Modeling Narrative Structure in Advertisement Videos

PI: Adriana Kovashka
Affiliation: University of Pittsburgh, Computer Science Department, School of Computing and Information
Mailing address: 210 S. Bouquet Street, Sennott Square Room 5325, Pittsburgh, PA 15260, USA
Email address: kovashka@cs.pitt.edu Phone number: 412-624-8852
Google sponsors, contacts: Jesse Berent (Google Zurich), Andrew Gallagher (Google Mountain View)

Video advertisements often use creative narrative structures. We will first examine the dynamics of these narratives, using different cues to plot how pacing and emotional intensity evolve over the course of the ad. We will use these plots to predict where in the ad a climax occurs. We will then examine narratives at a more semantic level, by predicting whether the ad has a protagonist, and what this protagonist desires. We will learn to associate low-level actions to high-level intent, with the help of humans who perform reasoning and rely on visual processing by our system.

1. Introduction

Video advertisements are powerful tools for affecting the public opinion. To achieve persuasive power, many ads explore creative narrative techniques. One classic technique is "Freytag's pyramid" where a story begins with exposition, followed by rising action, then climax, concluding with denouement [13, 4]. Often the events in the narrative relate to the experiences of one or many protagonists and their journey captured in the ad. We will analyze the narrative structure of ads at multiple levels. First, at the lower level we will analyze the dynamic nature of the story told in the ad. Second, at the higher level, we will model the role of characters in the ad.

Problem statement. We want to accomplish four concrete tasks: (1) plot how pacing progresses over the course of the ad; (2) plot how the emotional intensity in the ad progresses; (3) detect what type of central character we have in the ad, and whether he/she is a *protagonist*; (4) learn to predict what the protagonist wants, with the help of a human collaborator, by learning to map low-level actions to higher-level intents. Both understanding the dynamics of the ad, and the experiences of the protagonist, are instrumental in understanding an ad's message.

To determine how well the plots we generate reflect the dynamics of the ad, we will compare whether their shape agrees with the shape of a "ground-truth" plot obtained from captured human facial expressions. We will also verify whether a peak corresponds to human-labeled climactic moments.

We define "protagonist" as a central character whose journey and experiences the ad portrays. Based on examination of a random sample of 100 ads in our video ads dataset [5], we identified three types: ads without a central character, ads with a central character who simply delivers a message lecture-style, and ads with a protagonist.

Once we identify whether we have a protagonist, we will model the goals and desires of this protagonist. Some desires in the ads in our random sample are as follows, ranging from very concrete to abstract: build a crib, hunt ducks, get the chips, get a ball in a basket, perform well (for a singer), survive, become free from guilt, play by one's own rules, stay fresh, help sister be happy, be happy, etc. Inferring goals is a challenging task and to make initial progress, we will involve a human in the loop for training purposes. Examination of our random subset of video ads indicates that about half of the ads have a protagonist, and about two-thirds of those include the protagonist's journey towards achieving their desire. Note that the tasks of predicting dynamics and modeling protagonists and their desires are linked– the moment when the protagonist's desires come true often coincides with the climax of the ad.

Prior work. We assembled a dataset of 64,832 image ads and 3,477 video ads. These cover diverse categories, e.g. automobile ads, food and beverage ads, political ads, etc. Some of the annotations include the topic (product advertised), sentiment the ad provokes, binary labels for whether the ad is (a) exciting and (b) funny, and questions and answers to the question "What should I do according to this ad, and why should I do it?". We achieve 35.1% and 32.8% accuracy on the topic and sentiment prediction tasks, and 78.6%/78.2% on the funny/exciting tasks, using features extracted from a network fine-tuned on UCF101. Using pre-training on a subset of the YouTube-8M videos, we achieve a boost for topic and sentiment prediction (39.2% and 34.3%, resp.). Interestingly, there was not much benefit to modeling the speech in the videos, likely because automated tools for speech recognition were inaccurate.

Novelty. No prior work provides a method for understanding the narrative rhetoric of video ads; our [5] presents a simple baseline. Our work enables many applications, including understanding of bias in the media, generating human-like descriptions of media content for the visually impaired, better ad targeting, etc.

There is limited prior work in analyzing the narratives of videos. [3] create video stories out of consumer videos, while our task is more difficult since rather than compose existing elements, we must analyze unsegmented video and

infer what regions correspond to what parts of the story (e.g. climax). We are not aware of any work that models the associations humans harbor linking low-level actions and high-level intent, as we propose in Sec. 3. Related work models roles in film [12] and motivations in images [11].

2. Plotting dynamics

Video ads are carefully designed to guide the viewer's attention and provoke an emotional response [13]. Often the emotional charge evolves over the course of the ad [8]. The pacing of the ad might change as well, e.g. an ad might start out slower and build momentum before delivering its message. We are interested in modeling pacing and emotional intensity dynamics in their own right, but also as a cue for detecting the climax of an ad video.

Pacing. While there is no prior work on detecting climax in ads, existing approaches model the tempo and dynamics of general videos [7, 10, 1]. For example, [7] use cues like "motion intensity" and "audio pace" to detect action scenes. An action scene might be one that contains many short shots, indicating a fast pace of action. Another cue for an action scene is one that portrays complex motion, so [7] model complexity using entropy over the directions of motion vectors. [10] model the amount of disturbance as the fraction of moving pixels, and [1] the average magnitude of motion vectors of objects in a shot. Based on this work, we will separately plot: (1) the number of shot changes per second, (2) the magnitude of optical flow vectors, (3) the average speed of moving objects in a "tube" containing a single tracked object, (4) the number of moving pixels per second, (5) entropy over the directions of moving pixels, (6) the amplitude of sound, and (7) the number of spoken words. We will then extract the video subset corresponding to peaks in some combination of these plots (e.g. average or max).

Emotional intensity. The climactic part of an ad is likely one that humans have a strong emotional reaction to. In prior work, [8] capture human facial expressions in response to video ads, and discover that ad effectiveness is optimized when the brand is shown shortly before a positive emotion is elicited. Thus, we aim to predict the emotional responses that ads might provoke in the audience. We will plot the intensity of positive and negative emotions throughout the course of the ad. To predict emotionality, we will use [9, 2, 6] to predict the emotional charge of *images*. As before, we will identify climax as the region of emotional intensity in the ad. The benefit of using emotionality as a cue for climax is that we can leverage large-scale image data to model emotionality. In contrast, the number of video ads is limited, so we might not be able to directly train a climax detector.

Resources and evaluation. We model emotions using the image that the viewer sees. In contrast, [8] predict the emotions of the viewer from an image of her face. To test how well we have detected the emotional progress of the ad, using the ads in the dataset of [8], we will compare our emotion intensity plot to that of [8]. Treating moments of rising positive valence as "positives", we will compute the precision and recall of our method.

Next, we will compare to what extent our plots are a useful indicator of climax. We will first develop a dataset to study the climax detection problem. Using our collection of 3,477 ads, as well as an additional 2,000 we have downloaded but not annotated, we will ask annotators to mark the beginning and end of the part they deem most climactic. We will compute the temporal overlap of the video parts that five annotators marked, and consider as "climax" those frames that at least three annotators marked climactic. To make the climax detection task more challenging, we will require the system to say *if* there is climax in the video, and if so, where it is. We will feed to the system videos that may have been cut short. Thus, a simple baseline that predicts the penultimate or final shot of a video to be the climax will not do well, and we will be better able to judge if a system truly understands climax and the ad.

3. Modeling characters and what they desire

While the above are useful cues, whether we are at the punchline or climax of an ad is unlikely to be fully addressed by them. Narratives often cover the answers to basic questions known as the "five W's": *who, what, when, where, why*. We will focus on *who* the main character is, and *what* they want (alternatively *why* they are doing *what* they are in the ad). Commercials often appeal to desires that members of the audience might have, via a character the audience can relate to. For example, some commercials in our dataset appeal to the viewer's desire to feel more powerful, loved, attractive, etc. When that desire is achieved thanks to the product is the 'aha' moment that delivers the message of the ad. Ultimately we can use desire to also predict climax, but here we will analyze characters and desire as end-goals.

We first need to analyze the type of characters in ads, i.e. whether the concept of "protagonist's desire" even applies. An examination of 100 randomly chosen ads in our dataset indicates that ads can be broken down in the following way. Some ads feature a central character who has a certain desire, and the ad briefly traces whether this desire comes true or not. For example, in a popular Volkswagen ad, a child in a Darth Vader costume attempts to exert "the force" and magically turn on devices with his hands, but fails until his dad secretly uses his remote to turn the headlights of his Volkswagen on. In the Allegro English for Beginners ad, a Polish grandfather learns English so he can utter

the words "I am your grandpa" to his newborn grandson. The second type of ad features a central character but that character simply *tells* the audience about the qualities of the product, rather than *showing* through her own experience. A third type of ad is one where there are either no human or animated characters, or too many to identify a central protagonist. We will first train a neural net to predict in which of the above three cases an ad falls. We will use as cues the consistency of the same recognized face in many frames in the ad, whether or not a character primarily talks, and whether her body pose is static. We also need to determine *who* the protagonist is. In this project, we will simply assume that the character which most frequently appears in the video is the protagonist.

Next, we will focus on ads that have a protagonist, and examine what she wants. Inferring what a character wants is a complex task which requires solving reasoning, which is beyond the scope of this project. Thus, the proposed system will perform basic visual understanding, while the reasoning will temporarily be left to our human collaborator. The system will begin by detecting shot boundaries in the ad video, then pick a representative frame in each shot, and use an existing captioning algorithm to provide captions for these frames. These descriptions, without the frames, will be shown to a human helper. She can then attempt to infer what the protagonist wants, or can ask questions to help in finding the answer. Useful questions might be about the objects that the protagonist is interacting with, the type of setting she is in, any other people around, and what the objects/scenery/people look like (e.g. broken/fresh/luxurious, dark/bright, happy/sad/despondent/energetic, etc.). It is likely feasible for the machine to answer simple questions with some level of accuracy. We can use existing VQA systems or separate systems for different types of questions (e.g. objects, scenes) and train a classifier to adaptively choose which system to use. We will provide the human with confidence levels about each question, so she knows which answers to ignore. Equipped with answers to these questions, our human can perform reasoning about what the combination of objects/scenes/etc. *mean* about *what the protagonist wants*. For example, our helper might conclude that when you see a human running, then a shot of a bus, the human is probably trying to catch the bus. In contrast, if you see a human running followed by a shot of a lion, the human is probably trying to escape from the lion and save her life.

By involving a human in the reasoning loop, we can learn associations between (1) low-level actions and context, and (2) high-level intent. Video ads are an especially appropriate type of data for this since they are brief, so they must utilize associations that are simple to encode and quick to decode. The next step is to learn to make these associations like a human would. We likely need a lot of data to learn successful associations, i.e. why do '[A and B] implies F', and '[D and B] implies F', but '[E and B] does not imply F'? We can try swapping A/B with some other objects/scenes/attributes and obtain a label, e.g. this combination does/does not imply F. We can crowdsource such labels efficiently without any need for visual data, which is limited in the case of expensive-to-make video ads.

Resources and evaluation. We will develop a vocabulary of desires, and obtain annotations for each ad as options from that vocabulary. We will first collect free-form text, then manually quantize the results, discarding ads where we cannot obtain consistent responses from three annotators. We will compare (1) our human-machine collaboration (note the human helper does not see the ads so might produce an incorrect response), and (2) an existing method that predicts motivations in *images* [11], in terms of the extent to which they can retrieve the ground-truth responses.

References

[1] S. Benini, P. Migliorati, and R. Leonardi. A statistical framework for video skimming based on logical story units and motion activity. In *International Workshop on Content-Based Multimedia Indexing*, 2007.

[2] D. Borth, R. Ji, T. Chen, T. Breuel, and S.-F. Chang. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *ACM Multimedia*, 2013.

[3] J. Choi, T.-H. Oh, and I. So Kweon. Video-story composition via plot analysis. In *CVPR*, 2016.

[4] G. Freytag. *Freytag's technique of the drama: an exposition of dramatic composition and art*. Scholarly Press, 1896.

[5] Z. Hussain, M. Zhang, X. Zhang, K. Ye, C. Thomas, Z. Agha, N. Ong, and A. Kovashka. Automatic understanding of image and video advertisements. In *CVPR*, 2017.

[6] R. Kosti, J. M. Alvarez, A. Recasens, and A. Lapedriza. Emotion recognition in context. In *CVPR*, 2017.

[7] A. Liu, J. Li, Y. Zhang, S. Tang, Y. Song, and Z. Yang. An innovative model of tempo and its application in action scene detection for movie analysis. In *WACV*, 2008.

[8] D. McDuff, R. El Kaliouby, J. F. Cohn, and R. W. Picard. Predicting ad liking and purchase intent: Large-scale analysis of facial responses to ads. *IEEE Transactions on Affective Computing*, 6(3):223–235, 2015.

[9] K.-C. Peng, A. Sadovnik, A. Gallagher, and T. Chen. Where do emotions come from? predicting the emotion stimuli map. In *ICIP*, 2016.

[10] Z. Rasheed and M. Shah. Movie genre classification by exploiting audio-visual features of previews. In *ICPR*, 2002.

[11] C. Vondrick, D. Oktay, H. Pirsiavash, and A. Torralba. Predicting motivations of actions by leveraging text. In *CVPR*, 2016.

[12] C.-Y. Weng, W.-T. Chu, and J.-L. Wu. Rolenet: Movie analysis from the perspective of social networks. *IEEE Transactions on Multimedia*, 11(2):256–271, 2009.

[13] C. E. Young. *The advertising research handbook*. Ideas in Flight, 2008.

Data policy: We will share the climax and character desire annotations, and a database of associations that we learn, at a publicly accessible link.

Budget ($42,900 total): The PI has funding for understanding *image* ads as NSF awards 1566270 and 1718262, but video ads pose unique, different challenges. In particular, static images do not show a narrative evolve. We request:

- $27,600 in supporting a PhD student for 12 months, plus $13,800 in fringe benefits; and
- $1,500 in conference travel for the student.
- (The PI will use her startup funds for an estimated $10,000 in annotation costs on Amazon Mechanical Turk.)

4. Results of past projects funded by Google

A 2016 Google Faculty Research Award funded [5], our first paper on understanding image and video ads. The Google award funded the work in video ads. We assembled a dataset of 3,477 ads. About 1,000 of these were provided by our Google sponsor, and the rest were crawled using appropriate keywords. These video ads were annotated with the following labels on Amazon Mechanical Turk: Topic (e.g. Cars and automobiles, Safety), Sentiment (e.g. Cheerful, Amazed), Action/Reason (e.g. I should buy this car because it is pet-friendly.), binary "Is this ad funny?", "Is this ad exciting?", and "Is this ad in English?" labels, and worker ratings of the effectiveness of the ad. We showed baseline results for all tasks in [5]. We released our dataset (video IDs and annotations) for public access in April 2017. Our project page and dataset have been accessed **968 times** since the dataset's release (513 sessions from the United States from **22 different states**, 78 from Japan, 70 from China, 58 from India, and additional visits from **40 other countries**). Our work was presented as a spotlight at CVPR 2017.

The sponsors and the PI have communicated several times after the award was made, and the sponsors provided useful guidance. The PI visited Google Pittsburgh and gave a talk in 2016, but did not discuss details of this particular project as the paper was still under review. The PI and sponsors discussed opportunities for the PI's students to intern with the sponsors so that deeper collaboration can be established. One of the PI's students who was a co-author on [5] has already interned at Google twice, but his work was unrelated to this project.

Adriana Ivanova Kovashka

Assistant Professor in Computer Science at University of Pittsburgh
kovashka@cs.pitt.edu
http://people.cs.pitt.edu/∼kovashka
http://scholar.google.com/citations?hl=en&user=Dl949GoAAAAJ

EDUCATION

The University of Texas at Austin (Austin, TX): August 2008 - August 2014 – PhD, M.S.Comp.Sci.
Major: Computer Science, Concentration: Computer Vision / Artificial Intelligence

Pomona College (Claremont, CA): August 2004 - May 2008 – Bachelor of Arts
Majors: Computer Science and Media Studies, Minor: German

APPOINTMENTS

University of Pittsburgh, Computer Science Department – Assistant Professor – January 2015 - now.

UT Austin, Computer Science Department – Graduate Research Assistant – June 2009 - August 2014.

CONFERENCE PUBLICATIONS

- Bhavin Modi and Adriana Kovashka. "Confidence and Diversity for Active Selection of Feedback in Image Retrieval." To appear, *Proceedings of the British Machine Vision Conference* (BMVC), September 2017.
- Zaeem Hussain, Xiaozhong Zhang, Mingda Zhang, Keren Ye, Christopher Thomas, Zuha Agha, Nathan Ong, Adriana Kovashka. "Automatic Understanding of Image and Video Advertisements." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), July 2017. **(Spotlight)**
- Nils Murrugarra-Llerena and Adriana Kovashka. "Learning Attributes from Human Gaze." In *Proceedings of IEEE Winter Conference on Applications of Computer Vision* (WACV), March 2017.
- Debashis Ganguly, Mohammad H. Mofrad and Adriana Kovashka. "Detecting Sexually Provocative Images." In *Proceedings of IEEE Winter Conference on Applications of Computer Vision* (WACV), March 2017.
- Christopher Thomas and Adriana Kovashka. "Seeing Behind the Camera: Identifying the Authorship of a Photograph." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), June 2016.
- Siqi Liu and Adriana Kovashka. "Adapting Attributes by Selecting Features Similar across Domains." In *IEEE Winter Conference on Applications of Computer Vision* (WACV), March 2016.
- Adriana Kovashka and Kristen Grauman. "Attribute Adaptation for Personalized Image Search." In *Proceedings of the International Conference on Computer Vision* (ICCV), December 2013.
- Adriana Kovashka and Kristen Grauman. "Attribute Pivots for Guiding Relevance Feedback in Image Search." In *Proceedings of the International Conference on Computer Vision* (ICCV), December 2013.
- Adriana Kovashka, Devi Parikh, and Kristen Grauman. "WhittleSearch: Image Search with Relative Attribute Feedback." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), June 2012.
- Devi Parikh, Adriana Kovashka, Amar Parkash, and Kristen Grauman. "Relative Attributes for Enhanced Human-Machine Communication" (Invited paper). In *Proceedings of the AAAI Conference on Artificial Intelligence* (AAAI), July 2012.
- Adriana Kovashka, Sudheendra Vijayanarasimhan, and Kristen Grauman. "Actively Selecting Annotations Among Objects and Attributes." In *Proceedings of the International Conference on Computer Vision* (ICCV), November 2011.
- Adriana Kovashka and Kristen Grauman. "Learning a Hierarchy of Discriminative Space-Time Neighborhood Features for Human Action Recognition." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), June 2010.

- Sindhu Raghavan, Adriana Kovashka, and Raymond Mooney. "Authorship Attribution Using Probabilistic Context-Free Grammars." In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (ACL), Short Papers, July 2010.

JOURNAL PUBLICATIONS

- Adriana Kovashka, Devi Parikh, and Kristen Grauman. "WhittleSearch: Interactive Image Search with Relative Attribute Feedback." In *International Journal of Computer Vision* (IJCV), 2015.
- Adriana Kovashka and Kristen Grauman. "Discovering Attribute Shades of Meaning with the Crowd." In *International Journal of Computer Vision* (IJCV), 2015.

BOOK CHAPTERS AND SURVEYS

- Adriana Kovashka and Kristen Grauman, "Attributes for Image Retrieval," In *Visual Attributes (Advances in Computer Vision and Pattern Recognition)*, Springer, 2016.
- Adriana Kovashka, Olga Russakovsky, Kristen Grauman, Fei-Fei Li, "Crowdsourcing in Computer Vision. In *Foundations and Trends in Computer Graphics and Vision*, NOW Publishers, 2016.

FUNDING

- Title: "Decoding Video Advertisements", Role: PI, Funding agency: Google Faculty Research Awards, Amount: $41,361, Date awarded: February 8, 2016
- Title: "CRII: RI: Automatically Understanding the Messages and Goals of Visual Media", Role: PI, Funding agency: NSF, Amount: $174,590, Duration: June 1, 2016 to May 31, 2018, Date awarded: May 27, 2016
- Title: "Deeply Interactive Image Search: Learning Attribute Vocabularies and Spatial Support from People", Role: PI, Funding agency: University of Pittsburgh Central Research Development Fund (CRDF), Amount: $15,749, Duration: July 31, 2016 to June 30, 2018, Date awarded: June 24, 2016
- Title: "RI: Small: Modeling Vividness and Symbolism for Decoding Visual Rhetoric", Role: PI, Funding agency: NSF, Amount: $449,978, Duration: August 1, 2017 to July 31, 2020, Date awarded: July 27, 2017

PROFESSIONAL SERVICE

- **Area Chair**, CVPR 2018, WACV 2017, WACV 2016, ICVGIP 2016.
- **Tutorials Chair**, WACV 2018.
- **Chair / co-chair**, CVPR 2015, 2016, 2017 Doctoral Consortium.
- **Session Chair**, CVPR 2016, WACV 2016.
- **Judge**, LDV Vision Summit 2016 Computer Vision Challenge.
- **Organizer**, CVPR 2015 Workshop "Women in Computer Vision".
- **Organizer**, ECCV 2014 Workshop "Human-Machine Communication for Visual Recognition & Search".
- **Senior Program Committee Member**, IJCAI 2016.
- **Panelist**, National Science Foundation (NSF).
- **Program Committee / Reviewer**, CVPR 2013, 2015 (Outstanding Reviewer Award), 2016, 2017 (Outstanding Reviewer Award); ICCV 2013, 2015, 2017; ECCV 2012, 2014, 2016; ICLR 2018; NIPS 2014, 2015, 2016; IJCV 2013; TPAMI 2014, 2015, 2017; AAAI 2014; HCOMP 2012; UIST 2015.

PATENTS

- "Efficiently identifying images, videos, songs or documents most relevant to the user using binary search trees on attributes for guiding relevance feedback," Inventors: Kristen Grauman, Adriana Kovashka, Owner: Board of Regents, The University of Texas System, Issued: November 3, 2015
- "Efficiently identifying images, videos, songs or documents most relevant to the user based on attribute feedback," Inventors: Kristen Grauman, Adriana Kovashka, Devi Parikh, Owner: Board of Regents, The University of Texas System, Issued: March 22, 2016