10-2020

ViScene: A collaborative authoring tool for scene descriptions in videos

Rosiana NATALIE

Ebrima JARJUE

Hernisa KACORRI

Kotaro HARA

ViScene: A Collaborative Authoring Tool for Scene Descriptions in Videos

ROSIANA NATALIE, Singapore Management University, Singapore
EBRIMA JARJUE, University of Maryland, College Park, USA
HERNISA KACORRI, University of Maryland, College Park, USA
KOTARO HARA, Singapore Management University, Singapore

Audio descriptions can make the visual content in videos accessible to people with visual impairments. However, the majority of the online videos lack audio descriptions due in part to the shortage of experts who can create high-quality descriptions. We present ViScene, a web-based authoring tool that taps into the larger pool of sighted non-experts to help them generate high-quality descriptions via two feedback mechanisms—succinct visualizations and comments from an expert. Through a mixed-design study with $N = 6$ participants, we explore the usability of ViScene and the quality of the descriptions created by sighted non-experts with and without feedback comments. Our results indicate that non-experts can produce better descriptions with feedback comments; preliminary insights also highlight the role that people with visual impairments can play in providing this feedback.

Additional Key Words and Phrases: Scene description, visual impairment, video accessibility

ACM Reference Format:
Rosiana Natalie, Ebrima Jarjue, Hernisa Kacorri, and Kotaro Hara. 2020. ViScene: A Collaborative Authoring Tool for Scene Descriptions in Videos. In *Proceedings of ASSETS '20: ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '20).* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Videos that rely on visuals to convey information are inaccessible to people with visual impairments [2, 11, 17]. Audio descriptions (AD), verbal commentaries of the visual information in videos [6, 15], can increase their accessibility. However, providing high-quality AD is challenging because hiring professional video describers is costly and time-consuming [18]. Kobayashi found AD generated by minimally trained audio describer received low intelligibility [9]. Prior studies have explored technical solutions to minimize the cost of hiring professional describers and achieve good quality of audio description. For example, Kobayashi *et al.* designed an audio description authoring tool to streamline the audio description editing [8], and Yuksel *et al.* designed a machine learning-based tool to increase process efficiency [18]. However, no prior work investigated the efficacy of generating high-quality AD through collaboration between a novice audio describer and an expert—that be a sighted expert or blind users with lived expertise.

We investigate the efficacy of drawing from a larger pool of sighted non-experts to generate high-quality AD by eliciting experts' feedback through our AD authoring tool, called ViScene. As shown in Figure 1, ViScene is an interactive web-based system for people to collaboratively write scene descriptions (SD)—textual descriptions of scenes in a video that are converted into audio descriptions through text-to-speech (TTS).

Manuscript submitted to ACM

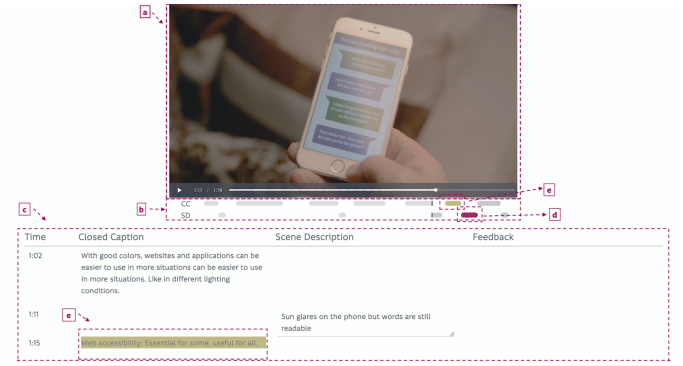

Fig. 1. ViScene's interface include : (a) the video; (b) close captions (CC) and scene descriptions (SD) bars; (c) a table with Time, CC, SD, and Feedback columns; (d) SD succinctness feedback; and (e) text highlighting to show CC/SD text-segment correspondence.

ViScene's interface includes the video pane, segment bars in parallel with the progress bar indicating where closed caption (CC) and scene descriptions appear in the video, and a table pane with details on the timing, close captions and descriptions, as well as segment-level feedback comments. ViScene has two feedback mechanisms—succinctness visualization and expert feedback. Non-experts can write the descriptions in segments where there is no audio (*i.e.,* no closed captions). Descriptions are rendered through Amazon Polly TTS service [1]. Once the ViScene receives the generated speech, it computes the length of the audio and visualizes it as a gray segment in the SD bar. A segment turns red if its duration overshoots the space between two adjacent CC segments, indicating the need for a more succinct description. Expert's comments per segment are shown on the right.

2 STUDY METHOD

We assessed the efficacy of supporting non-experts to provide SD with ViScene through a preliminary remote user study. We used Zoom for our study due to COVID-19 restrictions. We invited participants to two sessions, in which they were asked to use ViScene to create SDs for the same video. We recruited six participants (three females and three males) via a listserv. All participants were university students, aged 22 to 41 (*mean*=27.5, *SD*=7.29). None had prior experience in writing SDs. We used W3C's video (98s) explaining color with good contrast [16] because it: (i) was relatively short and (ii) had W3C-provided SD, which we could treat as a reasonable ground truth.

The study was a 2x2 mixed-design study with session as a repeated factor (*session 1* vs. *session 2*) and presence of expert feedback as a between-subjects factor (*with-feedback* vs. *without-feedback*). In the first session, we introduced what AD and SD are, the motivation of the research, and the task. We briefed participants on ViScene's interface; participants then used it to write SDs after watching the video at least once. Between *session 1* and *session 2*, a sighted research team member evaluated the quality of SDs using the codebook in Table 1. The evaluation took about fifteen to thirty minutes. After the evaluation, participants were invited after 1-2 days to revise the SDs they wrote. We randomly assigned the participants equally to each condition. For the three participants who were in the *without-feedback* condition, the feedback was not visible while they revised their SD. The other three in the *with-feedback* condition could see the expert feedback (Figure 1). At the end of each session, we administered participants with the SUS questionnaire to assess the usability of ViScene. We analyzed the quality of SDs and ViScene's usability.

SD Quality. As there is no established method to assess the SD's quality, we triangulated results from three approaches: (i) assessment of similarity between ground truth and participant-generated SDs; (ii) codebook-based

	Audio Description Property / Feature
Descriptive	Descriptive, objective and accurate [5, 13]
Succinct	Succinct, The description fitted into the natural pause in the video's dialogue. [3]
Learning	Prioritized to the intended learning and enjoyment outcomes [5]
Equal*	Equal access requires that the meaning and intention of the program be conveyed. [5]
Sufficient	Contain sufficient amount of information (i.e., not too much, not too less) [12]
Interest	Allowing the blind people gain interest in the content provided [17]
Confusion	Prevent the blind people to elicit their confusion [17]

Table 1. Evaluation codebook used by a sighted and a blind researcher to assess the quality of audio descriptions generated by the participants; the Equal code (marked with *) was the only one not used by the blind evaluator.

evaluation by a trained sighted evaluator; and (iii) codebook-based evaluation by a blind evaluator. Both the sighted and blind evaluators were co-authors of the paper. To quantify the semantic similarity between the ground truth SD and participant-generated SDs, we used embedding and cosine similarity [14]. We used an embedding technique called Doc2Vec [10, 19] that mapped a document—a set of words in SDs—into a dense vector ($\vec{v} \in \mathbb{R}^{300}$). We compute a cosine similarity [7] between vectors that represent ground truth SD and participant-generated SD. Cosine similarity, an oft-used document similarity metric, returned a value between [0, 1], where 1 indicated that two documents are similar.

We developed a codebook to evaluate the quality of SD via literature review [3, 5, 12, 13, 17]. Seven codes emerged (Table 1). In the assessment by a sighted evaluator, five codes (Descriptive, Succinct, Learning, Equal, Sufficient) were used to assess the quality of SDs. In the assessment by a blind expert, six codes (Descriptive, Succinct, Learning, Sufficient, Interest, Confusion) were used to assess the quality of SDs. The sighted evaluator and blind expert did not use the same codes because blind people could not justify the equality of the provided information without seeing the video (i.e., Equality code). We approved and counted SDs that possessed the characteristic described in the codebook.

Usability. We evaluated the usability of the interface with the System Usability Scale (SUS) that consists of ten 5-Likert scale questions [4]. The response to these items were mapped to numerical scores with a range of 0 to 100. We also recorded time taken by participants to complete the task to assess its complexity.

3 RESULTS

SD Quality. We observed an increase in the similarity between the ground truth and participant generated SD in the *with-feedback* condition but not in the *without-feedback* condition. The cosine similarities in the *with-feedback* condition were 0.750 (*SD*=0.09) and 0.804 (*SD*=0.06) in *session 1* and *session 2*, respectively. Likewise, the cosine similarities were 0.686 (SD=0.07) and 0.688 (SD=0.07) in the *without-feedback* condition.

Qualitative analysis by the sighted evaluator showed that all participants in the *with-feedback* condition satisfied all quality criteria after the revision. Figure 3 indicates that they improved the Learning, Equal, and Sufficient qualities in *session 2*. The results from the participants in the *without-feedback* condition were mixed; some improved the Succinctness and Learning, but Sufficient quality decreased. Only Succinct criterion was satisfied by all three participants.

The blind evaluator mentioned that all SDs were succinct and helped him to understand the learning objective of the video, but some SDs were not sufficiently descriptive. The blind evaluator expected the SD to describe all the scenes in more details. For example, the evaluator wanted to know what the actors were wearing, time of the day in the scene, and location (*e.g.,* outdoor *vs.* indoor). The missing information caused confusion and loss of interest in the video.

Usability, Completion Time, Participant Experience. The SUS score for *without-feedback* condition in *session 1* and *session 2* were 80 and 81.7, respectively. For *with-feedback* condition, the SUS scores were 85.8 and 86.7 in *session 1* and *session 2*. These figures indicate that there is no obvious changes in usability and the usability is *excellent*. The

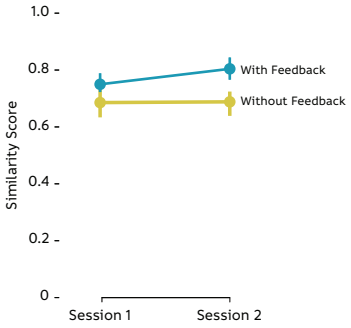

Fig. 2. A slopechart of SD similarity. Increase in the score shows that comments make SDs similar to the ground truth.

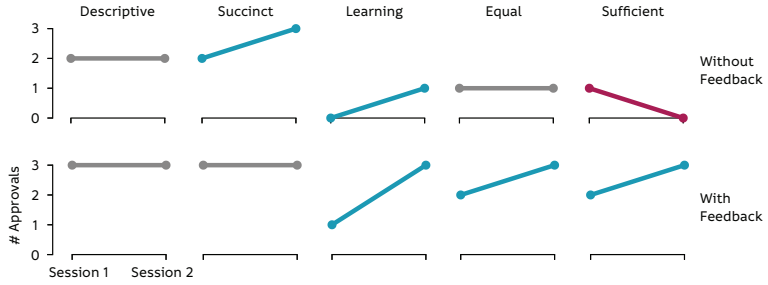

Fig. 3. Slope charts of five qualities of SDs in the first and second session for the with- and without-feedback groups. The qualities were assessed by a sighted author.

semi-structured interviews revealed that features, such as the succinctness visualization and the ability to hear the synthetic speech dynamically helped the participants to complete the task.

In *session 1*, the task completion time in both conditions *(with-/without-feedback)* were similar, but those in *without-feedback* condition spent slightly more time (*with-feedback*: mean=818.3s; *without-feedback*: 996.7s). In *session 2*, while those in *without-feedback* condition spent less time compared to *session 1*, the participants in the *with-feedback* condition took more time to complete the task (*with-feedback*: mean=1240s; *without-feedback*: mean=544.0s). Those in *with-feedback* condition made 3.48 changes on average in *session 2*. In contrast, those in *without-feedback* condition made 1 change on average, making minor changes like correcting grammatical errors.

4 DISCUSSION AND CONCLUSION

We showed the utility of comments from an experienced sighted evaluator in improving the SD quality. Preliminary results indicate that with the comments, participants were able to make their SD more similar to the W3C's ground truth (an increase in similarity by 7.2%). Both sighted and blind evaluators agreed that the generate SDs were succinct and prioritized to the intended learning and enjoyment outcomes of the video. While the sighted evaluator indicated five participants provided descriptive SDs, the blind evaluator noted that all participants' SDs had insufficient details. This indicates that for some criteria, sighted and blind evaluators have different sense of what is considered "good." The sighted evaluator approved the balance between succinctness and the level of detail. But the blind evaluator would've valued the detail more than succinctness. While the result suggests sighted non-experts can generate good SDs through the process described in this paper, more research is needed to design a scalable process that can generate even better SDs, particularly in *descriptive* dimension. Future work should also involve blind people other than one of the authors to assess the SD quality and investigate factors like types of videos to SD quality.

ACKNOWLEDGMENTS

This research was supported by the Singapore Ministry of Education (MOE) Academic Research Fund (AcRF) Tier 1 Grant and the National Research Foundation, Singapore under its International Research Centres in Singapore Funding Initiative. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore. Hernisa Kacorri is partially supported by the National Institute on Disability, Independent Living, and Rehabilitation Research (NIDILRR), ACL, HHS (90REGE0008).

REFERENCES

[1] Amazon. 2020. Amazon Polly. https://aws.amazon.com/polly/. Accessed: 2020-06-01.

[2] Jeffrey P Bigham, Ryan S Kaminsky, Richard E Ladner, Oscar M Danielsson, and Gordon L Hempton. 2006. WebInSight: making web images accessible. In *Proceedings of the 8th international ACM SIGACCESS conference on Computers and accessibility*. 181–188.

[3] Sabine Braun. 2011. Creating coherence in audio description. *Meta: Journal des traducteurs/Meta: Translators' Journal* 56, 3 (2011), 645–662.

[4] John Brooke. 1996. SUS: a "quick and dirty'usability. *Usability evaluation in industry* (1996), 189.

[5] Described and Captioned Media Program. 2020. Described and Captioned Media Program (DCMP). http://www.descriptionkey.org/quality_description.html. Accessed: 2019-03-19.

[6] Louise Fryer. 2016. *An introduction to audio description: A practical guide*. Routledge.

[7] Jiawei Han, Micheline Kamber, and Jian Pei. 2011. Data mining concepts and techniques third edition. *The Morgan Kaufmann Series in Data Management Systems* (2011), 83–124.

[8] Masatomo Kobayashi, Kentarou Fukuda, Hironobu Takagi, and Chieko Asakawa. 2009. Providing synthesized audio description for online videos. In *Proceedings of the 11th international ACM SIGACCESS conference on Computers and accessibility*. 249–250.

[9] Masatomo Kobayashi, Trisha O'Connell, Bryan Gould, Hironobu Takagi, and Chieko Asakawa. 2010. Are synthesized video descriptions acceptable?. In *Proceedings of the 12th international ACM SIGACCESS conference on Computers and accessibility*. 163–170.

[10] Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International conference on machine learning*. 1188–1196.

[11] Hisashi Miyashita, Daisuke Sato, Hironobu Takagi, and Chieko Asakawa. 2007. Making multimedia content accessible for screen reader users. In *Proceedings of the 2007 international cross-disciplinary conference on Web accessibility (W4A)*. 126–127.

[12] Meredith Ringel Morris, Jazette Johnson, Cynthia L Bennett, and Edward Cutrell. 2018. Rich representations of visual content for screen reader users. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–11.

[13] John M Slatin. 2001. The art of ALT: toward a more accessible Web. *Computers and Composition* 18, 1 (2001), 73–81.

[14] Pantulkar Sravanthi and B Srinivasu. 2017. Semantic similarity between sentences. *International Research Journal of Engineering and Technology (IRJET)* 4, 1 (2017), 156–161.

[15] World Wide Web Consortium (W3C). 2019. Audio Description of Visual Information. https://www.w3.org/WAI/media/av/description/. Accessed:2020-06-30.

[16] World Wide Web Consortium (W3C). 2020. Color with Good Contrast. https://www.w3.org/WAI/perspective-videos/contrast/. Accessed: 20 February 2020.

[17] Agnieszka Walczak and Louise Fryer. 2018. Vocal delivery of audio description by genre: measuring users' presence. *Perspectives* 26, 1 (2018), 69–83.

[18] Beste F Yuksel, Soo Jung Kim, Seung Jung Jin, Joshua Junhee Lee, Pooyan Fazli, Umang Mathur, Vaishali Bisht, Ilmi Yoon, Yue-Ting Siu, and Joshua A Miele. 2020. Increasing Video Accessibility for Visually Impaired Users with Human-in-the-Loop Machine Learning. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–9.

[19] Radim Řehůřek. 2020. Doc2vec paragraph embeddings. https://radimrehurek.com/gensim/models/doc2vec.html. Accessed: 2020-06-30.