpace system is composed by multiple views that can be divided into three parts. We will introduce how to build our framework for spatial exploration and multi-modal emotion analysis in this section.

3.1 Multi-level analysis in space

3.1.1 Processing Movie Scripts

Movie scripts are structured and contain rich content and a movie story consists of many events. We used a method proposed by Pavel et al. [16] to extract scenes from scripts and divided a story into individual events. We extracted scene headings, descriptions of each event, character names and dialogues using scripts automatically. We ignored elements like CUT TO, FADE TO and other technical notes for directing a movie.

In order to understand and explore stories clearly, we built our E-StorySpace by events. Firstly, we divided a story into clips according to technical notes like INT. or EXT. to determine the change of scenes. Then we used scene headings to cluster clips and formed event division. Taking the movie Titanic for example, we divided it into 97 events by scripts processing. Next, we divided scripts for each event into two parts: one part contains all the dialogues and the other part is the description (e.g. description of scene, action). Finally, we matched movie subtitles with dialogues to get a time range for different events. The processing of movie scripts is an important step that helps us to get the primary elements of a story.

3.1.2 Story Analysis in StorySpace

As we have mentioned above, we mapped a story in space according to scenes, characters and emotion. For the scenes of a movie, we extracted visual features from movies by deep neutral network

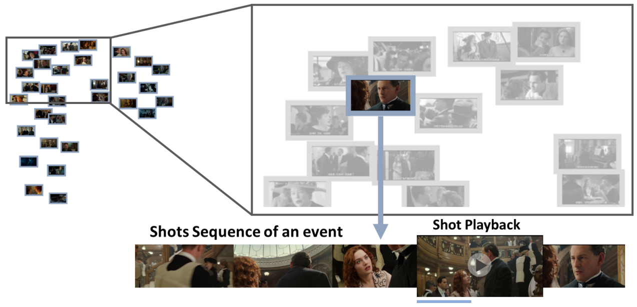


Figure 2: Spatial form for story analysis with scene. Users can select an event and watch corresponding shots.

VGG19 [21]. We referred to the method of M. Tapaswi's work [22] to get the visual similarity of each events and calculate distance matrix. We used tSNE to display the story of Titanic with scene. As shown in Fig.1(a), the rectangular key-frames represent events of the movie Titanic. Adjacent rectangles in space are identified as events with similar scenes visually and they are added with same border color. The images of each key-frame are chosen from the movie by calculating cumulative visual similarity. We used the method from PySceneDetect [1] to cut video clips in shots. As shown in Fig.2, users are free to zoom in the view and select an event to watch its shot sequence and play video clips to explore details of a movie.

For analysis of a story with character, we improved the method from Ma et al. [14]. We converted the characters in a story into nodes of social network and constructed a line between two characters if they are involved in the same event. We got character network by integrating the participation of all characters. Then we used Pagerank [15] to calculate centrality of each node in network and used it as the importance score of characters. We selected top 10 characters by centrality as the major characters. Then we extracted the their names from each event by processing scripts and construct a vector for each event (the value of elements in vector is 0 or 1, representing if character is involved in a event). We used Euclidean distance to calculate similarity of vectors. Finally, we mapped events in space by tSNE. As shown in Fig.3, we can see that each event is represented by a rose chart after zooming in on the view. The amount of fan in each rose chart indicates the number of characters in an event, and the color of fan represents different characters. The number in the center of rose chart is event number. And the radius of fan represents character's importance while the angle reflects the participation of a character. We calculated the participation of each character in different event based on their dialogues number in movie scripts. Fig.3 shows two detailed examples. From the enlarged view (a) we can see the cluster of events that only have one major character called Jack, and in the view (b) we can find that events with same characters are close. Three events that involve Rose and Cal are clustered in the left while the events involve Rose, Cal and Jack are relatively far away from them. In this kind of view, events with same characters are clustered closer. By mapping events of a story with characters, users can explore the story through characters that they are interested in. They can clearly know who are involved in the same events and compare events based on characters.

3.1.3 Emotional Analysis

We provided emotion analysis for a story by extracting multi-modal emotional data from a movie. Emotion is an important element for story analysis as it runs through the whole story and is always related to the plot of a story. The key that decides whether a story is attractive is to correctly convey the underlying emotions. Thus, exploring a story from an emotional perspective is a novel and effective way.

We described emotion in two ways, the intensity and type of emotion which are called arousal and valence in psychology. These two values are used to describe continuous emotion. Arousal is used to describe how strong an emotion is while valence depicts how positive or negative an emotion is. Firstly, we selected a frame from each second as key-frame. Key-frame is defined as the frame with the closest RGB histogram to the mean RGB histogram of the whole excerpt using the Manhattan distance. Secondly, we extracted feature of key-frames by VGG-19 [21] and took the feature of fc7 layer as output. Finally, these features were normalized and we used SVR regressive model for training. All the processes are based the continuous part of LIRIS-ACCEDE dataset [2, 12] which contains continuous induced valence and arousal self-assessments for 30 movies. In this process, we got two continuous values of emotion for each event and we turned these results into vectors. To better quantify the similarity between events in emotional dimension, we used the dynamic time warping algorithm (DTW) [19] which can find the best temporal alignment between two sequences and derive a more comprehensive similarity score. We mapped the events on two views in Fig.1 (c) and (d). As enlarged view (d) shown in Fig.4, we used color to represent the range of emotion in valence. The size of each point means the duration of an event. We also provided continuous emotion line at the bottom in time view of Fig.1, the color behind emotion line is linked to the selected events.

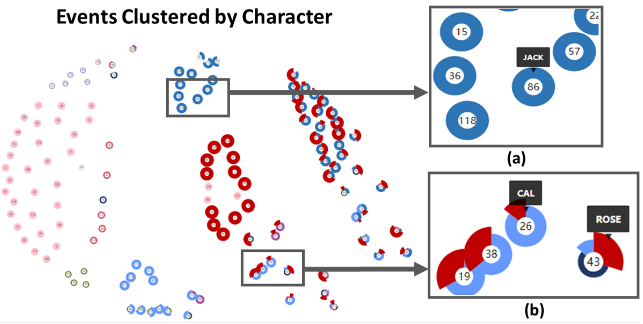


Figure 3: Spatial form for story analysis with characters. Two examples of enlarged views.

3.2 Multi-View Collaborative Analysis

Besides the StorySpace View in Fig.1 that has four switchable views, we also provide Fig.11(e) and Fig.11(f) in E-aided view and the time view. E-aided view is built for emotion auxiliary analysis. We modeled emotion of a story from movie scripts, audio of dialogues and facial expression of characters, combining text, audio and image for multi-modal analysis. E-StorySpace also derives sentiment (negative, positive) of dialogues. We analyzed sentiment by a simple Native Bayes classifier trained on movie reviews [13]. As shown in Fig1.1(e), dialogues from selected event are colored by sentiment analysis (positive, neutral, negative). We integrated emotion recognition from speech and facial expression for analyzing emotion of main characters. We referred the state-of-the-art method (acc=73%) on *Kaggle*[2] to perform facial expression recognition of a movie. ResNet18 [6] neutral network is adopted for feature extraction trained with FER-2013 dataset. We used OpenSmile [4] to extract audio feature and SVM [7] to classify audio and output emotional label. View in Fig.1(f) shows the character's emotion based on audio and facial recognition. Time view supports StorySpace view from perspective of time and it will get changed with the operation

[1] https://pyscenedetect.readthedocs.io/en/latest/

[2] https://www.kaggle.com/sivlemx/facial-expression-of-emotion

of users in StorySpace view. All views in E-StorySpace are related with user's operation and they work together to better display a story.

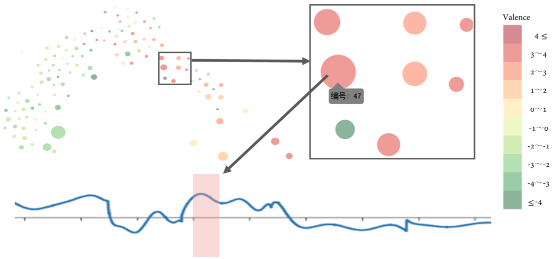


Figure 4: Story analysis with the valence of emotion. Events in the space are corresponding to the continuous emotion line by color.

4 Evaluation

In this section, we conducted experiment to evaluate E-StorySpace. We were interested in assessing the validity and learnability of spatial form for story visualization. We conducted task-based user study to investigate whether our spatial form of StorySpace can help people to explore a story and how emotion can help people for story analysis. We designed multiple tasks related to the plot of movie Titanic, and the question form is asked by WHERE, WHO, WHEN and WHAT. We recruited a total of 20 participants (10 female, 10 male) with different study background. Participants were aged between 18 and 28. 12 were undergraduates, 8 were graduates.

4.1 Procedure

Firstly, all participates were introduced to basic operation of E-StorySpace and the definition of emotion. They were also instructed how to use E-StorySpace, such as switching views, selecting events, playing video shots and so on. The process of instruction is less than 5 minutes. We designed 16 tasks for all participants, 10 tasks were about story exploration, 6 tasks were about emotion analysis. In order to ensure the fairness and accuracy of tasks, we ensured that all participants haven't watched the experimental movie before and they should finished tasks in the same environment.

All the tasks are proposed by the form of question and participants need to answer them one by one. Note that the order of questions is randomly assigned to participants. Participants looked for answers based on the view of E-StorySpace and they were free to switch between different views. They were also provided interactive operations such as zooming in and out, playing back video clips of the movie and selecting elements. The time limit for all tasks w+as set to 16 minutes, 1 minute for one task. We expected these tasks to serve as understanding stories and acquainting how spatial form and emotion work for story exploration and analysis.

After finishing all tasks, participants were asked to make subjective assessments for our system. Moreover, 5-point Likert scale for evaluation is presented to participants (5: very good, 1: very bad).

4.2 Results

The average correct score of all participants for the entire task is 86.72 (100 in total) (σ = 9.6). The raw scores ranged from 70.75 to 100. The average time taken to complete the whole task is 7.66 minutes per participant ($\sigma = 1.46$, min = 5.42, max = 11.12). The average assessment from all participants is 4.125 ($\sigma = 0.78$, min = 3, max = 5). The overall task performance was satisfactory, and participants were able to correctly finish more than 85% of the questions on average. The subjective feedback from them were positive as well.

We have checked individual tasks, each task took 23.28 sec ($\sigma = 8.6$) on average per participants, indicating that the time limit (1 minute per task) was appropriate. The average percentage of participants who answered each question correctly was 0.82 ($\sigma = 0.6$, min = 0.69, max = 1.0). The results show that participants were able to finish our tasks based on E-StorySpace easily and quickly .

Table 1: Some questions in our tasks.

No.	Task Description	Avg.Time
01	How many characters are there?	20 sec
02	Which event has the most positive emotion?	12 sec
03	Which events were happened on Keldysh?	14 sec
04	When and where did Jack and Rose meet?	35 sec
05	What's emotion of Rose in No.73 event?	19 sec
06	What happened when Jack and Andrew met?	38 sec

According to these results, we selected representative questions of task and summarized them in Table1. This set of questions reflect functions of E-StorySpace from a perspective. Our system is beneficial to comparative tasks and joint consideration of scenes, characters, emotions. In a follow-up survey with 5-point Likert scale subjective questions (5: very good, 1: very bad), participants indicated that they are able to use spatial form to answer the questions of the whole task (mean = 4.38, $\sigma = 0.60$) and using emotion to explore stories is efficient and novel (mean = 4.31, $\sigma = 0.58$). We also asked all participants to give their feedback about the difficulty of each question by 5-point Likert scale (5: very difficult, 1:very easy). They indicated that questions about time are rather difficult to answer (mean = 3.75, $\sigma = 0.83$), such as the question 4 and 6 in the Table1. And the difficulty of all questions are appropriate (mean = 2.68, $\sigma = 0.92$).

5 Discussion & Conclusion

Spatial form for story exploration Our framework provides a spatial form for story exploration with scene, character and emotion views. Spatial form is novel and efficient when people do comparative analysis and relationship exploration of a story. However, there are still some problems in our current work based on the experiment. We use a spatial form to present a story instead of time form, and the elements in space may lost the information of time. So participants' performance on the time-related task is relative not well.

Emotion for story analysis Using emotion to analyze a story is a nice way according to most participants in our experiment. Users can compare the emotion between different events in StorySpace view and they can explore events with similar emotion or specific emotion. We use discrete type, continuous intensity and type to describe emotion. Although the combination of emotion and spatial visualization in current work is not intuitive for all users, as they may need to use the time view for assistant, our method is more comprehensive than previous work on emotion analysis.

In this paper, we proposed a multi-view framework with spatial form for story exploration and emotional analysis. We divided a story into events and mapped them into space from different perspectives. People can do comparative analysis and relationship exploration based on our approach of story visualization. We extracted emotional feature from image, audio and script of a movie automatically and used a continuous and discrete way to represent emotion. Experiments demonstrated that our E-StorySpace can effectively help users to locate elements, compare events and analyze emotion in a story.

For future work, we plan to improve our E-StorySpace with more efficient spatial-temporal visualization that represents emotion and elements of a story in more intuitive way. We would like to extend our framework to various types of movie stories and compare it with other methods in a well-designed experiment.

References

[1] B. Bach, C. Shi, N. Heulot, T. Madhyastha, T. Grabowski, and P. Dragicevic. Time curves: Folding time to visualize patterns of temporal evolution in data. *IEEE Trans Vis Comput Graph*, 22(1):559–568, 2016.

[2] Y. Baveye, E. Dellandra, C. Chamaret, and L. Chen. Deep learning vs. kernel methods: Performance for emotion prediction in videos. In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, pp. 77–83, Sep. 2015. doi: 10.1109/ACII.2015.7344554

[3] T. Chen, Y. Wang, S. Wang, and S. Chen. Exploring domain knowledge for affective video content analyses. In *Proceedings of the 25th ACM International Conference on Multimedia*, MM '17, pp. 769–776. ACM, 2017.

[4] F. Eyben, M. Wöllmer, and B. Schuller. Opensmile: The munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, pp. 1459–1462. ACM, New York, NY, USA, 2010. doi: 10.1145/1873951.1874246

[5] A. Hanjalic and L. Q. Xu. Affective video content representation and modeling. *IEEE Transactions on Multimedia*, 7(1):143–154, 2005.

[6] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[7] M. A. Hearst. Support vector machines. *IEEE Intelligent Systems*, 13(4):18–28, July 1998. doi: 10.1109/5254.708428

[8] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim. Joint fine-tuning in deep neural networks for facial expression recognition. In *IEEE International Conference on Computer Vision*, pp. 2983–2991, 2015.

[9] N. W. Kim, B. Bach, H. Im, S. Schriber, M. Gross, and H. Pfister. Visualizing nonlinear narratives with story curves. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):595–604, Jan 2018. doi: 10.1109/TVCG.2017.2744118

[10] K. Kurzhals, M. John, F. Heimerl, P. Kuznecov, and D. Weiskopf. Visual movie analytics. *IEEE Transactions on Multimedia*, 18(11):2149–2160, 2016.

[11] Y. Lan, S. Wei, R. Liu, and Y. Zhao. Creating video summarization from emotion perspective. In *IEEE International Conference on Signal Processing*, pp. 1112–1117, 2017.

[12] T. Li, Y. Baveye, C. Chamaret, E. Dellandréa, and L. Chen. Continuous arousal self-assessments validation using real-time physiological responses. In *Proceedings of the 1st International Workshop on Affect &#38; Sentiment in Multimedia*, ASM '15, pp. 39–44. ACM, New York, NY, USA, 2015. doi: 10.1145/2813524.2813527

[13] S. Loria. Textblob: Simplified text processing. *Secondary TextBlob: Simplified Text Processing*, 2014.

[14] C. Ma, Y. J. Liu, G. Zhao, and H. Wang. Visualizing and analyzing video content with interactive scalable maps. *IEEE Transactions on Multimedia*, 18(11):2171–2183, 2016.

[15] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November 1999. Previous number = SIDL-WP-1999-0120.

[16] A. Pavel, D. Goldman, B. Hartmann, and M. Agrawala. Sceneskim: Searching and browsing movies using synchronized captions, scripts and plot summaries. pp. 181–190, 11 2015. doi: 10.1145/2807442.2807502

[17] S. Poria, E. Cambria, R. Bajpai, and A. Hussain. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, 37:98–125, 2017.

[18] L. Qiang, C. Bingjie, and Z. Haibo. Storytelling by the storycake visualization. *Visual Computer*, 33(46):1–12, 2017.

[19] S. Salvador and P. Chan. Toward accurate dynamic time warping in linear time and space. In *Intelligent Data Analysis*, pp. 561–580, 2007.

[20] L. Shixia, W. Yingcai, W. Enxun, L. Mengchen, and L. Yang. Storyflow: tracking the evolution of stories. *IEEE Trans Vis Comput Graph*, 19(12):2436–2445, 2013.

[21] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[22] M. Tapaswi, M. Buml, and R. Stiefelhagen. Storygraphs: Visualizing character interactions as a timeline. 06 2014. doi: 10.1109/CVPR.2014.111

[23] J. Tarvainen, J. Laaksonen, and T. Takala. Film mood and its quantitative determinants in different types of scenes. *IEEE Transactions on Affective Computing*, PP(99):1–1, 1949.

[24] B. Xu, Y. Fu, Y. G. Jiang, B. Li, and L. Sigal. Heterogeneous knowledge transfer in video emotion recognition, attribution and summarization. *IEEE Transactions on Affective Computing*, 9(2):255–270, 2015.

[25] S. Zhalehpour, Z. Akhtar, and C. E. Erdem. Multimodal emotion recognition based on peak frame selection from video. *Signal Image Video Processing*, 10(5):827–834, 2016.

[26] S. Zhao, H. Yao, X. Jiang, and X. Sun. Predicting discrete probability distribution of image emotions. In *IEEE International Conference on Image Processing*, pp. 2459–2463, 2015.

[27] S. Zhao, H. Yao, X. Sun, X. Jiang, and P. Xu. Flexible presentation of videos based on affective content analysis. In S. Li, A. El Saddik, M. Wang, T. Mei, N. Sebe, S. Yan, R. Hong, and C. Gurrin, eds., *Advances in Multimedia Modeling*, pp. 368–379. Springer Berlin Heidelberg, 2013.