Harnessing A.I. for Augmenting Creativity: Application to Movie Trailer Creation

John R. Smith[1], Dhiraj Joshi[1], Benoit Huet[2] Winston Hsu[3], Jozef Cota[1*]
[1]IBM T. J. Watson Research Center,[2] EURECOM, [3] National Taiwan University
jsmith@us.ibm.com,djoshi@us.ibm.com,Benoit.Huet@eurecom.fr,whsu@ntu.edu.tw,jcota@us.ibm.com

ABSTRACT

In this paper, we describe the *first-ever* machine human collaboration at creating a *real* movie trailer (officially released by 20^{th} Century Fox). We introduce an intelligent system designed to understand and encode patterns and types of emotions in horror movies that are useful in trailers. We perform multi-modal semantics extraction including audio visual sentiments and scene analysis and employ a statistical approach to model the key defining components that characterize horror movie trailers. The system was applied on a full-length feature film, "Morgan" released in 2016 where the system identified 10 moments as best candidates for a trailer. We partnered with a professional filmmaker who arranged and edited each of the moments together to construct a comprehensive trailer completing the entire processing as well as the final trailer assembly within 24 hours. We discuss disruptive opportunities for the film industry and the tremendous media impact of the AI trailer. We confirm the effectiveness of our trailer with a very supportive user study. Finally based on our close interaction with the film industry, we also introduce and investigate the novel paradigm of tropes within the context of movies for advancing content creation.

CCS CONCEPTS

• **Information systems** → **Multimedia content creation**; • **Computing methodologies** → *Video summarization*;

KEYWORDS

Automatic Trailer Generation; Computational Creativity

1 INTRODUCTION

The multimedia and vision communities have witnessed tremendous advances in the field of image and video understanding in the last two decades. Besides semantics, researchers have also made active contributions in highly challenging domains such as aesthetics, emotions, and sentiments in the audio-visual space. If we observe the spectra of research in multimedia content analysis in the last decade or so, we can see a shift in focus towards solving more creative and subjective problems with large quantities of data. Computational creativity lies along the far reaches of this space that researchers have begun to tread only recently.

One of the biggest beneficiaries of research in computational creativity within the context of audio-visual analysis can be the media and entertainment industry that produces thousands of TV shows and movies every year. Typical shooting ratios for TV shows and movies can vary anywhere from 10:1 to 100:1. In other words, this means that in order to produce a two hour movie about 200 hours of video is shot, recorded, manually watched, selected and curated. Video understanding technology that can analyze hours and hours of footage and identify "good" content which can in turn significantly enhance the creative process. With thousands of movies coming out every year[1], one of the key tasks for the producers is creation of trailers to advertise them. A trailer, besides being an advertising platform for a movie and its cast can be one of the most important determinants in the reception, popularity, and ultimately the success of the movie. Movie trailers are growing more important in the marketing of films because in the past they were typically confined to theaters and screened during the previews for upcoming attractions [12]. Now, they are seen by much wider audiences across the globe via social media platforms such as YouTube, Vimeo, and Facebook months before a film is released.

Over the years, trailers have been seen worthy of study. According to the noted film expert, Dr Kriss Ravetto-Biagioli, of Edinburgh University, "trailers play on emotional impact, set up situations, and produce an overall feel, both aesthetic and emotional, very quickly". From a creativity standpoint, movie trailers typically involve high levels of human cognitive effort. The reasons for this are multi-fold; The trailer-maker has to select which scenes from the film are the most pertinent to attract viewer's attention, The trailer should reveal some of the plot but not give away any key information that will spoil the film for anyone who has not previously watched it, The tone of the trailer has to be true to the genre of the film (for example one would not want to create a comedic trailer for a film that is actually more dramatic in nature), and finally the pacing is important to sustain the viewers intrigue and interest.

The above are just some examples of the many highly creative decisions that have to be made for a filmmaker who is working on a movie trailer. Since the creation of a movie trailer requires a great deal of human effort, they also typically require a great deal of time and cost. Teams have to sort through hours of footage and manually select each and every potential candidate moment. This process is expensive and time consuming taking anywhere between 10 and 30 days to complete. We demonstrate that it is possible to accomplish some of the above tasks with intelligent multimedia analysis while leaving the final creative decision making to the human.

*This work was performed while Benoit Huet and Winston Hsu were visiting IBM T.J. Watson Research.

MM '17, October 23–27, 2017, Mountain View, CA, USA

ACM ISBN 978-1-4503-4906-2/17/10…$15.00
https://doi.org/10.1145/3123266.3127906

[1]http://www.the-numbers.com/movies/year

In this paper, we present a novel and first of the kind application of multimedia analysis to creative trailer making for a real feature film. We demonstrate how a team of machine and human can effectively accomplish the highly creative task of creating a trailer that is traditionally extremely labor-intensive. We tackle the complex and creative process of selecting scenes for trailers especially in the domain of horror-thriller movies. The system has been trained on horror trailers from the top 100 horror movies by segmenting out each scene from the trailers and performing audio-visual analysis including visual sentiment and scene analysis of visual key-frames, audio analysis of the ambient sounds (such as the character's tone of voice and the musical score) to understand the sentiments associated with each of those scenes, and an analysis of each scene's composition (such as the location of the shot, the image framing, and the lighting) to categorize the types of locations and shots that traditionally make up suspense/horror movies. In summary, the key contributions of this paper are listed below:

- We demonstrate the **world's first ever** joint computer-human effort for creating a trailer that has been officially released, recognized, and received tremendous media attention. This was performed for a full-length feature film, Morgan released in September 2016 by 20^{th} Century Fox[2].
- We make a case for the fact that the current state-of-the-art multimedia AI technologies have already arrived at the stage where they can significantly enhance creative processes such as creating trailers. We leverage some of the the best available multimedia analysis tools and demonstrate how they can be effectively used to accomplish a highly creative task that is traditionally very labor intensive.
- We rigorously evaluate user reception of our Augmented Intelligence (AI) trailer produced with joint computer and human effort versus the official 20^{th} Century Fox Trailer for the same film with an extensive user study of more than 100 participants from all over the world.
- Inspired by our interaction with the movie industry, we introduce another complex multimodal ontology, "tropes", employed commonly in movies. We characterize tropes and show their utilities in film production by mining frequent visual elements across 140 films. We also discuss emerging opportunities for this open research problem and a new paradigm for contextual understanding.

In the light of our current work, we discuss about several related creative tasks in the movie and TV show industry that can significantly benefit from augmented intelligence (AI).

2 RELATED WORK

2.1 Multimedia Analysis for Trailers

Movie and in general video summarization or abstraction has been studied within the multimedia community for a few decades [14, 21, 35]. There are however a number of major differences between a summary and a trailer. The first relates to the intent. An abstract aims at giving the viewer a complete overview of the original content without watching it entirely, while trailers aim at attracting the viewer to see the entire content. Hence, while the summary will reveal the plot and the end, the trailer will attempt to keep it unspoiled. Similarly, summaries tend to follow the original timeline while teasers almost systematically break the temporal order so as to not divulge the underlying narrative.

Trailers did not receive the same attention as video summaries did from the multimedia analysis research community. Among the few works addressing the issue, the work of Smeaton *et al.* [33] focuses on action movies and studies specifically the visual motion level throughout the movie and the detection of specific audio cues (speech, music, silence, etc..) in order to describe and select individual sequences for creating a trailer. Another approach, proposed in [16] in the context of TV program, focuses on finding textual correspondences between sentence in the program summary (provided by the Electronic Program Guide) and the program's closed caption. Among the short-comings inherent to such an approach, there is the strong requirement for a textual summary to be prepared prior to having the trailer made. Other approaches are focusing on the identification of salient events to determine key moments in a movie. In [19], an approach combining visual motion features, audio energy and affective word analysis is proposed. Salient video segments are selected by fusing the individual confidence score along the three modality.

The MediaEval Benchmarking Initiative for Multimedia Evaluation has hosted a task aimed at identifying violent scenes which has evolved (since 2015) into a task addressing the Emotional Impact of movies task [32]. Approaches tackling this challenge focus on the prediction of valence and arousal scores at the level of short excerpts (time window and/or segments) and employ a wide range of features, ranging from visual to audio, to train various regression models, ranging from Support Vector [5] to deep RNNs [23].

2.2 Audio Visual Sentiment Analysis

Visual sentiment and emotion analysis within the context of images was explored extensively in the last two decades [15]. Traditional visual sentiment modeling was treated as a multi-class classification problem (using insights from psychology literature) and well curated controlled datasets such as IAPS [20]. IAPS consists of a diverse set of photos showing animals, people, activities, and nature, and has been categorized in valences (positive, negative, or no emotions) along various emotional dimensions [37]. More recently with the availability of very large-scale image datasets and advent of powerful deep learning approaches, newer philosophies to sentiment analysis are being explored [3, 6]. Both [3, 6] explore the association between sentiments and adjective-noun pairs (e.g. happy dog) wherein the hypothesis is that an adjective-noun pair evokes a specific mix of sentiments and the mapping is learned from large image datasets.

Within the audio domain, emotion recognition has been studied for speech [29] and music [17]. One of the most prominent recent works that focuses on audio-based sentiment is the OpenSMILE project [9, 11]. OpenSMILE provides an open-source audio feature extractor that incorporates features from music and speech within the framework of emotion recognition. A survey of joint audio visual affect recognition methods and spontaneous expressions is presented in [38]. In one of the earlier works [31], audio visual emotion recognition was studied for facial and vocal expressions. In a recent work, fusion of audio, visual and textual clues has been

[2]http://www.imdb.com/title/tt4520364/

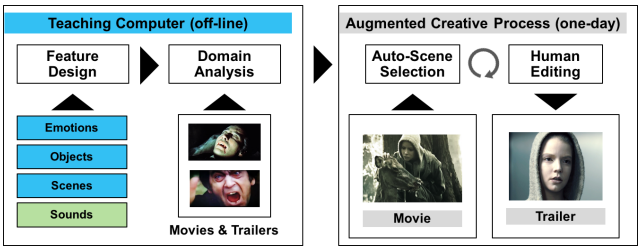

Figure 1: The high-level architecture of the Intelligent Multimedia Analysis driven Trailer Creation Engine.

proposed for sentiment analysis in [26]. In addition to these works, the annual AVEC challenge is largely devoted to studying mood and emotion (esp. depression) within the context of audio visual analysis [36].

2.3 Film Datasets

Datasets are critical for image/video learning. However, for film analysis, there are very limited datasets available due to copyright concerns. MovieQA dataset [34] aims at evaluating video comprehension for multiple modalities. There are 400 movies with parts or video segments, subtitles, plots, etc. It is mainly for question and answering along with its 15,000 multiple choice questions. The MovieBook Dataset [40] aims for movie and book alignment. The authors collected 11 movies with subtitles, shots, and alignment annotations. MPII Movie Description dataset (MPII-MD) [27] aims for movie descriptions from video content for visually impaired people. It contains 68K sentences and video snippets from 84 movies. However, the videos are generally short segments without dialogues. In short, we are unable to leverage existing datasets for trailer generation. We however obtained 100 trailers from the horror genre in order to train our system. Additionally, we will demonstrate how to mine film production rules, in a new paradigm, by leveraging video segments from MovieQA dataset [34].

3 TECHNICAL APPROACH

3.1 High-Level Architecture

A high-level architecture of our system is shown in Figure 1. The system incorporates both domain-specific and cross-domain knowledge as depicted in the figure. We leverage state-of-the-art research in audio and visual modeling and analysis in a wide spectrum of domains from web photos to news audio in order to create a set of diverse audio-visual representation for movie segments. Thus our representation incorporates a wide variety of affective signals and semantics learned from common world images and video. At the same time, we adapt these representations to learn domain-specific knowledge specific to the horror-thriller genre of movies. This forms the core of the multimedia learning involved. The augmented creative process involves application of domain specific learning to a given movie (as shown in Figure 1) to come up with a selection of a few scenes that are used by the human editor to create a trailer. In the following sections, we will describe the individual components that constitute the system.

3.2 Audio Visual Segmentation

The first step in creating a comprehensive understanding of a movie or a trailer was to break it into audio/visual snippets each of which bears a coherent theme. Each of the snippet can then be used as an individual entity for audio-visual modeling.

For this work, we performed audio and visual segmentation independently and later reconciled these segments to form composite pieces of the movie story as a whole. Visual shot-boundary detection was performed and for simplicity each shot was represented by a visual key-frame for further visual feature extraction. Audio segmentation was performed using OpenSmile [9]. For each audio segment a full fledged emotional vector representation was extracted using OpenEAR [10] an OpenSmile extension dedicated to audio emotion recognition. In totality, for the full movie Morgan we obtained a total of 1935 visual shots and 676 audio segments. Reconciliation of the audio visual segments was performed as follows: we aggregated visual sentiment features for all key-frames that fall within an audio segment to form a composite visual sentiment feature.

3.3 Audio Sentiment Analysis

Audio, whether speech, music or sounds conveys a key message concerning what is happening in a scene. Even with your eyes closed you can guess with a pretty good accuracy whether a person talking is sad or happy, or if a musical track is cheerful, peaceful or even suspenseful. Horror and thriller movies, as any other genre, do exhibit scenes with carefully chosen audio tracks in order to place the viewer in the right mood for the current or sometime upcoming scene. The intelligent selection of scenes for the creation of a movie trailer requires the identification of which sentiment is expressed within every scene. Indeed, when constructing the trailer of a horror/thriller, a director is going to favor certain scene versus others based on the emotion they communicate to the viewer.

In the work presented here, we employed OpenSmile [11] for performing audio analysis, and in particular the OpenEAR [10] extension which provides an efficient framework for emotion recognition. The strength of this framework is in the set of models provided which have been trained on six datasets for recognizing emotion. OpenEAR provides a variety of high level characteristics extracted from the audio track. There is almost as many feature sets are there are datasets used for its training (Berlin Speech Emotion Database [4], eNTERFACE [24], Airplane Behaviour Corpus [28], Audio Visual Interest Corpus [30], Belfast Sensitive Artificial Listener [8] and Vera-Am-Mittag [13]). Some emotional classes, like Anger, Disgust, Fear, Happiness, Sadness and Neutral are available multiple times, but might capture different audio signal properties for making the classification as they were trained on independent training data. Other emotional states, (such as Aggressive, Boredom, Cheerful, Intoxicated, Nervous, Surprise and Tired) are defined uniquely with a label. The aforementioned emotion are assigned a probability score which corresponds to the class prediction of the model with respect to some input audio signal. In addition to the discrete high level feature listed above, OpenEAR allows for two continuous dimensional features, namely valence and activation (also referred to as arousal in the literature), both in the range from -1 to +1, to define the emotion contained in the input signal. A total of 18 audio

Figure 2: Sentibank output for a frame from the movie, Morgan (courtesy - 20^{th} Century Fox). The relevant adjective noun-pairs (anps) in terms of sentiments they evoke are labeled in green. We also show the emotion vector corresponding to the anp with the highest score.

sentiment features are computed using OpenEAR for each audio segment extracted from a video (a trailer or a movie).

3.4 Visual Sentiment Analysis

While audio is a key component for setting the tone of a scene, visual imagery is equally if not more decisive especially in the context of horror films. Directors often choose the visual composition of scenes carefully to convey the sentiments of fear and suspense often through powerful imagery. Visuals such as somber scenes, scary forests, and haunted houses can evoke fear and suspense quite strongly (Figure 2).

In order to understand the visual sentiment structure of a movie scene, we need to create a holistic representation over a spectrum of emotions. In this work, we employed Sentibank [3] to extract visual sentiment information from movie key-frames. Sentibank is a departure from traditional visual sentiment modeling paradigm. Traditionally sentiment prediction was treated as a multi-class classification problem and careful sentiment labeling was an imperative part of the process. On the contrary Sentibank solely relies on crowd-sourced information and constructs associations between sentiments and adjective-noun pairs (e.g. happy dog). These associations are constructed by learning from large image datasets from the web. Sentiment prediction is an indirect consequence of adjective-noun pair classification wherein the hypothesis is that an adjective-noun pair evokes a specific mix of sentiments (along the 24 dimensions of the Plutchik's wheel of emotions [3]). An image is first classified into an adjective-noun pair category and is then assigned sentiment scores specific to that category. Our choice of Sentibank was based on the fact that it's one of the most recent state-of-the-art visual sentiment modeling methodology. For our purposes, we used a Sentibank API to obtain top 5 adjective-noun pairs (anps) for each key-frame in a trailer or movie. We then constructed a visual sentiment feature using the 24 sentiment scores corresponding to the sentiment distribution for the highest ranked anp. In order to create a composite visual sentiment representation for an entire segment, we computed a dimension-wise mean across all the key-frames in the segment.

[3]https://goo.gl/aYOwDA

Figure 3: Selected scenes from scary (top), tender (center), and suspenseful (bottom) moments from movie *The Omen, 1976* **(courtesy - 20^{th} Century Fox).**

3.5 Visual Attributes: Places and Scenes

We observe that the production teams often manipulate the atmosphere and the aesthetic factors in films by using certain visual composition rules. For example. in "horror" movies, we often have scenes with dark colors, complex backgrounds, a huge face, etc. Readers can see some of such scene examples in Figure 3 and 4. (A more advanced investigation into this will be presented in Section 7.2). For modeling essential scene oriented visual attributes for trailer generation, we adopted Places 205 CNN model [39] as the main visual feature because the network essentially models places (or locations) and emphasizes more the global context in scene composition. We experiment with features from different layers including softmax, fc7, and fc6.

3.6 Multimodal Scene Selection - Experiments and Results

Having extracted high level audio and visual features from a set of hundred horror trailers available on-line, the Augmented Intelligence (AI) system analyzed all the data gathered and computed a statistical model capturing the prominent characteristics exhibited by trailers from this specific movie genre. Implementation wise, Principal Component Analysis (PCA) was applied to the features extracted from the collection of horror trailers. The three dominant dimensions resulting from this analysis of the multimodal data were believed to capture the main characteristics important in horror movie trailers. This served as the basis to identify trailer worthy scenes from a movie of the same genre. In effect, each scene of a movie is projected onto the 3 dimensional space obtained through PCA and the scenes with the highest response, those that best match the requirement of such trailers, as selected as candidate for making the trailer. Through audio-visual inspection of the scenes responding strongly along each of those axes, we perceived that the three major dimensions corresponded in some fashion to *scary*, *tender*, and *suspenseful* moments.

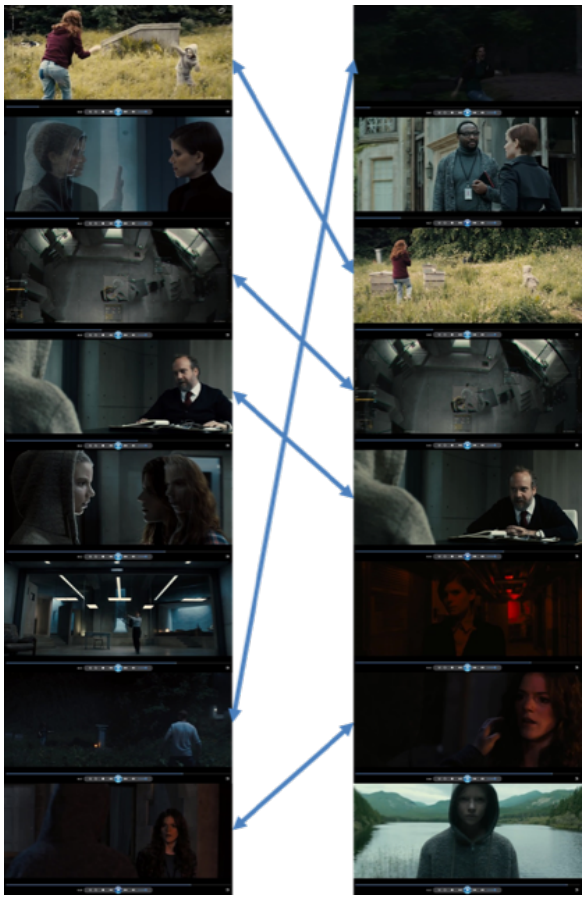

Figure 4: Selected Scenes from the Morgan Trailers arranged in timely fashion from top to bottom. The A.I. Trailer is shown on the left while the official 20^{th} Century Fox Trailer is on the right. Arrows highlight common scenes used in both trailers. (Courtesy - 20^{th} Century Fox)

In order to validate our approach, we analyzed an award winning 1976 motion picture *The Omen*[4] also produced by 20^{th} Century Fox. Some *scary*, *tender*, and *suspenseful* moments from the movie *The Omen* are shown in Figure 3.

In Figure 4 we present a qualitative evaluation of our AI based Morgan trailer vis-a-vis one of the professionally produced Fox trailers for Morgan (a quantitative evaluation will be presented in Section 5). The figure depicts eight scenes from two trailers for the Morgan movie. Frames representing scenes from the movie are organized temporally (from top to bottom) following the original ordering of each trailer with the Augmented Intelligence one on the left and the one made by 20^{th} Century Fox on the right. One can see certain scenes selected through our Multimedia Analysis framework (left) are also present in the original movie trailer (right). Common scenes from both trailers are highlighted by arrows connecting them. It is important to notice that the scene selection process resulted from the analysis of movie trailers of the same cinematographic genre but did not include any footage from Morgan. It is interesting to see how, after learning key multimedia characteristics and semantics extracted automatically from a large set of example trailers, it was possible to identify pertinent scenes from an unseen movie. Moreover, this clearly indicates that scenes selected for making a movie trailer are not chosen randomly from the entire movie and that is it possible to model such a selection process using Augmented Intelligence.

[4]http://www.imdb.com/title/tt0075005/

4 ROLE OF THE FILMMAKER

The movie trailer industry is a $200 million a year industry [7]. Movie trailers that are created for large budget (studio films) are typically created in *trailer houses*. A trailer house is a post production facility that is specifically geared and specialized in creating movie trailers. Sometimes they may be hired to create several versions of a trailer and conduct focus groups to see if they appeal to mass audiences. Sometimes a movie studio may actually contract the trailer to be produced by several trailer houses and then select the trailer that they think is the best match for the film. In recent years with the emergence of social media platforms, there may be even several trailers released that would accompany or precede the release of a film, whereas in the past word of mouth played a bigger role [18]. The role of social media in film marketing has risen in prominence, in which the average film has its own website, most likely its own Facebook page, and possibly numerous versions of the trailer.

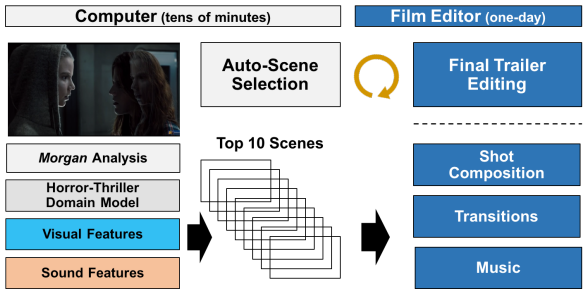

Figure 5: The roles played by the computer and the human in the augmented creative trailer making process.

The creative role a film-maker within our context was to look through the footage provided by our multimedia analysis engine and refine that into a finished trailer. Figure 5 captures the roles of the computer and human in the augmented creative trailer making process. Our system (the computer) provided domain and cross-domain multimedia analysis to come up with selected scenes for the movie. However the final creative touch which involved composition of shots, creating good transitions and overlaying the official Fox soundtrack for Morgan movie was provided by our film-maker colleague. In the current scenario, this entailed sifting through the ten scenes that had a running time of two minutes and fifty-eight seconds and then selecting and re-arranging the best shots and cutting it to a version that was 1 minute and nineteen seconds.

The structures of trailers vary. However most films adhere to something called three act structure (unless they are an experimental film). Three act structure essentially means that there's a

beginning (where the characters are introduced), and what's called "an inciting incident" which puts the protagonist into a forward trajectory in which they have to face an obstacle/or come head to head with the antagonist. There's the second act which puts the main character/protagonist in a deeper state of conflict. Then finally there's the third act which is the climax and resolution of the story. This is classical story telling in the form of narrative fiction and the vast majority of studio produced films strictly adhere to this. A typical trailer introduces audience to an idea of what the story is about, who some of the main characters are, and what obstacles they may face in the story while ideally not revealing too much of what occurs in the third act i.e. which is the resolution of the conflict. The order and the flow of these clips is extremely important in terms of the quality of the trailer. Fractions of a second can change the feel of how an audience perceives a scene.

In this work, no deliberate attempt was made to learn the three act structure of trailers. However, our system selected a diverse enough assembly/reel of scenes for our film-maker colleague to work on. He performed further cutting and mixing of the scenes so that an audience feels intrigued by the story but paid special heed not to reveal or spoil anything with the plot of the story. Our system eliminated the need for the decision process for which scenes should be selected in the first place which would have taken considerable time and effort. This expedited the process substantially. The final human editing of the trailer was completed within a span of 8 hours[5].

5 EVALUATION OF THE AI TRAILER

The previous section reported and covered the direct benefits and experiences of a professional film maker in creating a trailer using the video footage "hand-picked" by our system. A different yet very important test of our performance is to measure user perception of the trailer created using Augmented Intelligence (AI). In order to do so, we set up two anonymized user studies for (1) our AI trailer and (2) an official trailer created by 20^{th} Century FOX for the movie Morgan. Survey participants were shown a trailer and were asked a total of 10 questions including questions about their age and gender and 8 questions to assess what they felt about the trailer. Questions included assessing user's interest in horror films, their rating of the trailer, their feelings (suspense, fear) evoked by the trailer, and whether they would be interested in watching the film Morgan after watching the trailer. In the spirit of Turing Test, we also asked a question about whether they feel AI was used in the making of the trailer in question. Participants only filled in one of the 2 surveys, either for the AI trailer or for the official one.

In terms of demographics, participants of the survey cover a wide range of ages from 18 to 60 from many different geographic locations (all continents are represented although North America, Europe and Asia dominate the data collected). A total of 80 and 54 responses were obtained for AI and Fox trailers survey respectively. In order to evenly compare responses from both surveys as well as represent both genders equally, we focus our further analysis on 50 participants from each survey (25 male and 25 female) who were the first responders from each gender.

[5]Readers can view the complete AI trailer at https://www.youtube.com/watch?v=gJEzuYynaiw

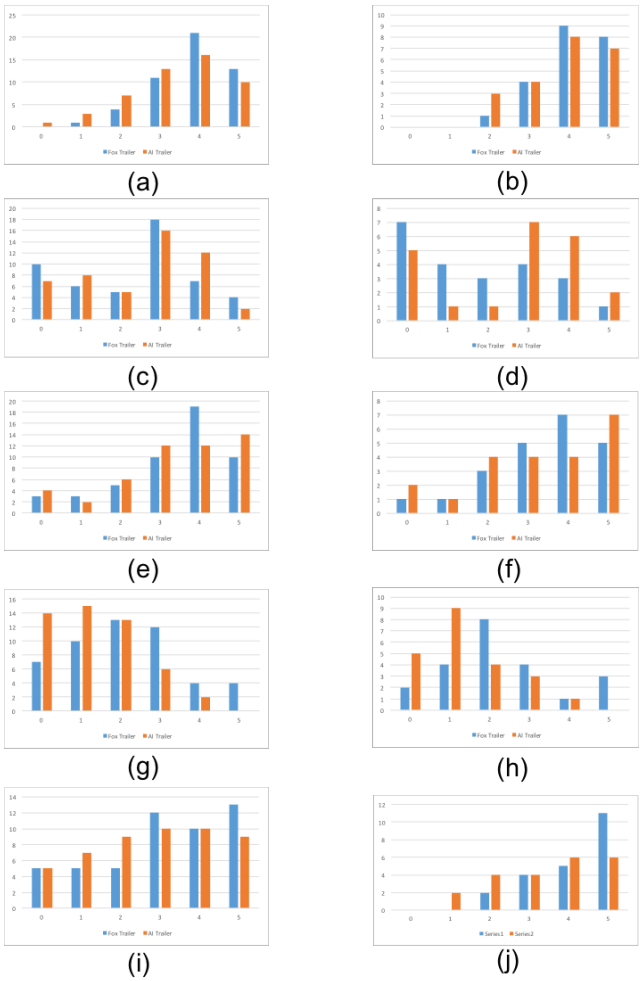

Figure 6: Comparing distributions of user ratings for five different questions for (left panel) all participants, (right panel) horror-movie fans. Ratings range between 0 (very low) to 5 (very high). Fox Trailer response is in blue while AI trailer response is shown in orange. Questions compared are: (a-b) Give the trailer you just saw a rating, (c-d) Does the trailer evoke feeling of fear in you, (e-f) Does the trailer evoke feeling of suspense in you, (g-h) Do you think this trailer gives away too much of the movie (spoilers), (i-j) Would you want to watch the movie Morgan after watching the trailer.

In Figure 6, we show the distributions of user responses across the two trailers for five of the questions most pertinent to our analysis. For each question, we plot the distribution for all participants (left panel) and those who answered "Yes" to the question about whether they watch horror films (right panel). From the figure, we see several interesting trends.

- Overall trailer rating trends (a) and (b) for both Fox and AI trailers are bell shaped (modes at 4) and in general horror-movie fans assign higher ratings to both trailers (b).
- The AI trailer appears to evoke more fear in viewers who watch Horror movies than the Fox trailer (c-d). This is also

true for the feeling of suspense but the effect is somewhat less pronounced (e-f).
- Clearly the AI trailer reveals less of the story than the Fox trailer (g-h). This is a consequence of the fact that only the first 80% of the movie was considered for selection for the trailer in order to avoid spoiling the end in the trailer.
- Studies about both trailers indicate that horror fans are more likely to see the movie after seeing either trailer than non horror fans (i-j). This is somewhat expected as the question about being motivated to watch Morgan is somewhat of a personal preference.

We further analyze the differences in the overall trailer rating distributions for both trailers by performing a two sample, two tailed T-test wherein the null hypothesis is that the means of the two distributions are same. The significance level α is set at 0.05 and a p-value of 0.06 is obtained thus indicating that we do not have sufficient evidence to reject the null hypothesis. In other words, this means that the two distributions come from a population with the same mean and that at population level, differences are marginal. What is also interesting to report is that while for the actual AI trailer, 50% people answered "Yes" to whether they thought AI was used in the creation of this trailer, for the Fox trailer, this percentage was 54% thus indicating that there was no credible evidence for users to differentiate between the two in terms of AI usage. In other words, the AI trailer looked as human-produced as the official Fox trailer.

6 MEDIA IMPACT

Our AI trailer made a tremendous impact in media which indicates a substantial interest from the general public about the creation of this technology and its potential impact for the future of the film industry. Following the release of this trailer there were articles in over 100 publications including Fortune, Entertainment Weekly, Popular Science, TechCrunch, Mashable, Engadget, Refinery 29, Just Jared, Buzzfeed, Fast Company, Business Insider, NY Daily News and Ad Week, where it was the most popular story.

A social campaign shared the story via Facebook, Twitter and Instagram has garnered 1.6M+ views. There have been 4K mentions of the activation on Twitter. Also on social media it was shared by Entertainment Weekly (5.6M followers), 20^{th} Century Fox (2M followers) and Engadget (1.92M followers). It also ranked in the Top 20 trending articles on Reddit's Futurology. Most importantly the YouTube video featuring the AI trailer (released by 20^{th} Century Fox on August 31^{st} 2016) has been viewed more than 2.9 million times about a million of which were received within the first 48 hours of its release. There are several reasons for the overwhelmingly positive response by the world. Firstly, this was the first of its kind accomplishment. Our AI trailer literally disrupted the movie industry and brought to light that AI can and should be increasingly incorporated even for creative tasks such as making trailers. Secondly, it comes at a time when there is an increasingly active interest in artificial intelligence and its future potential with respect to all walks of life.

7 CREATIVITY BY CONTEXTUAL UNDERSTANDING – TROPES

In this section, we shift our discussion to yet another very useful creative device employed in movie industry that we learned about from our previous interaction and demonstrate how they can be modeled and understood using modern multimedia methodologies. Current state-of-the-art solutions in content analysis mostly rely on low-level features or visual recognizers (e.g., objects, scenes, etc.). However, via this cross-disciplined collaboration for AI trailer, we learned that there are specific "recipes"– *tropes* – in creative works such as films, television, comics, etc. Here, we take an initiative to introduce tropes for the first time to the multimedia community, identify open research problems in this domain, and demonstrate how they will advance research in content analysis, contextual understanding, and even computer vision. In a pilot study, we will show how to mine frequent visual elements across movie genres for approximating tropes.

7.1 Tropes in Film Production

A *trope* is a storytelling device or a shortcut for describing situations the storyteller can reasonably assume the audience will recognize. Beyond actions, events, activities they are tools that the art creator uses to express ideas to the audience. In other words, tropes convey a concept to the audience without needing to spell out all the details and are frequently used as the ingredients for film/TV production recipes. For example, "Heroic Sacrifice" is a frequent trope defined as a character saves others from harm and is killed as a result[6]. A film usually contains hundreds of tropes orchestrated intentionally by the production team. Some others include "Bittersweet Ending" , "Going Cold Turkey", "Kick the Dog", etc.

Similarly, in sports videos, there are corresponding tropes for describing the salient contexts in the major events. For example, "buzzer beater": A shot in the final seconds of a game (right before the buzzer sounds) and results in a win or overtime, and "Hail Mary pass": a very long forward pass in American football, made in desperation with only a small chance of success. It is widely believed that tropes are of audiences' great interest.

Rich film tropes are (weakly) annotated in "tvtropes"[7], which is a wiki-like community service focusing on various conventions found in creative works such as films, television, comics, etc. For example, in our analysis, 367 films among the 408 in MovieQA dataset [34] are annotated with tropes (manually annotated by the film community). There are totally 42,359 tropes (11,773 unique) among the 367 films. The number of tropes per film is 115 (average), 84 (media), 627 (max), and 3 (min). The occurring frequency per trope is 3.60 (average), 203 (max), 1 (min). Among them, 266 tropes occur more than 20 times and 814 tropes occur more than 10 times in the 367 films. For example, in film "Interstellar", trope "Oh Crap!"[8], which is often used to reveal the moment when the character realizes something bad is about to happen – intending to create tension in the film.

[6] http://tvtropes.org/pmwiki/pmwiki.php/Main/HeroicSacrifice
[7] http://tvtropes.org/
[8] http://tvtropes.org/pmwiki/pmwiki.php/Main/OhCrap

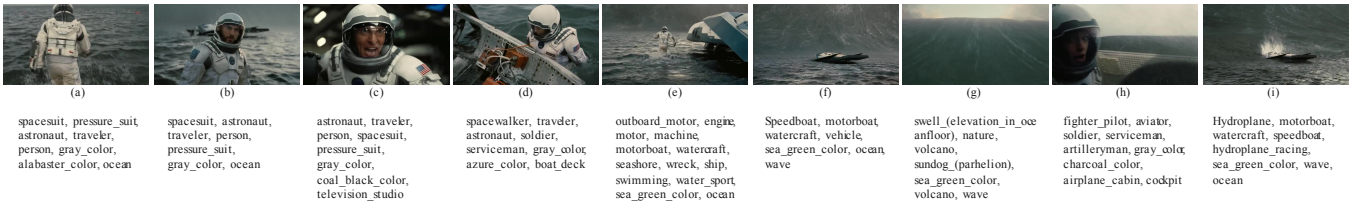

Figure 7: Selected keyframes from a video segment, which was labeled with trope "Oh, Crap!" in film "Interstellar". Corresponding annotations by a visual recognition engine are shown below. Meaningful objects, scenes, and colors can be detected by current technologies. However, the situation (and the tension) – danger is imminent – can be perceived by humans but cannot be detected by current solutions. See more details in 7.1. Best seen in color and in PDF. (Courtesy - MovieQA Dataset).

See sampled key-frames in Figure 7. There are rich descriptions regarding instances of tropes for films[9] such as:

> *Oh, Crap!*: After landing on Miller's planet, Brand notices what appear to be mountains in the distance, until Cooper realizes that the "mountains" are actually waves. However, they also notice that the waves are receding, so nobody panics... until they look behind them and see the ones approaching.

Being able to recognize or synthesize tropes will augment creativity in a very substantial way – having further capability for contextual understanding beyond literally recognizing only appearances. For example, state-of-the-art visual recognition can annotate the keyframes in Figure 7 with proper objects, places, and colors. However, it is very challenging to understand the context "Oh, Crap!" – the danger is approaching. We argue that tropes will be another challenging ontology for understanding the (multimodal) context, metaphors, sentiments, intension, etc., in videos.

At the moment, "trope understanding" poses a brand-new and open research problem for the research community. As a first step, we attempt to align the numerous tropes to video segments in films via cross-domain alignment (e.g., [40][22]) using trope descriptions (exemplified above) and the multimodal recognition results (e.g., the methods proposed in this work). Because of limited training data, we are investigating zero-shot learning [25] for contextual recognition. In the following section, instead, we propose "weak tropes" for mining salient visual factors for film production.

7.2 Weak Tropes for Film Analysis

Motivated by tropes, commonly adopted practices in the creative works, we aim to discover the major "visual elements" as (computational) *weak tropes* – i.e., the combination of dominant colors, objects, and scenes for films across genres. As shown in Figure 7, certain colors (azure, sea green), scenes (ocean, cockpit, wave), and objects (spacesuit, astronaut, watercraft) could be used to exemplify a trope. Besides analyzing the possible visual elements for tropes in a quantitative manner, we also investigate the feasibility of synthesizing (weak) tropes for parameterizing visual elements in a generative manner. We conducted an investigation over video segments from 140 films with average duration of 4293 seconds (i.e., 72 minutes) per film, from the MovieQA dataset [34]. Among them, videos are labeled with multiple genres per film including Drama, Adventure, Thriller, Romance, Comedy, Action, Fantasy, Crime, Sci-Fi, Mystery, Family, Biography, Horror, History, Music, War, etc. The top 10 genres with ratios (%) are listed in Figure 8. We annotate the keyframes via a state-of-the-art visual recognition system, which can provide dozens of thousands labels across colors, objects, and scenes. Some of the results are illustrated in Figure 7.

As we found that salient (and meaningful) annotations from keyframes are sparse (in dozens), we argue to construct *association rules* [1] for mining the effective weak tropes. The intuition is that the production team will manipulate certain visual elements to shoot a film. We need to find the frequent items X (among visual annotations), frequently occurring patterns in the films across all the keyframes F, with certain *supports* , and measure its correlation with the genre types Y in terms of the *confidence*. From a computational perspective, we wish to measure the frequent weak tropes that co-occur a lot with the genres Y. We believe it will be even more interesting if we can associate that with other factors such as directors, casts, ratings, studios, etc.

Computing over 137,672 keyframes across 140 films, we found that weak tropes (visual elements) do saliently exist within genres. For example, the visual elements with the highest confidence score for genre drama is {person, astronaut, spacesuit, traveler} (0,915), {indigo_color, discotheque, nightclub} (0.675) for thriller, {person, dressing_room, beauty_salon, claret_red_color} (0.832) for romance , {pink_color, female} (0.818) for comedy, {spacesuit, astronaut} (0.958) for sci-fi, etc. Readers will agree that these findings align with how we perceive movies from the respective genres (in terms of their visual depiction). It also confirms the practices of using certain coherent visual elements (as weak tropes) to orchestrate the atmosphere for films.

For visualizing the elements across (visual recognizer) categories (e.g., colors, objects, and places), we show the major (with highest confidence scores) annotations (recognition outputs) in Figure 8. Each row represents the genre ordered by its proportion in the 140 films. Each column shows the top ranked visual annotations for the category. In the cell, each annotation x comes with the confidence score (in the parentheses), which means that x, along with other visual elements, is in the frequent items with the top confidence score. For example, in row #3, indigo_color (0.675) is the top ranked visual element among the frequent items {indigo_color,

[9]See the descriptions from http://tvtropes.org/pmwiki/pmwiki.php/Film/Interstellar. Also see the video segment from https://www.dropbox.com/s/hzazgk2wv8887oy/Oh-Crap.mp4?dl=0

discotheque, nightclub} for thriller. It means that for colors, indigo_color, coal_black_color, black_color, are the confidently representative colors adopted in the thriller genre. It is also interesting to observe that some genres generally employ specific color tones; for example, claret red, pale yellow, and pink for romance, and pink, claret red, light brown, and Tyrian purple for comedy.

Similarly, salient objects are frequently associated with different genres (cf. column 3 in Figure 8); for example boxing_ring, witness_box, man_with_shaven_head, motor_vehicle, mustachio, etc., for genre crime (row #8), v.s., girl, child, waitress, platinum_blonde, couple for romance (row #4). Similarly for the scenes (places, 4th column). For example, discotheque, engine_room, rock_climbing, airplane_cabin, elevator_shaft, etc., for genre action (row #6) v.s. home theater, dressing_room/beauty_salon, bar/pub, nursing_home, etc., for comedy (row #5).

With weak tropes, we attempt to understand production rules for the film industry for approximating tropes in a computational manner. We can potentially leverage these as additional ingredients for trailer generation, review prediction, and even generating synthesized tropes based on current generative deep neural networks.

Genre	colors	objects	places
#1 Drama (54.3%)	coal_black_color (0.865), black_color (0.827), gray_color (0.809), sage_green_color (0.805)	astronaut/spacesuit (0.915), furniture (0.801), building (0.789), device (0.785), religion_related (0.783)	hospital_room (0.805), hotel_room (0.801), kitchen (0.798), dressing_room (0.775)
#2 Adventure (30.0%)	black_color (0.953), coal_black_color (0.948)	land_dweller (0.975), central_dweller (0.974), spacesuit/astronaut (0.970), primitive_man (0.940), coal_miner/laborer (0.932)	aquarium (0.975), underwater (0.975), catacomb/grotto/cavern (0.954), corn_field (0.951)
#3 Thriller (27.8%)	indigo_color (0.675), coal_black_color (0.649), black_color (0.623)	President_of_the_United_States (0.649), official (0.633), stock_trader (0.578), soul_patch_facial_hair (0.550), underclassman (0.543)	discotheque (nightclub) (0.675), conference_center (0.636), television_studio (0.634), home_theater (0.613)
#4 Romance (27.1%)	claret_red_color (0.832), pale_yellow_color (0.810), pink_color (0.752)	girl (0.793), child (0.733), waitress (0.733), platinum_blond (0.714), couple (0.07)	dressing_room/beauty_salon (0.832), wedding_celebration (0.810), clothing_store (0.732), pub (0.691)
#5 Comedy (25.0%)	pink_color (0.818), claret_red_color (0.801), light_brown_color (0.800), Tyrian_purple_color (0.799)	female/woman (0.818), girl (0.801), waitress (0.786), couple (0.680)	home_theater (0.793), dressing_room/beauty_salon (0.792), bar/pub (0.725), nursing_home (0.712)
#6 Action (22.1%)	coal_black_color (0.757), ultramarine_color (0.725), black_color (0.700)	android (0.844), device (0.844), robot (0.772), oxygen_mask/aviator (0.764), plate/shield (0.764), armor (0.757), laser (0.755)	discotheque (0.755), engine_room (0.749), rock_climbing (0.728), airplane_cabin (0.701), elevator_shaft (0.697)
#7 Fantasy (22.1%)	black_color (0.896), coal_black_color (0.871)	land_dweller (0.924), fish (0.871), nature (0.871), animal (0.871), primitive_man (0.860)	underwater/aquarium (0.924), corn_field (0.884), bamboo_forest (0.866), jail_cell (0.864)
#8 Crime (17.9%)	black_color (0.465), coal_black_color (0.442), gray_color (0.441), jade_green_color (0.433)	boxing_ring (0.465), witness_box (0.447), compartment (0.447), man_with_shaven_head (0.446), stock_trader (0.441), motor_vehicle (0.434), mustachio/beard (0.420)	food_court (0.567), archive/server_room (0.483), boxing_ring (0.465), dressing_room (0.446), parking_garage (0.443)
#9 Sci-Fi (17.9%)	coal_black_color (0.957), black_color (0.957), gray_color (0.932), ultramarine_color (0.667)	spacesuit/astronaut (0.958), headdress/helmet (0.878), oxygen_mask/aviator (0.779), dashboard/electrical_device (0.758), fighter_pilot (0.741), robot (0.737)	cockpit (0.741), science_museum (0.667), discotheque (0.655), music_studio (0.608)
#10 Mystery (14.3%)	black_color (0.436), coal_black_color (0.420), reddish_brown_color (0.397)	toilet (0.436), device (0.420), railcar/vehicle (0.415), matrix (0.410), official (0.398)	office (0.457), jail_cell (0.436), elevator (0.415), conference_center (0.398)

Figure 8: Showing "weak tropes" characterized by salient colors, objects, and places across genres. Readers can observe that visual elements characteristic to different genres discovered via frequent itemset mining correspond to our perception of these genres. See more details in Section 7.2. Best seen in PDF.

8 DISCUSSION AND APPLICATIONS OF COMPUTATIONAL CREATIVITY

We have demonstrated the power of modern multimedia technology for very highly creative tasks such as trailer making as well as analyzed tropes, another creative device employed in movies. Here we lay out certain ideas and applications that can benefit from the synergy of computational creativity, people (as editors, consumers, and providers of data), and the massive data available today.

(1) **Intelligent Indexing and Curation for Documentaries and TV shows**: Like in movies, documentaries and TV shows also suffer from very high shooting ratios and require costly manual intervention to create final productions. Multimedia analysis could directly help with the organization and sorting of massive amounts of such video footage. Intelligent systems can also be built to learn domain specific information about them. For e.g. producers may like to see raw footage that is most different in content from that went into finished production of a certain show, or may want to view all comic/sad scenes, or the most surprising scene in a certain show.

(2) **Personalization of Media Content**: Researchers have begun to explore ways of learning user interests from their publicly available multimedia data on the social media forums.A compelling application of creating such interest profiles is to automatically identify prominent interest groups in a population (such as animal lovers, nature lovers etc.) and tailoring the semantic/emotional content of film trailers and TV shows to target interest communities.

(3) **Video Hyperlinking**: Video hyperlinking refers to the creation of interconnection between video sequence sharing related content. Fine grained multimodal video analysis approaches such as the ones described in this paper are required to address this challenging task with limited or no human effort. This research domain is receiving increasing attention from the multimedia community and beyond [2].

Second screen applications, displaying additional or complementary information about content visualized on the main screen are among the compelling applications encompassed by such emerging technologies. Video hyperlinking is genre agnostic so if a movie director wants to provide additional content to a scene or to an object/person it is sure to create additional value.

9 CONCLUSION

In this paper, we presented the great potential of intelligent multimedia technology in augmenting the highly creative task of making a movie trailer. We performed analysis on the genre of horror thriller movies and produced a trailer for a major 20^{th} century Fox production, Morgan. To the best of our knowledge, this is the *first ever* real collaboration between researchers in multimedia and the movie industry to jointly accomplish this highly manual and creative task for a real film. We demonstrated the tremendous value of AI as part of the creation process, while focusing on the edition of a

movie trailer, in terms of time and effort reduction. We evaluated the quality of our AI trailer with an extensive user study. Our AI trailer has been viewed around 3M times on YouTube. Finally, we explored applications of multimedia technology to another new creative paradigm, tropes that is commonly used in movies. This research investigation is the first of many into what we hope will be a very promising area of machine and human creativity especially in the arena of creative film editing. We're very excited about pushing the possibilities of how AI can augment the expertise and creativity of individuals.

ACKNOWLEDGMENTS

The authors would like to thank 20^{th} Century Fox for this great collaboration that lead to creation of the world's first joint human and machine made trailer for a full length feature film Morgan.

REFERENCES

[1] Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. 1993. Mining Association Rules Between Sets of Items in Large Databases. In *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data (SIGMOD '93)*. ACM, New York, NY, USA, 207–216.

[2] George Awad, Jonathan Fiscus, Martial Michel, David Joy, Wessel Kraaij, Alan F. Smeaton, Georges Quénot, Maria Eskevich, Robin Aly, Gareth J. F. Jones, Roeland Ordelman, Benoit Huet, and Martha Larson. 2016. TRECVID 2016: Evaluating Video Search, Video Event Detection, Localization, and Hyperlinking. In *Proceedings of TRECVID 2016*. NIST, USA.

[3] Damian Borth, Rongrong Ji, Tao Chen, Thomas Breuel, and Shih-Fu Chang. 2013. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *Proc. ACM Multimedia*. ACM, 223–232.

[4] Felix Burkhardt, Astrid Paeschke, M. Rolfes, Walter F. Sendlmeier, and Benjamin Weiss. 2005. A database of German emotional speech.. In *INTERSPEECH*. ISCA, 1517–1520.

[5] Shizhe Chen and Qin Jin. 2016. RUC at MediaEval 2016 Emotional Impact of Movies Task: Fusion of Multimodal Features. In *Working Notes Proceedings of the MediaEval 2016 Workshop, Hilversum, Netherlands, October 20-21, CEUR-WS.org*.

[6] Tao Chen, Damian Borth, Trevor Darrell, and Shih-Fu Chang. 2014. DeepSentiBank: Visual Sentiment Concept Classification with Deep Convolutional Neural Networks. *CoRR* abs/1410.8586 (2014). http://arxiv.org/abs/1410.8586

[7] David Crookes. 2011. The Science of the Trailer. (Aug 2011). http://www.independent.co.uk

[8] Ellen Douglas-Cowie, Roddy Cowie, Ian Sneddon, Cate Cox, Orla Lowry, Margaret McRorie, Jean-Claude Martin, Laurence Devillers, Sarkis Abrilian, Anton Batliner, Noam Amir, and Kostas Karpouzis. 2007. The HUMAINE Database: Addressing the Collection and Annotation of Naturalistic and Induced Emotional Data.. In *ACII* (2007-09-05) *(Lecture Notes in Computer Science)*, Ana Paiva, Rui Prada, and Rosalind W. Picard (Eds.), Vol. 4738. Springer, 488–500.

[9] Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller. 2013. Recent developments in opensmile, the munich open-source multimedia feature extractor. In *Proc. ACM Multimedia*. ACM, 835–838.

[10] Florian Eyben, Martin Wöllmer, and Björn Schuller. 2009. OpenEAR; Introducing the munich open-source emotion and affect recognition toolkit. In *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*. 1–6.

[11] Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proc. ACM Multimedia*. ACM, 1459–1462.

[12] Stephen Garrett. 2012. The Art of First Impressions: How to Cut a Movie Trailer. (Jan 2012). http://filmmakermagazine.com

[13] Michael Grimm, Kristian Kroschel, and Shrikanth Narayanan. 2007. Support Vector Regression for Automatic Recognition of Spontaneous Emotions in Speech.. In *ICASSP (4)*. IEEE, 1085–1088.

[14] Benoit Huet and Bernard Merialdo. 2006. *Automatic Video Summarization*. Springer Berlin Heidelberg, Berlin, Heidelberg, 27–42.

[15] Dhiraj Joshi, Ritendra Datta, Elena Fedorovskaya, Quang-Tuan Luong, James Z. Wang, Jia Li, and Jiebo Luo. 2011. Aesthetics and Emotions in Images: A Computational Perspective. *IEEE Signal Processing Magazine* 28, 5 (2011), 94–115.

[16] Yoshihiko Kawai, Hideki Sumiyoshi, and Nobuyuki Yagi. 2007. Automated production of TV program trailer using electronic program guide. In *CIVR*.

[17] Youngmoo E Kim, Erik M Schmidt, Raymond Migneco, Brandon G Morton, Patrick Richardson, Jeffrey Scott, Jacquelin A Speck, and Douglas Turnbull. 2010. Music emotion recognition: A state of the art review. In *Proc. ISMIR*. 255–266.

[18] David Kirby. 2016. The Role Of Social Media In Film Marketing. (June 2016). www.huffingtonpost.com

[19] Petros Koutras, Athanasia Zlatintsi, Elias Iosif, Athanasios Katsamanis, Petros Maragos, and Alexandros Potamianos. 2015. Predicting audio-visual salient events based on visual, audio and text modalities for movie summarization. In *Proc. ICIP*. IEEE, 4361–4365.

[20] Peter J Lang, Margaret M Bradley, and Bruce N Cuthbert. 1997. International affective picture system (IAPS): Technical manual and affective ratings. *NIMH Center for the Study of Emotion and Attention* (1997), 39–58.

[21] Rainer Lienhart, Silvia Pfeiffer, and Wolfgang Effelsberg. 1997. Video Abstracting. *Commun. ACM* 40, 12 (Dec. 1997), 54–62.

[22] D. Lin, S. Fidler, C. Kong, and R. Urtasun. 2014. Visual Semantic Search: Retrieving Videos via Complex Textual Queries. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*. 2657–2664.

[23] Ye Ma, Zipeng Ye, and Mingxing Xu. 2016. THU-HCSI at MediaEval 2016: Emotional Impact of Movies Task. In *Working Notes Proceedings of the MediaEval 2016 Workshop, Hilversum, Netherlands, October 20-21, CEUR-WS.org*.

[24] Olivier Martin, Irene Kotsia, Benoit M. Macq, and Ioannis Pitas. 2006. The eNTERFACE'05 Audio-Visual Emotion Database.. In *ICDE Workshops*, Roger S. Barga and Xiaofang Zhou (Eds.). IEEE Computer Society, 8.

[25] Mohammad Norouzi, Tomas Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrea Frome, Greg Corrado, and Jeffrey Dean. 2014. Zero-Shot Learning by Convex Combination of Semantic Embeddings. *International Conference on Learning Representations (ICLR)* (2014).

[26] Soujanya Poria, Erik Cambria, Newton Howard, Guang-Bin Huang, and Amir Hussain. 2016. Fusing audio, visual and textual clues for sentiment analysis from multimodal content. *Neurocomputing* 174 (2016), 50–59.

[27] Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. 2015. A Dataset for Movie Description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[28] BjÃűrn Schuller, Dejan Arsic, Gerhard Rigoll, Matthias Wimmer, and Bernd Radig. 2007. Audiovisual Behavior Modeling by Combined Feature Spaces.. In *ICASSP (2)*. IEEE, 733–736.

[29] Björn Schuller, Gerhard Rigoll, and Manfred Lang. 2003. Hidden Markov model-based speech emotion recognition. In *Multimedia and Expo, 2003. ICME'03. Proceedings. 2003 International Conference on*, Vol. 1. IEEE, I–401.

[30] BjÃűrn W. Schuller, Ronald MÃijller, Florian Eyben, JÃijrgen Gast, Benedikt HÃűrnler, Martin WÃűllmer, Gerhard Rigoll, Anja HÃűthker, and Hitoshi Konosu. 2009. Being bored? Recognising natural interest by extensive audiovisual integration for real-life application. *Image Vision Comput.* 27, 12 (2009), 1760–1774.

[31] Nicu Sebe, Ira Cohen, Theo Gevers, and Thomas S Huang. 2006. Emotion recognition based on joint visual and audio cues. In *Proc. ICPR*, Vol. 1. IEEE, 1136–1139.

[32] Mats Sjöberg, Yoann Baveye, Hanli Wang, Vu Lam Quang, Bogdan Ionescu, Emmanuel Dellandréa, Markus Schedl, Claire-Hélène Demarty, and Liming Chen. 2015. The MediaEval 2015 Affective Impact of Movies Task.. In *MediaEval*.

[33] Alan F. Smeaton, Bart Lehane, Noel E. O'Connor, Conor Brady, and Gary Craig. 2006. Automatically Selecting Shots for Action Movie Trailers. In *Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval (MIR '06)*. ACM, New York, NY, USA, 231–238.

[34] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2016. MovieQA: Understanding Stories in Movies through Question-Answering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[35] Ba Tu Truong and Svetha Venkatesh. 2007. Video Abstraction: A Systematic Review and Classification. *ACM Trans. Multimedia Comput. Commun. Appl.* 3, 1, Article 3 (Feb. 2007).

[36] Michel Valstar, Björn Schuller, Kirsty Smith, Florian Eyben, Bihan Jiang, Sanjay Bilakhia, Sebastian Schnieder, Roddy Cowie, and Maja Pantic. 2013. AVEC 2013: the continuous audio/visual emotion and depression recognition challenge. In *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*. ACM, 3–10.

[37] Victoria Yanulevskaya, Jan C van Gemert, Katharina Roth, Ann-Katrin Herbold, Nicu Sebe, and Jan-Mark Geusebroek. 2008. Emotional valence categorization using holistic image features. In *Proc. ICIP*. IEEE, 101–104.

[38] Zhihong Zeng, Maja Pantic, Glenn I Roisman, and Thomas S Huang. 2009. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 31, 1 (2009), 39–58.

[39] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. 2014. Learning Deep Features for Scene Recognition using Places Database. In *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger (Eds.). 487–495.

[40] Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV) (ICCV '15)*. 19–27.